ROBERT COEBERGH VAN DEN BRAAK

COLON CANCER

THE CLINICAL SIGNIFICANCE OF MOLECULAR BIOMARKERS

Colon cancer - the clinical significance of molecular biomarkers

ISBN: 978-94-6295-826-5

Coverdesign & Lay out: Wendy Schoneveld || www.wenziD.nl
Printed by: Proefschriftmaken || Proefschriftmaken.nl

Printing this thesis has been financially supported by:
ABN Amro Bank, Erasmus MC University Medical Center, Department of Surgery, Blaak & Partners B.V., ChipSoft, Erasmus University Rotterdam, Erbe Nederland B.V., QP&S n.v., Nederlandse Vereniging voor Gastroenterologie, Raadsheeren B.V., Rabobank Rotterdam Medicidesk, De research manager, Servier Nederland Farma, Sirtex Medical Europe GmbH, Maag Lever Darm Stichting, Fonds NutsOhra, KWF Kankerbestrijding, Bayer B.V., Gericall

COLON CANCER
THE CLINICAL SIGNIFICANCE OF MOLECULAR BIOMARKERS

COLONCARCINOOM
DE RELEVANTIE VAN MOLECULAIRE BIOMARKERS

Proefschrift
ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof. dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

woensdag 31 januari 2018 om 15:30 uur

door

Robertus Rudolphus Johannes Coebergh van den Braak
geboren te Eindhoven

Erasmus University Rotterdam

Doctoral committee

Promoter	Prof. dr. J.N.M. IJzermans
Other members	Prof. dr. J.A. Foekens Prof. dr. J.P. Medema Prof. dr. M. Koopman
Copromoter	Prof. dr. J.W.M. Martens

TABLE OF CONTENTS

1

INTRODUCTION

Colorectal cancer is the second most common malignancy in the Western world after non-melanoma skin cancer.[1] In Europe, colorectal cancer has an annual incidence of 450,000 cases.[2] In the Netherlands, over 15,500 patients were diagnosed with colorectal cancer and over 5,000 patients with colorectal cancer died in 2015.[3] Most cases of colorectal cancer are sporadic, with hereditary forms being 3-5% of all cases.[4, 5]

As in many other solid cancers, the extent of disease progression is a very important determinant of prognosis in colorectal cancer. The most commonly used staging system is the TNM classification of malignant tumours (TNM), developed and maintained by the Union for International Cancer Control (UICC). In the Netherlands, the 5th version of the TNM classification is used for colorectal cancer.[6] The T category indicates the invasiveness of the primary tumour into the bowel wall, while the N category indicates the number of regional lymph nodes containing cancerous cells. The M stage indicates the absence or presence of distant dissemination to other organs. These three indicators are combined into a tumour stage ranging from stage I to stage IV (**Table 1**).

Tumour stage is also used to determine the treatment approach. Patients with stage I-III colorectal cancer (i.e. no distant metastases) should be considered for treatment with curative intent by radical resection of the primary tumour with en-bloc resection of the draining lymph nodes. In rectal cancer, surgery is preceded by (chemo) radiotherapy for patients with an intermediate or high risk carcinoma and is not followed by systemic therapy.[7] In colon cancer, the focus of this thesis, adjuvant systemic therapy may be considered in patients with stage III or high risk stage II disease.[8-10] However, to date it has not been established convincingly which individual

TABLE 1. THE 5TH TNM CLASSIFICATION FOR COLORECTAL CANCER

Stage	T	N	M
0	Tis	N0	M0
1	T1	N0	M0
	T2	N0	M0
2	T3	N0	M0
	T4	N0	M0
3	Tis-T4	N1-2	M0
4	Tis-T4	N0-2	M1

Tis=the tumour has not grown beyond the mucosa; T1=the tumour is confined to the submucosa; T2=the tumour has grown into (but not through) the muscularis propria; T3=the tumour has grown into (but not through) the serosa; T4=the tumour has penetrated through the serosa and the peritoneal surface extending directly into other nearby structures and/or causing a perforation of the bowel; N0=no lymph nodes contain tumour cells; N1=there are tumour cells in up to 3 regional lymph nodes; N2=there are tumour cells in 4 or more regional lymph nodes; M0=no metastasis to distant organs; M1=metastasis to distant organs.

patients with stage II may benefit from adjuvant chemotherapy. Current guidelines list five high-risk characteristics in stage II colon cancer: T4, a poorly differentiated tumour, lymphovascular or perineural invasion, tumour perforation and inadequate sampling of lymph nodes defined as the sampling of <10 or 12 lymph nodes depending on the guideline.[9-11] However, no prospective trials have demonstrated the predictive value of these features.[12] In the Netherlands, 10-15% of all patients with 'high risk' stage II receive adjuvant chemotherapy.[11, 13] In Stage IV colorectal cancer, 10-20% of the patients are eligible for curative treatment using a combination of local and systemic treatment options. The rest of the patients with stage IV colorectal cancer can be offered palliative (non-curative) treatment.[14]

The overall 5-year recurrence free survival rate in colorectal cancer is estimated to be 95% in stage I, 80-85% in stage II and 60-70% in stage III disease.[15] Similar, the 5-year disease specific survival rate is higher in stage I-II (91%) compared to stage III (72%) and stage IV (13%).[1] Although one may conclude that TNM staging holds considerable prognostic value, it should be noted that there are profound individual differences in clinical outcome within each stage. For example, one in five patients with stage II will develop recurrence of disease and these patients may thus be considered undertreated. Despite this recurrence rate, the small benefit of treating all stage II patients to prevent recurrence is too small to outweigh the hazards of chemotherapy for all patients which underlines the need for reliable criteria to identify the patients at risk.[16, 17] The introduction of adjuvant chemotherapy in stage III colon cancer has reduced the recurrence rate from 50-60% to 30-40% suggesting adjuvant chemotherapy may be effective in 20% of all patients with stage III colon cancer.[8] However, this also means 80% is over- or mistreated since 30-40% will still develop recurrent disease despite receiving adjuvant chemotherapy and roughly half of the patients with stage III colon cancer will never develop recurrence disease, irrespective of chemotherapy.

Therefore, additional factors are needed to complement the tool box of clinicians to come to a more tailor-made treatment for each individual patient. Currently, most effort in this field of research focuses on molecular biomarkers, characteristics enabling reliable disease identification. Well-maintained patient cohorts with both clinical data and biomaterial are key to conduct this type of research.

Due to a lack such patient cohorts, the MATCH study was initiated in 2007. The MATCH study (MicroArray and proteomics Technologies to analyse Colorectal cancer and Hepatic Metastases) was designed and set up to identify prognostic markers to predict colorectal cancer behaviour and response to treatment especially as far as liver metastasis is concerned using state of the art technology. The aim was to include a homogeneous population of approximately 1,000 patients with colorectal cancer

without metastatic disease and treated with curative surgical intent. In a collaborative project with the hospitals in Rotterdam-Rijnmond and neighbouring regions both detailed clinical data as well as high quality fresh frozen tissue samples of the resected primary tumours were collected. Eight hospitals (Maasstad Hospital, Rotterdam; Reinier de Graaf Hospital, Delft; Elisabeth-Tweesteden Hospital, Tilburg; Albert-Schweitzer Hospital, Dordrecht; IJsselland Hospital, Capelle a/d IJssel; Ikazia Hospital, Rotterdam; and Sint Franciscus Hospital, Rotterdam) ultimately included over 2,500 patients from the 1st of July 2007 onwards. This large number appeared to be necessary to include the estimated population of 1,000 patients without introducing a large variance in patient and tumour characteristics. All patient data and samples used in this thesis were collected as part of the MATCH study unless stated otherwise.

BIOLOGICAL SYSTEMS

The TNM staging system insufficiently reflects the clinical course of individual patients and does not represent the great variance in the biological behaviour of colorectal cancer. To fully understand the expansion of a malignant tumour, more insight is needed in the process of gene expression and the consecutive events within cancer cells. Briefly, the genetic and epigenetic processes can be split into four subsequent steps or layers of activity.

The first layer is the genome, which consists of DNA (Deoxyribonucleic acid) molecules. The human genome comprises of 23 chromosome pairs with a total of 3 billion base pairs, and contains all information required to grow and develop, to function and to reproduce. On estimate, the human genome encodes 20,000-25,000 protein-coding genes (1-1.5% of the total DNA).[18] Some of the regions that do not code protein encode noncoding RNA molecules (discussed below).[19] Other parts of the DNA sequence play structural roles in chromosomes such as centromeres, which have an important role during DNA replication, and telomeres, which protect the end of the chromosome from fusion with neighbouring chromosomes and from deterioration.[20, 21]

The second layer is the epigenome, a system that comprises all chemical compounds that regulate the accessibility and gene expression without changing the DNA sequence.[18] Epigenetic mechanisms are involved in the majority of biological processes in the human body including cell and tissue identity. During the transformation of multipotent cells (i.e. stem cells) to a specific type of cell, epigenetic changes silence genes involved in alternative lineages and genes necessary to stay pluripotent.[22] In the large intestine, undifferentiated progenitor cells arising from stem cells located at the base of the crypt move upward while dividing multiple times. During this upward migratory process, cells differentiate into various cell types of the epithelial layer of the large intestine such as Paneth cells, goblet cells and

enterocytes.[23] The effect of epigenetic marks can be long-lasting and irreversible which is essential in tasks such as memory or behaviour, but can also be reversible which is useful to regulate adaptations and development throughout life.[24] Epigenetic phenomena can be categorized into two main categories, DNA methylation and histone modification. Methyl groups can attach to segments of DNA molecules, generally turning gene activity and subsequently protein production off. Histone proteins form spool-like structures to enable DNA to be wound up into chromosomes. A variety of chemical tags attached to these histone proteins can be discerned by proteins in cells. These chemical tags determine whether that region of DNA should be exploited or ignored in that cell, thus affecting its functionality.

The third layer is the transcriptome, the collection of all transcripts which are RNA (RiboNucleic Acid) molecules.[18] The transcription of DNA into RNA molecules is the first step in gene expression, the process of carrying out an instruction encoded in the DNA. The transcriptome can be divided into messenger RNA (mRNA), which codes for proteins, and non-coding RNA, which include amongst others transfer RNAs, ribosomal RNAs, long non-coding RNAs and microRNAs. MicroRNAs are thought to function in mRNA silencing and post-transcriptional regulation of gene expression.[25, 26]

The fourth and last layer is the proteome, the entire set of proteins expressed by the genome.[27] Proteins are involved in virtually every process within cells, and have structural and mechanical functions. Proteins are produced through translation, a process during which the mRNA molecule binds to a ribosome, which then produces a chain of amino-acids based on the sequence of the mRNA molecule. When completed, this chain is folded to form the actual protein. During or after the synthesis, proteins can be modified in a process called posttranslational modification which often affects protein activity.

ONCOGENESIS

Oncogenesis, the formation of a cancer, is the transformation of normal cells to cancer cells. This process is characterized by changes at cellular, genetic and epigenetic levels. These changes disturb the balance between cell proliferation and programmed cell death (apoptosis) thus leading to abnormal cell numbers. In 2000, the biological properties of malignant tumour cells underlying this imbalance were described as the hallmarks of cancer.[28] In 2011, the same authors revisited, refined and extended these hallmarks, and gave an overview of the possible therapeutic approach per hallmark.[29] This rapid expansion of the concept of cancer hallmarks illustrate the complexity of cancer and the pace at which progress is made (**Figure 1**).

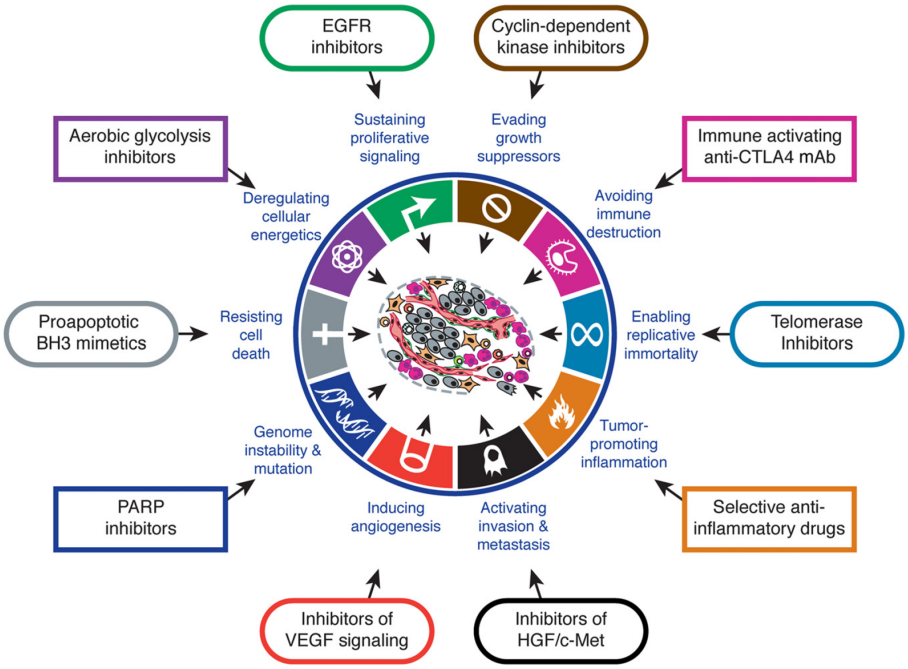


FIGURE 1. THE REVISITED HALLMARKS OF CANCER
(adapted from Weinberg et al.)[28]

Two important categories of genes involved in oncogenesis are oncogenes and tumour suppressor genes. Oncogenes are genes which stimulate the development of cancer while tumour suppressor genes would inhibit the development of cancer. The activation of an oncogene generally requires only one hit while loss of function of a tumour suppressor gene requires inactivation of both alleles of a gene known as the "two hit model".[30] These hits can occur on the level of the epigenome (hypermethylation) or genome (mutations). Mutations can be categorized in small- and large-scale genetic alternations to the nucleotide sequence of the genome. Small-sized genetic aberrations include substitutions, insertions or deletions of one (Single Nucleotide Polymorphism [SNP]) or a few nucleotides (**Figure 2**). Large-scale aberrations include deletions, insertions, duplications, inversions and translocations of (part of) a chromosome arm (**Figure 3**). Mutations can cause a loss- or gain-of-function, can produce a gene product which acts antagonistically to the wild-type allele, can be lethal or can restore or rescue the original phenotype.
Furthermore, mutations can give rise to fusion genes, i.e. hybrid genes formed from parts of two previously separate genes, that may produce fusion transcripts and fusion proteins with altered functionality.[31-33]

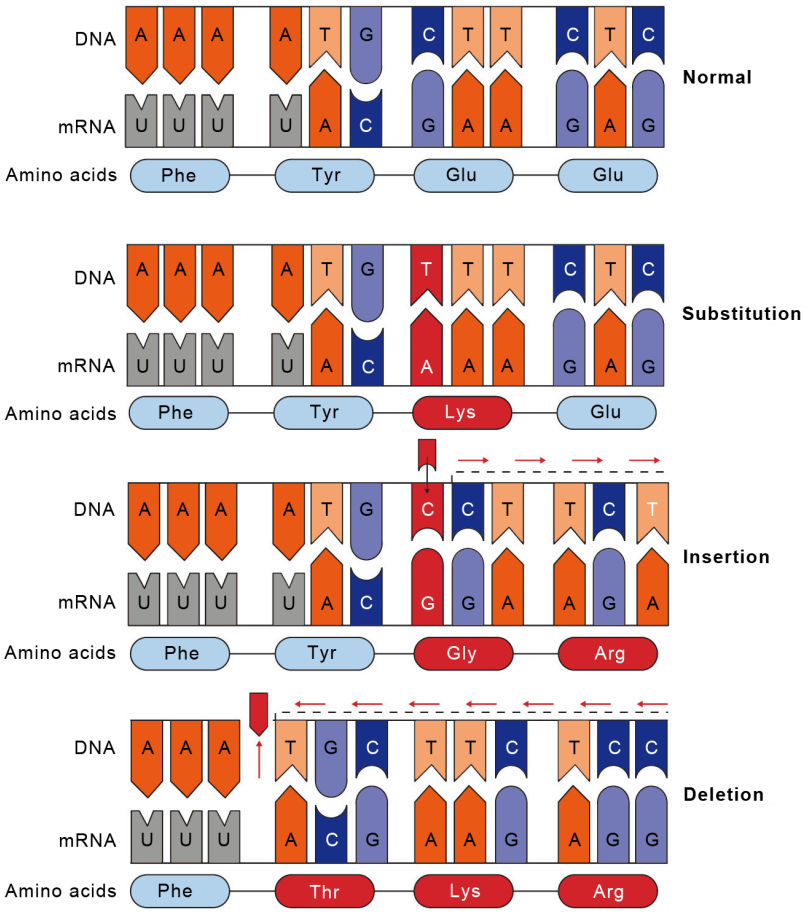


FIGURE 2. SMALL-SCALE MUTATIONS AND SUBSEQUENT CHANGES IN THE AMINO ACID SEQUENCE

In colorectal cancer, the majority of the cancers develop through the classical 'adenoma-carcinoma sequence' which was first described by Fearon and Vogelstein (**Figure 4**).[34] These cancers derive from adenomatous polyps which arise from normal the colon epithelial layer due to cellular inactivating mutations of the *APC* tumour suppressor gene which leads to activation of the Wnt signalling pathway. In the classical model, this early event is followed by activating mutations of the *KRAS* oncogene, downregulation of *SMAD4* tumour suppressor gene through either loss of chromosome 18q or inactivating mutations, and inactivating mutations of the *TP53* tumour suppressor gene.[35-37] These cancers often display chromosomal instability, which is characterized by large-scale mutations giving rise to DNA copy number

variations and/or aneuploidy, meaning cells containing an abnormal number of (part of) the affected chromosomes.[38] Mechanisms behind this type of genomic instability involve defects in chromosomal segregation, centromere and telomere dysfunction, and deficiencies in DNA damage response.[39-44]

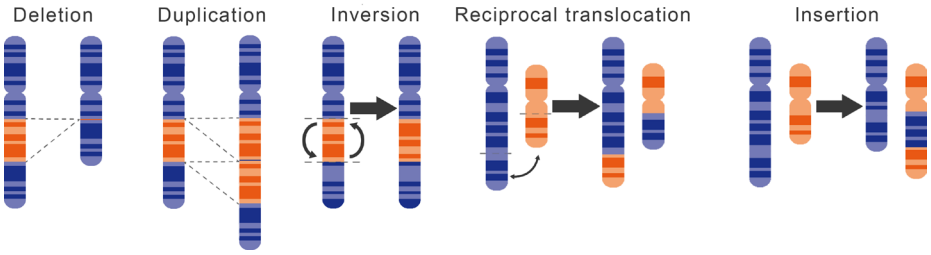


FIGURE 3. LARGE-SCALE MUTATIONS

A minority of the colorectal cancers (10-15%) develop through alternative genetic and epigenetic alterations. Instead of chromosomal instability, these tumours display microsatellite instability (MSI), a result from inactivating mutations or promoter hypermethylation of DNA mismatch repair genes. [46, 47] Due to inactivation of these mismatch repair genes, tumours accumulate many small deletion or insertion mutations in microsatellites, elements of 1-6 nucleotides located throughout the genome. The repetitive nature of these sequences makes them vulnerable to these types of replication errors, which are normally repaired by the mismatch repair system. Tumours with MSI are mostly diploid in contrast to tumours with chromosomal instability, but accumulate 10-100 times more mutations.[48] MSI is observed in 12% of all sporadic colorectal cancers and virtually in all patients with Lynch syndrome, one of the two types of hereditary colorectal cancer.[49]

Tumours displaying MSI mostly contain epigenetic and genetic changes in the Wnt signalling pathway other than alterations in the *APC* gene.[50] Activating *BRAF* oncogene mutations are often and almost exclusively restricted to sporadic microsatellite instable colorectal cancers, although *KRAS* mutations do occur in a minority of the cases.[51] Importantly, mutations in the *KRAS* and *BRAF* gene are mutually exclusive.[52] Lastly, mutations in one or more genes such as *TGFBR2*, *IGF2R* and *BAX* provide a *TP53*-independent mechanism of progression to carcinoma in these tumours (**Figure 4**).[45] Colorectal cancers are also instable at an epigenomic level, which can be global hypomethylation or widespread CpG island hypermethylation at specific gene promoters called the CpG Island Methylator Phenotype (CIMP).[53] (Hyper)methylation

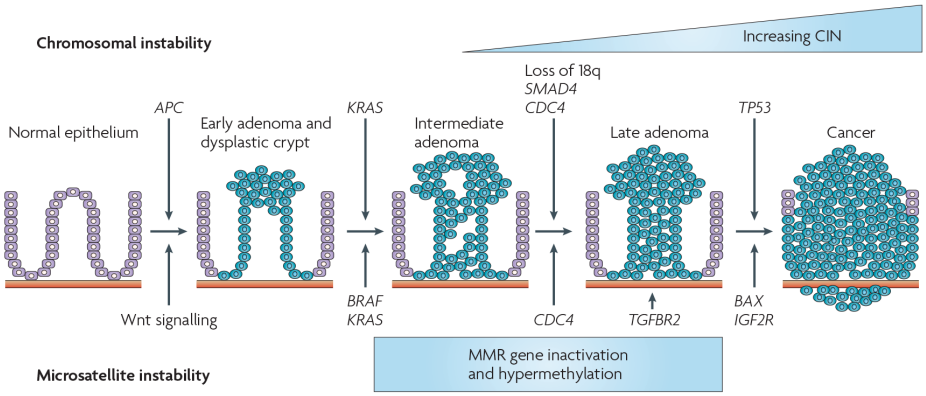


FIGURE 4. THE ADENOMA-CARCINOMA SEQUENCE
(adapted from Walther et al.)[45]

at these promoter regions leads to gene inactivation, which is thought to be the initiating event in these colorectal cancers. This phenomenon is present in almost all tumours with aberrant methylation of *MLH1*, the key mismatch repair gene in sporadic microsatellite instable colorectal cancers.[54]

CONSENSUS MOLECULAR SUBTYPES

In 2015, an international consortium published "The consensus molecular subtypes of colorectal cancer".[55] The consortium used a total of 18 CRC gene expression-based data sets (n=4,151 patients) to show marked interconnectivity between six independent classification systems which resulted in four subtypes of colorectal cancer (Consensus Molecular Subtype [CMS] 1 to 4). Each subtype is characterized by distinctive biological features (**Figure 5**). Some of these features predominantly occur within one subtype such as MSI while other features such as *KRAS* mutations are less linked to a single subtype (**Figure 5**).

In an aggregated survival analysis which included patients with stage I-IV colorectal cancer who underwent divergent treatments, patients with a CMS4 tumour had a worse relapse-free and overall survival compared to patients with a CMS1-3 tumour (**Figure 6**). However, due to the above-mentioned heterogeneity the relevance of the CMS classification regarding clinical outcome and prediction of response to therapy in clinically relevant subgroups needs to be refined.

CMS1 MSI immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high, hypermutation	SCNA high	Mixed MSI status, SCNA low, CIMP low	SCNA high
BRAF mutations		*KRAS* mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration, TGF-β activation, angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

FIGURE 5. THE BIOLOGICAL FEATURES ASSOCIATED WITH THE DIFFERENT CMS GROUPS
CIMP=CpG island methylator phenotype; MSI=microsatellite instability; SCNA=somatic copy number alterations. (adapted from Guinney et al.)[55]

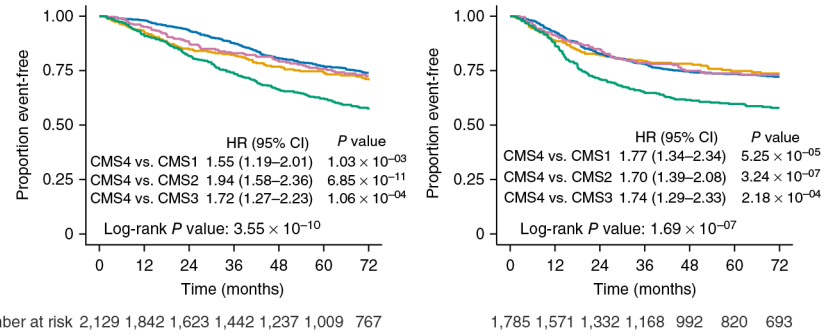


FIGURE 6. CMS1 (YELLOW), CMS2 (BLUE), CMS3 (PINK) AND CMS4 (GREEN) WITH KAPLAN-MEIER SURVIVAL ANALYSIS IN THE AGGREGATED COHORT FOR OVERALL SURVIVAL (N = 2,129) (LEFT), RELAPSE-FREE SURVIVAL (N = 1,785) (RIGHT)
On the x axis is the time in months and on the y axis the portion of patients without an event (death [left] of relapse of disease [right]). (adapted from Guinney et al.)[55]

BIOMARKERS IN CLINICAL PRACTICE

Currently, MSI is the only molecular biomarker used in stage I-III colorectal cancer to guide clinical decision making and inform the patient on the prognosis of the disease.[11] Patients with a microsatellite instable tumour have a better prognosis compared to patients with a microsatellite stable tumour.[56] Furthermore, patients with microsatellite instable stage II-III colon cancer have very little to no benefit from chemotherapy.[57, 58] In contrast, recent studies have shown very promising results for immunotherapy in patients with treatment-refractory progressive metastatic microsatellite instable colorectal cancer.[59]

OUTLINE OF THIS THESIS

As mentioned, 15-20% of patients with stage II colon cancer will develop recurrence of disease after curative surgery. Against the background of the information described above, the aim of this thesis was to identify biomarkers in colon cancer tissue at any molecular level that may help to identify these patients, thereby enabling a more personalized treatment.

