MULTI-OMICS DATA INTEGRATION TOWARDS BIOMARKERS FOR COLORECTAL ADENOMA-TO-CARCINOMA PROGRESSION

Małgorzata Anna Komór

ISBN: 978-94-6380-585-8

Cover: Małgorzata Komór inspired by Femme, Picasso
Layout and printing: Proefschriftmaken.nl

The printing of this thesis was financially supported by the Netherlands Cancer Institute.

This research was financially supported by Dutch Cancer Society.

Multi-omics data integration towards biomarkers for colorectal adenoma-to-carcinoma progression

Multi-omics data-integratie op weg naar biomarkers voor colorectaal adenoom tot carcinoom progressie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 12 december 2019 des middags te 12.45 uur

door

Małgorzata Anna Komór

geboren op 10 november 1990
te Warschau, Polen

Promotoren: Prof. dr. G.A. Meijer
Prof. dr. C.R. Jiménez

Copromotor: Dr. R.J.A. Fijneman

TABLE OF CONTENTS

Chapter 1 General introduction 9

Chapter 2 Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression 27

Chapter 3 Consensus molecular subtypes classification of colorectal adenomas 73

Chapter 4 Identification of differentially expressed splice variants by the proteogenomic pipeline Splicify 113

Chapter 5 Alternative splicing as a source of candidate biomarkers for early detection of colorectal cancer 139

Chapter 6 Proteins in stool as biomarkers for non-invasive detection of colorectal adenomas with high risk of progression 173

Chapter 7 Summary and Future Perspectives 203
Nederlandse Samenvatting 213

Appendum Curriculum vitae 219
List of publications 220
Acknowledgements 221

“In science, we must be interested in things, not in persons.”
Maria Skłodowska-Curie

Chapter 1

GENERAL INTRODUCTION

COLORECTAL CANCER

EPIDEMIOLOGY

Colorectal cancer (CRC) is the fourth most common cancer type worldwide with over 1 million new cases estimated in 2018[1]. It is the second most common cause of cancer-related deaths accounting for over 800 000 deaths annually[1]. Importantly, when CRC is detected early, the disease has a high cure rate. The 5-year survival rate of localized CRC is 90%, while it decreases for regional and distant disease to 71% and 14%, respectively[2].

PATHOLOGY

The normal colorectal mucosa is composed of the epithelium, lamina propria and muscularis mucosae[3]. The colonic epithelium consists of a single layer of cells that form a protective barrier between the host and the lumen of the colon. The epithelial cells are organized in the crypts, at the base of which stem cells divide and subsequently differentiate and migrate towards the top of the crypts, where they undergo apoptosis and eventually are shed into the lumen[4]. The differentiated epithelial cells are absorptive cells (which absorb water and nutrients) and secretory cells like goblet cells (which produce mucus that covers the mucosal surface) and endocrine cells. The lamina propria functions as a connective tissue scaffold between the epithelial crypts and the muscularis mucosae and consists of capillaries, myofibroblasts and immune cells, the majority of which being plasma cells. The muscularis mucosae is a layer of smooth muscle separating epithelium and lamina propria from submucosa[3].

Colorectal cancer arises from the normal epithelial cells due to DNA aberrations that cause altered cellular behavior, commonly referred to as the hallmarks of cancer[5, 6]. It is a lengthy process which includes formation of a benign, i.e. pre-invasive precursor lesion called adenoma[7]. Colorectal adenomas are defined as clearly delimited epithelial dysplasia. Once the tumor invades the submucosa is it considered a cancer[3]. In response to invasion of genomically foreign tumor cells, an inflammatory response occurs, referred to as desmoplasia, which involves stroma activation. At the same time, invading tumor cells change their shape, i.e. epithelial-mesenchymal transition. Even though it is well established that an adenoma may undergo malignant transformation and become a cancer, i.e. adenoma-to-carcinoma progression, it does not mean that all adenomas will progress to cancers. Actually, the prevalence of adenomas in the large intestine is much higher than the incidence of cancer[8, 9], implying that the majority of adenomas will never progress[10]. Based on the prevalence of focal cancer in endoscopically removed adenomas, it is estimated that only 5% of adenomas will eventually become cancers[11]. Adenoma features associated with presence of a focus of cancer are size, villous architecture and grade of dysplasia[11, 12]. Currently, adenomas larger than 1 cm and/or with a villous component and/or with high-grade dysplasia are referred to as “advanced adenomas” and are considered to be clinically relevant precursors of CRC. However, these features alone are not precise predictors of the malignant progression[13]. Identification of adenoma features that cause progression to malignancies is challenging due to the fact that adenomas detected during colonoscopy are completely removed and their natural history is interrupted.

DETECTION

Early stage colorectal cancer has a high cure rate, which is why early detection of the disease is crucial to reduce CRC mortality rates. Therefore, many countries have introduced population-wide screening programs to detect CRC at the early stages, being stool-based tests like guaiac-based fecal occult blood test or fecal immunochemical test (FIT), or direct visualization tests, like sigmoidoscopy or colonoscopy[14, 15]. FIT is the test used in the Netherlands, where people between 55 and 75 years of age take the test biennially. Once high levels of protein hemoglobin are detected by FIT, the participants are referred to follow-up colonoscopy to determine if they have cancer or precursor lesions[16]. In Poland the CRC screening program bases on colonoscopy, where individuals at the age of 55-64 years old are offered a single screening colonoscopy[17, 18]. The main advantages of FIT are non-invasiveness, low costs, high uptake and high sensitivity in identification of CRC (~79%) while among disadvantages are limited sensitivity in detection of advanced adenomas (~27%)[19, 20]. Colonoscopy is considered the gold standard in identification of adenomas and CRC; its main advantages are high diagnostic accuracy and feasibility of removal of the adenomas, while its drawbacks are high costs, low uptake and possible complications[19]. Another available methodology for CRC detection is molecular stool testing, where besides the levels of protein hemoglobin, other molecular markers are measured, like hypermethylated promoter CpG islands (NDRG4 and BMP3) and mutant KRAS in the Cologuard® test[21]. The Cologuard test has a higher sensitivity and a lower specificity in CRC detection, next to the higher costs when compared to FIT test[21]. Another approach currently in development is designing an antibody-based assay similar to FIT which would detect additional protein markers complementary to protein hemoglobin with the aim to improve FIT’s sensitivity without increasing the costs of the test significantly[22].

The variety of methodologies used in the national screening programs and no clear recommendation for one test[14, 15] show that there is still a clinical need for a better non-invasive test that will detect CRC and its relevant precursor lesions with the higher accuracy.

Currently, detection rate of advanced adenomas next to CRCs is used as an intermediate endpoint in CRC screening programs[15, 23], as removal of pre-malignant lesions during colonoscopy is an approach to decrease CRC incidence and mortality rates[24]. Given that not all of the advanced adenomas will progress to cancer and that advanced adenomas are very common in the large intestine in the elderly, the current strategy leads to overdiagnosis and overtreatment[10, 13, 25]. Incorporation of a more specific definition of an adenoma at an increased risk of progression to cancer could decrease the burden of overdiagnosis and overtreatment. Moreover, using advanced adenomas as intermediate endpoint yields great demand on the disease surveillance[26], even further increasing the burden of overtreatment and overdiagnosis[10]. Identification of biomarkers for adenomas at increased risk of progression to cancer may facilitate choosing better time points for CRC surveillance for individuals with adenomas.

MOLECULAR CHARACTERIZATION

GENOMIC INSTABILITY

As colorectal cancer is heterogeneous on the molecular level, it is often classified based on its global genomic (genomic instability) or epigenetic status (CpG island methylation phenotype)[19, 27]. Based on the genomic instability approximately 85% of CRCs are classified as chromosomal instable (CIN) and 15% as microsatellite instable (MSI)[28]. Chromosomal instability is characterized by acquisition of DNA copy number aberrations, when cancers gain or lose whole or large fractions of chromosomes. Microsatellite instability is characterized by deficiency in mismatch repair mechanisms which leads to acquisition of somatic mutations throughout the whole genome[28]. As for the adenomas, approximately 3% exhibit MSI[29].

In colorectal cancers and adenomas the DNA copy number aberrations obtained due to CIN exhibit a non-random pattern[30-34]. Seven chromosomal copy number aberrations have been identified as colorectal cancer-associated events (CAEs); gains of chromosomal arms 8q, 13q and 20q and losses of chromosomal arms 8p, 15q, 17p and 18q[30, 35]. Gain of chromosome arm 20q and loss of chromosomal arm 18q are the most frequent DNA copy number aberrations occurring in 67% and 49% of the CRC cases, respectively[36]. With the accuracy of 78%, the presence of at least two of these CAEs enabled distinction of an adenoma with a focus of cancer from a non-malignant adenoma[30]. Therefore, adenomas with at least two out of the seven CAEs are marked as high risk of progressing to malignancy, further referred to as high-risk adenomas[30]. The molecularly-defined “high-risk adenoma” definition is independent of the morphology-defined “advanced adenoma” and only 23-36% of advanced adenomas classify as high-risk adenomas based on their DNA copy number profile[35]. Studies on the non-random DNA copy number aberrations in colorectal adenomas and cancers lead to identification of potential CRC driver events located in the amplified regions, which play a major role in adenoma-to-carcinoma progression[30, 37-41]. Functional studies of candidate oncogenes from the 20q region indicated that AURKA and TPX2 promote 20q amplicon-driven adenoma-to-carcinoma progression[37]. This means that the non-random DNA copy number aberrations in fact influence biological processes within cells, through which they facilitate colorectal tumorigenesis.

SOMATIC MUTATIONS AND DISRUPTION OF SIGNALING PATHWAYS

The majority of CRCs (~81%) acquire truncating mutation in APC gene, typically already at the transition from normal epithelium to colorectal adenoma[7, 27]. The APC mutation leads to increased activity of Wnt signaling pathway and consequently increase in the proliferation rates[42]. However, activation of Wnt signaling has been observed in over 90% of CRCs, as it is not solemnly dependent on loss-of-function of APC, but can also be caused by other aberrations, e.g. an activating mutation of CTNNB1, a downstream component of the Wnt signaling pathway[27]. The second most common somatic mutation in CRC is a loss-of-function mutation of the well-known tumor suppressor gene TP53 (60%[27]), which is often due to the loss of chromosomal arm 17p and which has been associated with the transition from adenoma to cancer[7, 42]. Other frequent somatic mutations playing a role in colorectal carcinogenesis concern RAS-MAPK, PI3K and TGFβ signaling pathways and include KRAS (43%), BRAF (~10%), NRAS (9%), PIK3CA (18%) and SMAD4 (10%)[27, 42]. The low mutation frequencies in CRC display the heterogeneity of this disease and show that single mutated genes cannot be used as biomarkers for CRC.

CONSENSUS MOLECULAR SUBTYPES OF COLORECTAL CANCER

Next to the genetic classification, stratification of CRC patients based on their mRNA expression profiles has been pursued in multiple studies with the aim to facilitate clinical translation and improve precision medicine[43-49]. The multiple existing classifications were later combined into the consensus molecular subtypes (CMS) of CRC[50]. A consensus RNA expression-based classifier was provided that classifies CRCs into four CMS groups (CMS1-4). CMS1 consists of approximately 14% of CRCs and is associated with MSI, BRAF mutation, DNA promoter hypermethylation and immune infiltration. CIN is a feature characteristic of CMS2-4 classes. CMS2 is the most common CRC subtype (37%) and follows the canonical CRC carcinogenesis, including activation of Wnt and Myc pathways. Approximately 13% of CRCs represent the CMS3 subtype, which is characterized by dysregulated metabolism and KRAS mutation. Finally, the CMS4 subtype (23%) is described as a mesenchymal, stroma-rich group that is associated with poor prognosis[50]. The CMS classification was established to reconcile the differences between the existing classification algorithms and was widely approved by the scientific community[50]. Nevertheless, due to its limited stability, the CMS algorithm did not succeed to remain the final classifier representing the CRC heterogeneity and was followed by other methods[51-53].

COLORECTAL CANCER PROTEOMICS AND PROTEOGENOMICS

Cancer is caused by molecular alterations in DNA, thereby affecting the transcriptome and subsequently the proteome of the cancer cell. Proteomics is widely used to identify candidate biomarkers that reflect molecular aberrations characteristic for CRC[54]. Proteins are a promising source of biomarkers as they are responsible for processes occurring in the cancer cells and because protein detection using e.g. an antibody-based assay can be implemented in the clinical practice. And so, a number of protein biomarkers were identified using proteomic characterization of CRC cell lines, tissues or human body-fluids[54]. Cell surface proteomics lead to identification of SLC2A1 and PRNP as candidate biomarkers for detection of CRC and high-risk adenomas through molecular imaging[55]. Tissue secretome proteomics revealed MCM5, TIMP1 and LCN2 together with others as biomarkers for CRC screening in stool or blood[56, 57]. And stool proteomics identified multiple combinations of four proteins that can outperform hemoglobin in detection of CRCs and advanced adenomas[22]. Currently, follow-up validation studies are being performed to translate these results into clinical applications.

As proteins play a functional role in the cell, next to the biomarker studies global proteome profiling holds promise to provide additional insights into CRC biology. Indeed, proteome profiling was shown to outperform transcriptome profiling in gene function prediction based on co-expression network analysis[58]. Currently, proteogenomics, i.e. comprehensive integration of genomics and proteomics data, is used to improve our knowledge on CRC biology and identify potential oncogenes and tumor suppressors. Until recently, combining DNA and RNA data to study DNA copy-number driven gene-dosage effect and identify potential tumor drivers has been performed in CRC[39, 59]. However, only for a limited number of candidate drivers functional assays confirmed their oncogenic potential[37, 40, 60]. Therefore, in such studies addition of the protein layer was introduced, by e.g. The Cancer Genome Atlas and Clinical Proteomic Tumor Analysis Consortium, as it provides information about which chromosomal aberrations lead to functional consequences[61]. Gene-dosage effects of HNF4A, SRC and TOMM34, located on frequently gained chromosome arm 20q, were observed also on protein level indicating their driver role in CRC[61]. Additionally, in colon cancer focal deletion on chromosome 18q was observed to cause a decrease in the protein expression of a well-known tumor suppressor SMAD4. Additionally, RB1 was identified as a driver and a potential therapeutic target through integration of DNA copy number, protein expression and phosphoproteomics data[62]. Proteome profiling can also complement mutation analysis, as it identified SOX9 as an oncogene in colon cancer discordant with the somatic mutation inaccurately classifying this gene as a tumor supressor[62]. In conclusion, proteogenomics is a powerful approach to identify genomic aberrations that lead to functional consequences[63].

ALTERNATIVE SPLICING IN CANCER

The human transcriptome is far more complex than the protein-coding genome as approximately 95% of multi-exon transcripts undergo alternative splicing[64]. As a consequence, a single gene can be transcribed into a variety of isoforms which, when translated into proteins, may differ in structure, location, and function. Recently, to identify protein features associated with alternative splicing, Exon ontology was established[65] by mapping protein features derived from existing ontologies and databases back to genomic exons. In this way, skipping of a certain exon can be associated with losing protein domain carrying catalytic, binding, receptor or transporter activity or protein region containing subcellular localization signal, post-translational modifications or structural features[65]. Functional impact of alternative splicing was also analyzed in cancer, which revealed that alternative splicing often affects protein domains from families frequently mutated in cancer and seem to be mutually exclusive with mutations in cancer drivers, indicating that a number of isoforms carry driver-like properties and facilitate carcinogenesis[66]. Indeed, abnormally spliced RNA plays a role in tumor progression and metastasis, and has been shown to affect each of the biological processes commonly referred to as the hallmarks of cancer[67]. For instance, incorporation of an alternative 5' splice site of BCL2L1 causes a switch from a pro- to an anti-apoptotic isoform in cancer[68]. Usage of an alternative 3' splice site of VEGFA leads to a shift from an anti- to a pro-angiogenic isoform in cancer[69]. Splicing factors, i.e. proteins which play a direct role in splicing regulation and isoform expression, can develop oncogenic activity, e.g. due to aberrant expression or somatic mutations[67]. For instance, SF3B1 is one of the most commonly mutated splicing factors in cancer[70]. Recurrent mutations affecting this gene have been found in leukemia, melanoma and in pancreatic, breast and bladder cancer. In chronic lymphocytic leukemia, mutations in this splicing factor contribute to tumor progression, poor patient survival and poor chemotherapy response[71, 72]. Overexpression of another splicing factor, SRSF1, was observed in many tumor types including breast[73], colon, thyroid, small intestine, kidney, lung, liver and pancreas[74] and was proven to lead to oncogenic activity[67, 75-77]. Transcription of SRSF1 is directly regulated by MYC, a well-known oncogenic transcription factor. Through activation of SRSF1, MYC can affect alternative splicing of a subset of SRSF1 target genes and contribute to carcinogenesis[78]. In colorectal cancer, SRSF1 causes inclusion of exon 4 in RAC1, generating a Rac1b isoform that contributes to CRC cell survival[79, 80].

As aberrant splicing accompanies tumor progression, splice variants may provide a promising source of potential biomarkers for CRC.

ALTERNATIVE SPLICING ANALYSIS

Alternative splicing occurs in the RNA, and therefore global alternative splicing discovery is mostly performed using RNA sequencing (RNA-seq), often with Illumina technology. There are two main approaches to identify alternative splicing using the RNA-seq data; transcriptome assembly and identification of individual splicing events[81].

Transcriptome assembly can be reference-based, i.e. with the use of a reference genome, or de novo, which is often performed for unannotated species[82]. The reference-based approach consist of the following steps; splice-aware RNA read alignment, which allows mapping substrings of RNA reads through multiple loci, building a graph from RNA reads that represents all possible isoforms, where nodes are reads and edges are formed if there is an overlapping substring between two reads, and finally traversing this graph to assemble the transcripts. Reconstruction of full transcriptomes is difficult as it bases on assembling together very short RNA reads and therefore requires a lot of assumptions to complete the task. Moreover, in RNA-seq data read depth varies tremendously depending on the expression of the genes; hence, assembly of the lowly expressed transcript is very challenging. Additionally, multiple transcripts of the same gene share exons, which makes it difficult to resolve the transcripts unambiguously and assign RNA reads to the right isoform and quantitative analysis of the transcript expression is challenging. As transcriptome assembly requires a number of assumptions, there are multiple algorithms available which often result in very limited overlap of the identified transcriptomes or very limited number of isoforms observed on protein level[82-84]. Currently, instead of attempting to assemble full transcripts from short RNA sequencing reads, it is possible to perform full-length transcript identification using long read sequencing, like Oxford Nanopore[85] or PacBio Iso-Seq[86]. Long read sequencing improves the quality of the identified transcripts, but also entails higher costs than standard short read sequencing.

An alternative to the transcriptome assembly is identification of individual splicing events, which is focused on the spliced region only and does not attempt to reconstruct whole isoforms. The alternatively splicing events are skipped exon, mutually exclusive exons, alternative 3’ or 5’ splice sites and retained introns. The alternatively spliced events are often analyzed in a differential setting when differences between two conditions are quantified based on the exon-exon or exon-intron junction reads and are independent of gene expression. Inclusion level (Ψ – “percent spliced in”), i.e. number of inclusion specific reads divided by the total number of inclusion- and exclusion-specific reads in the region, is then compared between conditions and reported when the difference is statistically significant[81]. Methods for differential splicing analysis report even up to 85% of events being validated with RT-qPCR[87], while their main drawback is lack of information on the whole transcript.

IDENTIFICATION OF PROTEIN ISOFORMS

RNA sequencing allows studying the complexity of transcriptomes. There is a lot of evidence for alternative splicing on the RNA level; however, for many of the isoforms it is not known whether they are translated into functional proteins. This knowledge is important to understand the biological consequences of alternative splicing. Additionally, identification of splice variants translated into proteins provides novel candidate biomarkers, as protein isoforms have significant potential as biomarkers to increase the accuracy of diagnosis, prognosis or therapy prediction of the disease[88]. Protein isoforms can be studied on the proteome level with the use of in-depth tandem mass spectrometry (LC-MS/MS).

In the discovery proteomics experiment proteins are first fragmented to peptides using enzymatic digestion, most often by trypsin, and MS/MS spectra of peptides are measured. Then, peptide identification is performed by matching observed spectra to theoretical spectra established in sillico based on a reference protein sequence database[89]. This approach has been proven suitable for high-throughput protein identification[89]. Nevertheless, in this approach protein discovery is limited to the proteins in the database while still a number of high quality measured spectra are not identified. Proteogenomics improves this methodology by enriching the database with novel discoveries driven by genomics. These comprise alternative splice variants but also DNA-driven structural rearrangements or single nucleotide variants[90].

Proteogenomics is a field that can be used to identify potential tumor drivers, protein isoforms and novel protein biomarkers.

BIOINFORMATICS AND DATA ANALYSIS IN TRANSLATIONAL RESEARCH

The aim of translational research is to move basic biological discoveries into the patient-care setting. High throughput molecular profiling experiments like next generations sequencing or tandem mass spectrometry allow to measure vast amounts of molecules, like DNA, RNA or protein, in a short timeframe[91, 92]. Currently generation of large patient datasets from high throughput profiling experiments becomes a routine task, while interpretation of the data is much more complex. The aim of bioinformatics is to extract insights from these data using novel or established algorithms, statistical analysis as well as domain knowledge[93]. The field has developed new and complex ways to efficiently and effectively mine the data, which includes best practices for RNA-seq or proteogenomic data analysis[90, 94]. Nevertheless, a crucial step in evaluation of the bioinformatics discoveries is experimental validation of the results before considering them for translational medicine[93]. And so, quantitative and qualitative differences on RNA level can be validated with e.g. RT-qPCR, while validation of proteomics experiments often bases on antibodies and is performed with the use of Western blotting, immunohistochemistry or ELISA-like assays. This requires collaboration and good communication within the multidisciplinary teams of bioinformaticians, molecular biologists and clinicians.

THE AIM AND OUTLINE OF THIS THESIS

Early detection of colorectal cancer (CRC) is crucial to reduce CRC mortality rates. The fecal immunochemical test (FIT) is a non-invasive CRC screening test that detects human protein hemoglobin in stool. Although FIT is beneficial in its current form, its performance is still suboptimal and needs to be further improved. The aim of this thesis was to discover novel protein biomarkers to improve early detection of colorectal cancer.

As CRC is considered curable at early stages, it may be beneficial to detect its precursor lesions (adenomas)[95]. However, as only 5% of adenomas will eventually progress to cancer[11], detection of all adenomas in CRC screening would lead to overdiagnosis and overtreatment. There is a need for a better definition of adenomas at increased risk of progressing to cancer, which should be considered a better target for colorectal cancer screening. As "advanced adenoma" definition is still not specific enough[13], in this thesis we consider high-risk adenomas as clinically relevant precursors of CRC. In chapter 2 we set out to characterize high-risk adenomas in comparison to low-risk adenomas by in depth molecular profiling to identify gene and protein expression differences between them and to identify putative drivers of adenoma-to-carcinoma progression. We examined if high-risk adenomas resemble cancers in terms of biological processes when compared to low-risk adenomas, lending support to the assumption that high-risk adenomas are the more relevant precursors of CRC. In chapter 3, we examined if colorectal adenomas can be classified into CRC subtypes according to the CMS classification, to evaluate if adenomas carry the molecular signature of their future subtype and whether specific CMS classes are related to the presence of specific DNA copy number aberrations associated with risk of progression to malignancy.

Previously, functional studies on cancer-associated events lead to identification of AURKA as one of the putative oncogenes promoting 20q amplicon-driven adenoma-to-carcinoma progression[37]. It has also been shown that AURKA has indirect impact on pro- and anti-apoptotic alternative splicing, a common aberrant splicing event accompanying tumor progression[68, 77]. Due to this link between non-random copy number changes and cancer-associated alternative splicing, we set out to investigate alternative splicing as potential source of tumor-specific biomarkers for early detection of CRC and high-risk adenomas. As the current CRC screening method bases on detection of protein hemoglobin, we aimed to investigate proteins translated from alternatively spliced RNA so that in the future such a protein biomarker could be implemented in the similar way as FIT. In chapter 4, we developed a computational proteogenomic pipeline, Splicify, for identification of splice variants that are differential between two conditions and that are translated to protein isoforms. We applied Splicify to RNA sequencing and mass spectrometry data obtained from colorectal cancer cell line SW480, before and after siRNA-mediated down-modulation of the splicing machinery, in particular splicing factors SF3B1 and SRSF1, to present the utility of the method in a controlled setting. In chapter 5, we applied Splicify to colorectal tissue data obtained from colorectal cancers, adenomas and normal colon samples, to identify differential protein isoforms that may serve as potential biomarkers for CRCs and clinically relevant adenomas.

As we aim to improve of the non-invasive screening methodology, it is crucial to evaluate if the individuals with clinically relevant lesions, i.e. high-risk adenomas and CRCs, can be identified based on the molecular composition of their stool. It has been previously shown that several protein combinations outperform hemoglobin in discriminating CRC from control samples[22]. In chapter 6 of this thesis, we examined if individuals with high-risk adenomas can be distinguished from controls based on abundance of proteins identified in their stool samples.

Finally, summary and general discussion of the results described in this thesis are presented in chapter 7.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin 2017;67:177-193.
3. Ponz de Leon M, Di Gregorio C. Pathology of colorectal cancer. Dig Liver Dis 2001;33:372-88.
4. Clevers H. The intestinal crypt, a prototype stem cell compartment. Cell 2013;154:274-84.
5. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57-70.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646-74.
7. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759-67.
8. Lieberman DA, Weiss DG, Bond JH, et al. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs Cooperative Study Group 380. N Engl J Med 2000;343:162-8.
9. Imperiale TF, Wagner DR, Lin CY, et al. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. N Engl J Med 2000;343:169-74.
10. Kalager M, Wieszczy P, Lansdorp-Vogelaar I, et al. Overdiagnosis in Colorectal Cancer Screening: Time to Acknowledge a Blind Spot. Gastroenterology 2018;155:592-595.
11. Shinya H, Wolff WI. Morphology, anatomic distribution and cancer potential of colonic polyps. Ann Surg 1979;190:679-83.
12. Muto T, Bussey HJ, Morson BC. The evolution of cancer of the colon and rectum. Cancer 1975;36:2251-70.
13. Brenner H, Hoffmeister M, Stegmaier C, et al. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. Gut 2007;56:1585-9.
14. Vleugels JL, van Lanschot MC, Dekker E. Colorectal cancer screening by colonoscopy: putting it into perspective. Dig Endosc 2016;28:250-9.
15. U. S. Preventive Services Task Force, Bibbins-Domingo K, Grossman DC, et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2016;315:2564-2575.
16. The National Institute for Health and Environment TN. Colon Cancer Screening. [Cited 04 January 2019] https://www.rivm.nl/bevolkingsonderzoek-darmkanker
17. The Ministery of Health P. Colorectal cancer prevention. [Cited 04 January 2019] https://www.gov.pl/web/zdrowie/profilaktyka-raka-jelita-grubego1 2016.
18. Kaminski MF, Kraszewska E, Rupinski M, et al. Design of the Polish Colonoscopy Screening Program: a randomized health services study. Endoscopy 2015;47:1144-50.
19. Kuipers EJ, Grady WM, Lieberman D, et al. Colorectal cancer. Nat Rev Dis Primers 2015;1:15065.
20. Lee JK, Liles EG, Bent S, et al. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Ann Intern Med 2014;160:171.

21. Imperiale TF, Ransohoff DF, Itzkowitz SH. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med 2014;371:187-8.
22. Bosch LJW, de Wit M, Pham TV, et al. Novel Stool-Based Protein Biomarkers for Improved Colorectal Cancer Screening: A Case-Control Study. Ann Intern Med 2017;167:855-866.
23. Robertson DJ, Lee JK, Boland CR, et al. Recommendations on Fecal Immunochemical Testing to Screen for Colorectal Neoplasia: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2017;152:1217-1237.e3.
24. Pan J, Xin L, Ma YF, et al. Colonoscopy Reduces Colorectal Cancer Incidence and Mortality in Patients With Non-Malignant Findings: A Meta-Analysis. Am J Gastroenterol 2016;111:355-65.
25. Kaminski MF, Regula J. Adenoma Detection Race at Colonoscopy: The Good and the Bad. Gastroenterology 2015;149:273-4.
26. Hassan C, Quintero E, Dumonceau JM, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2013;45:842-51.
27. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-337.
28. Rajagopalan H, Nowak MA, Vogelstein B, et al. The significance of unstable chromosomes in colorectal cancer. Nat Rev Cancer 2003;3:695-701.
29. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell 1996;87:159-70.
30. Hermsen M, Postma C, Baak J, et al. Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. Gastroenterology 2002;123:1109-19.
31. Meijer GA, Hermsen MA, Baak JP, et al. Progression from colorectal adenoma to carcinoma is associated with non-random chromosomal gains as detected by comparative genomic hybridisation. J Clin Pathol 1998;51:901-9.
32. Douglas EJ, Fiegler H, Rowan A, et al. Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. Cancer Res 2004;64:4817-25.
33. Camps J, Grade M, Nguyen QT, et al. Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. Cancer Res 2008;68:1284-95.
34. Hirsch D, Camps J, Varma S, et al. A new whole genome amplification method for studying clonal evolution patterns in malignant colorectal polyps. Genes Chromosomes Cancer 2012;51:490-500.
35. Carvalho B, Diosdado B, Terhaar Sive Droste JS, et al. Evaluation of Cancer-Associated DNA Copy Number Events in Colorectal (Advanced) Adenomas. Cancer Prev Res (Phila) 2018;11:403-412.
36. De Angelis PM, Clausen OP, Schjolberg A, et al. Chromosomal gains and losses in primary colorectal carcinomas detected by CGH and their associations with tumour DNA ploidy, genotypes and phenotypes. Br J Cancer 1999;80:526-35.
37. Sillars-Hardebol AH, Carvalho B, Tijssen M, et al. TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. Gut 2012;61:1568-75.
38. Ried T, Knutzen R, Steinbeck R, et al. Comparative genomic hybridization reveals a

specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. Genes Chromosomes Cancer 1996;15:234-45.

39. Carvalho B, Postma C, Mongera S, et al. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. Gut 2009;58:79-89.
40. de Groen FL, Krijgsman O, Tijssen M, et al. Gene-dosage dependent overexpression at the 13q amplicon identifies DIS3 as candidate oncogene in colorectal cancer progression. Genes Chromosomes Cancer 2014;53:339-48.
41. Diosdado B, van de Wiel MA, Terhaar Sive Droste JS, et al. MiR-17-92 cluster is associated with 13q gain and c-myc expression during colorectal adenoma to adenocarcinoma progression. Br J Cancer 2009;101:707-14.
42. van Lanschot MCJ, Bosch LJW, de Wit M, et al. Early detection: the impact of genomics. Virchows Arch 2017;471:165-173.
43. Budinska E, Popovici V, Tejpar S, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol 2013;231:63-76.
44. De Sousa EMF, Wang X, Jansen M, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med 2013;19:614-8.
45. Marisa L, de Reynies A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013;10:e1001453.
46. Perez-Villamil B, Romera-Lopez A, Hernandez-Prieto S, et al. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. BMC Cancer 2012;12:260.
47. Roepman P, Schlicker A, Tabernero J, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int J Cancer 2014;134:552-62.
48. Sadanandam A, Lyssiotis CA, Homicsko K, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med 2013;19:619-25.
49. Schlicker A, Beran G, Chresta CM, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med Genomics 2012;5:66.
50. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine 2015;21:1350-6.
51. Isella C, Terrasi A, Bellomo SE, et al. Stromal contribution to the colorectal cancer transcriptome. Nat Genet 2015;47:312-319.
52. Calon A, Lonardo E, Berenguer-Llergo A, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat Genet 2015;47:320-329.
53. Alderdice M, Richman SD, Gollins S, et al. Prospective patient stratification into robust cancer-cell intrinsic subtypes from colorectal cancer biopsies. J Pathol 2018;245:19-28.
54. de Wit M, Fijneman RJ, Verheul HM, et al. Proteomics in colorectal cancer translational research: biomarker discovery for clinical applications. Clin Biochem 2013;46:466-79.
55. de Wit M, Jimenez CR, Carvalho B, et al. Cell surface proteomics identifies glucose transporter type 1 and prion protein as candidate biomarkers for colorectal

adenoma-to-carcinoma progression. Gut 2012;61:855-64.
56. de Wit M, Kant H, Piersma SR, et al. Colorectal cancer candidate biomarkers identified by tissue secretome proteome profiling. J Proteomics 2014;99:26-39.
57. Fijneman RJ, de Wit M, Pourghiasian M, et al. Proximal fluid proteome profiling of mouse colon tumors reveals biomarkers for early diagnosis of human colorectal cancer. Clin Cancer Res 2012;18:2613-24.
58. Wang J, Ma Z, Carr SA, et al. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. Mol Cell Proteomics 2017;16:121-134.
59. Haan JC, Labots M, Rausch C, et al. Genomic landscape of metastatic colorectal cancer. Nat Commun 2014;5:5457.
60. Camps J, Pitt JJ, Emons G, et al. Genetic amplification of the NOTCH modulator LNX2 upregulates the WNT/beta-catenin pathway in colorectal cancer. Cancer Res 2013;73:2003-13.
61. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513:382-7.
62. Vasaikar S, Huang C, Wang X, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell 2019.
63. Ruggles KV, Krug K, Wang X, et al. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics 2017.
64. Pan Q, Shai O, Lee LJ, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413-5.
65. Tranchevent LC, Aube F, Dulaurier L, et al. Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res 2017;27:1087-1097.
66. Climente-Gonzalez H, Porta-Pardo E, Godzik A, et al. The Functional Impact of Alternative Splicing in Cancer. Cell Rep 2017;20:2215-2226.
67. Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene 2014;33:5311-8.
68. Boise LH, Gonzalez-Garcia M, Postema CE, et al. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 1993;74:597-608.
69. Ladomery MR, Harper SJ, Bates DO. Alternative splicing in angiogenesis: the vascular endothelial growth factor paradigm. Cancer Lett 2007;249:133-42.
70. Sveen A, Kilpinen S, Ruusulehto A, et al. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene 2016;35:2413-27.
71. Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet 2011;44:47-52.
72. Oscier DG, Rose-Zerilli MJ, Winkelmann N, et al. The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. Blood 2013;121:468-75.
73. Anczuków O, Akerman M, Cléry A, et al. SRSF1-Regulated Alternative Splicing in Breast Cancer. Mol Cell 2015;60:105-17.
74. Karni R, de Stanchina E, Lowe SW, et al. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol 2007;14:185-93.
75. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev 2010;24:2343-64.
76. Ladomery M. Aberrant alternative splicing is another hallmark of cancer. Int J Cell Biol 2013;2013:463786.

77. Moore MJ, Wang Q, Kennedy CJ, et al. An Alternative Splicing Network Links Cell Cycle Control to Apoptosis. Cell 2010;142:625-36.
78. Das S, Anczukow O, Akerman M, et al. Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. Cell Rep 2012;1:110-7.
79. Goncalves V, Henriques AF, Pereira JF, et al. Phosphorylation of SRSF1 by SRPK1 regulates alternative splicing of tumor-related Rac1b in colorectal cells. Rna 2014;20:474-82.
80. Matos P, Jordan P. Increased Rac1b expression sustains colorectal tumor cell survival. Mol Cancer Res 2008;6:1178-84.
81. Dvinge H, Kim E, Abdel-Wahab O, et al. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer 2016;16:413-30.
82. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet 2011;12:671-82.
83. Steijger T, Abril JF, Engstrom PG, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 2013;10:1177-84.
84. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature 2014;509:575-81.
85. Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 2018;15:201-206.
86. Gordon SP, Tseng E, Salamov A, et al. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One 2015;10:e0132628.
87. Shen S, Park JW, Lu ZX, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A 2014;111:E5593-601.
88. Mischak H, Apweiler R, Banks RE, et al. Clinical proteomics: A need to define the field and to begin to set adequate standards. Proteomics Clin Appl 2007;1:148-56.
89. Duncan MW, Aebersold R, Caprioli RM. The pros and cons of peptide-centric proteomics. Nat Biotechnol 2010;28:659-64.
90. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11:1114-25.
91. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333-51.
92. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature 2016;537:347-55.
93. Londin ER, Barash CI. What is translational bioinformatics? Appl Transl Genom 2015;6:1-2.
94. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13.
95. Haug U, Knudsen AB, Lansdorp-Vogelaar I, et al. Development of new non-invasive tests for colorectal cancer screening: the relevance of information on adenoma detection. Int J Cancer 2015;136:2864-74.

Chapter 2

Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression

Malgorzata A Komor, Meike de Wit, Jose van den Berg, Sanne R Martens de Kemp, Pien M Delis-van Diemen, Anne S Bolijn, Marianne Tijssen, Tim Schelfhorst, Sander R Piersma, Davide Chiasserini, Joyce Sanders, Christian Rausch, Youri Hoogstrate, Andrew P Stubbs, Mark de Jong, Guido Jenster, Beatriz Carvalho, Gerrit A Meijer, Connie R Jimenez, Remond JA Fijneman

In collaboration with the NGS-ProToCol consortium

INTERNATIONAL JOURNAL OF CANCER 2019;EPUB AHEAD OF PRINT

Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression

Malgorzata A. Komor[1,2], Meike de Wit[1], Jose van den Berg[1], Sanne R. Martens de Kemp[1,2], Pien M. Delis-van Diemen[1], Anne S. Bolijn[1], Marianne Tijssen[1], Tim Schelfhorst[2], Sander R. Piersma[2], Davide Chiasserini[2], Joyce Sanders[1], Christian Rausch[1], Youri Hoogstrate[3], Andrew P. Stubbs[4], Mark de Jong[5], Guido Jenster[3], Beatriz Carvalho[1], Gerrit A. Meijer[1], Connie R. Jimenez[2], and Remond J.A. Fijneman[1], In collaboration with the NGS-ProToCol Consortium[6]

[1]Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[2]Oncoproteomics Laboratory, Amsterdam UMC, Vrije Universiteit Amsterdam, Medical Oncology, Amsterdam, The Netherlands
[3]Department of Urology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands
[4]Department of Bioinformatics, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands
[5]GenomeScan, Leiden, The Netherlands
[6]See Appendix for consortium members

Removal of colorectal adenomas is an effective strategy to reduce colorectal cancer (CRC) mortality rates. However, as only a minority of adenomas progress to cancer, such strategies may lead to overtreatment. The present study aimed to characterize adenomas by in-depth molecular profiling, to obtain insights into altered biology associated with the colorectal adenoma-to-carcinoma progression. We obtained low-coverage whole genome sequencing, RNA sequencing and tandem mass spectrometry data for 30 CRCs, 30 adenomas and 18 normal adjacent colon samples. These data were used for DNA copy number aberrations profiling, differential expression, gene set enrichment and gene-dosage effect analysis. Protein expression was independently validated by immunohistochemistry on tissue microarrays and in patient-derived colorectal adenoma organoids. Stroma percentage was determined by digital image analysis of tissue sections. Twenty-four out of 30 adenomas could be unambiguously classified as high risk ($n = 9$) or low risk ($n = 15$) of progressing to cancer, based on DNA copy number profiles. Biological processes more prevalent in high-risk than low-risk adenomas were related to proliferation, tumor microenvironment and Notch, Wnt, PI3K/AKT/mTOR and Hedgehog signaling, while metabolic processes and protein secretion were enriched in low-risk adenomas. DNA copy number driven gene-dosage effect in high-risk adenomas and cancers was observed for *POFUT1*, *RPRD1B* and *EIF6*. Increased POFUT1 expression in high-risk adenomas was validated in tissue samples and organoids. High POFUT1 expression was also associated with Notch signaling enrichment and with decreased goblet cells differentiation. In-depth molecular characterization of colorectal adenomas revealed *POFUT1* and Notch signaling as potential drivers of tumor progression.

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

Key words: colorectal adenoma, adenoma-to-carcinoma progression, POFUT1

Abbreviations: CAE: cancer-associated event; CRC: colorectal cancer; EIF6: eukaryotic translation initiation factor 6; GSEA: gene set enrichment analysis; HRA: high-risk adenoma; IHC: immunohistochemistry; LC–MS/MS: liquid chromatography tandem mass spectrometry; LRA: low-risk adenoma; MSI: microsatellite instability/instable; MSS: microsatellite-stable; POFUT1: protein O-fucosyltransferase 1; RNA-seq: RNA sequencing; RPRD1B: regulation of nuclear pre-mRNA domain containing 1B; TMA: tissue microarray; WGS: whole genome sequencing

Conflict of interest: G.A.M. has research collaborations with Exact Sciences and Sysmex for other studies regarding early detection of colorectal cancer. The companies provide materials, equipment or (sample) analyses. A patent application has been submitted based on results of this study, with M.A.K., B.C., G.A.M., C.R.J. and R.J.A.F. as inventors.

Grant sponsor: Cancer Center Amsterdam; **Grant sponsor:** Center for Translational Molecular Medicine; **Grant number:** NGS-ProToCol/03O-402; **Grant sponsor:** KWF Kankerbestrijding; **Grant number:** 2013-6025; **Grant sponsor:** SU2C-DCS International Translational Cancer Research Dream Team Grant; **Grant numbers:** SU2C-AACR-DT1415, MEDOCC; **Grant sponsor:** ZonMw; **Grant number:** 91116017

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/ijc.32627

History: Received 12 Jun 2019; Accepted 11 Jul 2019; Online 14 Aug 2019

Correspondence to: Remond J.A. Fijneman, The Netherlands Cancer Institute, Department of Pathology, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, Tel.: +31-20-5121725, E-mail: r.fijneman@nki.nl

What's new?

Removal of colorectal adenomas is an effective strategy to reduce colorectal cancer (CRC) mortality rates. However, as only a minority of adenomas progress to cancer, such strategies may lead to overtreatment. While high-risk adenomas, defined by specific DNA copy number aberrations, have an increased risk of progression, the mechanisms underlying colorectal adenoma-to-carcinoma progression remain unclear. This molecular characterization of colorectal adenomas, CRCs, and normal adjacent colon samples demonstrates that biological processes inherent to CRC are already more active in high-risk adenomas compared to low-risk adenomas. Moreover, the findings highlight POFUT1 and Notch signaling as potential drivers of colorectal tumor development.

Introduction

Colorectal adenomas are benign precursor lesions of colorectal cancer (CRC) that arise from normal epithelium.[1] The prevalence of adenomas in the large intestine is much higher than the incidence of cancer,[2,3] implying that the majority of adenomas will never progress to CRC.[4] In clinical practice, adenomas detected during colonoscopy are completely removed, and consequently the natural history of disease is disrupted. Based on the prevalence of focal cancer in endoscopically removed adenomas, it is estimated that only 5% of adenomas will eventually progress to CRC.[5,6] Currently, adenomas larger than 1 cm and/or with a villous component and/or with high-grade dysplasia are referred to as "advanced adenomas" and are considered to be clinically relevant precursors of CRC. However, incidence studies of both advanced adenomas and CRCs suggest that these features alone are not precise predictors of the malignant progression.[7]

Cancer is caused by molecular alterations in DNA, thereby affecting gene expression at RNA and protein level. The "advanced adenoma" definition neglects molecular changes that accompany adenoma-to-carcinoma progression. In multiple cancer types, the progression of dysplastic epithelial premalignant lesions, like colorectal adenomas, has been associated with acquisition of genomic instability.[8,9] This often concerns chromosomal instability, which affects about 85% of CRCs.[10] Studies on chromosomal instability in colorectal adenomas and cancers led to identification of nonrandom chromosomal aberrations and potential CRC driver events, which play a major role in adenoma-to-carcinoma progression.[11–18] Seven chromosomal copy number aberrations have been identified as colorectal cancer-associated events (CAEs); gains of chromosomal arms 8q, 13q and 20q and losses of chromosomal arms 8p, 15q, 17p and 18q. With the accuracy of 78%, the presence of at least two of these CAEs enabled distinction of an adenoma with a focus of cancer from a nonmalignant adenoma.[11] Therefore, adenomas with at least two out of the seven CAEs are marked as high risk of progressing to malignancy, further referred to as high-risk adenomas (HRAs).[11] We recently observed that only 23–36% of advanced adenomas classify as HRAs based on their DNA copy number profile.[19]

The aim of the present study was to characterize adenomas at low and high risk of progressing to cancer by molecular profiling at DNA, RNA and protein level, allowing to examine the biological processes in which these adenomas differ and to discover putative drivers of early colorectal tumor development.

Materials and Methods

Tissue data

Fresh frozen tissue material from 30 CRCs, 30 adenomas and 18 normal colorectal mucosa samples was collected at the Department of Pathology of the Amsterdam University Medical Center (VUmc) in Amsterdam, as described previously.[20] Collection, storage and use of tissue and patient data were performed in compliance with the "Code for Proper Secondary Use of Human Tissue in the Netherlands" (https://www.federa.org/). All normal samples were adjacent to colorectal neoplasia; four normal colon samples were adjacent to adenomas and cancers, six to colorectal adenomas and eight to CRC. All normal samples were obtained from the furthest point from colorectal neoplasia within the surgically resected material and judged as 100% normal by an expert pathologist. In our study all adenomas were larger than 1 cm in size to allow sampling of fresh frozen material for research purposes from tissues that were collected for routine diagnostics. Therefore, all of the adenomas used in our study were "advanced adenomas." For each sample, one tissue piece was cut into serial sections that were alternatingly used for DNA, RNA and protein isolation in the order DNA–RNA-protein-(…)-DNA–RNA-protein, to obtain the most comparable molecular profiles on DNA, RNA and protein level.

Genomics data

Low-coverage whole genome sequencing (WGS) data for the adenomas and RNA sequencing (RNA-seq) data for colorectal adenomas and cancers were obtained in our previous study.[20] For the normal adjacent colon sample collection, DNA and RNA isolation, low-coverage WGS and RNA-seq was performed as previously described for adenomas and cancers.[20] Raw sequencing data were made available through the European Genome-Phenome Archive (https://ega-archive.org/, EGAS00001002854). DNA copy number aberration identification in CRCs and normal adjacent colon samples was performed as described previously for the adenomas.[20]

Mass spectrometry proteomics data

Sample preparation for liquid chromatography tandem mass spectrometry proteomics (LC–MS/MS) was performed as previously described,[21] with some modifications (Supplementary Materials and Methods). Mass spectrometry was performed on a Q Exactive-HF mass spectrometer (Thermo Fisher, Bremen, Germany) using a data independent acquisition mass spectrometry protocol. The data independent acquisition mass spectrometry method consisted of a MS1 scan from 400 to 1,000 m/z at 15,000 resolution (AGC target of 3×10^6 and 50 ms injection time). For MS2, 24 variable size DIA segments were acquired at 30,000 resolution (AGC target 3×10^6 and auto for injection time). The data independent acquisition mass spectrometry method included 20 windows of 20 m/z, 2×40 m/z and 2×60 m/z. Collision energy was set at 28%. The spectra were recorded in centroid mode. The default charge state for the MS2 was set to 3.

RNA-seq data analysis

RNA-seq data preprocessing was performed as described previously,[20] now using human genome build hg19 (USCS RefSeq hg19, gencode v19 annotation). RNA-seq data were subjected to differential expression analysis, cellular decomposition (ESTIMATE[22] algorithm), gene set enrichment analysis (GSEA)[23] and gene-dosage effect analysis (Supplementary Materials and Methods).

Proteomics data analysis

An in-house spectral library was established using LC–MS/MS data derived from CRCs, colorectal adenomas and normal adjacent colon samples (manuscript in preparation), which was used in Spectronaut[24] to identify mass spectra. Protein groups were identified, quality control was performed and protein expression data was subjected to differential expression analysis, GSEA[23] and gene-dosage effect analysis (Supplementary Materials and Methods).

Quantification of tumor-stroma and goblet cells

Fresh-frozen tissue sections taken "before" and "after" the tissue sections used for DNA, RNA and protein isolation were stained with hematoxylin and eosin, and scanned using Aperio AT2 Scanner (Leica Biosystems Imaging, Amsterdam, The Netherlands). The digital images were used for stroma and goblet cells quantification (Supplementary Materials and Methods).

Immunohistochemical staining of tissue microarrays and patient-derived colorectal adenoma organoids

Candidate drivers of adenoma-to-carcinoma progression were selected for immunohistochemical (IHC) validation of protein expression in colorectal tissues using tissue microarrays (TMAs), and in cultures of epithelial cells using sections of patient-derived colorectal adenoma organoids. Candidates were selected using the following criteria: higher expression in HRAs when compared to low-risk adenomas (LRAs); and higher intensity in CRCs when compared to normal colon according to the Human Protein Atlas (www.proteinatlas.org).[25] See Supplementary Materials and Methods for details on IHC and patient-derived organoids.

Data availability

Raw sequencing data were made available through the European Genome-Phenome Archive (https://ega-archive.org/, EGAS00001002854). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium *via* the PRIDE partner repository with the accession identifier PXD012254.

Results

Molecular characterization of LRA and HRA

With the aim to characterize colorectal adenomas in the context of colorectal tumor progression, we have performed low-coverage WGS, genome-wide RNA-seq and tandem mass spectrometry proteomics (LC–MS/MS) on 30 colorectal adenomas,[20] 30 CRCs and 18 adjacent normal colon tissues (see Fig. 1 for an overview of the analyses applied in the entire study and Supplementary Table S2 for clinical information on the samples). Using low-coverage WGS we determined DNA copy number aberrations in the samples. Within the adenomas, nine HRAs were identified based on the presence of at least two CAEs. To obtain a robust representation of LRAs, only microsatellite-stable (MSS) lesions that carried none of the CAEs were included. Two adenomas were microsatellite-instable (MSI), two adenomas carried only one CAE, and for two adenomas the calling of CAEs remained inconclusive,[20] leaving 15 MSS adenomas with no CAEs that were classified as LRAs (Supplementary Fig. S1*a* and Table S3). No significant associations were observed for risk of progression and pathological adenoma features like size, grade of dysplasia or histology (Table S4). CRCs showed the well-known nonrandom pattern of chromosomal instability with CAEs being the most frequent, next to gain of chromosome 7 and loss of chromosome 14 (Fig. S1*a*). As six CRCs had previously been identified as MSI,[20] the DNA copy number frequency for MSI CRCs and MSS CRCs were examined separately, revealing less chromosomal aberrations in MSI CRCs (Fig. S1*b*). No chromosomal aberrations were observed in the normal adjacent colon samples (Fig. S1*a*).

To explore the biological processes playing a role in colorectal tumor progression, the tissue samples were analyzed by RNA-seq and LC–MS/MS. Mass spectrometry analysis lead to identification of 5,080 protein groups in the whole data set and 4,903 in the group of HRAs and LRAs (false discovery rate ≤0.01). Among the adenomas, one HRA was identified as an outlier due to low protein group number and highly differing expression profile from the rest of the adenoma samples (Fig. S2) and was excluded from further proteomic analyses. Dimensionality reduction of the RNA and protein expression data allowed to clearly discern adenomas from CRCs and normal adjacent colon tissues (Figs. S3*a* and S3*c*) while HRAs and LRAs were indistinguishable (Figs. S3*b* and S3*d*).

Differential gene expression analysis between the HRAs and LRAs revealed 298 genes with higher and 125 genes with lower expression in HRAs (Table S5). Differential protein expression analysis revealed 78 proteins with higher and 86 with lower

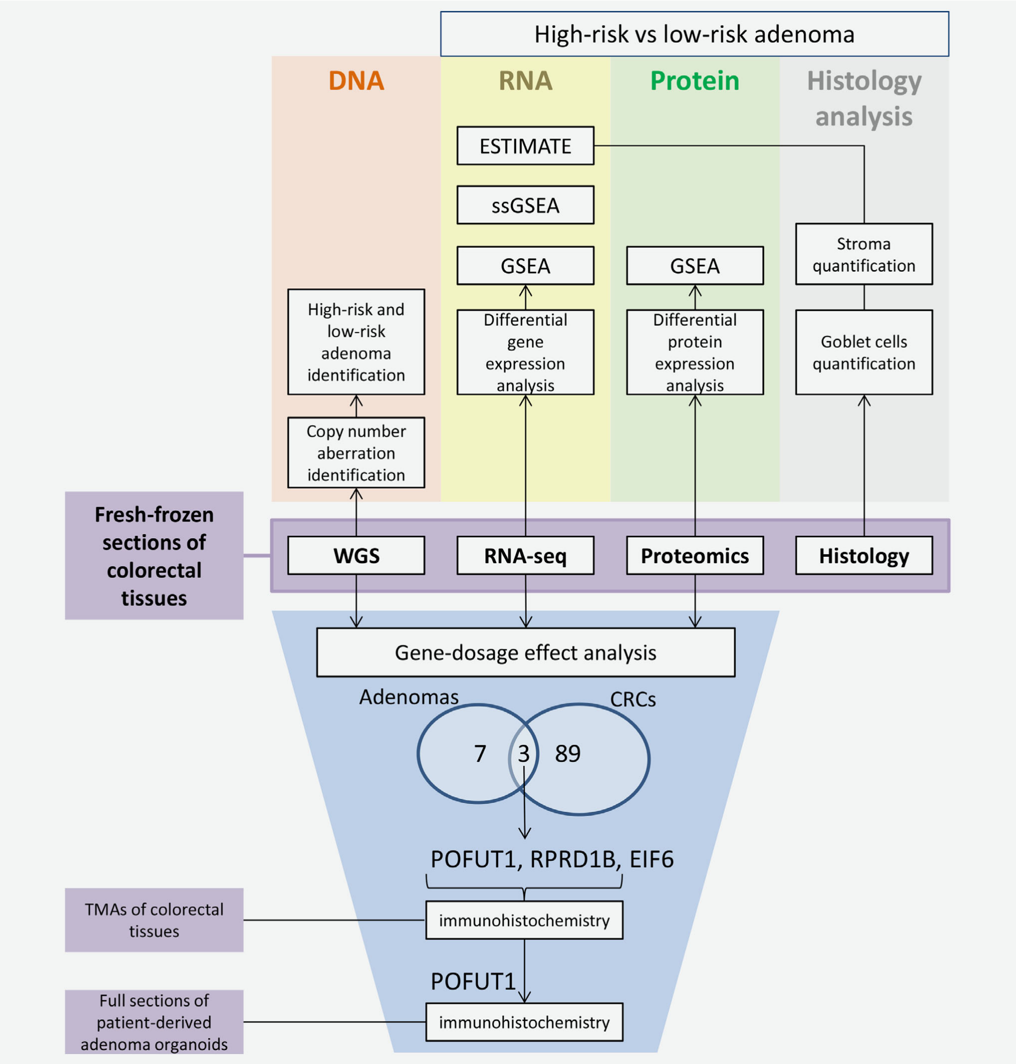

High-risk vs low-risk adenoma
DNA
RNA
Protein
Histology analysis
ESTIMATE
ssGSEA
GSEA
GSEA
Stroma quantification
High-risk and low-risk adenoma identification
Differential gene expression analysis
Differential protein expression analysis
Goblet cells quantification
Copy number aberration identification
Fresh-frozen sections of colorectal tissues
WGS
RNA-seq
Proteomics
Histology
Gene-dosage effect analysis
Adenomas
CRCs
7
3
89
POFUT1, RPRD1B, EIF6
TMAs of colorectal tissues
immunohistochemistry
POFUT1
Full sections of patient-derived adenoma organoids
immunohistochemistry

Figure 1. Fresh-frozen tissue fragments of colorectal cancers (*n* = 30), colorectal adenomas (*n* = 30) and normal adjacent colon samples (*n* = 18) were used for low-coverage WGS, RNA-seq, tandem mass spectrometry proteomics and histology analysis. DNA copy number aberration identification and HRA and LRA stratification was performed using the low-coverage WGS data. RNA-seq and proteomics data were used for differential gene/protein expression analysis and GSEA. Additionally, single sample GSEA and ESTIMATE algorithm, which calculate the enrichment of stromal and immune gene signatures, were used on the RNA expression data set. Stroma quantification was performed on sections originating from the same tissue fragments as used for the molecular profiling data to validate the results of the expression analysis. Stroma percentage was compared between HRA and LRA and correlated with the stromal score of the ESTIMATE algorithm. Next, DNA copy number driven gene-dosage effect analysis was performed. Ninety-two and ten genes were identified to correlate in terms of DNA copy number, RNA and protein expression in CRCs and adenomas, respectively. Three genes, *POFUT1*, *RPRD1B* and *EIF6*, were overlapping between adenomas and cancers and were observed to be amplified and overexpressed in HRAs and CRCs. Validation of *POFUT1* and *RPRD1B* by immunohistochemical staining was performed in TMAs of the formalin-fixed, paraffin-embedded tissue pieces and for *POFUT1* also in full sections of patient-derived adenoma organoids. Additionally, goblet cell quantification was performed on the sections of colorectal adenomas and association with *POFUT1* expression and risk of progression was identified. [Color figure can be viewed at wileyonlinelibrary.com]

expression in HRAs (Table S6). Fourteen genes were differentially expressed on both RNA and protein level, with 9 genes higher and 5 lower expressed in HRAs (Table S7). To gain further insights into the global differences between the adenomas, we performed GSEA with hallmark gene signatures (molecular signature database[26]) on lists of genes and proteins ranked according to differences in the expression between HRAs and LRAs (Fig. 2). Processes that were more prominent in HRAs on RNA and protein level were related to proliferation, immune response and stroma development. Additionally, a number of signaling pathways were enriched in HRAs either only on the RNA (KRAS-signaling up, Hedgehog-, WNT-, IL2-STAT5-, NOTCH-signaling' or protein level (PI3K/AKT/mTOR-, mTORC1-signaling). The processes more prominent in LRAs compared to HRAs were identified on the protein level and included "protein secretion" and the metabolic gene sets (Fig. 2).

To put the GSEA group-level differences between HRAs and LRAs in context of progression toward CRC, we performed single-sample GSEA on RNA level in adenomas and cancers using the hallmark gene sets (Fig. S4). Seven gene sets were significantly differential between HRAs and LRAs ($p \leq 0.05$, Fig. 3). In six cases, the single-sample GSEA score increased through colorectal tumor progression, with the lowest score in LRAs and the highest in CRCs. These include "Notch-" and "Hedgehog-signaling" together with immune- and stroma-related gene sets, like "epithelial-mesenchymal transition." For "heme metabolism," the single-sample GSEA score decreased through colorectal tumor progression (Fig. 3).

Characterization of LRA and HRA tumor microenvironment

As GSEA revealed increased stroma and immune processes in HRAs, we examined the differences in tumor microenvironment between HRAs and LRAs. By applying the ESTIMATE algorithm[22] on RNA expression data, enrichment scores for stromal and immune signatures were calculated in each sample reflecting the expression of stroma- and immune-related genes (Fig. S5). A significant increase of stromal score was identified in HRAs when compared to LRAs ($p = 0.012$). An even more significant increase was observed between MSS cancers and HRAs ($p = 5.7e^{-5}$). In terms of the immune score, even though a gradual increase from LRAs through HRAs to MSS cancers was identified, the differences between the groups were insignificant ($p = 0.096$ and 0.98, respectively). MSI cancers had significantly higher immune score than MSS cancer ($p = 0.021$, Fig. S5).

To morphologically confirm the differences in the amounts of stroma between the HRAs and LRAs, we performed stroma quantification on hematoxylin and eosin-stained slides by digital image analysis (Fig. 4*a*). One sample could not be analyzed due to excessive tissue folds. The amount of stroma in HRAs (median = 40.89) was significantly higher than in LRAs (median = 27.20, $p = 0.002$, Fig. 4*b*). Stroma percentage calculated by image analysis also positively correlated with the ESTIMATE stromal score from the RNA expression analysis (Fig. 4*c*). This indicates that the expression differences between HRAs and LRAs in stromal and immune pathways are associated with the morphological differences in the amount of stroma in the tissue samples.

Candidate drivers of adenoma-to-carcinoma progression

Next to identification of differences in tumor microenvironment, we investigated DNA copy number driven gene-dosage effect to reveal changes between HRAs and LRAs driven by the aberrations in the epithelial cells (Fig. 1). Pairwise correlation analysis was performed between DNA copy number, RNA and protein expression for colorectal adenomas and CRCs. In the cancers, 92 genes were positively correlated among the data types (Fig. S6 and Table S8). Chromosome 20 was associated with the largest global expression changes on RNA and protein level with 28 genes (~30%), including *HNF4A*, *TOMM34* and *RPRD1B*, which were previously described to be gained and overexpressed in CRC cell lines and tissues.[27,28] Gene-dosage effect was also identified for *DIS3*, which is located on chromosome 13 and often gained in CRC.[27,29] Other genomic regions with the highest number of perturbed genes considered almost all chromosomes involved in the CAEs.

In the adenomas, positive and significant correlations between DNA copy numbers, RNA and protein expression were identified for 10 genes (Fig. S6 and Table S9). As HRAs are characterized by presence of CAEs, potential drivers of early colorectal tumor progression are expected to reside on the CAE-defined chromosomes. Gene-dosage effect was identified for two genes from chromosome arm 8p; however, these genes were associated both with gains and losses in the HRA group (Fig. S1*a*) and consequently, higher and lower gene and protein expression when compared to LRAs. For the genes located on the CAE-related chromosome 20, *POFUT1*, *RPRD1B* and *EIF6*, gene-dosage effect was associated with only gains (Fig. S1*a*) and overexpression in HRAs when compared to LRAs (Fig. 5). We performed gene-level overlap analysis between gene-dosage effects in CRCs and in adenomas to identify genes prominent for both HRAs and CRCs. The analysis revealed *POFUT1*, *RPRD1B* and *EIF6*, implying that the gain of chromosome arm 20q and expression of these three genes play an important role in both HRAs and CRCs. For all of these three genes DNA copy number, RNA and protein expression increased gradually from normal adjacent colon, through LRAs and HRAs to CRCs (Fig. 5). *POFUT1*, *RPRD1B* and *EIF6* reside on neighboring cytogenetic bands—20q11.21, 20q11.23 and 20q11.22, respectively. Moreover, significant positive correlations were identified between these genes on DNA, RNA and protein level, suggesting their coamplification and coexpression (Fig. S7).

To validate gene-dosage effect of *POFUT1*, *EIF6* and *RPRD1B* in colorectal tumors, we evaluated the relation between DNA copy numbers, RNA and protein expression of these genes in The Cancer Genome Atlas (TCGA) Provisional CRC data set.[30,31] Gene-dosage effect was confirmed for each of these three genes in this data set on both RNA (n = 382) and protein level (n = 90), as gene and protein expression was higher when the DNA copy of the gene was gained or amplified (Figs. S8–S10).

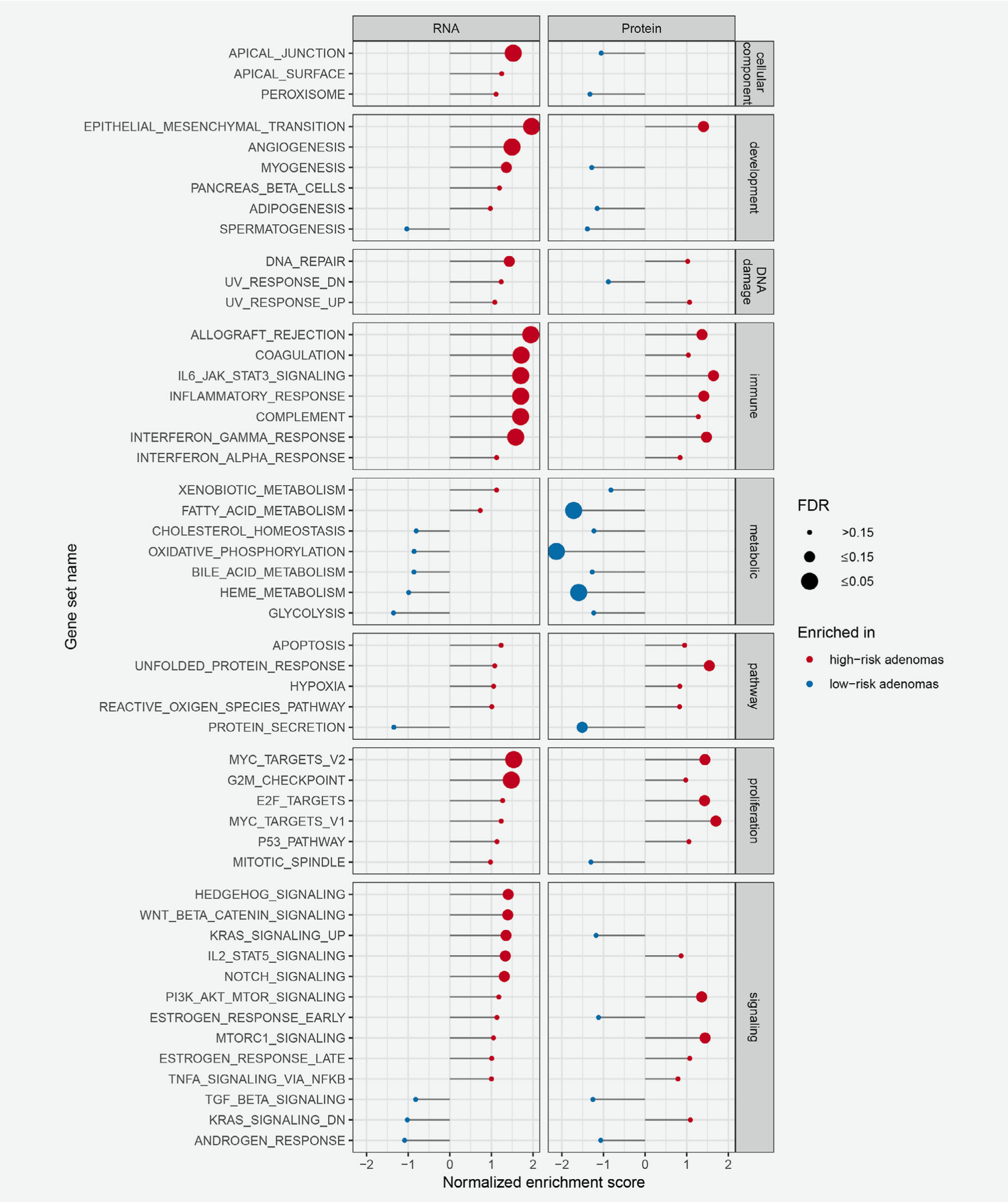


Figure 2. Legend on next page.

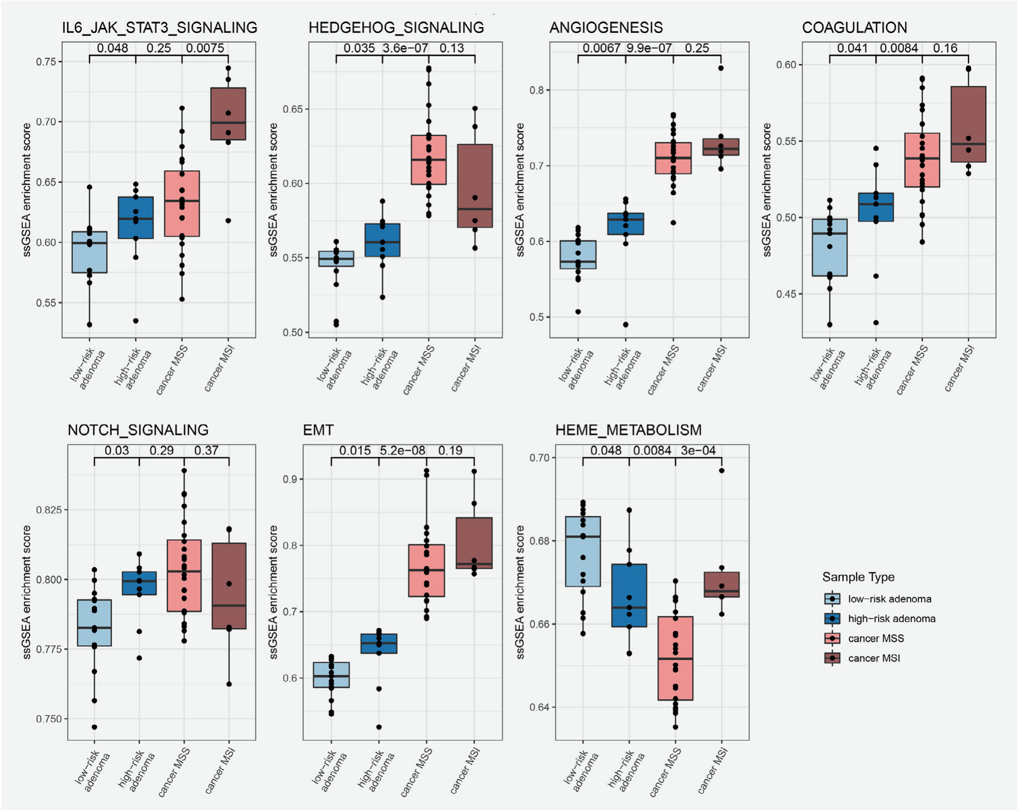


Figure 3. Single sample gene set enrichment scores represented per sample type; LRAs, HRAs and CRCs. Gene sets with significant differences in enrichment scores between HRA and LRA ($p \leq 0.05$) were selected for this figure. [Color figure can be viewed at wileyonlinelibrary.com]

Validation of increased POFUT1 expression in HRAs

To verify whether protein expression of *POFUT1*, *RPRD1B* and *EIF6* is increased in CRCs and HRAs compared to LRAs and normal colon tissue, we aimed to evaluate their expression by immunohistochemistry (IHC) using TMAs obtained from the same samples as were used for the molecular profiling. Data in the Human Protein Atlas[25] indicated that the expression of EIF6, as measured by IHC, is already high in normal colon tissue, leaving little room to detect increased EIF6 protein expression in adenomas and CRCs. Therefore, TMAs were stained for POFUT1 and RPRD1B, while EIF6 was discarded from IHC analysis.

Within the TMA cores of colorectal tissues, RPRD1B was observed mainly in the nuclei of epithelial cells (Fig. S8), the staining confirmed increasing protein expression of RPRD1B in HRAs and CRCs as observed in the molecular profiling data (Fig. 5*c*). Nevertheless, several LRAs and normal adjacent colon samples exhibited high intensity of RPRD1B staining (Fig. S11

Figure 2. Gene set enrichment analysis results in the differential analysis between HRA and LRA, on RNA and protein level, as measured by RNA-seq and mass spectrometry proteomics. Genes or proteins were ranked based on their fold change and p-value, with genes/proteins significantly overexpressed in HRAs on top of the list. GSEA was performed on the ranked list using hallmark gene sets. Gene sets enriched in HRAs are marked red, and gene sets enriched in LRAs are marked blue. The size of the dot reflects the significance of the enrichment (false discovery rate ≤0.15). For a subset of the signaling pathways, like Hedgehog, Wnt and Notch, GSEA on the protein level could not be determined since the number of proteins from these gene sets identified by LC–MS/MS was too small. [Color figure can be viewed at wileyonlinelibrary.com]

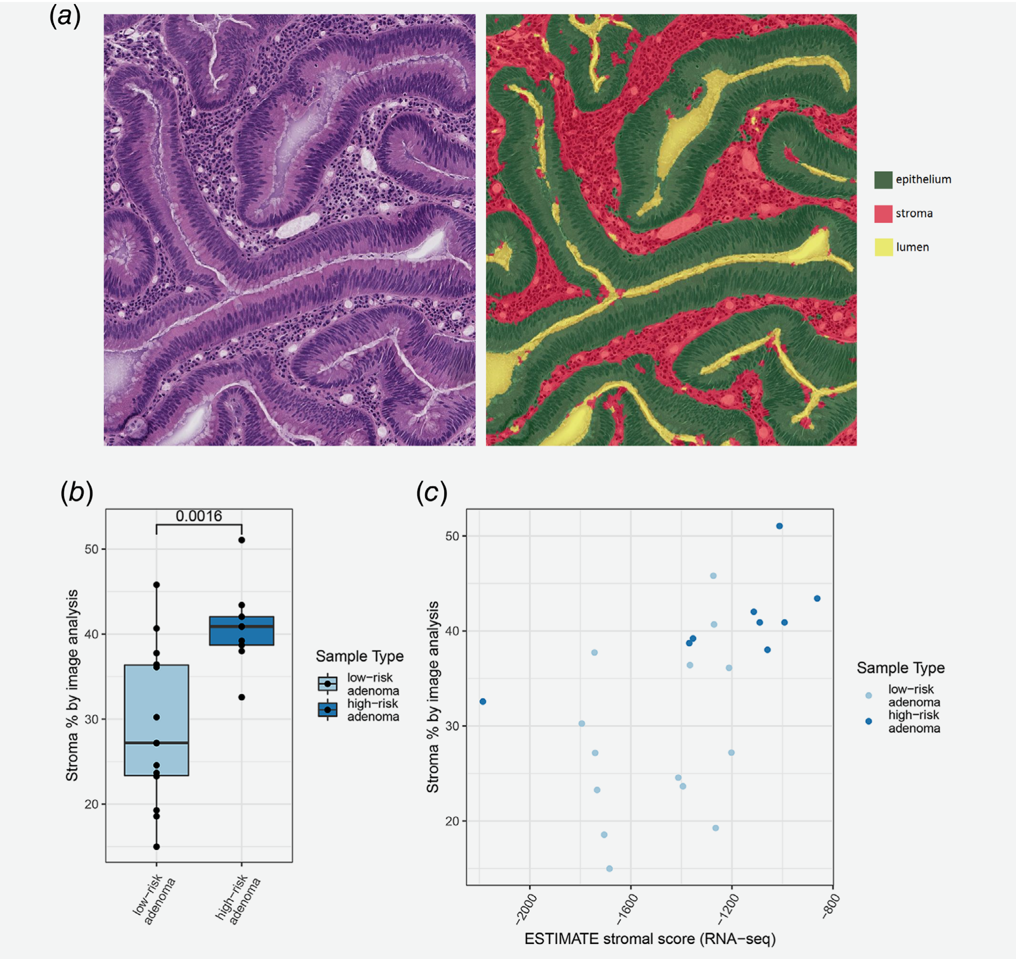


Figure 4. Stroma quantification on hematoxylin and eosin-stained slides. (*a*) Representative image of assigning class to area on the slide; stroma, epithelium or lumen. Each class was quantified by calculating the size of its area. (*b*) Significant difference in stroma percentage between HRA and LRA, as calculated by the image analysis. (*c*) Significant positive correlation identified between stroma percentage measured by image analysis and ESTIMATE stromal score. [Color figure can be viewed at wileyonlinelibrary.com]

and Table S10). Therefore, the difference in RPRD1B expression measured by IHC between LRAs and HRAs was not significant (*p* = 0.197; Table S10). Comparisons of CRCs with HRAs to LRAs and of CRCs with HRAs to LRAs with normal colon samples yielded significant differences (*p* = 0.017 and 0.003, respectively; Table S10).

POFUT1 immunohistochemical staining was predominantly observed in the cytoplasm of epithelial cells, the staining showed gradual increase of POFUT1 expression through different stages of colorectal tumor progression (Figs. 6*a* and 6*b*), thereby verifying the molecular profiling data (Fig. 5*b*). High levels of POFUT1 expression measured by IHC were more frequent in HRAs compared to LRAs, in HRAs and cancers compared to LRAs and in HRAs and cancers compared to LRAs and normal adjacent colon (Tables S11*a* and S11*b*). POFUT1 expression was also significantly associated with grade of dysplasia (Table S11*b*). Interestingly, POFUT1 expression was lower in MSI than in MSS cancers on both RNA and protein level (Figs. 5*b* and 6), suggesting its specific role for chromosomal instability tumors. Previously, depletion of *POFUT1* was shown to play a role in differentiation

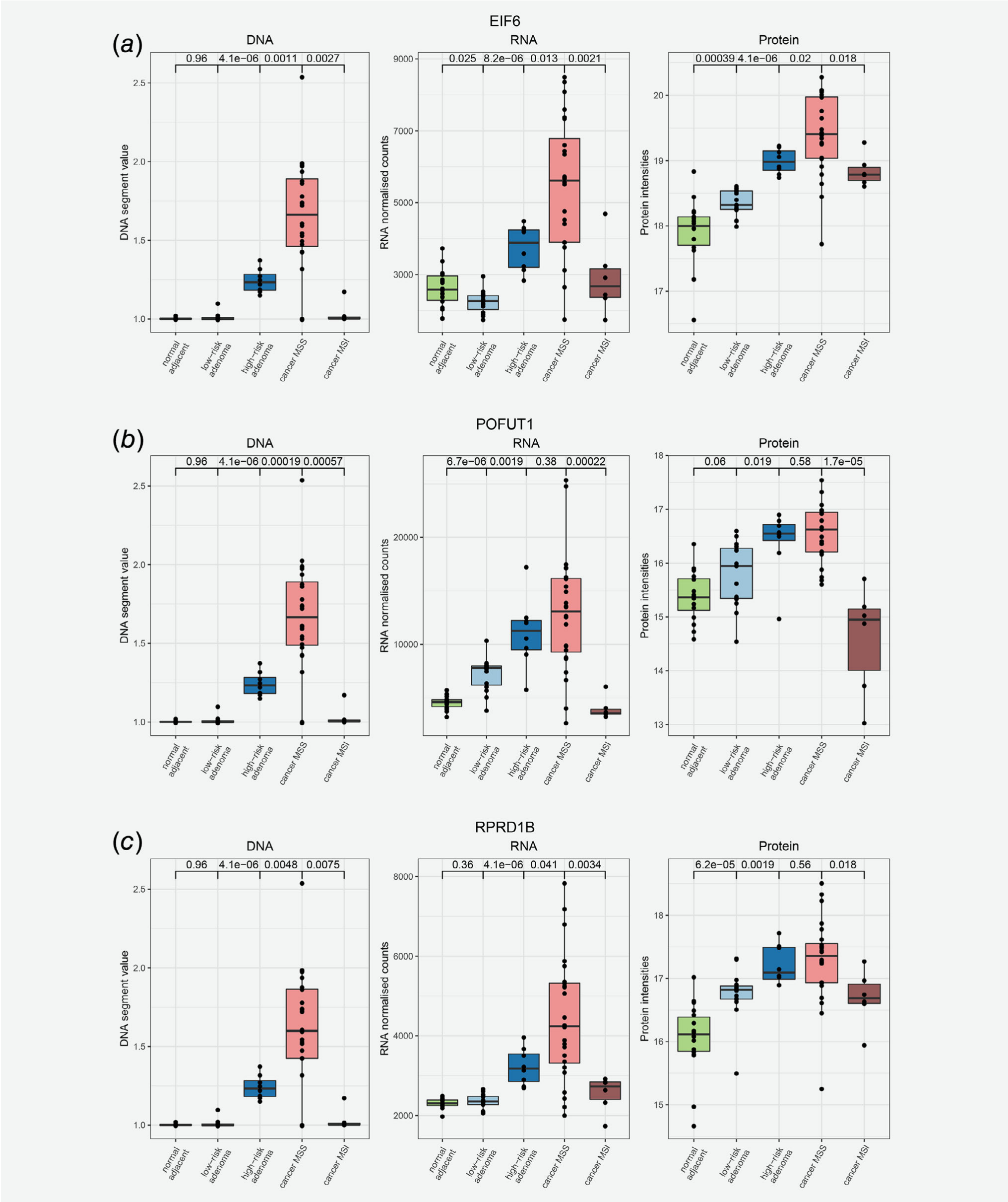


Figure 5. Proteogenomic representation of the potential drivers of colorectal tumors. DNA copy number, RNA and Protein expression (as measured by mass spectrometry proteomics) were plotted for EIF6 (*a*), POFUT1 (*b*) and RPRD1B (*c*) for each sample among different stages of colorectal tumor development: normal adjacent colon, LRAs, HRAs and CRCs. Correlating, gradual increase in DNA copy number and RNA and Protein expression was observed for each of these three genes. [Color figure can be viewed at wileyonlinelibrary.com]

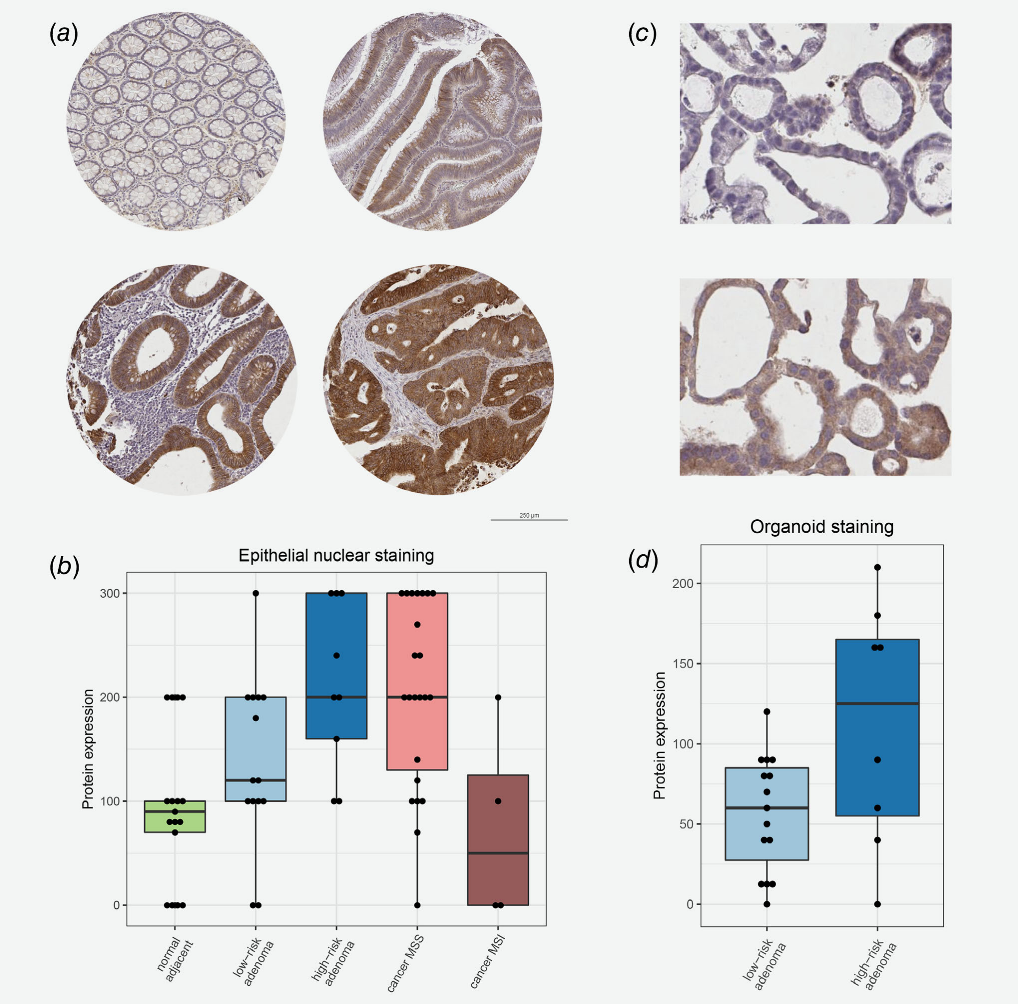


Figure 6. Immunohistochemical staining of POFUT1 in colorectal tissues and patient-derived organoids. (*a*) Representative POFUT1 staining in different tissue sample type. Top left: normal adjacent colon; top right: LRA; bottom left: HRA; bottom right: CRC. (*b*) POFUT1 expression as measured by a product of epithelial cytoplasmic staining intensity (negative = 0, weak = 1, moderate = 2 or strong = 3) and percentage of the cells stained positively (0–100%) was plotted for each tissue sample among different stages of colorectal tumor development. See Table S11 for group comparisons and statistical testing. (*c*) Representative images of POFUT1 staining in LRA organoid (top) and HRA organoid (bottom). (*d*) POFUT1 expression in epithelial cytoplasm plotted in HRA and LRA organoids, as measured by a product of epithelial cytoplasmic staining intensity (negative = 0, weak = 1, moderate = 2 or strong = 3) and percentage of the cells stained positively (0–100%). See Table S13 for group comparisons and statistical testing. [Color figure can be viewed at wileyonlinelibrary.com]

of the proliferative epithelial cells into goblet cells through inactivation of Notch signaling.[32] Therefore, we quantified the amount of goblet cells in the adenomas using hematoxylin and eosin-stained sections to examine this finding in the context of risk of progression. No association of the amount of goblet cells with dysplasia or other pathological features was identified (Table S11*b*). Lower amounts of goblet cells were significantly associated with high POFUT1 expression ($p = 0.017$; Table S11*a*) and high risk of progression ($p = 0.007$; Table S11*b*), implying that also in our study *POFUT1* is linked to goblet cell differentiation and indicating its role in early colorectal tumor development.