The quality of tissue sampling and tissue storage is pivotal for successful translational research. In **Chapter 2** we report the quality of a random set of fresh frozen samples from the MATCH study. The excellent preservation of RNA and DNA in fresh frozen tissue samples allows for high-throughput methods such as RNA sequencing to be used to reveal the presence and quantity of RNA in tumour samples. The interpretability of genomics data strongly depends on the bioinformatics pipeline used to process the data. In **Chapter 3** we present a new normalization method 'GeTMM' (Gene length corrected Trimmed Mean of M-values), which generates normalized RNA sequencing data suitable for both between-sample normalization as well as within-sample analyses. We compare GeTMM with existing normalization methods with respect to distributions, effect of RNA quality, subtype-classification, recall of differentially expressed genes and correlation to RT-qPCR data. In **Chapter 4** we investigate the interconnectivity of the tumour stage and tumour biology in (reflected by the Consensus Molecular Subtypes (CMS)) in colorectal cancer, and explore the added value of this knowledge in patients with stage II colon cancer. In **Chapter 5** a systematic analysis of oncogenic fusions is presented, which resulted in the identification of several known and novel fusion genes. We introduced some of these fusion products in cell lines to assess the biological potential of these fusion genes to drive malignant development. In **Chapter 6** we report a study on the prognostic value of the Spleen Tyrosine Kinase (*SYK*) gene, which has been posed as a marker for predicting both poor and favourable outcome in various epithelial malignancies including colorectal cancer. In **Chapter 7** the validation of a metastasis-specific microRNA signature and the underlying biology of these microRNAs is described. In **Chapter 8 and 9** the design, proceedings, governance, opportunities, and pitfalls of three nationwide cohort studies designed to facilitate research by generating and sharing standardized high quality data is presented.

REFERENCES

1. Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin 2017;67:177-93.
2. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer 2013;49:1374-403.
3. Cijfers over kanker. 2016. (Accessed 1th November at www.cijfersoverkanker.nl.)
4. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med 2003;348:919-32.
5. Jasperson KW, Tuohy TM, Neklason DW, et al. Hereditary and familial colon cancer. Gastroenterology 2010;138:2044-58.
6. Sobin LH, Fleming ID. TNM Classification of Malignant Tumors, fifth edition (1997). Union Internationale Contre le Cancer and the American Joint Committee on Cancer. Cancer 1997;80:1803-4.
7. van Gijn W, Marijnen CA, Nagtegaal ID, et al. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer: 12-year follow-up of the multicentre, randomised controlled TME trial. Lancet Oncol 2011;12:575-82.
8. Moertel CG, Fleming TR, Macdonald JS, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. Ann Intern Med 1995;122:321-6.
9. Benson AB, 3rd, Schrag D, Somerfield MR, et al. American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. J Clin Oncol 2004;22:3408-19.
10. Labianca R, Nordlinger B, Beretta GD, et al. Early colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2013;24 Suppl 6:vi64-72.
11. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249-64.
12. Andre T, de Gramont A, Vernerey D, et al. Adjuvant Fluorouracil, Leucovorin, and Oxaliplatin in Stage II to III Colon Cancer: Updated 10-Year Survival and Outcomes According to BRAF Mutation and Mismatch Repair Status of the MOSAIC Study. J Clin Oncol 2015;33:4176-87.
13. Buttner S, Lalmahomed ZS, Coebergh van den Braak RR, et al. Completeness of pathology reports in stage II colorectal cancer. Acta Chir Belg 2017;117:181-7.
14. Siriwardena AK, Mason JM, Mullamitha S, et al. Management of colorectal cancer presenting with synchronous liver metastases. Nat Rev Clin Oncol 2014;11:446-59.
15. Elferink MA, de Jong KP, Klaase JM, et al. Metachronous metastases from colorectal cancer: a population-based study in North-East Netherlands. Int J Colorectal Dis 2015;30:205-12.
16. Giannakis M, Ng K. To Treat or Not to Treat: Adjuvant Therapy for Stage II Colon Cancer in the Era of Precision Oncology. J Oncol Pract 2017;13:242-4.
17. Quasar Collaborative G, Gray R, Barnwell J, et al. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. Lancet 2007;370:2020-9.
18. National Institutes of Health DoHHS. Help me Understand Genetics. United States of America: U.S. National Library of Medicine; 2017.
19. Consortium EP, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799-816.

20. Nugent CI, Lundblad V. The telomerase reverse transcriptase: components and regulation. Genes Dev 1998;12:1073-85.

21. Pidoux AL, Allshire RC. The role of heterochromatin in centromere function. Philos Trans R Soc Lond B Biol Sci 2005;360:569-79.

22. Barrero MJ, Boue S, Izpisua Belmonte JC. Epigenetic mechanisms that regulate cell identity. Cell Stem Cell 2010;7:565-70.

23. Roostaee A, Benoit YD, Boudjadi S, et al. Epigenetics in Intestinal Epithelial Cell Renewal. J Cell Physiol 2016;231:2361-7.

24. Moosavi A, Motevalizadeh Ardekani A. Role of Epigenetics in Biology and Human Diseases. Iran Biomed J 2016;20:246-58.

25. Ambros V. The functions of animal microRNAs. Nature 2004;431:350-5.

26. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281-97.

27. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347:1260419.

28. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57-70.

29. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646-74.

30. Knudson AG, Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 1971;68:820-3.

31. Kumar-Sinha C, Kalyana-Sundaram S, Chinnaiyan AM. Landscape of gene fusions in epithelial cancers: seq and ye shall find. Genome Med 2015;7:129.

32. Mertens F, Johansson B, Fioretos T, et al. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 2015;15:371-81.

33. Stransky N, Cerami E, Schalm S, et al. The landscape of kinase fusions in cancer. Nat Commun 2014;5:4846.

34. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759-67.

35. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell 1996;87:159-70.

36. Fearon ER. Molecular genetics of colorectal cancer. Annu Rev Pathol 2011;6:479-507.

37. Fumagalli A, Drost J, Suijkerbuijk SJ, et al. Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. Proc Natl Acad Sci U S A 2017;114:E2357-E64.

38. Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. Nature 1997;386:623-7.

39. Bardelli A, Cahill DP, Lederer G, et al. Carcinogen-specific induction of genetic instability. Proc Natl Acad Sci U S A 2001;98:5770-5.

40. Ganem NJ, Godinho SA, Pellman D. A mechanism linking extra centrosomes to chromosomal instability. Nature 2009;460:278-82.

41. Tatsumoto N, Hiyama E, Murakami Y, et al. High telomerase activity is an independent prognostic indicator of poor outcome in colorectal cancer. Clin Cancer Res 2000;6:2696-701.

42. Engelhardt M, Drullinsky P, Guillem J, et al. Telomerase and telomere length in the development and progression of premalignant lesions to colorectal cancer. Clin Cancer Res 1997;3:1931-41.

43. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396:643-9.

44. Chae YK, Anker JF, Carneiro BA, et al. Genomic landscape of DNA repair genes in cancer. Oncotarget 2016;7:23312-21.

45. Walther A, Johnstone E, Swanton C, et al. Genetic prognostic and predictive markers in colorectal cancer. Nat Rev Cancer 2009;9:489-99.

46. Bettington M, Walker N, Clouston A, et al. The serrated pathway to colorectal carcinoma: current concepts and challenges. Histopathology 2013;62:367-86.

47. Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. Science 1993;260:816-9.

48. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-7.

49. Boland CR, Thibodeau SN, Hamilton SR, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248-57.

50. Krausova M, Korinek V. Wnt signaling in adult intestinal stem cells and cancer. Cell Signal 2014;26:570-9.

51. Parsons MT, Buchanan DD, Thompson B, et al. Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification. J Med Genet 2012;49:151-7.

52. Rajagopalan H, Bardelli A, Lengauer C, et al. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature 2002;418:934.

53. Toyota M, Ahuja N, Ohe-Toyota M, et al. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A 1999;96:8681-6.

54. Samowitz WS. The CpG island methylator phenotype in colorectal cancer. J Mol Diagn 2007;9:281-3.

55. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

56. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 2005;23:609-18.

57. Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol 2010;28:3219-26.

58. Des Guetz G, Schischmanoff O, Nicolas P, et al. Does microsatellite instability predict the efficacy of adjuvant chemotherapy in colorectal cancer? A systematic review with meta-analysis. Eur J Cancer 2009;45:1890-6.

59. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med 2015;372:2509-20.

Z.S. Lalmahomed, R.R.J. Coebergh van den Braak, M.H.A. Oomen, S.P. Arshad, P.H.J. Riegman, J.N.M. IJzermans, on behalf of the MATCH study group.

2

MULTICENTER FRESH FROZEN TISSUE SAMPLING IN COLORECTAL CANCER: DOES THE QUALITY MEET THE STANDARDS FOR STATE OF THE ART BIOMARKER RESEARCH?

CELL AND TISSUE BANKING 2017

ABSTRACT

The growing interest in the molecular subclassification of colorectal cancers is increasingly facilitated by large multicenter biobanking initiatives. The quality of tissue sampling is pivotal for successful translational research. This study shows the quality of fresh frozen tissue sampling within a multicenter cohort study for colorectal cancer (CRC) patients. Each of the seven participating hospitals randomly contributed ten tissue samples, which were collected following Standard Operating Procedures (SOP) using established techniques. To indicate if the amount of intact RNA is sufficient for molecular discovery research and prove SOP compliance, the RNA integrity number (RIN) was determined. Samples with a RIN < 6 were measured a second time and when consistently low a third time. The highest RIN was used for further analysis. 91% of the tissue samples had a RIN ≥ 6 (91%). The remaining six samples had a RIN between 5 and 6 (4.5%) or lower than 5 (4.5%). The median overall RIN was 7.3 (range 2.9–9.0). The median RIN of samples in the university hospital homing the biobank was 7.7 and the median RIN for the teaching hospitals was 7.3, ranging from 6.5 to 7.8. No differences were found in the outcome of different hospitals ($p = 0.39$). This study shows that the collection of high quality fresh frozen samples of colorectal cancers is feasible in a multicenter design with complete SOP adherence. Thus, using basic sampling techniques large patient cohorts can be organized for predictive and prognostic (bio)marker research for CRC.

INTRODUCTION

Colorectal cancer (CRC) is the second most common malignancy in the Western World.[1] As in all cancer research, there is a strong trend towards molecular subclassification of CRC.[2] The studies conducted to identify these molecular and clinically relevant markers demand large numbers of patients with accurate long-term clinical data combined with high quality tissue samples to be able to use state of the art techniques.[3,4] Subsequently, the standard enclosed formalin-fixed paraffin-embedded tissue can be used to develop assays for daily clinical practice. Therefore, large multicenter biobanking initiatives are needed to facilitate these research efforts.[5,6] However, 10% of the fresh frozen tissue samples collected for research purposes are unsuitable for molecular analyses. This is due to multiple non-modifiable factors such as tissue type, intrinsic patient factors, warm ischemia time (extraction of the resection specimen after ligation of the large vessels) and modifiable factors such as cold ischemia time (tissue transport from the operating theatre to the pathology lab), the conservation (fixation/stabilization) method, subsequent transport and the storage of the tissue samples.[7,8] The RNA Integrity Number (RIN), first described in 2006, is currently a common standard used to assess tissue quality.[9] This method became well accepted to measure the SOP adherence of quality in tissue banking.[10]

The current study assessed the tissue quality of the MATCH study, a multicenter cohort study in the region of Rotterdam, the Netherlands, enrolling patients with CRC and obtaining fresh frozen tissue samples in one university hospital with experience in tissue sampling and storage by dedicated personnel, and in six non-university teaching hospitals that are not used to nor standardly equipped and staffed for routine fresh frozen tissue sampling.

MATERIAL AND METHODS

MATCH STUDY DESIGN

The MATCH study is an ongoing multicenter cohort study including adult patients with CRC undergoing curative surgery. The participating centers include one university hospital (Erasmus University Medical Center) and six non-university teaching hospitals (Elisabeth-Tweesteden hospital, IJsselland hospital, Ikazia hospital, Maasstad hospital, Reinier de Graaf Hospital, Franciscus Gasthuis). The MATCH study was approved by the Medical Ethical Board of the Erasmus University Medical Center, Rotterdam, the Netherlands (MEC-2007-088). All patients provide written informed consent for the collection of longterm clinical data and storage of tissue samples. The study is an

integrated approach using clinical patient care in non-university hospitals with university-based facilities for tissue and data storage. The rationale of this study was to identify subtypes of colorectal cancer, related prognostic markers and outcome of treatment. Liver metastases was defined as primary outcome defining a good or dismal outcome of disease progression as liver involvement has been demonstrated to be the main factor to determine long term outcome.

CLINICAL DATA

Medical specialists of departments of Surgery, Pathology, Gastroenterology, Radiology and Medical oncology were consulted. Clinical data included reports of colonoscopy, radiology and pathology, as well as surgical reports and postoperative complications. A standard case record was created in a web based multicenter access database. The follow-up of these patients was standardized in all hospitals following an intensive follow-up schedule according the national CRC guidelines.[11]

TISSUE SAMPLING

All tissue samples were handled following a Standard Operation Procedure (SOP) provided by the study team at the start of the study. In short, resection specimens were transported (at room temperature without any conservation fluids) from the operating theatre to the pathology department, immediately following removal of the specimen from the patient. At the pathology department the specimen was handled at room temperature and within two hours after resection samples were snap-frozen as described below. When the 2 h time limit was exceeded, no tissue samples were taken.

Macroscopically, one to four tumor samples and one to two healthy colon tissue samples of 0.5–1 cm^3 were taken by the pathologist. Tissue sampling for the MATCH study was not allowed to interfere with the standard pathology routine needed for clinical practice. Tumor and normal tissue were stored in labeled cryovials and snap frozen in liquid nitrogen or dry-ice.[12] Samples were then stored at lowtemperature refrigerators (-80˚C) in the hospital of primary surgery and in batches transported to the central tissue bank (-196 ˚C liquid nitrogen barrels) at the university hospital. Of all new tissue specimens stored in the central bank, on a yearly base 2% is tested for quality, by determining the RNA integrity.[10,13]

TISSUE QUALITY ASSESSMENT

To assess the tissue quality of the samples collected in the MATCH study, we randomly selected 10 tissue samples per participating hospital, representing about 4% of the entire collection. Samples that were exposed to neoadjuvant chemotherapy and/or

radiotherapy were excluded as this may damage tissue resulting in failure of analysis. RNA quality was determined by measuring of the RIN.[9,14] For RNA isolation, 10–20 tissue slides of 10 lm were cut. One slide was colored by hematoxylin and eosin (H&E) stain for morphological confirmation of the diagnosis. For RNA extraction, the slides were put in a Qiazol Lysis buffer and shaken for ten seconds to homogenize the tissue. RNA was then extracted using the miRNeasy Mini Kit (Qiagen, Hilden, Germany) according to the method suggested by the manufacturer. The integrity of RNA was measured by the Bioanalyser (Agilent Technologies, Santa Clara, CA, USA) using the lab-on-a-chip, RNA 6000 nano assay. This is an automated system based on electrophoretic separation. The RIN is directly calculated by applying an algorithm on the ratio of 18S/ 28S ribosomal RNA bands. A tissue sample with a RIN of ≥ 6 is believed to be of good quality (**Figure 1a**).[15] Samples with a RIN < 6 (**Figure 1b**) were measured a second and if consistently low a third time. When the RIN was still low, the case was discussed with the technician to see if any deviation from protocol (e.g. during the freezing procedure or sample preparation) could explain the low RIN. When samples were measured multiple times, the highest RIN was used for further analysis.

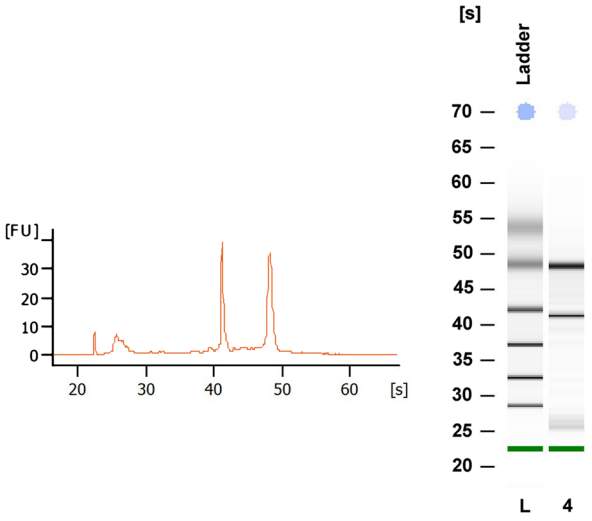


FIGURE 1A. IMAGE INTACT RNA (RIN 9.0), OBTAINED FROM THE ELECTROPHEROGRAM AND VIRTUAL GEL

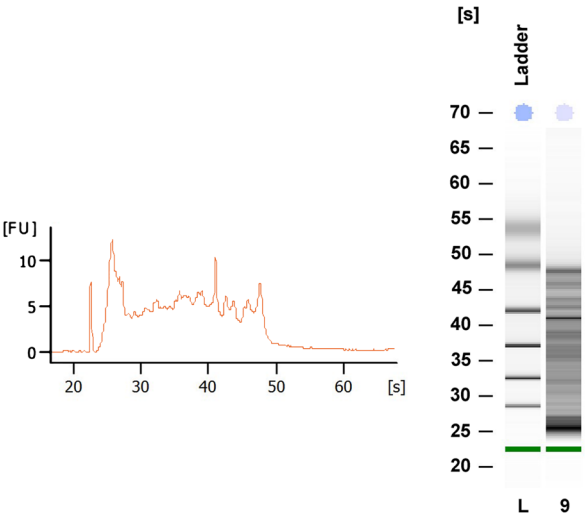


FIGURE 1B. IMAGE PARTIALLY DEGRADED RNA (RIN 3.3), OBTAINED FROM THE ELECTROPHEROGRAM AND VIRTUAL GEL

STATISTICAL ANALYSIS

Statistical analyses were performed using SPSS (IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.). Categorical data were described as frequencies with percentages and continuous data as median with the range. The Chi square test was used to compare categorical data, for continuous data the One-way ANOVA test was used. A p value less than 0.05 was considered to be statistically significant.

RESULTS

In total, 70 random samples were selected for analysis out of the 1700 samples collected in the study period 1st October 2007–1st January 2013. During the workup and data quality check, three samples were excluded leaving a total sample size of n = 67. Two tissue samples were exposed to neoadjuvant radiation therapy and one tissue sample was too small.

Out of the 67 samples, two samples were analyzed two times (3.0%) and seven samples three times (10.4%). The median overall RIN of all samples was 7.3 (range 2.9–9.0). The majority (n = 61) of the tissue samples had a RIN ≥ 6 (91%). The remaining six samples had a RIN between 5 and 6 (4.5%) or lower than 5 (4.5%) (**Figures 2 and 3**).

Three of the seven samples that were measured three times had a RIN < 5 and were discussed with the technician. However, the low RIN could not be attributed to protocol deviations. The median RIN for a center specialized in tissue sampling (university hospital) was 7.7 and the median RIN for teaching hospitals without a wide experience in this field ranged from 6.5 to 7.8 (**Table 1**). The overall median RIN of the non-university teaching hospitals (median RIN = 7.3) did not differ significantly with the median RIN of the university hospital (p = 0.39) (**Figure 4**). When using the specialized university hospital as a reference, the median RIN of one non specialized teaching hospital (hospital 6) had a significantly lower median RIN than the university hospital (p = 0.02). However, a median RIN of 6.5 is still well above the cut-off of 6. Interestingly, the range of RIN for the non-university teaching hospitals tended to be larger than the range of RIN if the university hospital (**Figure 3**).

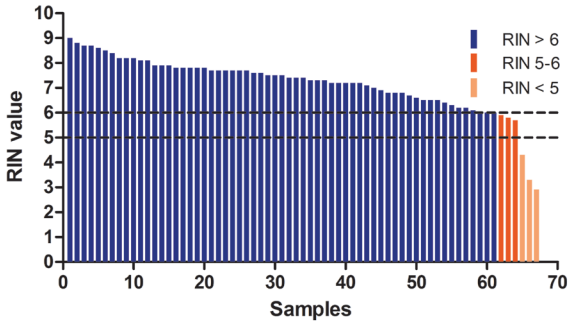


FIGURE 2. THE RIN DISTRIBUTION IN 67 SAMPLES

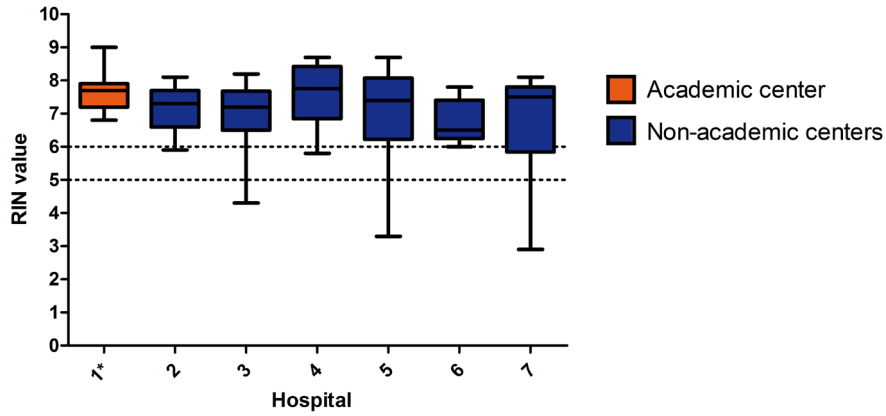


FIGURE 3. BOX PLOT WITH THE RIN PER HOSPITAL

TABLE 1. MEDIAN RNA INTEGRITY NUMBER PER HOSPITAL

Hospital	Number of samples	Median RIN	Range	P-value
1: University hospital	10	7.7	6.8 - 9	0.391
2	9	7.3	5.9 - 8.1	
3	10	7.2	4.3 – 8.2	
4	10	7.8	5.8 – 8.7	
5	10	7.4	3.3 – 8.7	
6	9	6.5	6 – 7.8	
7	9	7.5	2.9 -8.1	
All Samples	67	7.3	2.9-9	

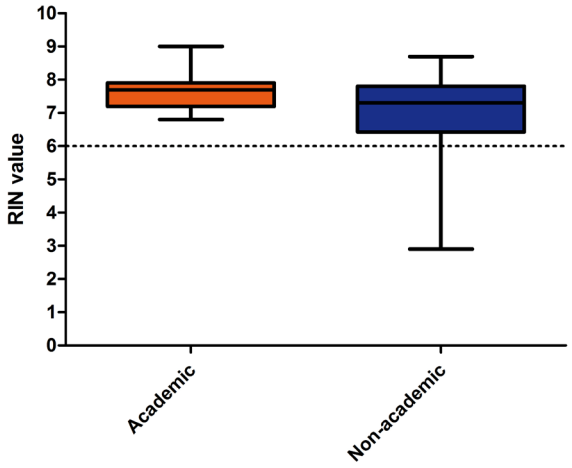


FIGURE 4. BOX PLOT WITH THE RIN FOR THE UNIVERSITY HOSPITAL AND NON-UNIVERSITY HOSPITALS

DISCUSSION

This study shows that the collection of high quality fresh frozen samples of CRC is feasible in a multicenter design including hospitals for which fresh frozen tissue sampling is not part of the daily routine. In our study, 91% had a RIN ≥ 6 and thus can be used for highly demanding gene array assays.

The RIN was developed and published in 2006 to meet the need for a reliable standard to estimate the integrity of RNA samples.[9] A comparison study comparing a subjective evaluation of the electropherogram, the 28S–18S peaks ratio and the RIN showed a superior result for the manual and RIN method over the ratio method.[15] Nowadays, the RIN is widely used to quantify the RNA quality of samples and select samples for expression analyses. However, the cut-off used to select 'high quality' samples varies in literature, ranging from a RIN of 5–7. These cut-offs can be based on the recommendations in a manufacturer manual or on the experience of a lab.[16-19] At our hospital, we use a RIN of ≥ 6 as the cut-off which qualified 91% of the samples as high quality samples. When samples repeatedly have a RIN < 6, they may be excluded to prevent a transcript specific bias, or analytical or bioinformatics steps specifically dealing with the low quality samples should be included in the methodology.[20,21] Furthermore, samples with a RIN < 6 can still be used for RT-qPCR applications in which only short amplicons are analyzed.