To further corroborate the role of *POFUT1* in the pathogenesis of CRC in an independent series, expression of POFUT1 was investigated in a cohort of patient-derived colorectal adenoma organoids. First, we performed low-coverage WGS and based on the presence of two or more CAEs revealed 8 HRA and 15 LRA organoids in the series (Table S12). Next, IHC staining of the organoids was performed to evaluate POFUT1 expression in the neoplastic cells. Also in the organoids, POFUT1 was mainly observed in the cytoplasm and high POFUT1 expression was associated with HRAs ($p = 0.008$; Table S13 and Figs. 6*c* and 6*d*), confirming its potential role in early colorectal tumor development.

Discussion

Studying the natural history of colorectal adenomas, including progression to cancer, is challenging because adenomas are removed when detected during colonoscopy. Yet, there is a need for better understanding of the biology of adenomas that progress to CRC. We set out to molecularly characterize adenomas at high risk of progressing to CRC and to identify putative drivers of this process. *POFUT1* was found to be amplified and overexpressed in HRAs and CRCs when compared to LRAs and adjacent normal colon epithelium. POFUT1 overexpression was successfully validated by immunohistochemical staining on TMAs and in patient-derived colorectal adenoma organoids, indicating that POFUT1 plays a role in colorectal adenoma-to-carcinoma progression. Additionally, high POFUT1 expression and high risk of progression to cancer were associated with a decrease in goblet cell differentiation.

The novelty of the current study is multi-omics analysis of colorectal adenomas at high and low risk of progressing to cancer, in the context of CRCs and normal adjacent colon samples. Comprehensive analysis of high throughput DNA, RNA and protein profiling data of the same samples has not been performed yet for colorectal adenomas, while it did provide additional insights in CRC.[27,28] On RNA and/or protein level, the enrichment of gene sets and pathways were identified to be increasing through different stages of colorectal tumor development, from normal colon, through LRA and HRA to CRC. These included pathways known to play a role in or accompany colorectal carcinogenesis like Hedgehog, Notch, KRAS, PI3K/AKT/mTOR or Wnt signaling, proliferation, epithelial-mesenchymal transition or immune activation.[33] This suggests that a lot of processes inherent to cancer are already more active in HRAs compared to LRAs. Conversely, gene sets enriched in LRAs when compared to HRAs, like protein secretion, fatty-acid or heme metabolism, decreased in CRC, consistent with previous observations.[34] Fourteen genes were identified to be differentially expressed between HRAs and LRAs on both RNA and protein level. Among upregulated genes/proteins in HRAs, genes of both epithelial and stromal origins were found. This included *HNF4A*, a transcriptional activator of epithelial differentiation[35] that is located on chromosomal arm 20q, previously shown to be amplified and activated in the majority of CRCs[28] and studied as a prognostic biomarker for this disease.[36] An unexpected result was the overexpression of multiple tumor microenvironment-related genes/proteins in HRAs, including collagens, fibronectin, vimentin, immunoglobulins or calprotectin. While a broad range of stroma proportion has been reported in CRC,[37] this is far less evident in adenomas. It has been shown that stromal genes can be expressed by epithelial cells, which typically occurs in association with invasion, a phenomenon referred to as epithelial-mesenchymal transition.[35] Nevertheless, by definition, stroma invasion is a process characteristic to cancer and not yet occurring in adenomas. We have performed stroma quantification by image analysis on adenoma tissue sections originating from the same tissue fragments that were used for molecular profiling, and observed a significant increase in stroma percentage in HRAs compared LRAs. Our data indicate that differential expression of the stroma genes between HRAs and LRAs is likely due to the differences in the stroma proportion. Even though significant, the variation in the amount of stroma in the adenomas is certainly not as big as in CRCs.[37]

To identify putative drivers of adenoma-to-carcinoma progression from the epithelial cells, we examined DNA-driven aberrations in the colorectal tumors. Combining DNA and RNA data to study gene-dosage effect has been performed in CRC[18]; however, only for a limited number of potential candidates functional assays confirmed their oncogenic potential.[14,29,38] Addition of the protein layer provides insight into which chromosomal aberrations lead to functional consequences.[28] Despite the high depth of the proteomics measurement in the present study with over 5,000 protein groups detected in total, adding the protein layer can be also limiting, in terms of the number of proteins measured overall and subsequently considered in the analysis. In our study, gene-dosage effect analysis in CRCs led to the identification of 92 genes, a subset of which has previously been described, including *HNF4A*,[28] *TOMM34*,[28] *DIS3*[29] or *RPRD1B*.[27]

In the adenomas, the CAE-driven gene-dosage effect analysis yielded potential drivers of colorectal tumor progression that are already amplified and overexpressed in HRAs—*POFUT1*, *RPRD1B* and *EIF6*. The three genes are located on neighboring cytobands of chromosome arm 20q, which is the most frequently amplified chromosomal arm in CRC.[18,28]

POFUT1 is a fucosylation factor that activates Notch through addition of fucose groups,[39] a process required for the canonical Notch signaling.[32,40] In our study, *POFUT1* was amplified and overexpressed while Notch signaling was enriched in HRAs and CRCs, when compared to LRAs. High expression of POFUT1 in HRAs and CRCs was validated using immunohistochemical staining of TMAs and adenoma-derived organoids. Recently, *POFUT1* overexpression was shown to have oncogenic activity in CRC through activation of *NOTCH1* signaling, and consequently affecting proliferation, invasion and migration.[41] Additionally, depletion of *POFUT1* or Notch signaling was shown to be associated with converting proliferative cells into goblet cells.[32,42] Indeed, in the present study, low numbers of goblet cells were significantly associated with high-risk status and high POFUT1 expression in adenomas, indicating that in HRAs *POFUT1* and

Notch signaling play a role in increased proliferation and decreased differentiation. Altogether this suggests that *POFUT1* through the Notch signaling pathway is a putative driver of adenoma-to-carcinoma progression. Further functional studies on adenoma preclinical models are needed to confirm this hypothesis.

RPRD1B is overexpressed in many tumor types and has been shown to have an oncogenic activity by regulating the transcription of cyclin D1[43] and other Wnt targets,[44] consistent with the significant enrichment of Wnt signaling in HRAs demonstrated by GSEA in the present study. *RPRD1B* was proven to accelerate tumorigenesis by promoting cell proliferation and invasion.[43,44] Altogether, this suggests that *RPRD1B* may play a role in colorectal tumor progression through enhanced Wnt signaling. Although the TMA IHC analyses did not validate differences in RPRD1B expression levels between LRA and HRA, its predominant staining of neoplastic cells combined with the molecular profiling data suggest that *RPRD1B* should also be considered as a putative driver of colorectal tumor development.

EIF6 is a translation initiation factor that plays a role in ribosome complex formation and protein synthesis downstream of PI3K/Akt/mTOR signaling pathway.[45,46] It is overexpressed in multiple tumor types,[47,48] including CRC, where expression of EIF6 has been shown to increase from normal colon, through adenoma to CRC.[49] Functional studies on *EIF6* suggest its oncogenic activity through increasing cancer cell motility and invasion.[50,51] The fact that we identified significant enrichment of PI3K/Akt/mTOR signaling in HRAs when compared to LRAs, suggests that *EIF6* and PI3K/Akt/mTOR signaling play a role in adenoma-to-carcinoma progression. Additionally, the transcription of *EIF6* has been shown to be regulated by *NOTCH1*,[51] consistent with Notch signaling enrichment in HRAs and CRCs.

Individuals with a history of colorectal neoplasia carry an increased risk of developing CRC in the future and therefore are enrolled in the colonoscopy-based surveillance programs.[52] As removal of nonmalignant precursor lesions during colonoscopy is an approach to decrease CRC incidence and mortality rates,[53] currently, detection of advanced adenoma is an indication to shorten the interval for the follow-up surveillance colonoscopy.[52] The high prevalence of advanced adenomas in an elderly population leads to a substantial burden on endoscopic capacity.[52] Moreover, given that not all advanced adenomas eventually progress to cancer, frequent surveillance colonoscopies in patients with these lesions lead to overdiagnosis and overtreatment.[4] In our study, we have shown that HRAs, in contrast to LRAs, in a number of aspects resemble CRCs on molecular level, while they represent only approximately 30% of the advanced adenomas.[19] Introduction of a more specific definition of adenomas associated with risk of future CRC development may significantly improve the CRC surveillance programs and reduce patient burden. Additional studies are still needed to evaluate if patients with HRAs indeed have higher CRC incidence and mortality rate compared to patients with advanced adenomas, and whether POFUT1 can be used as biomarker to identify HRAs in the surveillance setting.

In our study, we performed multi-omics characterization of colorectal adenomas in the context of colorectal tumor development. We focused on conventional chromosomal instability adenomas, the most prevalent precursors of CRC,[10] as MSI adenomas are relatively rare with a prevalence of only 3%.[54] MSI CRCs were included in our analyses, which frequently differed from MSS CRCs in terms of gene expression and GSEA, confirming the distinct etiology of MSS and MSI CRCs. *POFUT1*, *RPRD1B* and *EIF6* were identified as putative drivers of adenoma-to-carcinoma progression. In light of what is known about the roles these genes play in carcinogenesis, our results imply that the transition from LRAs to HRAs involves the interplay of Wnt, Notch and PI3K/AKT/mTOR signaling pathways. As such, our study shows that biological processes inherent to CRC are already more active in HRAs than in LRAs. Moreover, our study emphasizes the key role that specific DNA copy number alterations play in progression from premalignancy to cancer, indicating that in comparison to the generally used morphology-based concept of "advanced adenoma," the molecular CAE-based concept of HRA is a more specific marker to define risk of progressing to CRC.

Acknowledgements

This study was performed within the framework of the CTMM (Center for Translational Molecular Medicine) research program, project NGS-ProToCol (03O-402) and was supported by Dutch Cancer Society (KWF Kankerbestrijding), project number 2013-6025, Cancer Center Amsterdam and SU2C-DCS International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415, MEDOCC). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. Proteomics infrastructure was supported by ZonMW (91116017). The authors thank ServiceXS (GenomeScan, Leiden) for sequencing the NGS-ProToCol samples and the Genomics Core Facility (Netherlands Cancer Institute, Amsterdam) for additional sequencing. The authors also thank Erik Hooijberg and Marjolijn Mertz for their help with image analysis and Margriet Lemmens for her assistance in the data acquisition.

References

1. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990;61:759–67.
2. Lieberman DA, Weiss DG, Bond JH, et al. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs Cooperative Study Group 380. *N Engl J Med* 2000;343:162–8.
3. Imperiale TF, Wagner DR, Lin CY, et al. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N Engl J Med* 2000;343:169–74.
4. Kalager M, Wieszczy P, Lansdorp-Vogelaar I, et al. Overdiagnosis in colorectal cancer screening: time to acknowledge a blind spot. *Gastroenterology* 2018;155:592–5.
5. Muto T, Bussey HJ, Morson BC. The evolution of cancer of the colon and rectum. *Cancer* 1975;36:2251–70.
6. Shinya H, Wolff WI. Morphology, anatomic distribution and cancer potential of colonic polyps. *Ann Surg* 1979;190:679–83.
7. Brenner H, Hoffmeister M, Stegmaier C, et al. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. *Gut* 2007;56:1585–9.
8. Heselmeyer K, Schrock E, du Manoir S, et al. Gain of chromosome 3q defines the transition from severe

dysplasia to invasive carcinoma of the uterine cervix. *Proc Natl Acad Sci U S A* 1996;93:479–84.

9. Ried T, Just KE, Holtgreve-Grez H, et al. Comparative genomic hybridization of formalin-fixed, paraffin-embedded breast tumors reveals different patterns of chromosomal gains and losses in fibroadenomas and diploid and aneuploid carcinomas. *Cancer Res* 1995;55:5415–23.
10. Rajagopalan H, Nowak MA, Vogelstein B, et al. The significance of unstable chromosomes in colorectal cancer. *Nat Rev Cancer* 2003;3:695–701.
11. Hermsen M, Postma C, Baak J, et al. Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology* 2002;123:1109–19.
12. Meijer GA, Hermsen MA, Baak JP, et al. Progression from colorectal adenoma to carcinoma is associated with non-random chromosomal gains as detected by comparative genomic hybridisation. *J Clin Pathol* 1998;51:901–9.
13. Douglas EJ, Fiegler H, Rowan A, et al. Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res* 2004;64:4817–25.
14. Sillars-Hardebol AH, Carvalho B, Tijssen M, et al. TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* 2012;61:1568–75.
15. Camps J, Grade M, Nguyen QT, et al. Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. *Cancer Res* 2008;68:1284–95.
16. Hirsch D, Camps J, Varma S, et al. A new whole genome amplification method for studying clonal evolution patterns in malignant colorectal polyps. *Genes Chromosomes Cancer* 2012;51:490–500.
17. Ried T, Knutzen R, Steinbeck R, et al. Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosomes Cancer* 1996;15:234–45.
18. Carvalho B, Postma C, Mongera S, et al. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut* 2009;58:79–89.
19. Carvalho B, Diosdado B, Terhaar Sive Droste JS, et al. Evaluation of cancer-associated DNA copy number events in colorectal (advanced) adenomas. *Cancer Prev Res* 2018;11:403–12.
20. Komor MA, Bosch LJ, Bounova G, et al. Consensus molecular subtypes classification of colorectal adenomas. *J Pathol* 2018;246:266–76.
21. Bosch LJW, de Wit M, Pham TV, et al. Novel stool-based protein biomarkers for improved colorectal cancer screening: a case-control study. *Ann Intern Med* 2017;167:855–66.
22. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;4:2612.
23. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545–50.
24. Bruderer R, Bernhardt OM, Gandhi T, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen treated 3D liver microtissues. *Mol Cell Proteomics* 2015;14:1400–10.
25. Uhlen M, Zhang C, Lee S, et al. A pathology atlas of the human cancer transcriptome. *Science* 2017;357:eaan2507.
26. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The molecular signatures database hallmark gene set collection. *Cell Syst* 2015;1:417–25.
27. Wang J, Mouradov D, Wang X, et al. Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity. *Gastroenterology* 2017;153:1082–95.
28. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;513:382–7.
29. de Groen FL, Krijgsman O, Tijssen M, et al. Gene-dosage dependent overexpression at the 13q amplicon identifies DIS3 as candidate oncogene in colorectal cancer progression. *Genes Chromosomes Cancer* 2014;53:339–48.
30. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;2:401–4.
31. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;6:pl1.
32. Guilmeau S, Flandez M, Bancroft L, et al. Intestinal deletion of Pofut1 in the mouse inactivates Notch signaling and causes entero-colitis. *Gastroenterology* 2008;135:849–60. e6.
33. Sillars-Hardebol AH, Carvalho B, de Wit M, et al. Identification of key genes for carcinogenic pathways associated with colorectal adenoma-to-carcinoma progression. *Tumour Biol* 2010;31:89–96.
34. Carvalho B, Sillars-Hardebol AH, Postma C, et al. Colorectal adenoma to carcinoma progression is accompanied by changes in gene expression associated with ageing, chromosomal instability, and fatty acid metabolism. *Cell Oncol (Dordr)* 2012;35:53–63.
35. Vellinga TT, den Uil S, Rinkes IH, et al. Collagen-rich stroma in aggressive colon tumors induces mesenchymal gene expression and tumor cell invasion. *Oncogene* 2016;35:5263–71.
36. Chellappa K, Robertson GR, Sladek FM. HNF4α: a new biomarker in colon cancer? *Biomark Med* 2012;6:297–300.
37. Fijneman RJ, Carvalho B, Postma C, et al. Loss of 1p36, gain of 8q24, and loss of 9q34 are associated with stroma percentage of colorectal cancer. *Cancer Lett* 2007;258:223–9.
38. Camps J, Pitt JJ, Emons G, et al. Genetic amplification of the Notch modulator LNX2 upregulates the WNT/beta-catenin pathway in colorectal cancer. *Cancer Res* 2013;73:2003–13.
39. Li Z, Han K, Pak JE, et al. Recognition of EGF-like domains by the Notch-modifying O-fucosyltransferase POFUT1. *Nat Chem Biol* 2017;13:757–63.
40. Shi S, Stanley P. Protein O-fucosyltransferase 1 is an essential component of Notch signaling pathways. *Proc Natl Acad Sci U S A* 2003;100:5234–9.
41. Du Y, Li D, Li N, et al. POFUT1 promotes colorectal cancer development through the activation of Notch1 signaling. *Cell Death Dis* 2018;9:995.
42. van Es JH, van Gijn ME, Riccio O, et al. Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature* 2005;435:959–63.
43. Lu D, Wu Y, Wang Y, et al. CREPT accelerates tumorigenesis by regulating the transcription of cell-cycle-related genes. *Cancer Cell* 2012;21:92–104.
44. Zhang Y, Liu C, Duan X, et al. CREPT/RPRD1B, a recently identified novel protein highly expressed in tumors, enhances the beta-catenin. TCF4 transcriptional activity in response to Wnt signaling. *J Biol Chem* 2014;289:22589–99.
45. Golob-Schwarzl N, Schweiger C, Koller C, et al. Separation of low and high grade colon and rectum carcinoma by eukaryotic translation initiation factors 1, 5 and 6. *Oncotarget* 2017;8:101224–43.
46. Biffo S, Manfrini N, Ricciardi S. Crosstalks between translation and metabolism in cancer. *Curr Opin Genet Dev* 2018;48:75–81.
47. Rosso P, Cortesina G, Sanvito F, et al. Overexpression of p27BBP in head and neck carcinomas and their lymph node metastases. *Head Neck* 2004;26:408–17.
48. Miluzio A, Oliveto S, Pesce E, et al. Expression and activity of eIF6 trigger malignant pleural mesothelioma growth in vivo. *Oncotarget* 2015;6:37471–85.
49. Sanvito F, Vivoli F, Gambini S, et al. Expression of a highly conserved protein, p27BBP, during the progression of human colorectal cancer. *Cancer Res* 2000;60:510–6.
50. Pinzaglia M, Montaldo C, Polinari D, et al. eIF6 over-expression increases the motility and invasiveness of cancer cells by modulating the expression of a critical subset of membrane-bound proteins. *BMC Cancer* 2015;15:131.
51. Benelli D, Cialfi S, Pinzaglia M, et al. The translation factor eIF6 is a Notch-dependent regulator of cell migration and invasion. *PLoS One* 2012;7:e32047.
52. Hassan C, Quintero E, Dumonceau JM, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) guideline. *Endoscopy* 2013;45:842–51.
53. Pan J, Xin L, Ma YF, et al. Colonoscopy reduces colorectal cancer incidence and mortality in patients with non-malignant findings: a meta-analysis. *Am J Gastroenterol* 2016;111:355–65.
54. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell* 1996;87:159–70.

Appendix

NGS-ProToCol consortium members

Natasja Dits, Department of Urology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands.

René Böttcher, Department of Urology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands.

Annemieke C. Hiemstra, Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands.

Bauke Ylstra, Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.

Daoud Sie, Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.

Evert van den Broek, Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands.

Nicole van Grieken, Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.

David van der Meer, GenomeScan, Leiden, The Netherlands.

Floor Pepers, GenomeScan, Leiden, The Netherlands.

Eric Caldenhoven, Lygature, Utrecht, The Netherlands.

Bart Janssen, GenomeScan, Leiden, The Netherlands.

Wilbert van Workum, GenomeScan, Leiden, The Netherlands.

Stef van Lieshout, Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.

Chris H. Bangma, Department of Urology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands.

Geert van Leenders, Department of Pathology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands.

Harmen J.G. van de Werken, Department of Urology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands; and Computational Biology Center, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands.

SUPPLEMENTARY MATERIALS AND METHODS

SAMPLE PREPARATION FOR MASS SPECTROMETRY PROTEOMICS

Proteins were isolated from snap-frozen tissue samples. Twenty tissue sections of 16 µm thickness were lysed in 30µl of reducing sample buffer per mg of tissue (NuPAGE™ lithium dodecyl sulfate (LDS) Sample Buffer, supplemented with DTT 0.1M, Thermo Fisher, Bremen, Germany), thoroughly vortexed for 1 minute, heated at 70°C for 10 minutes and sonicated (3 times 20 seconds with 20 second intervals). Lysates were centrifuged for 10 minutes at 20,000 x g upon which supernatants were transferred to a new tube.

Equal amounts of proteins (~50µg) were loaded from each sample using block randomization across the three patient groups and gels (equal number of normal adjacent colon, adenoma and CRC samples in each gel). Proteins were separated on precast 4-12% gradient SDS-PAGE gels (Invitrogen, Carlsbad, USA). The gels were fixed in 50% ethanol containing 3% phosphoric acid, washed and stained overnight with Coomassie R-250. Gels were washed in ultrapure water (Merck Millipore, Billerica, MA, USA) and stored at 4°C until further processing. Each sample lane was cut from the gel as a single band and subjected to protein digestion as previously described[1, 2]. Peptides were extracted and the volume of the peptide fractions was reduced to 50 µl in a vacuum centrifuge to eliminate ACN from the solution. Peptide extracts were then desalted as an extra cleanup step using Oasis HLB cartridges (Waters Chromatography B.V, Etten-Leur, The Netherlands). Peptide eluates were dried in a vacuum centrifuge and re-dissolved in 4% acetonitrile + 0.5% Trifluoroacetic acid + 0.02% retention time peptides (iRT, Biognosys, Schlieren, Switzerland). Peptides were separated by an Ultimate 3000 nanoLC system (Dionex LC-Packings, Amsterdam, The Netherlands), equipped with a 50 cm x 75 µm ID fused silica column custom packed with 1.9 µm 120 Å ReproSil Pur C18 aqua (Dr Maisch GMBH, Ammerbuch,Entringen, Germany), as described previously[2]. After injection, peptides were trapped at 6 µl/min on a 10 mm × 100 µm ID trap column packed with 5 µm 120 Å ReproSil Pur C18 aqua at 2% buffer B (buffer A: 0.5% acetic acid in ultrapure water; buffer B: 80% ACN + 0.5% acetic acid in ultrapure water) and separated at 300 nl/min in a 10–40% buffer B linear gradient in 90 min (120 min inject-to-inject). LC-MS/MS runs were performed using block randomization across sample types, injecting in alternating order of normal colon-adenoma-cancer samples. Any instrument performance drift was equally distributed over all sample groups, thereby minimizing group bias by experiment design[3]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[4] partner repository with the accession identifier PXD012254.

RNA SEQUENCING DATA ANALYSIS

Differential gene expression analysis was performed between high-risk and low-risk adenomas using DESeq2[5]. Differentially expressed genes were obtained with the following filtering; absolute log2 fold change≥0.6 and adjusted p-value≤0.05. Regularized logarithmic transformation of expression values was performed, Euclidian distance between samples was calculated and the RNA expression data was visualized using the multidimensional scaling algorithm. Additionally, normalized counts as well as RPKMs were obtained for the whole expression matrix. Cellular decomposition analysis was performed using ESTIMATE[6] algorithm on the RPKM expression-based matrix. Group comparison of the results was performed with the Mann-Whitney test and p-values were obtained.

Proteomics data analysis

Separate searches were performed for differential expression analysis between high-risk and low-risk adenomas, and for obtaining protein expression matrix across all sample types (see Supplementary Table 1 for the Spectronaut parameters). Quality control was performed by comparing the total number of protein groups identified in each sample and by the multidimensional scaling algorithm on Euclidian distance of protein expression profiles. A sample was considered an outlier if its number of total protein groups was below the range of the average number of protein groups identified per sample in the whole dataset +/- 2 standard deviations of the number of protein groups identified per sample in the whole dataset, and was removed from the proteomic dataset. Protein group intensities were log2 transformed and median-centering normalization was performed. Differentially expressed proteins between high-risk and low-risk adenomas were identified using limma[7], Benjamini-Hochberg correction was used to calculate p-value adjusted for multiple hypothesis testing (absolute log2 fold change≥0.6 and p-value≤0.05). Euclidian distance and multidimensional scaling algorithm were used for data visualization.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed after differential expression analysis between high-risk and low-risk adenomas on both RNA and protein level. Genes were ranked according to log10 transformed p-value with the sign opposite to the log2 fold change (-sign(log2FC)*log10(p-value)). The ranked list was submitted to GSEA[8] and the collections of hallmark gene sets from Molecular Signature Database v6.0 (MSigDB) were used[9]. Significant gene sets were extracted based on an FDR threshold of ≤0.15. Single sample GSEA (ssGSEA) analysis was performed only for the RNA expression data using GSVA package[10] on normalized counts for adenomas and cancer samples, using the collection of hallmark gene sets from MSigDB[9]. P-values for group comparison were obtained using Mann-Whitney test.

DNA-RNA-protein correlation analysis for gene-dosage effect identification

Pairwise Spearman correlations were calculated between DNA copy number, RNA and protein expression for the genes occurring in all three datasets. The analysis was performed for the adenomas and the cancers separately. On DNA level, for each gene a segment value of its chromosomal location was assigned. On RNA level normalized counts were used and on protein level log2 transformed protein group intensities. Correlation coefficients (R_S) and p-values were obtained. FDR was calculated with the Benjamini-Hochberg method. Significant correlations were identified, when correlation coefficient was ≥ 0.5 and FDR values were either ≤0.25 or ≤0.1 in all 3 pairwise comparisons for adenomas and cancers, respectively. Pearson correlation was also calculated between the genes of interest *EIF6, POFUT1* and *RPRD1B* for each experiment and correlation matrix was plotted using "PerformanceAnalytics" R package.

Validation of gene-dosage effect was performed using cBioPortal (https://www.cbioportal.org/) TCGA Provisional dataset of colorectal adenocarcinomas[11, 12]. DNA copy number (n = 616), RNA expression from RNA sequencing (n = 382) and protein expression measured by mass spectrometry proteomics (n = 90) were used. In the plot tab of cBioPortal RNA or Protein expression for each gene, POFUT1, EIF6 or RPRD1B, were plotted against the available GISTIC copy number of the same gene; deep deletion, diploid, gain, amplification. For group comparisons p-values were obtained with the use of Mann-Whitney test.

Quantification of tumor-stroma and goblet cells

Digital images were analyzed with HALO software (v2.1, Indica Labs) to accurately determine tumor-stroma percentage. Random Forest classifier was used with the classes "epithelium", "stroma" and "lumen". The classifier was trained on the manually selected areas representing each class and applied on the whole tissue section. The results were evaluated by an expert pathologist. Only the stroma and epithelium areas were extracted from the results for further analysis, while the lumen area was excluded. Tumor-stroma percentage was obtained by dividing the stroma area by the sum of stroma and epithelium area.

The proportion of goblet cells in each sample was estimated by an expert pathologist, using the H&E-stained slides. The categories were few (0-20%), moderate (21-50%) and many goblet cells (more than 50%).

Immunohistochemical staining of tissue microarrays and patient-derived colorectal adenoma organoids

Tissue microarrays (TMA) for NGS-ProToCol samples were obtained as described previously[13]. Briefly, three tissue core biopsies of 0.6 mm in diameter were punched from morphologically representative areas of the Formalin-fixed, paraffin-embedded (FFPE) donor blocks and transferred into TMA recipient paraffin blocks using 3DHISTECH TMA Master (v1.14, #dHISTECH Ltd., Budapest, Hungary). The TMA represent 29 CRCs, 9 high-risk adenomas, 14 low-risk adenomas and 24 normal adjacent colon samples. TMA sections (4 µm) were deparaffinized by xylene and rehydrated with a decreasing alcohol series. TMAs were stained using HPA antibodies directed against POFUT1 (HPA054519) and RPRD1B (HPA066290).

For POFUT1 staining, antigen retrieval was performed by microwave heating in citric acid (10mM, pH6.0) and endogenous peroxidase quenching in 0.3% H_2O_2/methanol (30 minutes). Primary rabbit polyclonal monospecific antibody directed against human POFUT1 (1:75, 1hour, Atlas Antibodies, Stockholm, Sweden) was incubated at room temperature for 1 hour. For RPRD1B staining, antigen retrieval was performed by autoclave heating in citric acid (10mM, pH6.0) and endogenous peroxidase quenching in 0.3% H_2O_2/methanol (30 minutes). Primary rabbit polyclonal monospecific antibody directed against human RPRD1B (1:600, overnight, Atlas Antibodies, Stockholm, Sweden) was incubated at 4^oC overnight. Secondary anti-rabbit antibodies (BrightVision, Immunologic, Duiven, The Netherlands) were incubated for 30 minutes at room temperature. Secondary antibodies were visualized by liquid diaminobenzidine substrate chromogen system. Incubation without primary antibody served as negative control. TMAs were scanned using Aperio AT2 Scanner (Leica Biosystems Imaging). TMA scoring was performed with the use of Slide Score (www.slidescore.com) by an expert pathologist. For POFUT1, staining intensity (negative=0, weak=1, moderate=2 or strong=3) and percentage of the area stained (0-100%) were scored in the cytoplasm of epithelial cells within each tissue core. For RPRD1B, staining intensity (negative=0, weak=1, moderate=2 or strong=3) and percentage of the area stained (0-100%) were scored in the nuclei of epithelial cells within each tissue core. Protein expression values were obtained by multiplying staining intensity by the percentage of the area stained. Maximum values were selected from the replicate cores per sample. ROC analysis with Youden statistics were used to define the best threshold to distinguish cases from controls[14], expression values were dichotomized based on the best threshold and p-values were obtained with Fisher exact test.

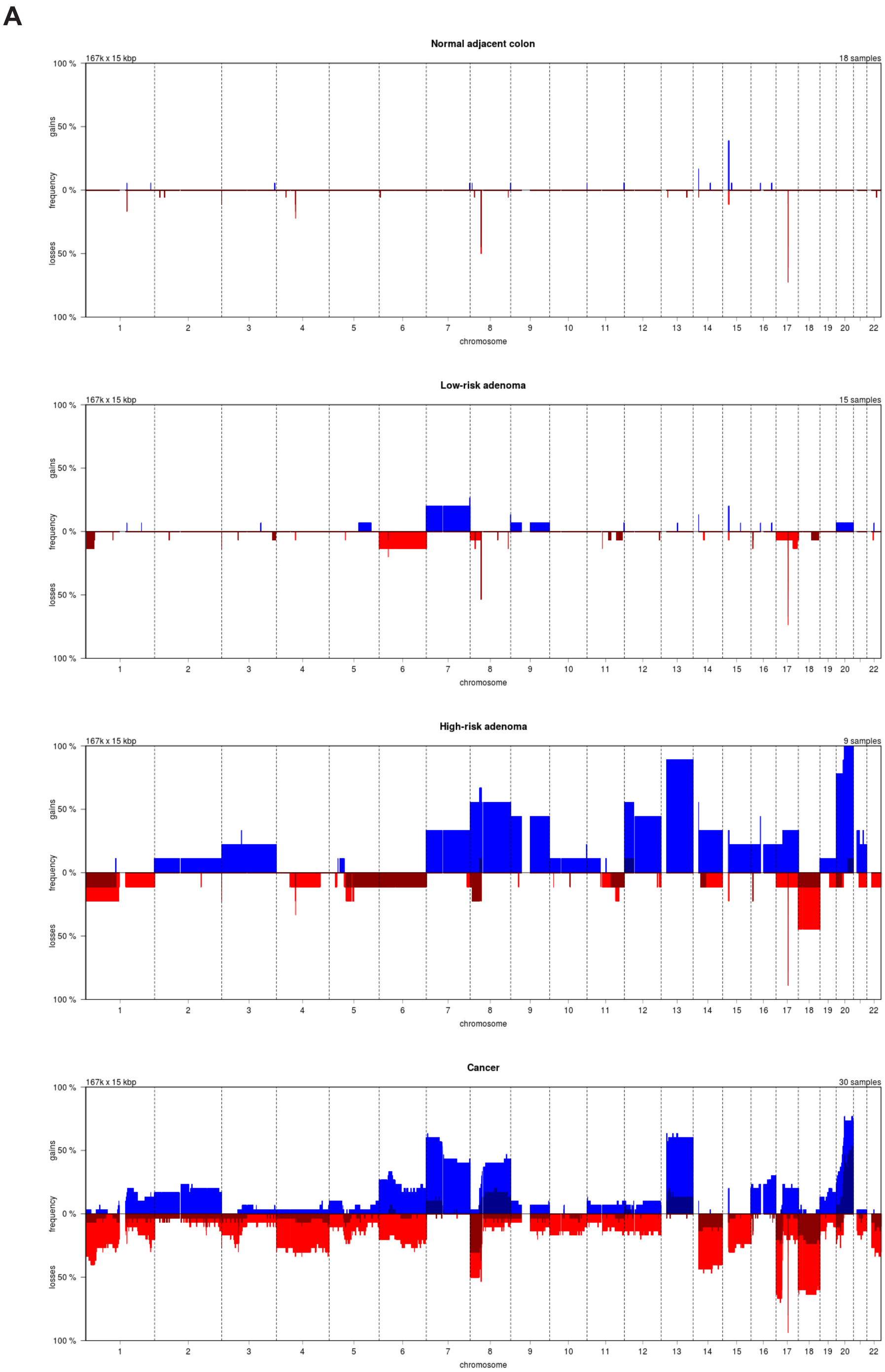
Colorectal adenoma-derived organoids were obtained from individuals participating in the Dutch colorectal cancer screening program, who underwent a colonoscopy procedure. All participants gave informed consent for the establishment of organoids and their use in molecular research. Colorectal organoids were obtained and cultured as described previously[15] with a few modifications. Organoids were cultured in Matrigel® Growth Factor Reduced Basement Membrane Matrix, Phenol Red-Free (Corning) and passaged approximately once per week by enzymatic digestion using TrypLE™ Express Enzyme ((1X), phenol red., 12605010, Thermo Fisher, Bremen, Germany) or by mechanical disruption by stretched glass pipets. Droplets of organoid-containing Matrigel, dispensed in pre-warmed (37 °C) 24-wells culture plates, were overlain with complete crypt culture medium (Advanced DMEM/F12 (Invitrogen) containing growth factors 20% R-Spondin conditioned medium, 10% Noggin conditioned medium, 1× B27(Invitrogen), 1,25 mM n-Acetyl Cysteine(Sigma), 10 mM Nicotinamide(Sigma), 50 ng/ml human EGF(Peprotech), 10 nM Gastrin(Sigma), 500 nM A83-01(Tocris), 3 µM SB202190(Cayman Chemicals), 10 nM Prostaglandin E2(Sigma), and 100 µg/ml Primocin(Invitrogen), see [15, 16]). All conditioned media were produced in-house from the cell lines 293T-HA-RspoI-Fc, HEK293-mNoggin-Fc and L-Wnt3a, as described previously [17]. These cell lines were kindly provided by Prof. Kuo (293T-HA-RspoI-Fc) from the Leland Stanford Junior University, Palo Alto, USA and by Prof. Clevers (HEK293-mNoggin-Fc and L-Wnt3a) from the Hubrecht Institute, Utrecht, the Netherlands. Genomic DNA was isolated from cell pellets using the ReliaPrep™ gDNA Tissue Miniprep System (Z6011, Promega), according to the manufacturers protocol Standard Protocol for Animal Tissue. Low-coverage whole genome sequencing and DNA copy number aberration identification was performed, as described previously[18].

To perform the immunohistochemical staining of the organoids, the organoid sections were prepared. For each organoid, pellets were prepared by spinning down for 5 minutes at 2000 rpm, the supernatant was discarded and the pellet was washed with cold phosphate-buffered saline (PBS). The pellet was then fixated overnight at 4^{o}C with formaldehyde and afterwards mounted in a gel matrix using Cytoblock system (Thermo Fisher, Bremen, Germany). The cell suspension was transferred to a transit cassette and imbedded in paraffin. Sections were made from the paraffin blocks. Organoids were first incubated for 5 minutes in cold PBS, to remove traces of Cultrex matrix, and then stained using antibodies directed against POFUT1 (HPA054519) and RPRD1B (HPA066290) as performed for the TMAs. Scoring and statistical analysis was performed as for the TMAs.

REFERENCES

1. Piersma SR, Warmoes MO, de Wit M, et al. Whole gel processing procedure for GeLC-MS/MS based proteomics. Proteome Sci 2013;11:17.
2. Piersma SR, Fiedler U, Span S, et al. Workflow comparison for label-free, quantitative secretome proteomics for cancer biomarker discovery: method evaluation, differential analysis, and verification in serum. J Proteome Res 2010;9:1913-22.
3. Pham TV, Piersma SR, Oudgenoeg G, et al. Label-free mass spectrometry-based proteomics for biomarker discovery and validation. Expert Rev Mol Diagn 2012;12:343-59.
4. Vizcaino JA, Csordas A, del-Toro N, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res 2016;44:D447-56.
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
6. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications 2013;4:2612.
7. 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.
8. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
9. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems 2015;1:417-425.
10. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 2013;14:7.
11. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401-4.
12. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
13. Goos JA, Coupe VM, Diosdado B, et al. Aurora kinase A (AURKA) expression in colorectal cancer liver metastasis is associated with poor prognosis. Br J Cancer 2013;109:2445-52.
14. 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.
15. Sato T, Vries RG, Snippert HJ, et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 2009;459:262-5.
16. van de Wetering M, Francies Hayley E, Francis Joshua M, et al. Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. Cell 2015;161:933-945.
17. Jung P, Sato T, Merlos-Suarez A, et al. Isolation and in vitro expansion of human colonic stem cells. Nat Med 2011;17:1225-7.
18. Carvalho B, Diosdado B, Terhaar Sive Droste JS, et al. Evaluation of Cancer-Associated DNA Copy Number Events in Colorectal (Advanced) Adenomas. Cancer Prev Res (Phila) 2018;11:403-412.

SUPPLEMENTARY FIGURES

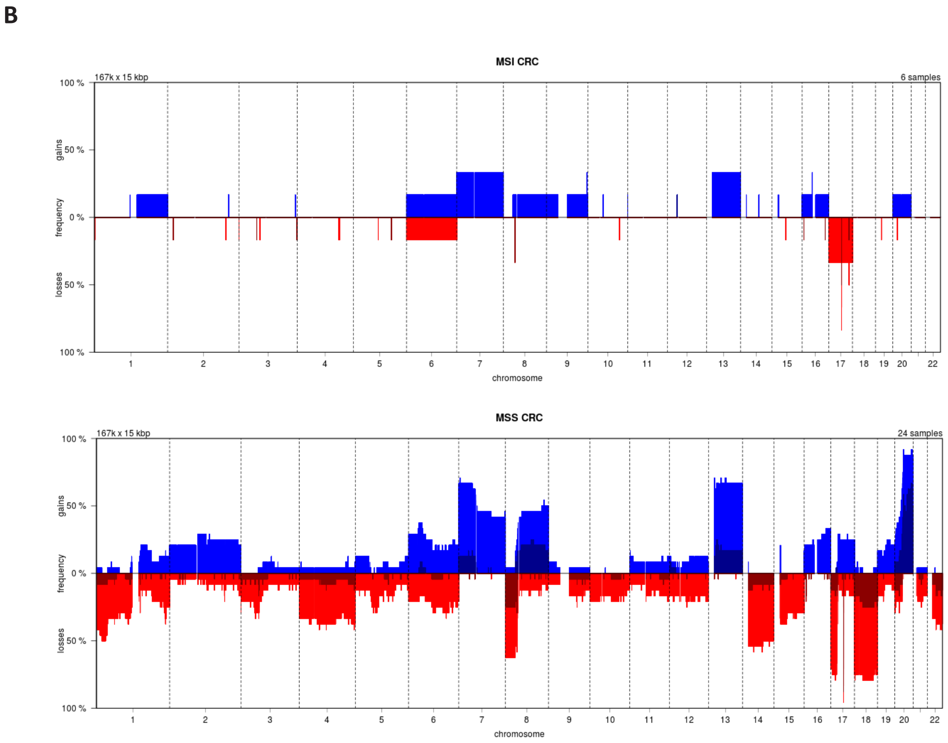


Normal adjacent colon

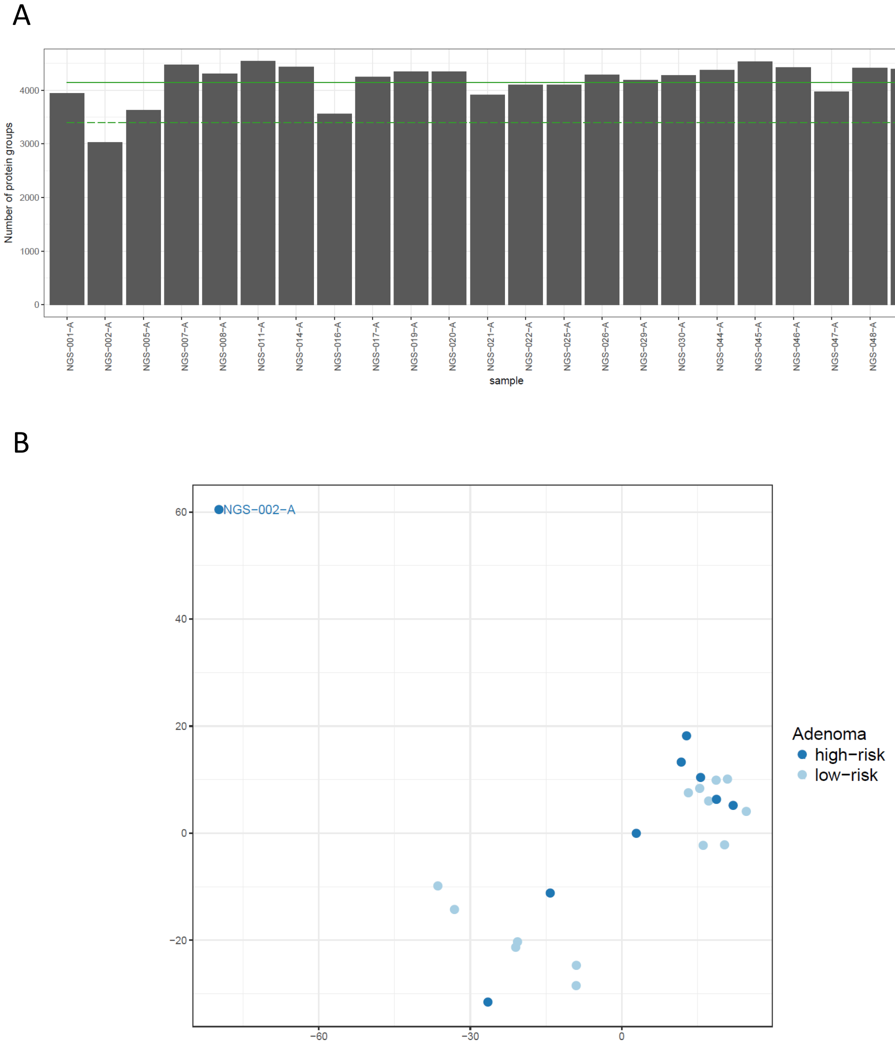
Low-risk adenoma

High-risk adenoma

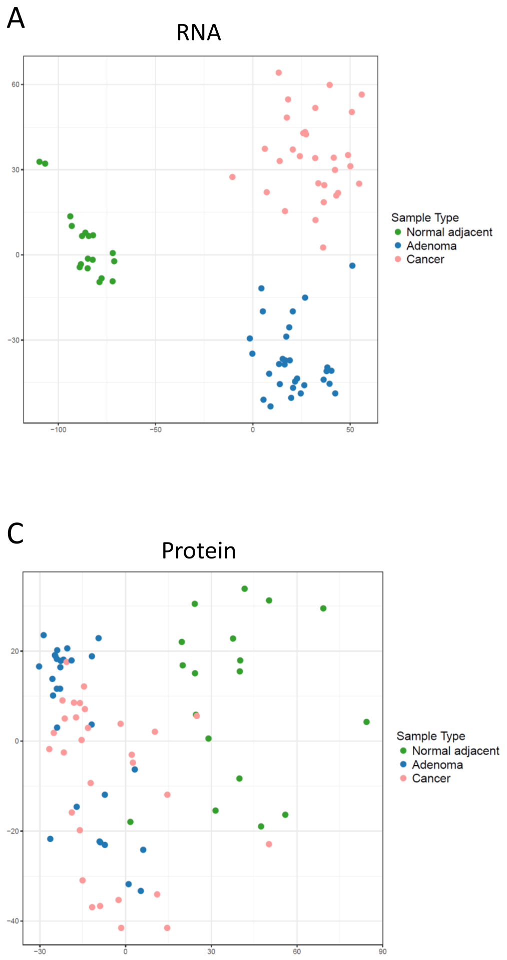
Cancer



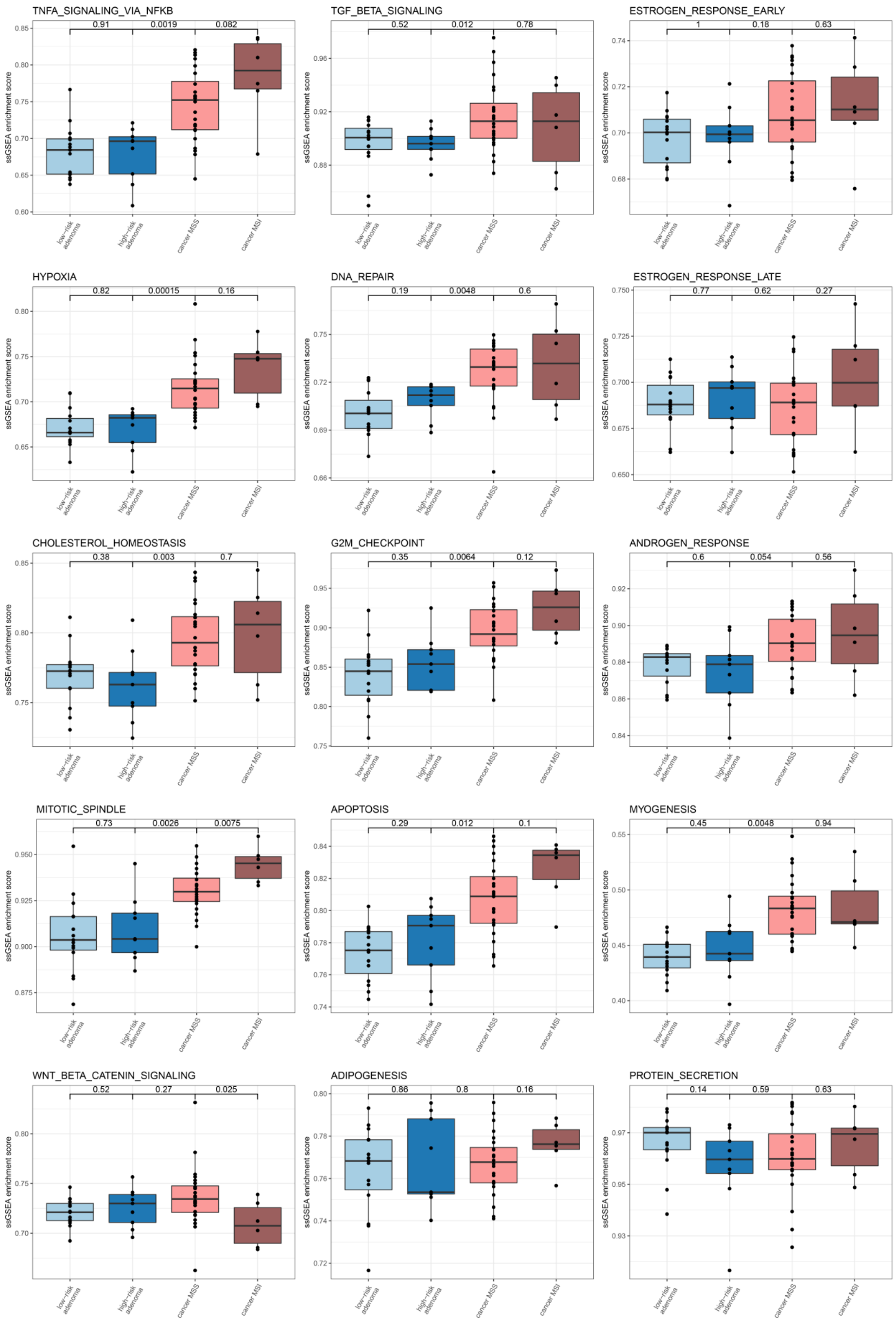
Supplementary Figure 1. Frequency plots of DNA copy number aberrations. **A.** From top, normal adjacent colon (n=18), low-risk adenomas (n=15), high-risk adenomas (n=9) and cancers (n=30) **B.** MSI cancers (n=6) and MSS cancers (n=24).

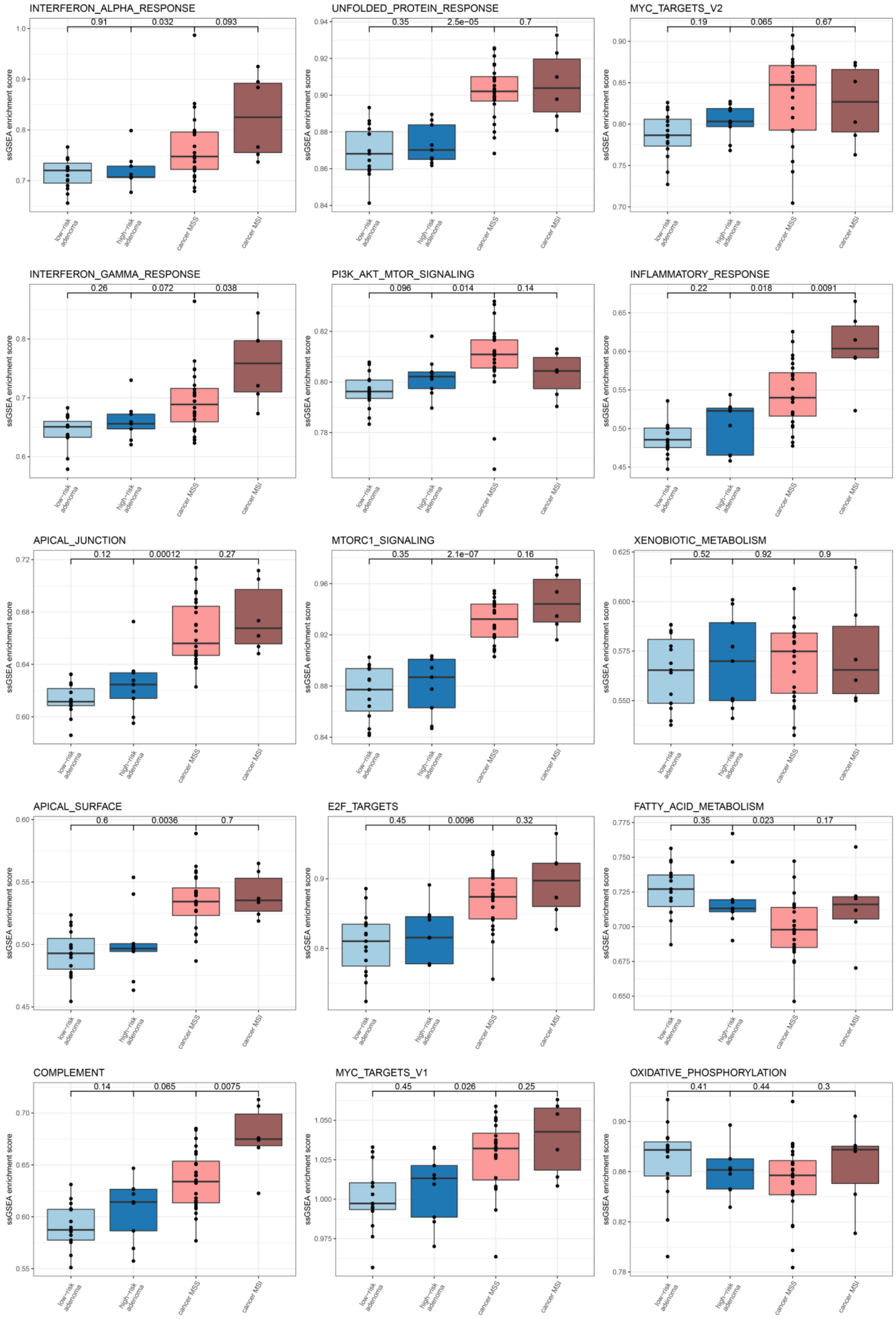


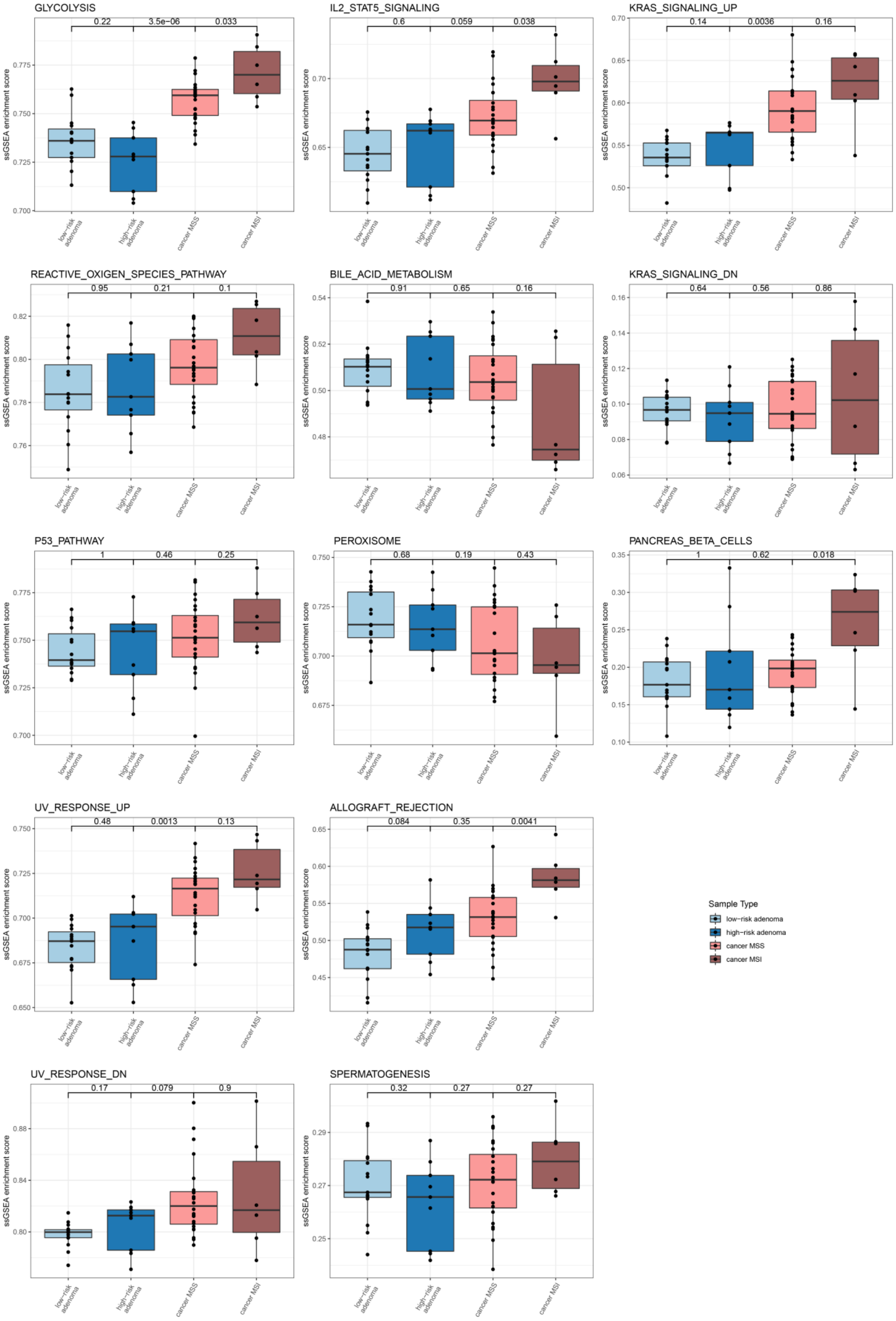
Supplementary Figure 2. Quality assessment of proteomics data. **A**. Number of protein groups identified in each adenoma sample. Mean number of protein groups is highlighted with green solid line, mean minus 2 standard deviations of the number of protein groups is highlighted with green dashed line. The outlier sample (NGS-002-A) is the only one with the number of protein groups below the dashed line. **B.** Multidimensional scaling of adenoma samples based on the protein expression. The same outlier is highlighted.



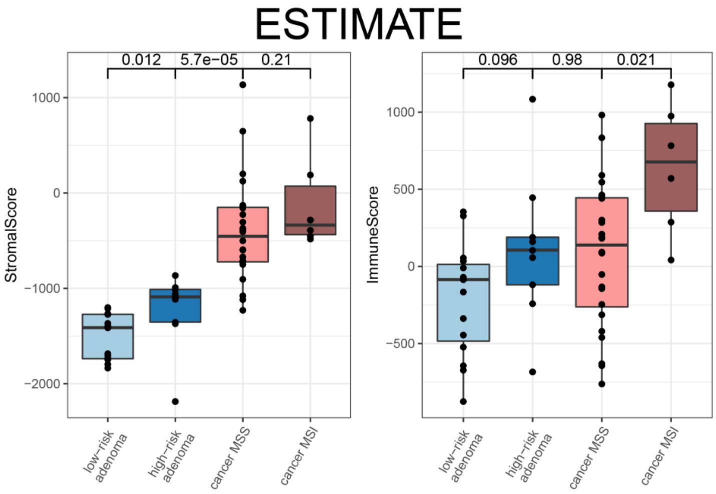
Supplementary Figure 3. Multidimensional scaling of the RNA and protein expression profiles. **A**. Visualization of normal adjacent colon samples (green), adenomas (blue) and cancers (pink) based on the RNA expression profile. **B.** Visualization of high-risk (dark blue) and low-risk adenomas (light blue) based on the RNA expression profile. **C.** Visualization of normal adjacent colon samples, adenomas and cancers based on the protein expression profile. **D.** Visualization of high-risk and low-risk adenomas based on the protein expression profile.


TNFA_SIGNALING_VIA_NFKB
0.91
0.0019
0.082
0.85
0.80
0.75
0.70
0.65
0.60
TGF_BETA_SIGNALING
0.52
0.012
0.78
0.96
0.92
0.88
ESTROGEN_RESPONSE_EARLY
1
0.18
0.63
0.74
0.72
0.70
0.68
HYPOXIA
0.82
0.00015
0.16
0.80
0.75
0.70
0.65
DNA_REPAIR
0.19
0.0048
0.6
0.75
0.72
0.69
0.66
ESTROGEN_RESPONSE_LATE
0.77
0.62
0.27
0.750
0.725
0.700
0.675
0.650
CHOLESTEROL_HOMEOSTASIS
0.38
0.003
0.7
0.85
0.80
0.75
G2M_CHECKPOINT
0.35
0.0064
0.12
0.95
0.90
0.85
0.80
0.75
ANDROGEN_RESPONSE
0.6
0.054
0.56
0.92
0.90
0.88
0.86
0.84
MITOTIC_SPINDLE
0.73
0.0026
0.0075
0.950
0.925
0.900
0.875
APOPTOSIS
0.29
0.012
0.1
0.84
0.82
0.80
0.78
0.76
0.74
MYOGENESIS
0.45
0.0048
0.94
0.55
0.50
0.45
0.40
WNT_BETA_CATENIN_SIGNALING
0.52
0.27
0.025
0.80
0.75
0.70
ADIPOGENESIS
0.86
0.8
0.16
0.80
0.78
0.76
0.74
0.72
PROTEIN_SECRETION
0.14
0.59
0.63
0.97
0.95
0.93
ssGSEA enrichment score
low-risk adenoma
high-risk adenoma
cancer MSS
cancer MSI

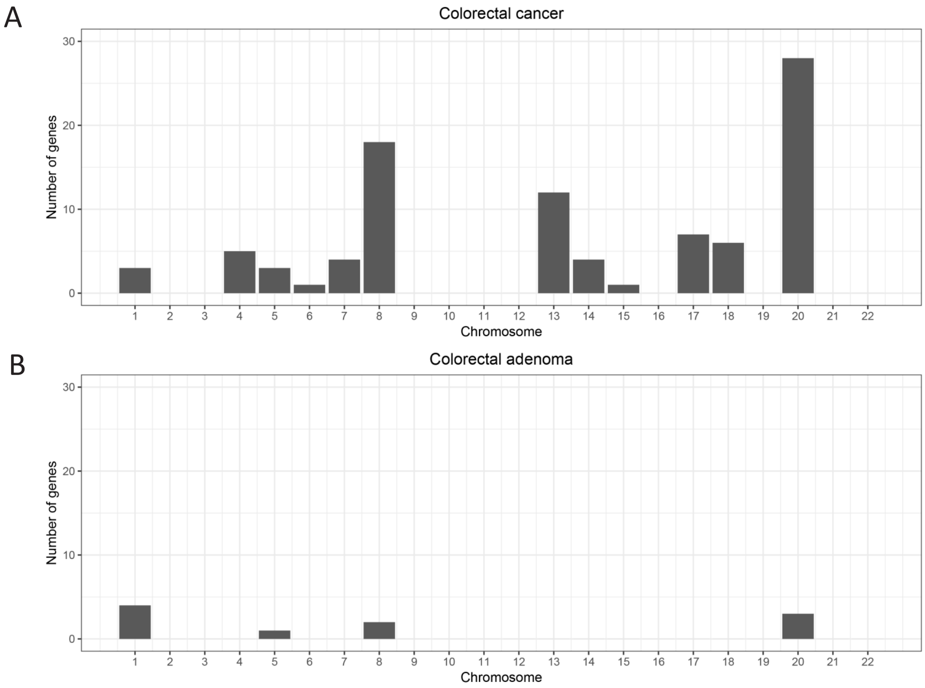

INTERFERON_ALPHA_RESPONSE
0.91
0.032
0.093
UNFOLDED_PROTEIN_RESPONSE
0.35
2.5e−05
0.7
MYC_TARGETS_V2
0.19
0.065
0.67
INTERFERON_GAMMA_RESPONSE
0.26
0.072
0.038
PI3K_AKT_MTOR_SIGNALING
0.096
0.014
0.14
INFLAMMATORY_RESPONSE
0.22
0.018
0.0091
APICAL_JUNCTION
0.12
0.00012
0.27
MTORC1_SIGNALING
0.35
2.1e−07
0.16
XENOBIOTIC_METABOLISM
0.52
0.92
0.9
APICAL_SURFACE
0.6
0.0036
0.7
E2F_TARGETS
0.45
0.0096
0.32
FATTY_ACID_METABOLISM
0.35
0.023
0.17
COMPLEMENT
0.14
0.065
0.0075
MYC_TARGETS_V1
0.45
0.026
0.25
OXIDATIVE_PHOSPHORYLATION
0.41
0.44
0.3
ssGSEA enrichment score
low−risk adenoma
high−risk adenoma
cancer MSS
cancer MSI



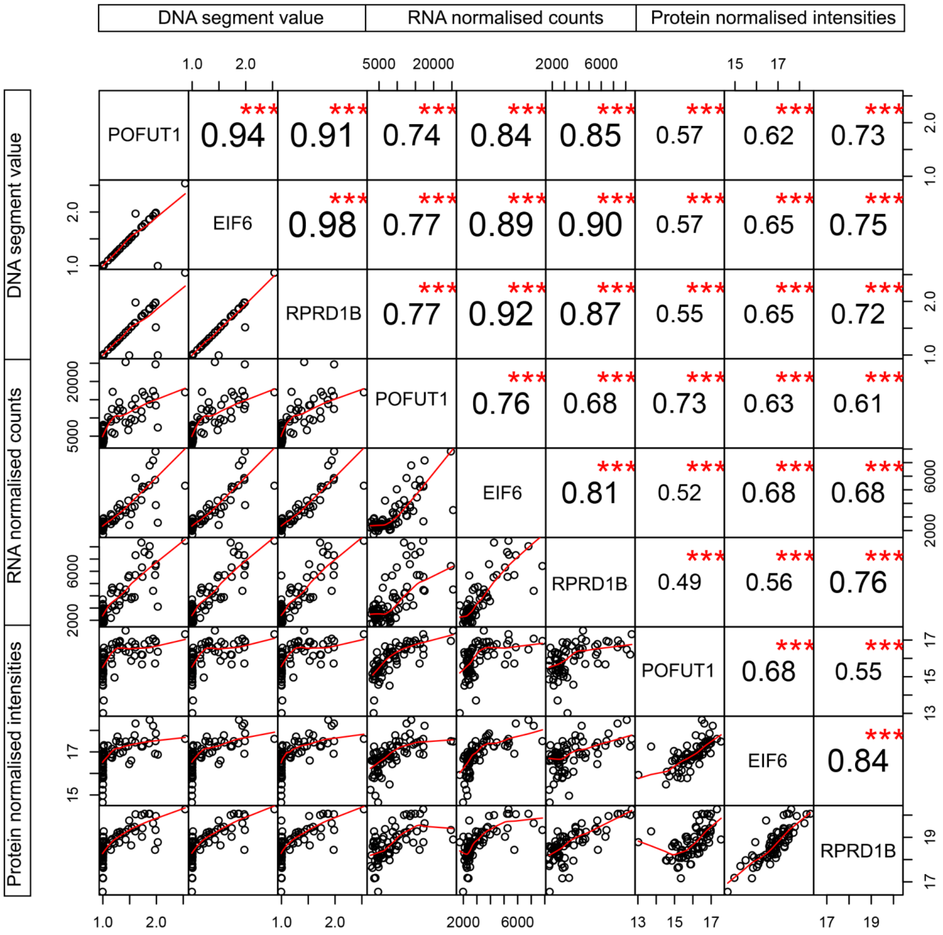
Supplementary Figure 4. Single sample GSEA analysis in low-risk adenomas, high-risk adenomas and CRCs (MSS and MSI CRCs presented separately). All gene sets with insignificant difference between low-risk and high-risk adenomas are presented based on the Mann-Whitney test.



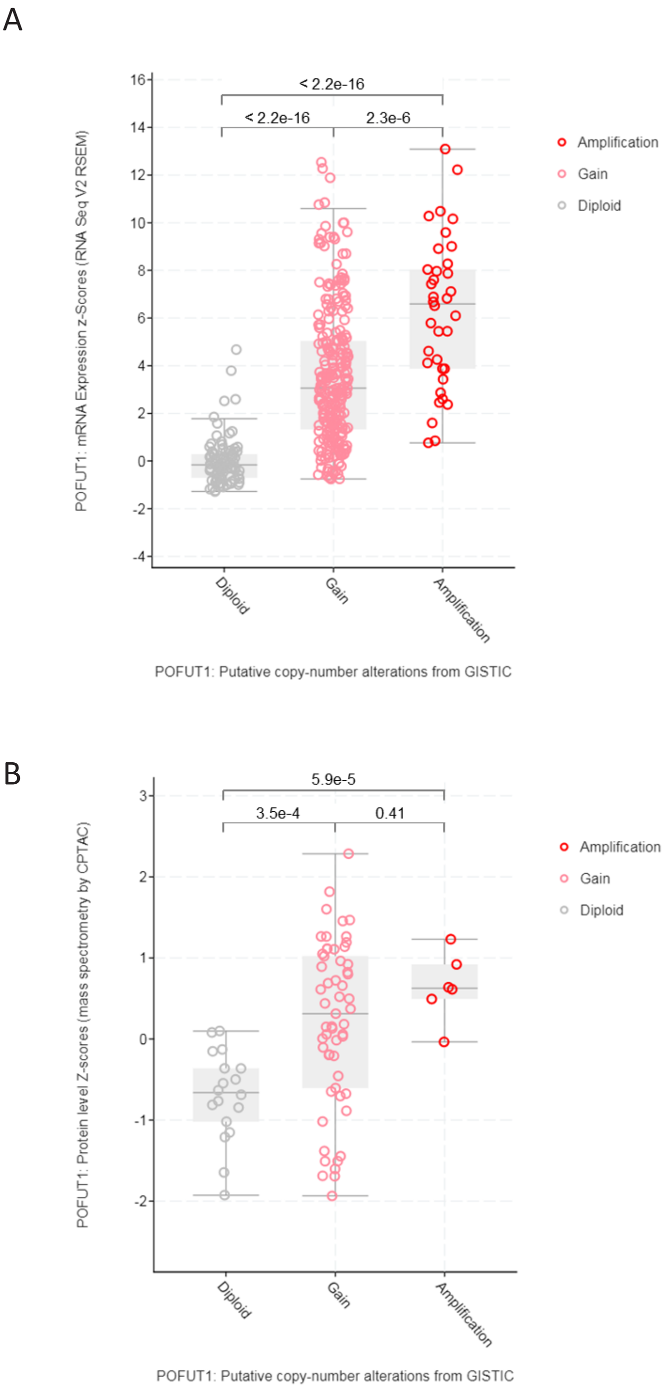
Supplementary Figure 5. Cellular decomposition based on RNA expression data. Stromal and immune enrichment scores as calculated by the ESTIMATE algorithm in low-risk and high-risk adenomas and cancers (MSS and MSI CRCs presented separately). P-values were obtained with the Mann-Whitney test.



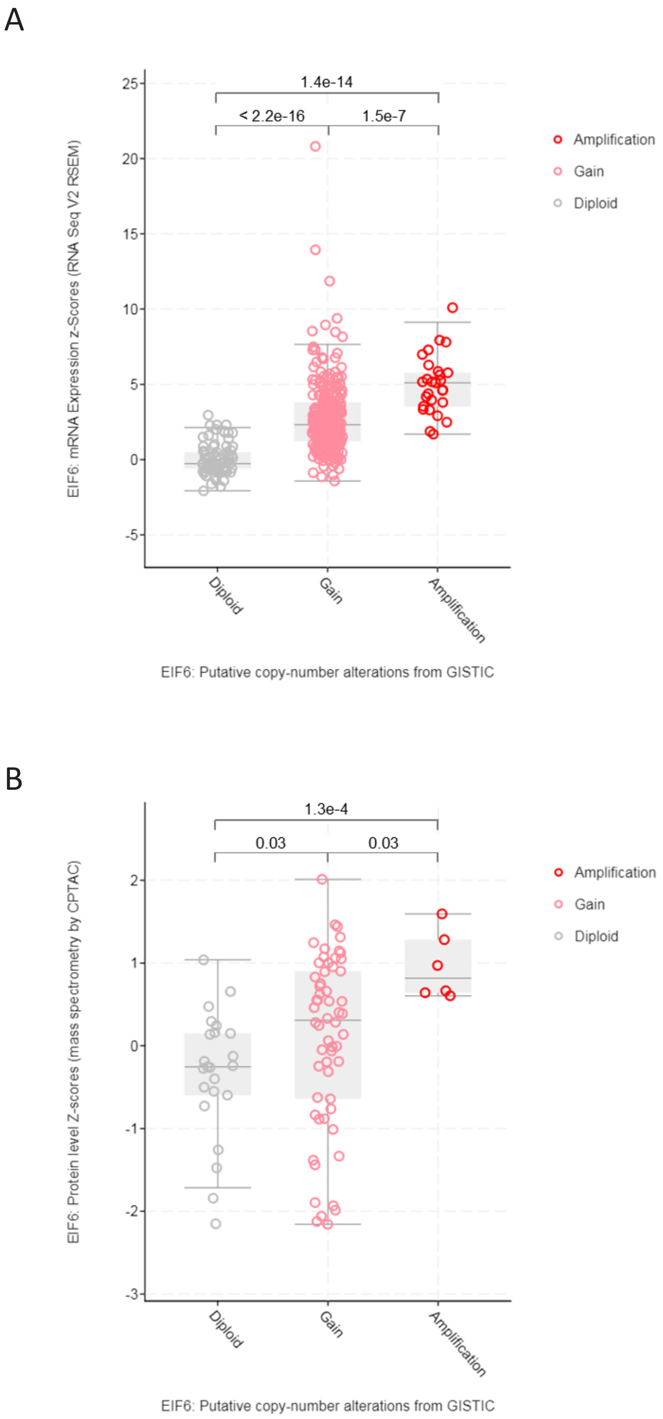
Supplementary Figure 6. DNA copy number driven gene dosage effect in cancers (**A**) and adenomas (**B**). Pairwise correlation analysis was performed between DNA copy number, RNA and protein expression. Significantly correlating genes (FDR ≤ 0.1 or 0.25 for cancers and adenomas, respectively) on DNA, RNA and protein level were identified and grouped per chromosome they reside on. The number of correlating genes was plotted per chromosome.



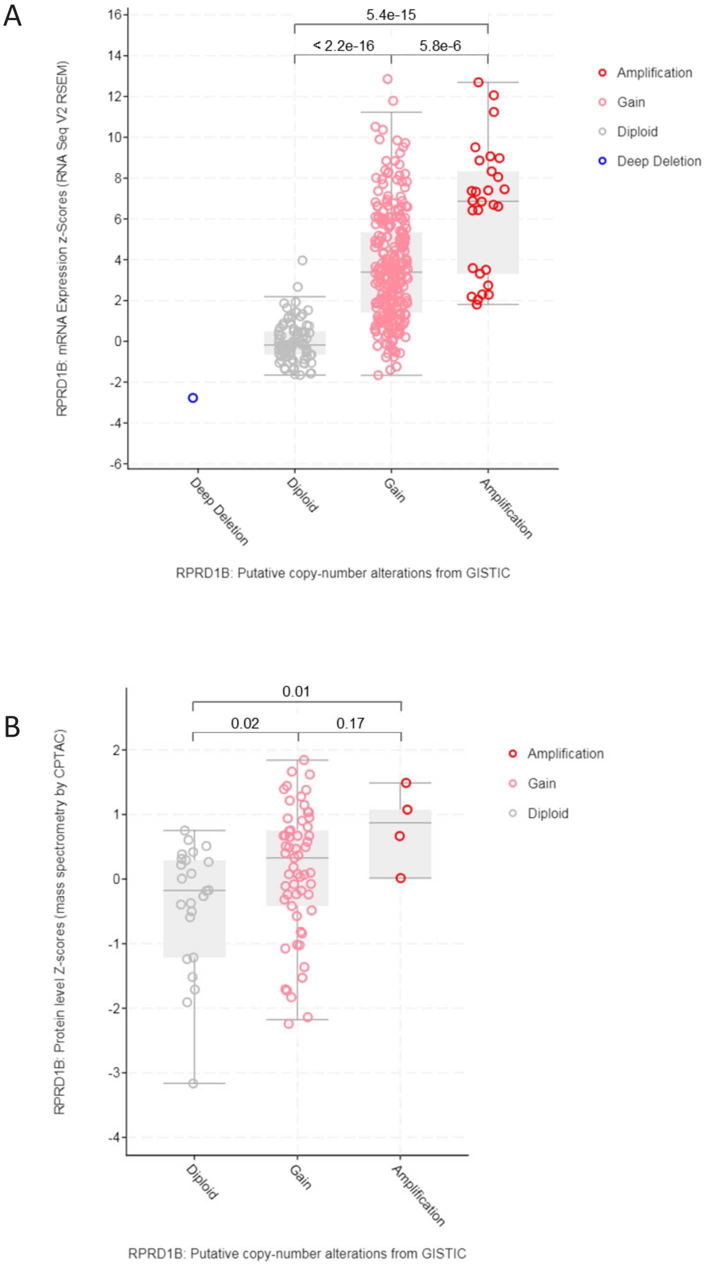
Supplementary Figure 7. Pearson correlation analysis between DNA segment value, RNA normalized counts and normalized protein intensities of EIF6, POFUT1 and RPRD1B. The correlation analysis was performed on all the samples, including normal adjacent colon, adenoma and cancer samples. Bottom left matrix presents bivariate scatter plots with a fitted line. Top right displays correlation coefficient and significance level, where "***" means p-value ≤ 0.001.