The quality of RNA expression in tissue samples is dependent on multiple factors such as tissue type, intrinsic patient factors, warm and cold ischemia time, the fixation method and the storage of the tissue samples. While tissue type and intrinsic patient factors cannot be modified, other factors (i.e. ischemia time, fixation method and the storage of samples) can be influenced. The RIN can be used to determine large influences during the pre-analytical phase. Smaller differences can be assessed based on RNA expression analyses.[22] For fresh frozen samples, the most important factor appears to be the ischemia time and freeze thawing effects after freezing. A recent review specifically addressing the effect of cold ischemia on RNA stability concluded that in most studies only minimal changes in the RIN were observed (≤10%) during a cold ischemia times of 1–6 h.[23] One outlier reported a significantly decreased RIN of 44% in samples with a cold ischemia time of 1.5 h compared to samples with a cold ischemia time of 10 min.[18] However, the 28S:18S ratios did not significantly differ.[18] Importantly, the definition of cold ischemia time differed between studies and often the cold ischemia time in the operating theatre was not taken into account. Furthermore, the effects of warm ischemia time are often ignored while they most likely interact with the effects of cold ischemia time. This may be explained by the fact that this factor is hard to reliably score and is considered to be a non-modifiable

factor since attempts to minimize warm ischemia time may affect patient care. Such nonmodifiable influences can only be documented to obtain a tool for determination of this influence.[24] Although we did not specifically assessed the association between ischemia time and the RIN in our study, the maximum cold ischemia time was 2 h since this was included in the SOP. Thus, the high percentage of high quality samples in our study is in line with the current literature. For the few samples with consistently low RIN values, no protocol deviations were found suggesting the low RIN was caused by non-modifiable factors.

Our study shows that SOP compliance was positive in all the cooperating hospitals and high quality fresh frozen tissue sampling is possible in a multicenter setting including both university and non-university hospitals. These findings support the feasibility of emerging large-scale 'fit-for-purpose' biobanks to facilitate the increasingly complex field of fundamental and translational cancer research.[5,6,25]

In conclusion, our study shows that the collection of high quality fresh frozen samples of CRC is feasible in a multicenter design and using basic sampling techniques. Thus, large patient cohorts can be organized for predictive and prognostic (bio)marker research for CRC.

REFERENCES

1. DeSantis CE, Lin CC, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2014. CA Cancer J Clin 2014;64:252-71.

2. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

3. Riegman PH, Bosch AL, Consortium OT. OECI TuBaFrost tumor biobanking. Tumori 2008;94:160-3.

4. Riegman PH, Dinjens WN, Oosterhuis JW. Biobanking for interdisciplinary clinical research. Pathobiology 2007;74:239-44.

5. Burbach JP, Kurk SA, Coebergh van den Braak RR, et al. Prospective Dutch colorectal cancer cohort: an infrastructure for long-term observational, prognostic, predictive and (randomized) intervention research. Acta Oncol 2016;55:1273-80.

6. Rose S. Huge Data-Sharing Project Launched. Cancer Discov 2016;6:4-5.

7. Boudou-Rouquette P, Touibi N, Boelle PY, Tiret E, Flejou JF, Wendum D. Imprint cytology in tumor tissue bank quality control: an efficient method to evaluate tumor necrosis and to detect samples without tumor cells. Virchows Arch 2010;456:443-7.

8. Qualman SJ, France M, Grizzle WE, et al. Establishing a tumour bank: banking, informatics and ethics. Br J Cancer 2004;90:1115-9.

9. Schroeder A, Mueller O, Stocker S, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 2006;7:3.

10. Morente MM, Mager R, Alonso S, et al. TuBaFrost 2: Standardising tissue collection and quality control procedures for a European virtual frozen tissue bank network. Eur J Cancer 2006;42:2684-91.

11. Lochhead P, Kuchiba A, Imamura Y, et al. Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication. J Natl Cancer Inst 2013;105:1151-6.

12. Mager SR, Oomen MH, Morente MM, et al. Standard operating procedure for the collection of fresh frozen tissue samples. Eur J Cancer 2007;43:828-34.

13. Chi Y, Zhou D. MicroRNAs in colorectal carcinoma--from pathogenesis to therapy. J Exp Clin Cancer Res 2016;35:43.

14. Schisterman EF, Faraggi D, Reiser B, Hu J. Youden Index and the optimal threshold for markers with mass at zero. Stat Med 2008;27:297-315.

15. Strand C, Enell J, Hedenfalk I, Ferno M. RNA quality in frozen breast cancer samples and the influence on gene expression analysis--a comparison of three evaluation methods using microcapillary electrophoresis traces. BMC Mol Biol 2007;8:38.

16. Asterand. RNA quality assurance using RIN (Internet) Detroit, MI: Asterand plc; 2006 (cited 2010 oct 3) Available from: http://wwwasterandcom/asterand/human_tissues/Asterand_RINpdf 2006.

17. Bao WG, Zhang X, Zhang JG, et al. Biobanking of fresh-frozen human colon tissues: impact of tissue ex-vivo ischemia times and storage periods on RNA quality. Ann Surg Oncol 2013;20:1737-44.

18. Hong SH, Baek HA, Jang KY, et al. Effects of delay in the snap freezing of colorectal cancer tissues on the quality of DNA and RNA. J Korean Soc Coloproctol 2010;26:316-23.

19. Viana CR, Neto CS, Kerr LM, et al. The interference of cold ischemia time in the quality of total RNA from frozen tumor samples. Cell Tissue Bank 2013;14:167-73.

20. Lauss M, Vierlinger K, Weinhaeusel A, Szameit S, Kaserer K, Noehammer C. Comparison of RNA amplification techniques meeting the demands for the expression profiling of clinical cancer samples. Virchows Arch 2007;451:1019-29.

21. Viljoen KS, Blackburn JM. Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity. BMC Genomics 2013;14:14.

22. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol 2014;12:42.

23. Grizzle WE, Otali D, Sexton KC, Atherton DS. Effects of Cold Ischemia on Gene Expression: A Review and Commentary. Biopreserv Biobank 2016.

24. Riegman PH, de Jong B, Daidone MG, et al. Optimizing sharing of hospital biobank samples. Sci Transl Med 2015;7:297fs31.

25. Kap M, Oomen M, Arshad S, de Jong B, Riegman P. Fit for purpose frozen tissue collections by RNA integrity number-based quality control assurance at the Erasmus MC tissue bank. Biopreserv Biobank 2014;12:81-90.

M. Smid*, R.R.J. Coebergh van den Braak*, H.J.G. van de Werken, J. van Riet, A. van Galen, V. de Weerd, M. van der Vlugt-Daane, S.I. Bril, Z.S. Lalmahomed, W.P. Kloosterman, S.M. Wilting, J.A. Foekens, J.N.M. IJzermans, J.W.M. Martens†, A.M. Sieuwerts†.

* These authors contributed equally; †These authors share senior co-authorship.

3

GENE LENGTH CORRECTED TRIMMED MEAN OF M-VALUES (GETMM) IMPROVES RNA-SEQ DATA PROCESSING FOR INTRA- AND INTERSAMPLE COMPARISONS

SUBMITTED

ABSTRACT

BACKGROUND

Current normalization methods for RNA-sequencing data allow either for intersample comparison to identify differentially expressed (DE) genes or for intrasample comparison for the discovery and validation of gene signatures. Most studies on optimization of normalization methods typically use simulated data to validate methodologies. We describe a new method, GeTMM, which allows for both inter- and intrasample analyses with the same normalized data set. We used actual (i.e. not simulated) RNA-seq data from 263 colon cancers to compare GeTMM with the most commonly used normalization methods (i.e. EdgeR, DESeq2 and TPM) with respect to distributions, effect of RNA quality, subtype-classification, recurrence score, recall of DE genes and correlation to RT-qPCR data.

RESULTS

We observed a clear benefit for GeTMM and TPM with regard to intrasample comparison while GeTMM performed similar to EdgeR and DESeq2 in intersample comparisons. Regarding DE genes, recall was found comparable among the normalization methods, while GeTMM showed the lowest number of false-positive DE genes. Remarkably, we observed limited detrimental effects in samples with low RNA quality.

CONCLUSIONS

We show that GeTMM outperforms established methods with regard to intrasample comparison while perfoming equivalent with regard to intersample normalization using the same normalized data. These combined properties enhance the general usefulness and comparability to public gene expression sources which provides an important advantage over existing normalization methods in the era of data sharing.

BACKGROUND

In recent years, the analysis of the transcriptome has switched from using microarrays to the potentially more powerful and informative massive parallel sequencing of cDNA (RNA-seq).[1] In RNA-seq, sequence reads are aligned to a reference genome, and the number of reads mapping to a feature – such as a gene – is a measure which is proportional to both the length and abundance of the said feature. Before performing downstream analyses, normalization has to be performed to correct for differences between sequencing runs (e.g. library size and relative abundances).

Current normalization methods allow for either inter- or intrasample comparison. The two most commonly used normalization methods when interested in DE genes between samples (intersample comparison) are EdgeR[2] and DESeq[3, 4], which show consistent good performance compared to other methods.[5-8] Notably, these methods do not correct the observed read counts for the gene length, which is theoretically irrelevant for intersample comparisons. However, this approach does not allow for intrasample comparison, because a longer gene will get more read counts compared to a shorter gene when expressed at equal levels. Thus, samples can seem highly correlated without correction when in fact the correlation is much lower after length correction (see **Supplementary Figure 1**), and in extremis can be correlated based on gene length instead of the expression levels. This problem extends to correlation based methods where for example a panel of genes of a sample is correlated to another sample, as is often done in hierarchical clustering (correlation is used as similarity metric). Furthermore, classifiers based on correlation of an established signature gene panel to a new sample such as the consensus molecular (CMS) subtypes in colorectal cancer will yield erroneous results without correcting gene expression levels for gene length.

The most commonly used normalization method that includes gene length correction is TPM (Transcripts-Per kilobase-Million)[9], as other methods like RPKM[1]/FPKM[10] (Reads/Fragments Per Kilobase per Million reads, respectively, proved to be inadequate and biased.[5, 6, 11, 12]

Ideally, a normalization method should generate a data set on which both between-sample and within-sample analyses can be performed. We therefore introduce GeTMM (Gene length corrected, Trimmed Mean of M-values), a novel normalization method combining gene-length correction with the normalization procedure TMM, as implemented in EdgeR, to allow both inter- and intrasample comparison with the same normalized data set. We used true (i.e. not simulated) RNA-seq data of a large cohort of primary tumors of 263 colon cancer patients, and normalized these data using our new method GeTMM, alongside EdgeR, DESeq2 and TPM.[6] We investigated

several properties of the normalized data sets with regard to distribution, effect of RNA quality, subtype-classification (i.e. the CMS classification)[13], a clinical recurrence score[14], recall of DE genes and correlation to RT-qPCR data generated from the same samples. The main objective of this study was to determine if GeTMM performs equivalent to the other normalization methods with regard to intersample analyses, and if and to what extent gene length correction influences intrasample analyses.

METHODS

The main objective of this study was to determine if GeTMM performs equivalent to the other normalization methods with regard to intersample analyses, and if and to what extent gene length correction influences intrasample analyses.

DESCRIPTION OF COHORT

Fresh-frozen tumor tissue of 263 colon cancer patients of the MATCH study, a multicenter observational cohort study, who underwent surgery in one of seven hospitals in the Rotterdam region, the Netherlands, were used. Inclusion criteria and additional clinical characteristics have been described.[15]

RNA ISOLATION, CDNA SYNTHESIS, QPCR AND RNA-SEQ

Detailed description of the RNA-isolation has been described previously[16, 17]; briefly, RNA was isolated from 30 µm sections using RNA-Bee® according to the manufacturer's instructions (Tel-Test inc., USA). Quality and quantity of RNA before and after genomic DNA (gDNA) removal and clean-up with the NucleoSpin RNA II tissue kit (Macherey-Nagel GmbH & Co. KG, Germany) were assessed with the Nanodrop ND-1000 (Thermo Scientific, Wilmington, USA) and the MultiNA Microchip Electrophoresis system (Shimadzu, Kyoto, Japan). RNA Integrity Numbers (RIN) were assessed using the MultiNA Microchip Electrophoresis system after gDNA removal and clean-up (**Supplementary Figure 2** evaluates the relation between Agilent's BioAnalyzer RIN value and the quality as measured by MultiNA). cDNA was generated from 1 µg total RNA with the RevertAid H Minus First Strand cDNA synthesis kit according to the manufacturer's instructions (Fermentas, St Leon-Rot, Germany). RT-qPCR was performed with the Mx3000P QPCR machine (Agilent Technologies, the Netherlands) using ABgene Absolute Universal or Absolute SYBR Green with ROX PCR reaction mixtures (Thermo Scientific, USA) according to the manufacturer's instructions. The intron-spanning assays to quantify levels of 33 transcripts by the delta-delta Cq method were assessed as described before[16, 17] and are summarized in **Supplementary Table 1**.

For RNA-seq, 500 ng of total RNA after gDNA removal, clean-up and removing ribosomal RNA using Ribo Zero (Illumina, USA), was used as input for the Illumina TruSeq stranded RNA-seq protocol (paired-end). Libraries were pooled and sequenced on Illumina HiSeq2500 (2x101bp) or NextSeq (2x76bp) instruments. Pool sizes and the amount of samples per run were determined based on the percentage of tumor cells estimated from histological examination.[15] We used the STAR algorithm[18] (version 2.4.2a) to align the RNA-seq data on the GRCh38 reference genome (settings are listed in the **Supplementary Methods**).
Gene annotation was derived from GENCODE Release 23 (https://www.gencodegenes.org/). To obtain exon specific counts for *CDK1* and *MKI67*, all unique HAVANA exons for each gene were extracted and used in FeatureCounts[19] with the following settings "–t exon", -O and –f. These settings, and the absence of –p (for paired-end counting), ensures that reads that overlap multiple exons are counted for each of these exons. This ensured all evidence for the presence of an exon was counted.

NORMALIZATION OF RNA-SEQ DATA

The raw read-counts of all samples were merged in a single read-count matrix. This matrix was used as input for the different normalization methods. The most commonly used RNA-seq normalization methods are implemented in EdgeR[2] and DESeq2.[3,4] Both these methods do not employ any gene length normalization since their aim is to identify differentially expressed genes between samples and thus assume that the gene length is constant across samples. The TPM method adds to the previously used RPKM - for single-end sequencing protocols - or its paired-end counterpart FPKM. TPM uses a simple normalization scheme, where the raw read counts of each gene are divided by its length in kb (Reads per Kilobase, RPK), and the total sum of RPK is considered the library size of that sample. Next, the library size is divided by a million, and that is used as scaling factor to scale each genes' RPK value. Thus, TPM does correct for gene length, but is lacking a sophisticated between-sample correction; it does not account for a possible small number of highly expressed genes, thus comprising a large portion of the total library size of that sample. DESeq2 and EdgeR address this problem by estimating correction factors that are used to rescale the counts (see reference 2 and 3 for more details). In short, EdgeR employs the Trimmed Means of M values (TMM)[2] in which highly expressed genes and those that have a large variation of expression are excluded, whereupon a weighted average of the subset of genes is used to calculate a normalization factor. DESeq2 also assumes most genes are not differentially expressed; here, for each gene the ratio of its read count in a sample over the geometric mean of that gene in all samples is calculated. The median of the ratios of all genes in a sample is used as

correction factor. Where EdgeR estimates a correction factor that is applied to the library size, the correction factor of DESeq2 is applied to the read counts of the individual genes.

Such normalized data are better comparable between samples, but still suffer from the inability to compare gene expression levels within a sample. To obtain a normalized data set that is equally suitable for between-samples and within-sample analyses, the following GeTMM method is proposed: first, the RPK is calculated for each gene in a sample: raw read counts / length gene (kb). In EdgeR, which uses TMM-normalization, normally the library size (total read count; RC) is corrected by the estimated normalization factor and scaled to per million reads, but in GeTMM the total RC is substituted with the total RPK (**Figure 1**).

$$\text{RPK/EdgeR Scaling} = \frac{\sum_{i}^{n} \frac{\text{raw RC gene i}}{\text{length gene i (kb)}} \times \text{TMM Norm. factor}}{10^6}$$

$$\text{GeTMM Normalized data gene i} = \frac{\text{RPK gene i}}{\text{Scaling}}$$

FIGURE 1. NORMALIZATION USING GETMM METHOD WITH N=NUMBER OF GENES AND I=GIVEN GENE I

Practically, this means calculating the RPK values and using these for input in EdgeR. The gene length is calculated using the annotation by gencode: the length of all exons with a unique exon_id annotated to the same gene_id is summed. DESeq2 only allows integers as input, thus the fractions generated by the gene length correction are rejected for input by DESeq2.

EdgeR and DESeq2 are available as R-packages (https://bioconductor.org/), and subsequent analyses were performed using R (v3.2.2). To obtain normalized data, the raw read count matrix (tab-delimited text file) was used as input. R commands to obtain normalized data are listed in the **Supplementary Methods**. After processing, read counts were log2-transformed (setting genes to NA when having 0 read counts). The CMS classification was performed using the "CMSclassifier" package (https://github.com/Sage-Bionetworks/CMSclassifier), using the single-sample prediction parameter. The Onco*type* DX®[14] recurrence score was performed as described (ref Clark-Lagone) for the RT-qPCR data, and using the RNA-seq normalized values as input for the algorithm.The signal-to-noise ratio was calculated as the (mean1 – mean2)/Sp, where Sp is the square root of the pooled variance Vp. This is calculated as Vp = [(n1 -1) V1 + (n2-1)V2]/(n1+n2-2), where V1 and V2 are the variance for each of the groups, and n1 and n2 the sample group sizes.

STATISTICS

Statistical tests were performed using R (v3.2.2) and are indicated in the main text, p-values were two-sided and p-values and FDRs were considered significant when below 0.05.

RESULTS

We used primary tumor tissue of a cohort of 263 colon cancer patients to generate RNA-seq data. We aligned these data to the human reference genome (GRCh38) and generated read counts per gene. This read count matrix was used for several normalization procedures: EdgeR[2], DESeq (version 2)[3] and TPM, in addition to a newly proposed method of gene length correction in combination with the normalization used by EdgeR - GeTMM. To validate the results, the same RNA used for generating the sequence libraries was also used for RT-qPCR analysis of 33 genes (see **Supplementary Table 1** for details).

DISTRIBUTION OF RNA-SEQ DATA

The library sizes (i.e. the number of mapped reads) of the samples ranged from 5.8 to 37.8 million (mean 16.0 million and median 14.2 million). Density plots were generated to get an overview of the read count distributions (**Figure 2**). Panel 2A shows the raw read counts (not normalized), which clearly shows a bimodal distribution after the initial peak at 0, with peaks at 1.1~1.4 log2-read counts and a broader peak at 7~10 log2-read counts. Similar bimodal distributions were seen after normalization by DESeq2 and EdgeR (**Figure 2B, 2C**), which both do not correct for gene length. Splitting the EdgeR normalized data by genes < 5 kb and those ≥ 5 kb (**Figure 2D**) shows that the bimodality is largely attributable to the gene length; as expected, longer genes generally have higher read counts. Methods employing correction for gene length - TPM and GeTMM - both show a more Gaussian distribution (**Figure 2E, 2F**).

COMPARISON TO RT-QPCR GENERATED DATA: INTERSAMPLE ANALYSIS

To evaluate how the different normalization methods affect downstream analysis, we measured the expression levels of 33 genes (of which 3 reference genes - *HMBS*, *HPRT1* and *TBP*) using RT-qPCR in the same RNA isolate as was used for sequencing. The RT-qPCR data were normalized using the reference genes and were considered as the gold standard to compare against. To assess the effect of the different normalization methods on intersample analysis, we correlated the normalized RNA-

seq data of the 30 genes to the RT-qPCR levels over all samples (**Figure 3, Supplementary Table 2** and **Supplementary Figure 3** for a detailed example). Overall, correlation coefficients for GeTMM were very comparable to the correlation

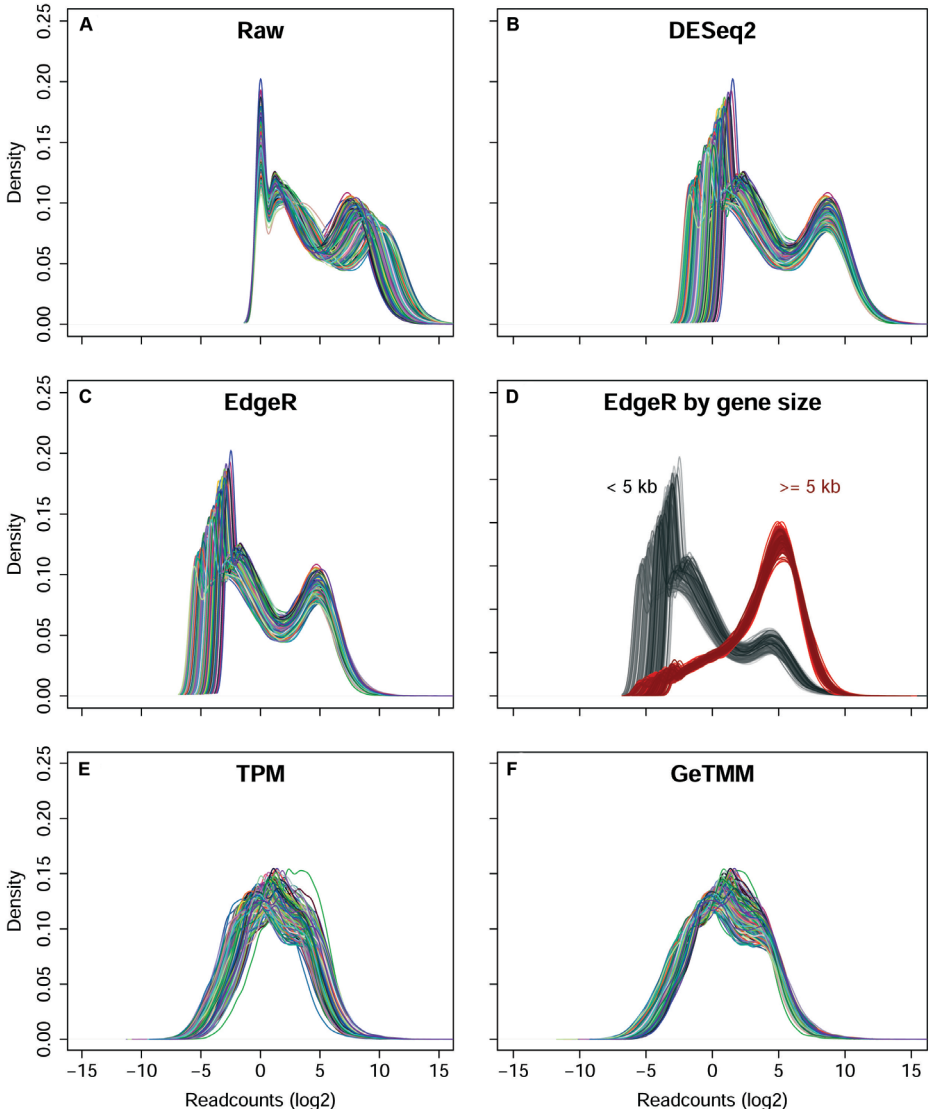


FIGURE 2. DENSITY PLOT BY NORMALIZATION METHOD

Each line corresponds to the distribution of expression levels in a sample. X-axis shows log2 of read counts. A-F respectively show the distribution without normalization, and normalization according to DESeq2, EdgeR, EdgeR by gene-size (black length < 5kb, red ≥ 5 kb), TPM and GeTMM.

coefficients for DESeq2 and EdgeR, and higher than the correlation coefficients for TPM (**Figure 3**). For most genes, DESeq2 had the highest correlation coefficients in absolute numbers, although the average and median difference with GeTMM showed very little difference in individual coefficients (0.014 and 0.008, respectively). Furthermore, no significant difference was observed between DESeq2, EdgeR and GeTMM (Mann-Whitney test, see **Supplementary Figure 4**) while TPM resulted in significantly lower coefficients compared to the other methods (p=0.02, p=0.04 and p=0.03 for DESeq2, EdgeR and GeTMM, respectively).

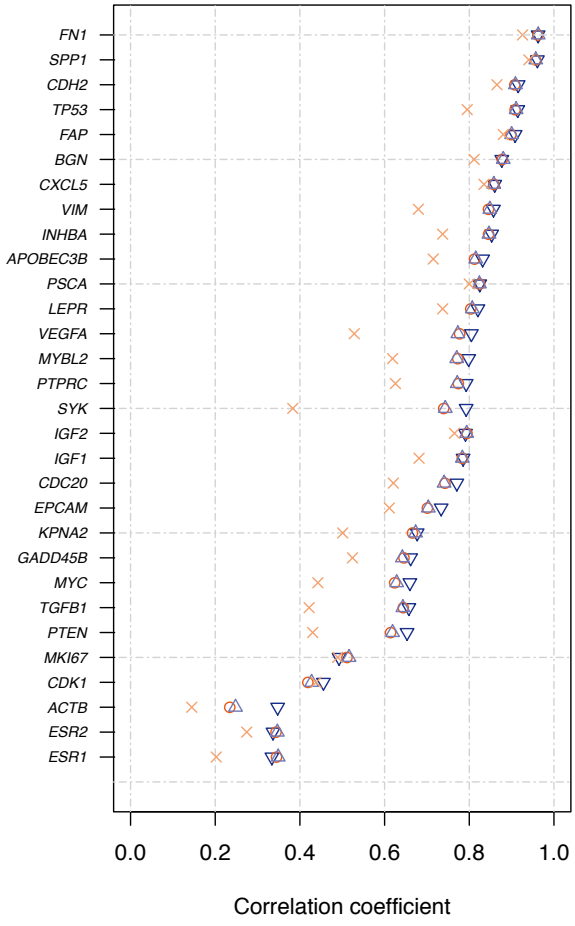


FIGURE 3. CORRELATION TO RT-QPCR DATA OF 30 GENES

Correlation coefficients (x-axis) of 30 genes comparing RNA-seq normalization methods to RT-qPCR generated data. Light orange cross: TPM, light blue triangle: GeTMM, dark orange circle: EdgeR and dark blue triangle: DESeq2.

The aim of this part of the study was not to appraise the correlation coefficients obtained using the RT-qPCR data but to use the RT-qPCR data as benchmark so the RNA-seq normalization procedures could be compared with each other. Nonetheless, we further investigated the five genes that showed an R<0.6; *MKI67*, *CDK1*, *ACTB, ESR1* and *ESR2*. The poor correlation of the latter 2 genes may be caused by the very low expression of these genes according to the RNA-seq data (median read count was just 22 for both *ESR1* and *ESR2*), indicating an insufficient sequencing depth for these genes. *ACTB* was the highest expressed gene of the 30 genes and had the lowest variance in 4 of 5 methods (0.25, 0.13, 0.16 and 0.16 for RT-qPCR, DESeq2, EdgeR and GeTMM, respectively), which may be the reason for the low correlation. For *CDK1* and *MKI67*, we re-analyzed all 263 samples to obtain the reads per exon. We observed a lower expression of exon 1 of *CDK1*, which may explain the poor correlation between the RT-qPCR and RNA-seq data as the RT-qPCR product spans exon 1 and 2 (**Figure 4A**). A similar analysis for *MKI67* did not show the same effect; here the

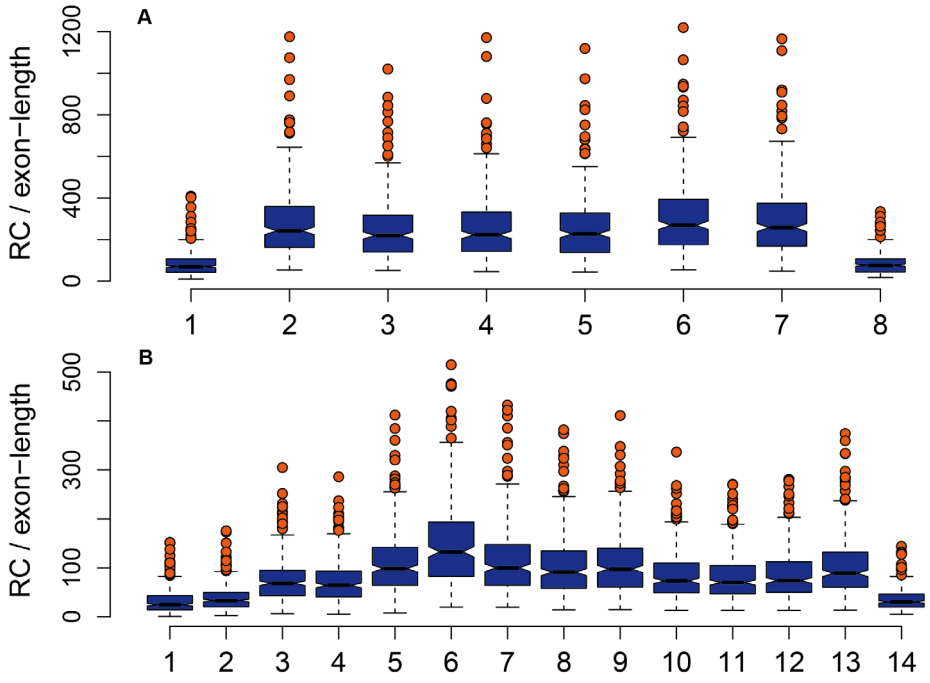


FIGURE 4. BOXPLOTS OF READ COUNTS PER EXON

A shows the expression levels in read counts per 100 bp for each exon in CDK1 (NB no additional normalization was performed). The whiskers extend to 1.5 IQR (interquartile range) above the third, or below the first quartile, with the median indicated by a horizontal line in the box. The notch indicates the 95% confidence interval of the median. **B** shows the same data for the *MKI67* gene.

RT-qPCR assay spans exon 10 to 11, which both showed similar expression levels as the overall gene expression level (**Figure 4B**). So unless transcript XM_006717864, which was the only truncated transcript of *MKI67* not covered by this RT-qPCR assay, is dominantly present in our sample cohort, we found no obvious explanation for this poor correlation.

COMPARISON TO RT-QPCR GENERATED DATA: INTRASAMPLE ANALYSIS

Previously[20], RNA-seq normalization methods were compared to RT-qPCR data in the MicroArray Quality Control (MAQC) and Sequence Quality Control SEQC effort[21], using an alternative setup; 996 genes were measured in a single sample by RT-qPCR and these were correlated to gene-expression levels as measured by RNA-seq of the same sample. To mimic the SEQC results, we repeated the analysis with the RT-qPCR data of the 30 genes, and calculated a Spearman's rank correlation coefficient between RT-qPCR and the different RNA-seq normalization methods for each of the samples, yielding 263 correlation coefficients per method (**Figure 5**).

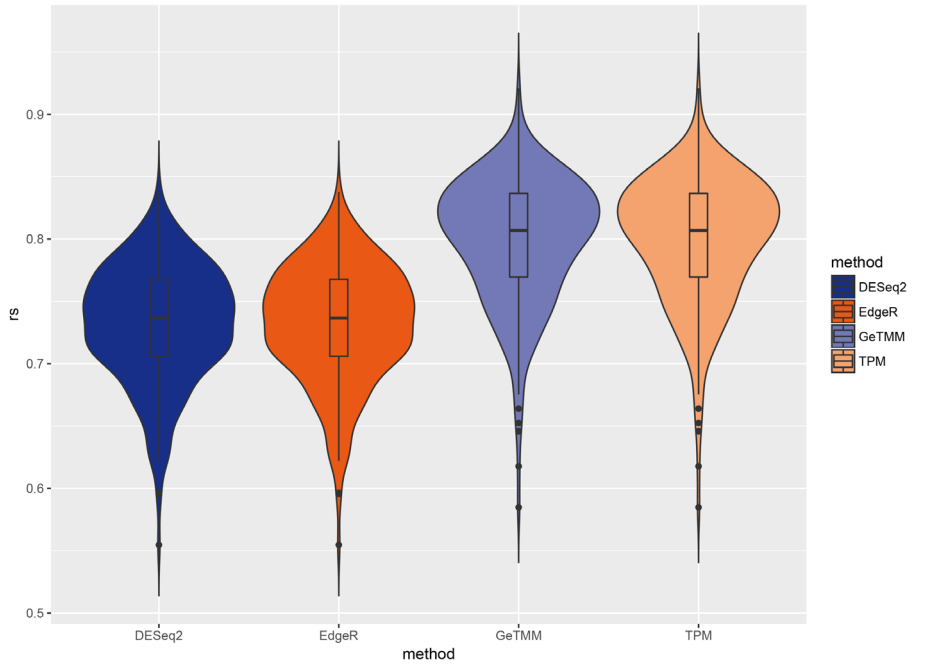


FIGURE 5. VIOLIN PLOTS RANK CORRELATION BY METHOD
Spearman rank correlation coefficients of 263 samples by correlating each method with RT-qPCR generated data.

GeTMM and TPM (the methods that include a gene length correction) both showed overall significant higher correlation to RT-qPCR data than DESeq2 and EdgeR (Mann-Whitney p<0.0001). GeTMM showed a higher correlation coefficient in 262 of the 263 cases.

THE PERFORMANCE OF GETMM IS NOT AFFECTED BY POOR RNA QUALITY

Next, we repeated the intersample correlation analysis with RT-qPCR data for the 76 samples that had an RNA integrity (RIN) value <7 after the cleanup procedure (median RIN 5.3), and compared these to an equally sized group of 76 samples with the highest RIN values (RIN>9, median RIN 9.5). The median library size of the low RIN group was slightly lower at 5.58 million versus 6.52 million for the high RIN group (Mann-Whitney p=0.02, see **Supplementary Figure 5A**). However, a principle component analysis using all expressed genes showed no separation of the low/high RIN groups, regardless of normalization method (**Supplementary Figure 5B-E**). Next, we correlated the RT-qPCR data to the RNA-seq data for each normalization method for the low and high RIN group separately, and compared the correlation coefficients between the groups. **Figure 6 A-D** shows a Bland-Altman difference plot for the four methods with the mean bias and p-value (Student's t-test under H0 that the difference is 0). Similar to the intersample comparison between RNA-seq and RT-qPCR in all samples, the result for GeTMM was similar to EdgeR and DESeq2, meaning the correlation coefficients were similar for the low and high RIN group. Normalization using TPM did result in significantly lower correlation coefficients in the low RIN group compared to the high RIN group (bias= -0.09477, p<0.0001), again indicating an advantage for GeTMM compared to TPM.

GETMM BEST RESEMBLES RESULTS OF DIFFERENTIAL EXPRESSION ANALYSIS USING RT-QPCR

The correlation of the different normalization methods to RT-qPCR data already showed that GeTMM performed equivalent to EdgeR and DESeq2, but outperformed TPM. To further study the effect of the different normalization methods on an intersample analysis in a biological relevant context, the expression of the genes in left sided and right sided colon tumors was examined, since tumors in the left and right hemicolon are known to be biologically different. In short, right-sided tumors are frequently hypermethylated, hypermutated, microsatellite instable and *BRAF*-mutated while left-sided tumors are frequently microsatellite stable and frequently carry an APC and *KRAS*-mutation.[22] This characteristic roughly divided our cohort in half (48% left-sided en 52% right-sided). We evaluated all 30 genes in the RT-qPCR data set by a standard t-test and after multiple testing correction (Benjamini-

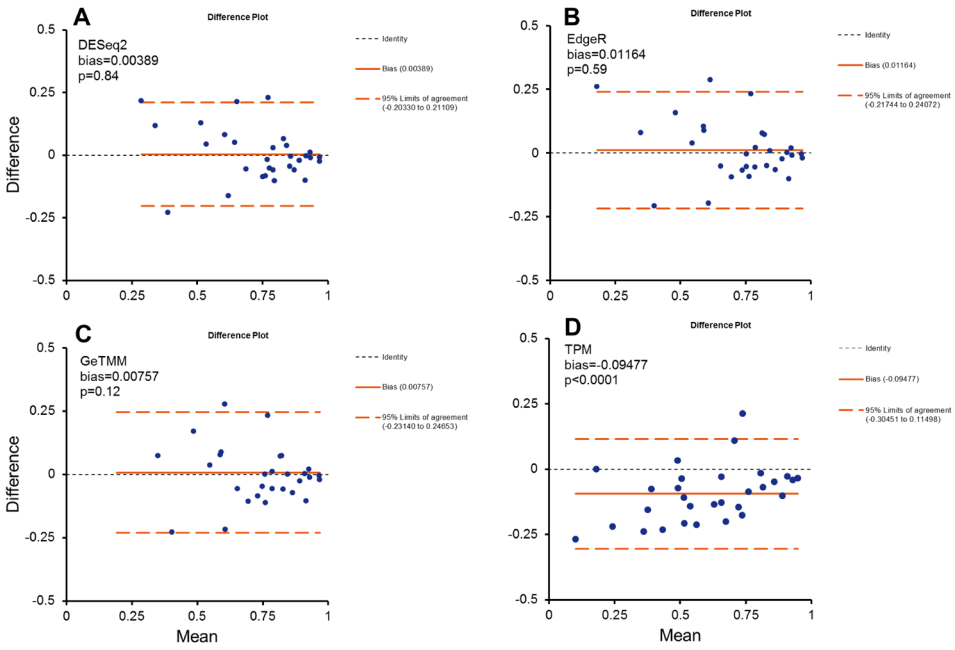


FIGURE 6. BLAND-ALTMAN PLOTS COMPARING SAMPLES WITH HIGH AND LOW RIN VALUES

Panel **A-D**: for each normalization method, a group of 76 samples with low RIN values (<7) was used to correlate expression data of 30 genes to RT-qPCR generated data. The same was performed for an equally sized high RIN sample group (>9) and the correlation coefficients were compared. X-axis shows the mean correlation, the y-axis the difference (high RIN – low RIN). The blue line indicates the bias (mean of all differences), the dashed light-blue lines show the 95% limits of agreement, the dashed black line at zero is the identity line (indicating no difference). The p-value is derived from a one-sample t-test.

Hochberg) 8 genes showed an FDR < 0.05: *MYBL2, MYC, EPCAM, SYK, APOBEC3B, SPP1, CDK1* and *IGF1*. Next, to check if the RNA-seq normalization methods showed differences in the amount of removal/compression of relevant biological variation, we calculated the Signal-to-Noise ratio (SNR) for these 8 genes. Again, GeTMM performed similar to EdgeR and DESeq2 showing very comparable SNRs, but outperformed TPM (see **Supplementary Table 3**). Next, the statistical tests implemented by EdgeR and DESeq2 were run on the respective data sets, while for TPM and GeTMM data, Student's t-tests were used on the 30 genes. **Figure 7** shows the results of comparing FDR adjusted p-values by normalization method. Out of the 22 genes that were not differentially expressed according to the RT-qPCR data, GeTMM had the lowest number of 'false positives' (5/22) compared to EdgeR (14/22), DESeq2 (7/22) and TPM (16/22). The recall was similar for all methods (4 out of 8 for EdgeR, and 3 out of 8 for the other methods).

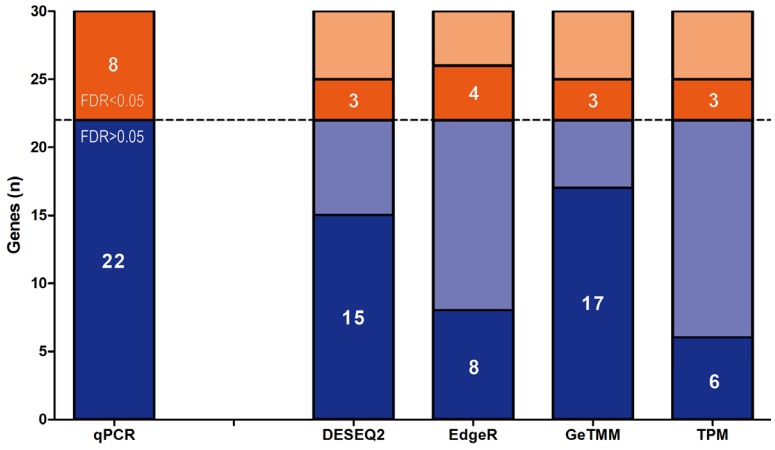


FIGURE 7. NUMBER OF DE GENES BETWEEN LEFT AND RIGHT SIDED TUMORS PER NORMALIZATION METHOD
RT-qPCR generated data were used as benchmark, showing 8 genes with FDR <0.05 (dark-grey) and 22 genes FDR>0.05 (black). For the RNA-seq normalization methods, black indicate true negatives (FDR>0.05, matches with RT-qPCR), white indicate false positives (FDR<0.05, not matching RT-qPCR), grey indicate true positives (FDR<0.05, matches RT-qPCR) and light-grey indicate false negatives (FDR>0.05, not matching RT-qPCR).

GENE LENGTH CORRECTION BENEFITS EDGER IN THE ONCOTYPE DX® RECURRENCE SCORE

An often used tool to estimate risk of recurrence in colon cancer is the Recurrence Score (RS) algorithm (Clark-Langone) of Onco*type* DX®[14], which uses a 7 cancer-gene panel. The RS was calculated for all samples, based on the RT-qPCR data as well as the RNA-seq normalized datasets (**Figure 8**). The distribution of the RT-qPCR generated scores are very similar to the scores generated using RNA-seq, except for the EdgeR derived RS. The overall lower scores will impact the RS evaluation, as the original RS is scaled such that negative scores will be set to zero. Using EdgeR, 41% of patients (n=109) would receive this score. Clearly GeTMM, which uses gene length correction on top of EdgeR normalization, improves the range and distribution of the RS scores.

GENE LENGTH CORRECTION IMPACTS CMS PREDICTION

Finally, the CMS classification was determined for each sample using data normalized by the different methods.[13] In this classification five possible groups are predicted: CMS1-4 and mixed/indeterminate. The type of classification is based on correlation of gene-signatures specific for each subtype to an individual sample, making this an intrasample-type analysis. Perfect agreement in the predicted CMS groups was seen between DESeq2 and EdgeR (both without gene length correction), and between

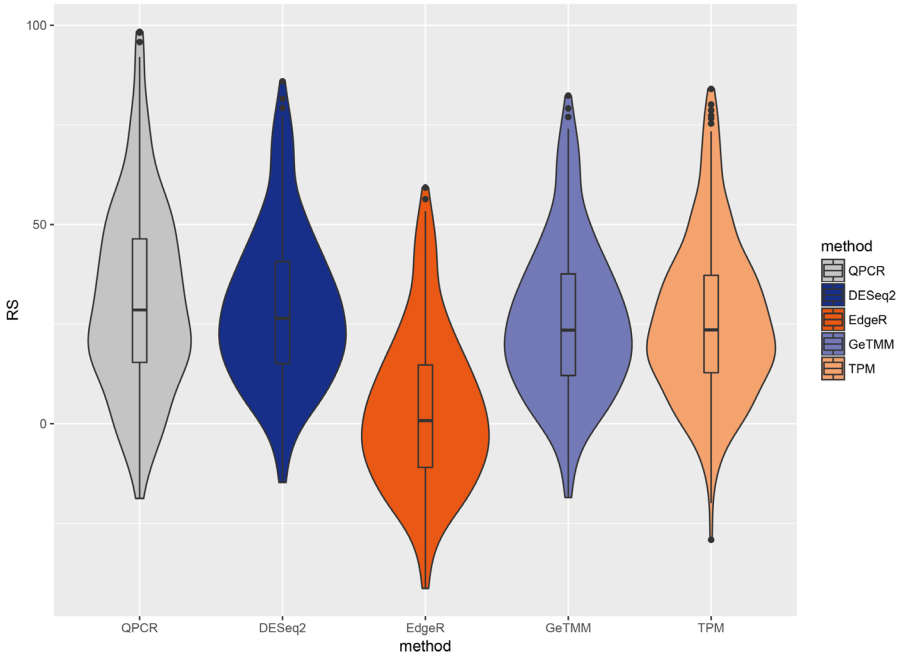


FIGURE 8. VIOLIN PLOTS RECURRENCE SCORE. THE ONCOTYPE DX® RECURRENCE SCORE (RS) OF 263 SAMPLES BY METHOD

TPM and GeTMM (both with gene length correction). However, gene length correction had a considerable impact on the prediction of the CMS groups: 40 samples (15.2%) were predicted in a different group when comparing EdgeR/DESeq2 and GeTMM/TPM (**Table 1**).

TABLE 1. PREDICTED CMS GROUP BY NORMALIZATION METHOD

	GETMM					
EdgeR	**CMS1**	**CMS2**	**CMS3**	**CMS4**	**Mixed/ indeterminate**	**Total**
CMS1	46	0	0	0	7	53
CMS2	0	127	0	0	5	132
CMS3	0	0	23	0	0	23
CMS4	0	1	0	5	4	10
Mixed/indeterminate	3	14	6	0	22	45
Total	49	142	29	5	38	263

DISCUSSION

The current study showed that GeTMM performed equivalent to the two most commonly used RNA-seq normalization methods DESeq2 and EdgeR (both use no gene length correction)[6-8], in intersample analyses while outperforming these methods in intrasample comparisons. Therefore, GeTMM generates a normalized data set directly suited for multiple endpoints. The effects of the different methods on the distribution of the gene expression data, samples with different RNA quality, subtype-classification, recurrence score, recall of differentially expressed genes and correlation to RT-qPCR data were assessed in a large cohort of real (i.e. not simulated) data, obtained from 263 primary colon tumors. Importantly, the current study focused on the application of RNA-Seq data for differential expression analysis between and within samples, thus not covering other applications such as the detection of fusion events, variant analysis and gene isoforms.[23] With regard to the latter, the normalization methods used in this study including GeTMM were not developed to distinguish possible isoforms, which requires estimating expression on a transcript level using more complex models and different statistics.[10, 24, 25] Thus, the investigated normalization methods may not be fully appropriate for such transcript level analyses.

The effect of gene length correction on downstream analysis is more important than it seems at first, when realizing that several frequently used standard analyses are vulnerable to gene length induced bias. Besides the theoretical example stated in the introduction, another example is e.g. in breast cancer, wherein the AIMS method[26] was developed to obtain a truly independent single sample classifier to robustly call molecular subtypes. Herein, subtype-specific genes are evaluated within each sample; e.g. when *GRB7* (a 532 bp transcript) is higher expressed than *BCL2* (a 239 bp transcript), it adds to the evidence for a HER2 subtype.[26] Without correcting for gene length, this prediction method will not work as intended on RNA-seq data as *GRB7* read counts will be about 2-fold higher compared to the *BCL2* read counts, when both genes are expressed at equal levels. Evaluating these intrasample-type analyses in the current study, GeTMM and TPM produced significantly better results compared to EdgeR and DESeq2 when correlating a set of genes measured by different methods within the same sample. A similar sort of analysis had been performed previously[20] using the data available from the MicroArray Quality Control (MAQC) effort, wherein more genes were measured by RT-qPCR, but only using two samples. In our study we used 263 samples, thus capturing the biological variation of gene expression levels much better. Regarding clinical applicability, this study showed that gene length correction influences the prediction of the subtypes (CMS) of colorectal cancer.[13]

Given the methodology of the CMS classifier, where the gene expression data of a single sample are correlated to a centroid of a set of genes that are specific to each of the 4 CMS groups, it makes more sense to use a normalization that includes a gene length correction, to avoid under- or overestimating the true expression levels of genes within a sample. Thus, assuming that the GeTMM classification reflects a more reliable prediction, 23 samples would change from a CMS group to mixed/indeterminate using a method without gene length correction, and 1 sample would change from CMS2 to CMS4. In calculating the recurrence score (Onco*type* DX®) EdgeR showed an overall much lower distribution and assigned almost half of the patients below a zero score. This was remedied by including a gene length correction (thus yielding GeTMM), resulting in scores very comparable and in the same range as the RT-qPCR generated scores. This illustrates the importance of using a normalization method like GeTMM, that results in a data set that is suited for both intersample as well as intrasample analyses.

Several metrics were used to evaluate the normalization methods, summarized in **Table 2**. In general, TPM is not sufficient to correct for between-sample differences. This echoes previously reported results using RPKM and FPKM normalization[5, 6, 11, 12], and it is reasonable to conclude that normalization by library size alone must be abandoned as viable method to detect DE genes between samples. DESeq2 and EdgeR differed only slightly with respect to distribution, correlation to RT-qPCR and sensitivity to RNA quality, and not at all with regard to the CMS classification. However, EdgeR seemed overly optimistic in identifying DE genes while DESeq2 is more conservative, a difference that was also observed by others.[8] Given the strong similarities between the data after normalization with DESeq2 and EdgeR, the differences in the reported DE genes are more likely a result of differences in the statistical tests employed by both methods than by the normalization itself.

The analyses using subsets of samples with a low or high RIN value showed remarkably little difference in downstream results. It appears that samples with a low RIN value may yield sequencing data suitable for expression analyses. Still, this

TABLE 2. SUMMARY OF RESULTS

NORMALIZATION METHOD	GENE LENGTH CORRECTION	DISTRIBUTION PER SAMPLE	INFLUENCE OF RIN	INTERSAMPLE CORRELATION	INTRASAMPLE CORRELATION
DESeq2	no	bimodal	no bias	++	+
EdgeR	no	bimodal	no bias	++	+
TPM	yes	normal	bias	-	++
GeTMM	yes	normal	no bias	++	++

conclusion may be very specific to the entire protocol that was used (RNA isolation, library prep etc.) and may therefore not be applicable to all studies and protocols. Still, a-priori disregarding samples with a low RIN value for sequencing could prove wasteful, though it is prudent to perform a robust QC on the generated sequencing data to spot failed samples.