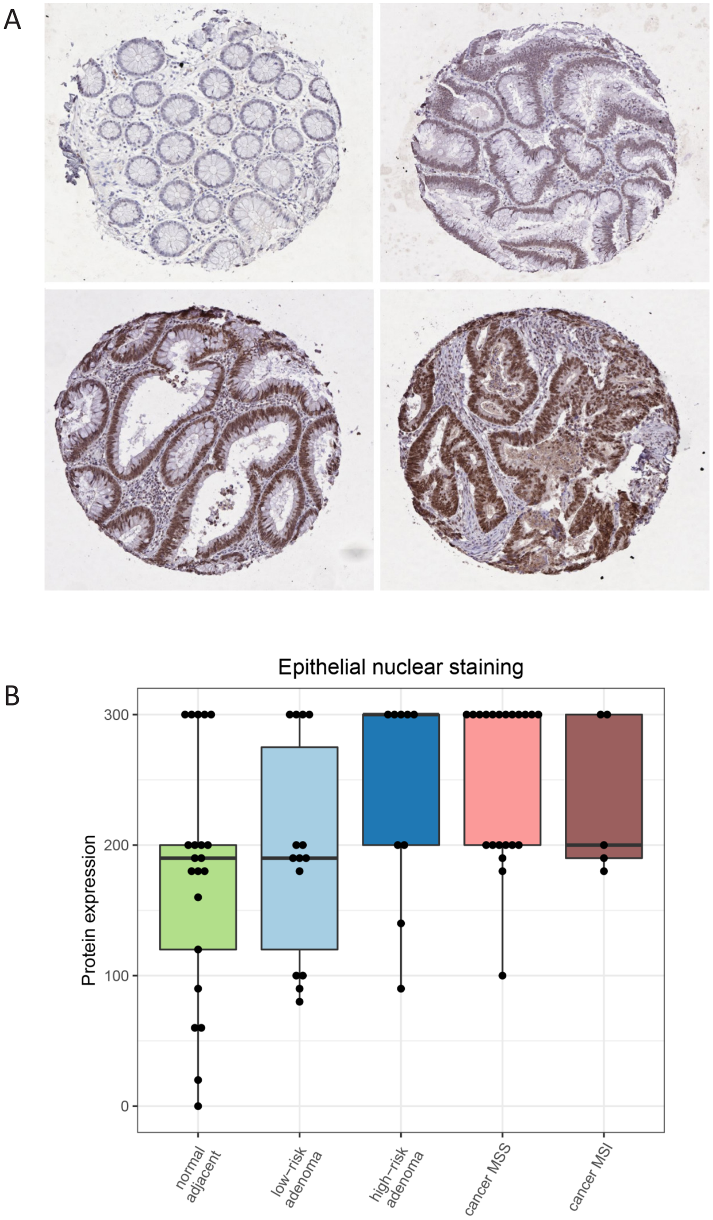
Supplementary Figure 8. DNA copy number driven gene dosage effect in the TCGA colorectal adenomacarcinomas for POFUT1 identified on RNA (A) and protein (B) level. Each dot represents a sample and is grouped according to the DNA copy number of POFUT1. RNA (A) and protein (B) expression values in a form of Z-scores are plotted per group.



Supplementary Figure 9. DNA copy number driven gene dosage effect in the TCGA colorectal adenomacarcinomas for EIF6 identified on RNA (A) and protein (B) level. Each dot represents a sample and is grouped according to the DNA copy number of EIF6. RNA (A) and protein (B) expression values in a form of Z-scores are plotted per group.



Supplementary Figure 10. DNA copy number driven gene dosage effect in the TCGA colorectal adenomacarcinomas for RPRD1B identified on RNA (A) and protein (B) level. Each dot represents a sample and is grouped according to the DNA copy number of RPRD1B. RNA (A) and protein (B) expression values in a form of Z-scores are plotted per group.



Supplementary Figure 11. Immunohistochemical staining of RPRD1B in colorectal tissues. **A.** Representative RPRD1B staining in different types of tissue samples. Top left: normal adjacent colon, top right: low-risk adenoma, bottom left: high-risk adenoma, bottom right: colorectal cancer. **B.** RPRD1B expression as measured by a product of epithelial nuclear staining intensity (negative=0, weak=1, moderate=2 or strong=3) and percentage of the cells stained positively (0-100%) was plotted for the different types of tissue samples: normal adjacent colon, low-risk adenomas, high-risk adenomas and CRCs (MSS and MSI CRCs presented separately). See Supplementary Table 10 for group comparisons and statistical testing.

SUPPLEMENTARY TABLES

Supplementary Table 2. Clinical and molecular characteristics of the study dataset. All adenomas were independent lesions without a focus of cancer.

PATIENT ID	SAMPLE ID	Sex	Age	Location	Tissue type	Microsatellite status	Risk of progression	Histology	Dysplasia	Differentiation Grade	Stage	Size (in mm)
NGS-001	NGS-001-A	Female	65	NA	colorectal adenoma	MSS	low-risk	villous	low grade			50
NGS-001	NGS-001-C	Female	65	NA	colorectal carcinoma	MSS				well/moderate	IV	70
NGS-001	NGS-001-N	Female	65	NA	normal adjacent colon							
NGS-002	NGS-002-A	Male	71	sigmoid	colorectal adenoma	MSS	high-risk	tubular	low grade			15
NGS-002	NGS-002-C	Male	71	sigmoid	colorectal carcinoma	MSS				well/moderate	III	45
NGS-002	NGS-002-N	Male	71	sigmoid	normal adjacent colon							
NGS-005	NGS-005-A	Female	65	cecum	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			20
NGS-005	NGS-005-C	Female	65	cecum	colorectal carcinoma	MSS				well/moderate	III	60
NGS-005	NGS-005-N	Female	65	cecum	normal adjacent colon							
NGS-062	NGS-062-A	Female	75	NA	colorectal adenoma	MSI		tubulovillous	high grade			10
NGS-062	NGS-062-C	Female	75	cecum	colorectal carcinoma	MSI				less/not	II	100
NGS-062	NGS-062-N	Female	75	cecum	normal adjacent colon							
NGS-016	NGS-016-A	Male	71	cecum	colorectal adenoma	MSS	high-risk	tubulovillous	high grade			45
NGS-016	NGS-069-C	Male	71	sigmoid	colorectal carcinoma	MSS				well/moderate	I	30
NGS-016	NGS-016-N	Male	71	cecum	normal adjacent colon							
NGS-061	NGS-061-A	Male	80	sigmoid	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			25
NGS-061	NGS-061-C	Male	80	NA	colorectal carcinoma	MSS				well/moderate	I or III	80
NGS-004	NGS-004-A	Male	73	sigmoid	colorectal adenoma	MSS	1 CAE	tubular	high grade			23
NGS-031	NGS-031-C	Male	69	colon descendens	colorectal carcinoma	MSS				well/moderate	I	30
NGS-007	NGS-007-A	Male	65	NA	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			40
NGS-033	NGS-033-C	Female	63	NA	colorectal carcinoma	MSS				well/moderate	II	75
NGS-020	NGS-020-A	Male	70	sigmoid	colorectal adenoma	MSS	high-risk	tubular	low grade			11
NGS-034	NGS-034-C	Female	69	colon transversum	colorectal carcinoma	MSS				well/moderate	II	45
NGS-006	NGS-006-A	Male	78	sigmoid	colorectal adenoma	MSS	No information	tubulovillous	low grade			14
NGS-035	NGS-035-C	Female	68	sigmoid	colorectal carcinoma	MSS				well/moderate	III	33
NGS-008	NGS-008-A	Male	66	sigmoid	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			20
NGS-036	NGS-036-C	Male	68	colon ascendens	colorectal carcinoma	MSI				NA	III	18
NGS-009	NGS-009-A	Male	59	rectum	colorectal adenoma	MSS	No information	villous	low grade			12
NGS-038	NGS-038-C	Male	81	NA	colorectal carcinoma	MSS				well/moderate	II	40
NGS-011	NGS-011-A	Female	72	rectum	colorectal adenoma	MSS	low-risk	tubulovillous	high grade			25
NGS-040	NGS-040-C	Female	85	colon transversum	colorectal carcinoma	MSS				well/moderate	I	30
NGS-013	NGS-013-A	Female	52	rectum	colorectal adenoma	MSS	1 CAE	villous	low grade			70
NGS-041	NGS-041-C	Male	67	sigmoid	colorectal carcinoma	MSI				well/moderate	II	35
NGS-014	NGS-014-A	Male	63	sigmoid	colorectal adenoma	MSS	high-risk	tubulovillous	high grade			35
NGS-042	NGS-042-C	Male	81	colon transversum	colorectal carcinoma	MSS				well/moderate	II	45
NGS-017	NGS-017-A	Female	60	sigmoid	colorectal adenoma	MSS	high-risk	tubulovillous	high grade			20
NGS-043	NGS-043-C	Male	81	sigmoid	colorectal carcinoma	MSS				well/moderate	II	34
NGS-019	NGS-019-A	Female	79	colon ascendens	colorectal adenoma	MSS	high-risk	tubular	low grade			15
NGS-050	NGS-050-C	Male	82	sigmoid	colorectal carcinoma	MSS				well/moderate	I	35

NGS-021	NGS-021-A	Female	72	colon transversum	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			18
NGS-051	NGS-051-C	Female	75	colon ascendens	colorectal carcinoma	MSI				less/not	III	50
NGS-022	NGS-022-A	Male	68	rectosigmoid	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			15
NGS-052	NGS-052-C	Female	89	NA	colorectal carcinoma	MSI				less/not	II	60
NGS-025	NGS-025-A	Male	79	sigmoid	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			20
NGS-054	NGS-054-C	Male	62	sigmoid	colorectal carcinoma	MSS				well/moderate	II	50
NGS-026	NGS-026-A	Female	74	rectum	colorectal adenoma	MSS	low-risk	tubulovillous	high grade			22
NGS-057	NGS-057-C	Male	60	sigmoid	colorectal carcinoma	MSS				well/moderate	IV	50
NGS-027	NGS-027-A	Male	67	colon descendens	colorectal adenoma	MSI		tubular	low grade			15
NGS-059	NGS-059-C	Female	88	sigmoid	colorectal carcinoma	MSS				well/moderate	I	35
NGS-029	NGS-029-A	Male	75	cecum	colorectal adenoma	MSS	low-risk	villous	high grade			45
NGS-060	NGS-060-C	Male	77	NA	colorectal carcinoma	MSI				less/not	II	80
NGS-030	NGS-030-A	Female	62	rectum	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			32
NGS-063	NGS-063-C	Female	67	cecum	colorectal carcinoma	MSS				well/moderate	III	55
NGS-044	NGS-044-A	Female	75	NA	colorectal adenoma	MSS	high-risk	tubulovillous	high grade			15
NGS-064	NGS-064-C	Male	75	cecum	colorectal carcinoma	MSS				well/moderate	IV	40
NGS-045	NGS-045-A	Female	81	rectum	colorectal adenoma	MSS	high-risk	tubulovillous	high grade			30
NGS-066	NGS-066-C	Female	78	cecum	colorectal carcinoma	MSS				well/moderate	II	30
NGS-046	NGS-046-A	Male	64	colon descendens	colorectal adenoma	MSS	low-risk	tubular	low grade			15
NGS-068	NGS-068-C	Male	67	rectum	colorectal carcinoma	MSS				well/moderate	I	37
NGS-047	NGS-047-A	Male	73	colon descendens	colorectal adenoma	MSS	high-risk	tubulovillous	low grade			25
NGS-070	NGS-070-C	Female	55	cecum	colorectal carcinoma	MSS				well/moderate	II	70
NGS-048	NGS-048-A	Female	51	sigmoid	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			22
NGS-071	NGS-071-C	Female	88	NA	colorectal carcinoma	MSS				well/moderate	I	60
NGS-049	NGS-049-A	Female	59	sigmoid	colorectal adenoma	MSS	low-risk	tubulovillous	low grade			21
NGS-072	NGS-072-C	Female	58	cecum	colorectal carcinoma	MSS				well/moderate	II	48
NGS-017	NGS-017-N	Female	60	sigmoid	normal adjacent colon							
NGS-030	NGS-030-N	Female	62	rectum	normal adjacent colon							
NGS-034	NGS-034-N	Female	69	colon transversum	normal adjacent colon							
NGS-051	NGS-051-N	Female	75	colon ascendens	normal adjacent colon							
NGS-052	NGS-052-N	Female	89	NA	normal adjacent colon							
NGS-072	NGS-072-N	Female	58	cecum	normal adjacent colon							
NGS-004	NGS-004-N	Male	73	sigmoid	normal adjacent colon							
NGS-014	NGS-014-N	Male	63	sigmoid	normal adjacent colon							
NGS-029	NGS-029-N	Male	75	cecum	normal adjacent colon							
NGS-031	NGS-031-N	Male	69	colon descendens	normal adjacent colon							
NGS-036	NGS-036-N	Male	68	colon ascendens	normal adjacent colon							
NGS-038	NGS-038-N	Male	81	NA	normal adjacent colon							
NGS-041	NGS-041-N	Male	67	sigmoid	normal adjacent colon							

Supplementary Table 3. Cancer associated events (CAEs) - DNA copy number aberrations and the risk of progression to cancer for the adenomas. See legend.

Legend:

Variable	Explanation
0	no copy number aberration
1	copy number aberration present
high	2 out of 7 aberrations present
low	0 out of 7 aberrations present
1 CAE	1 out of 7 aberrations present
MSI	MSI lesion, the CAEs definition does not apply
NA	not available

Sample name	Cancer associated events							Risk of progression
	8q	13q	20q	8p	15q	17p	18q	
NGS-001-A	0	0	0	0	0	0	0	low
NGS-002-A	0	1	1	0	0	0	0	high
NGS-004-A	0	1	0	0	0	0	0	1 CAE
NGS-005-A	0	0	0	0	0	0	0	low
NGS-006-A	0	NA	NA	0	0	0	NA	NA
NGS-007-A	0	0	0	0	0	0	0	low
NGS-008-A	0	0	0	0	0	0	0	low
NGS-009-A	0	NA	1	0	0	0	0	NA
NGS-011-A	0	0	0	0	0	0	0	low
NGS-013-A	0	0	0	0	0	0	1	1 CAE
NGS-014-A	1	1	1	0	0	0	1	high
NGS-016-A	1	1	1	0	0	0	1	high
NGS-017-A	0	1	1	0	0	0	0	high
NGS-019-A	1	1	1	0	0	0	0	high
NGS-020-A	1	1	1	0	0	0	0	high
NGS-021-A	0	0	0	0	0	0	0	low
NGS-022-A	0	0	0	0	0	0	0	low
NGS-025-A	0	0	0	0	0	0	0	low
NGS-026-A	0	0	0	0	0	0	0	low
NGS-027-A	0	0	0	0	0	0	0	MSI
NGS-029-A	0	0	0	0	0	0	0	low
NGS-030-A	0	0	0	0	0	0	0	low
NGS-044-A	0	0	1	1	0	0	1	high
NGS-045-A	0	0	1	1	0	0	1	high
NGS-046-A	0	0	0	0	0	0	0	low
NGS-047-A	1	1	1	0	0	0	0	high
NGS-048-A	0	0	0	0	0	0	0	low
NGS-049-A	0	0	0	0	0	0	0	low
NGS-061-A	0	0	0	0	0	0	0	low
NGS-062-A	0	0	0	0	0	0	0	MSI

Supplementary Table 4. Association between risk of progression and pathological adenoma features. P-values and odds ratios were obtained using Fisher exact test.

Comparison	**Risk of progression**		**Fisher exact test**	
	low	high	p-value	odds ratio
Dysplasia			0.099	4.6
high grade	3	5		
low grade	12	4		
Histology			0.13	6.4
tubulovillous/villous	14	6		
tubular	1	3		
Size (median = 21.5 mm)			1	0.7
≥ 21.5 mm	4	8		
< 21.5 mm	5	7		

Supplementary Table 7. Overlapping genes and proteins differentially expressed between colorectal adenomas at high and low risk of progressing to cancer.

Gene name	**Chromosomal arm**	**Gene**		**Protein**	
		log2FC	**adjusted p-value**	**log2FC**	**p-value**
S100A8	1q	1.15	0.032	1.83	0.002
S100A9	1q	1.37	0.001	1.23	0.005
SCPEP1	17q	0.69	0.029	0.78	0.002
ITGB3	17q	0.67	0.018	0.75	0.018
COL15A1	9q	1.13	0.005	0.73	0.023
NELFCD	20q	0.77	0.001	0.68	0.027
GLYCTK	3p	0.62	0.009	0.66	0.001
COL18A1	21q	0.87	0.019	0.64	0.045
HNF4A	20q	0.66	< 0.001	0.61	< 0.001
MIA3	1q	-0.73	0.018	-0.76	0.001
HSPA2	14q	-1.33	0.008	-0.95	0.023
MLPH	2q	-0.78	0.001	-1.34	0.001
CKB	14q	-0.98	0.039	-1.55	0.001
RAB27B	18q	-0.96	0.014	-2.23	0.008

Supplementary Table 8. Gene-dosage effect in CRCs. Significantly and positively correlated genes/proteins among DNA copy number, RNA and Protein expression are presented in the table.

		DNA vs RNA			RNA vs Protein			DNA vs Protein		
Gene name	**Chromosome**	**Correlation coefficient**	**p-value**	**FDR**	**Correlation coefficient**	**p-value**	**FDR**	**Correlation coefficient**	**p-value**	**FDR**
ACSS1	20	0.70	2.93E-05	3.61E-04	0.73	1.27E-05	2.77E-04	0.56	2.09E-03	8.37E-02
ADNP	20	0.92	1.45E-07	3.91E-05	0.75	5.49E-06	1.58E-04	0.71	3.05E-05	8.00E-03
AHCY	20	0.82	1.00E-06	3.91E-05	0.69	3.77E-05	6.54E-04	0.64	1.80E-04	2.29E-02
APMAP	20	0.87	5.58E-07	3.91E-05	0.83	8.96E-07	8.13E-05	0.80	1.23E-06	1.45E-03
CHMP4B	20	0.77	2.51E-06	5.46E-05	0.64	2.03E-04	2.21E-03	0.56	1.62E-03	7.33E-02
CRNKL1	20	0.90	2.58E-07	3.91E-05	0.70	3.21E-05	5.79E-04	0.63	3.56E-04	3.25E-02
CST3	20	0.79	1.41E-06	3.91E-05	0.69	4.54E-05	7.50E-04	0.62	3.32E-04	3.22E-02
CSTF1	20	0.91	2.18E-07	3.91E-05	0.54	2.77E-03	1.45E-02	0.58	1.08E-03	6.02E-02
DDX27	20	0.92	1.37E-07	3.91E-05	0.78	1.81E-06	8.62E-05	0.77	2.19E-06	1.45E-03
DSTN	20	0.87	5.08E-07	3.91E-05	0.74	6.91E-06	1.86E-04	0.72	1.38E-05	4.71E-03
EIF6	20	0.85	7.22E-07	3.91E-05	0.75	4.10E-06	1.29E-04	0.68	4.82E-05	1.04E-02
GSS	20	0.78	1.60E-06	4.06E-05	0.78	1.74E-06	8.60E-05	0.66	9.76E-05	1.71E-02
HNF4A	20	0.89	3.30E-07	3.91E-05	0.81	1.40E-06	8.42E-05	0.67	9.50E-05	1.71E-02
MAPRE1	20	0.87	5.00E-07	3.91E-05	0.59	6.87E-04	5.38E-03	0.56	1.61E-03	7.33E-02
MOCS3	20	0.94	1.81E-08	1.67E-05	0.70	6.55E-05	1.01E-03	0.72	4.11E-05	9.60E-03
NCOA5	20	0.87	4.83E-07	3.91E-05	0.54	2.67E-03	1.42E-02	0.55	1.96E-03	8.37E-02
NDRG3	20	0.73	8.31E-06	1.32E-04	0.76	5.83E-05	9.16E-04	0.65	1.38E-03	6.68E-02
NSFL1C	20	0.84	7.76E-07	3.91E-05	0.87	5.62E-07	8.13E-05	0.74	7.14E-06	3.00E-03
PIGU	20	0.91	1.80E-07	3.91E-05	0.68	1.79E-03	1.08E-02	0.66	2.64E-03	9.33E-02
POFUT1	20	0.75	5.13E-06	9.05E-05	0.80	1.63E-06	8.42E-05	0.58	1.07E-03	6.02E-02
PTK6	20	0.68	5.95E-05	6.22E-04	0.73	8.38E-05	1.17E-03	0.72	1.12E-04	1.76E-02
RPRD1B	20	0.96	0.00E+00	0.00E+00	0.79	1.59E-06	8.42E-05	0.75	4.97E-06	2.32E-03
RRBP1	20	0.81	1.11E-06	3.91E-05	0.74	6.25E-06	1.73E-04	0.61	4.81E-04	3.83E-02
RTFDC1	20	0.86	6.17E-07	3.91E-05	0.70	6.96E-05	1.04E-03	0.65	3.45E-04	3.22E-02
STAU1	20	0.93	5.78E-08	3.46E-05	0.77	2.19E-06	9.37E-05	0.77	2.41E-06	1.45E-03
TOMM34	20	0.85	7.17E-07	3.91E-05	0.79	1.50E-06	8.42E-05	0.64	1.76E-04	2.29E-02
VAPB	20	0.89	3.79E-07	3.91E-05	0.66	9.62E-05	1.26E-03	0.55	2.10E-03	8.37E-02
XRN2	20	0.94	1.99E-08	1.67E-05	0.61	4.08E-04	3.71E-03	0.53	2.70E-03	9.35E-02
CNDP2	18	0.63	2.52E-04	1.96E-03	0.82	9.81E-07	8.13E-05	0.59	6.87E-04	4.89E-02
GALNT1	18	0.81	1.14E-06	3.91E-05	0.63	2.76E-04	2.78E-03	0.56	1.54E-03	7.20E-02
LMAN1	18	0.68	5.28E-05	5.70E-04	0.68	5.53E-05	8.75E-04	0.58	8.85E-04	5.43E-02
NARS	18	0.82	9.51E-07	3.91E-05	0.76	2.83E-06	1.06E-04	0.63	2.72E-04	3.01E-02
TXNL1	18	0.82	1.02E-06	3.91E-05	0.62	3.66E-04	3.44E-03	0.63	2.37E-04	2.69E-02
USP14	18	0.79	1.41E-06	3.91E-05	0.73	1.03E-05	2.45E-04	0.58	8.76E-04	5.43E-02
ACOX1	17	0.58	9.97E-04	5.94E-03	0.84	8.00E-07	8.13E-05	0.54	2.57E-03	9.31E-02
ALKBH5	17	0.77	2.38E-06	5.27E-05	0.73	6.89E-03	2.90E-02	0.77	2.92E-03	9.71E-02
C1QBP	17	0.68	6.23E-05	6.45E-04	0.58	8.57E-04	6.29E-03	0.67	8.46E-05	1.61E-02
GLOD4	17	0.79	1.39E-06	3.91E-05	0.58	9.35E-04	6.75E-03	0.62	3.08E-04	3.16E-02
PFN1	17	0.74	5.22E-06	9.16E-05	0.53	3.19E-03	1.60E-02	0.63	2.86E-04	3.08E-02
TSR1	17	0.70	3.27E-05	3.93E-04	0.64	1.38E-03	8.92E-03	0.63	1.74E-03	7.79E-02
YWHAE	17	0.74	5.75E-06	9.86E-05	0.67	8.58E-05	1.19E-03	0.61	4.98E-04	3.83E-02
ANXA2	15	0.56	1.73E-03	9.01E-03	0.80	1.30E-06	8.42E-05	0.55	2.12E-03	8.37E-02
DLST	14	0.81	1.10E-06	3.91E-05	0.59	7.10E-04	5.50E-03	0.66	1.03E-04	1.73E-02
NEK9	14	0.75	3.92E-06	7.44E-05	0.66	1.86E-04	2.07E-03	0.57	2.02E-03	8.37E-02
PRMT5	14	0.69	3.38E-05	4.00E-04	0.77	3.30E-06	1.11E-04	0.55	2.35E-03	8.88E-02
TMX1	14	0.84	7.96E-07	3.91E-05	0.66	9.76E-05	1.28E-03	0.73	9.35E-06	3.57E-03
ALG5	13	0.80	1.21E-06	3.91E-05	0.75	6.02E-06	1.70E-04	0.59	9.06E-04	5.43E-02
DIS3	13	0.79	1.42E-06	3.91E-05	0.75	4.51E-06	1.33E-04	0.57	1.11E-03	6.02E-02
ERCC5	13	0.77	2.27E-06	5.06E-05	0.69	1.45E-04	1.71E-03	0.63	8.25E-04	5.43E-02
ESD	13	0.68	5.20E-05	5.64E-04	0.64	2.00E-04	2.18E-03	0.80	1.28E-06	1.45E-03
GTF2F2	13	0.73	7.64E-06	1.24E-04	0.55	2.27E-03	1.27E-02	0.56	2.07E-03	8.37E-02
HMGB1	13	0.76	3.22E-06	6.60E-05	0.80	1.24E-06	8.41E-05	0.77	2.14E-06	1.45E-03
IPO5	13	0.75	4.89E-06	8.70E-05	0.55	2.04E-03	1.18E-02	0.59	8.11E-04	5.43E-02
POLR1D	13	0.89	3.07E-07	3.91E-05	0.69	2.73E-04	2.77E-03	0.65	7.74E-04	5.41E-02
SUCLA2	13	0.71	1.68E-05	2.32E-04	0.89	3.22E-07	8.13E-05	0.76	3.00E-06	1.57E-03
SUGT1	13	0.76	3.22E-06	6.60E-05	0.60	5.16E-04	4.38E-03	0.61	5.04E-04	3.83E-02
TPP2	13	0.86	6.50E-07	3.91E-05	0.67	7.65E-05	1.13E-03	0.61	4.23E-04	3.70E-02
UCHL3	13	0.80	1.22E-06	3.91E-05	0.69	3.65E-05	6.39E-04	0.69	3.65E-05	9.02E-03
ARFGEF1	8	0.68	5.78E-05	6.13E-04	0.77	4.26E-06	1.29E-04	0.74	1.46E-05	4.71E-03
ATP6V1B2	8	0.75	3.86E-06	7.44E-05	0.73	9.67E-06	2.33E-04	0.70	2.49E-05	6.98E-03
CPNE3	8	0.80	1.33E-06	3.91E-05	0.51	4.69E-03	2.14E-02	0.53	2.93E-03	9.71E-02
CTSB	8	0.75	4.58E-06	8.39E-05	0.56	1.41E-03	9.02E-03	0.55	2.02E-03	8.37E-02

ESRP1	8	0.70	3.22E-05	3.88E-04	0.83	8.83E-07	8.13E-05	0.77	2.38E-06	1.45E-03
GGH	8	0.68	4.90E-05	5.34E-04	0.86	6.08E-07	8.13E-05	0.57	1.12E-03	6.02E-02
GSR	8	0.77	2.48E-06	5.42E-05	0.78	1.85E-06	8.62E-05	0.78	1.83E-06	1.45E-03
HSF1	8	0.74	6.15E-06	1.04E-04	0.68	1.28E-04	1.57E-03	0.65	3.45E-04	3.22E-02
LACTB2	8	0.54	2.57E-03	1.23E-02	0.73	1.05E-05	2.48E-04	0.65	1.34E-04	1.88E-02
PRKDC	8	0.74	5.66E-06	9.74E-05	0.73	7.77E-06	2.03E-04	0.54	2.36E-03	8.88E-02
PROSC	8	0.72	1.47E-05	2.08E-04	0.67	6.80E-05	1.03E-03	0.57	1.13E-03	6.02E-02
PUF60	8	0.81	1.10E-06	3.91E-05	0.76	2.87E-06	1.07E-04	0.61	5.10E-04	3.83E-02
RRS1	8	0.77	2.24E-06	5.06E-05	0.53	5.83E-03	2.56E-02	0.70	1.17E-04	1.76E-02
TCEA1	8	0.67	6.80E-05	7.02E-04	0.79	1.41E-06	8.42E-05	0.61	4.33E-04	3.71E-02
XPO7	8	0.80	1.19E-06	3.91E-05	0.55	2.02E-03	1.18E-02	0.64	1.93E-04	2.38E-02
YTHDF3	8	0.74	7.26E-06	1.20E-04	0.59	7.18E-04	5.54E-03	0.61	4.13E-04	3.69E-02
ZFAND1	8	0.73	7.90E-06	1.27E-04	0.58	4.36E-03	2.02E-02	0.60	2.94E-03	9.71E-02
ZNF706	8	0.76	2.83E-06	5.95E-05	0.77	1.95E-04	2.14E-03	0.70	1.18E-03	6.04E-02
HIBADH	7	0.79	1.51E-06	4.02E-05	0.78	1.60E-06	8.42E-05	0.59	8.29E-04	5.43E-02
NUDCD3	7	0.91	2.02E-07	3.91E-05	0.84	1.04E-04	1.33E-03	0.80	4.89E-04	3.83E-02
PPIA	7	0.72	1.47E-05	2.08E-04	0.65	1.58E-04	1.82E-03	0.57	1.16E-03	6.04E-02
SUN1	7	0.76	2.91E-06	6.06E-05	0.60	5.41E-04	4.51E-03	0.59	8.20E-04	5.43E-02
DNPH1	6	0.55	1.85E-03	9.55E-03	0.82	1.01E-06	8.13E-05	0.57	1.19E-03	6.04E-02
PRKAA1	5	0.61	4.33E-04	3.07E-03	0.65	1.46E-04	1.71E-03	0.65	1.56E-04	2.11E-02
SUB1	5	0.54	2.25E-03	1.11E-02	0.54	2.43E-03	1.33E-02	0.54	2.21E-03	8.57E-02
TARS	5	0.65	1.60E-04	1.37E-03	0.76	3.04E-06	1.09E-04	0.68	4.97E-05	1.04E-02
EIF4E	4	0.71	1.15E-05	1.72E-04	0.55	2.04E-03	1.18E-02	0.53	2.65E-03	9.33E-02
GALNT7	4	0.64	2.08E-04	1.67E-03	0.76	2.83E-06	1.06E-04	0.53	3.04E-03	9.83E-02
PPP3CA	4	0.52	3.24E-03	1.48E-02	0.71	2.12E-05	4.19E-04	0.54	1.98E-03	8.37E-02
RAP1GDS1	4	0.68	4.24E-05	4.75E-04	0.65	1.62E-04	1.85E-03	0.56	1.41E-03	6.74E-02
WDR1	4	0.66	1.17E-04	1.09E-03	0.76	2.91E-06	1.07E-04	0.60	5.53E-04	4.01E-02
CAPZB	1	0.58	8.76E-04	5.33E-03	0.61	4.13E-04	3.75E-03	0.71	1.71E-05	5.13E-03
EFHD2	1	0.57	1.17E-03	6.69E-03	0.82	9.99E-07	8.13E-05	0.57	1.17E-03	6.04E-02
PGD	1	0.64	2.25E-04	1.78E-03	0.66	9.62E-05	1.26E-03	0.53	2.72E-03	9.36E-02

Supplementary Table 9. Gene-dosage effect in colorectal adenomas. Significantly and positively correlated genes/proteins among DNA copy number, RNA and Protein expression are presented in the table.

		DNA vs RNA			**RNA vs Protein**			**DNA vs Protein**		
Gene name	**Chromosome**	**Correlation coefficient**	**p-value**	**FDR**	**Correlation coefficient**	**p-value**	**FDR**	**Correlation coefficient**	**p-value**	**FDR**
EIF6	20	0.67	1.22E-04	2.02E-02	0.71	2.46E-05	9.43E-04	0.75	5.11E-06	2.12E-02
POFUT1	20	0.71	2.42E-05	7.16E-03	0.80	1.98E-06	1.78E-04	0.62	6.06E-04	2.12E-01
RPRD1B	20	0.74	8.09E-06	4.19E-03	0.67	1.13E-04	2.89E-03	0.64	2.53E-04	1.77E-01
CCAR2	8	0.55	2.27E-03	6.60E-02	0.51	5.72E-03	4.23E-02	0.58	1.17E-03	2.43E-01
EPHX2	8	0.64	2.60E-04	2.76E-02	0.77	3.35E-06	2.21E-04	0.61	6.13E-04	2.12E-01
ALDH7A1	5	0.56	2.04E-03	6.47E-02	0.68	6.78E-05	1.98E-03	0.59	1.03E-03	2.42E-01
EPHB2	1	0.62	5.00E-04	3.40E-02	0.74	8.39E-06	4.20E-04	0.64	2.82E-04	1.77E-01
H6PD	1	0.64	2.78E-04	2.82E-02	0.73	1.50E-05	6.22E-04	0.59	1.02E-03	2.42E-01
RCC1	1	0.53	3.53E-03	7.88E-02	0.63	3.46E-04	6.19E-03	0.66	1.55E-04	1.77E-01
RCC2	1	0.73	1.18E-05	5.42E-03	0.60	6.94E-04	1.03E-02	0.60	8.13E-04	2.41E-01

Supplementary Table 10. Association between sample type and RPRD1B protein expression as measured by immunohistochemistry in tissue miscroarrays. RPRD1B expression was measured as a product of epithelial nuclear staining intensity (negative=0, weak=1, moderate=2 or strong=3) and percentage of the cells stained positively (0-100%).

Comparison	**RPRD1B expression**		**Fisher exact test**	
	0-190	200-300	p-value	odds ratio
high-risk adenoma	2	7	0.197	4.35
low-risk adenoma	8	6		
high-risk adenoma & CRC	7	28	0.017	5.12
low-risk adenoma	8	6		
high-risk adenoma & CRC	7	28	0.003	5.20
low-risk adenoma & normal adjacent colon	20	15		

Supplementary Table 11. A. Association of POFUT1 expression as measured by immunohistochemistry with sample type and with amount of goblet cells. POFUT1 expression is a product of staining intensity (negative=0, weak=1, moderate=2 or strong=3) and percentage of the area stained (0-100%). **B.** Association of POFUT1 expression and amount of goblet cells with adenoma features like risk of progression, grade of dysplasia, histology and size. P-values and odds ratio were onbtained with Fisher exact test.

Supplementary Table 11A

Comparisons	**POFUT1 expression**		**Fisher exact test**	
	0-200	240-300	p-value	odds ratio
high-risk adenoma & CRC	22	14	**0.039**	8
low-risk adenoma	13	1		
high-risk adenoma & CRC	22	14	**<0.001**	18.4
low-risk adenoma & normal adjacent colon	30	1		
few goblet cells	5	4	**0.017**	0
moderate-many goblet cells	13	0		

Supplementary Table 11B

Adenomas	**POFUT1 expression**		**Fisher exact test**		**Amount of goblet cells**		**Fisher exact test**	
	0-200	**240-300**	p-value	odds ratio	**few**	**moderate/ many**	p-value	odds ratio
Risk of progression			0.056	9.3			**0.007**	17.3
high	5	4			2	12		
low	13	1			7	2		
Dysplasia			**0.017**	16.6			0.066	6.8
high grade	3	4			5	2		
low grade	15	1			4	12		
Histology			0.539	0			1	0.5
tubulovillous/villous	14	5			8	11		
tubular	4	0			1	3		
Size (median = 21.5 mm)			0.64	1.8			0.68	1.6
≥ 21.5 mm	8	3			5	6		
< 21.5 mm	10	2			4	8		

Supplementary Table 12. Cancer associated events (CAEs) in patient-derived adenoma organoids - DNA copy number aberrations and the risk of progression for the adenomas. See legend.

Legend:

Variable	Explanation
0	no copy number aberration
1	copy number aberration present
high	2 out of 7 aberrations present
low	0 out of 7 aberrations present

Sample name	Cancer associated events							Risk of progression
	8q	13q	20q	8p	15q	17p	18q	
B16PON_C-002	0	0	0	0	0	0	0	low
B16PON_C-003	0	0	0	0	0	0	0	low
B16PON_C-004	0	0	0	0	0	0	0	low
B16PON_C-006	0	1	1	0	0	0	0	high
B16PON_C-008	1	1	0	0	0	0	0	high
B16PON_C-009	0	1	1	0	0	0	0	high
B16PON_C-010	0	0	0	0	0	0	0	low
B16PON_C-011	0	0	0	0	0	0	0	low
B16PON_C-015	0	0	0	0	0	0	0	low
B16PON_C-017	0	0	0	0	0	0	0	low
B16PON_C-018	0	0	0	0	0	0	0	low
B16PON_C-020	0	1	1	0	0	0	0	high
B16PON_C-024	0	0	0	0	0	0	0	low
B16PON_C-027	0	0	0	0	0	0	0	low
B16PON_C-028	0	0	0	0	0	0	0	low
B16PON_C-036	1	0	1	0	0	0	0	high
B16PON_C-040	0	0	0	0	0	1	1	high
B16PON_C-041	0	0	0	0	0	0	0	low
B16PON_C-045	1	1	1	0	0	0	0	high
B16PON_C-046	0	0	0	0	0	0	0	low
B16PON_C-047	0	0	0	0	0	0	0	low
B16PON_C-048	0	1	1	0	1	1	1	high
B16PON_C-049	0	0	0	0	0	0	0	low

Supplementary Table 13. Fisher exact test results for risk of progression and POFUT1 expression as measured by immunohistochemistry for patient-derived colorectal adenoma organoids. POFUT1 expression was measured as a product of epithelial cytoplasmic staining intensity (negative=0, weak=1, moderate=2 or strong=3) and percentage of the cells stained positively (0-100%).

Comparison	**POFUT1 expression**		**Fisher exact test**	
	0-120	160-210	p-value	odds ratio
high-risk adenoma	4	4	0.0079	Inf
low-risk adenoma	15	0		

Chapter 3

CONSENSUS MOLECULAR SUBTYPE CLASSIFICATION OF COLORECTAL ADENOMAS

Malgorzata A Komor, Linda JW Bosch, Gergana Bounova, Anne S Bolijn,
Pien Delis van-Diemen, Christian Rausch, Youri Hoogstrate,
Andrew P Stubbs, Mark de Jong, Guido Jenster, Nicole CT van Grieken,
Beatriz Carvalho, Lodewyk FA Wessels, Connie R Jimenez,
Remond JA Fijneman, Gerrit A Meijer

In collaboration with the NGS-ProToCol consortium

JOURNAL OF PATHOLOGY 2018;246:266-276

Consensus molecular subtype classification of colorectal adenomas

Malgorzata A Komor[1,2], Linda JW Bosch[1], Gergana Bounova[3], Anne S Bolijn[1], Pien M Delis-van Diemen[1], Christian Rausch[1], Youri Hoogstrate[4], Andrew P Stubbs[5], Mark de Jong[6], Guido Jenster[4], Nicole CT van Grieken[7], Beatriz Carvalho[1], Lodewyk FA Wessels[3,8], Connie R Jimenez[2], Remond JA Fijneman[1]*, Gerrit A Meijer[1]

In collaboration with the NGS-ProToCol Consortium:
Natasja Dits[4], Rene Bottcher[4], Annemieke C Hiemstra[1], Bauke Ylstra[7], Daoud Sie[7], Evert van den Broek[7], David van der Meer[6], Floor Pepers[6], Eric Caldenhoven[9], Bart Janssen[6], Wilbert van Workum[6], Stef van Lieshout[7], Chris H. Bangma[4], Geert van Leenders[10] and Harmen van de Werken[4]

[1] *Translational Gastrointestinal Oncology, Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands*
[2] *Oncoproteomics Laboratory, Department of Medical Oncology, VU University Medical Centre, Amsterdam, The Netherlands*
[3] *Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands*
[4] *Department of Urology, Erasmus Medical Centre Rotterdam, Rotterdam, The Netherlands*
[5] *Department of Bioinformatics, Erasmus Medical Centre Rotterdam, Rotterdam, The Netherlands*
[6] *GenomeScan, Leiden, The Netherlands*
[7] *Department of Pathology, VU University Medical Centre, Amsterdam, The Netherlands*
[8] *Department of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands*
[9] *Lygature, Utrecht, The Netherlands*
[10] *Department of Pathology, Erasmus Medical Centre Rotterdam, Rotterdam, The Netherlands*

**Correspondence to: RJA Fijneman, The Netherlands Cancer Institute, Department of Pathology, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. E-mail: r.fijneman@nki.nl*

Abstract

Consensus molecular subtyping is an RNA expression-based classification system for colorectal cancer (CRC). Genomic alterations accumulate during CRC pathogenesis, including the premalignant adenoma stage, leading to changes in RNA expression. Only a minority of adenomas progress to malignancies, a transition that is associated with specific DNA copy number aberrations or microsatellite instability (MSI). We aimed to investigate whether colorectal adenomas can already be stratified into consensus molecular subtype (CMS) classes, and whether specific CMS classes are related to the presence of specific DNA copy number aberrations associated with progression to malignancy. RNA sequencing was performed on 62 adenomas and 59 CRCs. MSI status was determined with polymerase chain reaction-based methodology. DNA copy number was assessed by low-coverage DNA sequencing ($n = 30$) or array-comparative genomic hybridisation ($n = 32$). Adenomas were classified into CMS classes together with CRCs from the study cohort and from The Cancer Genome Atlas ($n = 556$), by use of the established CMS classifier. As a result, 54 of 62 (87%) adenomas were classified according to the CMS. The CMS3 'metabolic subtype', which was least common among CRCs, was most prevalent among adenomas ($n = 45$; 73%). One of the two adenomas showing MSI was classified as CMS1 (2%), the 'MSI immune' subtype. Eight adenomas (13%) were classified as the 'canonical' CMS2. No adenomas were classified as the 'mesenchymal' CMS4, consistent with the fact that adenomas lack invasion-associated stroma. The distribution of the CMS classes among adenomas was confirmed in an independent series. CMS3 was enriched with adenomas at low risk of progressing to CRC, whereas relatively more high-risk adenomas were observed in CMS2. We conclude that adenomas can be stratified into the CMS classes. Considering that CMS1 and CMS2 expression signatures may mark adenomas at increased risk of progression, the distribution of the CMS classes among adenomas is consistent with the proportion of adenomas expected to progress to CRC.

© 2018 The Authors. *The Journal of Pathology* published by John Wiley & Sons Ltd on behalf of Pathological Society of Great Britain and Ireland.

Keywords: colon; rectum; neoplasia; adenoma; colorectal cancer

Received 6 March 2018; Revised 8 May 2018; Accepted 20 June 2018

No conflicts of interest were declared.

Introduction

Colorectal cancer (CRC) is heterogeneous in its molecular characteristics and its treatment response. Stratifying CRC patients into biologically and clinically distinct subtypes, based on gene expression profiles, has been performed in many studies, with the common aim of improving clinical precision [1–7]. Recently, a large effort was made by the CRC Subtyping Consortium to reconcile the differences between the multiple existing classifications and to derive consensus molecular subtypes (CMSs) of CRC [8]. A consensus RNA expression-based classifier was produced that classifies CRCs into four CMS groups. CMS1 includes ~14% of CRCs, and is associated with microsatellite instability (MSI), *BRAF* mutation, promoter hypermethylation, and immune infiltration. Chromosomal instability (CIN), the most common type of genomic instability in CRC, is a feature characteristic of CMS2–CMS4. CMS2 is the most prevalent CRC subtype (37%) and shows the hallmarks of canonical CRC carcinogenesis, including activation of the Wnt and Myc pathways. Approximately 13% of CRCs are in CMS3, characterised by dysregulated metabolism and *KRAS* mutation. Finally, CMS4 (23%) is described as a mesenchymal, stroma-rich group, associated with poor prognosis [8].

Most CRCs progress from normal epithelium, through a benign precursor adenoma, by accumulating genetic alterations in oncogenes and tumour suppressor genes [9]. However, adenomas are much commoner in the large intestine than cancers, and it is estimated that only 5% eventually progress to cancer [10]. Although it is evident that CMS signatures can be discerned at the CRC stage, the question remains of whether this would already be possible at the adenoma stage, and, if so, how the distribution of CMS classes would compare with that of CRCs.

A further question is whether adenomas with a high risk of progressing to cancer would differ in their CMS pattern from adenomas with a low risk of progression. In general, the progression of dysplastic epithelial premalignant lesions such as colorectal adenomas is associated with the acquisition of genomic instability. Often, this concerns aneuploidy or CIN, which marks ~85% of CRC cases [11]. CIN has been studied in CRC and its precursor lesions to identify non-random chromosomal aberrations and potential CRC driver events. In multiple studies, a distinct pattern has been observed in colorectal lesions with CIN, which has been shown to play a major role in adenoma-to-carcinoma progression [12–21]. Seven copy number aberrations have been identified as colorectal cancer-associated events (CAEs): gains of chromosomal arms 8q, 13q, and 20q, and losses of 8p, 15q, 17p, and 18q [12]. With an accuracy of 78%, adenomas with at least two of the seven CAEs can be identified as being at a high risk of progressing to malignancy; these are referred to as 'high-risk adenomas' [12]. Integration of these DNA copy number aberrations and RNA expression data led to the identification of putative oncogenes located in the amplified regions [22,23]. Functional studies of candidate oncogenes from the 20q region indicated that *AURKA* and *TPX2* promote 20q amplicon-driven adenoma-to-carcinoma progression [16]. This means that the non-random DNA copy number aberrations do, in fact, influence biological processes within cells, through which they facilitate colorectal tumourigenesis. The fact that these aberrations are present in some of the adenomas shows that the signal of malignant transformation can already be detected at a molecular level at the adenoma stage. This implies that gene expression profiles of colorectal adenomas may also carry information on the future CMS.

The present study therefore aimed to investigate whether the differentiation of colorectal epithelial neoplasia into CMS classes can already be recognised at the adenoma stage, and whether specific CMS classes are associated with the absence or presence of specific DNA copy number aberrations in colorectal adenomas that reflect a high risk of progressing to cancer.

Materials and methods

Sample collection

A total of 62 snap-frozen advanced adenomas and 59 CRCs were collected from two independent sample collections: Series 1 and Series 2 (described in supplementary material, Supplementary materials and methods). Clinical information is shown in Table 1. The collection, storage and use of tissue and patient data were performed in compliance with the Code for Proper Secondary Use of Human Tissue in the Netherlands [24].

DNA copy number analysis

For Series 1, copy number analysis by low-coverage whole genome sequencing was performed (supplementary material, Supplementary materials and methods and Table S1). Gains and losses of whole chromosomal arms were used for the identification of high-risk adenomas.

Table 1. Characteristics of sample Series 1 and Series 2 collected for this study

		Number of samples		
Characteristics		**Series 1**	**Series 2**	**Total**
Lesion	Adenoma	30	32	62
Histological type	Tubular	6	13	19
	Tubulovillous	20	16	36
	Villous	4	3	7
Dysplasia	High grade	10	8	18
	Low grade	20	24	44
Risk of progression	High	9	4	13
	Low	17	22	39
	No information	2	6	8
Microsatellite status	MSS	28	32	60
	MSI	2	0	2
Lesion	Carcinoma	30	29	59
Differentiation grade	Less/Not	4	2	6
	Well differentiated/ moderately differentiated	25	27	52
	No information	1	0	1
Stage	I	7	9	16
	II	13	10	23
	III	6	9	15
	IV	3	1	4
	I or III	1	0	1
Microsatellite status	MSS	24	23	47
	MSI	6	6	12

MSS, microsatellite-stable.

Samples were considered to have undetermined risk when the copy number aberrations were present but did not reach the probability cut-off of 0.5 ($n = 2$). For Series 2, DNA copy number data for 28 adenomas were obtained from the array-comparative genomic hybridisation (arrayCGH) analysis in an earlier study [22]. Samples were considered to have undetermined risk if the arrayCGH data were unavailable ($n = 4$) or only a minor part of the chromosomal arm was gained or lost ($n = 2$). For both series, adenomas with at least two of seven CAEs were labelled as high-risk [12].

MSI assay

Adenoma and carcinoma samples from both series were analysed for MSI with the MSI Multiplex System Version 1.2 (Promega, Madison, WI, USA; cat. no. MD1641) according to standard procedures, as described previously [25].

RNA sequencing (RNA-seq) and data preprocessing

Both series were subjected to RNA-seq and data preprocessing separately. Expression matrices were obtained for each series (supplementary material, Supplementary materials and methods and Table S1).

Batch effect removal with respect to The Cancer Genome Atlas (TCGA) CRC data

TCGA data served as a reference for performance of the analysis in the present study [15]. Expression values of 556 TCGA samples used in the original CMS classification were used for RNA-seq data normalisation and CMS classification (supplementary material, Supplementary materials and methods).

The batch effect was removed with M-Combat [26], separately for Series 1 and Series 2. In both cases, the TCGA dataset served as the reference, and Series 1 or Series 2 served as the normalised batch (Figure 1). Adenomas and cancers were kept together during the normalisation to avoid removal of the 'lesion-based' variance. TCGA data as the gold-standard reference dataset remained unchanged. All three datasets (Series 1, Series 2, and TCGA) were merged, and Series 1 and Series 2 formed the study dataset. Batch effect removal was evaluated by use of a multidimensional scaling algorithm on the Euclidian distance between the expression profiles of the samples. Evaluation of the preservation of the difference between adenomas and carcinomas was performed by the use of hierarchical clustering with complete linkage on the $\log_2$-transformed RPKMs of the top 30 and the top 1000 variable genes.

CMS classification

Ensembl IDs were translated to Entrez IDs with the biomaRt Bioconductor package [27]. The random forest CMS classifier [8] was applied on the merged dataset, including TCGA dataset, Series 1, and Series 2, and a CMS class was assigned when the posterior probability of a sample belonging to a subtype was ≥0.5. To obtain the original CMS labels for TCGA samples, the random forest CMS classifier was also applied to the whole CMS dataset downloaded from the CRC Subtyping Consortium Synapse website [8,28]. CMS labels for TCGA samples were extracted. To evaluate the results of the random forest CMS classifier, the single sample predictor (SSP) classification method [8] was applied to the adenomas from Series 1 and Series 2 before normalisation to the TCGA dataset. A CMS class was assigned according to the default settings (minCor = 0.15, minDelta = 0.06).

Validation set

To validate the results in an independent series of adenomas measured with a different platform, expression data from the Affymetrix Human Genome U133 Plus 2.0 Array of 45 colorectal adenomas and 36 CRCs (GSE20916) were downloaded from the Gene Expression Omnibus. This validation set will be referred as 'Series 3' [29]. The reference dataset chosen was the largest series of CRCs measured with the same methodology and used in the original CMS classification (GSE39582) [3,8]. See supplementary material, Supplementary materials and methods for details of data analysis and the CMS classification of Series 3.

Statistical analysis

The multinomial exact test was used to perform a goodness-of-fit test for the distributions of CMS classes in the adenomas in comparison with cancers from the study dataset, adenomas from the validation set, and cancers from the original CMS publication [8]. Contingency tables including adenomas classified as CMS2 and CMS3 were analysed; CMS1 and CMS4 were excluded because of the limited number of cases. Associations analysed were clinical features, risk of progression or occurrence of each of the seven CAEs separately. A relationship was considered to be significant if the P value was ≤0.05 (Fisher's exact test). Additionally, associations between CMS classes in CRCs and clinical features were analysed.

Gene set enrichment analysis (GSEA)

Prior to GSEA [30], an expression matrix after normalisation was extracted for CMS2 and CMS3 adenomas. Exponentiation with base 2 was applied, and values were rounded to integers to create count data. Differential gene expression analysis was performed with the Bioconductor package DESeq2 [31], and genes were sorted on the basis of $\log_2$ fold change, whereby genes upregulated in CMS2 adenomas were at the top of the list. (Fold change is defined as the ratio of test to reference expression level.) The $\log_2$ fold change-based ranked list was submitted to GSEA [30], and the collection of hallmark gene sets from Molecular Signature Database v6.0 was used [32]. Significant gene sets were extracted on the basis of a false discovery rate (FDR) threshold of ≤0.2. For the comparison of stroma and invasion signatures between adenomas and cancers,

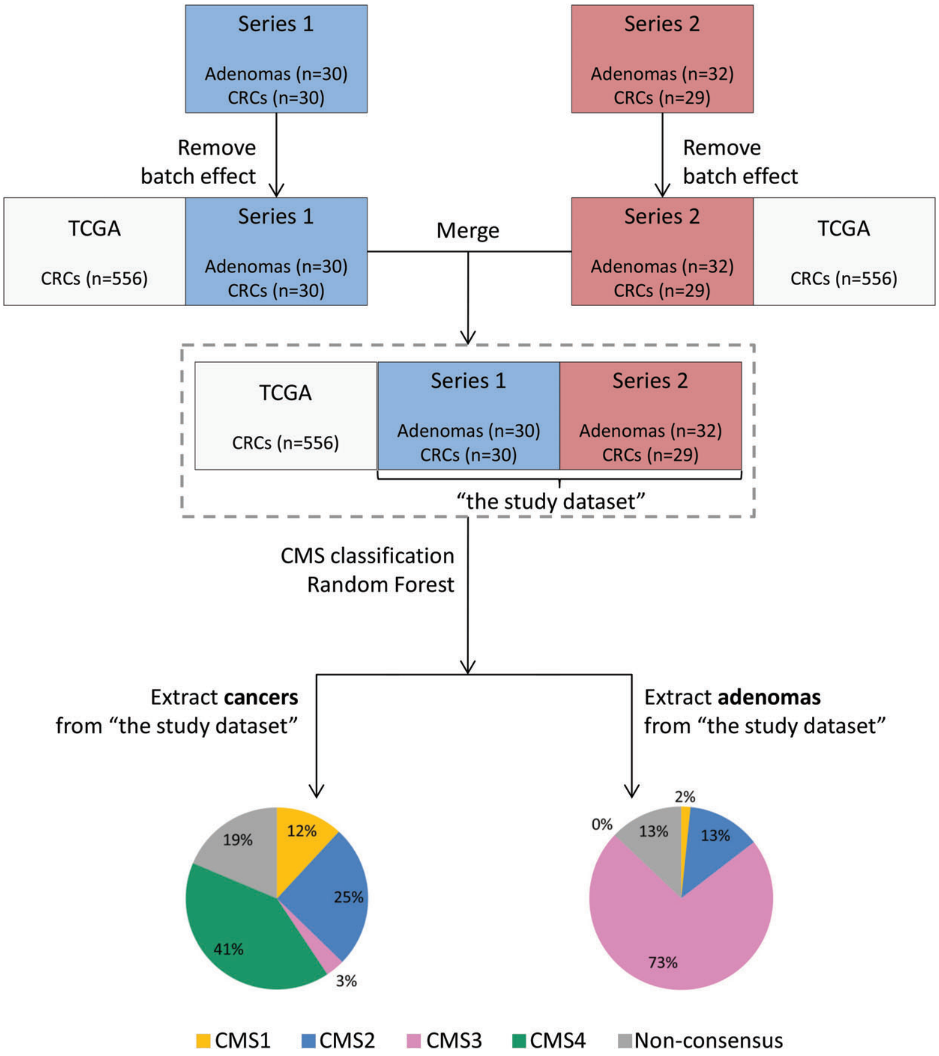


Figure 1. Overview of the data analysis approach. Both Series 1 and Series 2 were normalised separately to the TCGA CRC dataset via a batch effect removal method [27]. After normalisation, all three datasets were merged together. Series 1 and Series 2 form the 'study dataset'. CMS classification was applied to the merged dataset. The classes were obtained with the CMS random forest classifier, and assigned when the posterior probability of belonging to a CMS class was ≥0.5. Results of the classification were extracted for the CRCs and the adenomas from the study dataset. The pie charts represent the distribution of CMS classes for CRCs (left) and adenomas (right) for the study dataset.

the ESTIMATE algorithm [33] was used, as well as single-sample GSEA with the GSVA Bioconductor package [34], with the 'invasive front' and 'central tumour' signatures [35].

Results

CMS classification of the cancers and the adenomas

An overview of the data analysis is shown in Figure 1. Series 1, Series 2 and the TCGA dataset originated from different experiments, representing three separate batches that needed to be normalised (supplementary material, Figure S1A). To avoid a change in the original TCGA classification, the TCGA dataset remained unchanged and was used as a gold-standard reference for batch effect removal. Both Series 1 and Series 2 were successfully normalised to the TCGA dataset (supplementary material, Figure S1B). Hierarchical clustering based on expression values of the top 30 and top 1000 variable genes before and after batch effect removal showed that the normalisation did not remove

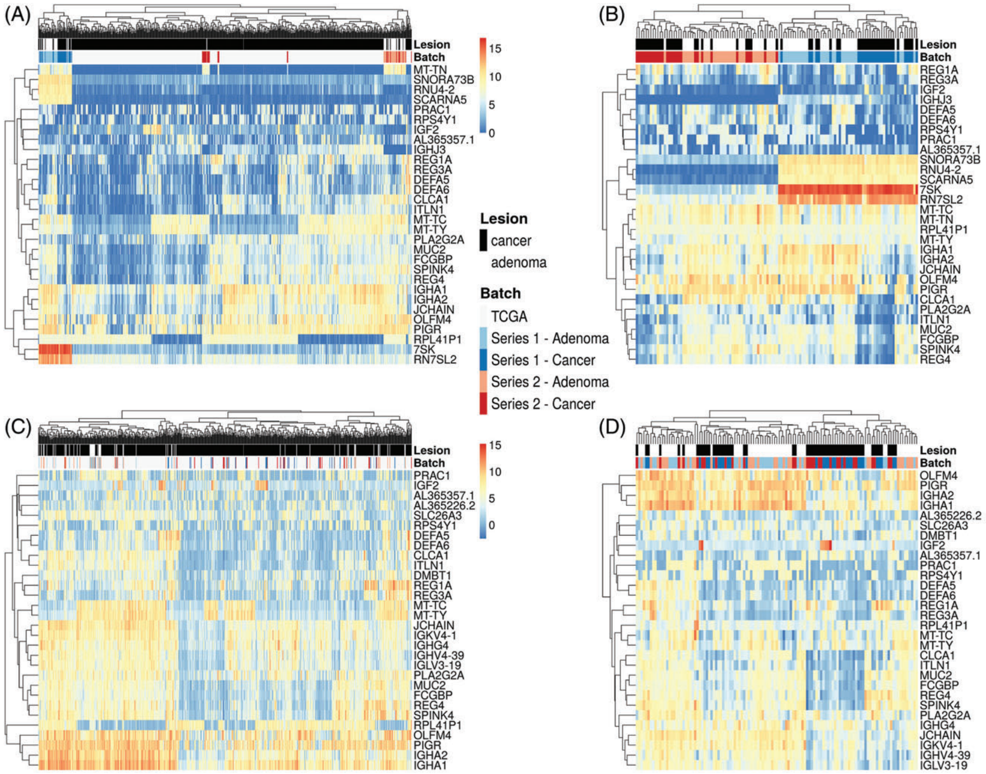


Figure 2. Hierarchical clustering based on gene expression profiles of the top 30 most variable genes. (A) Heatmap of all three datasets before batch effect removal. The batches corresponding to the TCGA dataset, Series 1 and Series 2 can be distinguished in the heatmap. (B) Heatmap before batch correction of the Series 1 and Series 2 study datasets only. Within the two batches, one can distinguish clusters enriched with adenomas and clusters enriched with cancers. (C) Heatmap of all three datasets after batch effect removal. Samples from the three experiments do not cluster together. (D) Heatmap of the Series 1 and Series 2 study datasets after batch effect removal. Clusters enriched with adenomas or cancers can still be distinguished, meaning that batch effect correction did not remove the variability between different lesions. The legend corresponds to all of the heatmaps in this figure.

the differences between the adenomas and the cancers, as the lesions could still be distinguished on the basis of their expression profiles (Figure 2; supplementary material, Figure S2). The variability between cancers and adenomas was thus preserved after batch effect removal.

On the basis of two tissue datasets, Series 1 and Series 2, we collected a cohort of 62 adenomas and 59 CRCs, referred to as the study dataset. To ensure proper classification of the adenomas, which constitute a different entity from CRCs, the CMS classification was applied to a merged dataset with carcinomas from the present study ($n=59$) and TCGA data ($n=556$); see Figure 1 for an overview of the data analysis approach. To evaluate whether the data analysis approach had an impact on the classification, the CMS labels obtained in this study for TCGA samples were compared with their original CMS labels [8]. The CMS labels of TCGA samples were reassigned in this study with an accuracy of 97%, corresponding to the previously reported overall accuracy of the random forest CMS classifier of 96% (supplementary material, Table S2) [8].

The CMS classification results of the study dataset were extracted. In total, 48 of 59 cancers were classified with a posterior probability of ≥0.5. Of these, seven were classified as CMS1, 15 as CMS2, two as CMS3, and 24 as CMS4 (Figure 1; Table 2; supplementary material, Table S3). Hence, the CMS4 mesenchymal subtype was the most prevalent in this dataset. Of the 12 samples of CRC with MSI, four were classified as CMS1, four were classified as CMS4, one was classified as CMS3, and three were not classified. Statistically significant associations of CMS classes with MSI status ($p=0.004$) and with differentiation grade ($p=0.006$) were observed, but no association with stage was identified ($p=0.235$; see supplementary material, Table S4, for MSI status and association analysis).

CMS subtype signatures were indeed expressed in the adenomas, and 54 of 62 samples were successfully classified with a probability threshold of ≥0.5. The

Table 2. Distribution of the CMS classes in cancers and adenomas from the study dataset and the validation set

	CMS1, *n* (%)	CMS2, *n* (%)	CMS3, *n* (%)	CMS4, *n* (%)	Non-consensus, *n* (%)
Study dataset (Series 1 and Series 2)					
Cancers	7 (12)	15 (25)	2 (3)	24 (41)	11 (19)
Adenomas	1 (2)	8 (13)	45 (73)	0 (0)	8 (13)
Validation set (Series 3)					
Cancers	5 (14)	7 (19)	1 (3)	18 (50)	5 (14)
Adenomas	1 (2)	5 (11)	28 (62)	0 (0)	11 (24)

vast majority of the adenomas, i.e. 45 samples (73%), were assigned to CMS3. Additionally, eight adenomas (13%) were subtyped as CMS2, representing the canonical CRC carcinogenesis. Only a single adenoma was classified as CMS1, being one of the two MSI adenomas identified in the whole dataset. No adenomas were subtyped as CMS4 (Table 2; supplementary material, Table S5). The distribution of CMS classes in the adenomas differed significantly from that in the CRCs from the study dataset ($p < 2.2 \times 10^{-16}$) and CRCs from the original CMS publication ($p < 2.2 \times 10^{-16}$) [8].

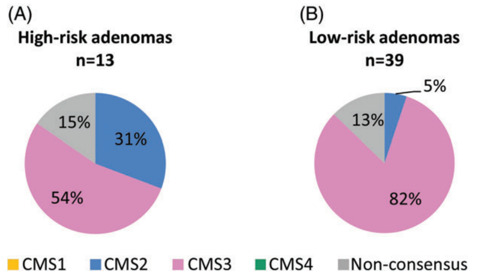


Figure 3. Distribution of CMS classes among adenomas at high risk and adenomas at low risk of progressing to cancer. (A) Distribution of CMS classes among 13 high-risk adenomas. (B) Distribution of CMS classes among 39 low-risk adenomas. No high-risk and low-risk adenomas were classified as CMS1 or CMS4.

CMS classification of adenomas, risk of progression, and biological characterisation

Adenomas from the study dataset were called high risk on the basis of the presence of at least two of seven specific DNA copy number aberrations: 8q, 13q and 20q gains, and 8p, 15q, 17p and 18q losses [12]. Adenomas with MSI were excluded, because a different genome instability process (i.e. not CIN) is involved. In total, 13 adenomas were called high risk and 39 were called low risk (Table 1; supplementary material, Table S6). No final calls could be made for the remaining eight.

Adenomas classified as CMS2 ($n = 8$) and CMS3 ($n = 45$) were the most prevalent; there were no CMS4 adenomas, and there was one adenoma classified as CMS1. Therefore, only differences between CMS2 and CMS3 adenomas were examined in terms of risk of progression, cancer-specific DNA copy number aberrations, clinical characteristics, and biological processes specific for each group. Examination of associations between CMS class and risk of progression revealed that CMS2 was significantly associated with high-risk adenomas and CMS3 with low-risk adenomas ($p = 0.025$; Figure 3). When each of the seven CAEs were examined, gain of 20q and loss of 18q were significantly associated with CMS2 ($p = 0.004$ and $p = 0.031$, respectively). No statistically significant associations were observed between CMS class and histological type ($p = 0.362$) and grade of dysplasia ($p = 0.389$), or between high-risk genotypic features and histological type ($p = 0.77$) and grade of dysplasia ($p = 0.079$; supplementary material, Table S7).

To explore associations of CMS2 and CMS3 adenomas with well-defined biological processes, we performed GSEA on the hallmark gene sets (Table 3) [30]. As expected, the gene sets enriched in CMS2 adenomas were involved in cell cycle and proliferation, including genes that are targets of E2F transcription factors, genes involved in G_2/M checkpoint, mitotic spindle assembly, the phosphoinositide 3-kinase (PI3K)–AKT–mammalian target of rapamycin (mTOR) pathway, and the Wnt–β-catenin signalling pathway, or genes regulated by MYC. These results are in line with the biological characterisation of CMS2 CRCs, which are known to be enriched with proliferation and cell cycle pathways [8]. Another gene set enriched in the CMS2 adenoma group was apical junction, this process also relates to increased proliferation. Additionally, CMS2 adenomas expressed genes involved in epithelial–mesenchymal transition, the transforming growth factor (TGF)-β signalling pathway, and the development of muscles, which are processes typically assigned to CMS4 CRCs, and genes involved in the immune response (coagulation), which are characteristic of CMS1 CRCs. Considering that the enrichment analysis in the original CMS CRC characterisation was performed by comparing each CMS class with the other three CMS classes, the fact CMS1-specific and CMS4-specific processes arose in the CMS2 versus CMS3 comparison does not represent a contadictory result, as a different analysis was performed in this study. On the other hand, the majority of gene sets enriched in CMS3 adenomas were metabolism-associated, including those involved in haem, fatty acid and sugar metabolism, which is in line with the original characterisation of the CMS3 'metabolic' CRC subtype.

To examine the differences between CMS2 and CMS3 adenomas in the context of CMS classes in CRC, 'stromal scores' and 'immune scores' from the ESTIMATE algorithm [33] and previously published 'invasive front'

Table 3. Gene sets enriched in CMS2 and CMS3 adenomas

Gene set	Process category	Size	Members in signal	Normalised enrichment score	*P* value	FDR
Gene sets enriched in CMS2 adenomas in comparison with CMS3 adenomas						
G_2M checkpoint	Proliferation	183	100	2.00	<0.001	<0.01
E2F targets	Proliferation	183	105	1.77	<0.001	0.01
MYC targets V2	Proliferation	57	78	1.57	0.005	0.02
Mitotic spindle	Proliferation	168	28	1.58	0.001	0.02
Epithelial–mesenchymal transition	Development	136	51	1.74	<0.001	0.01
Myogenesis	Development	103	42	1.67	0.001	0.01
PI3K–AKT–mTOR signalling	Signalling	85	38	1.62	0.004	0.01
Wnt–β-catenin signalling	Signalling	33	13	1.61	0.005	0.02
TGF-β signalling	Signalling	50	14	1.59	0.007	0.02
Coagulation	Immune	85	35	1.64	0.002	0.01
Apical junction	Cellular component	123	47	1.49	0.006	0.04
Gene sets enriched in CMS3 adenomas in comparison with CMS2 adenomas						
Protein secretion	Pathway	90	35	−1.78	<0.001	0.03
Glycolysis	Metabolic	169	45	−1.52	<0.001	0.08
Oxidative phosphorylation	Metabolic	194	85	−1.39	<0.001	0.13
Fatty acid metabolism	Metabolic	132	39	−1.35	0.017	0.13
Haem metabolism	Metabolic	144	33	−1.27	0.020	0.15
Oestrogen response late	Signalling	152	32	−1.30	0.020	0.15

Gene sets were grouped in process categories according to the original hallmark gene set grouping [32]. Size indicates number of genes in the gene set; members in signal indicates how many genes from the gene set contributed to the enrichment score. The statistical values, normalised enrichment score, *P* values and FDR were calculated with GSEA [30]. Gene sets enriched in CMS2 adenomas have positive enrichment scores, and gene sets enriched in CMS3 adenomas have negative enrichment scores.

and 'central tumour' signature enrichments were calculated [35] (supplementary material, Figure S5). As expected, 'stromal score' and tumour 'invasive front' signatures showed a high level of enrichment in CMS4 CRCs as compared with adenomas and other CMS CRC classes. The 'immune score' was enriched in CMS1 cancers as compared with CMS2–3 lesions, whereas the 'central tumour' signature showed similar results for all groups.

Validation in the independent series

Validation of the CMS classification results in colorectal adenomas was performed in an independent series – Series 3 (GSE20916) [29]. Series 3 consists of colorectal adenomas ($n = 45$) and cancers ($n = 36$) measured on the Affymetrix array. To perform a similar analysis as that used for the study dataset, CRCs from the GSE39582 dataset ($n = 566$) were chosen as the reference dataset for batch effect removal, normalisation, and CMS classification [3]. This reference dataset was the largest CRC series measured on the same platform as Series 3 and used in the original CMS classification publication [8]. CMS classes were extracted for CRCs and adenomas from Series 3 (Table 2; supplementary material, Table S8). CMS classification of colorectal adenomas in Series 3 confirmed the results obtained with the study dataset, with most adenomas being labelled as CMS3 ($n = 28$, 62%), none as CMS4, and a small number as CMS1 ($n = 1$, 2%) or CMS2 ($n = 2$, 11%) (Table 2; supplementary material, Table S8). In Series 3, the distribution of the CMS classes among adenomas differed significantly from that of the cancers from the same series ($p < 2.2 \times 10^{-16}$). No significant differences between the distribution of CMS classes among adenomas from the study dataset and those from the validation set were observed ($p = 0.13$).

Discussion

CMS classification constitutes an established consensus gene expression-based subtyping of CRC. We set out to determine whether this molecular classification is already present at the adenoma stage. Classification of adenomas according to CMS was achieved for 54 of 62 adenomas, in a group-wise analysis together with 59 CRCs from the study dataset and 556 TCGA CRC samples [15,36]. The results were validated in the independent series, in which 34 of 45 adenomas where classified with the same method; group-wise analysis including 36 CRCs from the same series and 566 CRCs from the reference dataset [3,29].

The distribution of CMS classes in adenomas differed significantly from that in CRCs, in both the study dataset and the validation dataset. The vast majority (73% and 62% for the study and validation sets, respectively) of adenomas were classified as the 'metabolic' CMS3 type, which was the least frequent CMS class among CRCs from the study dataset (3%). Multiple gene expression profiling studies of colorectal adenomas and CRCs have shown upregulated metabolism in adenomas. In particular, pathway analysis of genes overexpressed in adenomas in comparison with cancers revealed the same pathways that were dysregulated in CMS3, including fatty acid, amino acid and sugar metabolism [8,37,38]. It is evident that metabolic deregulation already occurs at the adenoma stage. In this study, GSEA comparing CMS2 and CMS3 adenomas confirmed enrichment of metabolic pathways in CMS3 adenomas. The results of this study imply that CMS3 is more representative of the adenoma than of the carcinoma stage. From the perspective of which adenomas have a risk of progressing to cancer, CMS3 may well represent low-risk adenomas, which was confirmed by the enrichment of low-risk adenomas in this class as defined by the presence of DNA copy number aberrations. As most adenomas never progress to cancer (95%), the observed frequency of CMS3 adenomas is consistent with this hypothesis.

Conversely, none of the adenomas from either the study dataset or the validation dataset were classified as the stroma-rich poor-prognosis CMS4 class. A process inherent to invasion and thus colorectal adenoma-to-carcinoma progression is activation of tumour stroma [21,39]. In fact, the tumour stroma represents an inflammatory response to foreign intruders, as well as being a scaffold for invading tumour cells. Mucosa of colorectal adenomas contains dysplastic epithelium as well as stroma (the lamina propria). In adenomas, this resembles the lamina propria of normal mucosa, being a framework of loose connective tissue, capillaries, myofibroblasts, and immune cells, and is quite different from the reactive stroma of cancers, which is the most prominent in CMS4 CRC. The lack of the mesenchymal subtype has also been observed for colorectal organoids, which are purely epithelial, and for patient-derived xenografts, in which the stroma is of mouse origin [40,41]. Multiple studies have shown that the CMS4 signature is mostly driven by stroma rather than epithelial cancer cells [41–43]. As the typical desmoplastic cancer stroma is, by definition, absent in adenomas, it is no surprise that no adenomas were classified as CMS4.

Regarding the CMS1 and CMS2 classes, the CMS classifier subtyped one of the adenomas with MSI as CMS1 and the second one as CMS3. MSI is rare in colorectal adenomas, with a prevalence of 3% overall [44], whereas approximately 15–20% of CRCs show MSI [45]. The observations in the present study are consistent with these data. When colorectal adenomas acquire MSI, they are considered to progress rapidly, leaving a small window of opportunity for them to be detected, resulting in the low frequency of MSI in colorectal adenomas. Not all adenomas with MSI were classified as CMS1, consistent with the observations made on CRCs with MSI, a subset of which were also classified as CMS3 [8]. Specific features that discriminate CMS1 CRCs with MSI and CMS3 CRCs with MSI have not been described yet. In the validation set, one adenoma was classified as CMS1 as well, but the MSI status of this adenoma is unknown. Eight of the adenomas were classified as CMS2 in the study dataset, and five in the validation set. From the perspective of adenoma-to-carcinoma progression, this is particularly interesting, as CMS2 represents canonical CRC carcinogenesis. Given that Wnt and MYC pathway activation occurs mostly in the transition from normal epithelium to adenoma, it may seem unexpected that CMS2 is not the predominant class within adenomas [46]. On the assumption that not the sequential order but the accumulation of mutations causes tumour progression, there must be more alterations in these adenomas to be classified as CMS2. Indeed, the enrichment of high-risk adenomas within CMS2 suggests that CMS2 adenomas might be closer to becoming malignant than those classified as CMS3. Additionally, the chromosomal gain of 20q and loss of 18q were found to occur more often in CMS2 adenomas. Gain of 20q is associated with a gene dosage effect of multiple genes [16], including *AURKA* and *TPX2*, which play a role in the G_2/M phase of the cell cycle [47]. This is consistent with the observed enrichment of the G_2/M checkpoint and mitotic spindle assembly gene sets in CMS2 adenomas. Another characteristic specific for adenoma-to-carcinoma progression and the CMS2 adenoma class is upregulation of pathways such as the cell cycle and epithelial differentiation [21]. In this study, GSEA confirmed that CMS2 adenomas have increased expression of genes involved in proliferation, the cell cycle and even epithelial–mesenchymal transition as compared with CMS3 adenomas. These findings are in line with CMS2 CRC characterisation as well as with the biological processes required for adenoma-to-carcinoma progression. Our results suggest that CMS2 adenomas, rather than CMS3 adenomas, may represent lesions at risk of becoming malignant. Owing to the lack of copy number information in Series 3, the association between risk of progression and CMS classification could not be further validated. Nevertheless, this association should be further investigated. Adenomas, once detected during colonoscopy, are completely removed, thereby interrupting their natural history in terms of either progressing to cancer or not. Currently, adenoma-to-carcinoma progression can only be studied *in vitro* by the use of, for example, organoid models. Although this has been done by perturbing frequently mutated cancer genes with prominent roles in CRC pathogenesis [48], relevant aspects of adenoma-to-carcinoma progression, including CIN, still remain to be incorporated in these model system studies.

The CMS classification of cancers revealed a relatively large number of CMS4 cases in the present series. Taking into account the different sample sizes of the current study and the original CMS publication, and given the variation in distributions of CMS classes among the six datasets from which the CMS classification originated [1–3,5–8], it may be that the CMS class distribution varies per dataset.

In the study dataset, we used large adenomas to sample fresh frozen material for research purposes, as well as routine tissue processing for diagnostics. Therefore, the majority (95%) of the adenomas were > 1 cm. Given the association of adenoma size with progression risk [49], the proportion of CMS3 could be even higher in smaller adenomas. The current study, however, does not allow conclusions to be drawn about the stage of development from normal epithelium to adenoma at which a CMS signature becomes detectable.

The present study focused on conventional adenomas, which are the most common precursors of CRC, especially in the context of CIN [50], representing the classic adenoma-to-carcinoma progression model. More recently, a serrated pathway has been introduced, with sessile serrated lesions being precursors of CRC [50]. The CMS classification of these lesions has already been presented besides the CMS classification of a small number of tubular adenomas, and resulted in a different distribution of the CMS classes from that observed in the current study [51]. However, given the highly selective composition of adenomas in this dataset and its considerable differences from our study cohort, significant variation in CMS classification is to be expected.

Technically, a combined analysis of the study dataset and the TCGA CRC series was performed to reduce the effect of the RNA-seq data normalisation on the CMS classification. Additionally, because of a further normalisation step implemented in the random forest CMS algorithm, combined analysis reduced the impact of the potentially different distribution of CMS classes in the study dataset from that in the original CMS training set. The concept of batch effect adjustment to a 'gold-standard' dataset, which the model was trained on, and classification by use of a merged dataset was previously introduced [26]. This approach proved to be appropriate for our research question by providing stability to the classifier in comparison with applying it on the study dataset alone (data not shown). The CMS classification of TCGA data performed in this study was not biased by our approach, as the original CMS labels for these samples were reassigned with an accuracy of 97%. Additionally, the CMS classification results for the adenomas were largely reproduced with the SSP CMS classifier (supplementary material, Tables S9 and S10). The SSP method is not sensitive to the composition of the dataset on which it is applied, so it did not require the context of a large series of CRCs or batch effect removal. Therefore, it is suitable for validation of the entire data analysis approach. The SSP method confirmed the CMS classes of adenomas to a large extent; however, in some cases, it lacked confidence in recognising CMS1 or CMS2 expression traits.

So far, classification of colorectal neoplasia has been morphology-based. Adenomas are classified on the basis of histological type, size and grade of dysplasia, whereas cancers are subtyped on the basis of grade of differentiation and stage. The CMS classification is an approach for molecular classification of cancers based on RNA expression. The present study has extended this approach to colorectal adenomas, and has demonstrated that CMS classification can be effectively applied to these lesions. In conclusion, colorectal adenomas proved to be heterogeneous in terms of CMS class, but with a different distribution from that of cancers. CMS3 turned out to be the most prevalent among the conventional adenomas, and our results indicate that it may represent mostly adenomas at low risk of progressing to CRC as compared with CMS1 or CMS2 adenomas. The frequency of CMS classes observed in adenomas is consistent with what could be expected on the basis of differences between adenomas and carcinomas, and on the proportion of adenomas expected to progress to cancer.

Acknowledgements

We thank ServiceXS (GenomeScan, Leiden) for sequencing the NGS-ProToCol sample Series 1 and the Genomics Core Facility (Netherlands Cancer Institute, Amsterdam) for additional sequencing. This study was performed within the framework of the Centre for Translational Molecular Medicine (CTMM) research programme, project NGS-ProToCol (03O-402), and was supported by Dutch Cancer Society (KWF Kankerbestrijding), project numbers 2013-6025 and 2013-5885.

Author contributions statement

MAK, LJWB, GB, YH, APS, GJ, BC, LFAW, CR, RJAF and GAM conceived the study and the experiments. NCTvG and GAM performed the histopathological review. LJWB, ASB, PDvD and MdJ performed experiments. MAK, GB and LFAW contributed to the design of the data analysis. MAK and CR performed the bioinformatics analysis. All authors were involved in writing the paper and gave final approval to the submitted and published versions.

References

1. Budinska E, Popovici V, Tejpar S, *et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol* 2013; **231:** 63–76.
2. De Sousa EMF, Wang X, Jansen M, *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med* 2013; **19:** 614–618.
3. Marisa L, de Reynies A, Duval A, *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* 2013; **10:** e1001453.
4. Perez-Villamil B, Romera-Lopez A, Hernandez-Prieto S, *et al.* Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer* 2012; **12:** 260.
5. Roepman P, Schlicker A, Tabernero J, *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer* 2014; **134:** 552–562.
6. Sadanandam A, Lyssiotis CA, Homicsko K, *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* 2013; **19:** 619–625.
7. Schlicker A, Beran G, Chresta CM, *et al.* Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics* 2012; **5:** 66.
8. Guinney J, Dienstmann R, Wang X, *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; **21:** 1350–1356.
9. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61:** 759–767.
10. Shinya H, Wolff WI. Morphology, anatomic distribution and cancer potential of colonic polyps. *Ann Surg* 1979; **190:** 679–683.
11. Rajagopalan H, Nowak MA, Vogelstein B, *et al.* The significance of unstable chromosomes in colorectal cancer. *Nat Rev Cancer* 2003; **3:** 695–701.
12. Hermsen M, Postma C, Baak J, *et al.* Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology* 2002; **123:** 1109–1119.