Lastly, this study uses RT-qPCR as standard so the RNA-seq normalization methods could be compared with each other. RT-qPCR is known for its precise and reproducible measurements and may have a bigger dynamic range compared to the usual coverage of sequence data. The downside is that RT-qPCR measures just a small part of the gene, may miss or be affected by splice-variants, and can be affected by SNPs in the primer regions. In that respect, the RNA-seq generated data may be nearer the mark of the actual expression level of a gene. In the future, RNA-seq may replace RT-qPCR as the gold standard for expression data, provided a well-founded normalization method is used.

This study shows that GeTMM produces a versatile normalized RNA-seq data set, appropriate for both inter- and intrasample comparisons. This quality of GeTMM should further enhance the capacity of RNA-seq as a solid method to explore and compare gene expression profiles, and thus may become increasingly interesting in the current era of data sharing efforts.

REFERENCES

1. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621-8.

2. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25.

3. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.

4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

5. Bullard JH, Purdom E, Hansen KD, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010;11:94.

6. Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 2013;14:671-83.

7. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 2013;14:R95.

8. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 2013;14:91.

9. Li B, Ruotti V, Stewart RM, et al. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 2010;26:493-500.

10. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28:511-5.

11. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 2009;4:14.

12. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012;131:281-5.

13. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

14. Clark-Langone KM, Sangli C, Krishnakumar J, et al. Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX Colon Cancer Assay. BMC Cancer 2010;10:691.

15. Kloosterman WP, Coebergh van den Braak RRJ, Pieterse M, et al. A Systematic Analysis of Oncogenic Gene Fusions in Primary Colon Cancer. Cancer Res 2017;77:3814-22.

16. Sieuwerts AM, Lyng MB, Meijer-van Gelder ME, et al. Evaluation of the ability of adjuvant tamoxifen-benefit gene signatures to predict outcome of hormone-naive estrogen receptor-positive breast cancer patients treated with tamoxifen in the advanced setting. Mol Oncol 2014;8:1679-89.

17. Sieuwerts AM, Meijer-van Gelder ME, Timmermans M, et al. How ADAM-9 and ADAM-11 differentially from estrogen receptor predict response to tamoxifen treatment in patients with recurrent breast cancer: a retrospective study. Clin Cancer Res 2005;11:7311-21.

18. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.

19. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923-30.

20. Li P, Piao Y, Shon HS, et al. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics 2015;16:347.

21. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol 2014;32:903-14.

22. Muller MF, Ibrahim AE, Arends MJ. Molecular pathological classification of colorectal cancer. Virchows Arch 2016;469:125-34.

23. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.

24. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.

25. Mehta S, Tsai P, Lasham A, et al. A Study of TP53 RNA Splicing Illustrates Pitfalls of RNA-seq Methodology. Cancer Res 2016;76:7151-9.

26. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J Natl Cancer Inst 2015;107:357.

3

SUPPLEMENTARY DATA

SUPPLEMENTARY METHODS

STAR ALGORITHM

The STAR algorithm (version 2.4.2a) was used to align the RNA-seq data on the GRCh38 reference genome. Settings were:

--outSAMstrandField intronMotif
--outFilterIntronMotifs RemoveNoncanonicalUnannotated
--chimSegmentMin 12
--chimJunctionOverhangMin 12
--alignSJDBoverhangMin 10
--alignMatesGapMax 200000
--alignIntronMax 200000
--outSAMtype BAM SortedByCoordinate
--outSAMunmapped Within
--alignEndsType Local
--chimOutType WithinBAM
--twopassMode Basic
--twopass1readsN -1
--quantMode GeneCounts

NORMALIZATION

The raw readcount matrix (tab-delimited text file) was used as input (x), in which the first column holds the geneID from Ensembl that are used as row names, the second column the gene length in kb and the remaining columns contain read counts of each sample.

```
# calculate RPK
rpk <- (x[,2:ncol(x)]/x[,1])
# remove length col in x
x <- x[,-1]
# for normalization purposes, no grouping of samples
group <- c(rep("A",ncol(x)))
```

EDGER

```
x.norm.edger <- DGEList(counts=x,group=group)
x.norm.edger <- calcNormFactors(x.norm.edger)
norm.counts.edger <- cpm(x.norm.edger)
```

RPK/EDGER

```
rpk.norm <- DGEList(counts=rpk,group=group)
rpk.norm <- calcNormFactors(rpk.norm)
norm.counts.rpk_edger <- cpm(rpk.norm)
```

TPM

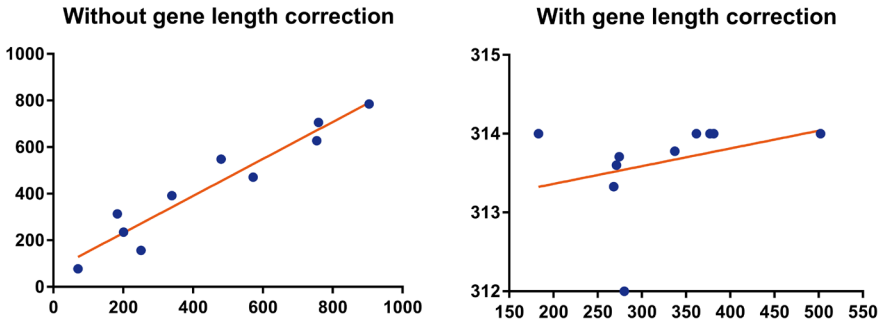
```
tpm = rpk
for (i in 1:ncol(rpk) ) {
        tpm[,i] <- rpk[,i]/(sum(rpk[,i])/1e6)
}
```

DESEQ2

```
# no group & no design implemented
colData = data.frame(group)
rownames(colData)=colnames(x)
dds<-DESeqDataSetFromMatrix(countData=x,colData=colData, design=~ 1)
dds <- estimateSizeFactors(dds)
sizefact <- sizeFactors(dds)
norm.counts.deseq <- counts(dds, normalized=TRUE)
```

After processing, read counts were log2-transformed (setting genes to NA when having 0 read counts).

SUPPLEMENTARY FIGURES



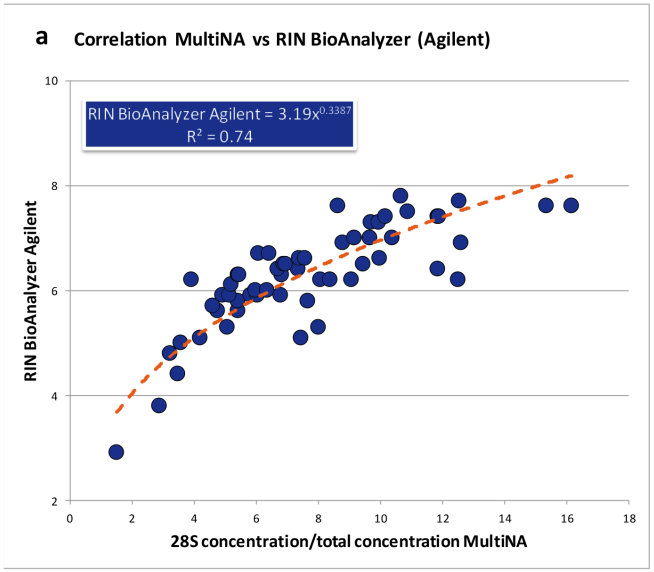
		READCOUNTS		RPK	
gene	**length (kb)**	**sample1**	**sample2**	**sample1**	**sample2**
gene1	0.25	70	78	280.00	312.00
gene2	0.5	251	157	502.00	314.00
gene3	0.75	201	235	268.00	313.33
gene4	1	183	314	183.00	314.00
gene5	1.25	339	392	271.20	313.60
gene6	1.5	572	471	381.33	314.00
gene7	1.75	480	549	274.29	313.71
gene8	2	754	628	377.00	314.00
gene9	2.25	759	706	337.33	313.78
gene10	2.5	904	785	361.60	314.00

SUPPLEMENTARY FIGURE 1. IMPACT OF GENE LENGTH CORRECTION ON CORRELATION

Simulated expression data of 10 genes in 2 samples. Correlation based on read counts show different results after correcting for gene length. RPK indicates reads per kilobase.

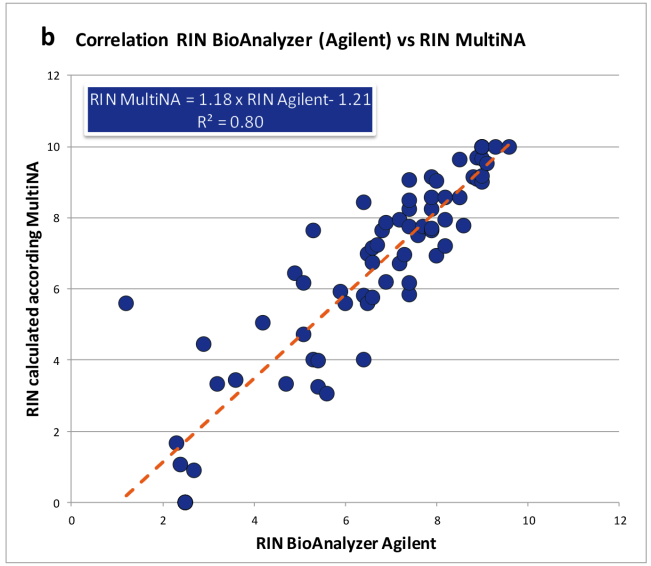
TRAINING: **[y=RIN BioAnalyzer, x=RIN MultiNA Conc. 28S/Total Conc]**

Training set: Total RNA isolated with RNA-B from 60 freshly frozen breast tumor tissues



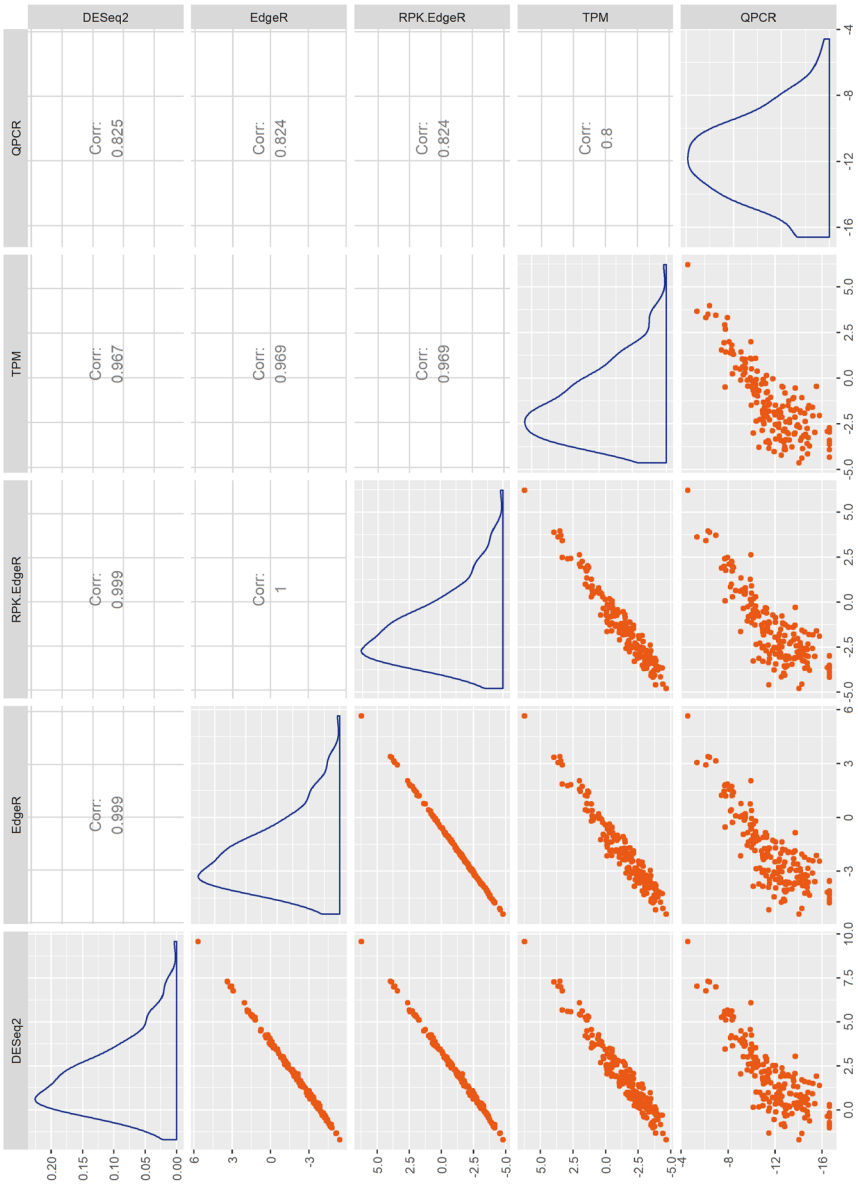
VALIDATION: **[RIN MultiNA =3.19*POWER(28S/total Conc*100,0.3387])**

Testing set: Total RNA isolated with RNA-B from 73 independent freshly frozen breast tumor tissues

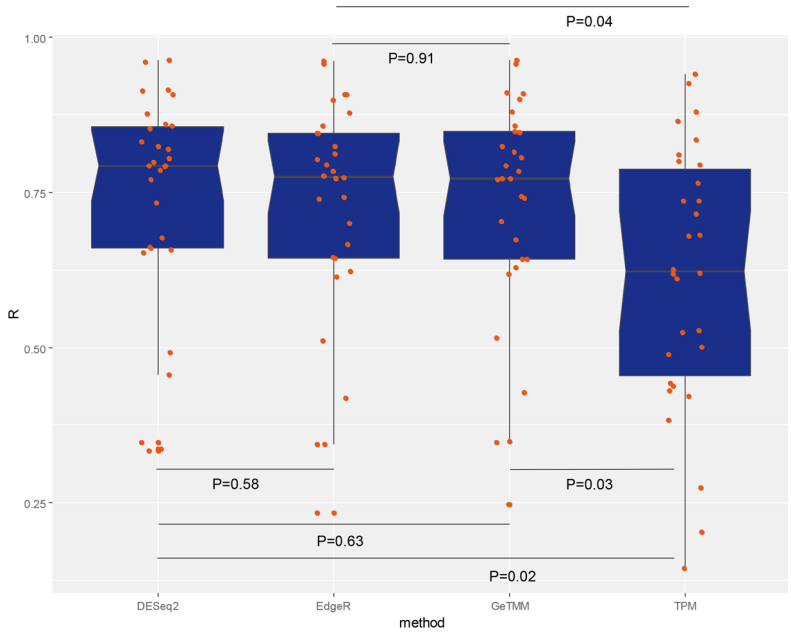


SUPPLEMENTARY FIGURE 2. CORRELATION RIN BIOANALYZER VS MULTINA

RIN values as measured by the Bioanalyzer (Agilent) were compared to the 28S/total concentration as measure by the MultiNA in a training set of 60 cases (**A**). The resulting trend line was validated in an independent cohort (**B**) of 73 cases.

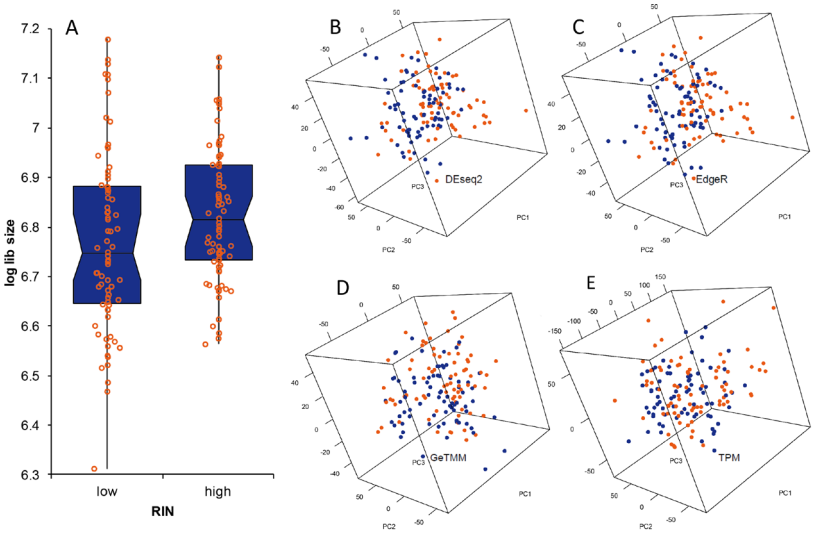


SUPPLEMENTARY FIGURE 3. EXPRESSION OF *PSCA* AND COMPARISON OF SEVERAL RNA-SEQ NORMALIZATION METHODS



SUPPLEMENTARY FIGURE 4. COMPARISON CORRELATION COEFFICIENTS BY METHOD

Boxplots show correlation coefficient of 30 genes, comparing 4 methods to RT-qPCR generated data. P-values are derived from the Mann-Whitney test.



SUPPLEMENTARY FIGURE 5. LIBRARY SIZE AND PCA PLOTS BY RIN

A shows the library size (log10) in samples with low RIN values (RIN<7) or high RIN (≥9). **B-E** show PCA plots, colored by samples with low RIN (red) or high RIN (blue), by normalization method.

SUPPLEMENTARY TABLES

SUPPLEMENTARY TABLE 1. DETAILS ON THE RT-QPCR ASSAYS

Approved Gene Symbol	Approved Gene Name	Method	Assay ID Applied BioSystems	F sequence
TBP	TATA box binding protein	SYBR		TTCGGAGAGTTCTGGGATTG
HPRT1	hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome)	SYBR		TATTGTAATGACCAGTCAACAG
HMBS	hydroxymethylbilane synthase	SYBR		CATGTCTGGTAACGGCAATG
CXCL5	chemokine (C-X-C motif) ligand 5 (CXCL5)	SYBR		CTGTGTTGAGAGAGCTGCGT
MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	SYBR		AGCAAGTGCAAGGTCAAATGG
ESR1	estrogen receptor 1	SYBR		ATCCTACCAGACCCTTCAGTG
VIM	vimentin	SYBR		CAGATTCAGGAACAGCATGTC
ESR2	estrogen receptor 2 (ER beta)	SYBR		CATGCTCCTGGCAACTACTTC
APOBEC3B	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B	SYBR		CGCCAGACCTACTTGTGCTA
PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	SYBR		CGGGAAGACAAGTTCATGTAC
CDC20	cell division cycle 20 homolog (S. cerevisiae)	SYBR		CTTCCCTGCCAGACCGTATC
SYK	spleen tyrosine kinase	SYBR		GCATCGACAAAGACAAGACAG
TGFB1	transforming growth factor, beta 1	SYBR		GCCCTGGACACCAACTATTG
CDK1	cell division cycle 2, G1 to S and G2 to M, Homo sapiens cyclin-dependent kinase 1 (CDK1)	SYBR		GCCGCCGCGGAATAAT
IGF2	insulin-like growth factor 2 (somatomedin A)	SYBR		GCGGCTTCTACTTCAGCAG
VEGFA	vascular endothelial growth factor A	SYBR		TACCTCCACCATGCCAAG
IGF1	insulin-like growth factor 1 (somatomedin C)	SYBR		TGGTGGATGCTCTTCAGTTC
KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)	SYBR		TTCCTGATGATGCTACTTCTC
ACTB	actin, beta	Taqman assay		AAGCCACCCCACTTCTCTCTAA
EPCAM	tumor-associated calcium signal transducer 1	Taqman assay		AGTTTGCGGACTGCACTTCA
PSCA	prostate stem cell antigen	Assay-on-demand	Hs00194665_m1	not shared by supplier
MKI67	antigen identified by monoclonal antibody Ki-67	Assay-on-demand	Hs00606991_m1	not shared by supplier
LEPR	leptin receptor	Assay-on-demand	Hs00900242_m1	not shared by supplier
PTPRC	protein tyrosine phosphatase, receptor type, C	Assay-on-demand	Hs00236304_m1	not shared by supplier
CDH2	cadherin 2, type 1, N-cadherin (neuronal)	Assay-on-demand	Hs00983062_m1	not shared by supplier
TP53	tumor protein p53	Assay-on-demand	Hs99999147_m1	not shared by supplier
FN1	fibronectin 1	Assay-on-demand	Hs00277509_m1	not shared by supplier
SPP1	secreted phosphoprotein 1	Assay-on-demand	Hs00959010_m1	not shared by supplier
MYC	v-myc avian myelocytomatosis viral oncogene homolog	Assay-on-demand	Hs00905030_m1	not shared by supplier
BGN	biglycan	Assay-on-demand	Hs00959141_g1	not shared by supplier
FAP	fibroblast activation protein, alpha	Assay-on-demand	Hs00990806_m1	not shared by supplier
INHBA	inhibin, beta A	Assay-on-demand	Hs01081598_m1	not shared by supplier
GADD45B	growth arrest and DNA-damage-inducible, beta	Assay-on-demand	Hs04188837_g1	not shared by supplier

R sequence	Probe context sequence	reference NM_code	exon boundery	product size (bp)	Specifics
ACGAAGTGCAATGGTCTTTAG		NM_003194.4	3 -> 4	94	Reference gene
GGTCCTTTTCACCAGCAAG		NM_000194.2	3/4 -> 7	192	Reference gene
GTACGAGGCTTTCAATGTTG		NM_000190.3	1 -> 4	139	Reference gene
GTTTTCCTTGTTTCCACCGTC		NM_002994.4	2 -> 3/4	218	
CTGTCCAAACTGCCTCACCA		NM_002466.3	2 -> 3	72	
GCCAGACGAGACCAATCATC		NM_000125.3	4 -> 5	186	66 kD variant
TCAGAGAGGTCAGCAAACTTG		NM_003380.3	4 -> 4/5	158	
GCTCTTGGCAATCACCCAAAC		NM_001291723.1	6 -> 7	221	
GCCACAGAGAAGATTCTTAGCC		NM_004900.4	5 -> 6	111	
CTCTATACTGCAAATGCTATCG		NM_001304717.2	7 -> 8	222	
CCAATCCACAAGGTTCAGGTAATA		NM_001255.2	5 -> 6	71	
GGATGGGAACCTGGAAGTTG		NM_003177.6	4 -> 6	193	
CGTGTCCAGGCTCCAAATG		NM_000660.5	2 -> 3	168	
CCTTCTCCAATTTTCTCTATTTTGGT		NM_001786.4	1 -> 2	86	variant 1+2
CAGGTGTCATATTGGAAGAAC		NM_000612.5	2/3 -> 4	214	
GGTACTCCTGGAAGATGTC		NM_001025366.2	1 -> 3	148	
GACAGAGCGAGCTGACTTG		NM_001111283.2	2 -> 3	191	
GCCCGGATTATGTTGTCTATG		NM_001320611.1	3 -> 5	187	
ATGCTATCACCTCCCCTGTGT	AGAATGGCCCAGTCCTCTCCCAAGTC	NM_001101.3	6 -> 6	69	
AATACTCGTGATAAATTTTGGATCCA	AAGGAGATCACAACGCGT	NM_002354.2	4/5 -> 5	72	
not shared by supplier	GCAGCCAGGCACTGCCCTGCTGTGC	NM_005672.4	1 -> 2	82	ONLY V1
not shared by supplier	AAGATCTTTCAGGAATAGCTGAAAT	NM_001145966.1	10->11	137	
not shared by supplier	AATTAATAGTTTCACTCAAGATGAT	NM_001003679.3	17->18	76	
not shared by supplier	AGAGGCTGAATTCCAGAGACTTCCT	NM_002838.4	26 -> 27	81	
not shared by supplier	CACCGTGGTCAAACCAATCGACTTT	NM_001792.4	9->10	78	
not shared by supplier	CACTAAGCGAGCACTGCCCAACAAC	NM_000546.5	8 -> 9	121	
not shared by supplier	ACCACTCTGGAGAATGTCAGCCCAC	NM_001306130.1	34->35	88	
not shared by supplier	GCAGACCTGACATCCAGTACCCTGA	NM_000582.2	5 -> 6	84	
not shared by supplier	AAACCAGCAGCCTCCCGCGACGATG	NM_002467.4	1 -> 2	87	
not shared by supplier	GGCATCCCCAAAGACCTCCCTGAGA	NM_001711.5	5 -> 6	65	
not shared by supplier	TTTCCAGGCAATGTGGTACTCTGAC	NM_004460.3	25 -> 26	67	
not shared by supplier	TTGCCGAGTCAGGAACAGCCAGGAA	NM_002192.2	2 -> 3	61	
not shared by supplier	GTCTCCTGGTCACGAACCCTCACAC	NM_015675.3	3 -> 4	66	

SUPPLEMENTARY TABLE 2. CORRELATION COEFFICIENTS FOR EACH METHOD COMPARED TO RT-QPCR

ENSG	NAME	LENGTH (KB)	DESEQ2	EDGER	RPK/EDGER	TPM
ENSG00000115414	*FN1*	15.601	0.963	0.962	0.963	0.926
ENSG00000118785	*SPP1*	4.578	0.961	0.957	0.957	0.941
ENSG00000170558	*CDH2*	5.709	0.915	0.908	0.909	0.865
ENSG00000141510	*TP53*	9.345	0.914	0.908	0.911	0.795
ENSG00000078098	*FAP*	3.252	0.908	0.899	0.9	0.88
ENSG00000182492	*BGN*	2.394	0.877	0.878	0.88	0.811
ENSG00000163735	*CXCL5*	2.534	0.86	0.858	0.858	0.835
ENSG00000026025	*VIM*	3.188	0.857	0.845	0.849	0.68
ENSG00000122641	*INHBA*	7.691	0.853	0.846	0.846	0.737
ENSG00000179750	*APOBEC3B*	2.823	0.832	0.812	0.815	0.715
ENSG00000167653	*PSCA*	1.566	0.825	0.824	0.824	0.8
ENSG00000116678	*LEPR*	14.642	0.82	0.803	0.807	0.737
ENSG00000112715	*VEGFA*	9.98	0.805	0.777	0.773	0.528
ENSG00000101057	*MYBL2*	3.445	0.798	0.773	0.77	0.619
ENSG00000081237	*PTPRC*	7.225	0.793	0.774	0.772	0.626
ENSG00000165025	*SYK*	5.16	0.793	0.74	0.743	0.383
ENSG00000167244	*IGF2*	14.849	0.791	0.795	0.794	0.765
ENSG00000017427	*IGF1*	15.78	0.785	0.784	0.784	0.681
ENSG00000117399	*CDC20*	2.047	0.771	0.743	0.741	0.621
ENSG00000119888	*EPCAM*	2.415	0.734	0.701	0.703	0.611
ENSG00000182481	*KPNA2*	2.766	0.677	0.666	0.673	0.501
ENSG00000099860	*GADD45B*	2.03	0.661	0.646	0.642	0.524
ENSG00000136997	*MYC*	8.663	0.66	0.624	0.629	0.442
ENSG00000105329	*TGFB1*	2.762	0.657	0.644	0.643	0.422
ENSG00000171862	*PTEN*	10.292	0.653	0.614	0.619	0.43
ENSG00000148773	*MKI67*	15.667	0.492	0.511	0.516	0.489
ENSG00000170312	*CDK1*	5.581	0.456	0.419	0.428	0.438
ENSG00000075624	*ACTB*	1.911	0.347	0.234	0.248	0.145
ENSG00000140009	*ESR2*	11.903	0.336	0.344	0.347	0.274
ENSG00000091831	*ESR1*	15.485	0.334	0.345	0.349	0.202

SUPPLEMENTARY TABLE 3. SIGNAL-TO-NOISE RATIOS

GENE	DESEQ2	EDGER	RPK/EDGER	TPM
MYBL2	0.586	0.562	0.560	0.305
MYC	0.541	0.512	0.505	0.218
EPCAM	0.330	0.283	0.279	-0.042
SYK	0.120	0.094	0.090	-0.170
APOBEC3B	0.119	0.106	0.103	-0.048
SPP1	-0.447	-0.453	-0.454	-0.510
CDK1	-0.122	-0.137	-0.141	-0.340
IGF1	-0.017	-0.027	-0.028	-0.146

R.R.J. Coebergh van den Braak, S. ten Hoorn, A.M. Sieuwerts, M. Smid, S.M. Wilting, J.W.M. Martens, J.A. Foekens, J.P. Medema, J.N.M. IJzermans, L. Vermeulen

4

INTERCONNECTIVITY BETWEEN TUMOR BIOLOGY AND TUMOR STAGE IN COLORECTAL CANCER

IN PREPARATION

ABSTRACT

INTRODUCTION

There are profound individual differences in clinical outcome within tumor stages of colorectal cancer. Tumor biology, in conjunction with the traditional TNM staging, is a promising way for predicting patient outcomes and treatment efficacy. This study was conducted to investigate the interconnectivity between tumor stage and tumor biology (reflected by the Consensus Molecular Subtypes (CMS)) in colorectal cancer, and explore the added value of this knowledge in patients with stage II colon cancer.

METHODS

The analyses were performed in a large series of colorectal cancer patients of which gene expression data was available that could be used to determine the CMS classification. The interconnectivity was assessed by investigating the association between CMS and TNM, and investigating differential gene expression between various TNM stages within and across disease subtypes. Furthermore, the prognostic value of CMS in stage II colon cancer patients in light of the issue of stage migration was evaluated.

RESULTS

CMS4 was more prevalent in advanced stages of disease, and in stage II colon cancer patients with inadequate lymph node assessment. The majority of genes that were differentially expressed between tumor stages were no longer differentially expressed when stratifying patients according to CMS subtype. CMS held prognostic value in stage II colon cancer patients with inadequate lymph node assessment.

CONCLUSION

The results of this study suggest considerable interconnectivity between tumor biology and tumor stage in colorectal cancer. This implies that TNM stage in addition to disease progression is a reflection of a distinct biological disease process. An important advantage of CMS compared to TNM is that it also provides insight in tumor biology, which might both predict the prognosis and specific treatment options. In particular, we show that CMS could help to guide treatment decisions in stage II colon cancers with inadequate lymph node assessment.

INTRODUCTION

As in most solid tumors, clinical decision making in colorectal cancer (CRC) is mainly driven by clinical and traditional pathological features. Although these features hold considerable prognostic and predictive value, one should acknowledge that there are profound individual differences in clinical outcome within for instance a single tumor stage. Currently, most efforts are directed towards the identification of prognostic and predictive biomarkers that complement the traditional features or identify a subgroup within a seemingly homogeneous subgroup.[1, 2] However, the traditional features may very well be interconnected with biomarkers that resemble tumor biology, which can hamper such biomarker studies.

The consensus molecular subtypes (CMS) classification is an upcoming stratification tool for CRC defining four groups (CMS 1-4) with distinct clinical features.[3] Specifically, patients with a CMS4 tumor ('the mesenchymal subtype') were associated with a worse relapse-free and overall survival compared to patients with a CMS1-3 in an aggregated cohort of patients with stage I-IV colorectal cancer.[3] In addition, CMS subtypes are associated with different responses to currently employed drugs.[4-6] Importantly, the distinct subtypes present with vastly different biological and molecular features. These range from frequencies of genetic driver events, the presence of microsatellite instability but also stromal composition.[3] Furthermore, we previously reported that the various colorectal cancer subtypes relate to distinct precursor lessions.[7] Hence, the CMS taxonomy offers an ideal framework to elucidate whether TNM solely resembles disease progression or biologically different entities that preferentially present with a specific stage of disease.

This study was conducted to investigate the interconnectivity between tumor stage and tumor biology in colorectal cancer, and to investigate the added value of this knowledge in patients with stage II colon cancer, a subgroup in which accurate prognostication and selection for adjuvant systemic treatment is still an unmet need.

METHODS

PATIENT SELECTION

Patients were selected from two sources. The first source were public data sets for which staging information and the CMS classification was available.[3] The second source was the MATCH study, an independent prospective multicenter observational cohort study that was initiated in 2007, in which patients undergoing curative surgery for CRC were enrolled. Informed consent was given for the collection of clinical data

and the storage and use of biobank samples for research purposes (Institutional Review Board number MEC 2007-088).

AFFYMETRIX ARRAY

A detailed description of sample collection and processing for the MATCH cohort was described previously.[8] In short, RNA was isolated from 30-mm sections taken from the frozen tumor tissue obtained at primary surgery. Only samples with an RNA integrity (RIN) value of at least 7.0 were included in the final analysis. Fragmentation of RNA, labeling, hybridization to Human Genome U133 Plus 2.0, microarrays scanning and the Affymetrix microarray analysis were performed following the manufacturer's protocol (Affymetrix).

CMS CLASSIFICATION

The consensus groups were identified using a consensus clustering approach. This approach included the construction of a network followed by network clustering using a Markov cluster algorithm[9], and cluster evaluation by repeating the first two steps 1,000 times. Then, the optimal number of clusters was determined using the weighted Silhouette width (R package 'WeightedCluster'). This approach generated four consensus molecular subtypes. A random forest algorithm[10] was used to define a 'group classifier' to classify sets of samples. Furthermore, a 'single-sample-predictor (SSP) classifier' was developed as an R package to classify future samples in an individual fashion.

The labels from the group classifier were used for the analyses in the aggregated data set.[3] For the survival analysis in the aggregated cohort of stage II colon cancer patients, the raw expression data were normalized in one batch using Robust Multi-array Average (RMA) and labeled using the SSP classifier.[3, 11]

BIOINFORMATIC ANALYSIS AND GENE SET ENRICHMENT ANALYSIS (GSEA)

The R2: Genomic Analysis and Visualization Platform (Academic Medical Center, Amsterdam, the Netherlands; http://r2.amc.nl) was used to identify differentially expressed genes between the different tumor stages and different CMS groups in the GSE39582 and TCGA data set.[12] For overall differentially expressed genes the ANOVA test (false discovery rate (FDR) corrected $p<0.05$) was used, for individual groups the limma-test (FDR corrected $p<0.05$). For each analysis, a random set of 200 tumors per group was sampled to correct for the effect of group size on the number of differentially expressed genes. Visualization of the genes that display significant differences between tumor stages in the whole group of colon cancers was done using a t-distributed stochastic neighbor embedding (t-SNE) algorithm.[13]

The basic idea of the t-SNE algorithm is to minimize the difference between specially defined conditional probability distributions that represent similarities, which is calculated for the data points in the high and low dimensional representations. The central assumption is that the conditional probabilities will be equal if the low dimensional mapped points in Y space correctly model the similarity structure of its higher dimensional counterparts in X.

STATISTICAL ANALYSIS

The Chi square test and Kruskal-Wallis test were used to assess differences in patient and tumor characteristics among the cohorts. The Kaplan-Meier method was used to estimate survival. Survival curves were compared using the log-rank test. Disease-free survival (DFS) times of >60 months were censored at 60 months. Lymph node assessment was considered inadequate when <10 lymph nodes were assessed or the number of assessed lymph nodes was unknown. Statistical analyses were performed using the SPSS statistical package version 21.

RESULTS

CONSENSUS MOLECULAR SUBTYPES ARE INTERCONNECTED WITH THE TNM CLASSIFICATION

To assess the interconnectivity between CMS and tumor stage, we investigated the association between CMS and tumor stage in a large aggregated cohort of colorectal cancer samples for which detailed staging information was available and which we classified previously in molecular subtypes (n=1,713).[14] Details on the cohort are listed in **Supplementary Table 1**. We observed that the prevalence of the poor-prognosis mesenchymal subtype (CMS4) increased with the extent of the disease (stage I 20.7%, stage II 23.6%, stage III 32.0% and stage IV 40.1%, p<0.001) (**Figure 1A**). As microsattelite instability (MSI) is considered to be a very distinct and well-defined phenotype that is mostly confined to CMS1 and is well known to infrequently present with metastatic disease, we also investigated the distribution of CMS in microsatellite stable (MSS) tumors.[15] We detected an even more striking trend regarding the prevalence of CMS4 cancers in each tumor stage (stage I 12.4%, stage II 27.4%, stage III 31.7% and stage IV 39.5%, p=0.001) (**Figure 1B**). The increased proportion of mesenchymal tumors was not dependent on altered frequencies of the epithelial type subtypes as we detected similar distributions when CMS2, representing the canonical colon cancer subtype with a pronounced epithelial-like expression signature and high levels of Wnt target genes, and CMS3 also characterized

by high expression of intestinal epithelial specific gene expression but unique because of a marked metabolic deregulation, were combined (**Figure 1C**). Importantly, the association between CMS4 and advanced stages of disease was consistently observed in the individual cohorts (**Supplementary Figure 1** and **Supplementary Table 2**). Together these data indicate that distinct TNM stages represent with different distributions of molecular subtypes. Most strikingly, the proportion of mesenchymal colon cancers increases with more advanced cancer stage.

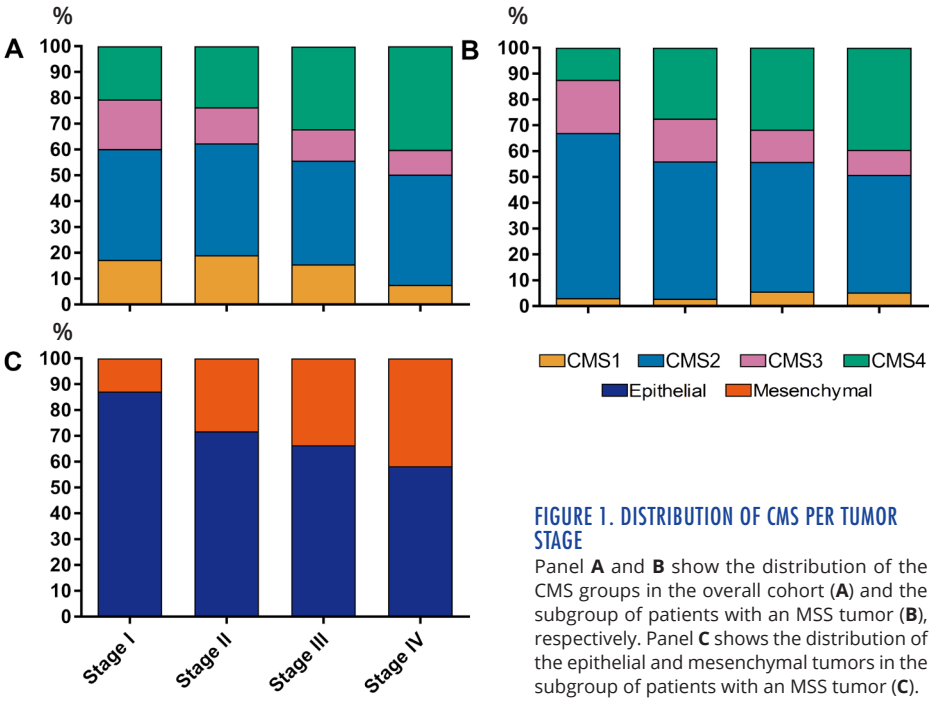


FIGURE 1. DISTRIBUTION OF CMS PER TUMOR STAGE
Panel **A** and **B** show the distribution of the CMS groups in the overall cohort (**A**) and the subgroup of patients with an MSS tumor (**B**), respectively. Panel **C** shows the distribution of the epithelial and mesenchymal tumors in the subgroup of patients with an MSS tumor (**C**).

TUMOR STAGE RESEMBLES TUMOR BIOLOGY RATHER THAN DISEASE PROGRESSION

In order to test the hypothesis that tumor stage does not only presents disease progression but also reflects molecular features, we investigated the changes in gene expression between distinct TNM stages and related these to the gene expression differences found between molecular subtypes. Initially, we performed a differential expression analysis (ANOVA) for individual genes on the MSS tumors. Details on the distribution of TNM and CMS is shown in **Supplementary Table 3**. We assessed the number of differentially expressed genes between tumor stages in the total cohort

consisting of all subtypes, and in the CMS groups separately, by ANOVA. As the total number of patients in CMS3, and especially CMS1 because we only analysed MSS cancers, was too small we excluded these subgroups for analysis.
The analysis revealed considerable gene expression differences in the total group between TNM stages (**Figure 2A**). However, when stratifying by CMS, the number of differentially expressed genes was significantly less in CMS2 and CMS4 compared to the overall analysis ($p=0.002$ and $p<0.001$, respectively) (**Figure 2A**). The far majority of the gene expression differences (98%) were observed when comparing the non-metastasized (stage I-II) with metastasized (stage III-IV) tumors while only marginal gene expression differences were observed when comparing stage I vs II and stage III vs IV. Of note, especially in CMS4 cancers no differentially expressed genes were detected between early stage and advanced disease. This indicates that the fundamental biology of these cancers does not change when these cancers progress and are primed for metastatic spread (**Figure 2A**). Visualization of the genes that display significant differences between tumor stages in the whole group of colon cancers (ANOVA $p< 0.05$; $n=2191$) using the t-SNE algorithm showed clear separation for the mesenchymal (CMS4) and the epithelial subtypes (CMS2/3) (**Figure 2B**). To more specifically investigate the association between CMS4 and more advanced tumor stages, we built a gene signature to segregate stage I-II vs stage III-IV cancers, and a gene signature to segregate epithelial (CMS2-3) vs mesenchymal (CMS4) subtypes using the top 100 differentially expressed genes. Both signatures were used to calculate a TNM III-IV and CMS4 signature score for each sample. Remarkably, the two scores were highly correlated ($r_s=0.85$, $p<0.001$) (**Figure 2C**), confirming marked interconnectivity between more advanced tumor stages and CMS4. These analyses demonstrate that the differences in gene expression between TNM stages of disease are in essence only a reflection of the relative abundance of the various subtypes, and do not reflect changes in tumor biology associating with disease progression.

CMS IS ASSOCIATED WITH STAGE MIGRATION IN STAGE II COLON CANCER

In order to employ these new insights to a clinically relevant problem we investigated the issue of stage migration in early stage colon cancer. In some cases patients that in fact have stage III disease are staged as stage II because a positive lymph node is missed during pathological assessment, for example due to a low number of assessed lymph nodes. This is of direct importance as stage III colon cancer patients are treated with adjuvant chemotherapy and sage II are not in most cases. However, stage II patients with high risk features, including inadequate number of lymfnodes assessed are offered adjuvant chemotherapy as well in some cases. Based on the

association between CMS and tumor stage, one may hypothesize that CMS4 may be informative in predicting stage III disease when a low number of lymph nodes is available for analysis.

To test this hypothesis, we first investigated the impact of the number of assessed lymph nodes on the chance of finding one or more positive lymph nodes in a cohort of colon cancers that underwent surgery from 1994-2001 when the number of lymph nodes was not yet widely implemented in pathological assesement protocols. This cohort contained 410 patients with stage II-III colon cancer for which basic

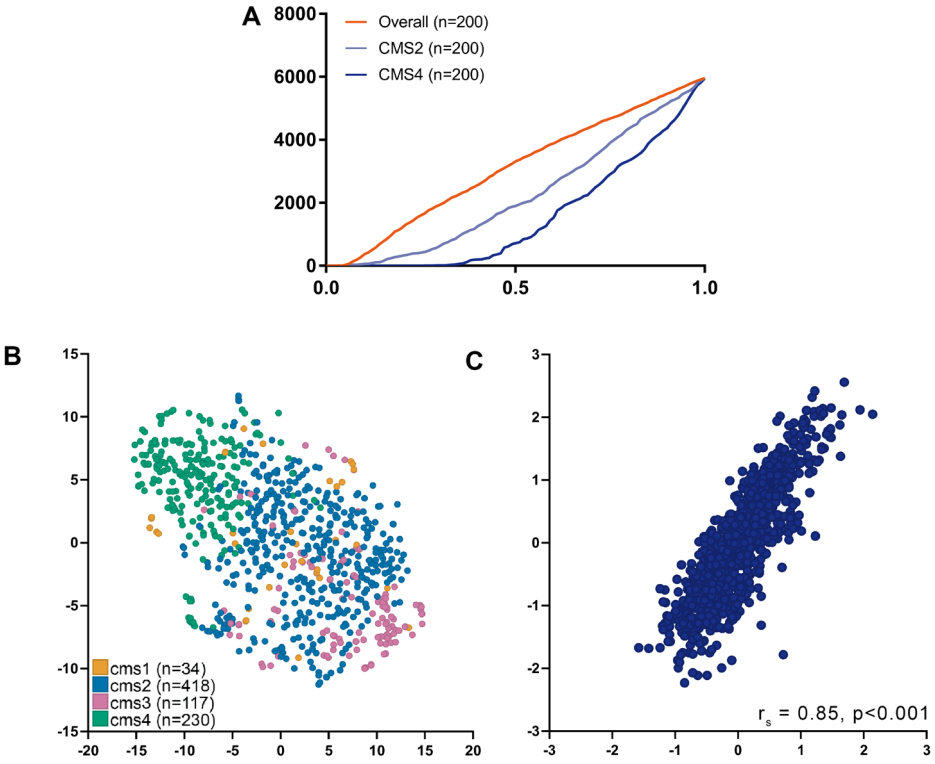


FIGURE 2. GENE EXPRESSION ANALYSIS

Panel **A** depicts the cumulative number of differentially expressed genes (y-axis) plotted against the p value used as cut-off to define differential expression (x-axis). The analysis showed considerable gene expression differences in the total group. When stratifying CMS, the number of differentially expressed genes was significantly less in CMS2 and CMS4 compared to the overall analysis (p=0.002 and p<0.001, respectively). Panel **B** is a visualization of the genes that display significant differences between tumor stages in the whole group using a t-SNE algorithm with clear separation of the mesenchymal (CMS4) and the epithelial subtypes (CMS2/3) (x- and y-axis are the conditional probabilities based on the higher and low dimensional representations, respectively). Panel C displays the correlation between the tumor stage III-IV (x-axis) and CMS4 (y-axis) signature score (r_s=0.85, p<0.001).

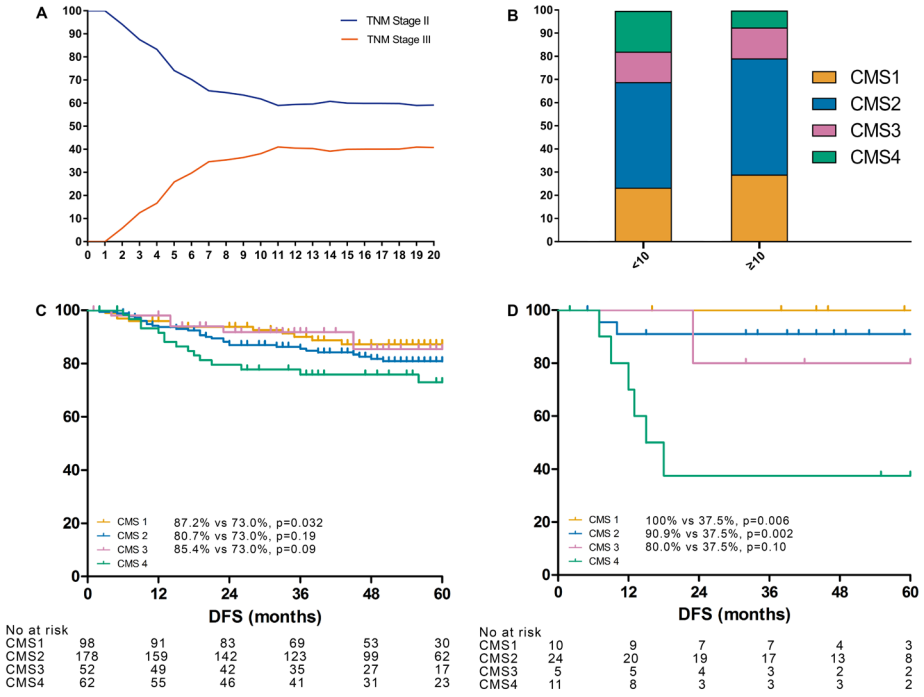


FIGURE 3. CMS AND STAGE II COLON CANCER

Panel **A** shows that the chance of finding a positive lymph node (y-axis) increases with an increasing number of assessed lymph nodes (x-axis), which plateaus after 10 lymph nodes. Panel **B** shows the distribution of CMS in stage II colon cancer (y-axis distribution in percentages) stratified for lymph node assessment with CMS4 being more prevalent in patients in whom <10 lymph nodes were assessed. Panel **C** and **D** display the disease-free survival (x-axis in months) of the total set of patients with stage II colon cancer (**C**), and the subset of patients with stage II colon cancer and <10 assessed LN (**D**) (y-axis survival in percentages).

characteristics are listed in **Supplementary Table 4**. This analysis showed that the percentage of stage III colon cancers increased with an increase of the number of assessed lymph nodes and plateaued at 10 lymph nodes (**Figure 3A**). We then investigated the association between CMS and the number of assessed lymph nodes, and found that CMS4 was more prevalent in patients with inadequate lymph node assessment compared to patients with adequate lymph node assessment (17.6% vs 7.3%, p=0.018 respectively) (**Figure 3B**). These results indicate that CMS may be associated with stage migration, is uncommon in properly staged stage II colon cancers, and may help to identify patients who are at risk to be understaged. More specifically, mesenchymal cancers are at higher risk for inadequate staging.