13. Meijer GA, Hermsen MA, Baak JP, *et al.* Progression from colorectal adenoma to carcinoma is associated with non-random chromosomal gains as detected by comparative genomic hybridisation. *J Clin Pathol* 1998; **51:** 901–909.
14. Douglas EJ, Fiegler H, Rowan A, *et al.* Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res* 2004; **64:** 4817–4825.
15. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487:** 330–337.
16. Sillars-Hardebol AH, Carvalho B, Tijssen M, *et al.* TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* 2012; **61:** 1568–1575.
17. Camps J, Grade M, Nguyen QT, *et al.* Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. *Cancer Res* 2008; **68:** 1284–1295.
18. Borras E, San Lucas FA, Chang K, *et al.* Genomic landscape of colorectal mucosa and adenomas in familial adenomatous polyposis. *Cancer Prev Res (Phila)* 2016; **9:** 417–427.
19. Hirsch D, Camps J, Varma S, *et al.* A new whole genome amplification method for studying clonal evolution patterns in malignant colorectal polyps. *Genes Chromosomes Cancer* 2012; **51:** 490–500.
20. Ried T, Knutzen R, Steinbeck R, *et al.* Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosomes Cancer* 1996; **15:** 234–245.
21. Sillars-Hardebol AH, Carvalho B, de Wit M, *et al.* Identification of key genes for carcinogenic pathways associated with colorectal adenoma-to-carcinoma progression. *Tumour Biol* 2010; **31:** 89–96.
22. Carvalho B, Postma C, Mongera S, *et al.* Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut* 2009; **58:** 79–89.
23. de Groen FL, Krijgsman O, Tijssen M, *et al.* Gene-dosage dependent overexpression at the 13q amplicon identifies DIS3 as candidate oncogene in colorectal cancer progression. *Genes Chromosomes Cancer* 2014; **53:** 339–348.
24. The Code of Conduct for the Use of Data in Health Research. [Accessed]. Available from: https://www.federa.org/codes-conduct
25. Belt EJ, Fijneman RJ, van den Berg EG, *et al.* Loss of lamin A/C expression in stage II and III colon cancer is associated with disease recurrence. *Eur J Cancer (Oxford, England: 1990)* 2011; **47:** 1837–1845.
26. Stein CK, Qu P, Epstein J, *et al.* Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics* 2015; **16:** 63.
27. Durinck S, Spellman PT, Birney E, *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009; **4:** 1184–1191.
28. Colorectal Cancer Subtyping Consortium (CRCSC). www.synapse.org. [Accessed 14 February 2017].
29. Skrzypczak M, Goryca K, Rubel T, *et al.* Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS ONE* 2010; **5**: e13091.
30. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; **102:** 15545–15550.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15:** 550.
32. Liberzon A, Birger C, Thorvaldsdóttir H, *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* 2015; **1:** 417–425.
33. Yoshihara K, Shahmoradgoli M, Martínez E, *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013; **4:** 2612.
34. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 2013; **14:** 7.
35. Dunne PD, McArt DG, Bradley CA, *et al.* Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. *Clin Cancer Res* 2016; **22:** 4095–4104.
36. NGS-ProToCol. Next Generation Sequencing from Prostate to Colorectal Cancer – Center for Translational Molecular Medicine (2014–2015). [Accessed]. Available from: http://www.ctmm.nl/en/projecten/translational-research-it-trait/ngs-protocol
37. Carvalho B, Sillars-Hardebol AH, Postma C, *et al.* Colorectal adenoma to carcinoma progression is accompanied by changes in gene expression associated with ageing, chromosomal instability, and fatty acid metabolism. *Cell Oncol (Dordr)* 2012; **35:** 53–63.
38. Pesson M, Volant A, Uguen A, *et al.* A gene expression and pre-mRNA splicing signature that marks the adenoma–adenocarcinoma progression in colorectal cancer. *PLoS ONE* 2014; **9**: e87761.
39. de Wit M, Carvalho B, Delis-van Diemen PM, *et al.* Lumican and versican protein expression are associated with colorectal adenoma-to-carcinoma progression. *PLoS ONE* 2017; **12**: e0174768.
40. Fujii M, Shimokawa M, Date S, *et al.* A colorectal tumor organoid library demonstrates progressive loss of niche factor requirements during tumorigenesis. *Cell Stem Cell* 2016; **18:** 827–838.
41. Isella C, Terrasi A, Bellomo SE, *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* 2015; **47:** 312–319.
42. Isella C, Brundu F, Bellomo SE, *et al.* Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat Commun* 2017; **8:** 15107.
43. Calon A, Lonardo E, Berenguer-Llergo A, *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* 2015; **47:** 320–329.
44. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell* 1996; **87:** 159–170.
45. Belt EJ, te Velde EA, Krijgsman O, *et al.* High lymph node yield is related to microsatellite instability in colon cancer. *Ann Surg Oncol* 2012; **19:** 1222–1230.
46. Rao CV, Yamada HY. Genomic instability and colon carcinogenesis: from the perspective of genes. *Front Oncol* 2013; **3**.
47. Asteriti IA, De Mattia F, Guarguaglini G. Cross-talk between AURKA and Plk1 in mitotic entry and spindle assembly. *Front Oncol* 2015; **5**.
48. Matano M, Date S, Shimokawa M, *et al.* Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat Med* 2015; **21:** 256–262.
49. Brenner H, Hoffmeister M, Stegmaier C, *et al.* Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. *Gut* 2007; **56:** 1585–1589.
50. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* 2007; **50:** 113–130.
51. Fessler E, Drost J, van Hooff SR, *et al.* TGFbeta signaling directs serrated adenomas to the mesenchymal colorectal cancer subtype. *EMBO Mol Med* 2016; **8:** 745–760.

*52. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 2014; **30:** 2114–2120.

*53. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)* 2009; **25:** 1754–1760.

*54. Picard Tools webpage. 2016. [Accessed 1 November 2016]. Available from: http://broadinstitute.github.io/picard/

*55. Scheinin I, Sie D, Bengtsson H, *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res* 2014; **24:** 2022–2032.

*56. van de Wiel MA, Brosens R, Eilers PH, *et al.* Smoothing waves in array CGH tumor profiles. *Bioinformatics (Oxford, England)* 2009; **25:** 1099–1104.

*57. Olshen AB, Venkatraman ES, Lucito R, *et al.* Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)* 2004; **5:** 557–572.

*58. van de Wiel MA, Kim KI, Vosse SJ, *et al.* CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics (Oxford, England)* 2007; **23:** 892–894.

*59. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 2013; **29:** 15–21.

*60. Harrow J, Frankish A, Gonzalez JM, *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012; **22:** 1760–1774.

*61. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* 2014; **30:** 923–930.

*62. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 2010; **26:** 139–140.

*63. Grossman RL, Heath AP, Ferretti V, *et al.* Toward a shared vision for cancer genomic data. *N Engl J Med* 2016; **375:** 1109–1112.

*64. Dai M, Wang P, Boyd AD, *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005; **33:** e175.

*Cited only in supplementary material.

SUPPLEMENTARY MATERIAL ONLINE

Supplementary materials and methods

Figure S1. Multidimensional scaling of the Euclidian distance between the gene expression profiles of all the samples from the study dataset with TCGA

Figure S2. Hierarchical clustering based on the gene expression profiles of the top 1000 most variable genes

Figure S3. Multidimensional scaling of the Euclidian distance between the gene expression profiles of all the samples for the validation set

Figure S4. Hierarchical clustering based on the gene expression profiles of the top 1000 most variable genes for the validation set

Figure S5. ESTIMATE scores and ssGSEA enrichment scores among CMS classes in adenomas and cancer

Table S1. Availability of the study data

Table S2. Comparison of the CMS classification of the TCGA data set in the current study to the original TCGA CMS labels

Table S3. CMS classification of colorectal cancers from the study dataset

Table S4. MSI samples in the study dataset: association between CMS classes in CRCs and differentiation grade, stage and MSI status

Table S5. CMS classification of adenomas from the study dataset

Table S6. Cancer-associated events (CAEs): DNA copy number aberrations and the risk of progression for adenomas in the study dataset

Table S7. Fisher exact test results for the association analysis in adenomas from the study dataset

Table S8. CMS classification of adenomas and cancers from the validation set

Table S9. CMS classification of colorectal adenomas from the study dataset performed with Single Sample Predictor

Table S10. Comparison of the CMS classification of colorectal adenomas by the study approach (random forest CMS classifier) and single sample predictor

SUPPLEMENTARY MATERIALS AND METHODS

SAMPLE COLLECTION

Series 1 originated from the NGS-ProToCol dataset[24], 60 snap-frozen colorectal tumors (30 colorectal polypoid adenomas and 30 colorectal carcinomas) were collected at the department of Pathology of the VU University Medical Center in Amsterdam, between 2011 and 2014. Excluded were patients below the age of 50, patients with Lynch syndrome, or patients who were known to have received radio- or chemotherapy before tumor removal. Samples were reviewed by an expert gastrointestinal pathologist and classified according to standard histopathological criteria. DNA and RNA were isolated from snap-frozen tissue pieces (Supplementary Materials and Methods). For Series 2, 32 colorectal polypoid adenomas and 29 colorectal carcinomas were collected at the department of Pathology of the VU University Medical Center in Amsterdam and described in a previous study[22]. RNA isolated from fresh frozen specimens of these samples was available.

DNA AND RNA ISOLATION FROM FRESH FROZEN TISSUE

Series 1: DNA and RNA were isolated from snap-frozen tissue pieces. For each piece a “before-” and “after-isolation” hematoxylin and eosin (HE)-slide was made. In between, tissue slides of 25 µm (for RNA isolation) or 15 µm (for DNA isolation) were cut. The HE slides were reviewed by an expert gastrointestinal pathologist. For most of the tissues (n=54) at least 70% of the tissue contained tumor cells. For six of them the tumor cell percentage was 60%. RNA was isolated from 30 to 40 25 µm slides using the miRNeasy Mini kit (QIAgen, Cat no 217004). The cut tissues were homogenized in 700 µl TRIzol (Invitrogen, Cat no 15596026), vortexed and incubated for 15-20 min at room temperature, followed by the manufacturer’s protocol. Genomic DNA was isolated from 15 to 25 15 µm slides, using the AllPrep DNA/RNA/miRNA Universal Kit (QIAgen, 80224), following the manufacturer’s protocol. DNA and RNA concentrations and purities were measured on a Nanodrop ND-1000 spectrophotometer (ThermoFisher Scientific). For the DNA samples also the double stranded DNA (dsDNA) concentrations were measured using the Qubit 3.0 Fluorometer using the dsDNA HS Assay Kit (Invitrogen, Cat no Q32851) DNA and RNA samples were also analyzed on a 0.8-1% agarose gel to confirm high molecular weight DNA or to check the RNA integrity.

LOW-COVERAGE WHOLE GENOME SEQUENCING AND DNA COPY NUMBER ANALYSIS

DNA from Series 1 was subjected to low-pass whole genome sequencing. The amount of dsDNA in the genomic DNA samples was quantified by using the Qubit, dsDNA HS Assay Kit (Invitrogen, Cat no Q32851). Up to 500 ng of dsDNA were fragmented by Covaris shearing to obtain fragment sizes of 160-180 bp. Samples were purified using 1.6X Agencourt AMPure XP PCR Purification beads according to manufacturer’s instructions (Beckman Coulter, Cat no A63881). The sheared DNA samples were quantified and qualified on a BioAnalyzer system using the DNA7500 assay kit (Agilent Technologies, cat no. 5067- 1506). With an input of maximum 1 µg sheared DNA, library preparation for Illumina sequencing was performed using the KAPA HTP Library Preparation Kit (KAPA Biosystems, Cat no KK8234). During library enrichment four to six PCR cycles were used to obtain enough yield for sequencing. After library preparation, the libraries were cleaned up using 1X AMPure XP beads. All DNA libraries were analyzed on a Caliper LabChip GX system with the HT DNA HiSense Reagent Kit (Caliper Life Sciences Inc, cat no. CLS760672) for determining the molarity. Up to 78 uniquely indexed samples were mixed together by equimolar pooling, in a final concentration of 10 nM, and subjected to sequencing on an Illlumina HiSeq 2500 machine in three lanes of a single read 65 bp run using V4 chemistry, according to manufacturer’s instructions. On average over 10M reads were obtained per adenoma sample.

Low quality reads and adapter sequences were trimmed with Trimmomatic version 3 to an average quality score for sliding window of 24 and quality of 26 both at the beginning and at the end of the sequences[26]. Reads were cropped to length of 50bp, shorter reads were removed. Trimmed reads were uniquely aligned to the human reference genome build hg19 using Burrows-Wheeler Aligner (“bwa aln”, allowing two mismatches and end-trimming of bases with qualities below 40, and “bwa samse” with default parameters)[27]. Reads identifiable as PCR duplicates were filtered out using Picard Tools MarkDuplicates version 2.7.1[28]. Read counting per bins, normalizations, corrections and filtering was done with Bioconductor package QDNAseq[29]. After median normalization, wave-correction was performed with an R package NoWaves[30]. Copy number segmentation was performed using Bioconductor package DNAcopy[31]. Gained and lost regions were identified using Bioconductor package CGHcall[32]. Copy number aberrations called with probability of more than 0.5 were taken along in further analysis.

RNA SEQUENCING AND DATA PRE-PROCESSING

Series 1: The NEBNext Ultra Directional RNA Library Prep Kit for Illumina with rRNA reduction was used to process the samples. The sample preparation was performed according to the protocol “NEBNext Ultra Directional RNA Library Prep Kit for Illumina” (NEB #E7420S/L and NEB #E6310S/L/X). Briefly, rRNA was reduced using RnaseH-based method. Then, fragmentation of the rRNA reduced RNA and a cDNA synthesis was performed. This was used for ligation with the sequencing adapters and PCR amplification of the resulting product. The quality and yield after sample preparation was measured with the Fragment Analyzer (Advanced Analytical). Clustering and DNA sequencing using the Illumina cBot and HiSeq 2500 was performed according manufacturer’s protocols. A concentration of 16.0 pM of DNA was used as input. HiSeq control software HCS v2.2.58 was used. Image analysis, base calling, and quality check was performed with the Illumina data analysis pipeline RTA v1.18.64 and Bcl2fastq v2.17. On average 67 million reads were obtained per sample.

Series 2: Quality and quantity of the total RNA was assessed by the 2100 Bioanalyzer using a Nano chip (Agilent). Total RNA samples having a RIN>8 were subjected to library generation. Strand-specific libraries were generated using the TruSeq Stranded mRNA sample preparation kit (Illumina Inc., Cat no RS-122-2101/2) according to the manufacturer's instructions (Illumina Inc., Cat no 15031047 Rev. E). Briefly, polyadenylated RNA from intact total RNA was purified using oligo-dT beads. Following purification the RNA was fragmented, random primed and reverse transcribed using SuperScript II Reverse Transcriptase (Invitrogen, Cat no 18064-014) with the addition of Actinomycin D. Second strand synthesis was performed using Polymerase I and RNaseH with replacement of dTTP for dUTP. The generated cDNA fragments were 3' end adenylated and ligated to Illumina Paired-end sequencing adapters and subsequently amplified by 12 cycles of PCR. The libraries were analyzed on a 2100 Bioanalyzer using a 7500 chip (Agilent), diluted and pooled equimolar into a 10 nM multiplex sequencing pool, containing 18 samples per pool. RNA sequencing was performed on an Illumina HiSeq V4 2500, using a 125 bases paired end run. On average 32 million reads were obtained per sample.

RNA-seq data preprocessing was performed for each series as follows. Low quality reads and adapter sequences were trimmed by Trimmomatic[26] version 3 to average quality score for sliding window of 24, and 26 for both leading and trailing part of the sequences. Minimum length was set to 36 bases. Mapping was performed with STAR aligner[34] version 2.4.2a to the human genome (USCS RefSeq hg38, annotation gencode v22[35]). Read counts per transcript were obtained with featureCounts from the Subread package v1.5.0-p2[36] with the gencode v22 annotation as reference. RPKM values were obtained with the use of rpkm function from the edgeR Bioconductor package version 3.12.1[37] and log2 transformed.

THE CANCER GENOME ATLAS CRC DATA

Gene expression data in the form of FPKM values were downloaded on 26.01.2017 from the NCI Genomic Data Commons (GDC) portal Release 1.4.1 for all the TCGA colorectal cancer samples (COAD and READ projects)[15,38]. The dataset was filtered for only primary tumors. 557 TCGA sample labels used by the CRC Subtyping Consortium in the original CMS classification were obtained from their Synapse instance[39,8] and used to filter the TCGA dataset; one TCGA label was missing from the GDC data portal. For the 556 TCGA samples FPKM values were log2 transformed forming a reference dataset for data normalization and CMS classification.

VALIDATION SET

CEL files were downloaded for datasets GSE20916 and GSE39582 from Gene Expression Omnibus and reprocessed to reflect current knowledge on the ensembl genome annotation. Probe sequences were realigned to the latest ensembl genome version using brainarray framework (first introduced by Dai et al.[43]), and re-normalization was performed using custom rma (affy). Ensembl IDs were translated to gene symbols with the use the latest edition of biomaRt version (ensembl 88, March 2017). The batch effect was removed with the use of M-Combat[40] as performed for Series 1 and 2 with the following changes; GSE39582 served as the reference dataset while Series 3 was the normalized batch. Evaluation of the batch effect removal and preservation of the differences between the adenomas and the cancers was performed with the multidimensional scaling algorithm on the Euclidian distance between the expression profiles and hierarchical clustering with complete linkage on the log2 expression values of the top 1000 variable genes (Supplementary Figure S3-4), respectively. Gene symbols were translated into Entrez ID with the use of the biomaRt Bioconductor package[41]. Reference dataset and Series 3 were merged after batch effect removal and CMS classification with the random forest algorithm was performed on the merged dataset. CMS classes were assigned as for the study dataset.

SUPPLEMENTARY FIGURES

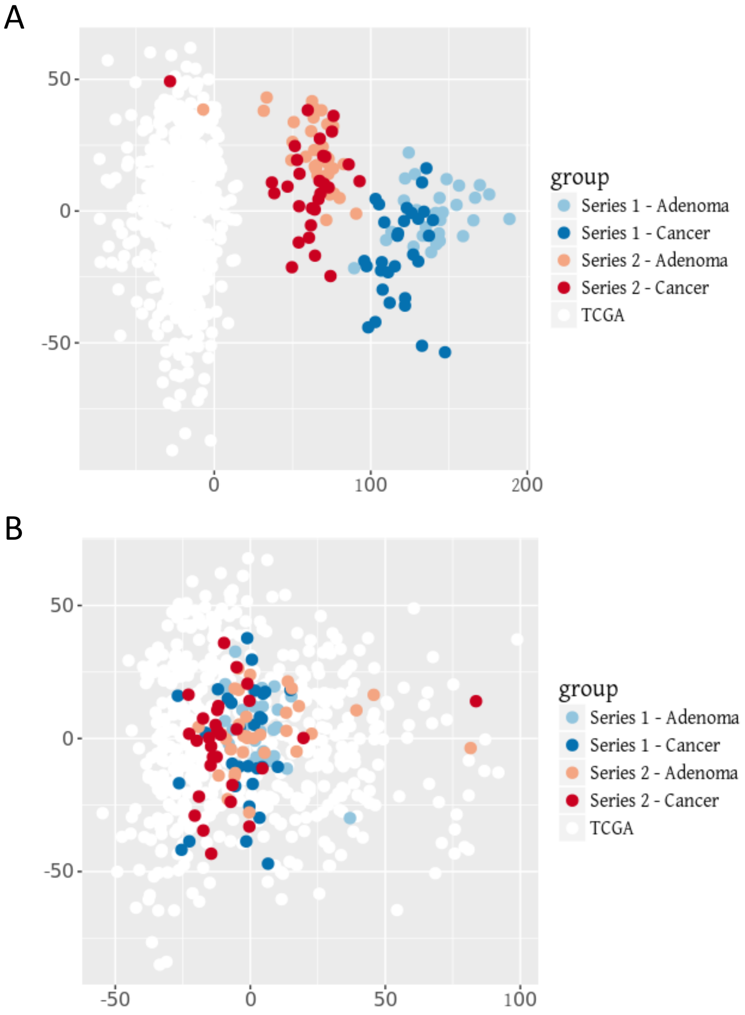
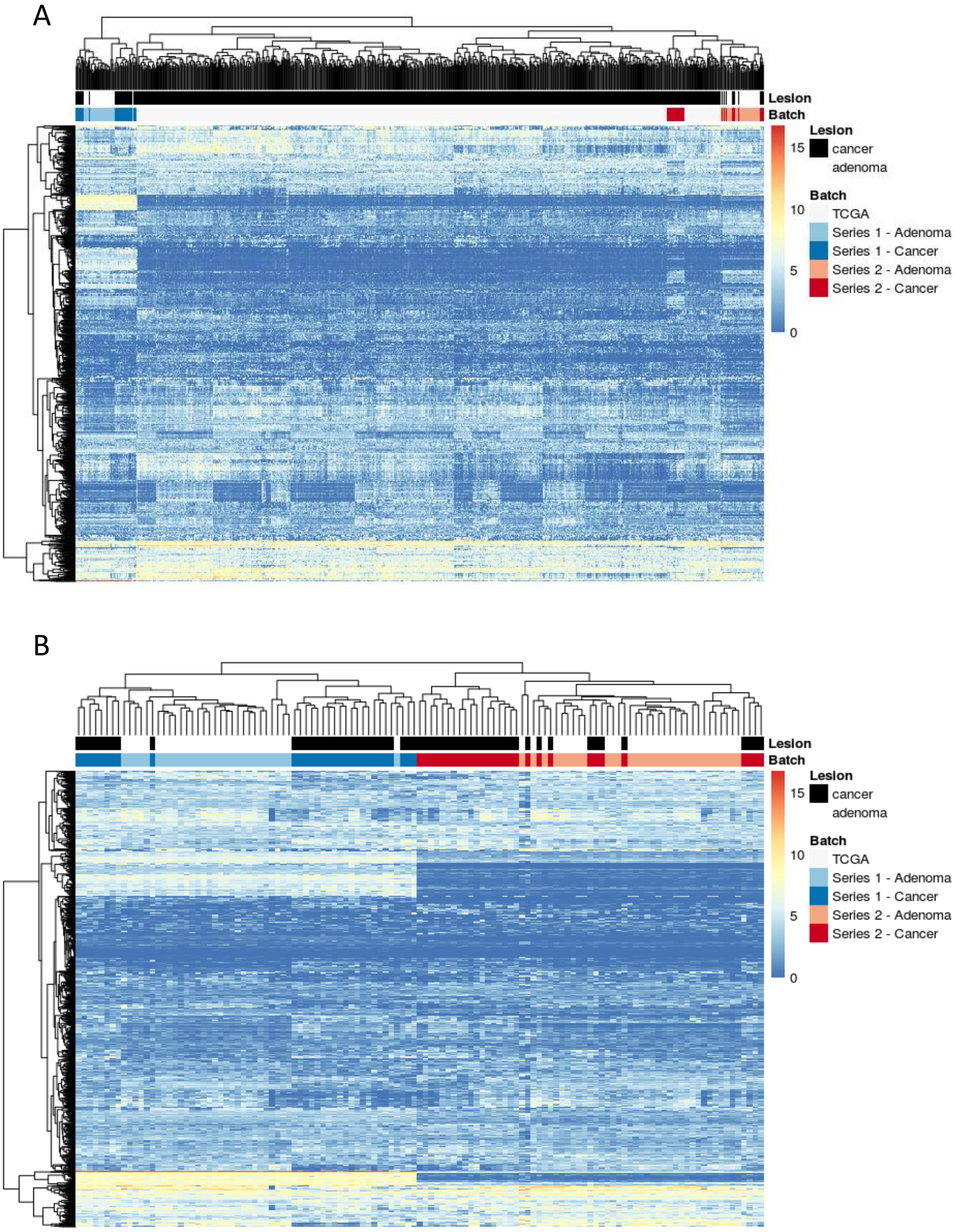


Figure S1. Multidimensional scaling of the Euclidian distance between the gene expression profiles of all the samples from the study dataset with TCGA A. Plot before batch effect removal. Three separate batches can be clearly distinguished, with white dots representing samples from the TCGA dataset, blue dots from the Series 1 and red dots from the Series 2. B. Plot after batch effect removal. The samples originating from different datasets cannot be distinguished by their location on the plot, indicating that the batch effect was removed.


A
Lesion
Batch
Lesion
cancer
adenoma
Batch
TCGA
Series 1 - Adenoma
Series 1 - Cancer
Series 2 - Adenoma
Series 2 - Cancer
15
10
5
0
B
Lesion
Batch
Lesion
cancer
adenoma
Batch
TCGA
Series 1 - Adenoma
Series 1 - Cancer
Series 2 - Adenoma
Series 2 - Cancer
15
10
5
0

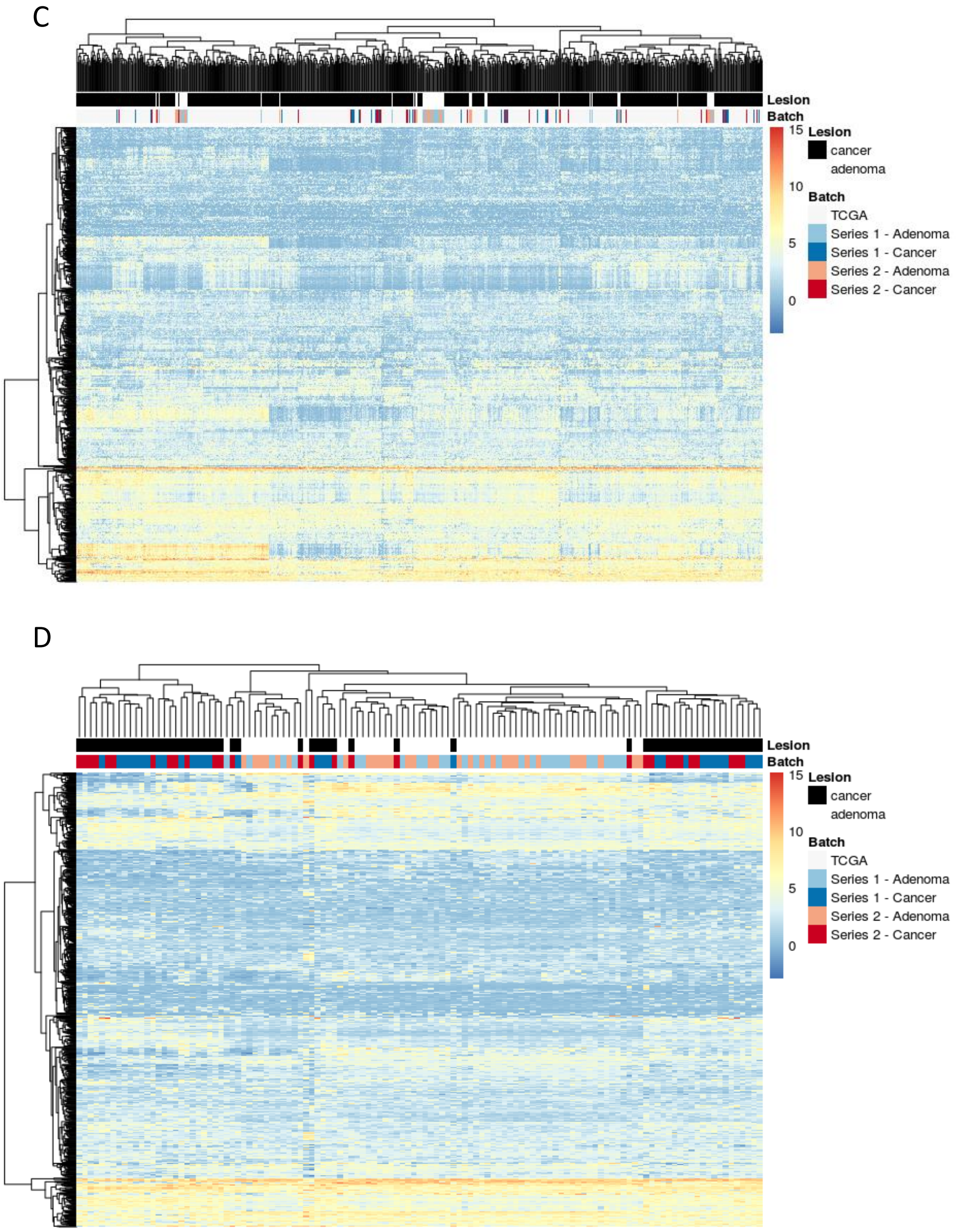


Figure S2. Hierarchical clustering based on the gene expression profiles of top 1000 most variable genes. A. Heatmap of all three datasets before batch effect removal. The batches corresponding to the TCGA dataset, Series 1 and Series 2 can be distinguished in the heatmap. B. Heatmap before batch correction of the Series 1 and Series 2 study datasets only. Next to the two batches, one can distinguish clusters enriched with adenomas and clusters enriched with cancers. C. Heatmap of all three datasets after batch effect removal. Samples from the three experiments do not cluster together. D. Heatmap of the Series 1 and Series 2 study datasets after batch effect removal. Clusters enriched with adenomas or cancers can still be distinguished, meaning that batch effect correction did not remove the variability between different lesions.

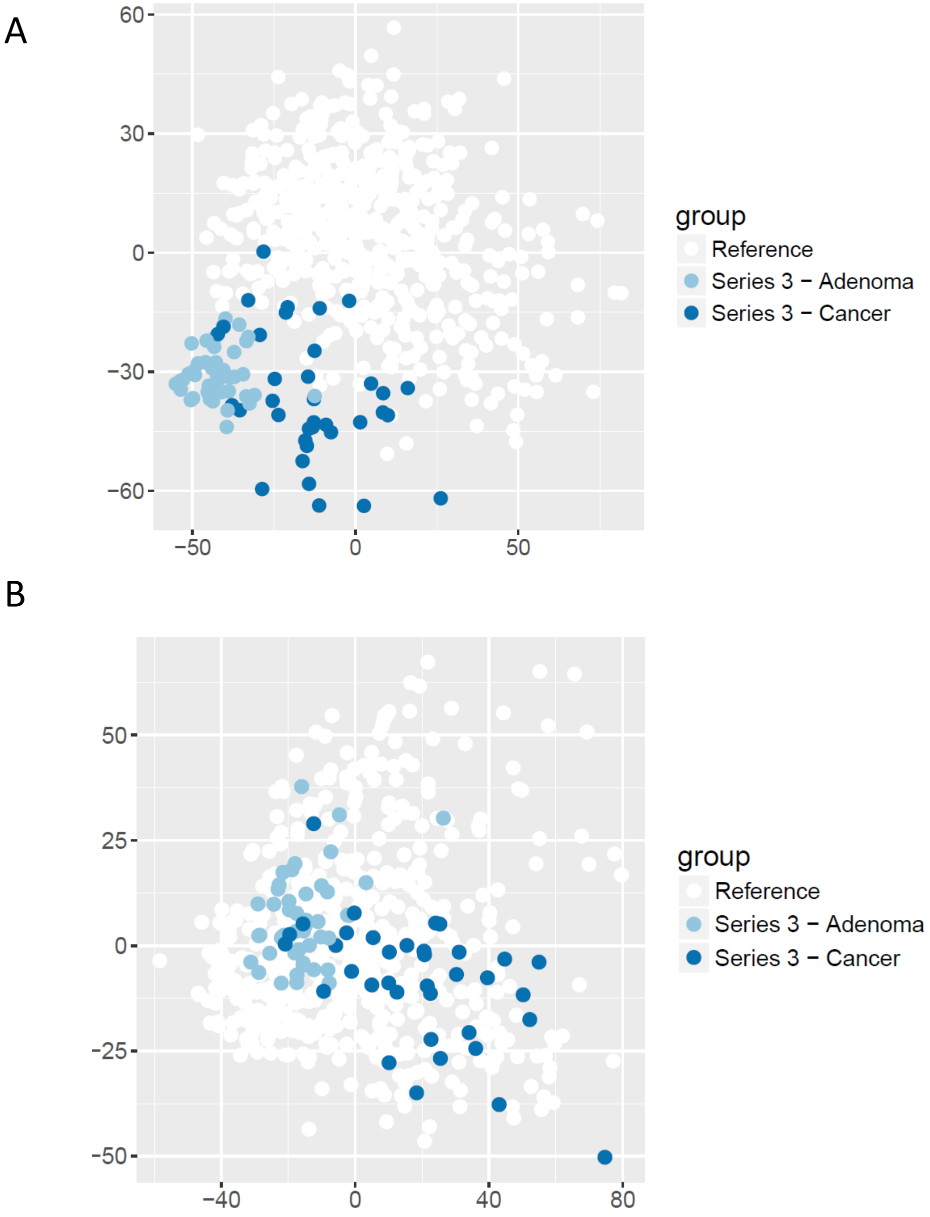
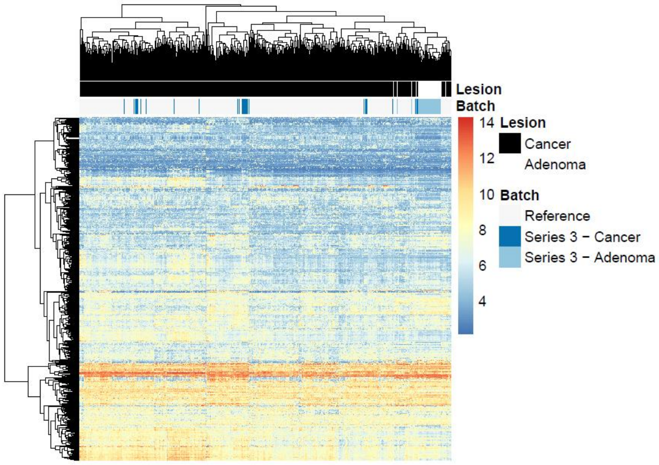
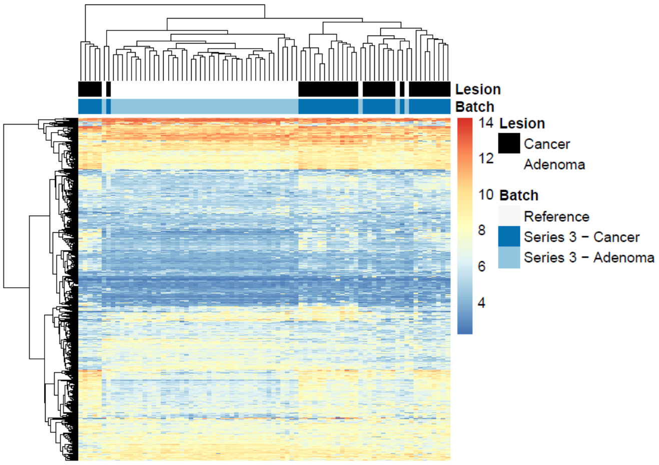


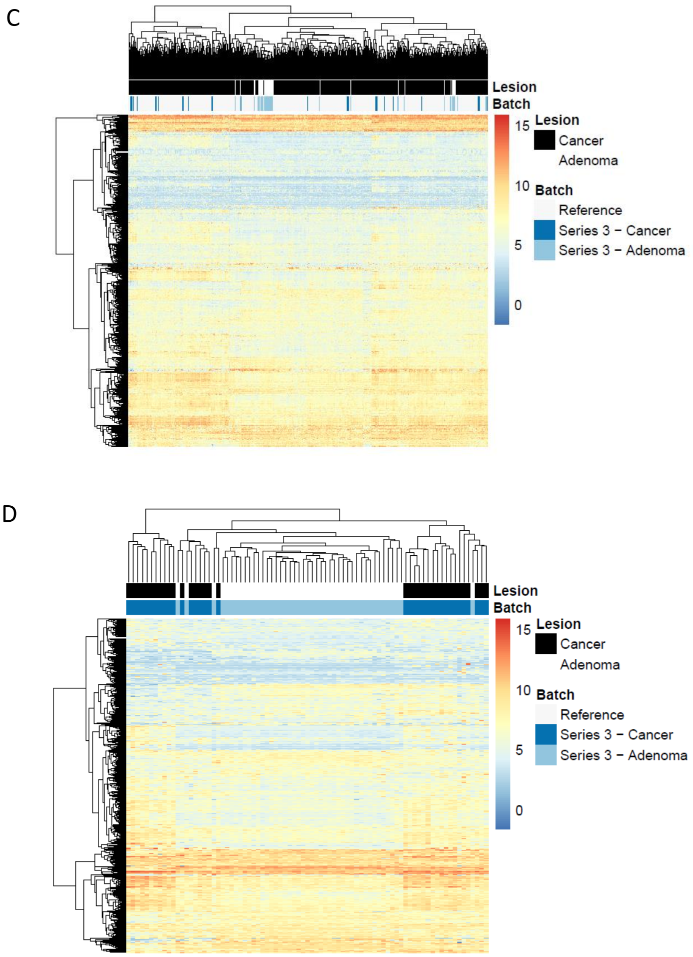
Figure S3. Multidimensional scaling of the Euclidian distance between the gene expression profiles of all the samples for the validation set Series 3 is the validation set with colorectal adenomas and cancers. Reference is the reference series with only colorectal cancers. A. Plot before batch effect removal. Two separate batches can be clearly distinguished, with white dots representing samples from the reference dataset and blue dots from the Series 3 B. Plot after batch effect removal. The samples originating from different datasets cannot be distinguished by their location on the plot, indicating that the batch effect was removed.

A


Lesion
Batch
14
12
10
8
6
4
Lesion
Cancer
Adenoma
Batch
Reference
Series 3 − Cancer
Series 3 − Adenoma

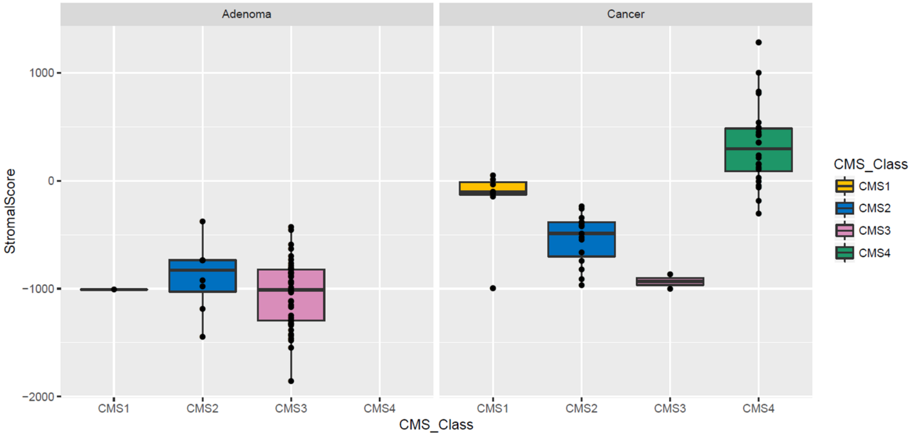
B


Lesion
Batch
14
12
10
8
6
4
Lesion
Cancer
Adenoma
Batch
Reference
Series 3 − Cancer
Series 3 − Adenoma

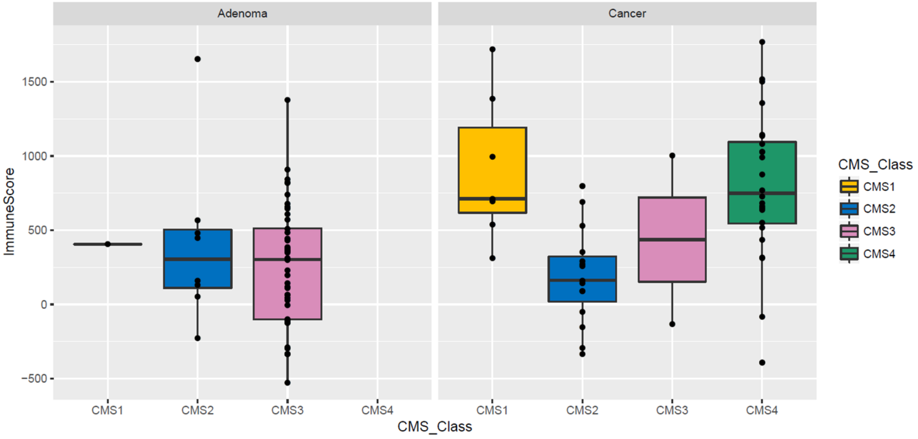


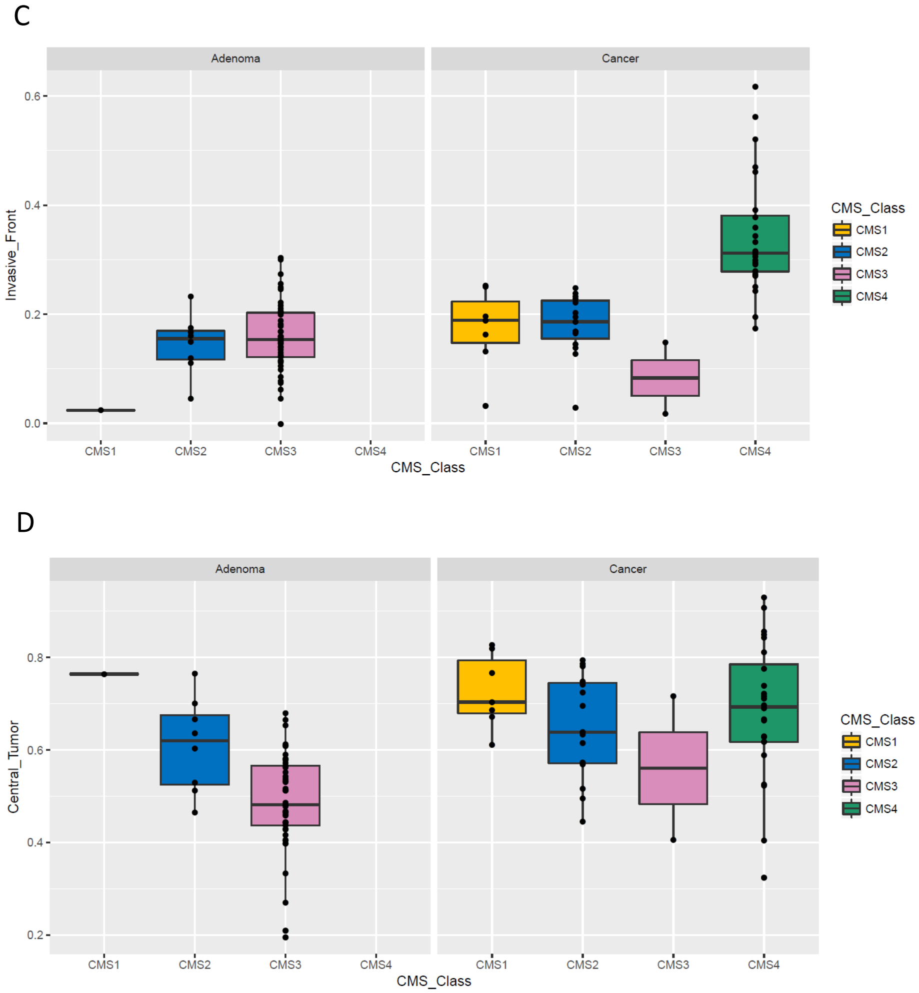
Supplementary Figure S4. Hierarchical clustering based on the gene expression profiles of top 1000 most variable genes. Reference is the reference dataset used for normalisation, Series 3 is the validation set. A. Heatmap of the two datasets before batch effect removal. The batches corresponding to the Reference and Series 3 can be distinguished in the heatmap. B. Heatmap before batch correction of the Series 3 only. Clusters enriched with adenomas and clusters enriched with cancers can be distinguished. C. Heatmap of the two datasets after batch effect removal. Samples from the two experiments do not cluster together. D. Heatmap of the Series 3 after batch effect removal. Clusters enriched with adenomas or cancers can still be distinguished, meaning that batch effect correction did not remove the variability between different lesions.

A


Adenoma
Cancer
StromalScore
1000
0
−1000
−2000
CMS1
CMS2
CMS3
CMS4
CMS1
CMS2
CMS3
CMS4
CMS_Class
CMS_Class
CMS1
CMS2
CMS3
CMS4

B


Adenoma
Cancer
ImmuneScore
1500
1000
500
0
−500
CMS1
CMS2
CMS3
CMS4
CMS1
CMS2
CMS3
CMS4
CMS_Class
CMS_Class
CMS1
CMS2
CMS3
CMS4



Supplementary Figure S5 ESTIMATE scores and ssGSEA enrichment scores among CMS classes in adenomas and cancer. “Stromal” (A) and “Immune (B) scores” were calculated with ESTIMATE algorithm and plotted per CMS group in colorectal adenomas and cancers. “Invasive front” (C) and “central tumor” (D) enrichment was calculated with ssGSEA algorithm.

SUPPLEMENTARY TABLES

Supplementary Table 1. Data availability for the data produced in this study. Raw data is available in European Genome-phenome Archive (EGA).

Series	Experiment	Lesion	EGA Dataset ID
1	low-coverage WGS	Adenoma	EGAD00001004092
1	RNA-seq	Adenoma	EGAD00001004055
1	RNA-seq	Cancer	EGAD00001004056
2	RNA-seq	Adenoma	EGAD00001004058
2	RNA-seq	Cancer	EGAD00001004059

Supplementary Table 2. Comparison of the CMS classification of the TCGA data set (n=556) in the current study to the original TCGA CMS labels.

CMS class obtained in this study

Original CMS class	**CMS1**	**CMS2**	**CMS3**	**CMS4**	Non-consensus	**Total**
CMS1	**63**	0	2	0	4	69
CMS2	0	**164**	2	0	13	179
CMS3	1	1	**69**	0	11	82
CMS4	1	7	0	**100**	28	136
Non-consensus	9	15	0	1	65	90
Total	74	187	73	101	121	556

Overall accuracy 96.59%

Supplementary Table 3. CMS classification of colorectal cancers from the study dataset. The analysis was performed on the merged dataset, including TCGA, Series 1 and Series 2, with the random forest CMS classifier. See the legend.

Legend

Column name	Explanation
Series	Sample series number
Sample name	Sample name
CMS1 posterior probability	posterior probability to be classified as CMS1
CMS2 posterior probability	posterior probability to be classified as CMS2
CMS3 posterior probability	posterior probability to be classified as CMS3
CMS4 posterior probability	posterior probability to be classified as CMS4
nearest CMS	CMS class with the highest posterior probability
predicted CMS	CMS class with the highest posterior probability that is equal to or more than 0.5

Series	Sample name	CMS1 posterior probability	CMS2 posterior probability	CMS3 posterior probability	CMS4 posterior probability	nearest CMS	predicted CMS
1	NGS-001-C	0.08	0.00	0.00	0.92	CMS4	**CMS4**
1	NGS-002-C	0.15	0.17	0.03	0.65	CMS4	**CMS4**
1	NGS-005-C	0.13	0.78	0.05	0.04	CMS2	**CMS2**
1	NGS-031-C	0.11	0.71	0.08	0.10	CMS2	**CMS2**
1	NGS-033-C	0.12	0.75	0.12	0.01	CMS2	**CMS2**
1	NGS-034-C	0.05	0.78	0.02	0.15	CMS2	**CMS2**
1	NGS-035-C	0.06	0.16	0.02	0.76	CMS4	**CMS4**
1	NGS-036-C	0.33	0.00	0.02	0.65	CMS4	**CMS4**
1	NGS-038-C	0.05	0.14	0.01	0.80	CMS4	**CMS4**
1	NGS-040-C	0.69	0.03	0.15	0.13	CMS1	**CMS1**
1	NGS-041-C	0.31	0.02	0.01	0.66	CMS4	**CMS4**
1	NGS-042-C	0.02	0.00	0.01	0.97	CMS4	**CMS4**
1	NGS-043-C	0.08	0.13	0.04	0.75	CMS4	**CMS4**
1	NGS-050-C	0.05	0.07	0.02	0.86	CMS4	**CMS4**
1	NGS-051-C	0.86	0.01	0.06	0.07	CMS1	**CMS1**
1	NGS-052-C	0.65	0.02	0.07	0.26	CMS1	**CMS1**
1	NGS-054-C	0.11	0.39	0.04	0.46	CMS4	**NA**
1	NGS-057-C	0.02	0.85	0.01	0.12	CMS2	**CMS2**
1	NGS-059-C	0.04	0.86	0.05	0.05	CMS2	**CMS2**
1	NGS-060-C	0.86	0.00	0.03	0.11	CMS1	**CMS1**
1	NGS-061-C	0.17	0.48	0.03	0.32	CMS2	**NA**
1	NGS-062-C	0.82	0.00	0.14	0.04	CMS1	**CMS1**
1	NGS-063-C	0.13	0.73	0.08	0.06	CMS2	**CMS2**
1	NGS-064-C	0.32	0.27	0.34	0.07	CMS3	**NA**
1	NGS-066-C	0.13	0.34	0.01	0.52	CMS4	**CMS4**
1	NGS-068-C	0.24	0.20	0.07	0.49	CMS4	**NA**
1	NGS-069-C	0.01	0.99	0.00	0.00	CMS2	**CMS2**
1	NGS-070-C	0.24	0.50	0.13	0.13	CMS2	**CMS2**
1	NGS-071-C	0.03	0.04	0.01	0.92	CMS4	**CMS4**
1	NGS-072-C	0.26	0.21	0.01	0.52	CMS4	**CMS4**
2	F11C	0.09	0.71	0.03	0.17	CMS2	**CMS2**
2	F12C	0.10	0.13	0.03	0.74	CMS4	**CMS4**
2	F14C	0.05	0.81	0.10	0.04	CMS2	**CMS2**
2	F15C	0.42	0.02	0.04	0.52	CMS4	**CMS4**
2	F27C	0.60	0.08	0.21	0.11	CMS1	**CMS1**
2	F28C	0.63	0.16	0.19	0.02	CMS1	**CMS1**
2	F1C	0.18	0.45	0.19	0.18	CMS2	**NA**
2	F30C	0.07	0.06	0.02	0.85	CMS4	**CMS4**
2	F33C	0.35	0.02	0.62	0.01	CMS3	**CMS3**
2	F34C	0.12	0.19	0.68	0.01	CMS3	**CMS3**
2	F35C	0.33	0.01	0.01	0.65	CMS4	**CMS4**
2	F36C	0.10	0.15	0.06	0.69	CMS4	**CMS4**
2	F39C	0.32	0.10	0.05	0.53	CMS4	**CMS4**
2	F46C	0.31	0.13	0.28	0.28	CMS1	**NA**
2	F51C	0.11	0.38	0.04	0.47	CMS4	**NA**
2	F3C	0.13	0.13	0.05	0.69	CMS4	**CMS4**
2	F53C	0.39	0.08	0.07	0.46	CMS4	**NA**
2	F54C	0.05	0.73	0.07	0.15	CMS2	**CMS2**
2	F57C	0.36	0.29	0.34	0.01	CMS1	**NA**
2	F58C	0.17	0.55	0.14	0.14	CMS2	**CMS2**
2	F6C	0.07	0.05	0.02	0.86	CMS4	**CMS4**
2	F29C	0.09	0.02	0.00	0.89	CMS4	**CMS4**
2	F18C	0.09	0.13	0.03	0.75	CMS4	**CMS4**
2	F25C	0.49	0.26	0.19	0.06	CMS1	**NA**
2	F31C	0.09	0.76	0.06	0.09	CMS2	**CMS2**
2	F67C	0.17	0.14	0.13	0.56	CMS4	**CMS4**
2	F8C	0.19	0.06	0.00	0.75	CMS4	**CMS4**
2	F10C	0.03	0.72	0.00	0.25	CMS2	**CMS2**
2	F19C	0.47	0.01	0.05	0.47	CMS1,CMS4	**NA**

Supplementary Table 4. MSI samples in the study dataset. Association between CMS classes in CRCs and differentiation grade, stage and MSI status.

Series	Sample name	Lesion
1	NGS-036-C	cancer
1	NGS-041-C	cancer
1	NGS-051-C	cancer
1	NGS-052-C	cancer
1	NGS-060-C	cancer
1	NGS-062-C	cancer
2	F15C	cancer
2	F19C	cancer
2	F25C	cancer
2	F33C	cancer
2	F35C	cancer
2	F57C	cancer
1	NGS-062-A	adenoma
1	NGS-027-A	adenoma

	CMS1	CMS2	CMS3	CMS4	p-value
Differentation grade					0.006
less/not	4	0	0	2	
well/moderate	3	15	2	21	
Stage					0.235
I	1	6	2	4	
II	4	3	0	12	
III	2	4	0	7	
IV	0	2	0	1	
Microsatellite status					0.004
MSI	4	0	1	4	
MSS	3	15	1	20	

Supplementary Table 5. CMS classification of adenomas from the study dataset. The analysis was performed on the merged dataset, including TCGA, Series 1 and Series 2, with the random forest CMS classifier. See the legend.

Legend

Column name	**Explanation**
Series	Sample series number
Sample name	Sample name
CMS1 posterior probability	posterior probability to be classified as CMS1
CMS2 posterior probability	posterior probability to be classified as CMS2
CMS3 posterior probability	posterior probability to be classified as CMS3
CMS4 posterior probability	posterior probability to be classified as CMS4
nearest CMS	CMS class with the highest posterior probability
predicted CMS	CMS class with the highest posterior probability that is equal to or more than 0.5

Series	**Sample name**	**CMS1 posterior probability**	**CMS2 posterior probability**	**CMS3 posterior probability**	**CMS4 posterior probability**	**nearest CMS**	**predicted CMS**
1	NGS-001-A	0.07	0.06	0.85	0.02	CMS3	**CMS3**
1	NGS-002-A	0.05	0.50	0.45	0.00	CMS2	**CMS2**
1	NGS-004-A	0.12	0.10	0.77	0.01	CMS3	**CMS3**
1	NGS-005-A	0.12	0.01	0.85	0.02	CMS3	**CMS3**
1	NGS-006-A	0.11	0.24	0.64	0.01	CMS3	**CMS3**
1	NGS-007-A	0.20	0.51	0.23	0.06	CMS2	**CMS2**
1	NGS-008-A	0.09	0.31	0.59	0.01	CMS3	**CMS3**
1	NGS-009-A	0.07	0.19	0.74	0.00	CMS3	**CMS3**
1	NGS-011-A	0.07	0.08	0.85	0.00	CMS3	**CMS3**
1	NGS-013-A	0.07	0.09	0.84	0.00	CMS3	**CMS3**
1	NGS-014-A	0.05	0.55	0.35	0.05	CMS2	**CMS2**
1	NGS-016-A	0.17	0.23	0.58	0.02	CMS3	**CMS3**
1	NGS-017-A	0.03	0.40	0.56	0.01	CMS3	**CMS3**
1	NGS-019-A	0.08	0.20	0.72	0.00	CMS3	**CMS3**
1	NGS-020-A	0.08	0.42	0.47	0.03	CMS3	**NA**
1	NGS-021-A	0.09	0.01	0.88	0.02	CMS3	**CMS3**
1	NGS-022-A	0.07	0.16	0.76	0.01	CMS3	**CMS3**
1	NGS-025-A	0.12	0.07	0.81	0.00	CMS3	**CMS3**
1	NGS-026-A	0.03	0.10	0.85	0.02	CMS3	**CMS3**
1	NGS-027-A	0.01	0.06	0.92	0.01	CMS3	**CMS3**
1	NGS-029-A	0.09	0.00	0.91	0.00	CMS3	**CMS3**
1	NGS-030-A	0.19	0.04	0.76	0.01	CMS3	**CMS3**
1	NGS-044-A	0.05	0.83	0.11	0.01	CMS2	**CMS2**
1	NGS-045-A	0.19	0.49	0.26	0.06	CMS2	**NA**
1	NGS-046-A	0.04	0.33	0.63	0.00	CMS3	**CMS3**
1	NGS-047-A	0.08	0.30	0.60	0.02	CMS3	**CMS3**
1	NGS-048-A	0.03	0.06	0.91	0.00	CMS3	**CMS3**
1	NGS-049-A	0.09	0.13	0.77	0.01	CMS3	**CMS3**
1	NGS-061-A	0.07	0.20	0.73	0.00	CMS3	**CMS3**
1	NGS-062-A	0.62	0.01	0.37	0.00	CMS1	**CMS1**
2	F20A	0.21	0.58	0.12	0.09	CMS2	**CMS2**
2	F23A1	0.12	0.06	0.81	0.01	CMS3	**CMS3**
2	F22A	0.15	0.08	0.74	0.03	CMS3	**CMS3**
2	F26A	0.15	0.24	0.51	0.10	CMS3	**CMS3**
2	F37A	0.06	0.44	0.48	0.02	CMS3	**NA**
2	F38A1	0.25	0.11	0.60	0.04	CMS3	**CMS3**
2	F38A2	0.09	0.11	0.77	0.03	CMS3	**CMS3**
2	F40A	0.03	0.33	0.59	0.05	CMS3	**CMS3**
2	F21A	0.07	0.10	0.83	0.00	CMS3	**CMS3**
2	F42A	0.01	0.15	0.84	0.00	CMS3	**CMS3**
2	F43A	0.09	0.01	0.90	0.00	CMS3	**CMS3**
2	F44A	0.10	0.03	0.86	0.01	CMS3	**CMS3**
2	F45A	0.11	0.06	0.78	0.05	CMS3	**CMS3**
2	F47A	0.05	0.09	0.86	0.00	CMS3	**CMS3**
2	F48A	0.26	0.27	0.21	0.26	CMS2	**NA**
2	F50A	0.06	0.33	0.61	0.00	CMS3	**CMS3**

2	F52A	0.01	0.67	0.30	0.02	CMS2	**CMS2**
2	F56A	0.05	0.12	0.83	0.00	CMS3	**CMS3**
2	F58A1	0.11	0.04	0.84	0.01	CMS3	**CMS3**
2	F59A	0.09	0.23	0.68	0.00	CMS3	**CMS3**
2	F60A	0.24	0.01	0.75	0.00	CMS3	**CMS3**
2	F61A	0.39	0.12	0.46	0.03	CMS3	**NA**
2	F62A	0.04	0.56	0.40	0.00	CMS2	**CMS2**
2	F64A	0.09	0.68	0.20	0.03	CMS2	**CMS2**
2	F66A	0.05	0.19	0.76	0.00	CMS3	**CMS3**
2	F68A	0.07	0.35	0.58	0.00	CMS3	**CMS3**
2	F25A	0.25	0.25	0.50	0.00	CMS3	**CMS3**
2	F17A	0.08	0.24	0.68	0.00	CMS3	**CMS3**
2	F65A	0.40	0.14	0.38	0.08	CMS1	**NA**
2	F7A	0.25	0.29	0.45	0.01	CMS3	**NA**
2	F63A	0.34	0.24	0.33	0.09	CMS1	**NA**
2	F23A2	0.13	0.11	0.73	0.03	CMS3	**CMS3**

Supplementary Table 6. Cancer associated events (CAEs) - DNA copy number aberrations and the risk of progression for the adenomas from the study dataset. See legend.

Legend:

Value	Explanation
0	no copy number aberration
1	copy number aberration present
high	2 out of 7 aberrations present
low	less than 2 out of 7 aberrations present
MSI	MSI lesion, the CAEs definition does not apply
NA	not available

Series	Sample name	Cancer associated events							Risk of progression
		8q	13q	20q	8p	15q	17p	18q	
1	NGS-001-A	0	0	0	0	0	0	0	**low**
1	NGS-002-A	0	1	1	0	0	0	0	**high**
1	NGS-004-A	0	1	0	0	0	0	0	**low**
1	NGS-005-A	0	0	0	0	0	0	0	**low**
1	NGS-006-A	0	NA	NA	0	0	0	NA	NA
1	NGS-007-A	0	0	0	0	0	0	0	**low**
1	NGS-008-A	0	0	0	0	0	0	0	**low**
1	NGS-009-A	0	NA	1	0	0	0	0	NA
1	NGS-011-A	0	0	0	0	0	0	0	**low**
1	NGS-013-A	0	0	0	0	0	0	1	**low**
1	NGS-014-A	1	1	1	0	0	0	1	**high**
1	NGS-016-A	1	1	1	0	0	0	1	**high**
1	NGS-017-A	0	1	1	0	0	0	0	**high**
1	NGS-019-A	1	1	1	0	0	0	0	**high**
1	NGS-020-A	1	1	1	0	0	0	0	**high**
1	NGS-021-A	0	0	0	0	0	0	0	**low**
1	NGS-022-A	0	0	0	0	0	0	0	**low**
1	NGS-025-A	0	0	0	0	0	0	0	**low**
1	NGS-026-A	0	0	0	0	0	0	0	**low**
1	NGS-027-A	0	0	0	0	0	0	0	MSI
1	NGS-029-A	0	0	0	0	0	0	0	**low**
1	NGS-030-A	0	0	0	0	0	0	0	**low**
1	NGS-044-A	0	0	1	1	0	0	1	**high**
1	NGS-045-A	0	0	1	1	0	0	1	**high**
1	NGS-046-A	0	0	0	0	0	0	0	**low**
1	NGS-047-A	1	1	1	0	0	0	0	**high**
1	NGS-048-A	0	0	0	0	0	0	0	**low**
1	NGS-049-A	0	0	0	0	0	0	0	**low**

1	NGS-061-A	0	0	0	0	0	0	0	**low**
1	NGS-062-A	0	0	0	0	0	0	0	MSI
2	F17A	0	0	0	0	0	0	0	**low**
2	F20A	0	NA	0	0	0	0	1	NA
2	F21A	0	0	0	0	0	0	0	**low**
2	F22A	0	0	0	1	0	1	0	**high**
2	F23A1	0	0	0	0	0	0	0	**low**
2	F23A2	0	1	0	0	0	1	0	**high**
2	F25A	0	0	0	0	0	0	0	**low**
2	F26A	0	0	0	0	0	0	0	**low**
2	F37A	0	0	0	0	0	0	0	**low**
2	F38A1	0	0	0	0	0	0	0	**low**
2	F38A2	0	0	0	0	0	0	0	**low**
2	F40A	0	0	0	0	0	0	0	**low**
2	F42A	NA	NA	NA	NA	NA	NA	NA	NA
2	F43A	NA	NA	NA	NA	NA	NA	NA	NA
2	F44A	0	0	0	0	0	0	0	**low**
2	F45A	NA	NA	NA	NA	NA	NA	NA	NA
2	F47A	0	0	0	0	0	0	0	**low**
2	F48A	0	NA	1	0	0	0	0	NA
2	F50A	0	0	1	0	0	0	1	**high**
2	F52A	NA	NA	NA	NA	NA	NA	NA	NA
2	F56A	0	0	0	0	0	0	0	**low**
2	F58A1	0	0	0	0	0	0	0	**low**
2	F59A	0	0	0	0	0	0	0	**low**
2	F60A	0	1	0	0	0	0	0	**low**
2	F61A	0	0	0	0	0	0	0	**low**
2	F62A	0	0	1	0	0	0	0	**low**
2	F63A	0	0	0	0	0	0	0	**low**
2	F64A	0	1	1	0	0	0	0	**high**
2	F65A	0	0	0	0	0	0	0	**low**
2	F66A	0	0	0	0	0	0	0	**low**
2	F68A	0	0	0	0	0	0	0	**low**
2	F7A	0	0	0	0	0	0	0	**low**

Supplementary Table 7. Fisher exact test results for the association analysis in adenomas from the study dataset. Associations analysed were CMS class, cancer associated events (DNA copy number aberrations), the risk of progression to CRC and clinical features.

Comparison	CMS2	CMS3	p-value	Odds ratio
High-risk	4	7	**0.025**	8.54
Low-risk	2	32		
8q gain	1	3	0.472	2.12
no 8q gain	6	39		
13q gain	3	7	0.140	3.62
no 13q gain	4	35		
20q gain	5	6	**0.004**	13.76
no 20q gain	2	36		
8p loss	1	1	0.270	6.40
no 8p loss	6	41		
15q loss	0	0	1.000	0.00
no 15g loss	7	42		
17p loss	0	2	1.000	0.00
no 17p loss	7	40		
18q loss	3	3	**0.031**	9.00
no 18q loss	4	39		

Comparison	High-risk	Low-risk	p-value	Odds ratio
Histological type			0.770	-
Tubular	4	11		
Tubulovillous	9	24		
Villous	0	4		
Dysplasia			0.079	3.77
High grade	7	9		
Low grade	6	30		
Comparison	**CMS2**	**CMS3**	**p-value**	**Odds ratio**
Histological type			0.362	-
Tubular	4	12		
Tubulovillous	4	26		
Villous	0	7		
Dysplasia			0.389	2.07
High grade	3	10		
Low grade	5	35		

Supplementary Table 8. CMS classification of adenomas and cancers from the validation set. The analysis was performed on the merged dataset, including reference dataset and Series 3, with the random forest CMS classifier. See the legend.

Legend

Column name	Explanation
Series	Sample series number
Sample name	Sample name
CMS1 posterior probability	posterior probability to be classified as CMS1
CMS2 posterior probability	posterior probability to be classified as CMS2
CMS3 posterior probability	posterior probability to be classified as CMS3
CMS4 posterior probability	posterior probability to be classified as CMS4
nearest CMS	CMS class with the highest posterior probability
predicted CMS	CMS class with the highest posterior probability that is equal to or more than 0.5

Series	Sample name	Lesion	CMS1 posterior probability	CMS2 posterior probability	CMS3 posterior probability	CMS4 posterior probability	nearest CMS	predicted CMS
3	GSM523287	Adenoma	0.21	0.25	0.46	0.08	CMS3	**NA**
3	GSM523288	Adenoma	0.15	0.53	0.3	0.02	CMS2	**CMS2**
3	GSM523293	Adenoma	0.19	0.44	0.26	0.11	CMS2	**NA**
3	GSM523294	Adenoma	0.12	0.26	0.56	0.06	CMS3	**CMS3**
3	GSM523295	Adenoma	0.14	0.11	0.74	0.01	CMS3	**CMS3**
3	GSM523300	Adenoma	0.11	0.12	0.76	0.01	CMS3	**CMS3**
3	GSM523301	Adenoma	0.2	0.13	0.66	0.01	CMS3	**CMS3**
3	GSM523302	Adenoma	0.11	0.43	0.46	0	CMS3	**NA**
3	GSM523307	Adenoma	0.27	0.05	0.67	0.01	CMS3	**CMS3**
3	GSM523308	Adenoma	0.04	0.25	0.71	0	CMS3	**CMS3**
3	GSM523309	Adenoma	0.04	0.68	0.28	0	CMS2	**CMS2**
3	GSM523315	Adenoma	0.13	0.37	0.45	0.05	CMS3	**NA**
3	GSM523319	Adenoma	0.2	0.15	0.64	0.01	CMS3	**CMS3**
3	GSM523320	Adenoma	0.06	0.59	0.35	0	CMS2	**CMS2**
3	GSM523321	Adenoma	0.12	0.18	0.69	0.01	CMS3	**CMS3**
3	GSM523322	Adenoma	0.12	0.25	0.63	0	CMS3	**CMS3**
3	GSM523327	Adenoma	0.36	0.1	0.45	0.09	CMS3	**NA**
3	GSM523328	Adenoma	0.06	0.42	0.52	0	CMS3	**CMS3**
3	GSM523329	Adenoma	0.33	0.07	0.6	0	CMS3	**CMS3**
3	GSM523333	Adenoma	0.07	0.47	0.46	0	CMS2	**NA**
3	GSM523334	Adenoma	0.14	0.14	0.72	0	CMS3	**CMS3**
3	GSM523335	Adenoma	0.08	0.65	0.25	0.02	CMS2	**CMS2**
3	GSM523336	Adenoma	0.11	0.41	0.47	0.01	CMS3	**NA**
3	GSM523340	Adenoma	0.05	0.18	0.77	0	CMS3	**CMS3**
3	GSM523341	Adenoma	0.11	0.27	0.62	0	CMS3	**CMS3**
3	GSM523342	Adenoma	0.15	0.12	0.73	0	CMS3	**CMS3**
3	GSM523346	Adenoma	0.16	0.2	0.63	0.01	CMS3	**CMS3**
3	GSM523347	Adenoma	0.09	0.38	0.51	0.02	CMS3	**CMS3**
3	GSM523348	Adenoma	0.11	0.35	0.52	0.02	CMS3	**CMS3**
3	GSM523353	Adenoma	0.1	0.11	0.79	0	CMS3	**CMS3**
3	GSM523354	Adenoma	0.06	0.58	0.33	0.03	CMS2	**CMS2**
3	GSM523355	Adenoma	0.31	0.02	0.66	0.01	CMS3	**CMS3**
3	GSM523356	Adenoma	0.11	0.18	0.71	0	CMS3	**CMS3**
3	GSM523361	Adenoma	0.12	0.07	0.81	0	CMS3	**CMS3**
3	GSM523362	Adenoma	0.1	0.1	0.8	0	CMS3	**CMS3**
3	GSM523363	Adenoma	0.19	0.35	0.45	0.01	CMS3	**NA**
3	GSM523367	Adenoma	0.13	0.13	0.74	0	CMS3	**CMS3**
3	GSM523374	Adenoma	0.34	0.4	0.23	0.03	CMS2	**NA**
3	GSM523375	Adenoma	0.07	0.18	0.75	0	CMS3	**CMS3**
3	GSM523376	Adenoma	0.1	0.11	0.77	0.02	CMS3	**CMS3**
3	GSM523381	Adenoma	0.11	0.11	0.78	0	CMS3	**CMS3**
3	GSM523383	Adenoma	0.56	0.1	0.3	0.04	CMS1	**CMS1**
3	GSM523384	Adenoma	0.2	0.32	0.44	0.04	CMS3	**NA**
3	GSM523385	Adenoma	0.21	0.34	0.4	0.05	CMS3	**NA**
3	GSM523386	Adenoma	0.15	0.1	0.75	0	CMS3	**CMS3**
3	GSM523283	Cancer	0.51	0.03	0.04	0.42	CMS1	**CMS1**
3	GSM523284	Cancer	0.08	0.02	0.02	0.88	CMS4	**CMS4**

3	GSM523285	Cancer	0.07	0.02	0.01	0.9	CMS4	**CMS4**
3	GSM523292	Cancer	0.51	0.17	0.23	0.09	CMS1	**CMS1**
3	GSM523296	Cancer	0.23	0.06	0.04	0.67	CMS4	**CMS4**
3	GSM523298	Cancer	0.07	0.84	0.04	0.05	CMS2	**CMS2**
3	GSM523303	Cancer	0.23	0.04	0.02	0.71	CMS4	**CMS4**
3	GSM523305	Cancer	0.13	0.1	0.04	0.73	CMS4	**CMS4**
3	GSM523306	Cancer	0.19	0.29	0.05	0.47	CMS4	**NA**
3	GSM523312	Cancer	0.29	0.02	0.07	0.62	CMS4	**CMS4**
3	GSM523313	Cancer	0.11	0.73	0.14	0.02	CMS2	**CMS2**
3	GSM523316	Cancer	0.16	0.08	0.03	0.73	CMS4	**CMS4**
3	GSM523317	Cancer	0.1	0.27	0.01	0.62	CMS4	**CMS4**
3	GSM523318	Cancer	0.72	0	0.07	0.21	CMS1	**CMS1**
3	GSM523323	Cancer	0.06	0	0.01	0.93	CMS4	**CMS4**
3	GSM523325	Cancer	0.11	0.15	0.02	0.72	CMS4	**CMS4**
3	GSM523326	Cancer	0.04	0.62	0.21	0.13	CMS2	**CMS2**
3	GSM523331	Cancer	0.59	0.05	0.19	0.17	CMS1	**CMS1**
3	GSM523332	Cancer	0.42	0.1	0.07	0.41	CMS1	**NA**
3	GSM523337	Cancer	0.24	0.35	0.09	0.32	CMS2	**NA**
3	GSM523339	Cancer	0.05	0.03	0.02	0.9	CMS4	**CMS4**
3	GSM523344	Cancer	0.07	0.05	0.01	0.87	CMS4	**CMS4**
3	GSM523345	Cancer	0.45	0.18	0.23	0.14	CMS1	**NA**
3	GSM523350	Cancer	0.15	0.01	0.01	0.83	CMS4	**CMS4**
3	GSM523351	Cancer	0.8	0.02	0	0.18	CMS1	**CMS1**
3	GSM523352	Cancer	0.26	0.17	0.5	0.07	CMS3	**CMS3**
3	GSM523357	Cancer	0.17	0.06	0.04	0.73	CMS4	**CMS4**
3	GSM523359	Cancer	0.09	0.04	0.02	0.85	CMS4	**CMS4**
3	GSM523366	Cancer	0.07	0	0.01	0.92	CMS4	**CMS4**
3	GSM523368	Cancer	0.23	0.55	0.1	0.12	CMS2	**CMS2**
3	GSM523370	Cancer	0.13	0	0.02	0.85	CMS4	**CMS4**
3	GSM523371	Cancer	0.19	0.6	0.21	0	CMS2	**CMS2**
3	GSM523372	Cancer	0.15	0.71	0.14	0	CMS2	**CMS2**
3	GSM523378	Cancer	0.13	0	0.01	0.86	CMS4	**CMS4**
3	GSM523379	Cancer	0.39	0.25	0.32	0.04	CMS1	**NA**
3	GSM523380	Cancer	0.12	0.79	0.03	0.06	CMS2	**CMS2**

Supplementary Table 9. CMS classification of colorectal adenomas from the study dataset performed with Single Sample Predictor. The classifier was applied on adenomas from Series 1 and Series 2 before batch effect removal.

Legend

Column name	Explanation
Series	Sample series number
Sample name	Sample name
minimum correlation to CMS1	minimum correlation of the gene expression profile to the CMS1 centroid
minimum correlation to CMS2	minimum correlation of the gene expression profile to the CMS2 centroid
minimum correlation to CMS3	minimum correlation of the gene expression profile to the CMS3 centroid
minimum correlation to CMS4	minimum correlation of the gene expression profile to the CMS4 centroid
median correlation to CMS1	median correlation of the gene expression profile to the CMS1 centroid
median correlation to CMS2	median correlation of the gene expression profile to the CMS2 centroid
median correlation to CMS3	median correlation of the gene expression profile to the CMS3 centroid
median correlation to CMS4	median correlation of the gene expression profile to the CMS4 centroid
maximum correlation to CMS1	maximum correlation of the gene expression profile to the CMS1 centroid
maximum correlation to CMS2	maximum correlation of the gene expression profile to the CMS2 centroid
maximum correlation to CMS3	maximum correlation of the gene expression profile to the CMS3 centroid
maximum correlation to CMS4	maximum correlation of the gene expression profile to the CMS4 centroid
SSP nearest CMS	CMS class with the highest correlation according to the single sample predictor
SSP predicted CMS	CMS class assigned with the single sample predictor

Numbers (and percentages)					
CMS1	**CMS2**	**CMS3**	**CMS4**	**Non-consensus**	**Total**
0 (0%)	3 (5%)	52 (84%)	0 (0%)	7 (11%)	62

Series	Sample name	minimum correlation to CMS1	minimum correlation to CMS2	minimum correlation to CMS3	minimum correlation to CMS4	median correlation to CMS1	median correlation to CMS2	median correlation to CMS3	median correlation to CMS4	maximum correlation to CMS1	maximum correlation to CMS2	maximum correlation to CMS3	maximum correlation to CMS4	SSP nearest CMS	SSP predicted CMS
1	NGS-001-A	-0.15	0.13	0.35	-0.50	-0.06	0.18	0.43	-0.48	-0.03	0.23	0.48	-0.41	CMS3	**CMS3**
1	NGS-002-A	-0.28	0.26	0.24	-0.40	-0.21	0.30	0.31	-0.38	-0.20	0.32	0.33	-0.33	CMS2	**NA**
1	NGS-004-A	-0.16	0.13	0.33	-0.45	-0.07	0.17	0.41	-0.43	-0.05	0.22	0.44	-0.37	CMS3	**CMS3**
1	NGS-005-A	-0.08	0.07	0.44	-0.54	0.00	0.12	0.53	-0.51	0.02	0.16	0.56	-0.44	CMS3	**CMS3**
1	NGS-006-A	-0.23	0.18	0.26	-0.40	-0.14	0.24	0.35	-0.39	-0.14	0.26	0.38	-0.30	CMS3	**CMS3**
1	NGS-007-A	-0.15	0.22	0.22	-0.53	-0.07	0.32	0.26	-0.51	-0.06	0.34	0.34	-0.46	CMS2	**CMS2**
1	NGS-008-A	-0.20	0.16	0.30	-0.41	-0.16	0.19	0.40	-0.40	-0.13	0.25	0.42	-0.31	CMS3	**CMS3**
1	NGS-009-A	-0.19	0.20	0.37	-0.54	-0.10	0.25	0.46	-0.53	-0.09	0.30	0.50	-0.47	CMS3	**CMS3**
1	NGS-011-A	-0.20	0.16	0.42	-0.53	-0.11	0.21	0.51	-0.51	-0.09	0.26	0.54	-0.45	CMS3	**CMS3**
1	NGS-013-A	-0.13	0.13	0.33	-0.53	-0.04	0.21	0.41	-0.51	-0.03	0.26	0.47	-0.44	CMS3	**CMS3**
1	NGS-014-A	-0.32	0.29	0.18	-0.37	-0.22	0.33	0.23	-0.34	-0.21	0.36	0.24	-0.30	CMS2	**CMS2**
1	NGS-016-A	-0.14	0.16	0.31	-0.54	-0.08	0.23	0.40	-0.51	-0.01	0.28	0.43	-0.43	CMS3	**CMS3**
1	NGS-017-A	-0.22	0.21	0.28	-0.44	-0.14	0.24	0.35	-0.42	-0.13	0.27	0.39	-0.35	CMS3	**CMS3**
1	NGS-019-A	-0.20	0.20	0.35	-0.50	-0.13	0.24	0.42	-0.48	-0.12	0.27	0.46	-0.43	CMS3	**CMS3**
1	NGS-020-A	-0.24	0.21	0.25	-0.42	-0.16	0.27	0.32	-0.39	-0.13	0.28	0.37	-0.35	CMS3	**NA**
1	NGS-021-A	-0.11	0.05	0.47	-0.51	0.00	0.09	0.57	-0.48	0.01	0.13	0.58	-0.41	CMS3	**CMS3**
1	NGS-022-A	-0.14	0.08	0.41	-0.44	-0.08	0.10	0.49	-0.42	-0.05	0.15	0.52	-0.35	CMS3	**CMS3**
1	NGS-025-A	-0.18	0.17	0.35	-0.50	-0.10	0.22	0.44	-0.50	-0.08	0.26	0.48	-0.42	CMS3	**CMS3**
1	NGS-026-A	-0.18	0.12	0.41	-0.49	-0.09	0.16	0.51	-0.48	-0.07	0.22	0.54	-0.39	CMS3	**CMS3**
1	NGS-027-A	-0.14	0.14	0.37	-0.56	-0.05	0.22	0.45	-0.55	-0.04	0.26	0.52	-0.48	CMS3	**CMS3**
1	NGS-029-A	0.03	-0.06	0.50	-0.55	0.13	-0.02	0.60	-0.50	0.13	0.05	0.63	-0.43	CMS3	**CMS3**
1	NGS-030-A	-0.03	-0.03	0.52	-0.52	0.05	0.00	0.61	-0.46	0.07	0.08	0.64	-0.41	CMS3	**CMS3**
1	NGS-044-A	-0.23	0.31	0.08	-0.45	-0.17	0.38	0.13	-0.41	-0.14	0.39	0.20	-0.38	CMS2	**CMS2**
1	NGS-045-A	-0.18	0.22	0.18	-0.43	-0.12	0.28	0.24	-0.41	-0.09	0.30	0.28	-0.35	CMS2	**NA**
1	NGS-046-A	-0.20	0.17	0.31	-0.45	-0.13	0.24	0.38	-0.43	-0.11	0.26	0.42	-0.37	CMS3	**CMS3**
1	NGS-047-A	-0.22	0.17	0.27	-0.40	-0.13	0.22	0.34	-0.38	-0.11	0.24	0.37	-0.29	CMS3	**CMS3**
1	NGS-048-A	-0.14	0.09	0.39	-0.48	-0.06	0.13	0.48	-0.46	-0.03	0.18	0.53	-0.39	CMS3	**CMS3**
1	NGS-049-A	-0.19	0.13	0.30	-0.38	-0.12	0.16	0.39	-0.36	-0.09	0.20	0.40	-0.28	CMS3	**CMS3**
1	NGS-061-A	-0.20	0.18	0.37	-0.51	-0.12	0.23	0.45	-0.50	-0.10	0.26	0.51	-0.43	CMS3	**CMS3**
1	NGS-062-A	0.17	-0.08	0.20	-0.52	0.24	0.02	0.28	-0.45	0.28	0.10	0.39	-0.39	CMS3	**NA**
2	F20A	-0.22	0.23	0.19	-0.46	-0.13	0.30	0.28	-0.42	-0.10	0.33	0.31	-0.39	CMS2	**NA**
2	F23A1	-0.12	0.07	0.46	-0.55	-0.01	0.12	0.58	-0.53	0.02	0.18	0.59	-0.46	CMS3	**CMS3**
2	F22A	-0.12	0.04	0.45	-0.47	-0.02	0.07	0.54	-0.43	-0.01	0.11	0.56	-0.37	CMS3	**CMS3**
2	F26A	-0.16	0.12	0.34	-0.46	-0.06	0.18	0.42	-0.44	-0.05	0.21	0.46	-0.39	CMS3	**CMS3**
2	F37A	-0.22	0.19	0.35	-0.49	-0.13	0.24	0.42	-0.46	-0.12	0.27	0.46	-0.41	CMS3	**CMS3**
2	F38A1	-0.03	-0.01	0.38	-0.46	0.06	0.04	0.47	-0.41	0.08	0.10	0.49	-0.35	CMS3	**CMS3**
2	F38A2	-0.10	0.03	0.40	-0.46	0.01	0.08	0.50	-0.42	0.02	0.11	0.51	-0.36	CMS3	**CMS3**
2	F40A	-0.18	0.12	0.38	-0.50	-0.07	0.17	0.47	-0.47	-0.06	0.21	0.52	-0.40	CMS3	**CMS3**
2	F21A	-0.15	0.14	0.43	-0.61	-0.03	0.20	0.53	-0.59	-0.02	0.26	0.60	-0.53	CMS3	**CMS3**
2	F42A	-0.20	0.15	0.42	-0.51	-0.09	0.20	0.50	-0.48	-0.09	0.25	0.54	-0.43	CMS3	**CMS3**
2	F43A	-0.05	-0.02	0.51	-0.52	0.07	0.00	0.61	-0.47	0.09	0.05	0.63	-0.41	CMS3	**CMS3**

2	F44A	-0.09	0.03	0.52	-0.57	0.03	0.08	0.62	-0.55	0.05	0.14	0.65	-0.47	CMS3	**CMS3**
2	F45A	-0.07	-0.03	0.49	-0.44	0.02	-0.01	0.59	-0.41	0.04	0.05	0.61	-0.34	CMS3	**CMS3**
2	F47A	-0.19	0.13	0.46	-0.54	-0.08	0.18	0.56	-0.51	-0.06	0.23	0.59	-0.47	CMS3	**CMS3**
2	F48A	-0.18	0.08	0.31	-0.31	-0.10	0.09	0.40	-0.30	-0.08	0.14	0.42	-0.23	CMS3	**CMS3**
2	F50A	-0.23	0.20	0.36	-0.51	-0.11	0.23	0.45	-0.48	-0.08	0.27	0.48	-0.44	CMS3	**CMS3**
2	F52A	-0.23	0.22	0.25	-0.49	-0.12	0.29	0.32	-0.45	-0.11	0.31	0.38	-0.40	CMS3	**NA**
2	F56A	-0.20	0.11	0.47	-0.53	-0.06	0.14	0.56	-0.46	-0.03	0.18	0.58	-0.45	CMS3	**CMS3**
2	F58A1	-0.08	0.08	0.45	-0.59	0.04	0.13	0.53	-0.56	0.07	0.18	0.57	-0.49	CMS3	**CMS3**
2	F59A	-0.21	0.17	0.40	-0.54	-0.10	0.22	0.49	-0.51	-0.08	0.26	0.54	-0.46	CMS3	**CMS3**
2	F60A	-0.09	0.05	0.42	-0.53	0.03	0.09	0.54	-0.51	0.04	0.15	0.57	-0.45	CMS3	**CMS3**
2	F61A	-0.06	0.03	0.35	-0.53	0.06	0.10	0.45	-0.49	0.09	0.16	0.50	-0.42	CMS3	**CMS3**
2	F62A	-0.29	0.28	0.29	-0.51	-0.17	0.32	0.38	-0.47	-0.16	0.34	0.42	-0.41	CMS3	**CMS3**
2	F64A	-0.31	0.28	0.24	-0.49	-0.16	0.34	0.32	-0.45	-0.15	0.35	0.36	-0.40	CMS2	**NA**
2	F66A	-0.16	0.10	0.44	-0.56	-0.03	0.15	0.55	-0.52	-0.02	0.20	0.60	-0.47	CMS3	**CMS3**
2	F68A	-0.19	0.15	0.36	-0.48	-0.09	0.20	0.44	-0.46	-0.08	0.24	0.48	-0.42	CMS3	**CMS3**
2	F25A	-0.08	0.12	0.34	-0.63	0.04	0.19	0.44	-0.60	0.06	0.27	0.53	-0.54	CMS3	**CMS3**
2	F17A	-0.22	0.19	0.39	-0.58	-0.11	0.27	0.47	-0.52	-0.09	0.32	0.53	-0.46	CMS3	**CMS3**
2	F65A	-0.05	0.03	0.37	-0.52	0.06	0.09	0.46	-0.47	0.07	0.15	0.50	-0.42	CMS3	**CMS3**
2	F7A	-0.17	0.14	0.33	-0.50	-0.06	0.21	0.40	-0.46	-0.05	0.24	0.46	-0.41	CMS3	**CMS3**
2	F63A	-0.14	0.01	0.26	-0.26	-0.04	0.02	0.33	-0.19	0.00	0.08	0.35	-0.12	CMS3	**CMS3**
2	F23A2	-0.15	0.11	0.37	-0.50	-0.05	0.16	0.46	-0.48	-0.01	0.19	0.50	-0.40	CMS3	**CMS3**

Supplementary Table 10. Comparison of the CMS classification of the colorectal adenomas by the study approach (random forest CMS classifier) and single sample predictor. The SSP method confirmed 48 out of 54 (89%) CMS labels assigned in this study. Five out of seven adenomas that were not classified ('non-consensus') by the SSP method were assigned a CMS label by the random forest CMS classifier: one adenoma was an MSI adenoma classified as CMS1, and four adenomas were classified as CMS2. From these four CMS2 adenomas, three adenomas were assigned to CMS2 as the nearest class by the SSP method, however, did not reach the correlation thresholds to be definitely classified (Supplementary Table 6).