TABLE 1. BASELINE CHARACTERISTICS OF PATIENTS WITH UNTREATED STAGE II COLON CANCER

	TOTAL		MATCH COHORT		GSE14333		GSE33113		GSE39582		P VALUE
	n=459		n=112		n=63		n=90		n=194		
Gender											
Female	216	47.1%	57	50.9%	31	49.2%	48	53.3%	80	41.2%	0.27
Male	243	52.9%	55	49.1%	32	50.8%	42	46.7%	114	58.8%	
Age											
median (interquartile range)	71	(63-78)	70	(63-76)	71	(63-78)	73	(60.8-79.3)	71	(71-78)	0.64
T~											
3	339	73.9%	107	95.5%	0	0.0%	81	90.0%	151	77.8%	0.001
4	50	10.9%	5	4.5%	0	0.0%	9	10.0%	36	18.6%	
missing	7	1.5%	0	0.0%	63	100.0%	0	0.0%	7	3.6%	
LN assessed											
median (range)	14	(1-46)	14	(5-28)	-		12	(1-46)	-		0.08
LN assessed*											
< 10	45	9.8%	14	12.5%	0	0.0%	31	34.4%	0	0.0%	<0.001
≥ 10	147	32.0%	98	87.5%	0	0.0%	49	54.4%	0	0.0%	
missing	10	2.2%	0	0.0%	63	100.0%	10	11.1%	194	100.0%	
Tumor location											
Left	205	44.7%	49	43.8%	30	47.6%	42	46.7%	84	43.3%	0.91
Right	254	55.3%	63	56.3%	33	52.4%	48	53.3%	110	56.7%	
MSI~											
MSS	276	60.1%	79	70.5%	0	0.0%	64	71.1%	133	68.6%	0.22
MSI	85	18.5%	28	25.0%	0	0.0%	26	28.9%	31	16.0%	
Missing	98	21.4%	5	4.5%	63	100.0%	0	0.0%	30	15.5%	
CMS											
1	99	21.6%	29	25.9%	15	23.8%	20	22.2%	35	18.0%	0.053
2	178	38.8%	52	46.4%	16	25.4%	31	34.4%	79	40.7%	
3	52	11.3%	11	9.8%	7	11.1%	9	10.0%	25	12.9%	
4	62	13.5%	5	4.5%	11	17.5%	16	17.8%	30	15.5%	
Mixed or inderminate	68	14.8%	15	13.4%	14	22.2%	14	15.6%	25	12.9%	

~ Comparison between the MATCH, GSE33113 and GSE39582 cohort. * Comparison between the MATCH and GSE33113 cohort

CMS HOLDS PROGNOSTIC VALUE IN STAGE II COLON CANCER WITH INADEQUATE LYMPH NODE ASSESSMENT

As prognostication is still an unmet need in stage II colon cancer, we investigated the prognostic value of CMS in stage II colon cancer in general, and specifically in patients in whom stage migration may be relevant. To this end, we selected patients with stage II colon cancer who were treated with surgery alone from the four cohorts for which information on age, location of the tumor (colon or rectum), tumor stage, adjuvant therapy and disease-free survival were available (MATCH, GSE14333[16], GSE33113[7] and GSE39582[12] cohort).

The aggregated cohort of stage II colon cancer consisted of 459 patients and included; [i] 112 patients from the MATCH cohort, [ii] 63 patients from the GSE14333 cohort, [iii] the entire GSE33113 cohort (n=90), and [iv] 194 patients of the GSE39582 cohort (patient selection is shown in **Supplementary Figures 2-4** and baseline characteristics are listed in **Table 1**). Overall, patients with a CMS4 tumor had a worse 5-year DFS compared to the other subtypes (CMS4 73.0% vs CMS1 87.2% p=0.032, CMS2 80.7% p=0.19, CMS3 85.4% p=0.09) (**Figure 3C**). We then analyzed the DFS of patients per subtype stratified for lymph node assessment in the two cohorts for which the number of lymph nodes was known (MATCH and GSE33113 cohort). In particular, in the subset of patients with inadequate lymph node assessment, CMS4 had a worse 5-year DFS rate (37.5%) compared to CMS1 (100%), CMS2 (90.9%) and CMS3 (80.0%) (p=0.006, p=0.002 and p=0.10, respectively) (**Figure 3D**). This suggests that CMS might be a classification strategy to assist TNM based classification when reliable determination of the tumor stage is not feasible.

DISCUSSION

CMS is a widely implemented and robust classification system for colorectal cancer which identifies four subtypes with distinguishing biological features and corresponding prognosis.[3] We therefore took advantage of this taxonomy to elucidate whether tumor stage resembles intrinsic features of tumor biology that are installed early during tumor development or disease progression. This study was conducted to investigate the interconnectivity between tumor stage and tumor biology in colorectal cancer, and investigate the added value of this knowledge in patients with stage II colon cancer.

The first observation that suggested interconnectivity between tumor biology and tumor stage was the increase of CMS4 patients with advancing stages of disease, which was observed consistently in the total group, the MSS patients and all individual

cohorts. These results may suggest that the poor prognosis (i.e. poor DFS en overall survival) for increased stages of disease is (in part) explained by the aggressive tumor biology of CMS4, given the poor prognosis of CMS4 compared to the other subtypes.[3] Within stage II, CMS4 was more prevalent in patients with inadequate lymph node assessment compared to patients with adequate lymph node assessment. These results illustrate the issue of stage migration in these patients, and suggest that the CMS classification may be useful to reduce the risk of understaging and undertreatment in patients with inadequate lymph node staging. Currently, inadequate lymph node assessment is considered to be a high risk factor in stage II colon cancer, and is subsequently used as an argument to offer adjuvant chemotherapy to these patients.[17-19] Interestingly, none of the five factors that are used to define high risk stage II colon cancer have been demonstrated to have the predictive value in prospective trials.[14] The survival analysis in patients with stage II colon cancer showed that CMS was of added value in patients with an uncertain lymph node status (<10 or unknown number of lymph nodes).[18-20] This suggests that CMS may help to identify patients at high risk of recurrence within the subset of patients with inadequate lymph assessment. However, these results should be validated in larger series given the relatively small number of patients. Given the distinct biological features of the four CMS groups, this classification may not only be helpful to identify high risk patients, but may also be used to select patients for specific treatments in contrast to the currently used high risk factors.[14] Patients with an MSI tumor (mostly CMS1) are known to have very little to no benefit from chemotherapy.[21, 22] However, these patients may very well benefit from immunotherapy as was shown in patients with heavily pre-treated metastasized colorectal cancer.[23] Patients with a CMS2 tumor were shown to be responsive to Oxaliplatin-containing chemotherapy while mesenchymal tumors (CMS4) seemed refractory to 5FU-based chemotherapy, suggesting CMS may also be used to select patients for specific chemotherapy regimens.[4, 5] Future prospective studies should be conducted to confirm these hints on CMS-specific drug sensitivity, as these findings originate from retrospective and therefore potentially biased studies.

A second observation in support of the interconnectivity between tumor stage and tumor biology was the marked decrease in differentially expressed genes between tumor stages when stratifying for CMS. This shows that almost all biological differences between tumor stages seem to be explained by CMS, which in turn strongly supports the hypothesis that different tumor stages are largely driven by intrinsic features of tumor biology that are installed early during tumor development rather than disease progression. Naxerova *et al.* showed that lymph node metastases (stage III) and distant metastases (stage IV) arose from different independent

subclones in the primary tumor in the majority of their cohort.[24] These findings further underline that different biological entities are likely to have different metastatic potential and corresponding prognosis. Still, our gene expression analysis suggested that the biological differences between non-metatasized (stage I-II) and metastasized (stage III-IV) disease largely exceeds the biological differences within these two groups (i.e. stage I vs II and stage III vs IV). In line with these observations, we previously uncovered that most of the published gene signatures that are generated by comparing patients with and without metastases mostly identify the same patients, namely those with a poor prognosis.[7] The majority of the patients who were marked as having a poor prognosis by the investigated prognostic gene signatures including the Oncotype Dx[25] were colon cancer subtype 3 (CCS3), the mesenchymal subtype corresponding with CMS4. Furthermore, none of these signatures held prognostic value after correcting for colon cancer subtype. Thus, CMS appears to identify the same high risk patients as these signatures, and has the advantage that it distinguishes four subtypes with extensive biological differences which helps in the understanding and specific targeting of these subgroups.

In line with the results from basic and translational studies that support the interconnectivity between tumor stage and tumor biology[7, 24], several clinical studies also provide support for the interconnectivity between tumor stage and tumor biology.[26-29] These studies, in which the clinical outcome in colorectal cancer patients with synchronous versus metachronous liver metastases was compared, showed that patients with synchronous and metachronous liver metastases had a similar overall survival after development of liver metastases. These findings were reported both in patients who underwent a radical metastasectomy and patients with irresectable disease who received palliative systemic therapy.[26-29] This suggests that tumor biology installed at an early moment in tumor development rather than the progression over time is the main determinant for prognosis in these patients.

This study provides substantial evidence to support the theory that tumor stage and the corresponding prognosis are largely driven by tumor biology. Therefore, CMS has the potential to be a major contributor to clinical decision making. Future efforts should focus on further substantiating these findings and the development of a clinically applicable CMS test.

REFERENCES

1. Dienstmann R, Mason MJ, Sinicrope FA, et al. Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. Ann Oncol 2017;28:1023-31.

2. Gray RG, Quirke P, Handley K, et al. Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. J Clin Oncol 2011;29:4611-9.

3. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

4. 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.

5. Song N, Pogue-Geile KL, Gavin PG, et al. Clinical Outcome From Oxaliplatin Treatment in Stage II/III Colon Cancer According to Intrinsic Subtypes: Secondary Analysis of NSABP C-07/NRG Oncology Randomized Clinical Trial. JAMA Oncol 2016;2:1162-9.

6. Dienstmann R, Vermeulen L, Guinney J, et al. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer 2017;17:79-92.

7. 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.

8. Kloosterman WP, Coebergh van den Braak RRJ, Pieterse M, et al. A Systematic Analysis of Oncogenic Gene Fusions in Primary Colon Cancer. Cancer Res 2017.

9. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30:1575-84.

10. Breiman L. Random Forests. Machine Learning 2001;45:5-32.

11. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249-64.

12. 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.

13. Jamieson AR, Giger ML, Drukker K, et al. Exploring nonlinear feature space dimension reduction and data representation in breast Cadx with Laplacian eigenmaps and t-SNE. Med Phys 2010;37:339-51.

14. Andre T, de Gramont A, Vernerey D, et al. Adjuvant Fluorouracil, Leucovorin, and Oxaliplatin in Stage II to III Colon Cancer: Updated 10-Year Survival and Outcomes According to BRAF Mutation and Mismatch Repair Status of the MOSAIC Study. J Clin Oncol 2015;33:4176-87.

15. Gelsomino F, Barbolini M, Spallanzani A, et al. The evolving role of microsatellite instability in colorectal cancer: A review. Cancer Treat Rev 2016;51:19-26.

16. 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.

17. Moertel CG, Fleming TR, Macdonald JS, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. Ann Intern Med 1995;122:321-6.

18. Benson AB, 3rd, Schrag D, Somerfield MR, et al. American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. J Clin Oncol 2004;22:3408-19.

19. Labianca R, Nordlinger B, Beretta GD, et al. Early colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2013;24 Suppl 6:vi64-72.

20. Oncoline. (Accessed at www.oncoline.nl.)

21. Des Guetz G, Schischmanoff O, Nicolas P, et al. Does microsatellite instability predict the efficacy of adjuvant chemotherapy in colorectal cancer? A systematic review with meta-analysis. Eur J Cancer 2009;45:1890-6.

22. Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol 2010;28:3219-26.

23. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med 2015;372:2509-20.

24. Naxerova K, Reiter JG, Brachtel E, et al. Origins of lymphatic and distant metastases in human colorectal cancer. Science 2017;357:55-60.

25. Clark-Langone KM, Sangli C, Krishnakumar J, et al. Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX Colon Cancer Assay. BMC Cancer 2010;10:691.

26. t Lam-Boer J, Van der Geest LG, Verhoef C, et al. Palliative resection of the primary tumor is associated with improved overall survival in incurable stage IV colorectal cancer: A nationwide population-based propensity-score adjusted study in the Netherlands. Int J Cancer 2016;139:2082-94.

27. van der Pool AE, Lalmahomed ZS, Ozbay Y, et al. 'Staged' liver resection in synchronous and metachronous colorectal hepatic metastases: differences in clinicopathological features and outcome. Colorectal Dis 2010;12:e229-35.

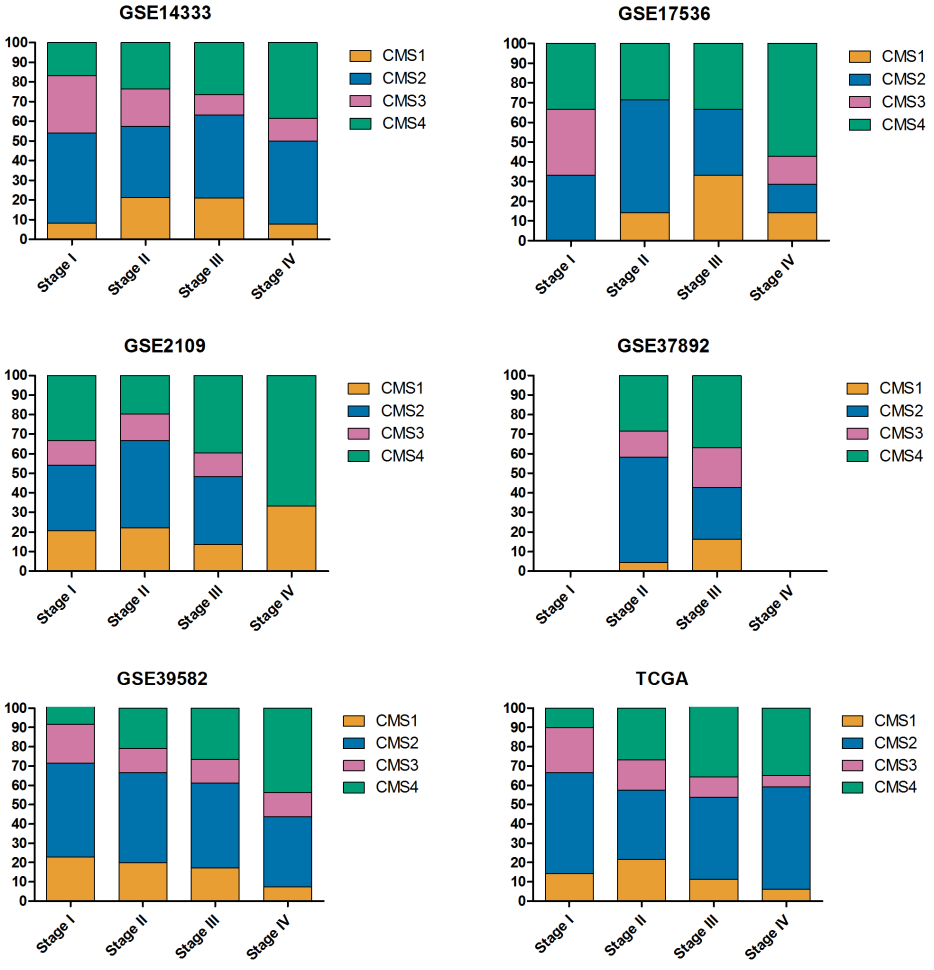
28. Mekenkamp LJ, Koopman M, Teerenstra S, et al. Clinicopathological features and outcome in advanced colorectal cancer patients with synchronous vs metachronous metastases. Br J Cancer 2010;103:159-64.

29. Rahbari NN, Carr PR, Jansen L, et al. Time of Metastasis and Outcome in Colorectal Cancer. Ann Surg 2017.

4

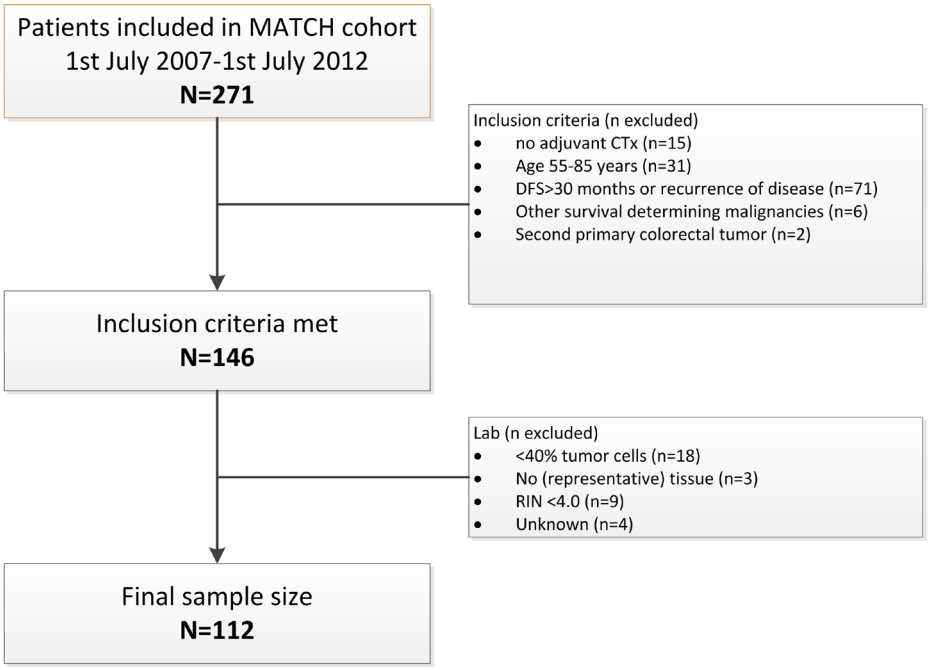
SUPPLEMENTARY DATA

SUPPLEMENTARY FIGURES



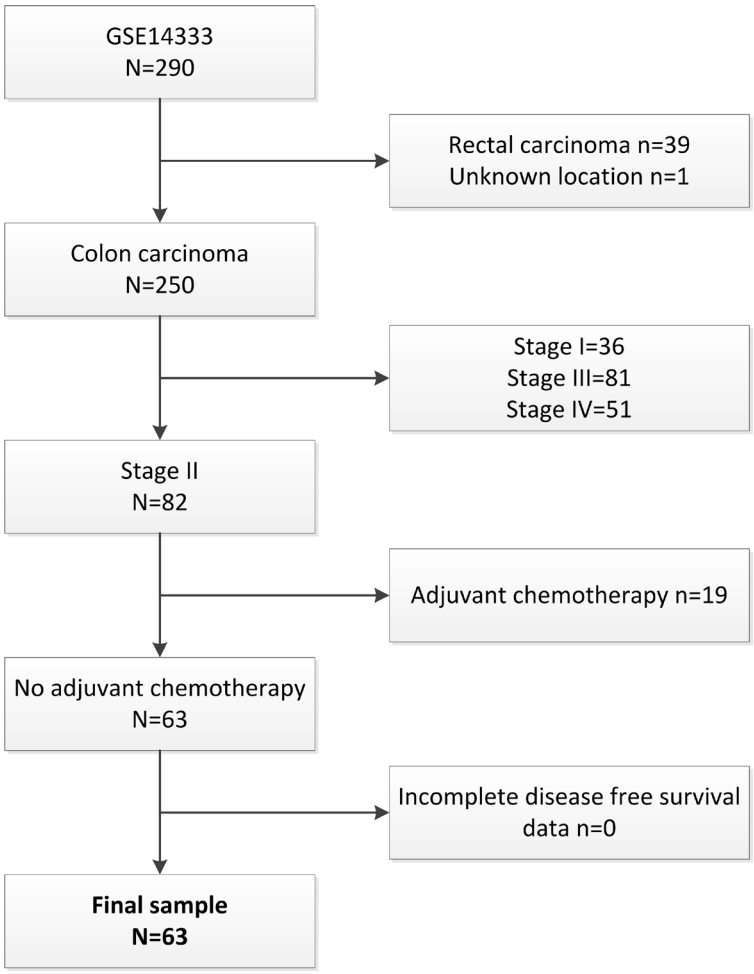
SUPPLEMENTARY FIGURE 1. DISTRIBUTION OF CMS PER TUMOR STAGE IN THE INDIVIDUAL COHORTS

The association between CMS4 and advanced stages of disease was consistently observed in the individual cohorts. Absolute numbers are listed in Supplemental Table 2.

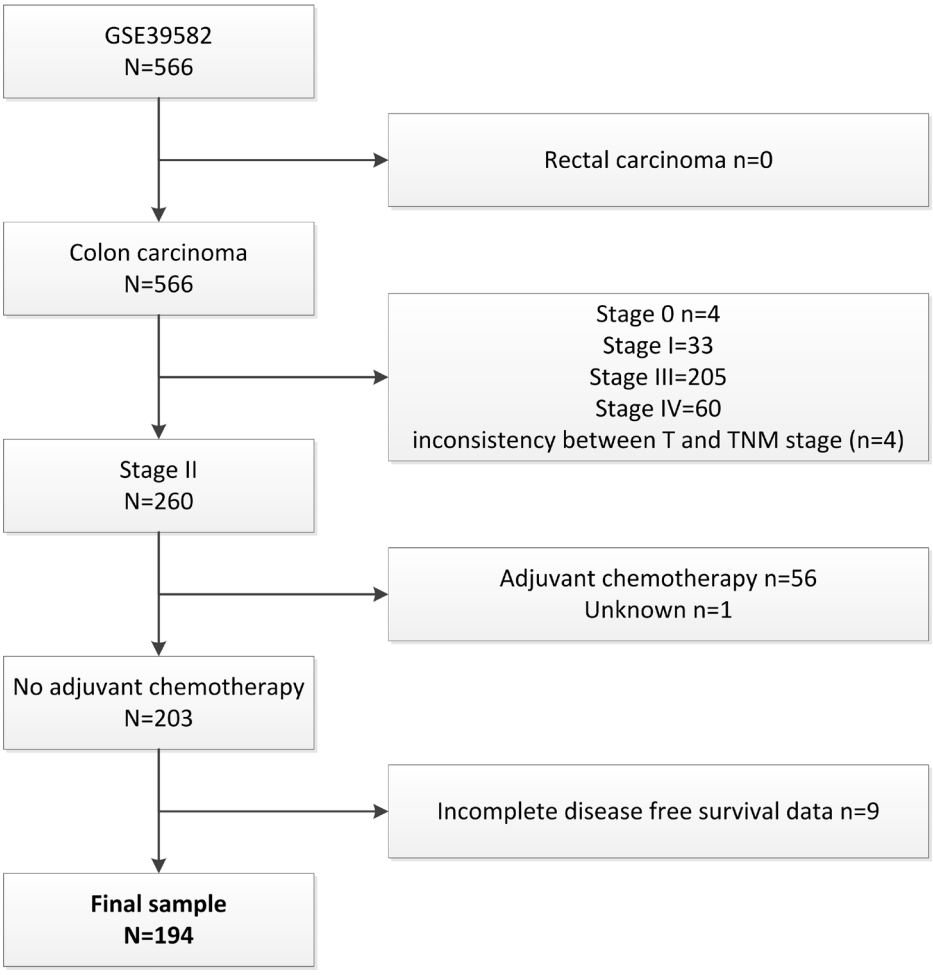


SUPPLEMENTARY FIGURE 2. PATIENT SELECTION IN THE MATCH STUDY

A total of 112 out of 271 patients with stage II colon cancer who were included in the first five years of the MATCH study were excluded because of receiving adjuvant chemotherapy (n=15), age <55 or >85 years (n=31), insufficient follow up (n=71), other survival determining malignancies (n = 6), a second primary colorectal tumor (n=2) and samples did not pass the quality checks in the lab(n=34)).



SUPPLEMENTARY FIGURE 3. PATIENT SELECTION IN THE GSE14333 COHORT
A total of 227 patients were excluded because of rectal cancer or unknown type (n=40), stage III-IV disease (n=168) and adjuvant chemotherapy (n=19).



SUPPLEMENTARY FIGURE 4. PATIENT SELECTION IN THE GSE39582 COHORT

A total of 372 patients were excluded due to stage 0, I, III or IV (n=302), inconsistency between T and TNM stage (n=4), adjuvant or unknown chemotherapy data (n=57), and incomplete DFS data (n=9).