The random forest CMS classifier

The single sample predictor CMS classifier	**CMS1**	**CMS2**	**CMS3**	**CMS4**	Non-consensus	**Total**
CMS1	**0**	0	0	0	0	0
CMS2	0	**3**	0	0	0	3
CMS3	0	1	**45**	0	6	52
CMS4	0	0	0	**0**	0	0
Non-consensus	**1**	**4**	0	0	2	7
Total	1	8	45	0	8	62

Chapter 4

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED SPLICE VARIANTS BY THE PROTEOGENOMIC PIPELINE SPLICIFY

Malgorzata A Komor, Thang V Pham, Annemieke C Hiemstra,
Sander R Piersma, Anne S Bolijn, Tim Schelfhorst, Pien M Delis-van Diemen,
Marianne Tijssen, Robert P Sebra, Meredith Ashby, Gerrit A Meijer,
Connie R Jimenez, Remond JA Fijneman

MOLECULAR & CELLULAR PROTEOMICS 2017;16:1850-1863

Identification of Differentially Expressed Splice Variants by the Proteogenomic Pipeline Splicify*[S]

Malgorzata A. Komor‡§ Thang V. Pham§, Annemieke C. Hiemstra‡, Sander R. Piersma§, Anne S. Bolijn‡, Tim Schelfhorst§, Pien M. Delis-van Diemen‡, Marianne Tijssen‡, Robert P. Sebra¶, Meredith Ashby‖, Gerrit A. Meijer‡, Connie R. Jimenez§, and Remond J. A. Fijneman‡**

Proteogenomics, *i.e.* comprehensive integration of genomics and proteomics data, is a powerful approach identifying novel protein biomarkers. This is especially the case for proteins that differ structurally between disease and control conditions. As tumor development is associated with aberrant splicing, we focus on this rich source of cancer specific biomarkers. To this end, we developed a proteogenomic pipeline, Splicify, which can detect differentially expressed protein isoforms. Splicify is based on integrating RNA massive parallel sequencing data and tandem mass spectrometry proteomics data to identify protein isoforms resulting from differential splicing between two conditions. Proof of concept was obtained by applying Splicify to RNA sequencing and mass spectrometry data obtained from colorectal cancer cell line SW480, before and after siRNA-mediated downmodulation of the splicing factors SF3B1 and SRSF1. These analyses revealed 2172 and 149 differentially expressed isoforms, respectively, with peptide confirmation upon knock-down of SF3B1 and SRSF1 compared with their controls. Splice variants identified included RAC1, OSBPL3, MKI67, and SYK. One additional sample was analyzed by PacBio Iso-Seq full-length transcript sequencing after SF3B1 downmodulation. This analysis verified the alternative splicing identified by Splicify and in addition identified novel splicing events that were not represented in the human reference genome annotation. Therefore, Splicify offers a validated proteogenomic data analysis pipeline for identification of disease specific protein biomarkers resulting from mRNA alternative splicing. Splicify is publicly available on GitHub (https://github.com/NKI-TGO/SPLICIFY) and suitable to address basic research questions using pre-clinical model systems as well as translational research questions using patient-derived samples, *e.g.* allowing to identify clinically relevant biomarkers. ***Molecular & Cellular Proteomics 16: 10.1074/mcp.TIR117.000056, 1850–1863, 2017.***

Approximately 95% of multiexon transcripts undergo alternative splicing, making the human transcriptome far more complex than the protein-coding genome (1). Because of alternative splicing, a single gene can be transcribed into a variety of isoforms which, when translated into proteins, will differ in structure, location, and function. Abnormally spliced RNA can cause or contribute to disease. Aberrant splicing is associated with tumor progression and metastasis, and has been shown to affect each of the biological processes commonly referred to as the hallmarks of cancer (2). Therefore, studying aberrant splicing may reveal additional insights into tumor biology and phenotype. For instance, usage of an alternative 5′ splice site of BCL2L1 causes a switch from a pro- to an antiapoptotic isoform in cancer and contributes to resisting cell death (3). Usage of an alternative 3′ splice site of VEGFA leads to a shift from an anti- to a proangiogenic isoform in cancer and induces angiogenesis (4). As aberrant splicing accompanies tumor progression, splice variants provide a promising source of clinically relevant biomarkers.

Splicing factors play a direct role in splicing regulation and isoform expression. Splicing factors can develop oncogenic activity, *e.g.* because of aberrant expression or somatic mutations, and through aberrant splicing lead to carcinogenesis (2). SF3B1 is a splicing factor required for the early spliceosome assembly and is also one of the most commonly mutated splicing factors in cancer (5). Recurrent mutations affecting this gene were found in leukemia, melanoma and in pancreatic, breast, and bladder cancer. Even though the spe-

---

From the ‡Translational Gastrointestinal Oncology, Department of Pathology, Netherlands Cancer Institute, Amsterdam, the Netherlands; §Oncoproteomics Laboratory, Department of Medical Oncology, VU University Medical Center, Amsterdam, the Netherlands; ¶School of Medicine at Mount Sinai, Institute for Genomics and Multiscale Biology, New York, New York; ‖Pacific Biosciences, Menlo Park, California

Received May 11, 2017

Published, MCP Papers in Press, July 26, 2017, DOI 10.1074/mcp.TIR117.000056

Author contributions: M.A.K., G.A.M., C.R.J., and R.J.F. designed research; M.A.K., A.C.H., S.R.P., A.S.B., T.S., P.M.D.-v.D., M.T., and R.P.S. performed research; M.A.K., T.P., R.P.S., and M.A. contributed new reagents/analytic tools; M.A.K., G.A.M., C.R.J., and R.J.F. analyzed data; M.A.K., A.C.H., S.R.P., A.S.B., R.P.S., G.A.M., C.R.J., and R.J.F. wrote the paper.

cific effects of these alterations on splicing are still to be explored, their features often suggest proto-oncogenic activity (6). In chronic lymphocytic leukemia, mutations in this splicing factor contribute to tumor progression, poor patient survival, and poor chemotherapy response (7, 8). Overexpression of another splicing factor, SRSF1, was observed in different tumor types including breast (9), colon, thyroid, small intestine, kidney, lung, liver, and pancreas (10) and was proven to lead to oncogenic activity (2, 11–13). Transcription of SRSF1 is directly regulated by MYC, a well-known oncogenic transcription factor. Through activation of SRSF1, MYC can affect alternative splicing of a subset of SRSF1 target genes and contribute to tumor development (14). For instance, in breast cancer upregulation of SRSF1 promotes transformation of mammary cells through abnormal splicing of BCL2L11 and BIN1 (15). In colorectal cancer (CRC),[1] SRSF1 causes inclusion of exon 4 in RAC1, generating a Rac1b isoform that contributes to cell survival (16, 17).

RNA-seq allows studying the complexity of transcriptomes. Although there is a lot of evidence for alternative splicing on the RNA level, for many of the isoforms it is still not known whether they are translated into proteins. This knowledge is crucial to understanding the biological consequences of alternative splicing, and toward identifying protein biomarkers that result from the translation of splice variants. Protein isoforms have significant potential as biomarkers to increase the accuracy of diagnosis, prognosis, or therapy prediction of the disease (18). Identification of disease-specific protein isoforms enables the discovery of biomarkers with better sensitivity and/or specificity.

Protein isoforms can be studied on the proteome level with the use of in-depth tandem mass spectrometry. Proteogenomics is a dynamic field integrating genomic and proteomic data (19). One of the focus areas in the field is to increase the knowledge of the human proteome and identify novel variant proteins resulting from single nucleotide variants or aberrant splicing (20, 21). The number of bioinformatics tools for performing proteogenomic analysis is rapidly increasing, including tools for proteogenomic database construction (22–27) or visualization of the peptides on the genome (28, 29). However, a number of these tools lack an automated, user-friendly downstream analysis after MS/MS identification to extract interesting outcomes. Moreover, the tools are often designed for single sample or single cohort analysis without the flexibility to perform a differential comparison between case and control groups on both RNA and protein level. To identify disease specific biomarkers resulting from aberrant splicing there is a need for a tool that will perform a differential group analysis.

Here we present a method to identify tumor-specific protein isoforms based on RNA-seq and mass spectrometry (LC-MS/MS) data. In this approach, RNA-seq analysis is used to perform quantitative isoform analysis and identify differential splice variants, and LC-MS/MS confirms translation of these variants into proteins. The method was applied to the CRC cell line SW480 upon downmodulation of the splicing machinery factors SF3B1 and SRSF1. In this way, a controlled setting was created that allowed to monitor changes in alternative splicing and consequently, to design a pipeline for proteogenomic analysis of spliced isoforms. The methodological novelty of this approach lies in differential analysis of alternative splicing between two groups in two molecular domains and could be applied in any comparative setting such as gene knock-down *versus* control or cancer *versus* healthy control.

EXPERIMENTAL PROCEDURES

Cell Culture, Gene Knock-down and Cell Viability Assay—SW480 cells cultured in Dulbecco's modified Eagle's medium (DMEM; Invitrogen, Bleiswijk, The Netherlands) containing 10% fetal bovine serum (FBS; Perbio Science, Etten-Leur, The Netherlands) were maintained in a humidified 5% CO_2 atmosphere at 37 °C. Twenty-four hours after seeding, cells were transfected in duplo with small interfering RNA (siRNA) pools against SF3B1 (siGENOME SF3B1 SMARTpool, M-020061-02; Thermo Fisher Scientific, Waltham, MA) and SRSF1 (siGENOME SRSF1 SMARTpool, M-018672-01), according to manufacturer's recommendations. A final siRNA concentration of 30 nM was obtained using DharmaFECT3 reagent (1:1000 dilution; T-2003-02, Thermo Fisher Scientific). A nontargeting siRNA pool (siGENOME Non-Targeting pool #2, D-001206-14) was used as negative control. Cell viability was determined after transfection using the MTT (3-(4,5-dimethylthiazolyl-2)-2,5-diphenyltetrazolium bromide; ICN Biomedicals, Solon, OH) assay, as described previously (30).

RNA Isolation and Quantitative Reverse Transcription PCR—Total RNA was isolated from viable cells, 48 h after siRNA transfection with siSF3B1 and the siNon-Targeting (siNT) control, and 72 h after transfection with siSRSF1 and its siNT control using Trizol reagent (15596; Invitrogen, Breda, The Netherlands) and the miRNeasy Mini Kit (217004; Qiagen, Venlo, the Netherlands), following the manufacturer's protocol. Concentrations and purities were measured on a Nanodrop ND-1000 spectrophotometer (Isogen, Ijsselstein, The Netherlands). cDNA was synthesized using the Iscript cDNA synthesis kit (170-8891; Bio-Rad Laboratories, Hercules, CA). Quantitative reverse transcription PCR (RT-qPCR) was performed using SYBR Green (4309155, Thermo Fisher Scientific), to monitor SF3B1 and SRSF1 knock-down efficiencies and to evaluate efficiency of alternative splicing for ADD3, CTNND1, RAC1, SYK, MKI67, and OSBPL3. Beta-2-Microglobulin (B2M) was used as a housekeeping reference gene. In brief, gene expression was measured using 2 μl of 10 ng/μl cDNA in a 25 μl SYBR Green reaction (see supplemental Table S1 for primers and conditions), as described previously (30).

cDNA Library Preparation and Illumina RNA Sequencing—cDNA libraries were prepared with the TruSeq Stranded mRNA LT sample Prep kit (RS-122-2101, Illumina, San Diego, CA) according to the TruSeq Stranded mRNA sample preparation guide (Part# 15031047, Revision E, October 2013). cDNA library quality control was per-

---

[1] The abbreviations used are: CRC, colorectal cancer; CCS, circular consensus sequencing; A3SS, alternative 3′ splice site; A5SS, alternative 5′ splice site; MXE, mutually exclusive exons; RI, retained intron; RNA-seq, RNA sequencing; RT-qPCR, quantitative reverse transcription PCR; SE, skipped exon; siNT, siNon-Targeting; siSF3B1, siRNA mediated downmodulation of SF3B1; siSRSF1, siRNA mediated downmodulation of SRSF1; SMART, Switching Mechanism at 5′ End of RNA Template; SMRT, single molecule real time.

formed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Sample libraries were diluted and pooled to obtain a final concentration of 10 nM. Sequencing was performed on an Illumina HiSeq V4 2500, using a 125 bases paired end run with an input of 16 pM cDNA. Quality assessment of RNA-seq data was performed with FastQC version 0.11.4 (31) with default settings and visualized with MultiQC version 0.9 (32) with default parameters.

Protein Isolation and Separation—Proteins were isolated at the same time points as RNA extraction. After thorough washing with PBS, cells were lysed in reducing sample buffer (NuPAGE LDS sample buffer, NP0008, Thermo Fisher Scientific; 65% Milli-Q, 25% 4*LDS, 10% 1 M DTT) to obtain an approximate protein concentration of 1 μg/μl. Cells were scraped and transferred to eppendorf tubes. After heating for 5–10 min at 99 °C and centrifugation for 1 min at 14,000 rpm aliquots of the samples were stored at −80 °C until further use. Approximately 35 μg protein from the supernatant was loaded on a NuPAGE Novex 4–12% Bis-Tris Protein Gel, 1.5 mm, 10-well (NP0335BOX; Thermo Fisher Scientific). Proteins were resolved at 150V for 1 h in 200 ml NuPAGE MES SDS Running buffer (NP0002; Thermo Fisher Scientific) supplemented with 0.5 ml NuPAGE antioxidant (NP0005; Thermo Fisher Scientific). The gel was placed in a container with fixing solution (50% ethanol, 46.5% Milli-Q and 3.5% phosphoric acid) for 15 min and stained with colloïdal Coomassie (48.4% Milli-Q, 34% methanol, 15% ammonium sulfate, 2.5% phosphoric acid, 0.1% Coomassie Brilliant Blue G-250 (20279; Thermo Fisher Scientific)) overnight and destained with multiple changes of Milli-Q water. Each gel lane was sliced in 10 slices.

Whole Gel In-gel Digestion—The in-gel digestion procedure was done as described previously (33) with the following changes: gel pieces were dried in a centrifugal evaporator (SpeedVac) for ~30 min and peptides were extracted with 100 μl 1% formic acid and two times 150 μl 5% formic acid/50% acetonitrile. Concentrated extracts were transferred to Millipore filters (Millex-HV Syringe driven filter unit, 0.45 μm, SLHVR04NL, Millipore), placed on autosampler vials and centrifuged for 5 min at room temperature in the centrifugal evaporator without vacuum.

LC-MS/MS—Peptides were separated by an Ultimate 3000 nanoLC-MS/MS system (Dionex LC-Packings, Amsterdam, The Netherlands) equipped with a 40 cm × 75 μm ID fused silica column custom packed with 1.9 μm 120 Å ReproSil Pur C18 aqua (Dr Maisch GMBH, Ammerbuch-Entringen, Germany). After injection, peptides were trapped at 10 μl/min on a 10 mm × 100 μm ID trap column packed with 5 μm 120 Å ReproSil Pur C18 aqua at 2% buffer B (buffer A: 0.5% acetic acid in MQ; buffer B: 80% ACN + 0.5% acetic acid in MQ) and separated at 300 nl/min in a 10–40% buffer B gradient in 60 min (90 min inject-to-inject). The nanoLC column was maintained at 50 °C using a column heater (Phoenix S&T, Chester, PA). Eluting peptides were ionized at a potential of +2 kVa into a Q Exactive mass spectrometer (Thermo Fisher Scientific). Intact masses were measured at resolution 70.000 (at *m/z* 200) in the orbitrap using an AGC target value of 3 × 106 charges. The top 10 peptide signals (charge-states 2+ and higher) were submitted to MS/MS in the HCD (higher-energy collision) cell (1.6 *m/z* isolation width, 25% normalized collision energy). MS/MS spectra were acquired at resolution 17.500 (at *m/z* 200) in the orbitrap using an AGC target value of 1 × 106 charges and an underfill ratio of 0.5%. Dynamic exclusion was applied with a repeat count of 1 and an exclusion time of 30 s.

Full Length Isoform Sequencing - Iso-Seq—RNA isolated from siSF3B1- and siNT-treated SW480 cells was subjected to full-length RNA single molecule real time (SMRT) sequencing called Iso-Seq (34). Briefly, RNA (RIN score of ~9.0 assessed by Agilent Bioanalysis) was amplified using the Clontech Switching Mechanism at 5′ end of RNA Template (SMART) technology which incorporates known sequence at both ends of the cDNA product in the first strand synthesis process without the need for conventional adapter ligation strategies. Four hundred eight nanograms of siSF3B1 and 352 ng siNT cDNA were used as input to the SMART cDNA amplification process to capture full-length, intact isoforms to be reverse transcribed and amplified into full-length cDNA representing the full transcriptome where the known sequences are used to complete SMRTbell library preparation using the cDNA products.

Once ample double stranded cDNA was synthesized, cDNA Iso-Seq sequencing libraries were prepared using the SMRTbell library preparation procedure resulting in a library containing molecular inserts that represent a single isoform per library molecule. These libraries were then size-selected to enrich for isoforms of interest by targeting a population of full-length transcripts to enhance coverage by loading individual size fractions on single SMRTcells. More specifically, the SageELF electrophoretic lateral fractionator instrument was used to separate independent fractions of library where isoforms that are 0–1 kbp, 1 kbp–2 kbp, 2 kbp–3 kbp, and 3 kbp–50 kbp were split into independent SMRTbell libraries for sequencing so that larger isoforms were not detrimentally dominated by smaller isoform library molecules during the sequencing process.

Finally, samples were sequenced using 6-hr movie collection on the PacBio RSII sequencer with two SMRTcells per cDNA size fraction. The RSII data yielded 523k to 750k subreads for each size fraction of the siNT sample, resulting in 66.8k to 98.3k CCS reads with up to 43k full length cDNA reads per size fraction. For siSF3B1, the RSII yield was 321k to 981k subreads for each size fraction, resulting in 47.5k to 97.3k CCS reads with up to 51.7k full-length cDNA reads per size fraction, using default Iso-Seq pipeline settings. Raw sequencing data was processed using Iso-Seq on PacBio SMRTportal (smrtanalysis v2.3.0) and ICE software (35) to predict low and high-quality isoforms and generate high resolution transcriptome references.

RNA-seq and LC-MS/MS Data Analysis Within the Proteogenomic Pipeline Splicify—The schematic overview of the proteogenomic pipeline, Splicify, is presented in Fig. 1*A*. Low quality reads and adapter sequences were trimmed by Trimmomatic (36) version 3 to average quality score for a 4-base wide sliding window of 20, both at the beginning and at the end of the sequences (ILLUMINACLIP: TruSeq3-PE.fa:2:30:10, LEADING:20, TRAILING:20, AVGQUAL:20, SLIDINGWINDOW:4:20). Because of the requirements of the further analysis (rMATS (37)) reads were processed to match length of 120 bp, shorter reads were discarded and longer reads were trimmed (CROP:120, MINLEN:120). Mapping was performed with the use of STAR aligner (38) version 2.4.2a to the human genome (USCS RefSeq hg19 annotation, as STAR option genomeDir) with the following parameters; outSAMtype: BAM SortedByCoordinate, readFilesCommand: zcat, runThreadN: 28, outSAMattributes: All. Differential splice variants were identified with rMATS version 3.2.5 using UCSC RefSeq hg19 GTF file as annotation in the unpaired analysis type (parameters; len: 120, t: paired, analysis: U). Significant events were extracted (FDR≤0.05). Both inclusion- and exclusion-isoforms of spliced genomic fragments were taken into account for further analysis. Nucleotide acid sequences of splicing regions (upstream and downstream exons with and without spliced fragment) were obtained and translated in forward frame to amino acid sequences. In this way, a database was obtained with protein sequences of potential splice variants that were all added to the human reference proteome database (Uniprot, release January 2014, no fragments, canonical and isoform, 42104 entries (39)) forming an enriched human protein database. Peptide identification was performed by MaxQuant 1.5.3.8 (40) with the use of the enriched human protein database. Enzyme specificity was set to trypsin and up to two missed cleavages were allowed. Cysteine carboxamidomethylation was treated as fixed modification and methionine oxidation and N-terminal acetylation as variable modifications. Peptide precursor ions were searched with a maximum mass deviation of 4.5 ppm and fragment ions with a maximum mass deviation of 20 ppm. Peptide and protein identifications were filtered at an FDR of 1% using the decoy database strategy. Common contaminants were included in the MS/MS search. Evidence and peptides files were taken along for further analysis. Peptides specific for splice variants were extracted. Additionally, human database of canonical proteins (Swissprot, canonical, 20197 entries) was used to detect which of the splice variants represented non-canonical isoforms. Peptide intensities were normalized to the average of the samples' medians and log 10 transformed. Imputation was performed on the normalized and transformed matrix, where missing values were imputed from the normal distribution of mean equal to minimal intensity observed and standard deviation equal to mean of standard deviations calculated for each peptide. Differential peptide expression analysis was performed with a Bioconductor package limma (41) and log10 fold changes and p values were obtained. Splicify is available at (https://github.com/NKI-TGO/SPLICIFY).

Isoform Identification with the Use of Full Length Transcripts—Redundant transcripts were removed by first aligning them to the human genome (hg19) with GMAP (42) and collapsing highly similar transcripts predicted across FASTA files from various size fractions with the software cupcake ToFU (v1.3). In these steps both BAM and GTF files were produced for each sample. Samples were chained, to standardize transcript IDs and merge the transcripts from both experiments. Details of the workflow can be found here (43). The merged file was used as input to rMATS instead of human reference annotation GTF file. In this way, the program can use the exon-exon and exon-intron junctions introduced by Iso-seq. Splice variants identified by rMATS were annotated by changing Iso-Seq transcript IDs into gene names based on genomic location, with the use of biomaRt Bioconductor package version 2.26.1(44). In case the Iso-Seq transcript was on the opposite strand than the gene, "otherstrand" was added to the gene symbol. In case there was no gene matching the coordinates of the transcript, "intergenic" was used as a gene symbol. The annotated output of rMATS was further processed as described in the RNA-seq and LC-MS/MS Data Analysis Within the Proteogenomic Pipeline section, with the exclusion of the quantification step.

RESULTS

Experimental Model System To Test the Proteogenomic Pipeline Design—The schematic overview of Splicify, the proteogenomic data analysis pipeline for identification of differential splice variants, is presented in Fig. 1*A*. To test the design of the proteogenomic pipeline, a model system needed to be established in which modulation of isoform changes could be controlled experimentally. For this purpose, the splicing factors SF3B1 and SRSF1, which play a key role in the splicing machinery, were downmodulated in the CRC cell line SW480, followed by RNA-seq-based transcriptomics and mass spectrometry-based proteomics analyses. A general overview of the experimental design is presented in Fig. 2.

The efficiency of siRNA-mediated downmodulation of SF3B1 and SRSF1 in SW480 CRC cells was determined by RT-qPCR, and reached on average up to a 50 and 40% reduction of mRNA expression for SF3B1 and SRSF1, respectively (supplemental Fig. S1). Cell viability was reduced by 10–30% by downmodulation of SF3B1 at 48 h after transfection, whereas no changes in cell viability were observed after the knockdown of SRSF1 at 72 h after transfection (data not shown). To assure that downmodulation of SF3B1 and SRSF1 resulted functionally in changes in expression of certain isoforms, monitoring of positive controls was included in the experiment. Skipped exons in ADD3 and CTNND1 were identified by literature search as positive controls for alternative splicing in colorectal cancer tissue compared with normal colon tissue (35). Indeed, RT-qPCR analysis for ADD3 exon 14 and CTNND1 exon 20 indicated that exclusion of these exons served as functional splicing controls for knock-down of SF3B1 and SRSF1, respectively (supplemental Fig. S2). These data demonstrate that a model system was established in which isoform switches can be modulated in a CRC cell line, suited to test the design of the proteogenomic pipeline.

Identification of Differentially Expressed RNA and Protein Isoforms by Applying the Proteogenomic Pipeline—To investigate alternative splicing in both the RNA and protein molecular domains, the transcriptome and the proteome of each sample were analyzed with RNA-seq and tandem mass spectrometry. Quality assessment of RNA-seq and LC-MS/MS data is available in supplemental Figs. S3–S5. Within the RNA-seq data analysis, isoforms were identified with the use of reads spanning exon-exon and exon-intron junctions. These splice-variant specific reads, together with reads mapping to the spliced fragment, were further quantified to distinguish differential events between two conditions. In the proteomics data analysis, exon-exon and exon-intron junction-spanning peptides and peptides mapping on the spliced fragment were used to confirm translation of the isoforms detected on the RNA level into proteins (Fig. 1*B*). The intensities of these peptides were used for quantification to identify differentially expressed protein isoforms. For details, see Fig. 1*A*.

Differential mRNA Isoforms Induced by Downmodulation of SF3B1 and SRSF1—Transcriptome analysis revealed a number of significantly differentially spliced events for siSF3B1 and siSRSF1 in comparison to their controls (Fig. 3*A*; see supplemental Tables S3–S12 for details of all the events), proving that manipulation of the splicing machinery resulted in differential splicing. Alternative splicing was more affected upon manipulation of SF3B1 compared with SRSF1, as the number of alternatively spliced events was larger for this splicing factor, for the events like skipped exon and mutually exclusive exons (Fig. 3*A*). This might be because of the different roles that these splicing factors play in the spliceosome complex. The significantly skipped exon events included the positive controls of alternative splicing, higher exclusion levels of ADD3 exon 14 upon downmodulation of SF3B1 and higher exclusion levels of CTNND1 exon 20 upon downmodulation of SRSF1 (supplemental Fig. S6). These data show that the intermediate mRNA results of the proteogenomic pipeline reproduced the expected outcome, and yielded information about hundreds (for SRSF1) to thousands (for SF3B1) of additional alternative splicing events.

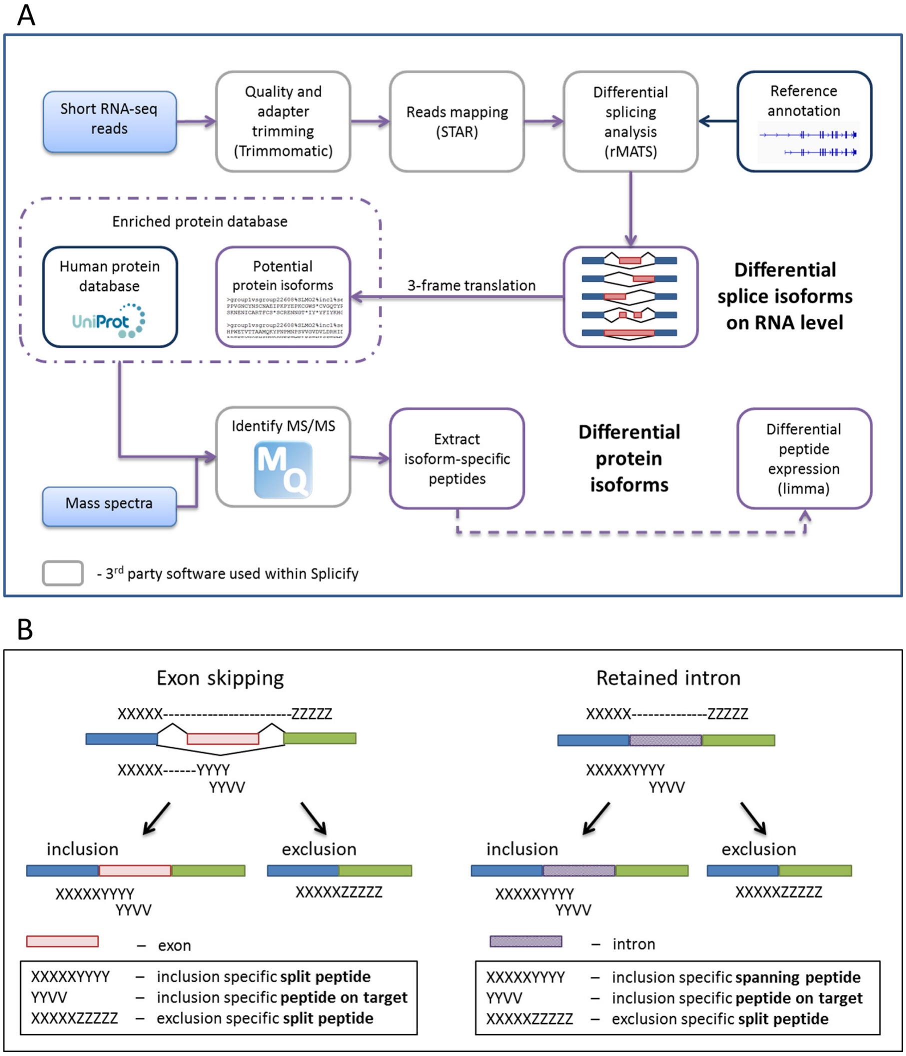

A
Short RNA-seq reads
Quality and adapter trimming (Trimmomatic)
Reads mapping (STAR)
Differential splicing analysis (rMATS)
Reference annotation
Enriched protein database
Human protein database
UniProt
Potential protein isoforms
3-frame translation
Differential splice isoforms on RNA level
Identify MS/MS
MQ
Extract isoform-specific peptides
Differential protein isoforms
Differential peptide expression (limma)
Mass spectra
- 3rd party software used within Splicify
B
Exon skipping
XXXXX-------------------------ZZZZZ
XXXXX------YYYY
YYVV
inclusion
exclusion
XXXXXYYYY
YYVV
XXXXXZZZZZ
– exon
XXXXXYYYY – inclusion specific split peptide
YYVV – inclusion specific peptide on target
XXXXXZZZZZ – exclusion specific split peptide
Retained intron
XXXXX--------------ZZZZZ
XXXXXYYYY
YYVV
inclusion
exclusion
XXXXXYYYY
YYVV
XXXXXZZZZZ
– intron
XXXXXYYYY – inclusion specific spanning peptide
YYVV – inclusion specific peptide on target
XXXXXZZZZZ – exclusion specific split peptide

FIG. 1. **Splicify, the proteogenomic pipeline for identification of differential splice variants.** *A*, The schematic overview of the Splicify data analysis. Within Splicify RNA-seq data analysis is performed by combining exemplar open-source RNA-seq analysis software, including quality and adapter trimming with Trimmomatic (36), reads mapping with STAR (38), differential splicing analysis with rMATS (37), where differential splice variants on RNA level are identified. These splice variants undergo 3-frame translation into potential protein isoform sequence database (FASTA). This database together with the human protein database from Uniprot (39) can be further used with MaxQuant (40), a search engine to identify MS/MS spectra originating from the same samples as RNA-seq reads. Downstream analysis of MaxQuant output is performed with the use of the results from RNA-seq analysis. Isoform-specific peptides are extracted and quantified and based on these peptides differential protein isoforms are identified. Splicify produces a final table with both RNA and protein isoform information. *B*, Example of peptides supporting translation of splicing events for skipped exon and retained intron. Split peptides map to both sides of an exon-exon junction, spanning peptides span exon-intron junctions (specific for inclusion variants for retained intron, alternative 3′ and 5′ splice sites) and peptides on target map to a spliced fragment.

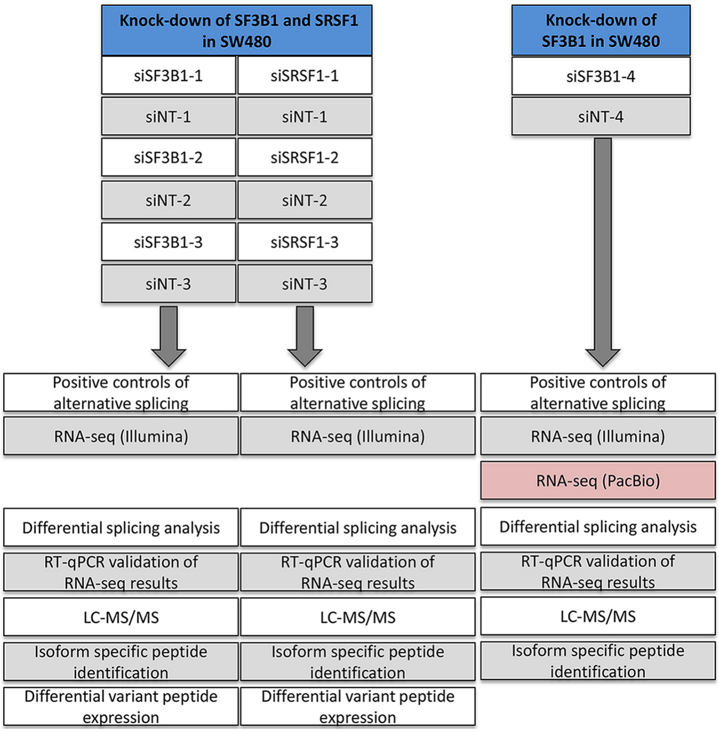

Knock-down of SF3B1 and SRSF1 in SW480
Knock-down of SF3B1 in SW480
siSF3B1-1
siSRSF1-1
siSF3B1-4
siNT-1
siNT-1
siNT-4
siSF3B1-2
siSRSF1-2
siNT-2
siNT-2
siSF3B1-3
siSRSF1-3
siNT-3
siNT-3
Positive controls of alternative splicing
Positive controls of alternative splicing
Positive controls of alternative splicing
RNA-seq (Illumina)
RNA-seq (Illumina)
RNA-seq (Illumina)
RNA-seq (PacBio)
Differential splicing analysis
Differential splicing analysis
Differential splicing analysis
RT-qPCR validation of RNA-seq results
RT-qPCR validation of RNA-seq results
RT-qPCR validation of RNA-seq results
LC-MS/MS
LC-MS/MS
LC-MS/MS
Isoform specific peptide identification
Isoform specific peptide identification
Isoform specific peptide identification
Differential variant peptide expression
Differential variant peptide expression

FIG. 2. **General overview of the experimental design and data analysis.** Downmodulation of splicing factors SF3B1 (48 h) and SRSF1 (72 h) was performed in CRC cell line SW480 three times. RT-qPCR of known splicing events obtained from literature (skipped exon in ADD3 and in CTNND1) were used as positive controls of alternative splicing to functionally verify that downmodulation of the splicing machinery caused differential splicing. The knock-downs and the paired non-targeting (NT) controls were subjected to RNA-sequencing and LC-MS/MS tandem mass spectrometry, followed by data analysis using the proteogenomic pipeline Splicify (see Fig. 1). Differential mRNA splice variants were identified and several candidates were validated with RT-qPCR. Isoform specific peptides were identified and differential expression of these peptides was performed. Downmodulation of SF3B1 was repeated in a separate experiment, including PacBio Iso-Seq sequencing of full length transcripts while excluding isoform-specific peptide quantitative analysis because of the lack of replicates.

To further validate our approach, four skipped exon splicing events were selected for confirmation by RT-qPCR, comprising SYK exon 7, RAC1 exon 4, OSBPL3 exon 9, and MKI67 exon 7 (Fig. 4, supplemental Table S2). These isoforms are also known as SYK(S) and SYK(L), Rac1b and MKI67 long and short isoforms. According to the RNA-seq analysis, all the events were differentially spliced upon downmodulation of SRSF1 whereas OSBPL3 and MKI67 were affected by downmodulation of SF3B1. The differences in the expression of inclusion and exclusion variants between downmodulation and controls were validated with RT-qPCR (supplemental Fig. S7–S9).

Differential Protein Isoforms Induced by Downmodulation of SF3B1 and SRSF1—All significant events identified on RNA level, comprising both exclusion and inclusion variants, were taken along for database construction for mass spectra identification. To prove that these splicing events are translated into proteins we searched for the peptides specific for the splice isoforms (Fig. 1*B*). Over 5070 and 370 isoform-specific peptides were identified for differential isoforms upon downmodulation of SF3B1 and SRSF1, respectively (Table I, see supplemental Fig. S10 for quality control of isoform-specific peptides). The differences in these numbers correspond to the sizes of the splice variant databases of the two experiments. Overall around 60% of the isoform-specific peptides turned out to map on target, peptides spanning exon-exon junction comprised around 40% and exon-intron junctions were identified far less frequently (Table II).

Based on all the isoform-specific peptides, 2172 and 149 isoforms on protein level were identified for siSF3B1 and siSRSF1, respectively (Table III). On average for ~15% of the

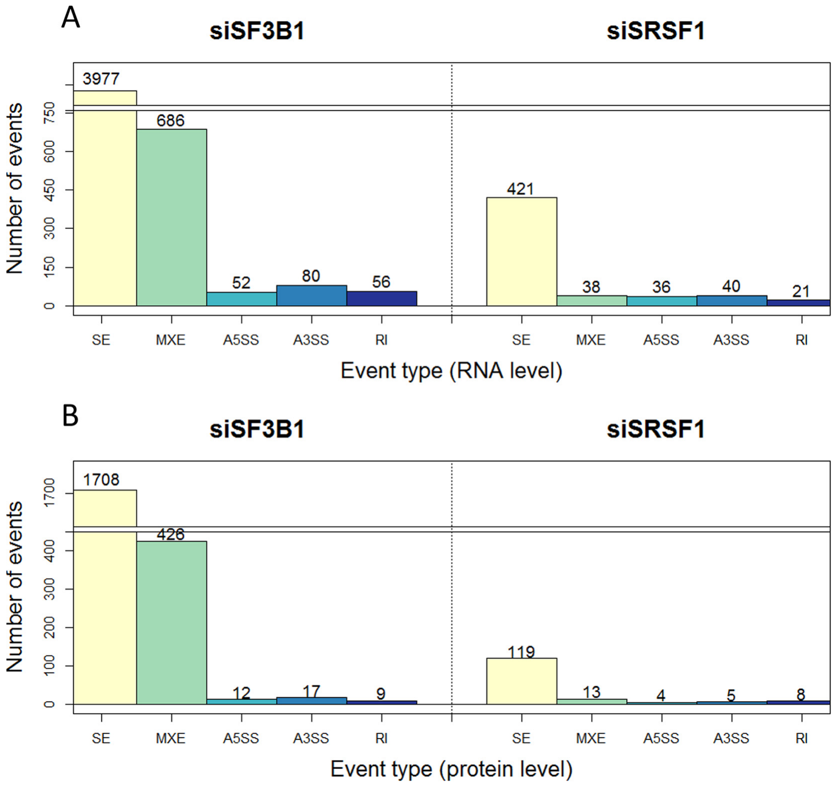


FIG. 3. **The number of splicing events identified on RNA and protein level upon knock-down of SF3B1 and SRSF1.** *A*, Number of significant alternatively spliced events on RNA level upon downmodulation of SF3B1 and SRSF1 *versus* their controls. *B*, The number of alternative splicing events for which at least one variant (inclusion/exclusion) was confirmed by identification of isoform-specific peptides. S.E. - skipped exon; MXE - mutually exclusive exons; A5SS - alternative 5′ splice site; A3SS - alternative 3′ splice site; RI - retained intron.

splicing events peptide confirmation was observed for both inclusion and exclusion variants of the same event. Most of these isoforms are considered canonical proteins based on the Swissprot canonical sequence database. Approximately 5 and 25% of the identified isoforms were classified as noncanonical for siSF3B1 and siSRSF1, respectively. A subset of peptides mapped to two or more isoforms, usually because of the overlapping exons between the different isoforms. More confirmation for inclusion variants was obtained than for exclusion variants, because of the longer sequences of the inclusion variants. Among the identified isoforms all categories of alternatively spliced events were represented, with most peptides supporting the skipped exon splicing category because of the predominance of this class already at the RNA level. Relatively, looking at the ratios of number of splicing events on RNA and protein level, mutually exclusive exons are more frequently detected (Fig. 3B). This is mainly because mutually exclusive exons do not have an exclusion variant as both isoforms include an additional exon, thereby increasing the overall fragment length and consequently the probability of peptide identification within the spliced region. Even though for the splicing controls ADD3 and CTNND1 no variant-specific peptides were detected, other events such as alternatively skipped exon in SYK, RAC1, OSBPL3, and MKI67 were confirmed on peptide level (supplemental Tables S13–S14).

Differential peptide expression analysis was performed for all of the splice-specific peptides and revealed that a subset of these peptides did significantly differ between splice factor knock-downs and controls, indicating concordant events between mRNA genomic and proteomic results (Table IV, supplemental Tables S13–S14). For both experiments around 65% of the significantly differentially expressed isoform-specific peptides showed concordant expression differences as observed on the RNA level. For instance, upon downmodulation of SF3B1 three split peptides spanning inclusion of

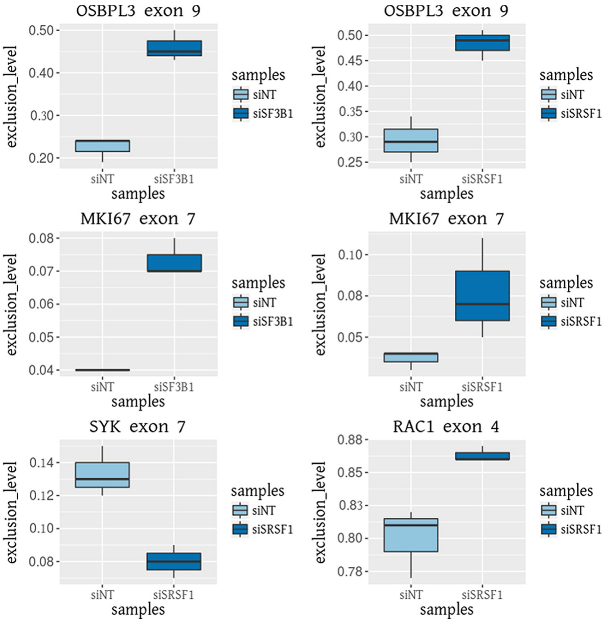


FIG. 4. **RT-qPCR validation of differential splicing events identified by RNA-seq data analysis with the proteogenomic pipeline, Splicify.** The exclusion isoforms of OSBPL3 exon 9 and MKI67 exon 7 are higher expressed upon downmodulation of SF3B1 and SRSF1. The inclusion isoform of SYK exon 7 and the exclusion isoform of RAC1 exon 4 are higher expressed upon downmodulation of SRSF1. Exclusion levels were calculated by dividing exclusion spanning reads by the sum of inclusion and exclusion spanning reads.

TABLE I

Overview of isoform-specific peptides identified upon knock-down of SF3B1 and SRSF1. The numbers of peptides specific for inclusion and exclusion isoforms are listed. Some peptides map to multiple isoforms, being inclusion-specific for one isoform and exclusion-specific for the other

Experiment	Isoform-specific peptides	Inclusion-specific peptides	Exclusion-specific peptides
siSF3B1 vs siNT	5079	4525	833
siSRSF1 vs siNT	374	309	87

TABLE II

Overview of categories of isoform-specific peptides identified upon knock-down of SF3B1 and SRSF1. Peptides on target map fully on the spliced fragment, spanning peptides span exon-intron junctions and split peptides span exon-exon junctions (see also Fig. 1B)

Experiment	On target	Spanning peptide	Split peptide
siSF3B1 vs siNT	3278	9	1794
siSRSF1 vs siNT	217	3	154

exon 9 in OSBPL3 and one split peptide supporting the exclusion of this exon were identified. Two of the inclusion specific peptides show significantly lower expression upon downmodulation of SF3B1 whereas the exclusion specific peptide indicates higher expression in comparison to the control (Fig. 5; supplemental Table S13). Another example is lower expression of the Rac1b isoform, resulting from the inclusion of exon 4 in RAC1 gene, upon downmodulation of SRSF1, which is in line with the current knowledge of the SRSF1 effect on alternative splicing of RAC1 in colorectal cancer (16). This result was detected in the proteogenomic pipeline at RNA level, both by RNA-seq and by RT-qPCR (Fig. 4; supplemental Fig. S9*B*). On protein level only inclusion specific peptides were identified. Even though the differences in peptide intensities between siSRSF1 downmodulation and the control were not significant, log10 fold changes suggest a similar effect as on RNA level (supplemental Fig. S11; supplemental Table S14).

Full-length Transcripts Validation—To examine if sequencing of full length transcripts can validate the isoforms identified within Splicify and enrich these results with novel transcripts, Iso-Seq was performed in SW480 cells upon downmodulation of SF3B1 and its siNT control (see Fig. 2). As Iso-Seq provides qualitative information, transcripts detected by this technique were used as the source of transcriptome variation instead of the human reference annotation, which could be further quantified upon mapping back the shorter, but higher density Illumina reads. On RNA level, within each alternative splicing category, the number of significantly dif-

Table III

Overview of mRNA splicing events confirmed by proteomics upon knock-down of SF3B1 and SRSF1. RNA isoforms were considered to be translated if there was at least one splice-specific peptide identified. For a subset of alternatively spliced events both inclusion and exclusion variants were confirmed by identification of splice-specific peptides. Based on the database of canonical proteins a small number of non-canonical proteins was identified

Experiment	Alternatively spliced events	Inclusion isoforms	Exclusion isoforms	Events with both isoforms	Non-canonical isoforms
siSF3B1 vs siNT	2172	2006	400	234	93
siSRSF1 vs siNT	149	128	47	26	36

Table IV

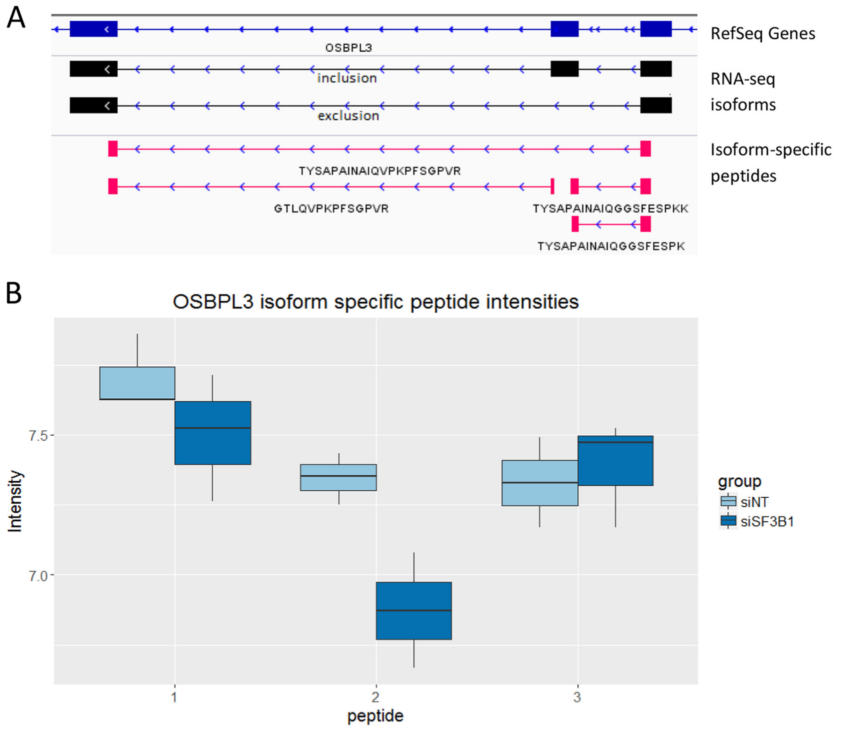
The number of isoform specific-peptides showing consistent or opposite expression changes as detected on RNA level. Isoform-specific peptides were filtered based on p value ≤0.1 and absolute value of log10 transformed fold change ≥0.5, to extract only the peptides differentially expressed between the two conditions; siRNA mediated down-modulation of a splicing factor and the non-targeting control. For inclusion-specific peptides, a peptide was labelled as "consistent" if the log10 fold change of the peptide expression showed the same direction of change as the Inclusion Level Difference for the RNA splice variant. For exclusion-specific peptides, a peptide was labelled as "consistent" if the direction of change was the opposite of the RNA-derived Inclusion Level Difference. As a subset of peptides maps to multiple isoforms, the percentages might exceed 100%

Experiment	Number (and percentage) of the isoform-specific peptides	
	Consistent	Opposite
siSF3B1 vs siNT	267 (65%)	157 (38%)
siSRSF1 vs siNT	16 (64%)	9 (36%)

ferential isoforms identified with the use of Iso-Seq data exceeded the results compared with the approach making use of the reference annotation (Fig. 6*A*; see supplemental Tables S15–S24 for details). There was a large overlap between detection of alternatively spliced events by Illumina-sequencing using the human reference annotation and the analysis that used Illumina reads with the Iso-Seq full length transcripts, thereby validating detection of alternatively spliced events by Splicify (Fig. 6*B*). Additionally, full length isoform sequencing revealed several novel events that were not detected with the standard Splicify approach before because of absence of these events in the reference genome annotation. The largest effect is noticed for detection of retained intron events, where rMATS uses a database of annotated retained introns instead of all the introns in the genome. On the protein level, the majority of the isoform-specific peptides were identified with both approaches (Fig. 6*C*). However, the protein database composed of the Iso-seq based findings increased the number of identified isoform-specific peptides compared with the use of the human reference annotation (supplemental Tables S25–S26). For example, three peptides supporting intron retention in FXR1 were identified by sequencing of full-length transcripts that included this intron, which therefore was included in the annotation file. Illumina short reads supported this event and provided quantitative proof that it is higher expressed upon downmodulation of SF3B1 compared with its control (Fig. 6*D*). These data indicate that to unravel differential splicing events more comprehensively, one should provide annotation files enriched with novel transcripts from *e.g.* transcriptome assembly tools or full-length transcript sequencing.

DISCUSSION

Splicify was designed to identify differentially expressed splice variants on RNA and protein level. Splicify was applied on CRC cell line SW480 upon downmodulation of splicing machinery and nontargeting controls. We showed that this method can successfully identify condition-specific aberrant splicing events on protein level, by performing comparative splice variant analysis on both RNA and protein level. A subset of the RNA-seq based results of Splicify was validated by RT-qPCR. This proved that the pipeline yielded real splice variants on RNA level. Additionally, applying Splicify using PacBio Iso-Seq full-length transcript sequencing confirmed the existence of the identified isoforms and increased the transcriptomic space to detect novel events. These were especially prevalent in the retained intron and alternative 3′ and 5′ splice site splicing events, where the overlap between Splicify with reference annotation and Splicify with Iso-Seq full length transcripts was smaller than for skipped exon and mutually exclusive exons splicing events. This shows that the reference annotation is still lacking several alternatively spliced isoforms which include whole or a part of the intronic sequence. A number of the novel events were also detected on protein level. This indicates that Splicify, next to the standard approach with the use of the human reference annotation, can also be applied using an alternative transcriptome annotation file that extends isoform identification with novel splicing events. On protein level, we identified several noncanonical isoforms, which is a valuable finding as it indicates their translation into proteins that may play a different functional role in comparison to their canonical counterparts. This is known for the Rac1b isoform, which has been shown to have a different functional role than the canonical RAC1 protein enhancing cell survival (45). Splicing of RAC1 is known to be dependent on SRSF1 activity, which was confirmed with the Splicify pipeline applied to the SW480 CRC cell line upon downmodulation of SRSF1. These data indicate that the results on protein level are in line with current literature. Other



Peptide number	Sequence	Incl_Excl	logFC	p-value	adj.p-value
1	GTLQVPKPFSGPVR	incl	0.202	0.214	0.864
2	TYSAPAINAIQGGSFESPK[K]	incl	0.473	0.027	0.864
3	TYSAPAINAIQVPKPFSGPVR	excl	-0.058	0.699	0.970

FIG. 5. **Splicing isoforms of OSBPL3 presented in two molecular domains.** *A*, Screenshot from IGV of the spliced region of OSBPL3 exon 9; in blue -RefSeq genes, in black - inclusion and exclusion variants identified with RNA-seq, in pink - inclusion and exclusion specific peptides identified in mass spectrometry. *B*, Peptide intensities upon downmodulation of SF3B1 and its control for two inclusion specific peptides and one exclusion specific peptide for exon 9 in OSBPL3. Peptide number on the *x* axis corresponds to the peptide sequence in the table. Intensities of the overlapping peptides TYSAPAINAIQGGCFESPK and TYSAPAINAIQGGCFESPKK were manually summed and annotated as TYSAPAINAIQGGCFESPK[K] in the table and as peptide number 2 on the figure. Differential peptide expression analysis was performed with limma with no imputation for all the isoform-specific peptides including the merged peptide TYSAPAINAIQGGCFESPK[K] instead of the two. Even though not all of the peptides are significantly up or downregulated, the signal is concordant with RT-qPCR and RNA-seq results, with higher exclusion and lower inclusion of the exon upon downmodulation of SF3B1.

interesting findings include the detection of differential splice variants of OSBPL3. These isoforms have been shown to be differentially expressed on RNA level in various tissues, indicating that the OSBPL3 splice variants might have different functionality (46). Translation of these splice-variants into proteins and differential protein isoform expression was now shown by Splicify. These results demonstrate that Splicify successfully identifies differentially expressed mRNA and protein isoforms.

Our findings include identification of several other biologically interesting isoforms that might be linked to SF3B1 and SRSF1 activity. For instance, SYK splice variants, SYK(S) and SYK(L), have been shown to play a role in breast, liver and colorectal cancers (47). Alternative splicing of SYK has been demonstrated to regulate colorectal cancer progression and sensitivity of CRC cells to chemotherapy (48). Here, identification of differential splicing of SYK upon downregulation of SRSF1 might indicate possible impact of SRSF1 on alterna-

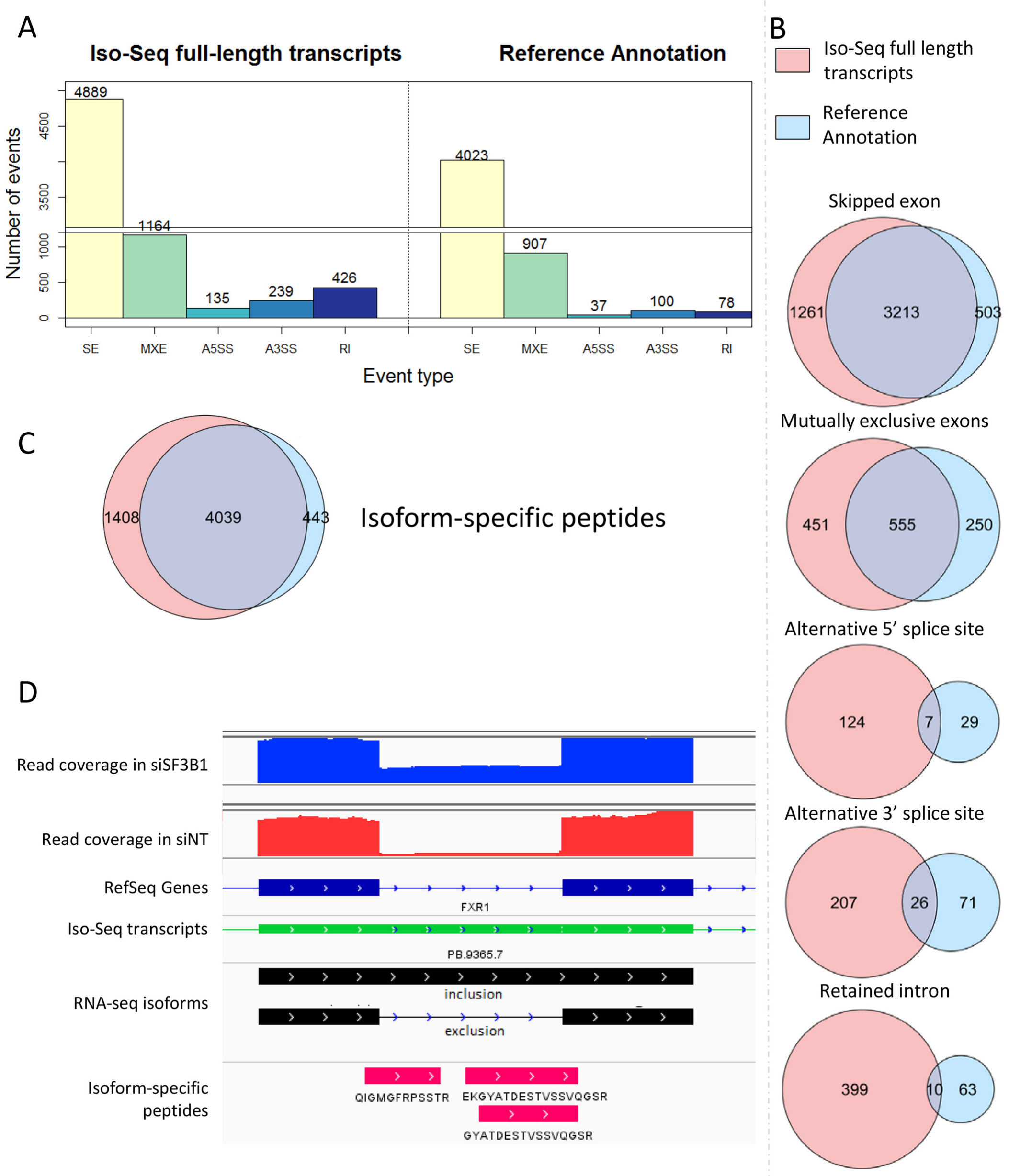


FIG. 6. **Comparison of the standard Splicify approach with the reference annotation and Splicify analysis with Iso-seq full length transcripts used as annotation.** *A*, Number of significant alternatively spliced events on RNA level for downmodulation of SF3B1 *versus* the non-targeting controls with the use of Iso-Seq full-length transcripts or Reference Annotation, S.E. - skipped exon, MXE - mutually exclusive exons, A5SS - alternative 5′ splice site, A3SS - alternative 3′ splice site, RI - retained intron. Illumina reads were quantified on alternatively spliced events originating from reference annotation or Iso-Seq full-length transcripts. *B*, Overlap analysis between alternatively spliced events

tive splicing of SYK and subsequently on colorectal cancer progression and chemotherapy resistance. Another interesting finding is identification of differential expression of MKI67 long and short isoforms upon modulation of SF3B1 as well as SRSF1 expression. It is speculated that MKI67 long isoform plays a role in cell differentiation by causing the cell to exit the cell cycle, whereas the short isoform leads to permanent cell cycle (49). Based on our findings, one could hypothesize that SF3B1 and SRSF1 might regulate cell proliferation through alternative splicing of MKI67. However, further studies are needed to support these statements. In addition to these examples, Splicify provided many other differentially spliced isoforms. Studies that aim to investigate gene function or biomarker utility could focus on splice events with peptide evidence, as these events confirm RNA translation that implies functional consequences. Moreover, different filtering approaches can be applied, *e.g.* based on fold change in RNA and protein expression, false discovery rates or the number of split peptides required.

The small number of protein isoforms that were detected compared with the results obtained based on analyses of RNA-seq data demonstrated the current struggles in the field of proteogenomics. There might be various reasons why many mRNA splice variants were not identified on protein level, including biological and technical ones. First, not all of the aberrant isoforms are translated into proteins. For instance, if there is a stop codon on the fragment that is alternatively spliced in, it will lead to degradation of the shorter transcript via nonsense-mediated decay. There are also splicing events called detained introns that may not exit the nucleus and therefore do not undergo translation (50). Another reason might be the kinetics of transcription and translation, concerning the siRNA mediated downmodulation. It is possible that although transcripts are already present on the RNA level, they might not be translated into proteins yet at the time of RNA and protein isolation. Also, low protein isoform count can be a result of post-translational modifications of the spliced regions, for instance phosphorylation, which requires alternative sample processing preceding mass spectrometry to obtain high resolution of phosphopeptide identifications. There are also technical issues that limit the identification of splice-specific peptides, especially for the exclusion variants. If one would exclude the peptides with missed cleavages, there can only be one split peptide spanning an exclusion variant. This peptide needs to have a suitable distribution of lysine and arginine so that it spans the junction, while also having the required length and physicochemical features to be identified by a mass spectrometer. Inclusion isoforms are identified more frequently because of their longer sequence and therefore higher probability to contain a suitable tryptic peptide within the fragment of interest.

All these issues explain the current advantage of RNA-seq over mass spectrometry in terms of performing quantitative analyses of splice fragments. The aberrant isoforms are often lower expressed than canonical proteins, which further complicates differential isoform expression analysis on protein level (5, 51). The 65% consistency of splice variant expression differences on RNA and protein level was expected in the context of multiple studies reporting modest correlation between RNA and protein expression (21, 52, 53). However, the qualitative information provided by mass spectrometry is highly valuable and crucial to determine what isoforms are translated into proteins. Detection of protein isoforms gives more confidence in the functional relevance of splice variants identified on RNA level, and enables to prioritize candidate biomarkers for further studies when identified in both molecular domains. In terms of biomarkers studies, Splicify can be applied in a clinically relevant setting, *e.g.* to compare a large series of cancer samples to healthy control tissues, and reveal differentially expressed isoforms. As the proteogenomic approach within Splicify is an unbiased first discovery step, these candidate biomarkers should be further quantified by *e.g.* multiple reaction monitoring or data independent acquisition, preferably both in human tissues and in relevant human body fluids for which a biomarker test is being developed (54–57). Ultimately, a highly robust approach of detecting these isoforms is necessary that could be applied in a clinical setting. For instance, antibodies targeting the spliced region could be incorporated into an immunoassay for testing large cohorts of human samples (56, 57). In conclusion, the output of proteogenomic analysis within Splicify provides answers to basic research and translational research questions, allowing identifying biologically and clinically relevant isoform-specific biomarkers.

DATA AVAILABILITY

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (58) partner repository with the dataset identifier PXD006486.

upon downmodulation of SF3B1 and its control, identified with reference annotation or Iso-Seq used as annotation. Overlap was defined by chromosome number and coordinates of the spliced fragment. In case of skipped exon, retained intron and alternatively spliced sites it was one fragment, in case of mutually exclusive exons, coordinates of both exons were taken into the overlap. *C*, Overlap analysis of splice-specific peptides identified with the databases based on the approach including reference annotation or Iso-Seq data. Differential splicing events were translated in 3-frame into potential proteins. These databases were used for mass spectra identification with MaxQuant. Splice-specific peptides were extracted from the MaxQuant output. Overlap analysis was performed based on unique peptide sequences. *D*, IGV screenshot of retained intron in FXR1 gene. Blue and red coverage plots represent Illumina reads for samples siSF3B1-4 and siNT-4, respectively. Below in dark blue - reference annotation, in green - Iso-Seq transcripts obtained from the same samples, in black - retained intron event identified with Iso-Seq and quantified by Illumina reads, in pink - 3 peptides spanning the exon-intron junction and supporting intron retention on protein level.

Acknowledgments—We thank Bo Han, Sarah Kingan, and Elizabeth Tseng from PacBio for the support with data analysis and David Cozijnsen for his contribution. The transcriptomics data was obtained at Leiden Genome Technology Center, VUmc Clinical Genetics, and Genome Core Facility at the NKI. We also thank HPC facility at the NKI for making the computational analysis possible.

* This work was supported by KWF Kankerbestrijding, project number 2013-6025 and PacBio AACR SMRT Grant Award. This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

** To whom correspondence should be addressed: The Netherlands Cancer Institute, Department of Pathology, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. Tel.: (+31) 20 5121732; E-mail: R.Fijneman@nki.nl.

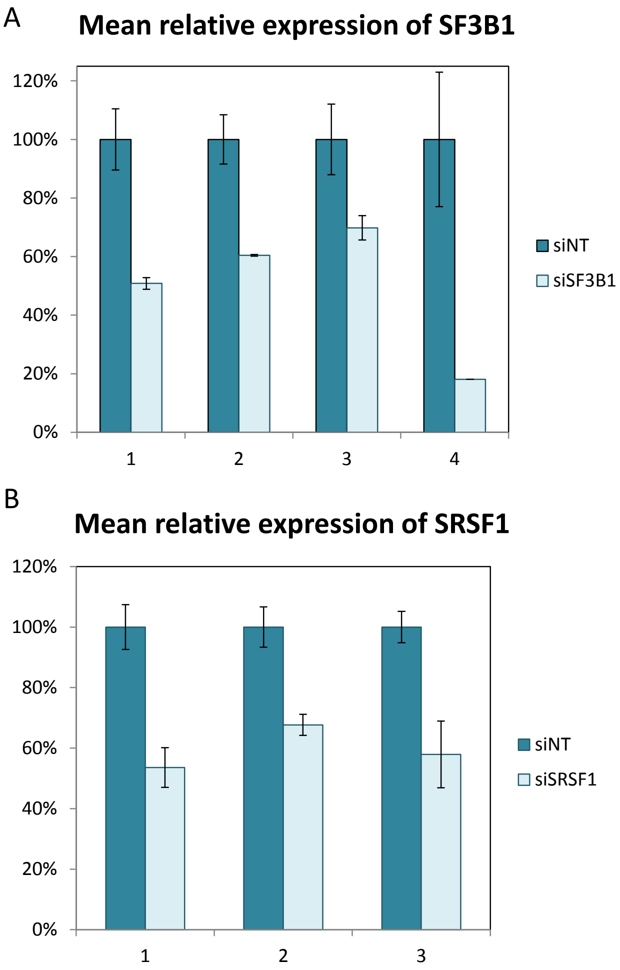
S This article contains supplemental material.

REFERENCES

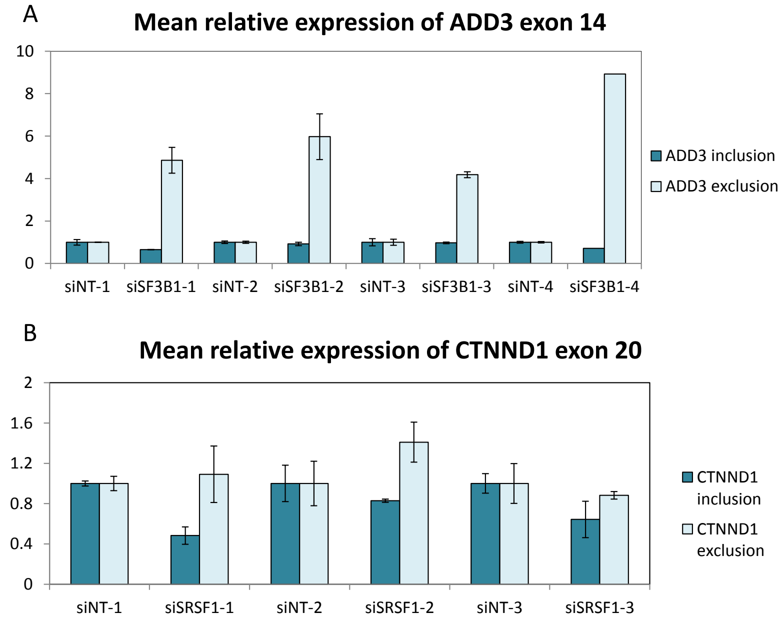
1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Gen.* **40,** 1413–1415
2. Oltean, S., and Bates, D. O. (2014) Hallmarks of alternative splicing in cancer. *Oncogene* **33,** 5311–5318
3. Boise, L. H., Gonzalez-Garcia, M., Postema, C. E., Ding, L., Lindsten, T., Turka, L. A., Mao, X., Nunez, G., and Thompson, C. B. (1993) bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* **74,** 597–608
4. Ladomery, M. R., Harper, S. J., and Bates, D. O. (2007) Alternative splicing in angiogenesis: the vascular endothelial growth factor paradigm. *Cancer Lett.* **249,** 133–142
5. Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., and Skotheim, R. I. (2016) Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* **35,** 2413–2427
6. Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016) RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16,** 413–430
7. Quesada, V., Conde, L., Villamor, N., Ordonez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Bea, S., Pinyol, M., Martinez-Trillos, A., Lopez-Guerra, M., Colomer, D., Navarro, A., Baumann, T., Aymerich, M., Rozman, M., Delgado, J., Gine, E., Hernandez, J. M., Gonzalez-Diaz, M., Puente, D. A., Velasco, G., Freije, J. M., Tubio, J. M., Royo, R., Gelpi, J. L., Orozco, M., Pisano, D. G., Zamora, J., Vazquez, M., Valencia, A., Himmelbauer, H., Bayes, M., Heath, S., Gut, M., Gut, I., Estivill, X., Lopez-Guillermo, A., Puente, X. S., Campo, E., and Lopez-Otin, C. (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Gen.* **44,** 47–52
8. Oscier, D. G., Rose-Zerilli, M. J., Winkelmann, N., Gonzalez de Castro, D., Gomez, B., Forster, J., Parker, H., Parker, A., Gardiner, A., Collins, A., Else, M., Cross, N. C., Catovsky, D., and Strefford, J. C. (2013) The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. *Blood* **121,** 468–475
9. Anczuków, O., Akerman, M., Cléry, A., Wu, J., Shen, C., Shirole, N. H., Raimer, A., Sun, S., Jensen, M. A., Hua, Y., Allain, F. H. T., and Krainer, A. R. (2015) SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol. Cell* **60,** 105–117
10. Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., and Krainer, A. R. (2007) The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat. Structural Mol. Biol.* **14,** 185–193
11. David, C. J., and Manley, J. L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Develop.* **24,** 2343–2364
12. Ladomery, M. (2013) Aberrant alternative splicing is another hallmark of cancer. *Int. J. Cell Biol.* **2013,** 463786
13. Moore, M. J., Wang, Q., Kennedy, C. J., and Silver, P. A. (2010) An Alternative Splicing Network Links Cell Cycle Control to Apoptosis. *Cell* **142,** 625–636
14. Das, S., Anczukow, O., Akerman, M., and Krainer, A. R. (2012) Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. *Cell Reports* **1,** 110–117
15. Anczukow, O., Rosenberg, A. Z., Akerman, M., Das, S., Zhan, L., Karni, R., Muthuswamy, S. K., and Krainer, A. R. (2012) The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nat. Structural Mol. Biol.* **19,** 220–228
16. Goncalves, V., Henriques, A. F., Pereira, J. F., Neves Costa, A., Moyer, M. P., Moita, L. F., Gama-Carvalho, M., Matos, P., and Jordan, P. (2014) Phosphorylation of SRSF1 by SRPK1 regulates alternative splicing of tumor-related Rac1b in colorectal cells. *RNA* **20,** 474–482
17. Matos, P., and Jordan, P. (2008) Increased Rac1b expression sustains colorectal tumor cell survival. *Mol. Cancer Res.* **6,** 1178–1184
18. Mischak, H., Apweiler, R., Banks, R. E., Conaway, M., Coon, J., Dominiczak, A., Ehrich, J. H., Fliser, D., Girolami, M., Hermjakob, H., Hochstrasser, D., Jankowski, J., Julian, B. A., Kolch, W., Massy, Z. A., Neusuess, C., Novak, J., Peter, K., Rossing, K., Schanstra, J., Semmes, O. J., Theodorescu, D., Thongboonkerd, V., Weissinger, E. M., Van Eyk, J. E., and Yamamoto, T. (2007) Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics. Clin. Appl.* **1,** 148–156
19. Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., Fenyo, D., Zhang, B., and Mani, D. R. (2017) Methods, tools and current perspectives in proteogenomics. *Mol. Cell. Proteomics* **16,** 959–981
20. Liu, S., Im, H., Bairoch, A., Cristofanilli, M., Chen, R., Deutsch, E. W., Dalton, S., Fenyo, D., Fanayan, S., Gates, C., Gaudet, P., Hincapie, M., Hanash, S., Kim, H., Jeong, S. K., Lundberg, E., Mias, G., Menon, R., Mu, Z., Nice, E., Paik, Y. K., Uhlen, M., Wells, L., Wu, S. L., Yan, F., Zhang, F., Zhang, Y., Snyder, M., Omenn, G. S., Beavis, R. C., and Hancock, W. S. (2013) A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J. Proteome Res.* **12,** 45–57
21. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J., and Liebler, D. C. (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387
22. Wang, X., and Zhang, B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29,** 3235–3237
23. Li, Y., Wang, X., Cho, J. H., Shaw, T., Wu, Z., Bai, B., Wang, H., Zhou, S., Beach, T. G., Wu, G., Zhang, J., and Peng, J. (2016) JUMPg: an Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. *J. Proteome Res.* **15,** 2309–2320
24. Wen, B., Xu, S., Sheynkman, G. M., Feng, Q., Lin, L., Wang, Q., Xu, X., Wang, J., and Liu, S. (2014) sapFinder: an R/Bioconductor package for detection of variant peptides in shotgun proteomics experiments. *Bioinformatics* **30,** 3136–3138
25. Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M. D., Clauser, K. R., Tabb, D. L., Mertins, P., Slebos, R., Erdmann-Gilmore, P., Li, S., Gunawardena, H. P., Xie, L., Liu, T., Zhou, J. Y., Sun, S., Hoadley, K. A., Perou, C. M., Chen, X., Davies, S. R., Maher, C. A., Kinsinger, C. R., Rodland, K. D., Zhang, H., Zhang, Z., Ding, L., Townsend, R. R., Rodriguez, H., Chan, D., Smith, R. D., Liebler, D. C., Carr, S. A., Payne, S., Ellis, M. J., and Fenyo, D. (2016) An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell. Proteomics* **15,** 1060–1071
26. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., and Bafna, V. (2014) Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **13,** 21–28
27. Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., and Jones, A. R. (2014) ProteoAnnotator–open source proteogenomics annotation software supporting PSI standards. *Proteomics* **14,** 2731–2741
28. Wang, X., Slebos, R. J., Chambers, M. C., Tabb, D. L., Liebler, D. C., and Zhang, B. (2016) proBAMsuite, a bioinformatics framework for genome-based representation and analysis of proteomics data. *Mol. Cell. Proteomics* **15,** 1164–1175
29. Askenazi, M., Ruggles, K. V., and Fenyo, D. (2016) PGx: putting peptides to BED. *J. Proteome Res.* **15,** 795–799
30. Sillars-Hardebol, A. H., Carvalho, B., Tijssen, M., Belien, J. A., de Wit, M., Delis-van Diemen, P. M., Ponten, F., van de Wiel, M. A., Fijneman, R. J., and Meijer, G. A. (2012) TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* **61,** 1568–1575

31. Andrews, S. (2015) FastQC a quality control tool for high throughput sequence data.
32. Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–3048
33. Piersma, S. R., Warmoes, M. O., de Wit, M., de Reus, I., Knol, J. C., and Jimenez, C. R. (2013) Whole gel processing procedure for GeLC-MS/MS based proteomics. *Proteome Sci.* **11,** 17
34. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138
35. Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I. V., Figueroa, M., Schilling, J. S., Chen, F., and Wang, Z. (2015) Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PloS One* **10,** e0132628
36. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120
37. Shen, S., Park, J. W., and Lu, Z. X. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E5593–E5601
38. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21
39. The Uniprot Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169
40. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
41. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47
42. Wu, T. D., and Watanabe, C. K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875
43. Magdoll (03/14/2017) https://github.com/Magdoll/cDNA_Cupcake/wiki/Cupcake-ToFU%3A-supporting-scripts-for-Iso-Seq-after-clustering-step. *GitHub*
44. Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protocols* **4,** 1184–1191
45. Singh, A., Karnoub, A. E., Palmby, T. R., Lengyel, E., Sondek, J., and Der, C. J. (2004) Rac1b, a tumor associated, constitutively active Rac1 splice variant, promotes cellular transformation. *Oncogene* **23,** 9369–9380
46. Collier, F. M., Gregorio-King, C. C., Apostolopoulos, J., Walder, K., and Kirkland, M. A. (2003) ORP3 splice variants and their expression in human tissues and hematopoietic cells. *DNA Biol.* **22,** 1–9
47. Krisenko, M. O., and Geahlen, R. L. (2015) Calling in SYK: SYK's dual role as a tumor promoter and tumor suppressor in cancer. *Biochim. Biophys. Acta* **1853,** 254–263
48. Ni, B., Hu, J., Chen, D., Li, L., Chen, D., Wang, J., and Wang, L. (2016) Alternative splicing of spleen tyrosine kinase differentially regulates colorectal cancer progression. *Oncol. Lett.* **12,** 1737–1744
49. Schmidt, M. H., Broll, R., Bruch, H. P., Finniss, S., Bogler, O., and Duchrow, M. (2004) Proliferation marker pKi-67 occurs in different isoforms with various cellular effects. *J. Cell. Biochem.* **91,** 1280–1292
50. Boutz, P. L., Bhutkar, A., and Sharp, P. A. (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Develop.* **29,** 63–80
51. Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Gen. Biol.* **14,** R70
52. Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., and Nilsson, P. (2009) Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10,** 365
53. Kosti, I., Jain, N., Aran, D., Butte, A. J., and Sirota, M. (2016) Cross-tissue analysis of gene and protein expression in normal and cancer tissues. *Scientific Reports* **6,** 24799
54. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**
55. Anderson, L., and Hunter, C. L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* **5,** 573–588
56. Carr, S. A., and Anderson, L. (2008) Protein Quantitation Through Targeted Mass Spectrometry: the Way Out of Biomarker Purgatory? *Clin. Chem.* **54,** 1749–1752
57. Rifai, N., Gillette, M. A., and Carr, S. A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotech.* **24,** 971–983
58. Vizcaino, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D456

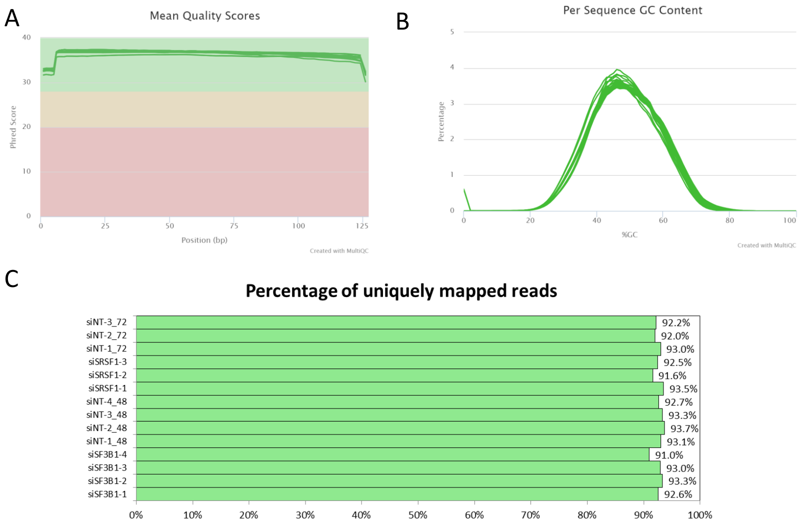
SUPPLEMENTARY FIGURES



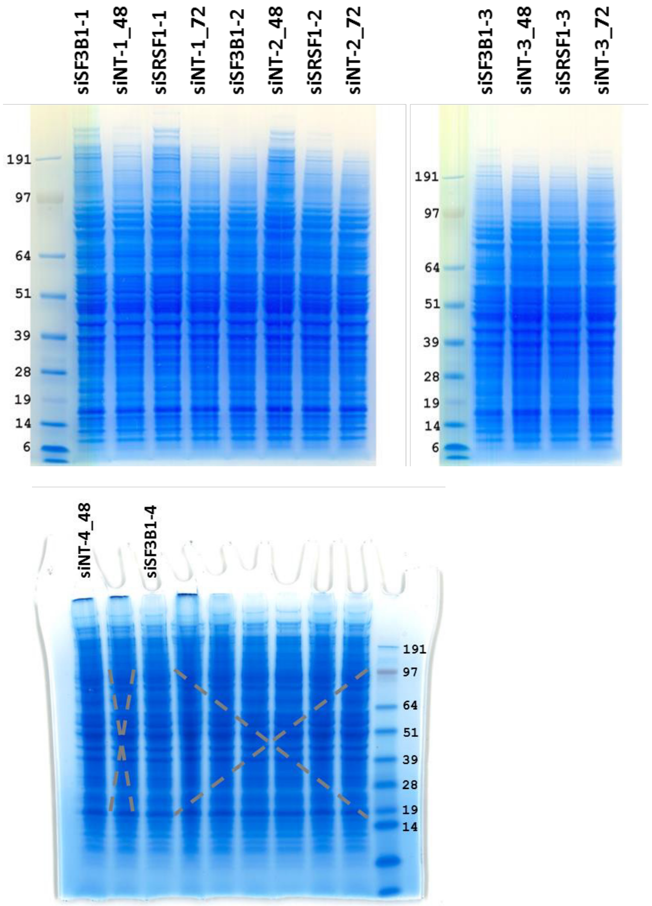
Supplementary Figure 1. Knock-down efficiency of SF3B1 and SRSF1 A Mean relative expression of SF3B1 upon siRNA-mediated down-modulation (siSF3B1) compared to the non-targeting control (siNT). RNA was harvested in 48 hours after transfection. The experiment was performed three times in technical duplicates. Then experiment was repeated the forth time in technical replicate for the control and as a single experiment for the siSF3B1 B Mean relative expression quantified by RT-qPCR of SRSF1 upon siRNA-mediated down-modulation (siSRSF1) compared to the control (siNT). RNA was harvested in 72 hours after transfection. The experiment was performed three times in technical duplicates.



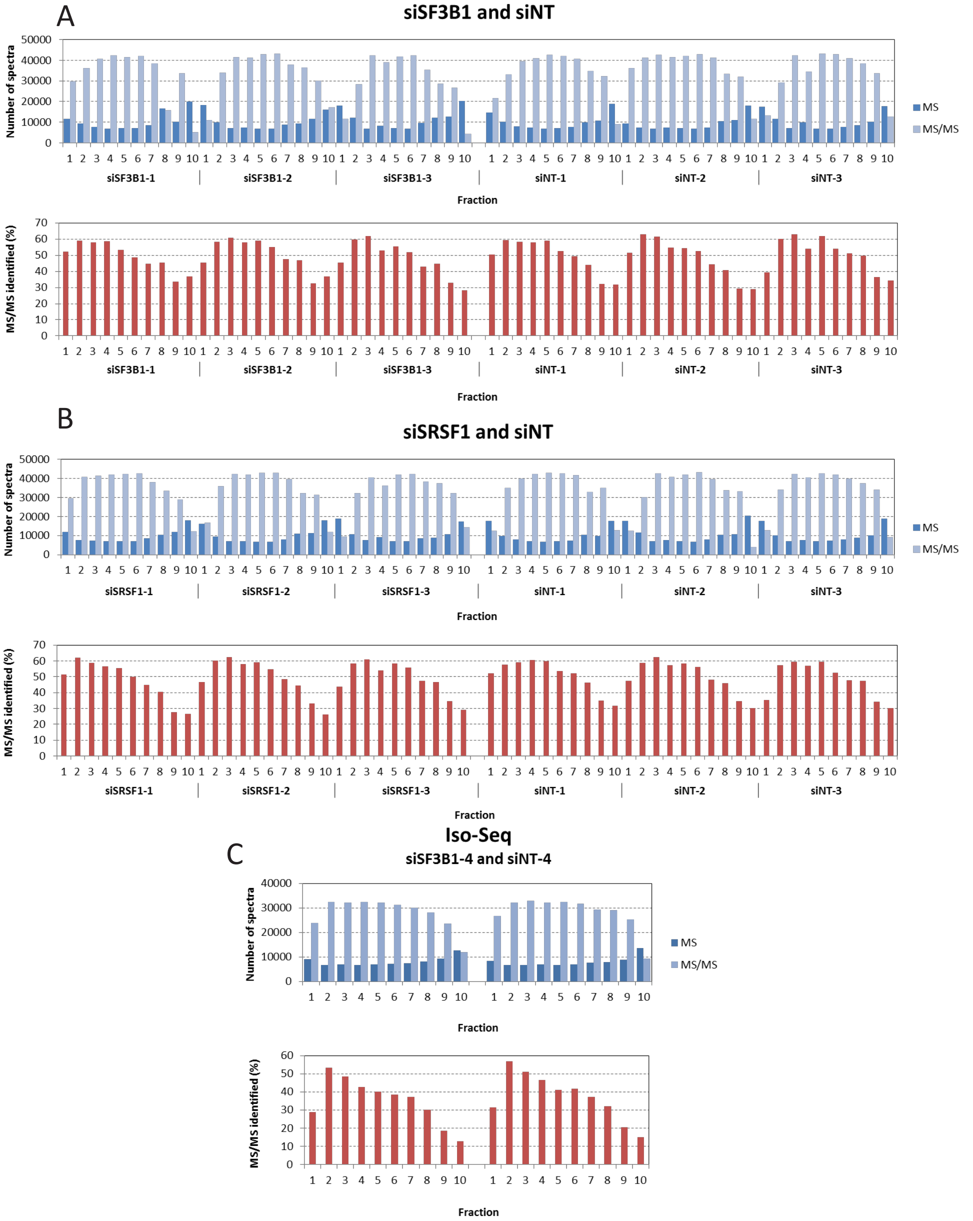
Supplementary Figure 2 Positive controls of alternative splicing A RT-qPCR quantification of the inclusion and exclusion variants of ADD3 exon 14 in siSF3B1 compared to the siNT control. Upon-down modulation of SF3B1 there is higher expression of the ADD3 isoform in which exon 14 is excluded. B RT-qPCR quantification of the inclusion and exclusion variants of CTNND1 exon 20 in siSRSF1 compared to the siNT control. Upon down-modulation of SRSF1 there is higher expression of the CTNND1 isoform in which exon 20 is excluded.



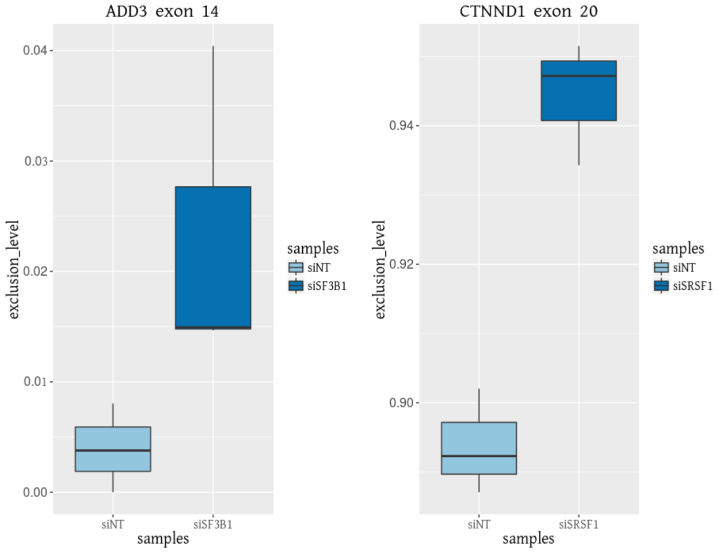
Supplementary Figure 3 Quality checks of RNA-seq data A The mean phred score value across each base position in the read was calculated with FastQC. Each line represents a sample that was subjected to RNA-seq. Plot was produced with MultiQC. B The average GC content of all the reads was calculated with FastQC. Each line represents a sample that was subjected to RNA-seq. Plot was produced with MultiQC. C Percentage of uniquely mapped RNA reads per sample were obtained from STAR output. The controls for siSF3B1 were labelled "siNT-x_48 " and the controls for siSRSF1 were labelled "siNT-x_72".



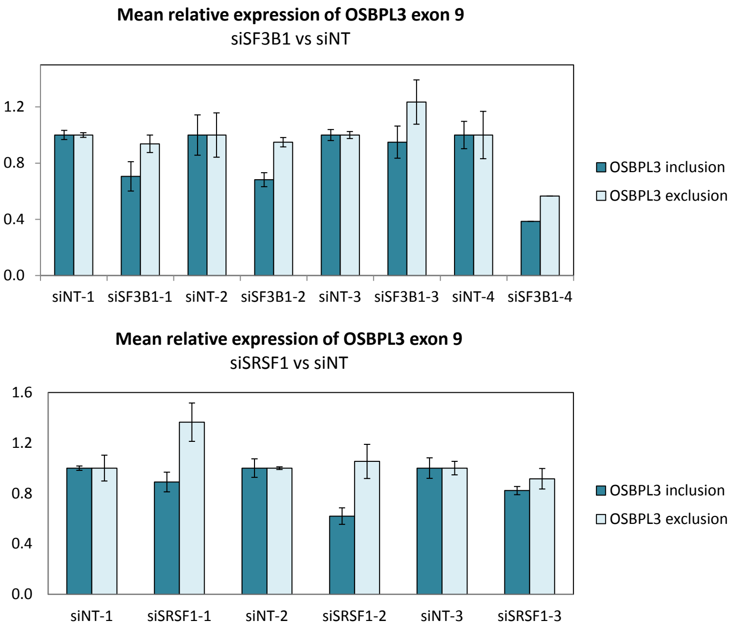
Supplementary Figure 4 The Coomassie-stained gels Coomassie-stained gels present the protein band pattern of all the samples subjected to mass spectrometry, indicating equal protein loads. Sample names are shown above each gel lane. Gel lanes not used are not annotated and are crossed out. The controls for siSF3B1 were labelled “siNT-x_48 “ and the controls for siSRSF1 were labelled “siNT-x_72”.



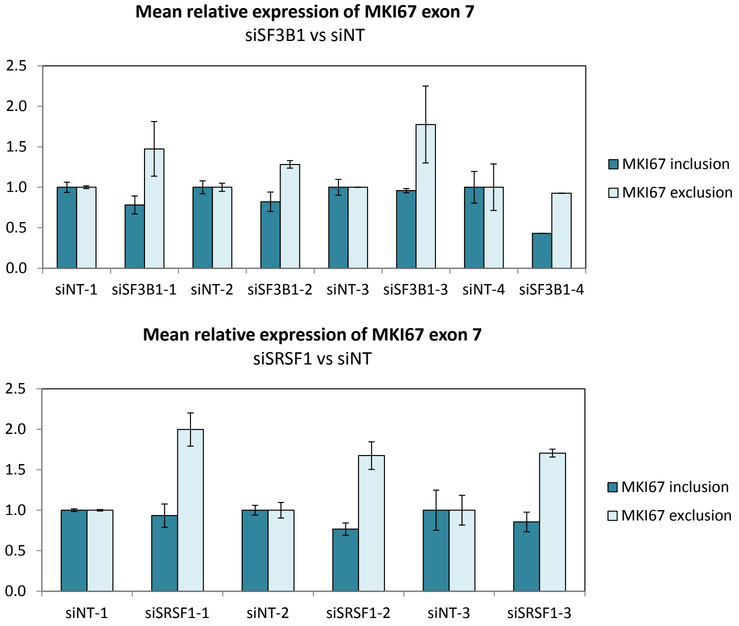
Supplementary Figure 5 Quality checks of mass spectrometry database search Bar plots represent the number of MS and MS/MS spectra identified and the percentage of MS/MS spectra identified in the database search performed by MaxQaunt. Numbers were obtained from MaxQuant summary files produced per each search; siSF3B1 and siNT (A), siSRSF1 and siNT (B) and Iso-Seq experiment (C).



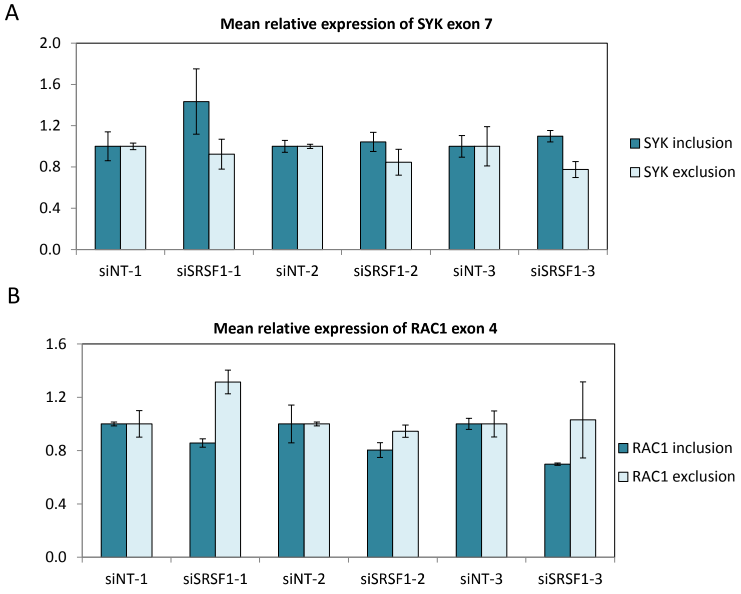
Supplementary Figure 6 Positive controls of alternative splicing identified with RNA-seq data. Exclusion isoforms for ADD3 exon 14 in siSF3B1 and its control (siNT) and for CTNND1 exon 20 for siSRSF1 and its control (siNT). Here, exclusion level is higher in siSF3B1 and siSRSF1 versus siNT, respectively, for both exclusion of exon 14 in ADD3 and exclusion of exon 20 in CTNND1. Exclusion level was calculated based on exclusion spanning reads divided by the sum of inclusion and exclusion spanning reads.



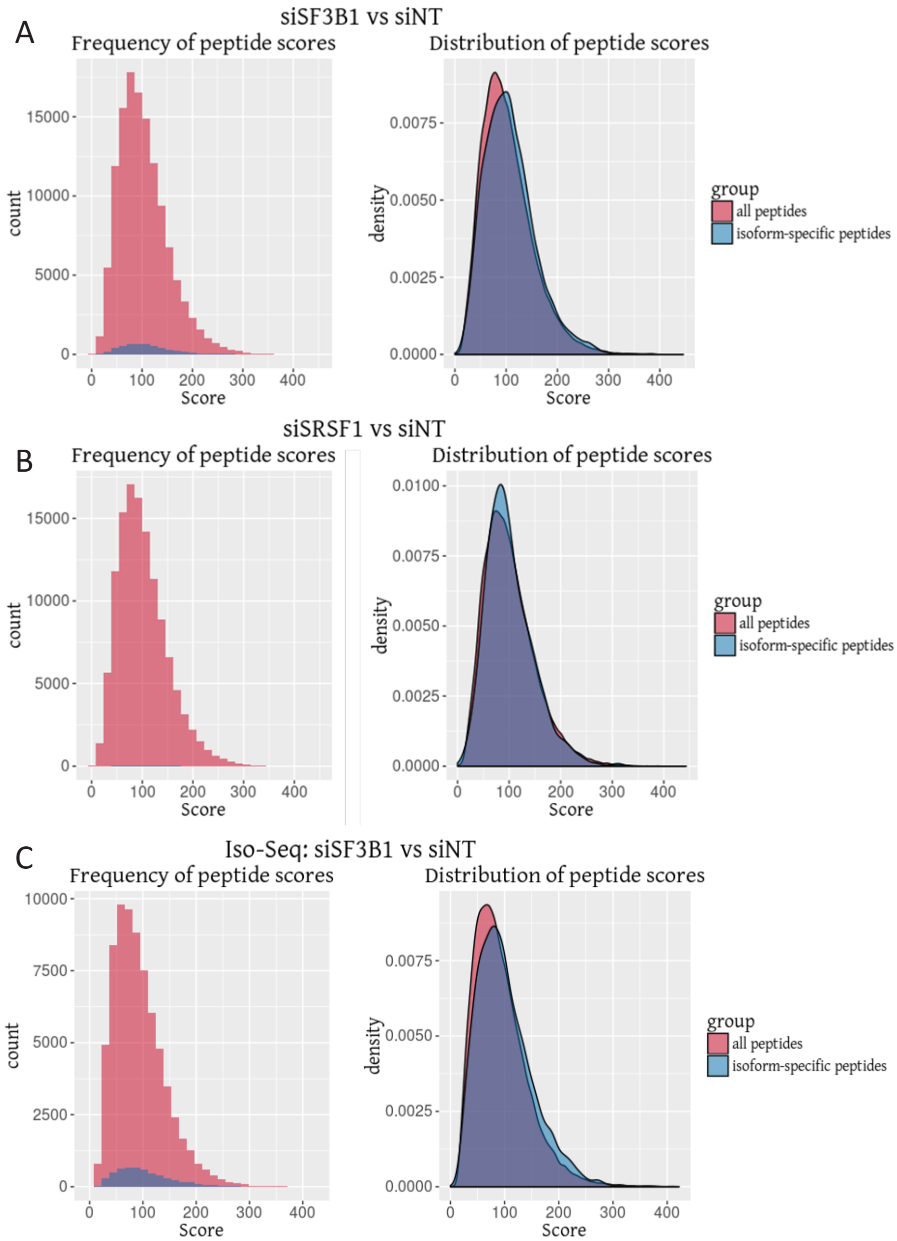
Supplementary Figure 7 RT-qPCR quantification of the inclusion and exclusion variants of OSBPL3 exon 9 upon knock-down of siSF3B1 and siSRSF1 compared to the control (siNT). Upon both down-modulation of SF3B1 or SRSF1 there is higher expression of the OSBPL3 isoform where exon 9 is excluded.



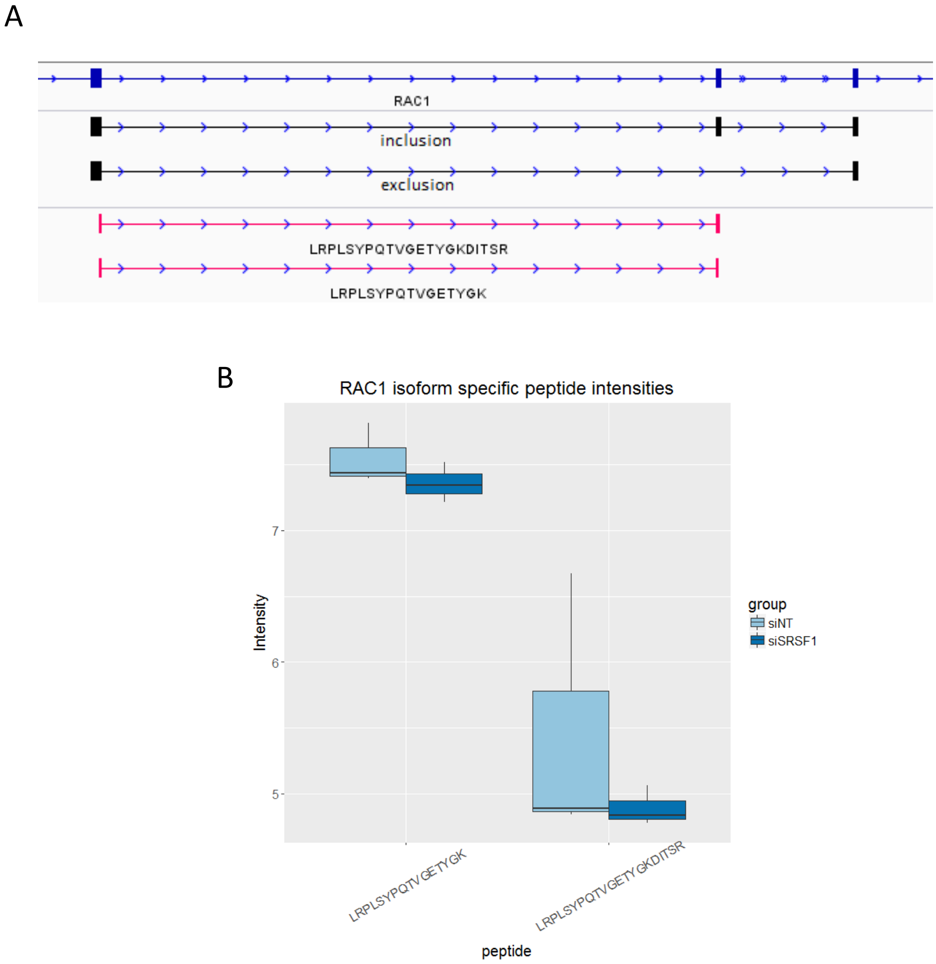
Supplementary Figure 8 RT-qPCR quantification of the inclusion and exclusion variants of MKI67 exon 7 in siSF3B1 and siSRSF1 in comparison to the control (siNT). Upon both down-modulation of SF3B1 or SRSF1 there is higher expression of the MKI67 isoform where exon 7 is excluded.



Supplementary Figure 9 RT-qPCR validation of skipped exon in SYK and RAC1 upon down-modulation of SRSF1 A RT-qPCR quantification of the inclusion and exclusion variants of SYK exon 7 in siSRSF1 in comparison to the control (siNT). Upon down-modulation of SRSF1 there is higher expression of the SYK isoform where exon 7 is included. B RT-qPCR quantification of the inclusion and exclusion variants of RAC1 exon 4 in siSRSF1 in comparison to the control (siNT). Upon down-modulation of SRSF1 there is higher expression of the RAC1 isoform where exon 4 is excluded.



Supplementary Figure 10 Comparison of peptide scores of isoform-specific peptides to all identified peptides Peptide scores (Andromeda scores) were calculated by MaxQuant and obtained from the peptide output file for each database search; siSF3B1 and siNT (A), siSRSF1 and siNT (B) and Iso-Seq experiment (C). The histogram represents the frequency (count) of each peptide score. Kernell density was calculated to obtain the distribution of peptide scores. The figures show that isoform-specific peptides are not scoring better or worse than the standard peptides, indicating that they were correctly identified.



Supplementary Figure 11 A An IGV screenshot of a fragment of RAC1 gene; in blue – RefSeq Gene, in black – skipped exon inclusion and exclusion variants identified in the RNA-seq data, in pink – split peptides supporting inclusion on the exon 4 in RAC1 gene. B Differences in peptide intensities between down-modulation of siSRSF1 and the control for the inclusion specific peptides for RAC1 isoform. Expression differences indicate that inclusion isoform is higher expressed in the control. Peptide expression confirms differential isoform expression obtained from the RNA-seq data.

SUPPLEMENTARY TABLES

Supplementary Table 1. Primer sequences for RT-qPCR quantification (Eurogentec, Belgium).

Gene	Description	Forward primer sequence	Reverse primer sequence	final primer conc (µM)	annealing temp (°C)
B2M	housekeeping	5'-TGACTTTGTCACAGCCCAAGATA-3'	5'-AATGCGGCATCTTCAAACCT-3'	0.5	60
SRSF1	splicing factor	5'-GTGGTTGTCTCTGGACTGCCTC-3'	5'-CCGTACAAACTCCACGACACC-3'	0.3	62
SF3B1	splicing factor	5'-GGAGTGGGCCTCGATTCTACA-3'	5'-GGCTTCTTCTGACCAAGCAAACT-3'	0.5	62
ADD3	Inclusion exon 14	5'-AGAGGACAATCGAACGTAAACAACAA-3'	5'-TGCGGTGACTGAGTTTGAGACTG-3'	0.3	59
ADD3	Exclusion exon 14	5'-ACCCATTTAGTCATCTCACAGAAGGA-3'	5'-GGAAAAGCTCATGGTTTTCTTCTAGG-3'	0.3	59
CTNND1	Inclusion exon 20	5'-ACACCCTTGATGCAGGACGA-3'	5'-CCTCATCATCCAAAACCAACACA-3'	0.3	60
CTNND1	Exclusion exon 20	5'-ACAACACCCTTGATGCAGAAGATTT-3'	5'-GGCACAATAGTCCAGCGAAGAA-3'	0.3	60
RAC1	Inclusion exon 4	5'-CCCTATCCTATCCGCAAACA-3'	5'-GGCAATCGGCTTGTCTTTGC-3'	0.3	60
RAC1	Exclusion exon 4	5'-CCTATCCGCAAACAGATGTGT-3'	5'-GGATACCACTTTGCACGGACAT-3'	0.3	60
SYK	Inclusion exon 7	5'-CCCATCCTGCGACTTGGTCA-3'	5'-GGGTGCAAGTTCTGGCTCAT-3'	0.3	60
SYK	Exclusion exon 7	5'-GAGTTCTTACTGTCCCATGTC-3'	5'-GGGAGGACGCAGGATGGGAA-3'	0.3	60
MKI67	Inclusion exon 7	5'-TTACAGGGGGAGACCCAACT-3'	5'-CCCTTCCCCTTGTTCTGGTC-3'	0.3	60
MKI67	Exclusion exon 7	5'-GACCCTGATGAGAGTGAGGGAA-3'	5'-AGAGGCGTATTAGGAGGCAA-3'	0.3	60
OSBPL3	Inclusion exon 9	5'-AATGCTCCAAAGACCTGGC-3'	5'-CCACCTCCTGTGCGATCTTT-3'	0.5	60
OSBPL3	Exclusion exon 9	5'-AATGCTCCAAAGACCTGGC-3'	5'-GGGACCTGGATGGCGTTGATA-3'	0.5	60

Supplementary Table 12. Alternative splicing events chosen for qRT-PCR validation, +/- indicate if the event was identified as significant by the proteogenomic analysis pipeline in an experiment.

Gene name	Skipped exon number	Skipped exon coordinates in hg19	siSF3B1 vs siNT	siSRSF1 vs siNT
SYK	7	93629412 - 93629481	-	+
RAC1	4	6438292 - 6438349	-	+
OSBPL3	9	24902818 - 24902911	+	+
MKI67	7	129913191 - 129914271	+	+

Chapter 5

ALTERNATIVE SPLICING AS A SOURCE OF CANDIDATE BIOMARKERS FOR EARLY DETECTION OF COLORECTAL CANCER

Malgorzata A Komor, Meike de Wit, Anne S Bolijn, Pien M Delis-van Diemen, Tim Schelfhorst, Sander R Piersma, Thang V Pham, Alexander Fish, Patrick Celie, Youri Hoogstrate, Mark de Jong, Guido Jenster, Beatriz Carvalho, Gerrit A Meijer, Connie R Jimenez, Remond JA Fijneman

In collaboration with the NGS-ProToCol consortium

IN PREPARATION

ABSTRACT

Background: Current strategies for early diagnosis of colorectal cancer (CRC) and its adenoma precursor lesions are largely based on the fecal immunochemical test (FIT) that detects the protein hemoglobin in stool. Although FIT is beneficial in its current form, its performance can be further improved using additional molecular markers. Cancer development is accompanied by alternative splicing, which results in expression of tumor-specific protein isoforms. The aim of this study was to identify proteins translated from alternatively spliced RNA that may serve as novel candidate biomarkers for early detection of CRC.

Materials and methods: Thirty CRCs, 30 advanced adenomas and 18 normal colon samples were subjected to RNA sequencing. From six patients a CRC, an adenoma and a normal colon tissue were analyzed by in-depth tandem mass spectrometry. The proteogenomic pipeline Splicify was applied to identify splice variants that were differentially expressed and translated into protein isoforms. Validation of differential splicing at the RNA level was performed globally in an independent mRNA sequencing series of 28 CRCs and 32 adenomas, and for a number of selected events by RT-qPCR.

Results: Comparative splicing analysis between CRCs and normal colon, between CRCs and adenomas and between adenomas and normal colon revealed 2876, 2285 and 1758 significant events, respectively. Translation of 916, 745 and 519 splicing events was confirmed by detection of isoform-specific peptides. These included known CRC isoforms of RAC1, KRAS and CTTN, along with novel candidates. Due to prominent quantitative differences at RNA and/or protein level, NT5C3A, EIF4H and PI4KB isoforms were further evaluated as candidate biomarkers for CRC. At the RNA level, NT5C3A isoform detected 80% and 72% of CRCs in the discovery and validation series, respectively.

Conclusions: Proteogenomics analysis of CRCs, adenomas, and normal colon yielded protein isoforms as novel candidate biomarkers for early detection of CRC, among which splice variants of NT5C3A.

INTRODUCTION

Early detection of colorectal cancer (CRC) is crucial for reducing CRC mortality rates, as stage I and II CRC is considered curable with a 5-year survival rate of approximately 90%[1]. In many countries, non-invasive CRC screening tests have been introduced like the fecal-immunochemical test (FIT) that detects the hemoglobin protein in stool and can be used as triage to colonoscopy[2, 3]. While the FIT is beneficial in its current form, its sensitivity in detection of CRCs (79%) or CRC precursor lesions, i.e. advanced adenomas (31%), can be further improved[4, 5]. Given that only approximately 5% of colorectal adenomas are estimated to progress to CRC[6], the improved screening test should detect not only CRCs but also this particular small fraction of adenomas that are likely to progress.

Molecular alterations accompanying colorectal carcinogenesis may serve as tumor-specific biomarkers for early detection of CRC[7]. Alternative splicing has been shown to play a role in each of the biological processes involved in carcinogenesis, commonly referred to as the hallmarks of cancer[8, 9] and a number of isoforms have been proven to play a role in CRC. For instance, Rac1b is a cancer-specific splice variant formed by inclusion of the additional exon four in the RAC1 gene, forming a constitutively activated protein[10, 11]. Rac1b is often overexpressed in CRC cells and contributes to cell survival[12]. KRAS4A, a KRAS isoform formed by exclusion of exon four, was shown to have a prognostic value for CRC[13]. Finally, an inclusion variant of CTTN, with additional exon 11, has been shown to impact CRC cell migration and invasion[14]. RAC1 and CTTN isoforms are the most frequent alternatively spliced events in CRC with a known functional impact[15]. As alternatively spliced RNA is often translated into proteins, tumor-specific protein isoforms may be a source of candidate biomarkers for early detection of CRC that, in contrast to RNA transcripts, could be detected by an antibody-based assay similar to FIT.

The aim of the present study was to characterize alternative splicing changes accompanying colorectal tumor development and to identify protein isoforms that could serve as novel biomarkers for early detection of CRC.

MATERIALS AND METHODS

Design of the entire study is presented in Figure 1.

SAMPLES

Discovery series: Fresh frozen tissue material from 30 colorectal advanced adenoma, 30 colorectal cancer and 18 normal colorectal mucosa samples were collected at the department of Pathology of the Amsterdam UMC (location VU University Medical Center) as described previously[16, 17] and used for RNA isolation. From this series, eighteen samples were selected for tandem mass spectrometry proteomics analysis

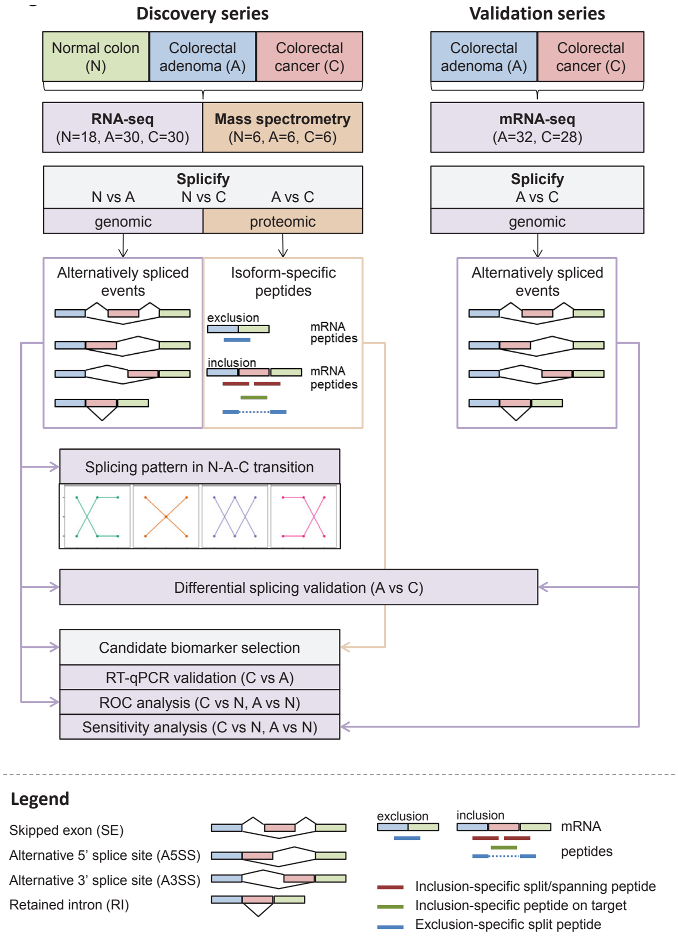

Discovery series
Normal colon (N)
Colorectal adenoma (A)
Colorectal cancer (C)
RNA-seq
(N=18, A=30, C=30)
Mass spectrometry
(N=6, A=6, C=6)
Splicify
N vs A
N vs C
A vs C
genomic
proteomic
Alternatively spliced events
Isoform-specific peptides
exclusion
mRNA
peptides
inclusion
mRNA
peptides
Validation series
Colorectal adenoma (A)
Colorectal cancer (C)
mRNA-seq
(A=32, C=28)
Splicify
A vs C
genomic
Alternatively spliced events
Splicing pattern in N-A-C transition
Differential splicing validation (A vs C)
Candidate biomarker selection
RT-qPCR validation (C vs A)
ROC analysis (C vs N, A vs N)
Sensitivity analysis (C vs N, A vs N)
Legend
Skipped exon (SE)
Alternative 5' splice site (A5SS)
Alternative 3' splice site (A3SS)
Retained intron (RI)
exclusion
inclusion
mRNA
peptides
Inclusion-specific split/spanning peptide
Inclusion-specific peptide on target
Exclusion-specific split peptide

Figure 1. A flow diagram of the study design. Discovery series consisted of fresh-frozen tissue pieces of normal colon, colorectal adenoma and colorectal cancer samples, from which 18, 30, 30 respectively were used for RNA-seq and 6, 6, 6 respectively for tandem mass spectrometry. Proteogenomic pipeline Splicify was used to identify differential splicing events at RNA and at protein level between normal colon and cancer, between normal colon and adenoma and between adenoma and cancer. Splicing pattern in the transition from normal colon, through adenoma to cancer was identified based on significant quantitative changes in RNA inclusion/exclusion levels of splice variants identified in the discovery series. Validation series consisted of fresh-frozen tissue pieces of 32 colorectal adenomas and 28 CRCs, for which mRNA sequencing was performed[16]. Genomic part of Splicify pipeline was used to identify differential splicing events between adenomas and cancers. Overlap analysis was performed with differential splicing events between adenomas and cancers identified in the discovery series for validation. Next, candidate biomarkers were selected based on the splicing results at RNA and protein level. RT-qPCR validation of quantitative differences between RNA inclusion/exclusion levels was performed in the adenoma and cancer samples from the discovery series. ROC analysis was performed for RNA inclusion/exclusion level for candidate biomarkers to evaluate their performance in detection of cancers and adenomas in comparison to normal colon samples. Thresholds for RNA inclusion/exclusion levels of candidate biomarkers were selected to reach specificity at ~95% and sensitivities for detection of cancers and adenomas were evaluated in the discovery and validation series. Four alternatively spliced events on RNA level were analyzed in this study: skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS) and retained intron (RI). Blocks represent exons and lines represent alternative junctions after splicing, indicating inclusion and exclusion variants of each alternatively spliced event on RNA level with isoform-specific peptides supporting these variants. Red line: inclusion-specific split or spanning peptides; green line: inclusion-specific peptide on target; and blue line: exclusion specific split peptide.

consisting of six patient-matched triplets of normal colon, adenoma and cancer. All normal samples were adjacent to colorectal neoplasia, i.e. resection margins of surgical specimens.

Validation series: Fresh frozen tissue material from 32 colorectal advanced polypoid adenomas and 28 colorectal carcinomas were collected at the department of Pathology of the Amsterdam UMC (location VU University Medical Center) and described in previous studies[16, 18]. RNA isolated from fresh frozen specimens of these samples and mRNA sequencing data were available.

RNA SEQUENCING

RNA isolation and RNA sequencing (RNA-seq) was performed as previously described for both discovery and validation series[16]. RNA-seq data are available through the European Genome-phenome Archive. For the discovery series (study ID: EGAS00001002854)[16], on average, 67 million reads were obtained per sample. For the validation series (dataset IDs: EGAD00001004058, EGAD00001004059)[16], on average 32 million reads were obtained per sample.

PROTEOMICS

Proteomics was performed for 18 samples derived from six patients from the discovery series, representing six patient-matched triplets of normal colon, adenoma and CRC. Twenty sections of 16 µm were cut from the snap-frozen tissue pieces, with frozen sections taken before and after for histopathological verification. Tissue sections were lysed in a 1:30 ratio, ~1 mg of tissue was lysed in 30µl of NuPAGE LDS sample buffer (Fisher Scientific, Landsmeer, the Netherlands) containing 100 mM DTT, thoroughly vortexed for 1 minute, heated at 70°C for 10 minutes and sonicated (3 cycles of 20 seconds on, 20 seconds off). Lysates were centrifuged for 10 minutes at 20,000 x *g* and the supernatant was transferred to a new tube and the pellet was discarded. Denatured samples were loaded on a 4-12% NuPAGE Bis-Tris polyacrylamide gel, and proteins were resolved at 200 V in MOPS buffer containing NuPAGE antioxidant for 55 minutes. The gel was fixed for 15 minutes in 50% ethanol/3% phosphoric acid, stained with colloidal Coomassie (0.1% Coomassie Brilliant Blue G-250, 30% methanol, 3% phosphoric acid, 15% ammonium sulfate), and destained overnight with Milli-Q water. Each gel lane was cut into 10 slices. The in-gel digestion and on-line liquid chromatography-tandem mass spectrometry (LC-MS/MS) procedures were performed as described previously[19]. Briefly, intact peptide mass spectra and fragmentation spectra were acquired on a Q Exactive mass spectrometer (Thermo Fisher, Bremen, Germany) in a data dependent acquisition mode. Intact masses were measured at resolution 70.000 (at m/z 200) in the orbitrap using an AGC target value of 3 × 106 charges. The top 10 peptide signals (charge-states 2+ and higher) were submitted to MS/MS in the HCD (higher-energy collision) cell (1.6 m/z isolation width, 25% normalized collision energy). MS/MS spectra were acquired at resolution 17.500 (at m/z 200) in the orbitrap using an AGC target value of 1 × 106 charges and an underfill ratio of 0.5%. Dynamic exclusion was applied with a repeat count of 1 and an exclusion time of 30 s.

RNA-SEQ AND PROTEOMICS DATA ANALYSIS

Discovery series: Splicify[19] was applied on RNA-seq and mass spectrometry data in the comparative settings: CRCs *versus* colorectal adenomas (C vs A) and CRCs *versus* normal colon samples (C vs N) and colorectal adenomas *versus* normal colon samples (A vs N). Differential splice variants on RNA and protein level were obtained for all comparisons using default settings, including MaxQuant protein identification. Mutually exclusive exons events were excluded due to high false positive rate for this kind of events. No imputation was applied for missing values in the proteomics data. Relative expression of inclusion and exclusion isoforms were calculated using RNA-seq data and identification of the RNA splice variants translated into proteins, i.e. supported by isoform-specific peptides (Figure 1).

RNA alternatively spliced events were categorized into groups showing the same pattern of FDR threshold and sign of “InclusionLevelDifference” (Supplementary Table 1, Figure 2A). Group 1 was defined by significant changes between CRC and normal colon and between adenoma and normal colon (FDR ≤ 0.05) and the same direction of change in both comparisons (sign(C vs N) = sign(A vs N)), while the comparison of cancers to adenomas was not significant (FDR > 0.05). Group 2 was defined by significant changes and the same direction of change for all comparisons (FDR ≤ 0.05, sign(C vs N) = sign(A vs N) = sign(C vs A). Group 3 was defined by significant changes between CRC and adenoma and between adenoma and normal colon (FDR ≤ 0.05) and different direction of change between these comparisons (sign(C vs A) ≠ sign(A vs N)). Group 4 was defined by significant changes between CRC and normal colon and between CRC and adenoma (FDR ≤ 0.05) and the same direction of change in both comparisons (sign(C vs N) = sign(C vs A)), while the comparison of adenoma to normal colon was not significant (FDR > 0.05, Supplementary Table 1). Overrepresentation analysis of parental genes, *i.e.* genes affected by alternative splicing, from each pattern group was performed with the use of hallmark gene sets from Molecular Signature Database (MSigDB)[20], p-values were obtained with the hypergeometric test for overlap significance and FDR values with the use of Benjamini-Hochberg correction.

Candidate biomarkers were selected based on the presence of peptides supporting the isoform higher expressed in CRC samples when compared to adenoma or normal colon samples. Additionally, detection of peptides for both inclusion and exclusion variants, concordance of quantitative differences between RNA and protein level were taken into account.

Validation series: RNA-seq data was analyzed with the use of Splicify[19] step1 (genomic part) for the differential splicing analysis between CRCs and adenomas. Overlap analysis between differential splice variants in the discovery series and validation series was performed based on the coordinates of spliced fragments, upstream and downstream exons and the direction of "InclusionLevelDifference". Overlap significance was evaluated with the use of a hypergeometric test, where the reference number of events for testing was defined per alternative splicing event as all identified events in the validation series of a particular kind (including not significant ones).

RT-qPCR VALIDATION

Reverse transcription and quantitative PCR (RT-qPCR) was performed using the iScript cDNA synthesis kit (1708891, Bio-Rad) and SYBR Green (4309155, Thermo Fisher Scientific), to measure expression of the splicing variants of NT5C3A, PI4KB and EIF4H in the RNA samples of cancers and adenomas from the discovery series. Glucuronidase beta (GUSB) was used as a housekeeping reference gene. In brief, gene expression was measured using 2 µl of 10 ng/µl cDNA in a 25 µl SYBR Green reaction (see Supplementary Table 2 for primers and conditions), as described previously[21]. Inclusion (or exclusion) level was obtained by division of inclusion (or exclusion) expression by the sum of inclusion and exclusion expression as quantified with RT-qPCR.

STATISTICAL ANALYSIS

In the discovery series receiver operating characteristic (ROC)[22] analysis was used to evaluate the performance of RNA inclusion/exclusion variants to discriminate cases, i.e. CRCs or adenomas, from controls, i.e. normal colon samples, by calculating partial area under the curve (pAUC) at the specificity of 95%-100%. A threshold for RNA inclusion/exclusion levels was defined to reach 94% specificity and evaluate sensitivity for cancers and adenomas at 94% specificity. Given that there are 18 controls, 94% specificity reflects 17 out of 18 controls classified correctly. Due to the lack of normal colon samples in the validation series, the threshold for RNA exclusion/inclusion level defined in the discovery series was applied for cancers and adenomas from validation series to obtain approximate sensitivity.

RESULTS

GLOBAL SPLICING CHANGES ACCOMPANY COLORECTAL TUMOR PROGRESSION

With the aim to identify splicing isoforms that play a role in colorectal tumor development, we have performed pairwise comparative splicing analysis for normal colon, colorectal adenoma and CRC samples (see Figure 1 for the study design). Differential splicing analysis between CRCs and normal colon samples (C vs N) revealed 2876 splicing events on RNA level ($FDR \leq 0.05$) in total. The vast majority of the alternative splicing events were skipped exon (SE, n = 2229), followed by alternative 5' splice site (A5SS, n = 259), alternative 3' splice site (A3SS, n = 241) and retained intron (RI, n = 147; Table 1; Supplementary Table 3A-D). To obtain high protein sequence coverage for isoform-specific peptide identification, in-depth proteomics was performed reaching over 9100 protein groups in total, and on average 32% of protein sequence coverage. With the use of the isoform-specific peptides detected in the mass spectrometry data, RNA events translated into proteins were identified. For SE events, 783 (35% of RNA events) had peptide confirmation for at least one of the inclusion or exclusion isoforms, while for A5SS, A3SS and RI, 47 (18%), 53 (22%) and 33 (22%) had peptide confirmation, respectively (Table 1; Supplementary Table 4). Differential splicing analysis between CRCs and adenomas (C vs A) revealed 2285 alternative splicing events, in particular 1838 SE, 160 A5SS, 150 A3SS and 137 RI (Table 1; Supplementary Table 5A-D). On the protein level, for 745 (33%) events there was at least one isoform-specific peptide identified, in particular 656 (36%) SE, 28 (18%) A5SS, 33 (22%) A3SS and 28 (20%) RI events had peptide confirmation (Table 1; Supplementary Table 6). For the comparison of colorectal adenomas to normal colon samples (A vs N), on the RNA level 1758 events were identified in total (SE = 1351, A5SS = 176, A3SS = 157, RI = 74, Supplementary Table 7A-D), from which 30% had peptide confirmation (Table 1; Supplementary Table 8). Overlap analysis on RNA level revealed that approximately 50% of the splicing events are common for at least two out of three comparisons (Supplementary Figure 1).

CHARACTERIZATION OF SPLICING PATTERNS IN COLORECTAL TUMOR DEVELOPMENT

To further study splicing changes that accompany the transition from normal colon to adenoma to CRC, we categorized the RNA alternatively spliced events into groups that follow the same quantitative pattern along the normal-adenoma-cancer sequence (Figure 2; Supplementary Tables 1 and 9). Parental genes of the splicing events of each pattern were subjected to an overrepresentation analysis assessing presence in the hallmark gene sets of MSigDB[20](Supplementary Table 10).

Group 1 represented neoplasia-specific splicing changes, i.e. splicing changes occurring at the normal-to-adenoma transition and maintained at the cancer stage. Genes that were spliced in the pattern represented by Group 1 were enriched in the “mitotic spindle”, “oxidative phosphorylation”, “myogenesis”, “apoptosis”, “UV response down” and “epithelial mesenchymal transition” gene sets. Group 2 represented the gradient pattern and consisted of events that gradually increase/decrease their expression through the transition from normal colon, through adenoma to CRC. Gene set overrepresentation analysis did not reveal any significant enrichment for this group. Group 3 represented adenoma-specific events and was enriched with genes from the “mitotic spindle” gene set. Finally, group 4 consisted of cancer-specific events and was enriched in genes from “E2F targets”, “Peroxisome” and “DNA repair” gene sets (Figure 2A, 2D, and Supplementary Table 10). The cancer-specific group 4 was the most prevalent one with 948 events in total (Figure 2B, Supplementary Table 9); and most frequent among three out of four alternative splicing event types, SE, A3SS and RI (Figure 2C). Group 1 was the second most common pattern in the dataset with 743 events (Figure 2B). Group 2 was the only group that required an event to be statistically significant in all three comparisons, which had an impact on its final size and overrepresentation analysis (Figure 2B-C; Supplementary Tables 9-10). Next, in each pattern group we evaluated which isoform; inclusion or exclusion, is higher expressed in normal colon in comparison to adenoma and/or cancer samples. In all four groups inclusion variant was the dominant one in normal colon, meaning that in most cases exclusion variant the aberrant isoform (Figure 2E).

Table 1. **Overview of alternatively spliced events on RNA and protein level**. Results are displayed separately for comparisons colorectal cancer *vs* normal colon, colorectal cancer *vs* colorectal adenoma and colorectal adenoma *vs* normal colon. Number of events on RNA level was determined by FDR threshold of ≤ 0.05. Number of events on protein level was obtained based on the presence of at least one isoform-specific peptide for inclusion and/or exclusion variant of an RNA event. The percentage corresponds to the fraction of RNA events with peptide confirmation.

Number of events					
	Skipped exon	**Alternative 5' splice site**	**Alternative 3' splice site**	**Retained intron**	**All events**
	Colorectal cancer versus normal colon				
RNA	2229	259	241	147	2876
Protein	783 (35%)	47 (18%)	53 (22%)	33 (22%)	916 (32%)
	Colorectal cancer versus colorectal adenoma				
RNA	1838	160	150	137	2285
Protein	656 (36%)	28 (18%)	33 (22%)	28 (20%)	745 (33%)
Colorectal adenoma versus normal colon					
RNA	1351	176	157	74	1758
Protein	450 (33%)	25 (14%)	29 (18%)	15 (20%)	519 (30%)

ADENOMAS EXPRESS A NUMBER OF ISOFORMS INHERENT TO COLORECTAL CANCER

To evaluate the results of Splicify in the context of current knowledge about alternative splicing in colorectal cancer, we selected skipped exon events for RAC1, KRAS and CTTN (Figure 3A) that have been described in literature to frequently occur in colorectal cancer and to play a role in carcinogenesis[10, 12, 13]. Interestingly, all three isoforms belonged to Group 1, representing neoplasia-specific events. At the RNA level, the isoforms expressed higher in cancer than in normal colon (inclusion for RAC1 and CTTN and exclusion for KRAS) were already prevalent in colorectal adenomas (Figure 3B). This indicates that alternative splicing of these genes may play a role in the transition from normal colon to colorectal adenomas, but not necessarily in the adenoma-to-carcinoma progression as there is no further increase between adenomas and cancers. Additionally, we evaluated quantitative differences of these isoforms on the protein level, by analyzing isoform-specific peptide intensities, for representative peptides per RNA splice variant higher expressed in CRC. For RAC1 and CTTN the inclusion-specific split peptide and peptide on target, respectively, had significantly increased intensities in the adenoma and cancer samples compared to normal colon (RAC1: p-value = 0.01 and 0.01, respectively; CTTN: p-value = 0.03 and 0.04, respectively), while the difference between cancers and adenomas was not significant (RAC1: p-value = 0.80; CTTN: p-value = 0.57; Figure 3C).

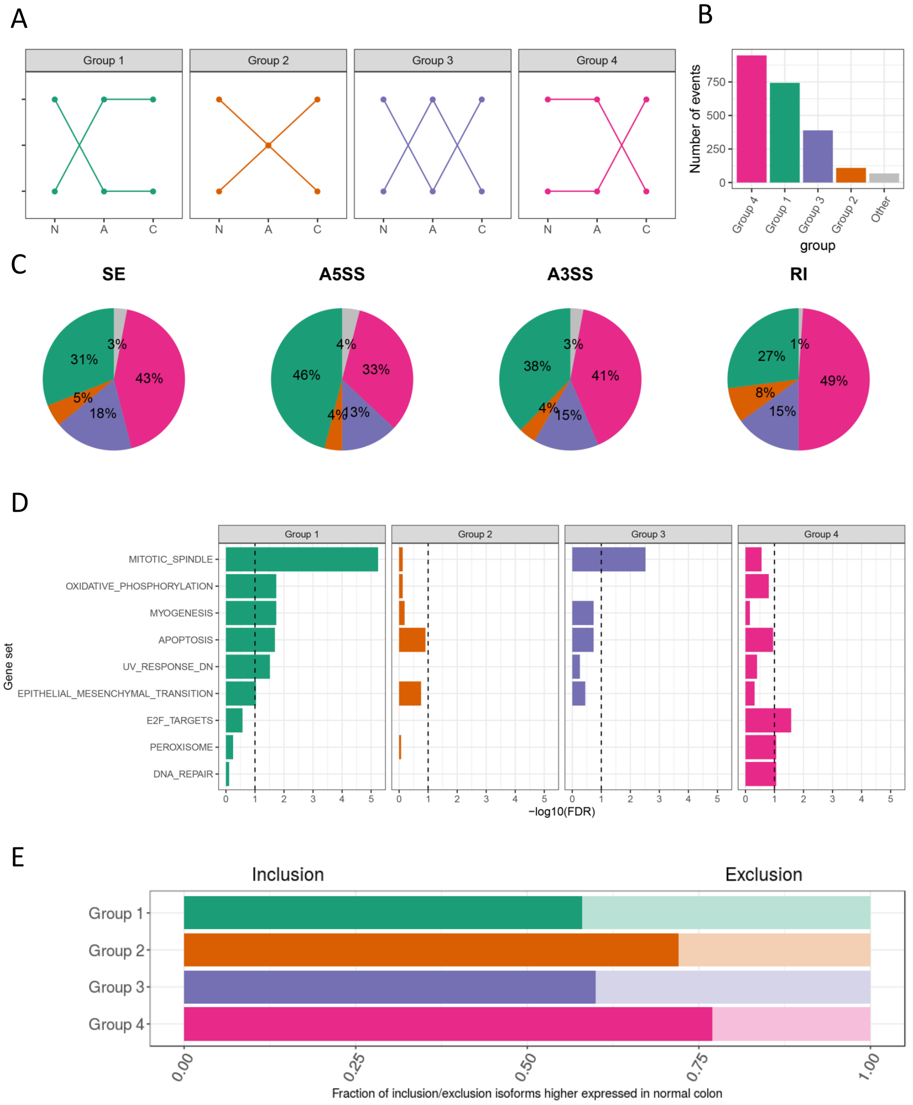


Figure 2. A. Patterns of alternative splicing in normal colon (N), colorectal adenoma (A) and colorectal cancer (C) samples. Changes among colorectal tumor development were classified into four distinct groups. B. Number of alternatively splice events in each group. Events not representing any of the patterns were classified as “Other”. C. Frequency of each group in different alternatively spliced events, skipped exon (SE), alternative 5’ and 3’ splice site (A5SS, A3SS) and retained intron (RI). D. Overrepresentation analysis of parental genes from each splicing pattern group in the hallmark gene sets of MSigDB. Significance threshold was applied at FDR ≤ 0.1 (highlighted with a dashed line). E. Fraction of alternatively spliced events per pattern group, for which RNA inclusion/exclusion variant was higher expressed in normal colon samples compared to adenomas and/or cancers.

For the KRAS isoform, the differences on the peptide level were not significant for all comparisons (A vs N: p-value = 0.24; C vs N: p-value = 0.37; C vs A: p-value = 0.80). These results indicate that expression of RAC1 and CTTN protein isoforms differs between normal colon and colorectal neoplasia and may have a functional role in the normal-to-adenoma transition, while KRAS isoforms may differ only on RNA level.

In silico validation of differential splicing between colorectal cancers and adenomas at the RNA level

For global validation of alternatively spliced events identified in this study, we performed the genomic part of the Splicify analysis for a cancer (n=28) *versus* adenoma (n=32) comparison using mRNA sequencing validation dataset. We identified in total 1247 differentially spliced events, which is 55% of the total number of spliced events observed in the discovery series (Supplementary Table 11A-D). Of the alternative splicing events identified in the discovery series, 26%, 21%, 18% and 31% of SE, A5SS, A3SS and RI events, respectively, were also observed in the validation series. After correcting for the overall lower number of events in the validation series these figures were 47%, 36%, 38% and 53% of the SE, A5SS, A3SS and RI events identified in the validation series that overlapped with the discovery series, respectively. Overlap analysis on the event level revealed significant overlaps for all event types (SE: p-value < $2.20e^{-308}$, A5SS: p-value = $2.11e^{-19}$, A3SS: p-value = $9.95e^{-19}$, RI: p-value = $6.26e^{-14}$, Supplementary Figure 2), thus validating the splicing changes in the transition from colorectal adenoma to cancer.

Selection of splice variants as candidate biomarkers for CRC

As only 5% of adenomas are estimated to progress to CRC, potential candidate biomarkers for CRC screening should be selected from Groups 2 and 4, where expression of splice variants increases from adenoma to cancer, and remains low in normal colon samples (Figure 2A). Candidate biomarkers were selected from the gradient and cancer-specific pattern groups 2 and 4, when at least one isoform-specific peptide was identified for the RNA variant higher expressed in cancer samples, which was the case for 12 and 66 events for Groups 2 and 4, respectively (Supplementary Tables 4, 6 and 8). Other selection criteria were either identification of isoform-specific peptides for both inclusion and/or exclusion variants or concordant quantitative differences on RNA and protein level. Based on these criteria, skipping of exon 2 in NT5C3A emerged as a candidate. It was differentially spliced between all three group comparisons and assigned to the gradient Group 2 (Figure 4A-B). The exclusion variant was higher expressed in colorectal adenomas and even higher in CRCs compared to normal colon samples. Two isoform-specific peptides were identified in the mass spectrometry data; one exclusion-specific split peptide and one inclusion-specific peptide on target (Supplementary Tables 4, 6 and 8). The exclusion-specific peptide intensity showed a similar pattern as RNA exclusion-level, while the inclusion-specific peptide intensity was opposite and significantly differential between cancer and normal colon samples (p-value = 0.02), as well as cancer and adenoma samples (p-value = $2.25e^{-3}$, Figure 4C). The quantitative differences at protein level provide strong evidence that NT5C3A splice variants may serve as potential biomarkers for CRC. As for other candidates, skipped exon 5 in EIF4H was assigned to the gradient group as well. The inclusion-level increased gradually from normal colon to colorectal adenoma and colorectal cancer samples (Figure 4B). Two split peptides were identified; exclusion- and inclusion-specific, and even though the differences in peptide intensities between groups were not statistically significant, they followed the pattern observed on the RNA level (Figure 4C).

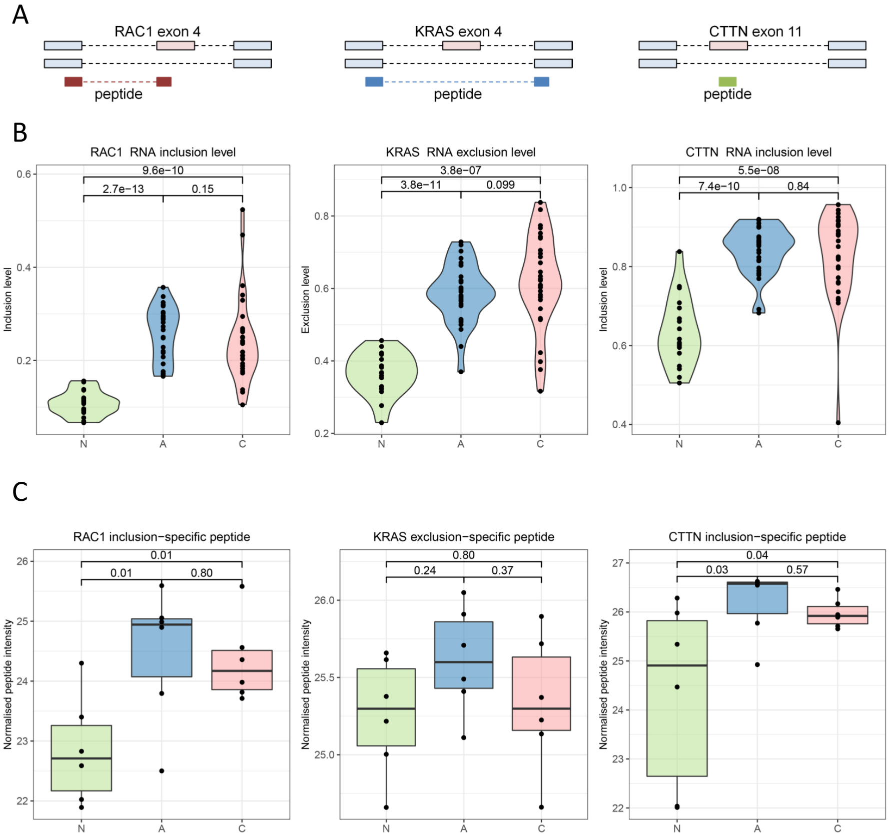


Figure 3. Known isoforms identified by Splicify. A. Schematic overview of skipped exon events for RAC1 (exon4), KRAS (exon 4) and CTTN (exon 11). For each event at least one isoform-specific peptide was identified, and one isoform-specific peptide was selected for the isoform higher expressed in CRC; inclusion-specific split peptide for RAC1, exclusion-specific split peptide for KRAS and inclusion-specific peptide on target for CTTN. B. Quantitative representation of RAC1, KRAS and CTTN isoforms in normal colon (N), colorectal adenoma (A) and cancer (C) samples on RNA level. Inclusion (exclusion) level was calculated by number of inclusion-specific (exclusion-specific) RNA reads divided by the sum of inclusion- and exclusion-specific RNA reads. P-values were obtained with the use of Mann-Whitney test. C. Quantitative representation of isoform-specific peptides (schematically shown in Figure 3A) in normal colon (N), colorectal adenoma (A) and cancer (C) samples. Normalized peptide intensities were plotted in each group, p-values were obtained from the Splicify analysis.

Other candidates from Group 2 with isoform-specific peptides identified for both inclusion and exclusion variants were skipped exon events in MYL6, COL6A3, CALD1, SPTAN1 and DYNC1I2. In the cancer-specific Group 4, we identified exclusion variant of skipped exon 4 in PI4KB to be more abundant in CRC compared to adenoma and normal colon samples. As an exclusion-specific peptide was identified in only one sample, differential expression analysis at peptide level was not feasible (Supplementary Figure 3). Next to PI4KB, in the cancer-specific splicing group, there were also candidates with peptides identified for both inclusion and exclusion variants; RI event for TGOLN2, A5SS events for PAIP1, CNBP and skipped exon events for STK39, TPD52L2, YBX3, MYO1B, TNC, ARFIP1 and HNRNPK. Quantitative differences on RNA level of the NT5C3A, EIF4H and PI4KB isoforms between colorectal cancer and adenoma samples were validated with the use of RT-qPCR (Supplementary Figure 4).

Next, we evaluated the potential diagnostic performance of NT5C3A, EIF4H and PI4KB splice variants, which were higher expressed in CRC than normal colon samples. As the RNA dataset encompassed a higher number of samples, we performed the analysis at RNA exclusion level for NT5C3A and PI4KB and inclusion level for EIF4H (Figure 5). Due to the high specificity required for a screening test, we evaluated partial area under the curve (pAUC) metrics at the specificity level of 95-100% for the receiver operating characteristic (ROC) curve and sensitivity at the specificity level of approximately 95%. For NT5C3A RNA exclusion level, a pAUC of 88% was observed for CRCs and 79.5% for adenomas. A threshold of RNA exclusion level was defined at 94% specificity (threshold = 0.348), and confusion matrices and sensitivity for CRCs (80%) and adenomas (60%) were obtained (Supplementary Table 12; Figure 6A). For EIF4H RNA inclusion level, lower performance was observed with a pAUC for CRC of 70.9% and 53.8% for the adenomas, while sensitivity at specificity level of 94% (threshold = 0.093) was only 43% for CRC and 10% for adenomas. Finally, PI4KB RNA exclusion level exhibited the best performance in CRC detection with pAUC of 93.2% and sensitivity of 97% at a specificity level of 94% (threshold = 0.746), while the pAUC for adenomas was 59% and sensitivity 33%.

Skipped exons for NT5C3A, EIF4H and PI4KB were also significantly differential in the validation series (Supplementary Table 11A), therefore we evaluated their potential diagnostic performance in the validation series as well. As the validation series did not contain controls, we applied thresholds for either RNA exclusion or inclusion levels as defined in the discovery series, where cancers where compared to normal colon samples. In this way, we obtained sensitivities for CRC and adenoma detection. For NT5C3A RNA exclusion level application of the same threshold revealed 73% and 28% sensitivity for CRCs and adenomas, respectively. For EIF4H RNA inclusion level sensitivities for CRC and adenomas were 64% and 16%, respectively. Lastly, for PI4KB RNA exclusion level the sensitivity of detection of CRC remained high with 86% while for adenomas it reached 22% (Supplementary Table 12; Figure 6B).

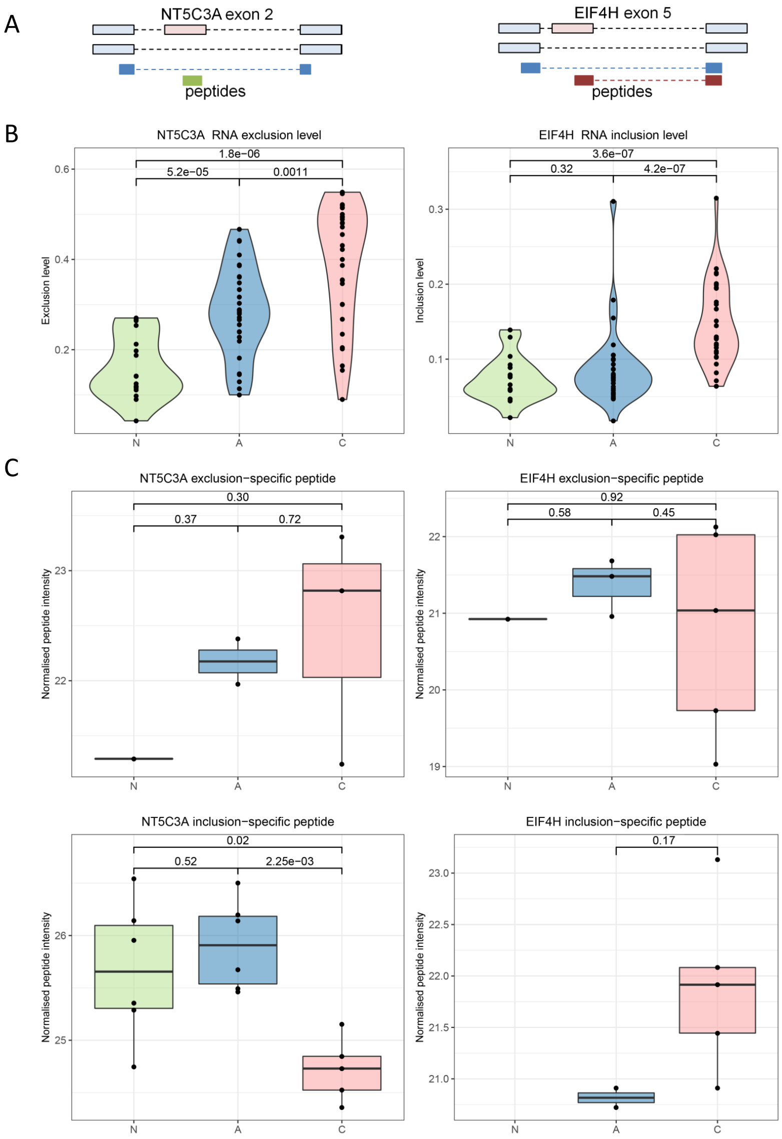


Figure 4. A. Schematic overview of skipped exon events for NT5C3A (exon 2) and EIF4H (exon 5). For each event two isoform-specific peptides were identified; both inclusion- and exclusion-specific peptides. B. Quantitative representation of NT5C3A and EIF4H isoforms in normal colon (N), colorectal adenoma (A) and cancer (C) samples on RNA level. Inclusion (exclusion) level was calculated by number of inclusion-specific (exclusion-specific) RNA reads divided by the sum of inclusion- and exclusion-specific RNA reads. P-values were obtained with the use of Mann-Whitney test. C. Quantitative representation of isoform-specific peptides in normal colon (N), colorectal adenoma (A) and cancer (C) samples. Normalized peptide intensities were plotted in each group, p-values were obtained from the Splicify analysis.

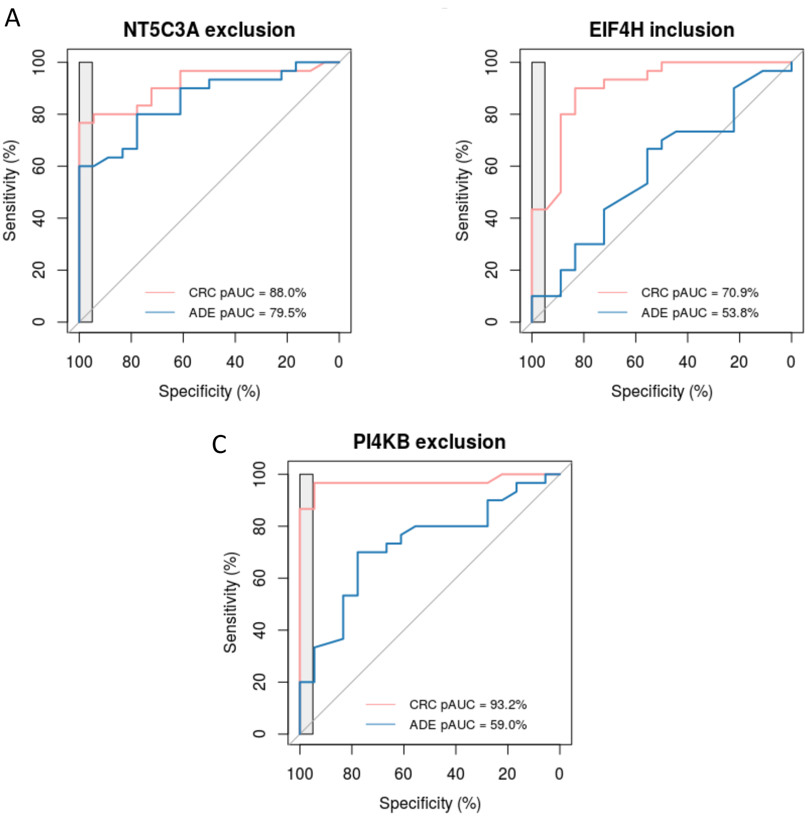


Figure 5. Performance of isoform inclusion/exclusion RNA levels as candidate biomarkers for CRC. Evaluation of ROC curves for NT5C3A RNA exclusion level (A), EIF4H inclusion level (B) and PI4KB exclusion level (C) for CRCs and adenomas (ADE) compared to normal colon samples. Partial AUCs (pAUC) were calculated at the specificity range of 95-100%.

Discussion

We aimed to characterize alternative splicing changes associated with colorectal cancer development, and to identify protein isoforms as novel candidate biomarkers for early detection of colorectal cancer. Our study showed that the transition from normal epithelium to adenoma and colorectal cancer is accompanied by changes in ratios of alternative spliced RNA variants, from which at least 30% also were observed at the protein level, proving their translation into possibly functional proteins. Differential RNA splice variants were validated in an independent mRNA sequencing dataset of colorectal cancers and adenomas. Additionally, differential splicing of NT5C3A, EIF4H and PI4KB between CRCs and adenomas was validated by RT-qPCR. Our study revealed that alternatively spliced variants provide a source of candidate biomarkers for early detection of CRC.

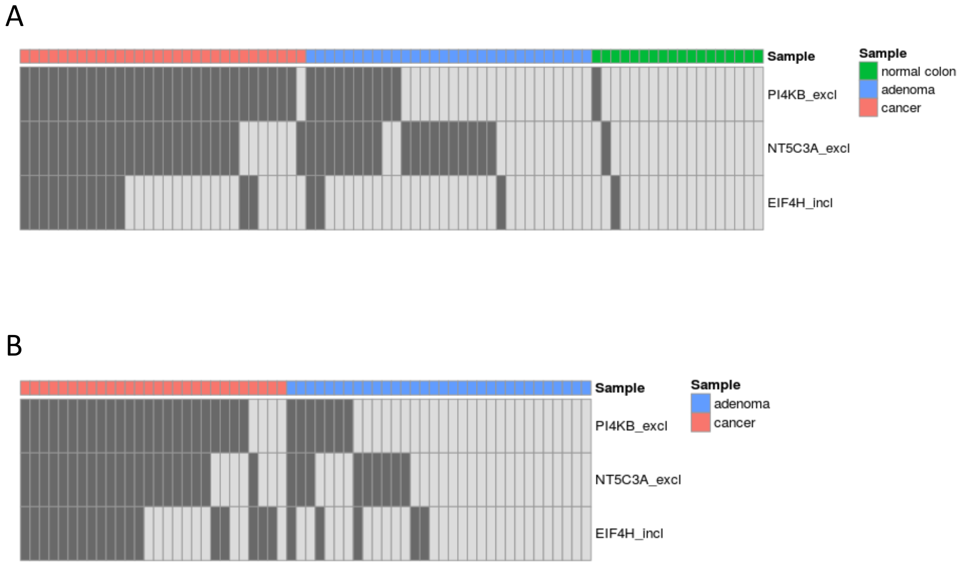


Figure 6. Distribution of candidate biomarker isoforms among CRC, adenoma and normal colon samples. Samples expressing RNA exclusion level of the isoforms PI4KB, NT5C3A and RNA inclusion level of the EIF4H isoform above thresholds defined at 95% specificity (threshold, = 0.746, 0.348 and 0.093, respectively) were marked with dark grey in the discovery (A) and the validation series (B). Samples were grouped by their type: colorectal cancer, adenoma and normal colon.

The classical approach to identify cancer-specific alternative splicing events has been to compare cancer samples to normal samples or healthy controls, which has been performed extensively e.g. for The Cancer Genome Atlas dataset[15, 23-25]. The novelty of the present study is two-fold: firstly, the incorporation of colorectal adenomas in a comparative splicing analysis of CRCs, adenomas and normal colon samples; and secondly, analysis of both RNA data and in-depth proteome data. This approach revealed alternative splicing switches to occur at each step of colorectal tumor development, showing four possible patterns of alternative splicing; early events occurring at the transition from normal to an adenoma (Group 1), gradient events (Group 2), adenoma-specific splice variants (Group 3) and cancer-specific events (Group 4; Figure 2A). In general, functional annotation of alternative splice variants is challenging as the available databases, like MSigDB[20, 26], do not characterize qualitative differences in transcripts and are mostly derived from gene expression profiling experiments. Yet, inclusion and exclusion variants of the same gene may differ in function, translation, and localization in the cell, or alternatively, differ only in the protein sequence without any functional consequences[15, 27]. We performed overrepresentation analysis of parental genes from each splicing pattern groups in the hallmark gene sets. “Mitotic spindle” was enriched in the neoplasia-specific and adenoma-specific splicing pattern sets indicating splicing regulation of the genes involved in cell division, a process involved in adenoma formation. Other processes affected by alternative splicing in adenomas and cancers were related to interactions of epithelial cells with tumor microenvironment ("epithelial-mesenchymal transition"), but also "apoptosis" and "oxidative phosphorylation", which are also associated with tumor development[28, 29]. No significant enrichment was identified for the gradient group, possibly due to its small size. The cancer-specific spliced genes were enriched in the cell-cycle related "E2F targets" and "DNA repair" biological processes, which are associated with malignant transformation as well. Additionally, "peroxisome"-related genes were enriched in the cancer-specific splicing group, which may indicate deregulation of fatty-acid or oxygen metabolism[30]. A previous study reporting on gene expression changes in adenoma-to-carcinoma progression observed enrichment of gene sets involved in cell cycle and chromosome binding and segregation in cancers compared to adenomas and conversely fatty acid metabolism enrichment in adenomas compared to cancers[18]. Our study indicates that in the adenoma-to-carcinoma progression these processes are not only regulated by quantitative changes in gene expression but also qualitative changed in the splicing of these genes. However, interpretation of these results remains challenging as only the annotation of the parental genes, and not splicing events, was considered in this analysis.

In this study isoforms were identified that are described in literature to play a role in colorectal cancer, namely isoforms of RAC1, KRAS and CTTN[10, 12, 13]. Interestingly, these three isoforms were highly expressed already in the non-malignant adenomas and belonged to Group 1. As adenomas already display alterations in cellular processes like e.g. cell proliferation or apoptosis, it is not a surprise that the anti-apoptotic Rac1b is highly expressed in these lesions. On the other hand, invasion is a feature restricted to cancer. Therefore, the high expression of the pro-invasive CTTN splice variant in pre-malignant adenomas was an unexpected finding. In the context of biomarker discovery for colorectal cancer detection, it is crucial to select protein isoforms that are not yet abundant in any of the adenomas, as this could lead to low specificity of the potential CRC screening test and subsequently to overdiagnosis and overtreatment[31]. Moreover, adenomas that do express cancer-specific isoforms should be further investigated for their risk of progressing to malignancy. Therefore, the knowledge of alternative splicing in the adenomas is vital for the candidate biomarker selection. We performed global validation of alternative splicing in an independent series of colorectal adenomas and cancers, as the focus of this study was on the splicing switch accompanying malignant transformation, and observed significant overlap for each type of splicing events. As different library preparation and sequencing depths were applied in the discovery series (total RNA, 64 million reads on average per sample) and the validation series (mRNA, 32 million reads on average per sample), the higher number of events identified in the discovery series was expected.

We focused on the gradient and cancer-specific splicing pattern groups to select candidate biomarkers for CRC detection. Alternative splicing of NT5C3A and

EIF4H were identified in the gradient group, thus differential between all three comparisons; cancer *versus* adenoma, cancer *versus* normal colon and adenoma *versus* normal colon, while alternative splicing of PI4KB was identified in the cancer-specific group, thus differential between cancer and normal colon as well as between cancer and adenoma. Differential splicing of all three isoforms NT5C3A, EIF4H and PI4KB between cancers and adenomas was validated by RT-qPCR in the discovery series and by mRNA-seq in the validation series.

Based on the RNA exclusion or inclusion level we evaluated the biomarker potential of these three isoforms in identification of CRCs and adenomas in the discovery and the independent validation series. EIF4H inclusion level did not perform well in identification of CRCs, with sensitivity of only 43% at the specificity level of 94% in the discovery series and 64% in the validation series and therefore does not hold promise as a candidate biomarker for CRC. On the other hand, PI4KB exclusion level was very sensitive in terms of identification of CRC with over 90% in the discovery series and 86% in the validation series, while the sensitivity for adenomas was 33% and 22%, respectively. Unfortunately, as only one isoform-specific peptide was identified for the PI4KB exclusion variant in the proteomics data, the quantitative analysis on protein level was not feasible. Therefore, there is not yet enough evidence to verify differential expression of PI4KB protein isoforms between cancers and the benign adenomas or normal colon samples. Based on the NT5C3A RNA exclusion level, 80% and 73% of CRCs and 60% and 28% of adenomas could be identified at the specificity level of 94%, in the discovery and validation series, respectively. As all adenomas analyzed were in fact advanced adenomas, the high sensitivity for these lesions is promising. This analysis indicates the potential of NT5C3A isoforms for the follow up biomarker research.

NT5C3A is a nucleotidase that catalyzes the dephosphorylation of pyrimidine and uridine 5’ monophosphates[32], and their analogs such as gemcitabine and cytosine arabinoside that are used to treat cancer[33]. In CRC, loss of NT5C3A DNA copy number and decrease in NT5C3A expression have also been shown to be associated with reduced production of active metabolites of 5-FU, another pyrimidine analog[34]. Mutations in NT5C3A causing its deficiency result in hemolytic anemia[33]. NT5C3A expresses four alternatively spliced isoforms, two of which were identified in this study; however, the functional impact of these two isoforms is still unknown. According to protein domain annotation by Exon ontology, the exclusion variant-specific sequence is predicted to carry a retention signal for endoplasmic reticulum and to form a transmembrane helix, while the inclusion-specific sequence constitutes an intrinsically unstructured polypeptide region[27]. Additionally, the exclusion variant, and not the inclusion variant, is annotated to contain a haloacid dehydrogenase-like domain[27], which is characteristic for phosphatases, including nucleotidases. This indicates that these isoforms may differ in their cellular localization and catalytic function. The exact role of NT5C3A inclusion and exclusion variants in colorectal tumor development should be further evaluated by functional studies.

To be able to use RNA splice variants in a stool-based test, the isoforms should be expressed in the protein form to be detectable in a similar assay that is currently used, i.e. an antibody-based assay like FIT. Due to the prominent quantitative differences on both RNA and protein level, NT5C3A isoforms are considered highly promising candidate protein biomarkers for colorectal cancer. However, the performance of NT5C3A isoforms still needs to be evaluated in stool samples. Due to the fact that the alternative start codon is incorporated in the exclusion isoform and subsequently the N-terminal part of the protein differs between the isoforms, generating isoform-specific antibodies targeting the N-terminal part of the protein for further validation may be feasible.

The modest number of protein isoforms identified in comparison to the RNA results, and limited concordance between RNA and protein expression, were in line with results described by us and others previously[19, 35, 36]. Briefly, these differences may be caused by biology, e.g. not all the transcripts are translated into proteins, or by technical challenges in proteomics, like stochastic samples or limitations of trypsin digestion patterns. Moreover, given that most of the isoforms higher expressed in CRC were exclusion variants, identification of isoform-specific peptides for these events was less probable than for inclusion variants. Meanwhile, the concordance of expression of splice variants on both RNA and protein level, as observed for NT5C3A, enables prioritization of the isoforms as candidate biomarkers for further studies. This study presents a protein biomarker discovery in a clinically relevant set of samples that is based on tumor-specific alternative splicing. NT5C3A, among others, was differentially spliced along colorectal tumor progression, and translated into differentially expressed protein isoforms. Therefore, NT5C3A is a promising candidate biomarker for early detection of colorectal cancer that warrants further evaluation.

Acknowledgements

This study was performed within the framework of the CTMM (Center for Translational Molecular Medicine) research program, project NGS-ProToCol (03O-402) and was supported by Dutch Cancer Society (KWF Kankerbestrijding, project number 2013-6025) and SU2C-DCS International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415, MEDOCC). Stand Up To Cancer is a programme of the Entertainment Industry Foundation administered by the American Association for Cancer Research. We thank ServiceXS (GenomeScan, Leiden) for the RNA sequencing. Cancer Center Amsterdam is acknowledged for support of the mass spectrometry infrastructure. The authors would like to thank Protein facility at the Netherlands Cancer Institute, in particular John de Widt for his help in hybridoma culture and antibody production. The authors would also like to thank Dr. Jaco C. Knol for critical reading of the manuscript.

REFERENCES

1. Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin 2017;67:177-193.
2. Vleugels JL, van Lanschot MC, Dekker E. Colorectal cancer screening by colonoscopy: putting it into perspective. Dig Endosc 2016;28:250-9.
3. U. S. Preventive Services Task Force, Bibbins-Domingo K, Grossman DC, et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2016;315:2564-2575.
4. Lee JK, Liles EG, Bent S, et al. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Ann Intern Med 2014;160:171.
5. de Wijkerslooth TR, Stoop EM, Bossuyt PM, et al. Immunochemical fecal occult blood testing is equally sensitive for proximal and distal advanced neoplasia. Am J Gastroenterol 2012;107:1570-8.
6. Shinya H, Wolff WI. Morphology, anatomic distribution and cancer potential of colonic polyps. Ann Surg 1979;190:679-83.
7. van Lanschot MCJ, Bosch LJW, de Wit M, et al. Early detection: the impact of genomics. Virchows Arch 2017;471:165-173.
8. Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene 2014;33:5311-8.
9. Sveen A, Kilpinen S, Ruusulehto A, et al. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene 2016;35:2413-27.
10. Esufali S, Charames GS, Pethe VV, et al. Activation of tumor-specific splice variant Rac1b by dishevelled promotes canonical Wnt signaling and decreased adhesion of colorectal cancer cells. Cancer Res 2007;67:2469-79.
11. Singh A, Karnoub AE, Palmby TR, et al. Rac1b, a tumor associated, constitutively active Rac1 splice variant, promotes cellular transformation. Oncogene 2004;23:9369-80.
12. Matos P, Jordan P. Increased Rac1b expression sustains colorectal tumor cell survival. Mol Cancer Res 2008;6:1178-84.
13. Abubaker J, Bavi P, Al-Haqawi W, et al. Prognostic significance of alterations in KRAS isoforms KRAS-4A/4B and KRAS mutations in colorectal carcinoma. J Pathol 2009;219:435-45.
14. Wang ZN, Liu D, Yin B, et al. High expression of PTBP1 promote invasion of colorectal cancer by alternative splicing of cortactin. Oncotarget 2017;8:36185-36202.
15. Climente-Gonzalez H, Porta-Pardo E, Godzik A, et al. The Functional Impact of Alternative Splicing in Cancer. Cell Rep 2017;20:2215-2226.
16. Komor MA, Bosch LJ, Bounova G, et al. Consensus molecular subtypes classification of colorectal adenomas. J Pathol 2018.
17. NGS-ProToCol. Next Generation Sequencing from Prostate to Colorectal Cancer - Center for Translational Molecular Medicine (2014-2015) http://www.ctmm.nl/en/projecten/translational-research-it-trait/ngs-protocol.
18. Carvalho B, Postma C, Mongera S, et al. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. Gut 2009;58:79-89.
19. Komor MA, Pham TV, Hiemstra AC, et al. Identification of Differentially Expressed Splice Variants by the Proteogenomic Pipeline Splicify. Mol Cell Proteomics 2017;16:1850-1863.
20. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems 2015;1:417-425.
21. Sillars-Hardebol AH, Carvalho B, Tijssen M, et al. TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. Gut 2012;61:1568-

75.
22. 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.
23. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-337.
24. Kahles A, Lehmann KV, Toussaint NC, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell 2018;34:211-224 e6.
25. Snezhkina AV, Krasnov GS, Zaretsky AR, et al. Differential expression of alternatively spliced transcripts related to energy metabolism in colorectal cancer. BMC Genomics 2016;17:1011.
26. Liberzon A, Subramanian A, Pinchback R, et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739-1740.
27. Tranchevent LC, Aube F, Dulaurier L, et al. Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res 2017;27:1087-1097.
28. Martin C, Connelly A, Keku TO, et al. Nonsteroidal anti-inflammatory drugs, apoptosis, and colorectal adenomas. Gastroenterology 2002;123:1770-7.
29. Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. Cell 2008;134:703-7.
30. Dahabieh MS, Di Pietro E, Jangal M, et al. Peroxisomes and cancer: The role of a metabolic specialist in a disease of aberrant metabolism. Biochim Biophys Acta Rev Cancer 2018;1870:103-121.
31. Kalager M, Wieszczy P, Lansdorp-Vogelaar I, et al. Overdiagnosis in Colorectal Cancer Screening: Time to Acknowledge a Blind Spot. Gastroenterology 2018;155:592-595.
32. Marinaki AM, Escuredo E, Duley JA, et al. Genetic basis of hemolytic anemia caused by pyrimidine 5’ nucleotidase deficiency. Blood 2001;97:3327-32.
33. Aksoy P, Zhu MJ, Kalari KR, et al. Cytosolic 5’-nucleotidase III (NT5C3): gene sequence variation and functional genomics. Pharmacogenet Genomics 2009;19:567-76.
34. Tong M, Zheng W, Li H, et al. Multi-omics landscapes of colorectal cancer subtypes discriminated by an individualized prognostic signature for 5-fluorouracil-based chemotherapy. Oncogenesis 2016;5:e242.
35. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513:382-7.
36. Wang X, Codreanu SG, Wen B, et al. Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. Mol Cell Proteomics 2018;17:422-430.

APPENDIX

Proof of concept: generation of isoform-specific antibodies

Based on the global proteogenomic analysis of differential splicing accompanying colorectal tumor progression, NT5C3A isoforms were selected as candidate biomarkers for follow up validation. To this end, mouse monoclonal antibodies were generated against NT5C3A inclusion protein isoform (clone 3D1), exclusion protein isoform (clones 1F8 and 4B4) and general protein sequence occurring in both isoforms (clone 1C7; Supplementary Figure 5; see Materials and Methods for Antibody generation). The antibodies were screened using Human Proteome Microarray[37] to analyze their specificity (Supplementary Figure 6). The general antibody 1C7 bound with the highest specificity to its target; NT5C3A. The inclusion-specific 3D1 antibody bound to six other proteins with a higher specificity than to the NT5C3A isoforms, while the exclusion-specific antibodies 4B4 and 1F8 bound as the best and the second best to the NT5C3A exclusion-variant protein, respectively. No overlap was observed between off-target binding of the antibodies targeting different sequence, implying that NT5C3A could be detected with high specificity in an ELISA assay using combination of one general and one isoform-specific antibody. Next, we tested the binding of the antibodies to their targets using Western blotting on purified exclusion and inclusion protein isoforms (Supplementary Figure 7; see Western Blotting for Materials and Methods). Exclusion protein isoform, due to the alternative start codon, has a longer N-terminal sequence and as a result is larger (~38 kDa) than the inclusion protein isoform (~34 kDa). Isoform-specific antibodies 3D1 and 1F8 showed specific binding to the inclusion and exclusion proteins, respectively, while the general antibody 1C7 bound to both isoforms (Supplementary Figure 7A). Exclusion-specific antibody 4B4 bound to both isoforms but displayed a stronger affinity for the exclusion variant. Next, we tested binding affinity of the antibodies to their targets using Surface Plasmon Resonance (Supplementary Figure 7 B-E; see Binding affinity using Surface Plasmon Resonance for Materials and Methods). The experiment confirmed binding of antibody 1C7 to both NT5C3A isoforms and binding of 3D1 to the inclusion-specific isoform only. The exclusion-specific 4B4 antibody bound only to the exclusion protein isoform, while exclusion-specific 1F8 antibody did not bind to any of the proteins. The discordance between 4B4 and 1F8 binding measured by Western Blotting and Surface Plasmon Resonance may be due to the denatured and native form of the protein, respectively. Next, we examined epitope specificity of the antibodies with the same methodology (Supplementary Figure 8) to evaluate if the antibodies can be used in an ELISA-like immunoassay. This time 1C7 was immobilized on the chip and coupled to each of the full-length protein isoforms. Binding of each antibody to the NT5C3A protein isoforms was measured. From all the combinations of 1C7 with either 4B4, 1F8 or 3D1 measured in the experiment, only the 1C7-3D1 combination worked on the inclusion protein isoform (Supplementary Figure 8A), indicating their utility in an ELISA-like assay, while no combination worked on the exclusion protein isoform (Supplementary

Figure 8B).

To this end, we have generated promising antibodies to detect NT5C3A isoforms with the aim of identifying CRCs and clinically relevant adenomas in a population-wide screening setting. As current population-wide screening programs are based on stool testing, from a logistic perspective implementation of a novel biomarker in the screening program is most feasible using stool-based testing. However, it is possible that NT5C3A may be also detectable in other body fluids like e.g. blood. Therefore, first it is crucial to evaluate if NT5C3A isoforms can be identified in stool or other body fluids. Ultimately, NT5C3A isoform detection should be performed in an antibody-based assay like ELISA where NT5C3A protein would be captured with the general antibody and separate isoforms quantified with the isoform-specific antibodies. In such a way, relative expression of isoforms could be quantified independently of the stool/sample composition. Once built, such an assay should be tested on stool and FIT samples of healthy individuals as well as individuals with colorectal adenomas and cancers to examine the diagnostic performance of NT5C3A in a screening setting.

MATERIALS AND METHODS

ANTIBODY GENERATION

Novel mouse monoclonal antibodies were generated against NT5C3A isoforms (inclusion and exclusion variants) as well as common NT5C3A sequence by CDI Laboratories (CDI-lab, USA) according to the CDI's Fast-Mab® workflow[37]. Three epitopes were selected from the NT5C3A protein sequence: inclusion-specific (MTNQESAVHVKMMPE), exclusion-specific (TKIIEMMPEFQKSSVR) and general (DGALRNTEYFNQLKDN). Antibodies were tested on the HuProt™ Human Proteome Microarray v3.0 as described previously[37]. Full-length purified proteins for both inclusion and exclusion variants were produced and spotted on the HuProt array to test the specificity of the antibodies against the NT5C3A isoforms.

WESTERN BLOTTING

Approximately 200 ng of purified protein samples were separated by gel electrophoresis using 10% SDS-polyacrylamide gel electrophoresis gels and transferred to polyvinylidene fluoride membranes (Bio-rad, Hercules, USA). Membranes were blocked with PBS containing 0.05% Tween-20 (Bio-rad, Hercules, USA) and 5% w/v dry milk powder, incubated with a 1/5000, 1/5000, 1/5000, 1/10 000 dilution of the 1C7, 3D1, 4B4 or 1F8 antibodies, respectively, in blocking solution overnight at 4°C, washed with PBS containing 0.05% Tween-20, incubated with horseradish peroxidase (HRP)-conjugated secondary antibody at room temperature for 30 minutes, and developed using ECL-Plus (GE Healthcare Amersham, The Netherlands).

BINDING AFFINITY USING SURFACE PLASMON RESONANCE

Interaction between the full length purified protein isoforms and the antibodies (clones 1C7, 4B4 and 3D1) was measured using Biacore T200 (GE Healthcare, USA). All the experiments were performed in PBS buffer with addition of 0.05% Tween. Protein G sensor chip (GE Healthcare, USA) was used for direct immobilization of the antibodies. Flow cell one was kept empty for blank subtraction, general NT5C3A antibody (clone 2A12, data not shown) was immobilized on the surface of flow cell two for positive control and two additional antibodies were immobilised on the surface of flow cells three and four subsequently for each experiment. The proteins were injected over all four flow cells in five increasing concentrations (2 µM, 4 µM, 8 µM, 16 µM and 32 µM) using single cycle kinetics protocol. Surface of the chip was regenerated with 10 mM Glycine-HCl pH 1.5 buffer after each experiment. Alignment and initial analysis of the blank subtracted data was done using Biacore T200 Evaluation software 3.0 (GE Healthcare, USA). Next GraphPad Prism 7.02 (GraphPad Software Inc., USA) was used to calculate the ratio of bound protein to immobilized antibody and to generate the final figures. Ratio was calculated using flowing equation:

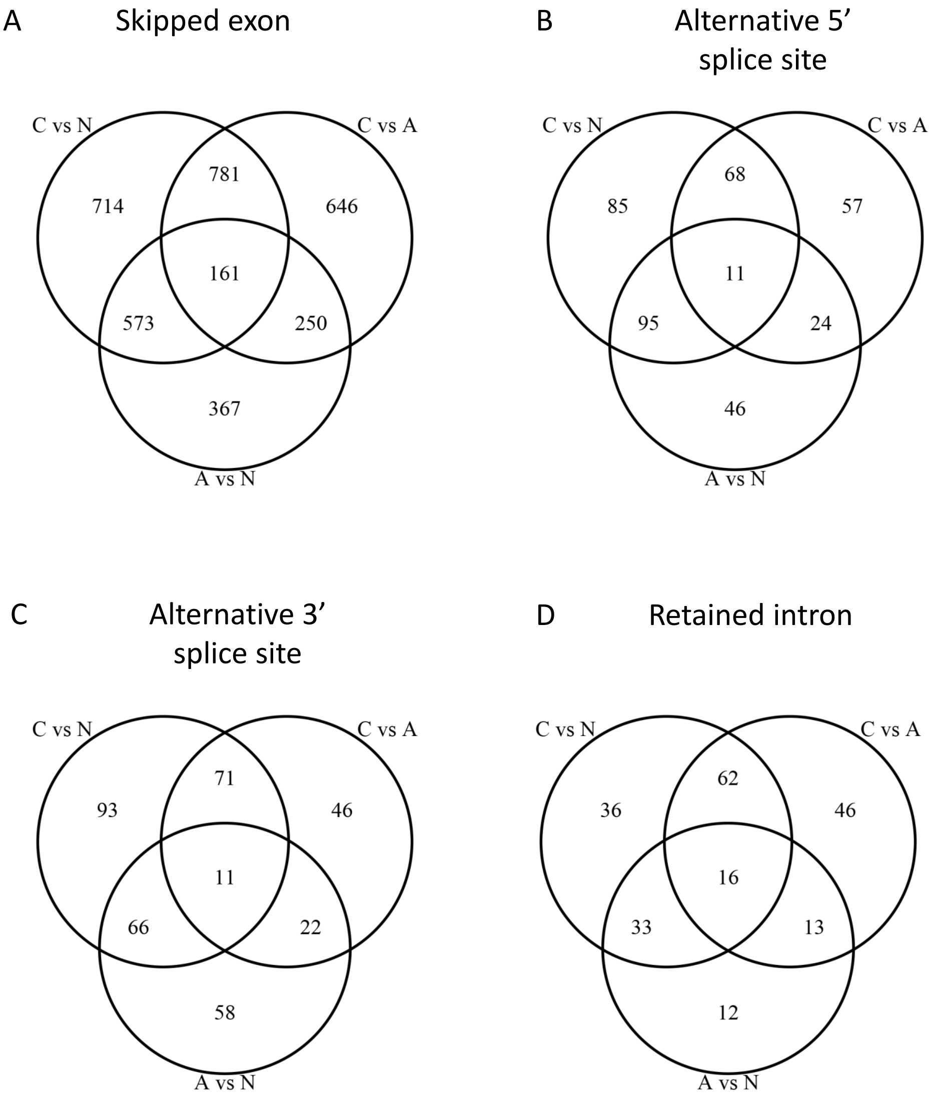
$$Ratio = \frac{RU_B \times MW_I}{RU_I \times MW_B}$$

Where:

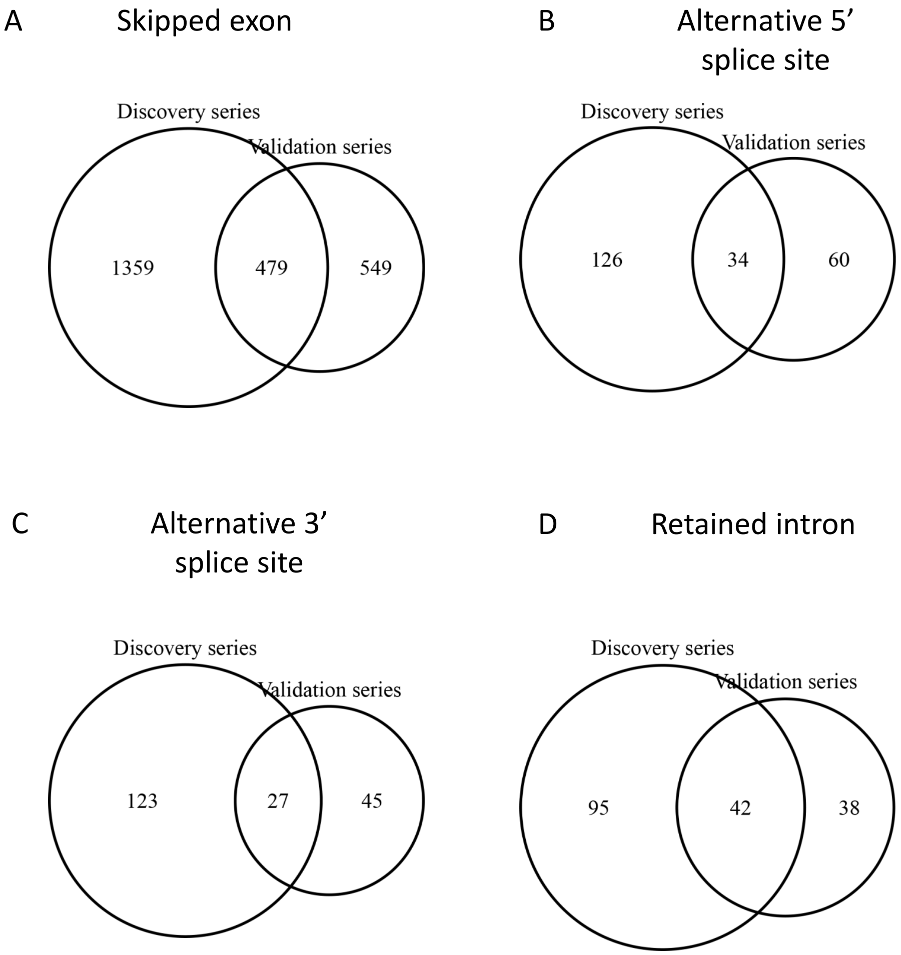
RU_B – binding response protein in response units (RU)
RU_I – total immobilization of the antibody in RU
MW_B – molecular weight of the protein
MW_I – molecular weight of the antibody

Next, Biacore T200 (GE Healthcare, USA) was used to study epitope specificity of the antibodies. All the experiments were performed in PBS buffer with addition of 0.05% Tween. CM5 sensor chip (GE Healthcare) was used for amine coupling immobilization of 1C7. Flow cell one was blocked directly with ethanolamine after activation for blank subtraction. Full-length purified protein isoforms were coupled to the separate flow cells with immobilized specific antibody (1C7). After the protein coupling all the antibodies were injected in separate cycles over the chip at single concentration of 0.1 mg/ml. Alignment and initial analysis of the blank subtracted data was done using Biacore T200 Evaluation software 3.0 (GE Healthcare, USA). GraphPad Prism 7.02 (GraphPad Software Inc., USA) was used to generate the final figures.

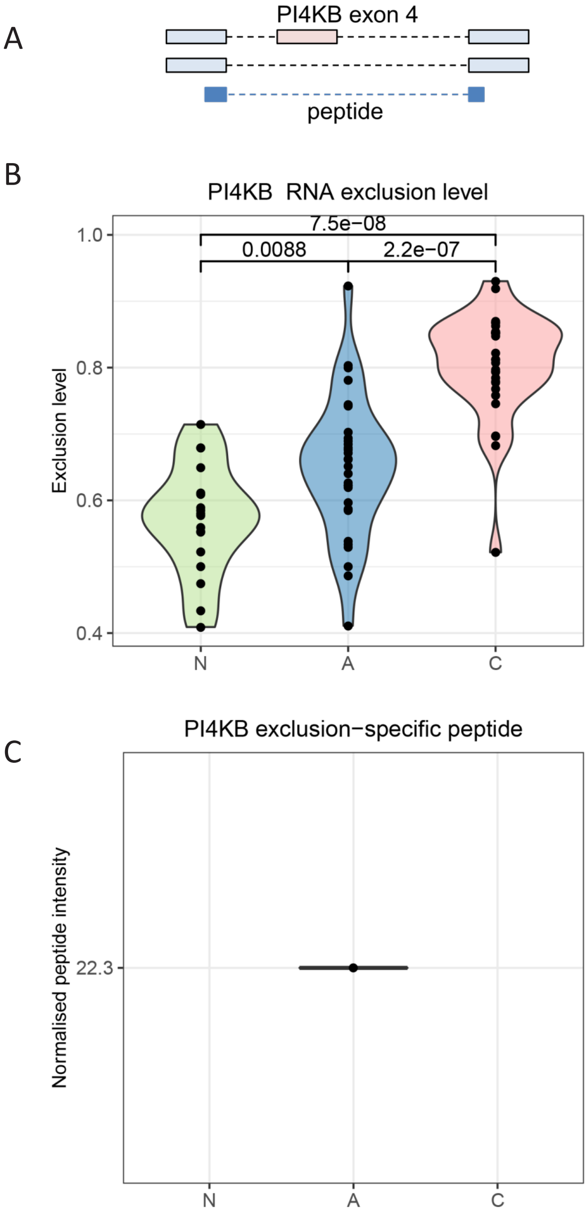
Supplementary Figures



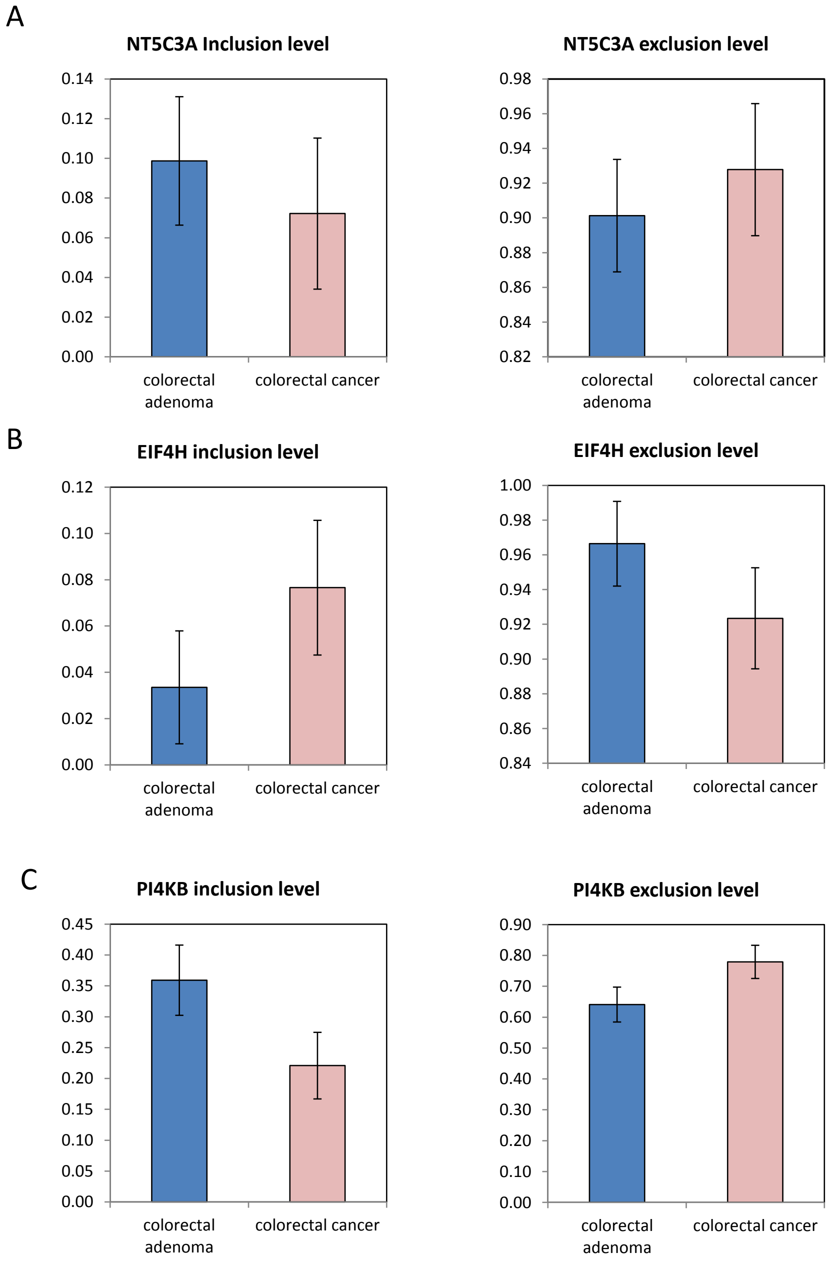
Supplementary Figure 1. Overlap analysis of the alternatively spliced events identified in the comparisons: colorectal cancer versus normal adjacent colon (C vs N), colorectal cancer versus adenoma (C vs A) and colorectal adenoma versus normal adjacent colon (A vs N). Overlap was calculated for skipped exon (A), alternative 5' splice site (B), alternative 3' splice site (C) and retained intron (D).



Supplementary Figure 2. Overlap analysis of the alternatively spliced events identified in the comparison of colorectal cancers and colorectal adenomas in the discovery series and the validation series. Overlap is presented for skipped exon (A), alternative 5’ splice site (B), alternative 3’ splice site (C) and retained intron (D).



Supplementary Figure 3. A. Schematic overview of skipped exon event for PI4KB (exon 4). For this event one exclusion-specific peptide was identified. B. Quantitative representation of PI4KB isoform in normal adjacent colon, colorectal adenoma and cancer samples on RNA level. Exclusion level was calculated by number of exclusion-specific RNA reads divided by the sum of inclusion- and exclusion-specific RNA reads. P-values were obtained with the use of Mann-Whitney test. C. Quantitative representation of exclusion specific peptide in normal adjacent colon, colorectal adenoma and cancer samples. Normalized peptide intensity was plotted. As the peptide was identified only in one sample, quantitative analysis was not feasible.

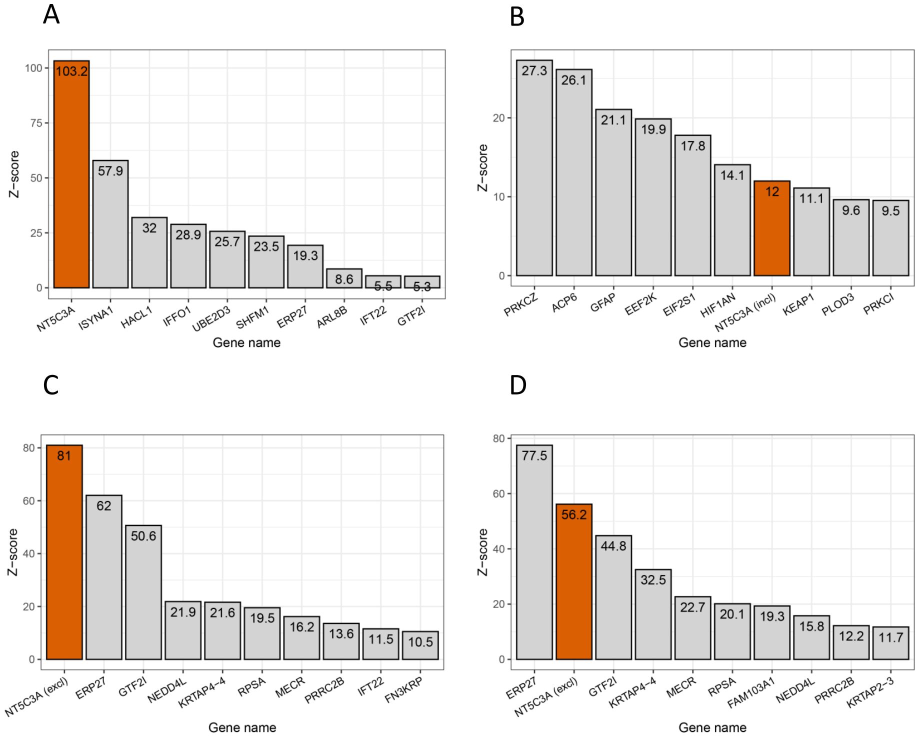


Supplementary Figure 4. RT-qPCR validation of alternative splicing of NT5C3A (A), EIF4H (B) and PI4KB (C) in colorectal adenoma and cancers from the discovery series. Quantification of the expression of inclusion and exclusion variants was performed with RT-qPCR. Inclusion (exclusion) level was obtained by division of inclusion (exclusion) expression by the sum of inclusion and exclusion expression. Median values of the inclusion (left) and exclusion levels (right) are plotted with standard deviations of 30 colorectal adenomas and 30 CRCs.

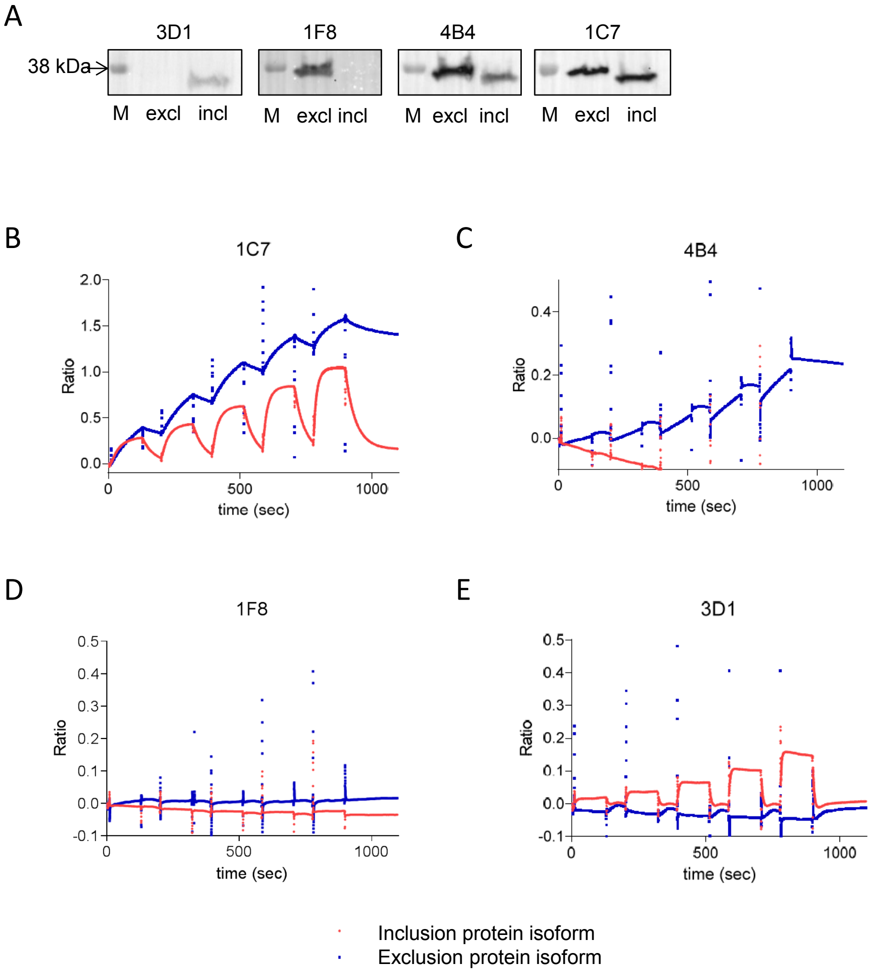
```
> Exclusion isoform
MRAPSMDRAAVARVGAVASASVCALVAGVVLAQYIFTLKRKTGRKTKIIEM/MPEFQKSSV
RIKNPTRVEEIICGLIKGGAAKLQIITDFDMTLSRFSYKGKRCPTCHNIIDNCKLVTDEC
RKKLLQLKEKYYAIEVDPVLTVEEKYPYMVEWYTKSHGLLVQQALPKAKLKEIVAESDVM
LKEGYENFFDKLQQHSIPVFIFSAGIGDVLEEVIRQAGVYHPNVKVVSNFMDFDETGVLK
GFKGELIHVFNKHDGALRNTEYFNQLKDNSNIILLGDSQGDLRMADGVANVEHILKIGYL
NDRVDELLEKYMDSYDIVLVQDESLEVANSILQKIL

> Inclusion isoform
MTNQESAVHVKM/MPEFQKSSVRIKNPTRVEEIICGLIKGGAAKLQIITDFDMTLSRFSYK
GKRCPTCHNIIDNCKLVTDECRKKLLQLKEKYYAIEVDPVLTVEEKYPYMVEWYTKSHGL
LVQQALPKAKLKEIVAESDVMLKEGYENFFDKLQQHSIPVFIFSAGIGDVLEEVIRQAGV
YHPNVKVVSNFMDFDETGVLKGFKGELIHVFNKHDGALRNTEYFNQLKDNSNIILLGDSQ
GDLRMADGVANVEHILKIGYLNDRVDELLEKYMDSYDIVLVQDESLEVANSILQKIL
```

TKIIEMMPEFQKSSVR – exclusion-specific epitope (1F8, 4B4)
MTNQESAVHVKMMPE – inclusion-specific epitope (3D1)
DGALRNTEYFNQLKDN – general epitope (1C7)
// – isoform-specific exon-exon junctions

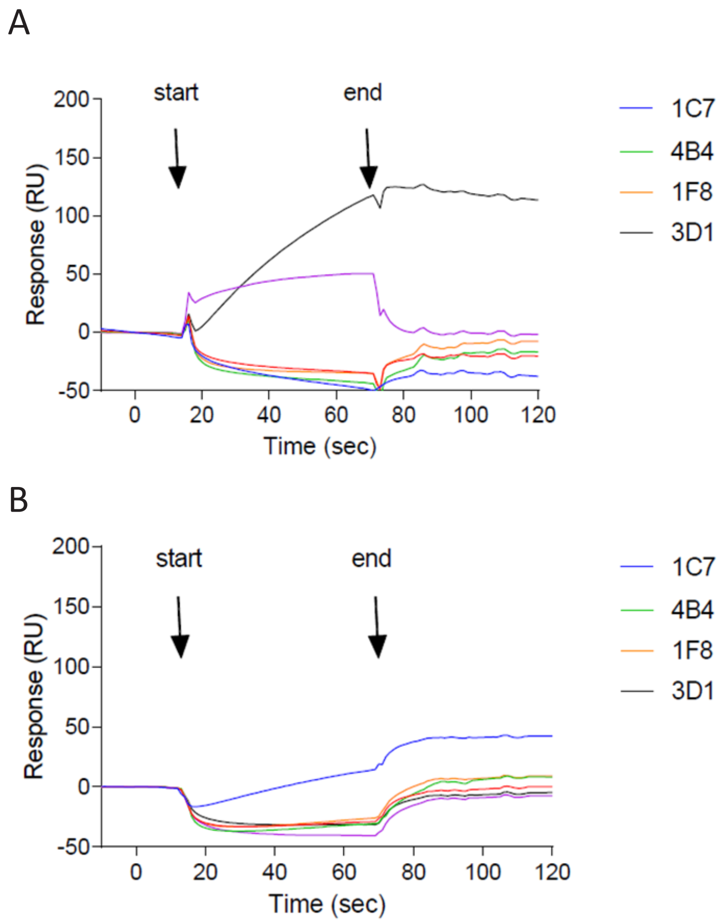
Supplementary Figure 5. Isoform-specific antibodies for NT5C3A. Epitopes selected for antibody generation portrayed on the sequences of the protein isoforms.



Supplementary Figure 6. Human Proteome Microarray screening results for antibodies 1C7 (A), 3D1 (B), 4B4 (C) and 1F8 (D). Top 10 targets of each antibody according to the highest binding affinity presented by Z-scores are presented in the figure.



Supplementary Figure 7. Purified full-length proteins for exclusion (excl) and inclusion (incl) isoforms were used to analyze binding of isoform-specific antibodies to its targets using Western Blotting (A) and Surface Plasmon Resonance (B-D). A. In the Western Blot marker is denoted by M. 3D1, an antibody raised against the inclusion variant, bound to the inclusion protein isoforms, while it did not bind to the exclusion protein. The exclusion specific antibody, 1F8, bound specifically only to the exclusion variant, while the other exclusion specific antibody 4B4 found to both isoforms. Finally, the general antibody, 1C7, bound to both proteins. In Surface Plasmon Resonance measured by Biacore the proteins were injected in five increasing concentrations (2 µM, 4 µM, 8 µM, 16 µM and 32 µM) indicated by the pattern in the plots. 1C7 (B) antibody bound to both inclusion and exclusion protein variants, 4B4 (C) only to the exclusion variant, 1F8 (D) to none of the proteins and 3D1 (E) to the inclusion variant.



Supplementary Figure 8. Epitope specificity of the antibodies. 1C7 antibody was immobilized on the surface of flow cells and full-length purified proteins, inclusion isoform (A) and exclusion isoform (B) were coupled to the antibody. The panels represent binding of the antibodies to the full-length purified protein isoforms.

SUPPLEMENTARY TABLES

Supplementary Table 1. Thresholds applied on alternatively spliced events on the RNA level to define splicing pattern groups.analysis of the alternatively spliced events identified in the comparisons: colorectal cancer versus normal adjacent colon (C vs N), colorectal cancer versus adenoma (C vs A) and colorectal adenoma versus normal adjacent colon (A vs N).

Comparison	**Group 1**		**Group2**		**Group3**		**Group4**	
	FDR	**InclusionLevel Difference**	**FDR**	**InclusionLevel Difference**	**FDR**	**InclusionLevel Difference**	**FDR**	**InclusionLevel Difference**
colorectal cancer vs normal colon (CvsN)	≤ 0.05	sign(CvsN) = sign(AvsN)	≤ 0.05	sign(CvsN) = sign(AvsN) = sign (CvsA)	-	sign(CvsA) ≠ sign(AvsN)	≤ 0.05	sign(CvsN) = sign(CvsA)
colorectal cancer vs colorectal adenoma (CvsA)	> 0.05		≤ 0.05		≤ 0.05		≤ 0.05	
colorectal adenoma vs normal colon (AvsN)	≤ 0.05		≤ 0.05		≤ 0.05		> 0.05	

Supplementary Table 2. Primers with conditions used for RT-qPCR

Gene	**Specific primer**	**Forward primer sequence**	**Reverse primer sequence**	**final primer concentration**	**annealing temperature (°C)**
GUSB	housekeeping	5’- GAAAATATGTGGTTGGAGAGCTCATT-3’	5’- CCGAGTGAAGATCCCCTTTTTA-3’	0.5	60
NT5C3A	Inclusion exon 2	5’-GTGGTGCTGGCTCAGTACAT-3’	5’-CACATGTACGGCAGACTCTTGA-3’	0.5	60
NT5C3A	Exclusion exon 2	5’-GTGGTGCTGGCTCAGTACAT-3’	5’-CTGGAATTCTGGCATCATCTCG-3’	0.5	60
EIF4H	Inclusion exon 5	5’-CGTGTGGACATTGCAGAAGG-3’	5’-TCTCGAGAGCTACCCATTCCT-3’	0.5	60
EIF4H	Exclusion exon 5	5’-CGTGTGGACATTGCAGAAGG-3’	5’-GTCATCCCTGAAGCCTCTGT-3’	0.5	60
PI4KB	Inclusion exon 4	5’-GCGCTCTAAGTCAGATGCCA-3’	5’-ATACTCTCGGTGCTGGAGGA-3’	0.5	60
PI4KB	Exclusion exon 4	5’-ATGCCACTGCCAGCATAAGT-3’	5’-GTCGAACAGGCTCATCCTCA-3’	0.5	60

Supplementary Table 9. Number of alternatively spliced events in each group. Percentages correspond to fraction of all alternative splicing events identified to be significant in at least two out of three comparisons.

	SE	**A5SS**	**A3SS**	**RI**	**All events**
Group1	554 (31%)	92 (46%)	64 (38%)	33 (27%)	743 (33%)
Group2	86 (5%)	7 (4%)	6 (4%)	10 (8%)	109 (5%)
Group3	319 (18%)	26 (13%)	25 (15%)	19 (15%)	389 (17%)
Group4	751 (43%)	66 (33%)	70 (41%)	61 (49%)	948 (42)
Other	55 (3%)	7 (4%)	5 (3%)	1 (1%)	68 (3%)
Total	1765	198	170	124	2257

Supplementary Table 12. Evaluation of the performance of NT5C3A RNA exclusion level, EIF4H RNA inclusion level and PI4KB RNA exclusion level as biomarkers for CRCs and adenomas in the study and validation series. Confusion matrices were built by application of the threshold for RNA exclusion level of 0.348, 0.093 and 0.746 for NT5C3A exclusion level, EIF4H inclusion level and PI4KB exclusion level, respectively, to obtain high specificity (94%) for CRCs and adenomas.

Study series												
Isoform:	**NT5C3A RNA exclusion level**				**EIF4H RNA inclusion level**				**PI4KB RNA exclusion level**			
	normal	cancer			normal	cancer			normal	cancer		
predicted control	17	6	specificity	94%	17	17	specificity	94%	17	1	specificity	94%
predicted case	1	24	sensitivity	80%	1	13	sensitivity	43%	1	29	sensitivity	97%
	normal	adenoma			normal	adenoma			normal	adenoma		
predicted control	17	12	specificity	94%	17	27	specificity	94%	17	20	specificity	94%
predicted case	1	18	sensitivity	60%	1	3	sensitivity	10%	1	10	sensitivity	33%
Validation series												
Isoform:	**NT5C3A RNA exclusion level**				**EIF4H RNA inclusion level**				**PI4KB RNA exclusion level**			
	cancer				cancer				cancer			
predicted control	7		sensitivity	73%	10		sensitivity	64%	4		sensitivity	86%
predicted case	21				18				24			
	adenoma				adenoma				adenoma			
predicted control	23		sensitivity	28%	27		sensitivity	16%	25		sensitivity	22%
predicted case	9				5				7			

Chapter 6

PROTEINS IN STOOL AS BIOMARKERS FOR NON-INVASIVE DETECTION OF COLORECTAL ADENOMAS WITH HIGH RISK OF PROGRESSION

Malgorzata A Komor, Linda JW Bosch, Veerle MH Coupe, Christian Rausch, Thang V Pham, Sander R Piersma, Sandra Mongera, Chris JJ Mulder, Evelien Dekker, Ernst J Kuipers, Mark A van de Wiel, Beatriz Carvalho, Remond JA Fijneman, Connie R Jimenez, Gerrit A Meijer, Meike de Wit

SUBMITTED

ABSTRACT

Screening to detect colorectal cancer (CRC) in an early or premalignant state is an effective method to reduce CRC mortality rates. Current stool-based screening tests, e.g. faecal immunochemical test (FIT), have a suboptimal sensitivity for colorectal adenomas and difficulty distinguishing adenomas at high risk of progressing to cancer from those at lower risk. We aimed to identify stool protein biomarker panels that can be used for the early detection of high-risk adenomas and CRC. Proteomics data (LC-MS/MS) were collected on stool samples from adenoma (n=71) and CRC patients (n=81) as well as controls (n=129). Colorectal adenoma tissue samples were characterized by low-coverage whole genome sequencing to determine their risk of progression based on specific DNA copy number changes. Proteomics data was used for logistic regression modelling to establish protein biomarker panels. In total, 15 of the adenomas (15.8%) were defined as high-risk of progressing to cancer. A protein panel, consisting of Hp, LAMP1, SYNE2 and ANXA6, was identified for the detection of high-risk adenomas (sensitivity of 53% at specificity of 95%). Two panels, one consisting of Hp and LRG1 and one of Hp, LRG1, RBP4 and FN1 were identified for high-risk adenomas and CRCs detection (sensitivity of 66% and 62%, respectively, at specificity of 95%). Validation of Hp as biomarker for high-risk adenomas and CRCs was performed using an antibody-based assay in FIT samples from a subset of individuals from the discovery series (n=158) and an independent validation series (n=795). The Hp protein was significantly more abundant in high-risk adenoma FIT samples compared to controls in the discovery (p-value=0.036) and the validation series (p-value=9e-5). We conclude that Hp, LAMP1, SYNE2, LRG1, RBP4, FN1 and ANXA6 may be of value as stool biomarkers for early detection of high-risk adenomas and CRCs.

INTRODUCTION

Colorectal cancer (CRC) remains a major health care problem, representing 6.1% percent of all cancers worldwide[1]. Early detection through population screening is an efficient method to reduce the burden of CRC and screening programs have been implemented in many countries[2]. Screening programs aim to detect CRC at a curable stage or when it is still at a precursor non-malignant stage (i.e. colorectal adenoma), and have been proven to reduce CRC mortality rates[3-5]. Most population screening programs use a faecal immunochemical test (FIT) as a triage test to colonoscopy[2]. In this setting all participants with a positive FIT are referred for colonoscopy, during which adenomas and early cancers can be diagnosed and removed.

The reported sensitivity of FIT depends on the study characteristics but is overall high for CRC (67-86%), and relatively low for colorectal adenomas (29-35%), leaving room for improvement[6-8]. It has been suggested that an increase in sensitivity for colorectal adenomas is the best approach to make CRC screening more cost-effective and efficient[9-11]. However, detecting all adenomas during screening is not the aim, as only approximately 5% of all adenomas are expected to develop into cancer[12]. Advanced adenomas, defined as adenomas with a size of ≥10 mm, a villous component of ≥25%, and/or high-grade dysplasia, are currently regarded as an intermediate endpoint for CRC in screening programs, since advanced adenomas are considered to carry a higher risk to develop into CRC than non-advanced adenomas[13-15]. Based on the fact that advanced adenomas are far more prevalent than CRC, not all advanced adenomas are expected to progress[12].Therefore, it is important to develop new screening tests directed at identification of those lesions with the highest risk of progression.

Cancer is caused by DNA alterations, including specific changes in DNA copy numbers. Gains of chromosomal arms 8q, 13q, and 20q, and losses of 8p, 15q, 17p, and 18q have been associated with adenoma-to-carcinoma progression (i.e. cancer associated events or CAEs) [16, 17]. Adenomas carrying two or more CAEs are considered at high risk of progression, i.e. high-risk adenomas[17]. Approximately 23%-36% of advanced adenomas and 1.7-4.8% of non-advanced adenomas were reported to be high-risk adenomas[18]. Based on the incidence of CRC, the molecularly-defined high-risk adenoma phenotype may better reflect the true progression risk than the advanced adenoma phenotype.

We have previously reported on stool protein biomarkers, which increased sensitivity compared to haemoglobin for detection of CRC and advanced adenomas[19]. In contrast to the previous study where the focus was on advanced adenomas, here a molecularly-defined intermediate endpoint was applied for biomarker discovery. In this study, we set out to further explore the same proteomics dataset for identification of protein biomarkers that are specifically suited for the detection of molecularly-defined high-risk adenomas.

MATERIALS AND METHODS

The design of the study is presented in Figure 1.

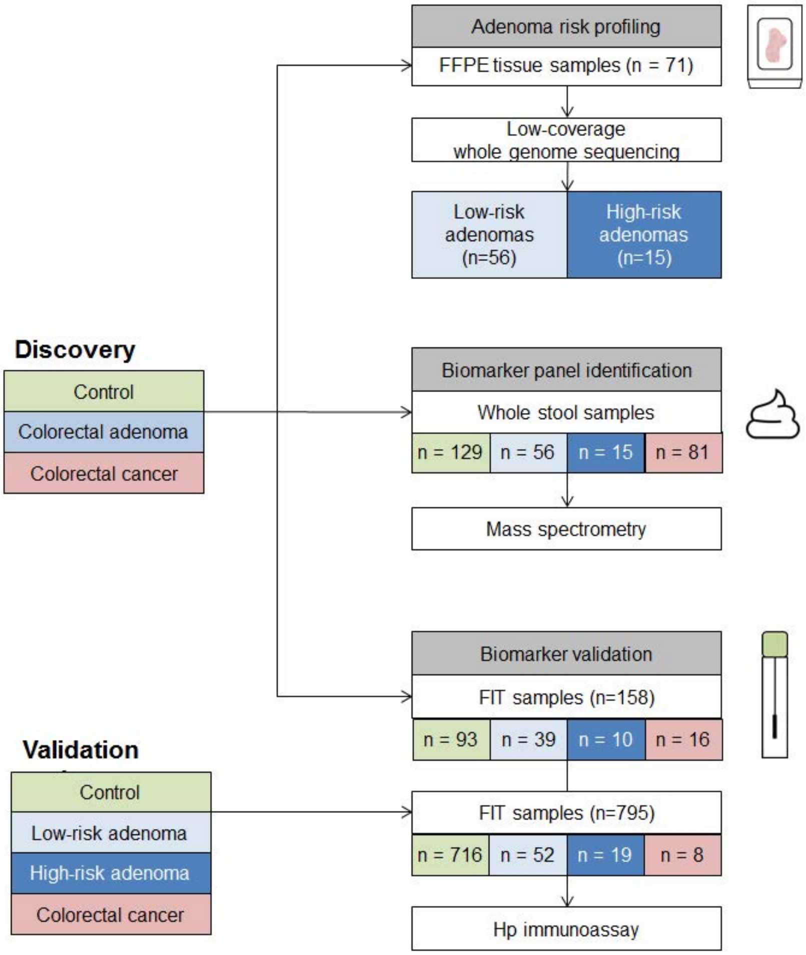


Figure 1. Overview of the design of this study. The discovery series consisted of control, colorectal adenoma and colorectal cancer (CRC) samples. FFPE tissue blocks were obtained from 71 adenoma patients and low-coverage whole genome sequencing was performed to identify DNA copy number aberrations. Fifteen high-risk adenomas were identified according to their DNA copy number profiles. Whole stool samples of individuals from the discovery series were used for mass spectrometry proteomics analysis. Proteins identified were used for biomarker panel identification for high-risk adenomas and high-risk adenomas together with CRCs. An immunoassay was applied on 158 FIT samples from the discovery series and 795 FIT samples from the validation series for biomarker validation, to evaluate quantitative difference of Hp between controls, low-risk adenomas, high-risk adenomas and CRCs.

SAMPLES

Informed consent was obtained from all subjects who provided stool and FIT samples. Collection, storage and use of patient-derived tissue and data were performed in compliance with the 'Code for Proper Secondary Use of Human Tissue in The Netherlands' Dutch Federation of Biomedical Scientific Societies[20].

Stool, tissue and FIT samples of the discovery series

For discovery, whole stool samples from 293 individuals diagnosed with CRC (n=81), advanced adenoma (n=40) or non-advanced adenoma (n=43) as most advanced lesion, and individuals without colorectal neoplasia (n=129) further referred to as "controls" were collected from a referral population that underwent colonoscopy at multiple centers in the Netherlands and Germany, between 2005 and 2012. Sample description and processing was previously described[19]. In total for 71 adenoma patients, formalin-fixed paraffin embedded (FFPE) tissue samples were available and requested from the pathology archive of the Amsterdam UMC, location VUmc, the Netherlands. In total, 95 tissue samples were retrieved as some individuals carried multiple adenomas.

From a subset of the individuals from the discovery series (n=162), FIT samples (OC-sensor, Eiken Chemical, Tokyo, Japan) were obtained prior to colonoscopy. These included patients diagnosed with CRC (n=17), high-risk adenoma (n=10) or low-risk adenomas (n=39) as most advanced lesion, and controls (n=96).

FIT samples of the validation series

Between June 2009 and July 2010, in a population-based screening study (COlonoscopy or COlonography for Screening (COCOS) trial) run in the Netherlands asymptomatic individuals were invited for primary colonoscopy screening[21, 22]. Screening participants allocated to the colonoscopy arm of the COCOS-trial were invited to collect a FIT sample (OC-sensor, Eiken Chemical, Tokyo, Japan) prior to their screening colonoscopy. FIT samples from 795 individuals diagnosed with CRC (n=8), high-risk adenomas (n=19) or low-risk adenomas (n=52) as most advanced lesion, or without colorectal neoplasia (n=716) were used for validation.

DNA copy number analysis using low-coverage whole genome sequencing

DNA was isolated from FFPE tissues with a column-based method (QIamp DNA microkit, Qiagen, Hilden, Germany) as described before[18, 23]. DNA copy number analysis (Supplementary Materials and Methods) and status for adenomas of the discovery and the validation series were reported previously[18], data are available in the European Genome and Phenome Archive (EGAS0000100295). If two or more of CAEs were present, an adenoma was classified as high-risk adenoma[17, 18]. Individuals with at least one high-risk adenoma were defined as high-risk.

LC-MS/MS data analysis

The tandem mass spectrometry (LC-MS/MS) data on the stool samples of the 293 individuals were readily available and described previously[19]. Protein identification was performed with MaxQuant[24] as described previously[19] with some adaptations (see Supplementary Materials and Methods).

PROTEIN BIOMARKER PANEL IDENTIFICATION WITH LOGISTIC REGRESSION

An overview of the data analysis approach is presented in Supplementary Figure 1. Proteins with higher abundance in cases (high-risk adenomas or high-risk adenomas and CRCs) compared to controls constituted input for selecting biomarker panels. Logistic regression analysis with Lasso regularization was used to identify biomarker panels consisting of two, three or four proteins that best distinguish cases from controls. A leave-one-out cross-validation procedure was applied to evaluate the performance of the model. Cross-validated logistic predictions were obtained. Receiver operating characteristic (ROC) analysis was used to evaluate the performance of protein panels to discriminate cases from controls by calculating partial area under the curve (pAUC) between specificity of 95%-100% and by calculating sensitivity at 95% specificity. The pAUC was compared to pAUC of haemoglobin (HBA1), p-values were obtained with the stratified bootstrap resampling of case/control labels of the individuals with 2000 permutations[25].

Haptoglobin quantification in FIT samples

FIT samples from both the discovery and validation series were analysed with an antibody-based assay (Figure 1). From the 162 FIT samples in the discovery series, four were excluded due to technical reasons (controls n=3, CRC n=1) leaving 158 samples for Hp quantification. Immunoassays for Hp employing a sandwich immunoassay format and electrochemiluminescence (ECL) detection were carried out on commercial instrumentation and multi-well plate consumables from Meso Scale Diagnostics, LLC (MSD), for more details see Supplementary Materials and Methods[26]. All samples were analysed in duplo and final analyses were performed on mean concentrations.

Fit values – correlation analysis

In the discovery series, Haemoglobin (HBA1 and HBB) and haptoglobin (Hp) protein abundance as determined by mass spectrometry were compared to FIT values in the same samples. Missing values were excluded from the analysis. Spearman correlation analysis was performed on normalized spectral counts of HBA1, HBB, Hp and FIT values, correlation coefficients (rho) and p-values were obtained.

RESULTS

CHARACTERIZATION OF CANCER ASSOCIATED EVENTS IN COLORECTAL ADENOMAS

In total 95 adenomas from 71 adenoma patients from the discovery series were available for CAE identification as was described before (Supplementary Figure 2)[18]. For a complete overview of the frequencies and the associations to adenoma histologic features see Supplementary Table 1. Two CAEs or more, indicating a higher risk of progression, were identified in 15.8% of all adenomas (n=15, further referred to as high-risk adenomas), in 36.4% (12/33) of advanced adenomas and in 4.8% (3/ 62) of non-advanced adenomas (Supplementary Table 1, Supplementary Figure 2).

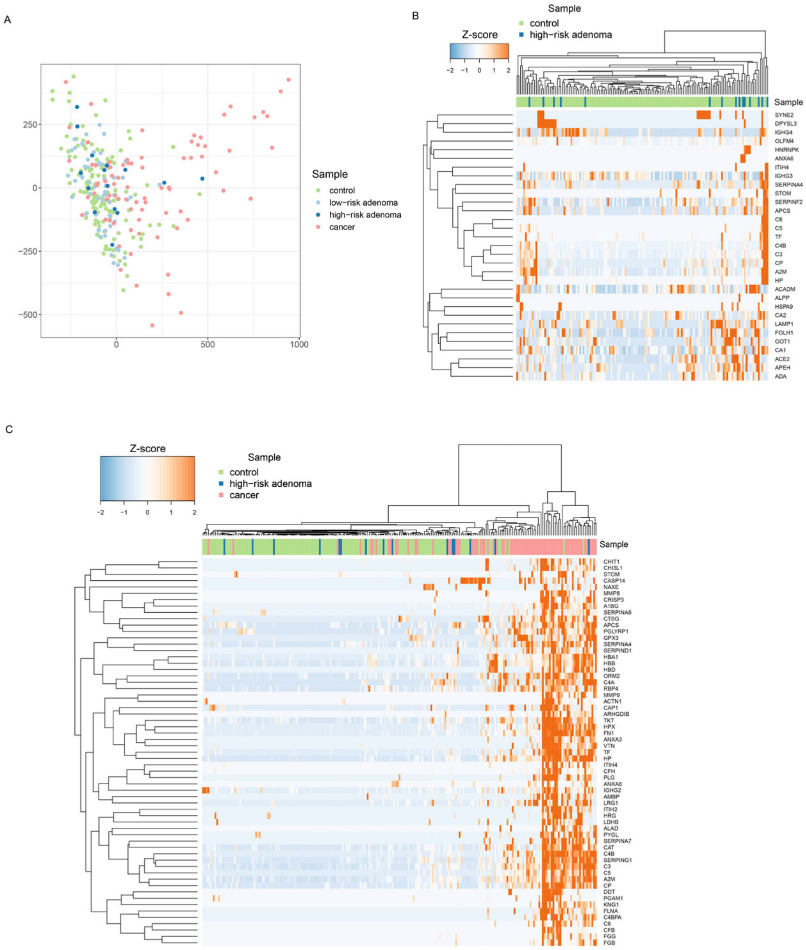


Figure 2. Proteomics profiling of human stool samples. A. Multidimensional scaling of protein expression profiles of stool samples derived from controls (n=129), individuals with low-risk adenomas (n=56), high-risk adenomas (n=15) and cancers (n=79). B. Hierarchical clustering of protein profiles of stool samples derived from high-risk adenomas and controls based on 31 proteins expressed higher in high-risk adenomas compared to the controls. C. Hierarchical clustering of protein profiles of stool samples derived from CRCs, high-risk adenomas and controls based on 61 proteins expressed higher in CRCs and high-risk adenomas compared to controls.

PROTEIN PROFILING AND SELECTION OF CANDIDATE BIOMARKERS

In the discovery series, proteomics profiling of all stool samples revealed 792 protein groups (FDR≤0.01, Supplementary Table 2). Correlation analysis was performed between FIT values obtained from a subsample of the same bowel movement and normalized spectral counts for haemoglobin, in particular for HBA1 and HBB separately. Significant positive correlations were identified for both HBA1 (rho=0.46, p-value<0.001) and HBB (rho=0.43, p-value<0.001, Supplementary Figure 3). Dimensionality reduction performed on the protein expression profiles distinguished stool samples from CRC patients from the ones with adenomas or controls (Figure 2A). To identify proteins that discriminate high-risk adenomas from controls, we performed differential protein expression analysis. This yielded 31 proteins more abundant in high-risk adenoma stool samples (log2 fold change>0 and p-value≤0.1, Figure 2B). Additionally, we have performed differential protein expression analysis to identify proteins differentiating all screen-relevant lesions, i.e. CRCs and high-risk adenomas, from controls. Application of the same threshold revealed 125 protein groups to be higher expressed in high-risk adenomas and CRCs. For further analysis, a more stringent threshold was applied (i.e. p-value≤0.05 and log2 fold change ≥2) and revealed 61 proteins more abundant in screen-relevant lesions compared to controls (Figure 2C). Significant overlap was identified between differentially expressed proteins from both analyses (p-value=1.47e^{-4}, hypergeometric test) with 13 proteins overlapping: CP, Hp, A2M, C3, C5, APCS, TF, ANXA6, C4B, C6, STOM, SERPINA4 and ITIH4.

BIOMARKER PANEL SELECTION FOR HIGH-RISK ADENOMAS

The proteomics dataset was further investigated to find biomarker panels of complementary proteins that would perform better than haemoglobin in distinguishing individuals with high-risk adenomas from controls and a combination of high-risk adenomas and CRCs from controls. Panels of two, three or four proteins were examined. To evaluate the diagnostic performance of each biomarker panel in the context of population screening, we compared its performance to haemoglobin, which is the protein currently used in CRC-screening by means of FIT. Since FIT values were not available for the whole dataset, the performance of the biomarker panel was compared to HBA1 quantified by LC-MS/MS as a substitute (for comparison to FIT see Supplementary Figure 4). The analysis was done on a partial AUC (pAUC) at the specificity level between 95%-100% and sensitivity was evaluated at 95% specificity, since high specificity is pivotal for the success of a population screening program.

First, we applied logistic regression with Lasso regularization on the 31 upregulated proteins in high-risk adenomas to identify a biomarker panel (see Supplementary Figure 1 for the data analysis overview). In the resulting regression model Hp, LAMP1, SYNE2 and ANXA6 were selected, while the models for three or two proteins were not built, as due to the Lasso regularization the coefficients for LAMP1, SYNE2 and ANXA6 shrunk to zero at the same time, meaning that the three proteins were excluded from the regression model at once. Then, the performance of the model was evaluated using leave-one-out cross-validation and an ROC analysis was used to compare to the performance of haemoglobin. In the cross-validation procedure only models based on four proteins were included (Figure 3). Despite the fact that the pAUC of the biomarker panel (pAUC=60.2%) was higher than for HBA1 (pAUC=54.5%), the difference was not significant. At the specificity level of 95% the biomarker panel could identify 8 out of 15 high-risk adenomas (sensitivity=54%, CI=[27%, 79%]), which was more than haemoglobin (sensitivity=13%, CI=[2%, 40%], see Table 1A). The markers most frequently selected in the cross-validation procedure were Hp, LAMP1, SYNE2, ANXA6, with a frequency of over 90%, indicating that these proteins have the most discriminative roles in the regression models (Figure 3B).

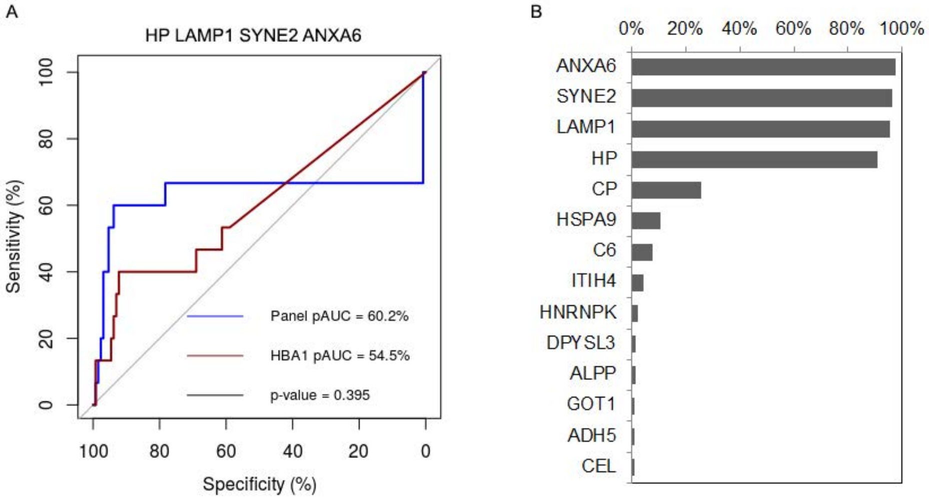


Figure 3. Biomarker panels from logistic regression analysis to identify high-risk adenomas and CRCs. A. ROC curve of the regression model using four biomarker panel (Hp, LAMP1, SYNE2 and ANXA6) to distinguish between stool samples from individuals with high-risk adenomas (n=15) and controls (n=129). ROC curve was obtained from logistic regression predictions from leave-one-out cross-validation analysis. Partial area under the curve (pAUC) was calculated for specificity of 95%-100% and compared to pAUC of haemoglobin to obtain the p-value. B. Frequency plot of biomarkers occurring in the regression models built during the cross-validation analysis to distinguish between the high-risk adenomas and controls. Four proteins were clearly selected more frequently by the Lasso regularization in the cross-validation analysis.

The model was also applied to low-risk adenomas. Here, five (9%, CI=[3%, 20%]) low-risk adenomas were classified as cases and 51 (91%) as controls, indicating that this biomarker panel has a high specificity for the identification of high-risk adenomas (see Supplementary Table 3).

Biomarker panel selection for high-risk adenomas and CRCs combined

Next, we performed the same analysis for the 61 up-regulated proteins in stool samples derived from individuals with high-risk adenomas and CRCs. The model with four protein biomarkers consisted of Hp, LRG1, RBP4 and FN1, the model with three features was not built as due to Lasso regularization the coefficients of FN1 and RBP4 shrunk to zero at the same time, and the model of two proteins consisted of Hp and LRG1. In the cross-validation procedure, the models of four and two proteins were evaluated (Figure 4). The cross-validated pAUCs of four (pAUC= 70.4%) and two (pAUC=71.1%) proteins models significantly outperformed haemoglobin (pAUC HBA1=62.7%, both p-value=0.007, Figure 4A, C). At the specificity level of 95% the four and two biomarker panels could identify 58 and 62 out of 94 cases, respectively (sensitivity=62% and 66%, CI=[51%, 72%] and [55%, 75%]), which was more than HBA1 (sensitivity=40%, CI=[30%, 51%], Table 1B). The most frequent proteins included in the four protein regression models in the cross-validation procedure were Hp, LRG1, RBP4 and FN1 with frequencies of over 90%, confirming their predictive characteristics and the stability of the model (Figure 4B). The model with two proteins always consisted of Hp and LRG1 in the cross-validation procedure, indicating their strongest predictive characteristics (Figure 4D).

The four and two protein models were also tested for identification of low-risk adenomas. The four protein panel classified 6 (11%, CI=[4%, 22%]) out of 56 low-risk adenomas as cases and 50 (89%) as controls, while the two protein panel classified 7 (13%, CI=[5%, 24%]) low-risk adenomas as cases and 49 (87%) as controls (Supplementary Table 3).

When focusing on the overlap of up-regulated proteins in both comparisons and the biomarker panels selected by Lasso regularization, Hp was the only protein present in all panels. This suggests that Hp might be a crucial component when distinguishing between high-risk adenomas and CRCs from controls.

VALIDATION OF HP EXPRESSION BY IMMUNOASSAY IN FIT SAMPLES

As Hp forms a complex with haemoglobin, we explored if the protein abundance as measured by mass spectrometry was correlated to FIT and/or haemoglobin (Supplementary Figure 3). As expected, we observed a strong correlation to HBA1 and HBB and a somewhat weaker correlation to FIT (Correlation coefficient 0.77, 0.67 and 0.55, respectively, p-value <0.001 for all comparisons). In line with this, Hp as a single marker did not outperform FIT (Supplementary Figure 5).

Nevertheless, as in the regression models Hp was consistently selected in all three markers panels, we further explored the Hp levels in two FIT cohorts. Using an immunoassay Hp quantification was successfully performed in FIT samples of a subset of individuals from the discovery series (n=158; 16 CRCs, 10 high-risk adenomas, 39 low-risk adenomas and 93 controls). A significantly higher concentration of Hp was identified in the high-risk adenoma samples compared to the controls (fold change=1.9, p-value=0.036, Figure 5A). Additionally, an independent validation series was used (Figure 5B), which consisted of 716 controls, 52 low-risk adenomas, 19 high-risk adenomas and 8 CRCs. Here a higher abundance of Hp in high-risk adenomas (fold change =15.9, p-value=$9e^{-5}$) and CRCs (fold change=42.6, p-value=$9.7e^{-5}$) compared to controls was confirmed. This confirms our findings by mass spectrometry and suggests that Hp can be applied as biomarker for high-risk adenomas and CRCs.

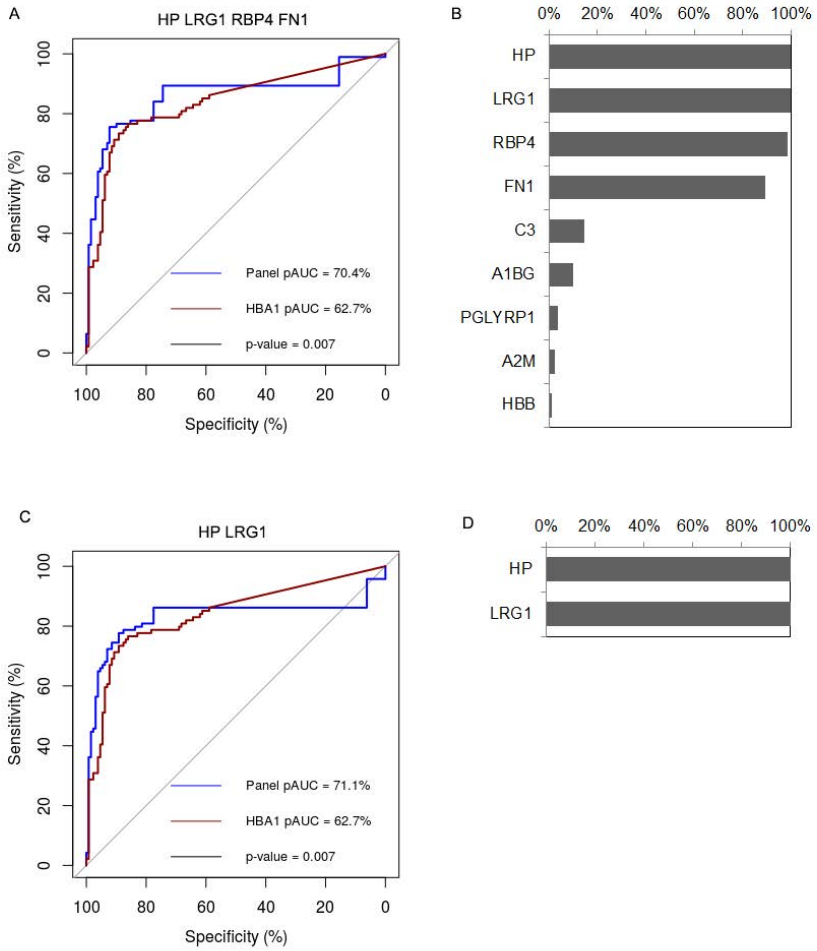


Figure 4. Biomarker panels from logistic regression analysis to identify high-risk adenomas and CRCs. A. ROC curve of the model based on the panel of four biomarkers (Hp, LRG1, RBP4 and FN1) for high-risk adenomas and CRCs (n=94) compared to controls (n=129). ROC curve was obtained from logistic regression predictions from the leave-one-out cross-validation analysis B. Frequency plot of biomarkers occurring in the regression models built during the cross-validation analysis to discriminate high-risk adenomas and CRCs from controls based on four proteins. Four proteins were clearly selected more frequently by the Lasso regularization in the cross-validation analysis. C. ROC curve of the model based on the panel of two biomarkers (Hp and LRG1) for high-risk adenomas and CRCs (n=94) compared to controls (n=129). ROC curve was obtained from logistic regression predictions from the leave-one-out cross-validation analysis D. Frequency plot of biomarkers occurring in the regression models built during the cross-validation analysis to discriminate high-risk adenomas and CRCs from controls based on two proteins. The same two proteins were consistently selected in the cross-validation analysis.

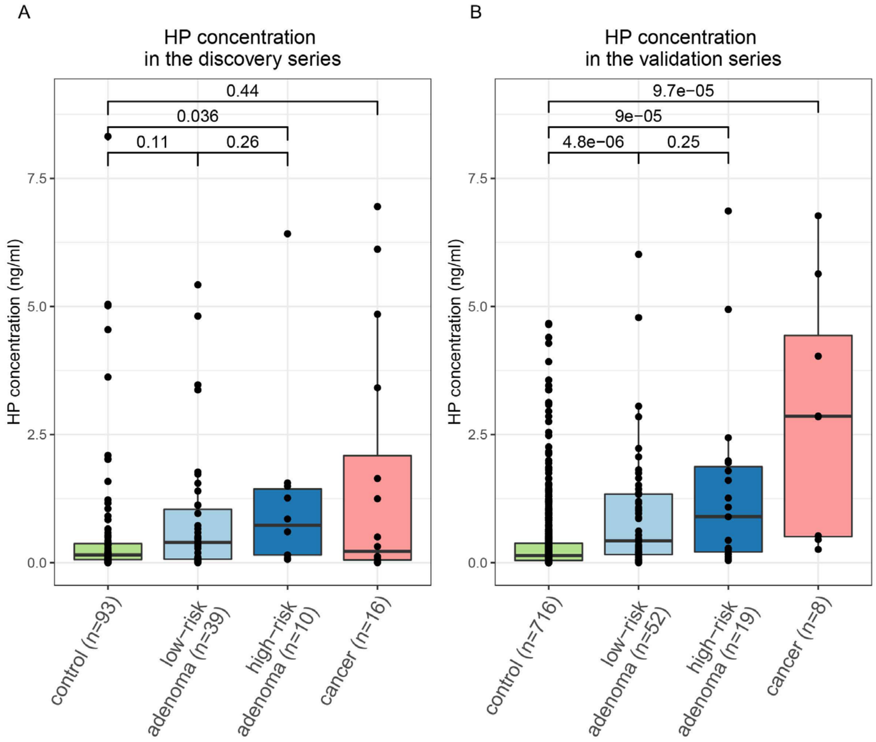


Figure 5. Validation of Hp protein expression with the use of an immunoassay. A. The discovery series B. The validation series.

DISCUSSION

It is well known that not all colorectal adenomas will progress to CRC. This underlines the importance to develop screening tests for the detection of specifically those adenomas that are at high risk of progressing to malignancy[27]. The widely used FIT is not optimal to detect such adenomas, and therefore additional biomarkers could aid to improve sensitivity for early detection of CRC. Proteins are an attractive category of molecules to be used as biomarkers for application in stool-based CRC screening, as they can be measured in small sample volumes with simple economic assays like FIT[28]. In the present study, we aimed to identify combinations of specific stool-based protein biomarkers that outperform haemoglobin in the detection of molecularly-defined high-risk adenomas and CRCs. Based on their DNA copy number profiles, adenomas were classified into lesions at low or high risk of progressing to cancer[16-18]. High-risk adenomas comprised 15.8% of all adenomas and 36.4% of the advanced adenomas. Using mass spectrometry proteomics on stool samples and regression modelling we selected marker panels consisting of up to four proteins that distinguish screen-relevant lesions, i.e. high-risk adenomas and CRCs, from controls. We have identified a biomarker panel of Hp, LAMP1, SYNE2 and ANXA6 for identification of high-risk adenomas and two biomarker panels; Hp and LRG1 as well as Hp, LRG1, RBP4 and FN1, for identification of high-risk adenomas and CRCs that outperformed haemoglobin. Since Hp was the single protein present in all three biomarker panels it was selected for further validation. To test its applicability in a screening setting we used antibody-based assays on FIT samples for the validation experiments. The higher concentration of Hp in high-risk adenomas and CRCs compared to controls was confirmed using an immunoassay in FIT samples of both the discovery series as well as a validation series.

Using mass spectrometry analysis of stool samples, we previously established protein panels that showed a higher sensitivity for advanced adenoma and CRC samples compared to haemoglobin[19]. In the present study, we performed subsequent statistical analyses to select alternative candidate biomarkers, including the most promising protein combinations that may improve the current stool-based CRC population screening in the detection of high-risk adenomas and CRCs. Statistical analysis of discovery mass spectrometry proteomics datasets on complex samples like stool are challenging due to missing data. Therefore two feature selection methods were used to select the best biomarker panels for identification of cases *vs* controls accounting for complexity of our dataset; the beta-binomial test[29] and Lasso regularization in the regression modelling[30]. The beta-binomial test was used for detection of proteins higher expressed in the cases than controls, while logistic regression with Lasso regularization was applied to select for the best combination of these higher expressed proteins to distinguish cases from the controls. Lasso regularization shrinks coefficients of less importance or correlating features to zero therefore achieving a sparser solution, i.e. smaller number of features in the final regression model. This method does not only avoid overfitting, but also performs feature selection of the best performing model.

A limitation of this study was the small number of molecularly-defined high-risk adenoma patients (n=15), which affected performance of the model built on only high-risk adenomas as cases. Based on our previous work it was anticipated that only a limited number of even the morphologically defined advanced adenomas would carry two or more CAEs[18]. However, the most relevant screening targets are CRCs as well as adenomas considered at high risk of progression. In line with this approach, combining CRCs and molecularly-defined high-risk adenomas increased the size of the set of cases, and improved the performance of the models. Moreover, in the discovery series FIT results were not available for all samples (162 out of 277 samples), which limited the possibilities of direct comparison of the marker panels to FIT performance, especially for the high-risk adenomas (n=10 with FIT available).

The marker panels in the discovery phase consistently contained haptoglobin (Hp), which as the haemoglobin-haptoglobin complex has been previously investigated as a biomarker for CRC[31]. The Hp-Hb complex has been suggested to render a more stable biomarker than Hb or Hp alone, and could therefore increase sensitivity for the more proximal lesions in the bowel[32]. This, however, was not confirmed in the current study (data not shown). It has been described that the sensitivity for CRC does not increase with the detection of an Hp-Hb complex compared to haemoglobin alone, but the sensitivity for adenomas does[33]. In this study, the sensitivity of the complex versus the single proteins could not be assessed. Nevertheless, using an antibody-based assay, higher abundance of Hp was confirmed in FIT samples of patients with high-risk adenomas and CRCs in the discovery series and in a much larger independent validation series. These findings underline the importance of Hp as a biomarker for screen-relevant lesions and hold promise for future application of Hp in CRC screening. Meanwhile, haemoglobin (HBA1, HBB or HBD) was not significantly differential between high-risk adenomas compared to controls and subsequently it was not selected in any of the biomarker panels, which is in line with the limited sensitivity of FIT for adenomas. Although one would expect that Hp is a marker of blood in the stool and therefore should not have complementary value to haemoglobin, our data suggest that Hp is of added value for the detection of high-risk adenomas. A possible explanation may be that the Hp protein detected in stool is not only derived from blood but may also be derived from the CRC or high-risk adenoma tissues. In line with this, Hp has been described to be expressed by colorectal cancer cells; both cell lines as well as within the tumour where its expression was associated with the stage of progression[34].

Next to Hp, LAMP1, SYNE2 and ANXA6 were selected in the analysis for high-risk adenomas and also LRG1, RBP4 and FN1 for the high-risk adenomas and CRCs. LAMP1 is a lysosome-associated membrane protein, which has been implicated in several tumour-promoting activities such as promotion of metastasis, drug resistance and cancer cell survival[35]. The gene coding for LAMP1 is located on chromosome 13q, gain of which is one of the seven CAEs used for classifying adenomas as high-risk. SYNE2 (or Nesprin 2) is a nuclear envelope protein that is involved in regulation of nuclear trafficking; even though its role in cancer is yet to be established there are indications that its presence is pivotal in the DNA damage response[36]. Since high-risk adenomas are characterized by chromosomal gains and losses, the upregulation of SYNE2 might be linked to these DNA aberrations. ANXA6 is present at the cell membrane and in the endosomal compartments, where it functions as a multifunctional scaffolding protein. In that position, ANXA6 can contribute to many different processes including cancer cell migration and invasion[37]. RBP4 has been linked to insulin resistance and it has been shown to be present in serum of breast cancer patients[38], and was previously described as a potential marker for colorectal advanced adenomas in stool[19]. FN1 is an extracellular matrix protein that is involved in cell adhesion and migration processes; it has been shown to be present in serum of patients with hepatocellular carcinoma and has been suggested as a biomarker for this disease[39]. Finally, LRG1 has been reported to be highly upregulated in CRC, both at the mRNA as well as at the protein level[40, 41]. An evident role in tumour development has been established for LRG1, as it stimulates proliferation and inhibition of apoptosis through regulating RUNX1 expression[40, 42]. In addition, the protein is secreted and may therefore end up in blood or stool. Indeed, increased protein levels of LRG1 in plasma have been reported for colorectal cancer and colon adenoma patients[40, 43, 44]. Altogether for the majority of these biomarker proteins their potential involvement in tumour biology has been demonstrated. Further investigation is needed to evaluate the diagnostic potential of these protein biomarkers in a CRC screening setting.

The present study is unique because a molecularly-defined intermediate endpoint was used for biomarker discovery, by applying chromosomal copy number alterations highly associated with colorectal adenoma-to-carcinoma progression. This is in contrast to the morphological features traditionally used to define the advanced adenoma intermediate endpoint. Our study resulted in the identification of novel protein biomarker panels with higher sensitivities for high-risk adenomas and CRCs than HBA1, which have plausible roles in colorectal carcinogenesis. These biomarkers have the potential to improve current FIT-based screening strategies.

ACKNOWLEDGEMENTS

The research was supported by a SU2C-DCS International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415, MEDOCC) and by the Dutch Cancer Society (KWF Kankerbestrijding), project number 2013-6025. Stand Up to Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. Support for these studies was also provided by the Dutch Digestive Foundation and VU University Medical Center, Cancer Center Amsterdam (proteomics infrastructure). The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. This article is based upon work from COST Action [CA17118], supported by COST (European Cooperation in Science and Technology). The authors would like to thank Genomics Core Facility at the Netherlands Cancer Institute for sequencing.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018.
2. Young GP, Rabeneck L, Winawer SJ. The Global Paradigm Shift in Screening for Colorectal Cancer. Gastroenterology 2019;156:843-851 e2.
3. Kerr J, Day P, Broadstock M, et al. Systematic review of the effectiveness of population screening for colorectal cancer. N Z Med J 2007;120:U2629.
4. Carroll MR, Seaman HE, Halloran SP. Tests and investigations for colorectal cancer screening. Clin Biochem 2014;47:921-39.
5. Zauber AG, Winawer SJ, O’Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 2012;366:687-96.
6. Lee JK, Liles EG, Bent S, et al. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Ann Intern Med 2014;160:171.
7. de Wijkerslooth TR, Stoop EM, Bossuyt PM, et al. Immunochemical fecal occult blood testing is equally sensitive for proximal and distal advanced neoplasia. Am J Gastroenterol 2012;107:1570-8.
8. Song LL, Li YM. Current noninvasive tests for colorectal cancer screening: An overview of colorectal cancer screening tests. World J Gastrointest Oncol 2016;8:793-800.
9. Haug U, Knudsen AB, Lansdorp-Vogelaar I, et al. Development of new non-invasive tests for colorectal cancer screening: the relevance of information on adenoma detection. Int J Cancer 2015;136:2864-74.
10. Imperiale TF, Kahi CJ. Cost-effectiveness of Future Biomarkers for Colorectal Cancer Screening: Quantified Futility or Call for Innovation? Clin Gastroenterol Hepatol 2018;16:483-485.
11. Lansdorp-Vogelaar I, Goede SL, Bosch LJW, et al. Cost-effectiveness of High-performance Biomarker Tests vs Fecal Immunochemical Test for Noninvasive Colorectal Cancer Screening. Clin Gastroenterol Hepatol 2018;16:504-512 e11.
12. Shinya H, Wolff WI. Morphology, anatomic distribution and cancer potential of colonic polyps. Ann Surg 1979;190:679-83.
13. Winawer SJ, Zauber AG, O’Brien MJ, et al. The National Polyp Study. Design, methods, and characteristics of patients with newly diagnosed polyps. The National Polyp Study Workgroup. Cancer 1992;70:1236-45.
14. Muto T, Bussey HJ, Morson BC. The evolution of cancer of the colon and rectum. Cancer 1975;36:2251-70.
15. Click B, Pinsky PF, Hickey T, et al. Association of Colonoscopy Adenoma Findings With Long-term Colorectal Cancer Incidence. JAMA 2018;319:2021-2031.
16. Carvalho B, Postma C, Mongera S, et al. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. Gut 2009;58:79-89.
17. Hermsen M, Postma C, Baak J, et al. Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. Gastroenterology 2002;123:1109-19.
18. Carvalho B, Diosdado B, Terhaar Sive Droste JS, et al. Evaluation of Cancer-Associated DNA Copy Number Events in Colorectal (Advanced) Adenomas. Cancer Prev Res (Phila) 2018;11:403-412.

19. Bosch LJW, de Wit M, Pham TV, et al. Novel Stool-Based Protein Biomarkers for Improved Colorectal Cancer Screening: A Case-Control Study. Ann Intern Med 2017;167:855-866.
20. Dutch Federation of Biomedical Scientific Societies. Code for Proper Secondary Use of Human Tissue in the Netherlands. http://www.federa.org/ 2011.
21. de Wijkerslooth TR, de Haan MC, Stoop EM, et al. Study protocol: population screening for colorectal cancer by colonoscopy or CT colonography: a randomized controlled trial. BMC Gastroenterol 2010;10:47.
22. Stoop EM, de Haan MC, de Wijkerslooth TR, et al. Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. Lancet Oncol 2012;13:55-64.
23. Voorham QJ, Carvalho B, Spiertz AJ, et al. Chromosome 5q loss in colorectal flat adenomas. Clin Cancer Res 2012;18:4560-9.
24. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26:1367-72.
25. 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.
26. Debad J, Glezer E, Wohlstadter J, et al. Clinical and biological applications of ECL. Electrogenerated chemiluminescence Marcel Dekker, Inc, New York, NY., 2004:359-396.
27. Sillars-Hardebol AH, Carvalho B, van Engeland M, et al. The adenoma hunt in colorectal cancer screening: defining the target. J Pathol 2012;226:1-6.
28. Bosch LJ, Carvalho B, Fijneman RJ, et al. Molecular tests for colorectal cancer screening. Clin Colorectal Cancer 2011;10:8-23.
29. Pham TV, Piersma SR, Warmoes M, et al. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. Bioinformatics 2010;26:363-9.
30. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1-22.
31. Karl J, Wild N, Tacke M, et al. Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers. Clin Gastroenterol Hepatol 2008;6:1122-8.
32. Sieg A, Thoms C, Luthgens K, et al. Detection of colorectal neoplasms by the highly sensitive hemoglobin-haptoglobin complex in feces. Int J Colorectal Dis 1999;14:267-71.
33. Vasilyev S, Smirnova E, Popov D, et al. A New-Generation Fecal Immunochemical Test (FIT) Is Superior to Quaiac-based Test in Detecting Colorectal Neoplasia Among Colonoscopy Referral Patients. Anticancer Res 2015;35:2873-80.
34. Marino-Crespo O, Cuevas-Alvarez E, Harding AL, et al. Haptoglobin expression in human colorectal cancer. Histol Histopathol 2019:18100.
35. Alessandrini F, Pezze L, Ciribilli Y. LAMPs: Shedding light on cancer biology. Semin Oncol 2017;44:239-253.
36. Kelkar P, Walter A, Papadopoulos S, et al. Nesprin-2 mediated nuclear trafficking and its clinical implications. Nucleus 2015;6:479-89.
37. Grewal T, Hoque M, Conway JRW, et al. Annexin A6-A multifunctional scaffold in cell motility. Cell Adh Migr 2017;11:288-304.

38. Jiao C, Cui L, Ma A, et al. Elevated Serum Levels of Retinol-Binding Protein 4 Are Associated with Breast Cancer Risk: A Case-Control Study. PLoS One 2016;11:e0167498.
39. Kim H, Park J, Kim Y, et al. Serum fibronectin distinguishes the early stages of hepatocellular carcinoma. Sci Rep 2017;7:9449.
40. Zhou Y, Zhang X, Zhang J, et al. LRG1 promotes proliferation and inhibits apoptosis in colorectal cancer cells via RUNX1 activation. PLoS One 2017;12:e0175122.
41. Choi JW, Liu H, Shin DH, et al. Proteomic and cytokine plasma biomarkers for predicting progression from colorectal adenoma to carcinoma in human patients. Proteomics 2013;13:2361-74.
42. Fijneman RJ, Anderson RA, Richards E, et al. Runx1 is a tumor suppressor gene in the mouse gastrointestinal tract. Cancer Sci 2012;103:593-9.
43. Ladd JJ, Busald T, Johnson MM, et al. Increased plasma levels of the APC-interacting protein MAPRE1, LRG1, and IGFBP2 preceding a diagnosis of colorectal cancer in women. Cancer Prev Res (Phila) 2012;5:655-64.
44. Zhang Q, Huang R, Tang Q, et al. Leucine-rich alpha-2-glycoprotein-1 is up-regulated in colorectal cancer and is a tumor promoter. Onco Targets Ther 2018;11:2745-2752.
45. Liu H, Sadygov RG, Yates JR, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004;76:4193-201.

Table 1. Confusion matrix for the cross-validated performance of the models of biomarker panels. Performance of the biomarker panel regression models were evaluated at 95% specificity and compared to haemoglobin. A. high-risk adenomas *versus* controls and B. high-risk adenomas and CRCs *versus* controls.

Table 1A

Protein(s)	**Control**	**High-risk adenoma**	**Sensitivity at 95% specificity** [95% confidence intervals]
Hp, LAMP1, SYNE2, ANXA6			
Predicted control	123	7	53% [27%, 79%]
Predicted high-risk adenoma	6	8	
HBA1			
Predicted control	123	13	13% [2%, 40%]
Predicted high-risk adenoma	6	2	

Table 1B

Protein(s)	**Control**	**High-risk adenoma or CRC**	**Sensitivity at 95% specificity** [95% confidence intervals]
Hp, LRG1, RBP4, FN1			
Predicted control	123	36	62% [51%, 72%]
Predicted high-risk adenoma or CRC	6	58	
Hp, LRG1			
Predicted control	123	32	66% [55%, 75%]
Predicted high-risk adenoma or CRC	6	62	
HBA1			
Predicted control	123	56	40% [30%, 51%]
Predicted high-risk adenoma or CRC	6	38	

SUPPLEMENTARY MATERIALS AND METHODS

DNA copy number analysis

In brief, isolated DNA was subjected to low-coverage whole-genome sequencing on a HiSeq 2000 (Illumina) in a 50-bp single-read modus using the Illumina Truseq Nano kit. Raw sequence reads were mapped to the human reference genome build GRCh37/hg19 and data was further analysed using QDNAseq, CGHcall, CGHregions[18]. Adenomas were characterized for gains of chromosomal arms 8q, 13q, and 20q, and losses of 8p, 15q, 17p, and 18q.

LC-MS/MS data analysis

Briefly, Swissprot human reference FASTA file was used as database (canonical and isoforms, obtained in October 2017, 20237 entries). Contaminants and reversed proteins were removed. Protein groups with a positive Andromeda score were extracted. Proteins were quantified by spectral counting[45]. Protein groups were excluded from further analysis if they had missing data for over 15% of the cases, i.e. 13 samples for high-risk adenomas or 80 samples for high-risk adenomas and CRCs. Euclidian distance between samples was calculated based on their protein expression profiles and proteomics data was visualized using the multidimensional scaling algorithm. Differential protein expression analysis was performed using the beta-binominal test[29], log2 fold changes and p-values were obtained. P-values adjusted for multiple hypothesis testing were obtained with the Benjamini-Hochberg correction. Differential analysis was performed for the following groups and the following thresholds were applied to select for proteins higher expressed in cases than in controls: stool samples from high-risk adenoma patients compared to samples from controls (log2 fold change>0 and p-value≤0.1), and stool samples from CRC and high-risk adenoma patients compared to samples from controls (log2 fold change ≥ 2 and adjusted p-value≤0.05). Clustering of the proteins higher expressed in cases than controls was performed using hierarchical clustering, where protein abundances were normalized to Z-scores. Subsequently, the Euclidean distance was used with ward linkage for samples and complete linkage for proteins.

Haptoglobin quantification in FIT samples

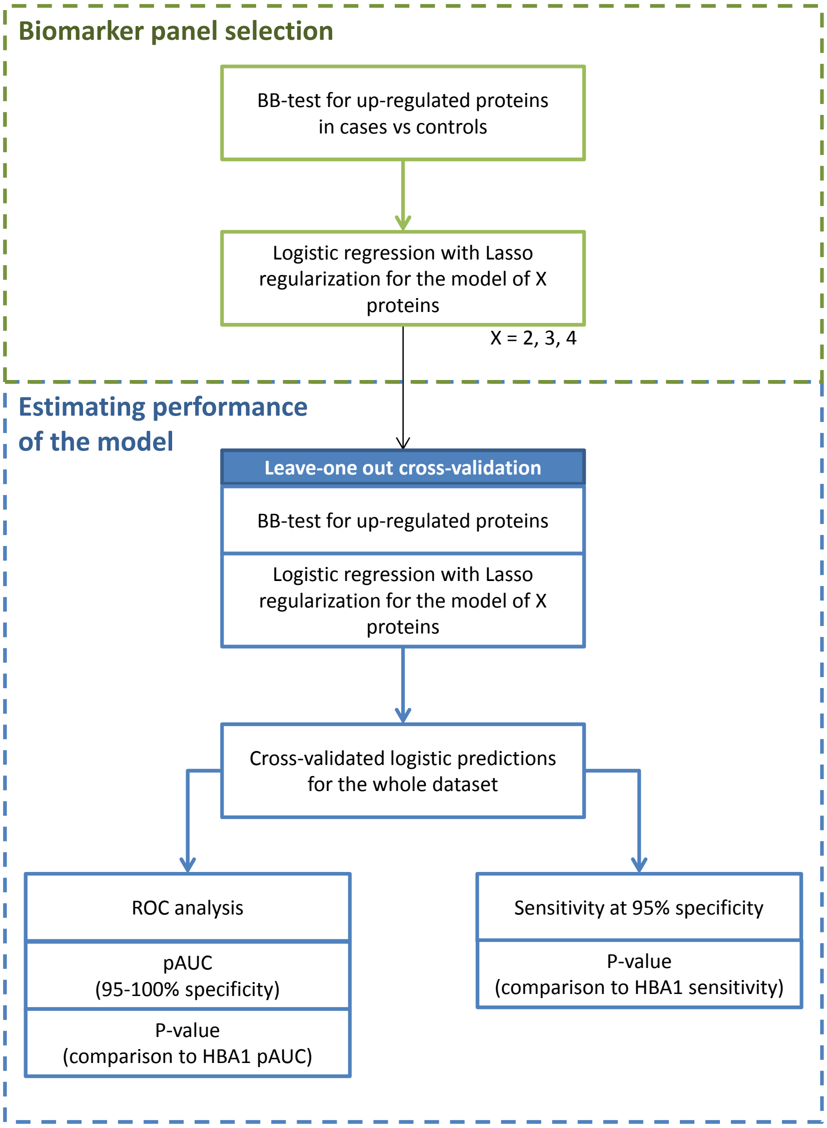
The immunoassay for Hp employed a sandwich immunoassay format and electrochemiluminescence (ECL) detection was carried out on commercial instrumentation and multi-well plate consumables from Meso Scale Diagnostics, LLC (MSD)[26]. The assay was run in MSD's U-PLEX format. The U-PLEX format employs 96-well plates, in which each well comprises a screen-printed carbon ink electrode coated with a generic 10-plex array of binding reagents.

The capture antibodies (goat polyclonal; MSD) were biotinylated with Sulfo-NHS-LC-Biotin (Thermo Fisher Scientific) and coupled to U-PLEX linkers via biotin-streptavidin binding. Detection antibodies (goat polyclonal; MSD) were conjugated to the MSD

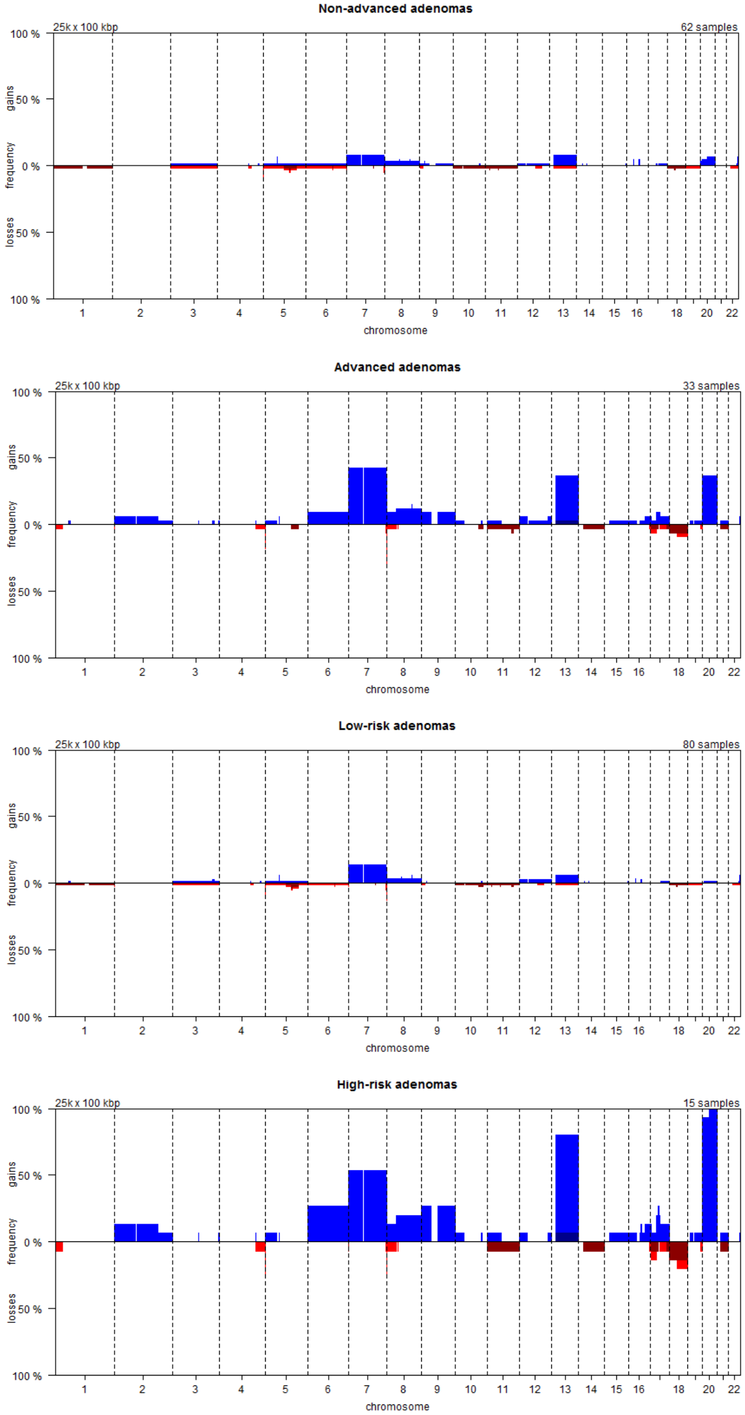
SULFO-TAG ECL label. The assay was run according to the following protocol using commercial diluents from MSD: (i) Capture antibody-linker conjugate, specific for the target, was prepared and used immediately or stored at 4°C. To each well of the U-PLEX plate, 50 µL of this was added. The plates were incubated for 1 hour at room temperature with shaking to allow the antibody arrays to assemble and then washed with 1X MSD Wash Buffer to remove excess unbound capture antibody. (ii) MSD Diluent 100 (99 µL) was combined with 1 µL of sample in each well of the plate, and the plates were incubated for 1 hour at room temperature with shaking to bind Hp in the sample to the capture antibody in the well. Each plate was calibrated with an 8-point standard curve of purified Hp (50 µL per well) prepared in the same diluent; all samples were run in duplicate. (iii) After washing the wells to remove the unbound sample, 25 µL of SULFO-TAG-labeled detection antibody (in MSD Diluent 100) was added and incubated for an additional hour at room temperature with shaking to complete the immunoassay sandwich. (iv) Plates were washed to remove the unbound detection antibody, then the wells were filled with 150 µL of 2X MSD Read Buffer T with surfactant. ECL was measured on an MSD SECTOR Imager 6000 plate reader. The plate reader applies a voltage to the electrodes in each well and quantitates the light emission from each array spot.

The relationship of ECL signal to calibrator concentration was fit to a 4-parameter logistic (4-PL) model with 1/Y2 weighting. Concentrations for the test samples were calculated by back-fitting ECL signals to the 4-PL fit for each plate.[26]

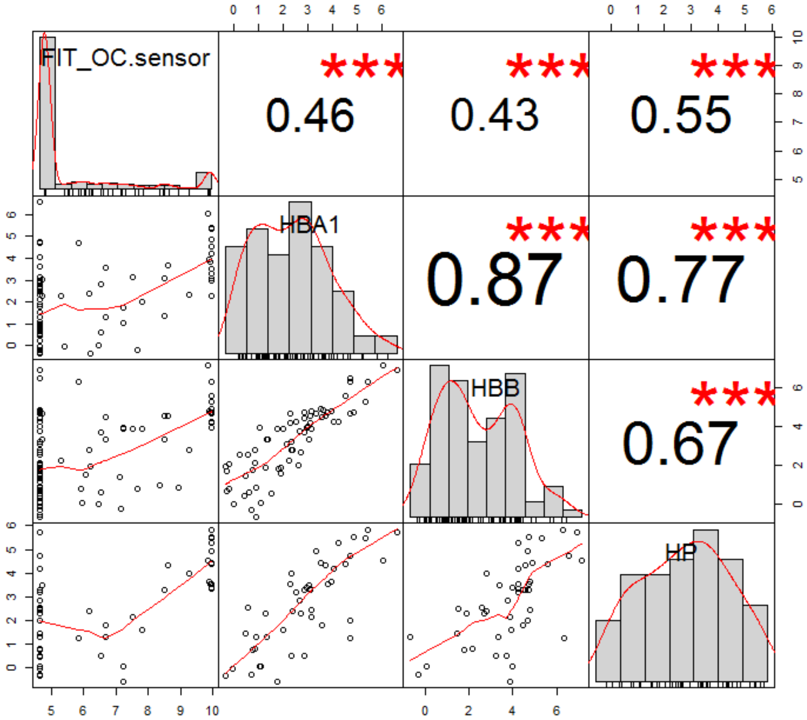
Supplementary Figures


Biomarker panel selection
BB-test for up-regulated proteins in cases vs controls
Logistic regression with Lasso regularization for the model of X proteins
X = 2, 3, 4
Estimating performance of the model
Leave-one out cross-validation
BB-test for up-regulated proteins
Logistic regression with Lasso regularization for the model of X proteins
Cross-validated logistic predictions for the whole dataset
ROC analysis
pAUC (95-100% specificity)
P-value (comparison to HBA1 pAUC)
Sensitivity at 95% specificity
P-value (comparison to HBA1 sensitivity)

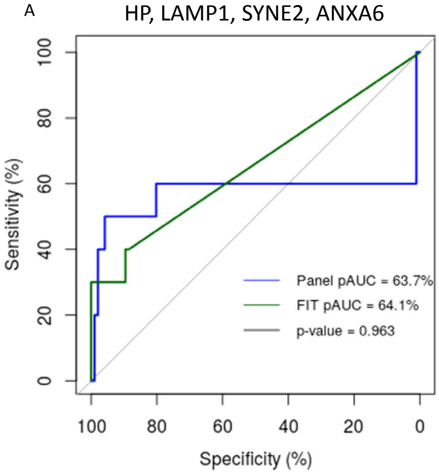
Supplementary Figure 1. Overview of the data analysis approach for the biomarker panel identification. Feature selection was performed using beta-binomial test (BB-test) in the comparative setting cases *vs* controls, in particular high-risk adenomas *vs* healthy controls and high-risk adenomas with CRCs *vs* healthy controls. Up-regulated proteins were selected using different thresholds for each comparison (see Materials and Methods). Logistic regression with Lasso regularization was applied to built a model based on X features (where X is either two, three or four features). The performance of the model was evaluated using leave-one-out cross-validation, where feature selection with BB-test and logistic regression with Lasso regularization were repeated. Cross-validated performance of the built models were evaluated with respect to hemoglobin (HBA1) at high specificity levels.



Supplementary Figure 2. Frequency plots of DNA copy number aberrations in the adenomas. Copy number aberrations are plotted per set; in non-advanced adenomas (n = 62), advanced adenomas (n = 33), low-risk adenomas (n = 80) and high-risk adenomas (n=15).

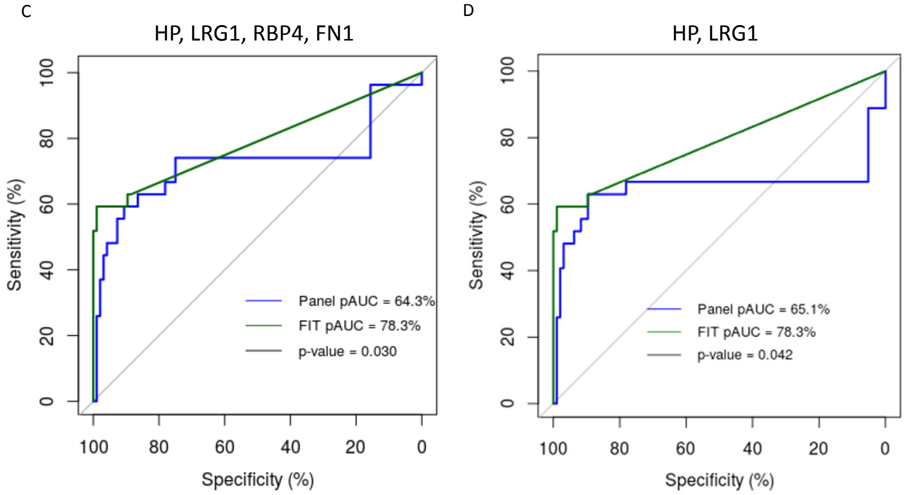


Supplementary Figure 3. Spearman correlation analysis of hemoglobin (HBA1, HBB) and haptoglobin (HP) spectral counts and FIT values. Logarithmic transformation was applied on spectral counts and FIT values. The correlation analysis was performed on all the samples for which FIT values were available, including healthy controls (n=96), low-risk adenomas (n=43), high-risk adenomas (n=10), unclassified adenomas (n=8) and CRCs (n=17). Bottom left matrix presents bivariate scatter plots with a fitted line. Top right displays correlation coefficient and significance level, where “***” means p-value ≤ 0.001.



B

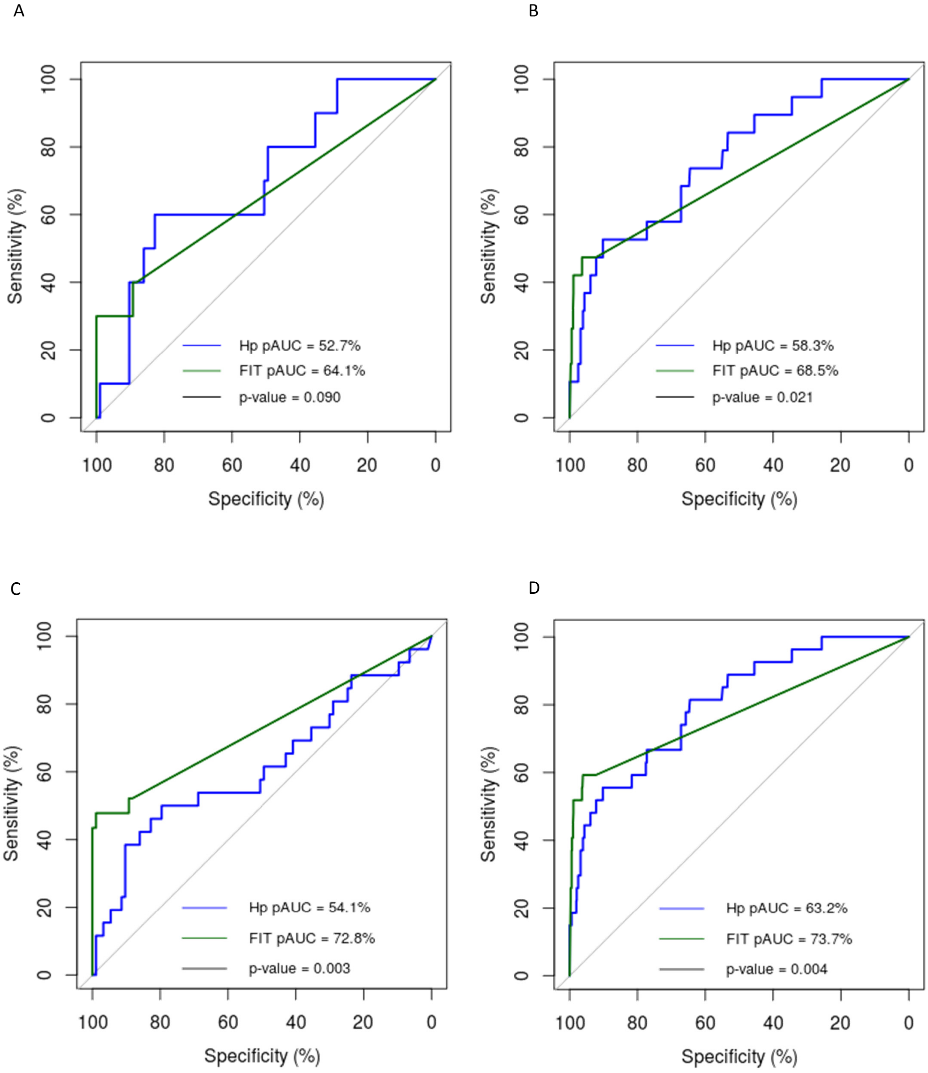
		Control	High-risk adenoma	Sensitivity
FIT	Predicted control	93	7	30%
	Predicted high-risk adenoma	3	3	
Model after CV	Predicted control	92	5	50%
	Predicted high-risk adenoma	4	5	



E

		Control	CRC + high-risk adenoma	Sensitivity
FIT	Predicted control	92	11	59%
	Predicted CRC + high-risk adenoma	4	16	
Model after CV	Predicted control	92	14	48%
	Predicted CRC + high-risk adenoma	4	13	

Supplementary Figure 4. Comparison of the biomarker panels to FIT values. The FIT data was available for healthy controls (n = 96), high-risk adenomas (n=10) and CRCs (n=17). The cross-validated performance of four proteins model was evaluated for high-risk adenoma identification (A-B), pAUC was calculated for ROC curve at the specificity level of 95-100% and compared to FIT values. Sensitivities of the model and FIT were evaluated at 95% specificity (B). For identification of high-risk adenomas and CRCs, four (C) and two (D) feature cross-validated models were evaluated with pAUC for ROC curves at the specificity level of 95-100%. Sensitivities for 95% specificity for both models resulted in the same sensitivity, which was compared to FIT (E).



Supplementary Figure 5. Comparison of the diagnostic performance of FIT and haptoglobin (Hp) measured with an antibody-based assay for high-risk adenomas (A-B) and high-risk adenomas with CRCs (C-D). ROC curves were obtained and pAUC at the specificity level of 95-100% were calculated separately for the study series (A, C) and the validation series (B, D). The study series consisted of 93 healthy controls, 10 high-risk adenomas and 16 CRCs while the validation series included 716 healthy controls, 19 high-risk adenomas and 8 CRCs.

SUPPLEMENTARY TABLES

Supplementary Table 1. Frequencies of cancer associated events and histologic features of the adenomas, n = number of samples. P-values were obtained using χ2 test or Fisher exact test, if the number of samples was smaller than five. Q-values were obtained using Benjamini-Hochberg correction.

	Gain						Losses								Risk of progression			
Total n=95	8q		13q		20q		8p		15q		17p		18q		low risk (≤ 1 CAE)		high risk (≥ 2 CAE)	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Adenoma	6/95 (6.3)	2.4-13.2	17/95 (17.9)		16/95 (16.8)	10.8-27.1	1/95 (1.1)	0.0-5.7	0/95 (0.0)	0.0-3.8	1/95 (1.1)	0.0-5.7	3/95 (3.2)	0.7-9.0	80/95 (84.2)	75.3-90.9	15/95 (15.8)	9.1-24.7
All																		
Advanced	4/33 (12.1)	3.4-28.2	12/33 (36.4)	20.4-54.9	12/33 (36.4)	20.4-54.9	1/33 (3.0)	0.1-15.8	0/33 (3.0)	0.0-10.6	1/33 (0.0)	0.1-15.8	2/33 (6.1)	0.7-20.2	21/33 (63.6)	45.1-79.6	12/33 (36.4)	20.4-54.9
Non-advanced	2/62 (3.2)	0.4-11.2	5/62 (8.1)	2.7-17.8	4/62 (6.5)	1.8-15.7	0/62 (0.0)	0.0-5.8	0/62 (0.0)	0.0-5.8	0/62 (0.0)	0.0-5.8	1/62 (1.6)	0.0-8.7	59/62 (95.2)	86.5-99.0	3/62 (4.8)	1.0-13.5
Non-adjusted p-value	p=0.2		**p=0.001**		**p<0.001**		p=0.4		n.a.		p=0.4		p=0.3		**p<0.001**		**p<0.001**	
Adjusted p-value (q-value)	q=0.5		**q=0.009**		**q=0.004**		q=0.7		n.a.		q=0.7		q=0.6		**q=0.003**		**q=0.003**	
Size																		
>= 10 mm	3/24 (12.5)	2.7-32.4	10/24 (41.7)	22.1-63.4	9/24 (37.5)	18.8-59.4	0/24 (0.0)	0.0-14.2	0/24 (0.0)	0.0-14.2	1/24 (4.2)	0.1-21.1	1/24 (4.2)	0.1-21.1	15/24 (62.5)	40.6-81.2	9/24 (37.5)	18.8-59.4
< 10 mm	3/71 (4.2)	0.9-11.9	7/71 (9.9)	4.1-19.3	7/71 (9.9)	4.1-19.3	1/71 (1.4)	0.0-7.6	0/71 (0.0)	0.0-5.1	0/71 (0.0)	0.0-5.1	2/71 (2.8)	0.3-9.8	65/71 (91.6)	82.5-96.8	6/71 (8.5)	3.2-17.5
Non-adjusted p-value	p=0.2		**p=0.001**		**p=0.004**		p=1.0		n.a.		p=0.2		p=1.0		**p=0.002**		**p=0.002**	
Adjusted p-value	q=0.5		**q=0.009**		**q=0.020**		q=1.0		n.a.		q=0.6		q=1.0		**q=0.012**		**q=0.012**	
Grade of dysplasia																		
High	0/3 (0.0)	0.0-70.8	3/3 (100.0)	29.2-100.0	1/3 (33.3)	0.8-90.6	0/3 (0.0)	0.0-70.8	0/3 (0.0)	0.0-70.8	0/3 (0.0)	0.0-70.8	1/3 (33.3)	0.8-90.6	2/3 (66.7)	9.4-99.2	1/3 (33.3)	0.8-90.6
Low	6/92 (6.5)	2.4-13.7	14/92 (15.2)	8.6-24.2	15/92 (16.3)	9.4-25.5	1/92 (1.1)	0.0-5.9	0/92 (1.1)	0.0-5.9	1/92 (0.0)	0.0-3.9	2/92 (2.2)	0.3-7.6	78/92 (84.8)	75.8-91.4	14/92 (15.2)	8.6-24.2
Non-adjusted p-value	p=1.0		**p=0.005**		p=0.4		p=1.0		n.a.		p=1.0		p=0.1		p=0.4		p=0.4	
Adjusted p-value	q=1.0		**q=0.021**		q=0.7		q=1.0		n.a.		q=1.0		q=0.3		q=0.7		q=0.7	
Histology																		
Tubular	2/72 (2.8)	0.3-9.7	8/72 (11.1)	4.9-20.7	6/72 (8.3)	3.1-17.3	0/72 (0.0)	0.0-5.0	0/72 (0.0)	0.0-5.0	0/72 (0.0)	0.0-5.0	1/72 (1.4)	0.0-7.5	67/72 (93.1)	84.5-97.7	5/72 (6.9)	2.3-15.5
(Tubulo)villous	4/23 (17.4)	5.0-38.8	9/23 (39.1)	19.7-61.5	10/23 (43.5)	23.2-65.5	1/23 (4.3)	0.1-21.9	0/23 (4.3)	0.1-21.9	1/23 (0.0)	0.0-14.8	2/23 (8.7)	1.1-28.0	13/23 (56.5)	34.5-76.8	10/23 (43.5)	23.3-65.5
Non-adjusted p-value	**p=0.029**		**p=0.005**		**p<0.001**		p=0.2		n.a.		p=0.2		p=0.1		**p<0.001**		**p<0.001**	
Adjusted p-value	q=0.111		**q=0.021**		**q=0.004**		q=0.6		n.a.		q=0.6		q=0.5		**q=0.003**		**q=0.003**	
Age																		
≥ 65 years old	2/51 (3.9)	0.5-13.4	9/51 (17.7)	8.4-30.1	9/51 (17.7)	8.4-30.1	1/51 (2.0)	0.0-10.4	0/51 (0.0)	0.0-7.0	1/51 (2.0)	0.0-10.4	1/51 (2.0)	0.0-10.4	42/51 (82.4)	69.1-91.6	9/51 (17.7)	8.4-30.9
< 65 years old	4/44 (9.1)	2.5-21.7	8/44 (18.2)	8.2-32.7	7/44 (15.9)	8.2-30.1	0/44 (0.0)	0.0-8.0	0/44 (0.0)	0.0-8.0	0/44 (0.0)	0.0-8.0	2/44 (4.6)	0.6-15.5	38/44 (86.4)	72.6-94.8	6/44 (13.6)	5.2-27.4
Non-adjusted p-value	p=0.4		p=1.0		p=1.0		p=1.0		n.a.		p=1.0		p=0.6		p=0.8		p=0.8	
Adjusted p-value	q=0.7		q=1.0		q=1.0		q=1.0		n.a.		q=1.0		q=0.9		q=1.0		q=1.0	
Gender																		
Male	5/57 (8.8)	2.9-19.3	11/57 (19.3)	10.0-31.9	10/57 (17.5)	8.7-29.9	1/57 (1.8)	0.0-9.4	0/57 (0.0)	0.0-6.3	1/57 (1.8)	0.0-9.4	1/57 (1.8)	0.0-9.4	48/57 (84.2)	72.1-92.5	9/57 (15.8)	7.5-27.9
Female	1/38 (2.6)	0.1-13.8	6/38 (15.8)	6.0-31.3	6/38 (15.8)	6.0-31.3	0/38 (0.0)	0.0-9.3	0/38 (0.0)	0.0-9.3	0/38 (0.0)	0.0-9.3	2/38 (5.3)	0.6-17.7	32/38 (84.2)	68.7-94.0	6/38 (15.8)	6.0-31.3
Non-adjusted p-value	p=0.4		p=0.8		p=1.0		p=1.0		n.a.		p=1.0		p=0.6		p=1.0		p=1.0	
Adjusted p-value	q=0.7		q=1.0		q=1.0		q=1.0		n.a.		q=1.0		q=0.9		q=1.0		q=1.0	

Supplementary Table 3. Performance of the biomarker panels in the dataset including low-risk adenomas at the specificity level of 95%.

HP, LAMP1, SYNE2, ANXA6	**Low-risk adenoma**	**Sensitivity**	**95% CI**
Predicted control	51	9.0%	[3%, 20%]
Predicted high-risk adenoma	5		
HP, LRG1, RBP4, FN1	**Low-risk adenoma**		
Predicted control	50	11%	[4%, 22%]
Predicted high-risk adenoma or CRC	6		
HP, LRG1	**Low-risk adenoma**		
Predicted control	49	13%	[5%, 24%]
Predicted high-risk adenoma or CRC	7		

Chapter 7

SUMMARY AND FUTURE PERSPECTIVES

SUMMARY

Colorectal cancer is a major health concern worldwide. The inverse relationship between the disease stage and survival emphasizes the importance of early detection of CRC, and thus implementation of the population-wide screening programs. In the Netherlands, the CRC screening is based on the FIT (fecal immunochemical test) as triage for colonoscopy, which identifies approximately 79% of CRCs and 27% of advanced adenomas. As hemoglobin, which is detected by FIT, is not specific for cancer, performance of FIT could be further increased with protein products of molecular alterations occurring in the tumor cells, which accompany colorectal carcinogenesis. The aim of this thesis was to identify candidate biomarkers to improve early detection of colorectal cancer. This included biomarkers to detect adenomas at an increased risk of progressing to cancer, and biomarkers to improve the performance of the current screening test.

Currently, in clinical practice, morphologically-defined advanced adenomas are considered relevant precursor lesions of colorectal cancer, while this definition alone is not a precise predictor of malignant transformation and leads to overdiagnosis and overtreatment. In the first two chapters of this thesis (chapters 2-3) we focused on molecular characterization of colorectal adenomas, in particular making use of DNA copy number aberrations to distinct adenomas at low risk of progressing to cancer from the ones at high risk, with the aim to characterize these adenomas on RNA and protein level in the context of malignant transformation. In chapter 2, we performed molecular profiling of colorectal tissue samples on DNA, RNA and protein level and presented comparative analysis of normal colon samples, advanced adenomas (including low-risk adenomas and high-risk adenomas) and CRCs. We showed that molecularly-defined high-risk adenomas are enriched in biological processes inherent to CRC when compared to low-risk adenomas. As in this study both low-risk and high-risk adenomas were advanced, this indicated that high-risk adenomas may be more relevant precursors of CRC than advanced adenomas. Moreover, we identified gene-dosage effect for three potential drivers of colorectal tumor development that play a role in the transition from low-risk to high-risk adenoma or cancer: EIF6, RPRD1B and POFUT1. In chapter 3 we performed CMS (Consensus Molecular Subtype) classification of colorectal adenomas to examine whether the CMS gene signature can distinct lesions at the premalignant stage. We demonstrated that there is heterogeneity in terms of the CMS classes among colorectal adenomas. CMS1 class was associated with microsatellite instability, CMS2 class with high-risk of progression and CMS3 class was associated with adenomas that are unlikely to progress to CRC. No CMS4 adenomas were identified. This study confirmed that adenomas differ on the molecular level, which provides insights into their underlying biology and indicates differential risk of progressing to cancer.

Next, we focused on novel protein biomarker identification for early detection of colorectal cancer. From the rich source of molecular alterations accompanying colorectal carcinogenesis, alternative splicing is one that often results in an alternative protein product. We studied alternative splicing with the aim to identify protein isoforms that can serve as candidate biomarkers in colorectal cancer screening. In chapter 4, we developed a computational proteogenomic pipeline, Splicify, for identification of splice variants that are differential between two conditions and that are translated to protein isoforms. And so, in contrast to other proteogenomic tools available, Splicify can be used for qualitative and quantitative analysis in the comparative setting disease *versus* healthy control. We showed the utility of the pipeline on CRC cell lines with down-modulation of splicing machinery. In chapter 5 we applied Splicify on colorectal tissue samples. We have shown that there is a switch in splicing between different stages of colorectal tumor progression; normal colon, colorectal adenomas and colorectal cancers, providing a source of promising candidate biomarkers. Due to the significant quantitative differences on RNA and protein level between normal colon, adenoma and cancer samples, we selected isoforms of NT5C3A for further validation. As FIT is a non-invasive test that detects protein hemoglobin in stool samples, a desirable solution would be a test that detects NT5C3A isoforms in stool as well with a similar technology, *i.e.* an antibody-based assay. To this end, we have developed a set of isoform-specific antibodies that will be used in the future to examine the biomarker potential of NT5C3A isoforms in stool and FIT samples.

Chapter 6 of this thesis describes protein biomarker discovery in a proteomics dataset of stool samples. In this chapter, we examined whether individuals with molecularly-defined high-risk adenomas can be distinguished from the ones without colorectal neoplasia, i.e. controls, based on abundance of proteins identified in their stool samples. We performed the analysis for the joint set of clinically relevant lesions: high-risk adenomas and colorectal cancers compared to controls. Biomarker panels, consisting of HP and LRG1 or HP, LRG1, FN1 and RBP4, were identified, which performed significantly better than hemoglobin alone in the identification of individuals with high-risk adenomas and CRCs, in this series. Based on its most significant predictive value, we selected HP for further validation in an independent series and confirmed its increased levels in FIT samples of individuals with high-risk adenomas or CRCs compared to controls.

FUTURE PERSPECTIVES

In this thesis we presented a number of proteins and protein isoforms playing a role in or accompanying adenoma-to-carcinoma progression. Functional analysis of these results is crucial to understand the underlying biology of colorectal tumor development. Also, validation in large clinical series of their usefulness as potential biomarkers for high-risk adenoma and/or colorectal cancer detection is necessary.

UNRAVELING THE BIOLOGY

Studying the biology of adenoma-to-carcinoma progression is challenging as adenomas, once detected during colonoscopy, are completely removed, thereby interrupting their natural history in terms of either progressing to cancer or not. Therefore, currently adenoma-to-carcinoma progression can only be studied using e.g. *in vitro* organoid models, which for instance has been done by perturbing frequently mutated genes in CRC [1]. In parallel to the work presented in this thesis, we have shown that by overexpressing oncogenic miR17-92 cluster, which was previously associated with colorectal adenoma-to carcinoma progression[2], adenoma organoids express a "more carcinogenic transcriptome". Studies on POFUT1, EIF6 and RPRD1B should be performed in the adenoma organoids in a similar manner to unravel how these proteins may be involved in colorectal tumor progression. For instance, based on our findings, we expect that overexpression of POFUT1 in colorectal adenoma organoids may increase their proliferation rate. Additionally, NT5C3A splicing was identified to be differential between different stages of CRC development (chapter 5). Manipulation of NT5C3A splicing in adenoma and carcinoma organoids and studying their phenotype will reveal whether NT5C3A isoforms directly affect adenoma-to-carcinoma progression or are a passenger product of other driver events.

As alternative splicing may have an impact on protein function, cancers benefit from expressing aberrantly spliced isoforms which play a role in cancer development or progression [3], with the well-known examples of anti-apoptotic isoform of BCL2L1 [4] or pro-angiogenic splice variant of VEGFA [5]. Recently, alternative splicing has been considered as a source of cancer-specific neoepitopes, as a number of splice variants are translated to proteins that in theory can be presented by MHC molecules. Using a proteogenomic approach, cancer-specific RNA splice variants can be identified with RNA sequencing and potential alternative splicing-derived neoepitopes with mass spectrometry. MHC presentation can be predicted for these isoform-specific peptides per patient knowing each patient's HLA type or can be measured with mass spectrometry-based immunopeptidomics [6-9]. However, even though a number of isoform-specific peptides are predicted to bind to the MHC molecules, it is not sufficient to prove that they can induce an immune response. Experimental validation, through e.g. T-cell screening needs to be performed to evaluate if alternative splicing-derived peptides are indeed neoepitopes and if they have potential to induce an immune response. To our knowledge, these experiments have not been performed so far for alternative splicing-derived peptides. In the colorectal tissue dataset presented in this thesis, a number of cancer-specific alternative splicing-derived peptides were indeed predicted to bind to MHC molecules (data not shown); however, further validation is crucial to draw meaningful conclusions from this analysis. In general, given that MSI CRC patients respond to immune checkpoint blockade better than MSS CRC patients [10] and isoforms identified in this project were shared between MSI and MSS CRCs, it is uncertain if alternative splicing can be as successful as the tumor mutational burden in guiding colorectal cancer immunotherapy treatment.

TRANSLATION TO CLINICAL PRACTICE

In this thesis we have provided additional evidence that, based on the molecular profiles, high-risk adenomas may be more relevant precursors of CRC than the currently used advanced adenomas (chapters 2-3). These findings, however, need further validation to incorporate the “high-risk adenoma” definition in the clinical practice. Molecularly-defined high-risk adenomas could be used as intermediate endpoints in CRC screening and surveillance; namely, in the detection of relevant CRC precursor lesions in population-wide screening programs; and as indicators of risk of metachronous lesions, thereby determining the frequency of follow-up colonoscopies after polypectomy in colorectal cancer surveillance.

The test used in the CRC screening program (FIT) is non-invasive as it relies on detection of hemoglobin in stool. Therefore, for straightforward implementation of novel biomarkers for early detection of CRC preferably they should be detectable in stool samples. Additionally, for the costs and logistics of the population-wide screening the biomarker should be detectable in small sample volumes, comparable to the ones used for FIT. Methods detecting aberrations in the DNA in stool (e.g. Cologuard [11]) often require a larger sample volume and are technically more challenging when compared to antibody-based protein detection. Therefore, the introduction of novel protein biomarkers appears to be more suitable for implementation to improve current population-wide screening programs.

In this thesis we have shown that with mass spectrometry we are able to identify over 9000 human proteins in tissue samples, while in stool we have identified almost 800 human proteins (chapters 5 and 6). The reasons for this difference in experiment depth are not only the differences in the proteome variety of these samples but also in the complexity of these samples. Therefore, the global discovery experiments in stool provide only “the tip of the iceberg” while targeted approaches like antibody-based assays are expected to be more sensitive. As the colorectal tissue-derived proteins POFUT1 and NT5C3A were not identified in the stool proteomics experiment (chapter 6), these biomarkers will need to be evaluated with an antibody-based approach. In particular, if quantitative differences of POFUT1 between low-risk adenomas, high-risk adenomas and CRCs remain in the stool samples, POFUT1 could improve the FIT by increasing the detection rate of adenomas at higher risk of progressing to cancer. The advantage of using alternative splicing as a source of potential biomarkers, e.g. NT5C3A isoforms, is the fact that the isoforms reflect not only quantitative but also qualitative changes specific for adenoma-to-carcinoma progression. Therefore, with the use of isoform-specific antibodies relative expression, in the form of ratios of one isoform compared to the general part of the protein, may be obtained. These ratios would not depend on the stool composition, e.g. stool density, and subsequently the thresholds applied to distinguish individuals with CRC from healthy ones would be more robust compared to the thresholds for hemoglobin in the FIT. In this thesis, next to NT5C3A isoforms, we have identified a number of other isoforms as candidate biomarkers that still should be evaluated for their abundance in stool (chapter 5). However, isoform-specific antibody generation is a costly and timely procedure that is not feasible to be performed globally for all candidates, posing a significant hurdle towards this necessary step of biomarker validation. Therefore, even though there is a potential in protein isoforms as biomarkers, it will take time before they will be considered for clinical implementation.

Individuals with a history of colorectal neoplasia carry an increased risk of developing CRC in the future and therefore are enrolled in the colonoscopy-based surveillance programs [12]. Introduction of the population-wide screening increased the detection rate of such individuals and subsequently the demand for CRC surveillance. Currently, detection of advanced adenoma is an indication to shorten the interval for the follow-up surveillance colonoscopy. The high prevalence of advanced adenomas in an elderly population leads to a substantial burden on endoscopic capacity. It is estimated that approximately 25% of endoscopic capacity is occupied by surveillance colonoscopies and the number is still increasing due to screening [12]. Moreover, given that not all advanced adenomas eventually progress to cancer, frequent surveillance colonoscopies in patients with these lesions lead to overdiagnosis and overtreatment. Introduction of a more specific definition of adenomas with an increased risk of progression to malignancy could not only reduce patient burden but also improve the cost-effectiveness of the CRC surveillance program. Incorporation of the molecularly-defined high-risk adenoma definition as an intermediate endpoint for colorectal cancer could be a solution, as only ~30% of advanced adenomas carry DNA copy number aberrations associated with adenoma-to-carcinoma progression [13]. First, studies need to be performed to evaluate if patients with high-risk adenomas indeed have higher CRC incidence and mortality rate than patients with advanced adenomas. Using health technology assessment modelling, alternative surveillance strategies should be proposed and compared to the current one. Next, POFUT1 could be evaluated for its performance as a biomarker to identify high-risk adenomas in the surveillance setting by *e.g.* immunoassay or mass spectrometry measurements of adenomas removed during colonoscopy procedure. Introduction of a protein biomarker in a surveillance program, instead of DNA copy number profiling, could decrease the cost of the test and subsequently improve the cost-effectiveness of the program.

Recently, a significant progress has been made in the field of non-invasive liquid biopsies using circulating tumor DNA (ctDNA) in blood for applications in cancer management, where most of the methods rely on detection of somatic mutations in cell free DNA. However, as the abundance of ctDNA correlates with tumor size and stage, so far, the sensitivities of mutations detected in ctDNA in early stage CRC need further improvement [14, 15], especially for individuals with colorectal adenomas [16]. A recently developed method holding promise for future clinical application in early detection of cancer in liquid biopsies is measuring fragmentation patterns of cell free DNA shed from cancer cells into the blood [17]. Another novel approach that improved CRC identification rate based on addition of protein biomarkers to the somatic mutations in the ctDNA assay [18]. This confirmed that in terms of early detection of CRC, proteins remain very promising as biomarkers. However, none of these methods were yet applied to evaluate detection rates for colorectal adenomas and their performance was lower for early stage CRC, while detection of these lesions is crucial for colorectal cancer screening and surveillance.

REFERENCES

1. Matano M, Date S, Shimokawa M, et al. Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. Nat Med 2015;21:256-62.
2. Diosdado B, van de Wiel MA, Terhaar Sive Droste JS, et al. MiR-17-92 cluster is associated with 13q gain and c-myc expression during colorectal adenoma to adenocarcinoma progression. Br J Cancer 2009;101:707-14.
3. Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene 2014;33:5311-8.
4. Boise LH, Gonzalez-Garcia M, Postema CE, et al. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 1993;74:597-608.
5. Ladomery MR, Harper SJ, Bates DO. Alternative splicing in angiogenesis: the vascular endothelial growth factor paradigm. Cancer Lett 2007;249:133-42.
6. Lundegaard C, Lamberth K, Harndahl M, et al. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 2008;36:W509-12.
7. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. Nat Protoc 2019;14:1687-1707.
8. Bassani-Sternberg M. Mass Spectrometry Based Immunopeptidomics for the Discovery of Cancer Neoantigens. Methods Mol Biol 2018;1719:209-221.
9. Reits EA, Hodge JW, Herberts CA, et al. Radiation modulates the peptide repertoire, enhances MHC class I expression, and induces successful antitumor immunotherapy. J Exp Med 2006;203:1259-71.
10. Kalyan A, Kircher S, Shah H, et al. Updates on immunotherapy for colorectal cancer. J Gastrointest Oncol 2018;9:160-169.
11. Imperiale TF, Wagner DR, Lin CY, et al. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. N Engl J Med 2000;343:169-74.
12. Hassan C, Quintero E, Dumonceau JM, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2013;45:842-51.
13. Carvalho B, Diosdado B, Terhaar Sive Droste JS, et al. Evaluation of Cancer-Associated DNA Copy Number Events in Colorectal (Advanced) Adenomas. Cancer Prev Res (Phila) 2018;11:403-412.
14. Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 2017;9.
15. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6:224ra24.
16. Myint NNM, Verma AM, Fernandez-Garcia D, et al. Circulating tumor DNA in patients with colorectal adenomas: assessment of detectability and genetic heterogeneity. Cell Death Dis 2018;9:894.
17. Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 2019.
18. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018;359:926-930.

NEDERLANDSE SAMENVATTING

SAMENVATTING

Colorectaal carcinoom is de verzamelnaam voor dikkedarmkanker en endeldarmkanker en is wereldwijd een groot gezondheidsprobleem. In Nederland zijn er ongeveer 15.000 nieuwe diagnoses per jaar. Wanneer colorectaal carcinoom in een laat stadium wordt ontdekt, bijvoorbeeld op het moment dat er al sprake is van uitzaaiingen naar andere organen (stadium IV), is de kans op genezing aanzienlijk kleiner dan wanneer de ziekte in een vroeg stadium (stadium I) wordt ontdekt. Op dit moment is de vijfjaarsoverleving van patiënten met stadium I en stadium IV colorectaal carcinoom respectievelijk 95% en 11%. Vroeg detectie van colorectaal carcinoom is dan ook enorm belangrijk.

De ontwikkeling van colorectaal carcinoom begint met de vorming van een adenoom (een goedaardige tumor). Een klein gedeelte van deze adenomen, circa 5%, zullen zich verder ontwikkelen tot kanker. Met behulp van het bevolkingsonderzoek darmkanker kunnen zowel vroege stadia van darmkanker als voorstadia van darmkanker, de adenomen, worden ontdekt. In Nederland wordt voor dit onderzoek gebruik gemaakt van de fecale immunochemische test (FIT), een test die bloed in de ontlasting kan detecteren, waarna bij een positieve test een kijkonderzoek van de endeldarm, dikke darm en het laatste stukje van de dunne darm (coloscopie) wordt aanbevolen. De FIT is in staat ongeveer 79% van de colorectaal carcinomen en 27% van de voortgeschreden adenomen te detecteren. Er is dan ook ruimte voor verbetering, bijvoorbeeld door gericht te zoeken naar biomarkers die direct gerelateerd zijn aan de moleculaire veranderingen in tumorcellen. In de kliniek wordt op basis van morfologische en histologische kenmerken (afwijkingen in de weefselstructuur) bepaald of een adenoom naar verwachting zal uitgroeien tot colorectaal carcinoom. Een hoog risico adenoom wordt ook wel voortgeschreden adenoom genoemd en wordt gezien als relevant voorstadium van colorectaal carcinoom. Deze morfologische en histologische kenmerken zijn echter onvoldoende om de overgang van goedaardige tumor naar darmkanker betrouwbaar te kunnen voorspellen.

Darmkanker ontstaat door veranderingen in het DNA van een cel. Deze veranderingen kunnen allerlei biologische processen in de cel verstoren waardoor deze zich kan ontwikkelen tot een kankercel. DNA vormt een code met informatie van meer dan 20 duizend genen. Deze genen worden eerst vertaald in messenger-RNA, afgekort als mRNA, wat op haar beurt weer de informatie bevat voor de vorming van eiwitten. Afwijkingen in het DNA van een kankercel kunnen daarom leiden tot afwijkende eiwitvarianten, of tot afwijkingen in de hoeveelheid (expressie) van gewone eiwitten. Op deze manier weerspiegelen deze eiwitten de veranderingen in de moleculaire processen die ten grondslag liggen aan het ontstaan van kanker en de verdere ontwikkeling van de ziekte. Men kan deze eiwitten gebruiken als zogenaamde ‘biomarkers’ in klinische testen, voor vroege opsporing van ziekte, om het verloop van de ziekte te voorspellen en om te bepalen welke (chemo) therapie het beste kan werken voor een bepaalde patiënt.

Het doel van het onderzoek beschreven in dit proefschrift was om kandidaat biomarkers te identificeren die bij kunnen dragen aan een verbeterde vroegtijdige opsporing van darmkanker. Daarbij hebben we gezocht naar twee soorten biomarkers. Ten eerste bekeken we biomarkers die specifiek adenomen kunnen identificeren met een verhoogd risico op progressie tot colorectaal carcinoom. Ten tweede, hebben we gekeken naar biomarkers die toegevoegd zouden kunnen worden aan het bevolkingsonderzoek om de prestatie van de huidige FIT test te verbeteren.

In hoofdstuk 2 en 3 van dit proefschrift hebben we ons daarom gericht op de karakterisatie van de moleculaire kenmerken van colorectale adenomen. In hoofdstuk 2 hebben we op DNA-, RNA- en eiwitniveau gezocht naar verschillen tussen normaal darmweefsel, tumorweefsel van voortgeschreden adenomen en darmkanker weefsel. Hierbij hebben we de voortgeschreden adenomen op basis van specifieke chromosomale afwijkingen onderverdeeld in ‘laag-risico’ en ‘hoog-risico’ op progressie van ziekte. We hebben aangetoond dat er in hoog-risico adenomen meer biologische processen aanwezig zijn die geassocieerd worden met darmkanker ten opzichte van laag-risico adenomen. Hoewel alle adenomen in deze studie vielen in de categorie ‘voortgeschreden adenomen’, geeft deze analyse aan dat de subset van hoog-risico adenomen de meest relevante voorloperstadia van darmkanker betreffen. Op basis van correlatie tussen chromosomale afwijkingen en expressie van genen op RNA- en eiwit-niveau hebben we drie genen geïdentificeerd die mogelijk een drijvende kracht zijn bij de transitie van een laag-risico naar een hoog-risico adenoom: EIF6, RPRD1B en POFUT1. In hoofdstuk 3 hebben we adenomen onderworpen aan de CMS-classificatie (Consensus Molecular Subtype) om te zien of in voorstadia van darmkanker al verschillen in CMS-subtypes aantoonbaar zijn. Daarmee hebben we de heterogeniteit van colorectale adenomen aangetoond. Adenomen van de CMS2-klasse zijn met name geassocieerd met kenmerken van hoog risico op progressie naar darmkanker, terwijl adenomen van de CMS3-klasse met name geassocieerd zijn met het laag-risico type. Deze studie bevestigt dat adenomen, hoewel morfologisch niet van elkaar te onderscheiden, op moleculair niveau sterk van elkaar verschillen.

Vervolgens hebben we ons gericht op de identificatie van nieuwe eiwit-biomarkers voor de vroege detectie van darmkanker. Hierbij hebben we gebruik gemaakt van het feit dat er vaak meerdere varianten van een gen tot expressie worden gebracht door ‘alternatieve splicing’, wat ook kan resulteren in eiwitvarianten die kunnen dienen als kandidaten voor nieuwe biomarkers voor de opsporing van darmkanker.

In hoofdstuk 4 hebben we een zogenaamde proteogenomische analyse pijplijn ontwikkeld, 'Splicify', voor identificatie van eiwitvarianten met differentiële expressie tussen twee condities, bijvoorbeeld tussen kankerweefsel en gezond weefsel. We hebben de toepasbaarheid van de Splicify pijplijn aangetoond door gebruik te maken van een darmkanker cellijn waarbij we het mechanisme van alternatieve splicing artificieel hebben onderdrukt. In hoofdstuk 5 hebben we vervolgens Splicify toegepast op normaal darmweefsel, adenomen en colorectaal carcinomen. We hebben aangetoond dat er veel verschillen zijn in expressie van eiwitvarianten tussen deze weefsels, wat heeft geleid tot identificatie van nieuwe veelbelovende kandidaat biomarkers. Een van deze nieuwe biomarkers betreft eiwitvarianten van NT5C3A, welke we hebben geselecteerd voor verdere validatie. Hiertoe hebben we een aantal antilichamen gegenereerd, die specifiek de ene of de andere eiwitvariant van NT5C3A kunnen herkennen. In de toekomst zullen deze antilichamen worden gebruikt om het biomarker-potentieel van NT5C3A-eiwitvarianten in ontlasting- en FIT-monsters te onderzoeken.

Hoofdstuk 6 van dit proefschrift beschrijft de ontdekking van biomarkers in een grootschalige eiwit dataset van ontlastingsmonsters. In dit hoofdstuk hebben we onderzocht of individuen met hoog-risico adenomen kunnen worden onderscheiden van gezonde controles op basis van eiwitten die zijn geïdentificeerd in hun ontlastingmonsters. We hebben de analyse uitgevoerd door de klinisch relevante set van hoog-risico adenomen en darmkankers te vergelijken met gezonde controles. Vergeleken met het eiwit hemoglobine dat nu wordt gedetecteerd in FIT-monsters, werden de eiwitten HP en LRG1 samen of de combinatie van HP met LRG1, FN1 en RBP4 geïdentificeerd als significant beter geschikt voor het opsporen van personen met hoog-risico adenomen en darmkanker. HP werd geselecteerd voor verdere validatie in een onafhankelijke serie FIT-monsters, waarbij we verhoogde eiwitniveaus hebben gedetecteerd in personen met hoog-risico adenomen of colorectaal carcinoom ten opzichte van gezonde controles. Hiermee hebben we de mogelijke toepasbaarheid van dit eiwit als biomarker bevestigd.

De resultaten beschreven in dit proefschrift zullen als basis dienen voor toekomstig onderzoek gericht op het ontrafelen van de biologische processen die ten grondslag liggen aan de ontwikkeling van colorectaal carcinoom. Tevens biedt dit onderzoek aanknopingspunten voor vervolgonderzoek naar veelbelovende biomarkers, voor potentieel gebruik in klinische toepassingen.

"I don't like to lose - at anything…
Yet I've grown most not from victories, but setbacks.
If winning is God's reward, then losing is how he teaches us."
Serena Williams

APPENDUM

CURRICULUM VITAE

Małgorzata Komór was born on November 10th 1990 in Warsaw, Poland. In 2009 she graduated from the XIV High School of Stanislaw Staszic in Warsaw where she followed an experimental mathematics program. In 2012 she received a Bachelor's degree in Bioinformatics and Systems' Biology from the University of Warsaw. During her internship at the Department of Bioinformatics of the Institute of Biochemistry and Biophysics, Polish Academy of Sciences in Warsaw she built a web interface to a database of small compounds. For her Bachelor thesis she conducted research with prof. dr hab. Paweł Golik at the Institute of Biochemistry and Biophysics, Polish Academy of Sciences in Warsaw on evolution of PPR proteins in *Schizosaccharomyces,* where she performed homology sequence analysis using Hidden Markov Models to identify PPR proteins in the genome sequence of *Schizosaccharomyces pombe*. In 2014 she obtained her double Master's degree*, cum laude* in Bioinformatics and System's Biology from University of Warsaw and Vrije Universiteit Amsterdam. She conducted research for her Master's thesis at the Department of Pathology, VU medical center under the supervision of dr. Evert van den Broek, dr. Remond Fijneman and dr. Sanne Abeln (Centre for Integrative Bioinformatics (IBIVU), Vrije Universiteit Amsterdam). The thesis title was "Transcriptome analysis of colorectal cancers: the quest for structural genomic rearrangements" and the project entailed integration of DNA copy number, RNA sequencing and mass spectrometry data of The Cancer Genome Atlas. At the end of her studies she worked as a student assistant at the "Genomics and Bioinformatics" course at the Amsterdam University College. In September 2014 she started her PhD project with dr. Remond Fijneman, prof. dr. Gerrit Meijer (Translational Gastrointestinal Oncology Group, Department of Pathology, Netherlands Cancer Institute) and prof. dr. Connie Jiménez (Oncoproteomics Laboratory, Amsterdam University Medical Centers, location VU medical center) on tumor-specific biomarkers for early detection of colorectal cancer. In parallel to her PhD studies, she was a board member of the Regional Student Group Netherlands of the International Society of Computational Biology. The results of her PhD project are described in this thesis.

LIST OF PUBLICATIONS

1. Komor, M. A., Pham, T. V., Hiemstra, A. C., Piersma, S. R., Bolijn, A. S., Schelfhorst, T., Delis-van Diemen, P. M., Tijssen, M., Sebra, R. P., Ashby, M., Meijer, G. A., Jimenez, C. R., and Fijneman, R. J. A. (2017) Identification of Differentially Expressed Splice Variants by the Proteogenomic Pipeline Splicify. *Molecular & cellular proteomics : MCP* **16**, 1850-1863

2. Komor, M. A., Bosch, L. J., Bounova, G., Bolijn, A. S., Delis-van Diemen, P. M., Rausch, C., Hoogstrate, Y., Stubbs, A. P., de Jong, M., Jenster, G., van Grieken, N. C., Carvalho, B., Wessels, L. F., Jimenez, C. R., Fijneman, R. J., and Meijer, G. A. (2018) Consensus molecular subtype classification of colorectal adenomas. *J Pathol* **246**, 266-276

3. Carvalho, B., Diosdado, B., Terhaar Sive Droste, J. S., Bolijn, A. S., Komor, M. A., de Wit, M., Bosch, L. J. W., van Burink, M., Dekker, E., Kuipers, E. J., Coupe, V. M. H., van Grieken, N. C. T., Fijneman, R. J. A., and Meijer, G. A. (2018) Evaluation of Cancer-Associated DNA Copy Number Events in Colorectal (Advanced) Adenomas. *Cancer Prev Res (Phila)* **11**, 403-412

4. Chen, S., Huang, V., Xu, X., Livingstone, J., Soares, F., Jeon, J., Zeng, Y., Hua, J. T., Petricca, J., Guo, H., Wang, M., Yousif, F., Zhang, Y., Donmez, N., Ahmed, M., Volik, S., Lapuk, A., Chua, M. L. K., Heisler, L. E., Foucal, A., Fox, N. S., Fraser, M., Bhandari, V., Shiah, Y. J., Guan, J., Li, J., Orain, M., Picard, V., Hovington, H., Bergeron, A., Lacombe, L., Fradet, Y., Tetu, B., Liu, S., Feng, F., Wu, X., Shao, Y. W., Komor, M. A., Sahinalp, C., Collins, C., Hoogstrate, Y., de Jong, M., Fijneman, R. J. A., Fei, T., Jenster, G., van der Kwast, T., Bristow, R. G., Boutros, P. C., and He, H. H. (2019) Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell* **176**, 831-843. e822

5. Komor, M. A., de Wit, M., van den Berg, J., Martens de Kemp, S. R., Delis-van Diemen, P. M., Bolijn, A. S., Tijssen, M., Schelfhorst, T., Piersma, S. R., Chiasserini, D., Sanders, J., Rausch, C., Hoogstrate, Y., Stubbs, A. P., de Jong, M., Jenster, G., Carvalho, B., Meijer, G. A., Jimenez, C. R., and Fijneman, R. J. A. (2019) Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression. *Int J Cancer*

ACKNOWLEDGEMENTS

There are things in life that do not come easy; like eating lunch at precisely 11:37 every day or completing a PhD thesis. Therefore, to accomplish those one needs persistence as well as input and support from many people.

Remond, there is no doubt we got here thanks to you. You created the idea behind this project and (with more or less freedom) let me work on it. You have taught me not only how to do science but also skills I never expected to learn: confidence in presenting my work, assertiveness towards my co-promotor and rewriting one sentence in ten different ways. I am just not sure yet in which aspect of my life I can apply the latter. But besides learning from you as a scientist, I also did enjoy working together. That is because of your sense of humor or because from the heated discussions we had I only remember the ones that went my way. Your enthusiasm every day drove my scientific curiosity and motivated me to work hard. I wish every PhD student a supervisor with so much passion for science as you have.

Connie, thank you for all your input in this project and in my growth as a PhD student. I appreciate how you always found the time for a scientific discussion, for giving me feedback on a manuscript or just for a chat. I admire your accomplishments, productivity and how you are always yourself. I enjoyed working together because of your scientific expertise, your honesty and advice.

Gerrit, this work would not have been done without your knowledge, innovative approach and management skills. If in a scientific discussion or informal chat, you always surprised me with your sharpness, point of view or way of handling conflict. I still have no idea if it was the power of arguments or some managerial tricks that made me go through all these FAIR data management tasks with a smile on my face. Thank you for your support, your optimism and for all I have learnt from you. It has been a pleasure to have you as my promotor.

I would like to share my gratitude to the members of the assessment committee for taking the time to review this thesis and join my defense.

Bea, even though not an official supervisor, you have guided me through this project as well. Thank you for sharing your knowledge and experience with me and for the atmosphere you create in this group.

Meike, Linda, next to thanking you for your scientific input in this work, I would also like to thank you for the emotional guidance during this PhD. At the end, my scientific journey is over, but my personal one isn't and I consider your contributions to both very valuable. Linda, you have encouraged me to learn and profit from negative feedback instead of fighting it and you taught me that the start of handling a problem is to just communicate it. Meike, next to your pivotal role in this project, you have been also an enormous support for me. Your experience with the same combination of (co)promotors and your personality made a perfect mix for me to ask for advice or joke around. It was not only fun but also very educational (also on other fronts than science) to share the office with you.

Mariska, since they put me in your room at the CCA you have been a friend. Thanks for answering all my PhD-related questions, for the fun in- and outside of the office and for always reading my mood.

Anne and Tim, an enormous amount of work presented in this thesis was done by you. However, you deserve to be acknowledged not only for the lab work, but also for the ease I had working with you. Regardless of how much you had to do for this project, I never heard you complain once. I always looked forward to working with you, because not only you are good at what you do but you also enjoy it. And as we had a lot of fun together also outside of work, I hope to continue this during a great defense party!

Pien, thank you for all the hard work you have put in this project and your patience in testing all the possible hypotheses I could think of. The most exciting findings of this thesis (in my subjective opinion) were validated by you and I am grateful for that.

I would also like to thank all the members (and ex-members) of the TGO group for the scientific discussions we had, for listening to me talk during all these group meetings, for always having the time and being eager to help me and for laughing at my jokes. My office-mates: Sanne, Meta, Willemijn and Alex, I enjoyed the conversations we had, especially the ones behind the closed doors. Other PhD students: Iris, Karlijn, Pieter and Carmen, good luck with your work, the sun is on the horizon! Marianne and Margriet, thank you for helping out in this project and always stepping in for Anne and Pien when needed. Pauline, thank you for being the second, very well organized Gerrit. Janneke, thank you for always speaking Dutch to me. And Evert, thank you for your supervision during my internship. Christian and Erik, I appreciate your feedback and help with the bioinformatics questions. Jan, Menno, Lana, thank you for all your help with the data management and the fun conversations. Annemieke, you did a lot for this project and paved the FAIR way for me, thank you! I would also like to thank the students I supervised, Domingo and Rosa, for their help in this project and the experience I gained from working with them. Finally, as my knowledge of the Dutch language is ever improving, it would be a lie to say that I am the sole author of the Dutch summary in this thesis. I would like to acknowledge Mariska, Meike, Iris, Menno, Willemijn and Remond for their help even though none of them agreed on how to call a biomarker or an advanced adenoma in Dutch.

My gratitude goes to the members of the OPL group. Sander, Thang, Jaco, you make the wheel spin. Thank you for sharing your knowledge and supporting me in this project. Frank, Franciska, Irene, Giulia, Aysa, Robin, Tessa, thank you for your feedback during group meetings, the great atmosphere in the group and the fun during HUPOs.

Sanne Abeln, Anton Feenstra and Jaap Heringa, thank you for giving me the knowledge from which I am still benefiting today and helping me with the start of my career.

I would like to share my gratitude with the great researchers from the NKI and VUmc. Hein te Riele and Lodewyk Wessels, thank you for being the members of my supervisory committee, for taking the time to follow my progress and for your feedback on the project. Lodewyk, Gergana Bounova, Daniel Vis and other members of the Computational Cancer Biology group, thank you for letting me learn from you and for your advice when I was stuck. I would also like to thank the members of the Department of Pathology (Jose and Joyce, thank you for teaching me some pathology) at the NKI, the NKI facilities; Research High Performance Computing, Genomics Core Facility, Core Facility Molecular Pathology & Biobanking and Protein Facility and many colleagues from the VUmc.

I would also like to acknowledge the collaborators from other institutes, all the members of the NGS-ProToCol consortium, including Guido, Youri and Andrew for your feedback, your help and great atmosphere at all the meetings. Members of the RNAAST meetings: Jacqueline, Victor, Anna, Rocco and Maxime, for the scientific discussion. And the Pacbio team, Meredith and Bobby, for the great collaboration.
But next to the remarkable group of colleagues, many people from my private life supported me during the last five years. I would like to thank all my friends and family members.

Marleen, you have been in this country with me from the start. I wouldn't go for this PhD, if I didn't have you as a friend. Thank you for always being there to listen to me talk about my work, for figuring out together how to start a life in a different country and for all the fun we have had over the years.

Wilma and Matthieu, you have been extremely supportive during the whole process. Thank you for giving me the perspective when I needed it and for celebrating every achievement of mine.

Kasia, since I remember you have always cheered me up and paved the way for me. Thank you for being the brave older sister, and for sharing your optimism and your advice with me.

Mom and Dad, you have raised me the way I am and to be honest you are probably not always very happy with the results. But getting here is a consequence of everything that you did and everything that you achieved in your lives. So, thank you for working hard, for being ambitious (which is a nice word for competitive), for always being curious, for your sense of humor and for your passion for travelling. I admire where you've come yourselves and I am proud of having you as my parents.

Rohan, you have been interested in this project from the beginning to the end and had so much patience listening to me talk about it that I am pretty sure you could defend this work yourself. Here, I would like to thank you for "being on the other side". Even though I might not like it, I appreciate that you always show me the other point of view. Thank you for introducing calmness in our home, for helping me to relax and for always taking the troubles away.