SUPPLEMENTARY TABLE 1. BASIC CHARACTERISTICS FOR THE AGGREGATED COHORT (N=1,713)

	TOTAL		GSE14333		GSE17536		GSE2109		GSE37892		GSE39582		TCGA	
	n=1713		n=157		n=42		n=287		n=127		n=557		n=543	
Gender														
Female	799	46.6%	64	40.8%	19	44.4%	148	51.6%	60	47.2%	252	45.2%	256	47.1%
Male	914	53.4%	93	59.2%	23	54.8%	139	48.4%	67	52.8%	305	54.8%	287	52.9%
Age														
Median (interquartile range)	68.0 (59.0-76.0)		-		67.5 (59.5-78.0)		-		69.0 (59.0-77.0)		68.0 (59.0-76.0)		68.0 (58.0-77.0)	
Missing	445	26.0%	157	100%			287	100%			1	0.2%		
TNM														
I	292	17.0%	27	17.2%	7	16.7%	122	42.5%	-	-	38	6.8%	98	18.0%
II	698	40.7%	57	36.3%	18	42.9%	89	31.0%	73	57.5%	256	46.0%	205	37.8%
III	546	31.9%	45	28.7%	10	23.8%	73	25.4%	54	42.5%	203	36.4%	161	29.7%
IV	177	10.3%	28	17.8%	7	16.7%	3	1.0%	-	-	60	10.8%	79	14.5%
MSI														
MSS	887	51.8%	-	-	-	-	-	-	-	-	436	78.3%	451	83.1%
MSI	153	8.9%	-	-	-	-	-	-	-	-	75	13.5%	78	14.4%
Missing	673	39.3%	135	100%	36	100%	261	100%	116	100%	46	8.3%	14	2.6%
CMS														
1	255	14.9%	22	14.0%	6	14.3%	51	17.8%	11	8.7%	91	16.3%	74	13.6%
2	653	38.1%	55	35.0%	14	33.3%	96	33.4%	49	38.6%	228	40.9%	211	38.9%
3	215	12.6%	23	14.6%	3	7.1%	33	11.5%	19	15.0%	67	12.0%	70	12.9%
4	425	24.8%	35	22.3%	13	31.0%	81	28.2%	37	29.1%	125	22.4%	134	24.7%
Mixed/indeterminate	165	9.6%	22	14.0%	6	14.3%	26	9.1%	11	8.7%	46	8.3%	54	9.9%

SUPPLEMENTARY TABLE 2. DISTRIBUTION OF CMS PER TUMOR STAGE IN THE INDIVIDUAL COHORTS

	GSE14333 (N=22 MIXED/ INDETERMINATE)					GSE17536 (N=6 MIXED/ INDETERMINATE)					GSE2109 (N=26 MIXED/ INDETERMINATE)				
	Stage I	**Stage II**	**Stage III**	**Stage IV**	**Total**	**Stage I**	**Stage II**	**Stage III**	**Stage IV**	**Total**	**Stage I**	**Stage II**	**Stage III**	**Stage IV**	**Total**
CMS 1	2	10	8	2	22	0	2	3	1	6	23	18	9	1	51
	8.30%	21.30%	21.10%	7.70%	16.30%	0.00%	14.30%	33.30%	14.30%	16.70%	20.70%	22.20%	13.60%	33.30%	19.50%
CMS 2	11	17	16	11	55	2	8	3	1	14	37	36	23	0	96
	45.80%	36.20%	42.10%	42.30%	40.70%	33.30%	57.10%	33.30%	14.30%	38.90%	33.30%	44.40%	34.80%	0.00%	36.80%
CMS 3	7	9	4	3	23	2	0	0	1	3	14	11	8	0	33
	29.20%	19.10%	10.50%	11.50%	17.00%	33.30%	0.00%	0.00%	14.30%	8.30%	12.60%	13.60%	12.10%	0.00%	12.60%
CMS 4	4	11	10	10	35	2	4	3	4	13	37	16	26	2	81
	16.70%	23.40%	26.30%	38.50%	25.90%	33.30%	28.60%	33.30%	57.10%	36.10%	33.30%	19.80%	39.40%	66.70%	31.00%
Total	24	47	38	26	135	6	14	9	7	36	111	81	66	3	261
	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

	GSE37892 (N=11 MIXED/ INDETERMINATE)					GSE39582 (N=46 MIXED/ INDETERMINATE)					TCGA (N=54 MIXED/ INDETERMINATE)				
	Stage I	**Stage II**	**Stage III**	**Stage IV**	**Total**	**Stage I**	**Stage II**	**Stage III**	**Stage IV**	**Total**	**Stage I**	**Stage II**	**Stage III**	**Stage IV**	**Total**
CMS 1		3	8		11	8	47	32	4	91	13	41	16	4	74
		4.50%	16.30%		9.50%	22.90%	19.90%	17.30%	7.30%	17.80%	14.40%	21.60%	11.20%	6.10%	15.10%
CMS 2		36	13		49	17	110	81	20	228	47	68	61	35	211
		53.70%	26.50%		42.20%	48.60%	46.60%	43.80%	36.40%	44.60%	52.20%	35.80%	42.70%	53.00%	43.10%
CMS 3		9	10		19	7	30	23	7	67	21	30	15	4	70
		13.40%	20.40%		16.40%	20.00%	12.70%	12.40%	12.70%	13.10%	23.30%	15.80%	10.50%	6.10%	14.30%
CMS 4		19	18		37	3	49	49	24	125	9	51	51	23	134
		28.40%	36.70%		31.90%	8.60%	20.80%	26.50%	43.60%	24.50%	10.00%	26.80%	35.70%	34.80%	27.40%
Total		67	49		116	35	236	185	55	511	90	190	143	66	489
		100.00%	100.00%		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

The association between CMS4 and advanced stages of disease was consistently observed in the individual cohorts. A visual representation can be found in Supplemental Figure 2.

SUPPLEMENTARY TABLE 3.

Distribution of CMS per tumor stage for the MSS tumors of the Marisa and TCGA cohort (n=88 mixed/indeterminate).

	Stage I	Stage II	Stage III	Stage IV	Total
CMS 1	3	9	16	6	34
	3.1%	2.9%	5.7%	5.3%	4.3%
CMS 2	62	163	141	52	418
	63.9%	53.1%	50.2%	45.6%	52.3%
CMS 3	20	51	35	11	117
	20.6%	16.6%	12.5%	9.6%	14.6%
CMS 4	12	84	89	45	230
	12.4%%	27.4%	31.7%	39.5%	28.8%
Total	97	307	281	114	799
	100.0%	100.0%	100.0%	100.0%	100.0%

SUPPLEMENTARY TABLE 4.

Characteristics extended GSE33113 cohort.

	Total n=410	
Gender		
Female	205	50%
Male	205	50%
Age		
median (interquartile range)	68	(59-77)
TNM		
2	246	60%
3	164	40%
Ln Assessed		
median (range)	11	(0-100)
Ln Assessed		
< 10	159	39%
≥ 10	225	55%
missing	26	6%
Tumor location		
Left	250	61%
Right	155	38%
Both	5	1%

W.P. Kloosterman*, R.R.J. Coebergh van den Braak*, M. Pieterse*, M.J. van Roosmalen*, A.M. Sieuwerts*, C. Stangl, R. Brunekreef, Z.S. Lalmahomed, S. Ooft, A. van Galen, M. Smid, A. Lefebvre, F. Zwartkruis, J.W.M. Martens, J.A. Foekens, K. Biermann, M.J. Koudijs, J.N.M. IJzermans†, and E.E. Voest†.

* These authors contributed equally; † These authors share senior co-authorship.

5

A SYSTEMATIC ANALYSIS OF ONCOGENIC GENE FUSIONS IN PRIMARY COLON CANCER

CANCER RESEARCH 2017

ABSTRACT

Genomic rearrangements that give rise to oncogenic gene fusions can offer actionable targets for cancer therapy. Here we present a systematic analysis of oncogenic gene fusions among a clinically well-characterized, prospectively collected set of 278 primary colon cancers spanning diverse tumor stages and clinical outcomes. Gene fusions and somatic genetic variations were identified in fresh frozen clinical specimens by Illumina RNA-sequencing, the STAR fusion gene detection pipeline, and GATK RNA-seq variant calling. We considered gene fusions to be pathogenically relevant when recurrent, producing divergent gene expression (outlier analysis), or as functionally important (e.g., kinase fusions). Overall, 2.5% of all specimens were defined as harboring a relevant gene fusion (kinase fusions 1.8%). Novel configurations of *BRAF*, *NTRK3*, and *RET* gene fusions resulting from chromosomal translocations were identified. An R-spondin fusion was found in only one tumor (0.35%), much less than an earlier reported frequency of 10% in colorectal cancers. We also found a novel fusion involving *USP9X-ERAS* formed by chromothripsis and leading to high expression of *ERAS*, a constitutively active RAS protein normally expressed only in embryonic stem cells. This *USP9X-ERAS* fusion appeared highly oncogenic on the basis of its ability to activate AKT signaling. Oncogenic fusions were identified only in lymph node–negative tumors that lacked *BRAF* or *KRAS* mutations. In summary, we identified several novel oncogenic gene fusions in colorectal cancer that may drive malignant development and offer new targets for personalized therapy.

INTRODUCTION

Colorectal cancer is the third most common malignant disease in men and second in women with an estimated yearly incidence of 1.35 million new cases associated with 694,000 annual deaths.[1,2] Within colorectal cancer, colon cancer and rectal cancer are considered two separate disease entities that are treated differently.[3] This article focuses on primary nonmetastatic colon cancer (stage I to III) to avoid the identification of genetic aberrations that are a result of neoadjuvant treatment (e.g., in case of rectal cancer). The classical driver mutations of colon cancer have been studied extensively and consist of constitutive activation of the WNT pathway by mutations in the tumor suppressor *APC*, inactivation of *TP53* and activation of RAS/MAPK pathways through mutation of *RAS* family members.[4] These key primary drivers are sufficient for the transformation of primary colon stem cells into genomically instable adenocarcinomas.[5,6] Besides these classical driver genes, large-scale genomic screening has revealed hypermutated and nonhypermutated colorectal cancer, which contain a different repertoire of mutated genes.[7] Nonhypermutated colorectal cancers mostly carry mutations in the classical driver genes *APC*, *TP53*, *KRAS*, *PIK3CA*, and *SMAD4* and form the majority of colorectal tumors. Hypermutated colorectal cancers often harbor genetic changes in DNA mismatch repair genes along with mutations in *BRAF*, *APC*, *TGFBR2*, and *ACVR2A*.[7]

Genomic instability is a frequent hallmark of colorectal cancer, particularly of nonhypermutated tumors. Profiling of genomic copy number aberrations has revealed numerous recurrent changes, located at known fragile sites (*FHIT*, *WWOX*) or targeting tumor suppressors (*APC*, *PTEN*, *SMAD4*).[7] Genomic instability can lead to the formation of fusion genes.[8-10] Fusion genes have attracted significant attention because they were identified as potential cancer-specific targets for treatment. Several fusion genes (e.g., *BCR-ABL*, *EML4-ALK*) are clinically used to select patients for treatment.[9] One of the most prevalent fusion genes described in colorectal cancer involve the R-spondin family members *RSPO2* and *RSPO3*.[11] R-spondin fusions can activate WNT signaling and are mutually exclusive with mutations in *APC*. Recent work showed that inhibition of RSPO3 fusions impairs tumor growth.[12] Other recurrent fusions in colorectal cancer contain the *TCF7L1* and *TCF7L2* genes, encoding TCF3 and TCF4 transcription factors, although their relevance for colorectal cancer development is currently unknown.[7,13,14] Finally, a variety of kinase fusions have been observed in colorectal cancer, such as those involving *BRAF* or receptor tyrosine kinases.[10,11,15,16] Despite the growing support for a role of gene fusions in colorectal cancer development and their potential therapeutic value, small sample sizes, differences in experimental approaches, and the low frequency of fusions have

resulted in conflicting results regarding their prevalence and relevance. We report a comprehensive and unbiased screening for gene fusions in a unique, clinically well-defined, and prospectively collected cohort of 278 primary stage I to III colon cancers. We found that 2.5% of colon cancers in our dataset contained an oncogenic gene fusion and we identified novel fusions, including an *USP9X-ERAS* fusion with strong oncogenic activity *in vitro*.

MATERIALS AND METHODS

SAMPLE COLLECTION

Patients were selected from the MATCH study, a prospective multicenter cohort study from 2007 onwards including adult patients undergoing curative surgery in one of seven hospitals in the Rotterdam region, the Netherlands (institutional review board number MEC-2007-088). All patients gave written informed consent for the storage and use of tissue samples for research purposes, and the collection of clinical data. The study has been conducted in accordance with the guidelines of the Declaration of Helsinki.

Only samples with at least 40% invasive tumor cells were included in the final analysis, with the number of samples per RNA sequencing run depending upon this percentage. DNA and RNA was isolated from 30-mm sections taken from the frozen tumor tissue obtained at primary surgery. Only samples with an RNA integrity value of at least 7.0 were included in the final analysis.

RNA-SEQUENCING

Total RNA (500 ng) from tumor samples was used as input for the Illumina TruSeq stranded RNA-seq protocol. Libraries were pooled and sequenced on Illumina HiSeq2500 or NextSeq instruments. We used the STAR fusion gene detection pipeline (version STAR-2.4.1) for analysis of RNA-seq data.[17]

A list of filtered junctions was annotated by adding fusion gene counts, donor gene counts, acceptor gene counts, overlap with protein domains and calculation of expression z-scores for the donor and acceptor genes relative to samples without the fusion. GATK RNA-seq variant calling best practices were used for somatic variant calling in RNA-seq data.

MATE-PAIR SEQUENCING AND ANALYSIS

Mate-pair library preparation was done using the Illumina Nextera Mate Pair library kit. Libraries were sequenced on NextSeq using 2*75 bp configuration. Discordant

read pairs were detected from BAM files using a custom analysis pipeline as described previously.[18] FREEC was used to detect copy number variations.[19]

WHOLE EXOME SEQUENCING

DNA was sequenced by GATC Biotech using Illumina's HiSeq protocol (paired-end 100 bp, captured regions according to SureSelect v5).

Raw sequence data for both normal and tumor DNA were mapped to the human reference genome GRCh37 using BWA (v 0.7.5a-r405).[20] BAM files were used for variant calling using GATK (v3.3.0) and Mutect (v1.1.6), followed by custom filtering steps.[21,22]

FUSION GENE EXPRESSION AND WESTERN BLOTTING

HEK293T cells or NIH-3T3-A14 cells were transfected with 2 mg of pBABE constructs containing fusion genes by the calcium phosphate method. NIH-3T3-A14 cells were obtained from Burgering and colleagues in 2014.[23] HEK293T cells were obtained from ATCC in the late 1980s. Both cell lines have only been tested and authenticated on the basis of their morphologic appearance. The cell lines were cultured for up to ten passages after thawing before use in experiments. Mycoplasma testing was done every three months using MycoAlert (Lonza). Cells were cotransfected with constructs encoding either MYC-tagged ERK1 or GFP-tagged AKT. Western blotting was done with rabbit polyclonal antibodies directed against phospho-AKT (Ser473; D9E, Cell Signaling Technology), phospho-p44/42 MAPK (ERK1/2, Thr202/Tyr204, Cell Signaling Technology), c-MYC (910E, Santa Cruz Biotechnology), and mouse monoclonal anti-α-Tubulin (CP06, Calbiochem). Detection was done with fluorescently labeled secondary antibodies (goat-anti-rabbit IgG (H+L) 800 CW and donkey-anti-mouse (680 RD) from LI-COR Biosciences).

DATA ACCESS

The sequencing data described in this study can be accessed through the European Genome Phenome Archive under accession number EGAS00001002197.

A detailed description of materials and methods can be found in the Supplementary Materials and Methods.

RESULTS

TRANSCRIPTOME SEQUENCING OF 278 PRIMARY COLON TUMORS

To identify rearranged transcripts we explored the transcriptomic profile of colon cancers, making use of a large prospectively collected cohort of primary tumor samples from patients with stage I to III colon cancer (Rotterdam MATCH study). Primary tumor samples were selected from our database, based on clinical and technical criteria (**Supplementary Figure 1**). The study cohort included stage I (n = 66), stage II (n = 115) and stage III (n = 97) pathologically confirmed adenocarcinomas (**Table 1**). Detailed clinical description including follow-up time, adjuvant therapy, disease outcome, tumor stage, lymph node status, histologic data, and patient details was collected (**Supplementary Table 1**).

Tumor samples with at least 40% invasive tumor cells (based on histologic examination) were sectioned and tissue slices were consecutively used for RNA and DNA isolation. Total RNA was extracted from tissue slices and used for the preparation of RNA-seq libraries subsequent to removal of abundant noncoding mRNAs (ribominus). Libraries were sequenced at a mean depth of 27M paired reads per library in either 2100*bp or 2*75 bp configuration (**Supplementary Table 1**). Next, the data were mapped to the human reference genome (GRCh37) to identify discordantly mapping reads indicating potential somatic fusion genes using STAR software and a custom annotation pipeline.[17] The pipeline parameters were set to achieve maximal sensitivity, leading to the prediction of 3 million raw potential fusion gene calls. These calls were subsequently filtered through a series of rational filtering steps (**Figure 1**), including read coverage, removal of paralogous gene sets, filtering against control data, prediction of in-frame fusion and recurrence among tumor samples.[16] Using these specific filtering criteria, we obtained a dataset of 75 fusion genes, which were subjected to experimental verification by RT-PCR (**Supplementary Figure 2; Supplementary Table 2**). A total of 22 out of 75 tested fusion genes were validated in the correct tumor specimen and were absent in the corresponding control tissue. We observed two cases where a *TFG-GPR128* fusion was present in both tumor and corresponding normal colon tissue. This fusion has previously been described in renal cell cancer and was later shown to be caused by a germline structural genomic variation.[24,25] For a *DLG1-BRAF* fusion, we also observed a weak RT-PCR product in the corresponding control tissue, which is likely a result of contamination of the control tissue with tumor cells. For 35 predicted fusion genes, we did not observe an RT-PCR product, indicating that these are either false positive calls or that the fusion gene RT-PCR detection assay was suboptimal. The remainder of fusion genes was not specific for the tumor sample. To assess whether our filtering strategy indeed enriched for true positive fusion genes, we also subjected

TABLE 1. PATIENT AND TUMOR CHARACTERISTICS

CHARACTERISTICS	ALL PATIENTS		LYMPH NODE NEGATIVE TUMORS		LYMPH NODE POSITIVE TUMORS		NO FUSION GENE		FUSION GENE	
	N=278	(%)	N=181	(%)	N=97	(%)	N=271	(%)	N=7	(%)
Gender										
Female	132	47.5%	92	50.8%	40	41.2%	127	46.9%	5	71.4%
Male	146	52.5%	89	49.2%	57	58.8%	144	53.1%	2	28.6%
Age										
Median (IQR)	68.2 (62.4 - 75.2)		70.2 (63.2 - 76.9)		70.0 (61.4 - 70.8)		68.1 (62.4-75.2)		71.1 (67.4-76.1)	
Tumor stage										
Stage I	66	23.7%	66	36.5%	-	-	65	24.0%	1	14.3%
Stage II	115	41.4%	115	63.5%	-	-	109	40.2%	6	85.7%
Stage III	97	40.2%	-	-	97	100%	97	35.8%	0	0%
T status										
T2	79	28.4%	66	36.5%	13	13.4%	78	28.8%	1	14.3%
T3	194	69.8%	110	60.8%	84	86.6%	190	70.1%	4	57.1%
T4	5	1.8%	5	2.8%	-	-	3	1.1%	2	28.6%
Nodal status										
N0[a]	148	53.2%	148	81.8%	-	-	143	52.8%	5	71.4%
N0[b]	33	11.9%	33	18.2%	-	-	31	11.4%	0	0%
N1	64	23.0%	-	-	63	64.9%	64	23.6%	0	0%
N2	33	11.9%	-	-	34	35.1%	33	12.2%	2	28.6%
N0	181	65.1%	181	100%	-	-	174	64.2%	7	100%
N+	97	34.9%	-	-	97	100%	97	35.8%	0	0%
Tumor grade										
Good	23	8.3%	16	8.8%	7	7.2%	23	8.5%	0	0%
Moderate	220	79.1%	152	84.0%	68	70.1%	213	78.6%	7	100%
Poor	24	8.6%	10	5.5%	14	14.4%	24	8.9%	0	0%
Unknown	11	4.0%	3	1.7%	8	8.2%	11	4.1%	0	0%
Adjuvant therapy										
No	182	65.5%	182	100%	1	1.0%	175	64.6%	7	100%
Yes	96	34.5%	-	-	95	99.0%	96	35.4%	0	0%
MSI-status										
MSS	217	78.1%	134	74.0%	83	85.6%	215	79.3%	2	28.6%
MSI	61	21.9%	45	24.9%	14	14.4%	56	20.7%	5	71.4%
Location										
Left	133	47.8%	84	46.4%	49	50.5%	130	48.0%	3	42.9%
Right	145	52.2%	97	53.6%	48	49.5%	141	52.0%	4	57.1%

Abbreviation: MSI, microsatellite instability. [a] Total lymph node yield ≥ 10. [b] Total lymph node yield < 10.

a series of 70 predicted fusions that did not pass our filtering steps to experimental verification by RT-PCR. Out of these 70 predicted fusions none could be confirmed (data not shown).

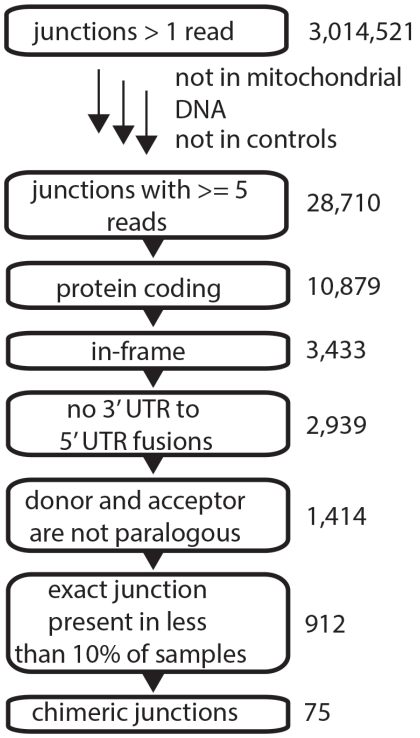


FIGURE 1. OVERVIEW OF FUSION GENE DETECTION AND FILTERING APPROACH

BRAF FUSIONS ARE RECURRENT AND PRESENT AT LOW FREQUENCY IN COLON CANCER

We next searched among the validated fusion transcripts for those, which were recurrent and in-frame. We identified three unique fusions (1.1% of 278 samples) involving the *BRAF* oncogene (*AGAP3-BRAF*, *TRIM24-BRAF*, *DLG1-BRAF*, **Figure 2A**; **Supplementary Figure 3**). *BRAF* fusions have been described in a variety of cancer types.[10,16] The structure of the *TRIM24-BRAF* fusion was identical to those reported previously, with exon 3 of *TRIM24* connected to exon 10 of *BRAF*.[15] The *AGAP3-BRAF* fusion contained a junction between exon 8 of *AGAP3* and exon 9 of *BRAF*, which is different from the exon 9-exon 9 configuration previously described.[15] The *DLG1-BRAF* fusion, containing a junction between exon 5 of *DLG1* and exon 9 of *BRAF*, is novel and extends the broad spectrum of known *BRAF* fusions in cancer.

We sequenced the genomes of the tumor samples with *BRAF* fusion genes using large-insert mate-pair sequencing (insert size 2.5 kb) to detect somatic structural variations that could account for *BRAF* fusion formation. In two cases (*AGAP3-BRAF* and *TRIM24-BRAF*) the fusion was caused by an inversion event, while the novel *DLG1-BRAF* fusion resulted from a reciprocal translocation between chromosomes 3 and 7 (**Figure 2A**).

All three fusion genes contained the entire C-terminal kinase domain of *BRAF* by fusion of exon 9 or 10 to their respective donor genes. We hypothesized that disconnection of the *BRAF* kinase domain from its N-terminal autoinhibitory domain leads to constitutive activation.[26] To assess whether our *BRAF* fusions can activate signaling pathways, we cloned the *DLG1-BRAF* and *AGAP3-BRAF* fusion genes and expressed them in HEK293 cells. Subsequent analysis for activation of ERK/MAPK signaling was performed by coexpression of ERK1. Protein analysis revealed a strong effect of BRAF fusion proteins on ERK1 phosphorylation, underscoring their role as oncogenes in colon cancer (**Figure 2B**). Although the effect of BRAF fusions on ERK1 phosphorylation appeared stronger than for the native BRAF protein, the effect was less strong than for BRAF carrying the activating V600E mutation.

IDENTIFICATION OF *NTRK3* AND *RET* KINASE FUSION GENES

To further evaluate the relevance of the remaining fusion genes that were verified by RT-PCR, we reasoned that fusions that lead to upregulation of the acceptor gene may be of particular importance.[27] Therefore, we analyzed the expression of our entire set of fusions and compared the expression of the donor and acceptor genes to all other tumor samples without such a fusion (outlier analysis, **Supplementary Table 2**).

An *EML4-NTRK3* fusion was among the top hits that resulted from this analysis with an expression Z score of 17.96 for the *NTRK3* gene. This fusion was formed through a reciprocal translocation that joined the 50 part of *EML4* (ending with exon 2) with the 30 exons of *NTRK3* (starting with exon 14) in a lymph node negative adenocarcinoma (**Figure 2A**; **Supplementary Figure 4A**). An additional *NTRK3* fusion (*ETV6-NTRK3*) has been reported previously in colon cancer and an *EML4-NTRK3* fusion has been observed in glioma.[11,28] By examining the expression of the individual exons across *ETV6-NTRK3*, we noticed that the fusion also leads to an increased expression of the exons encoding the tyrosine kinase domain, which is retained in the fusion transcript (**Supplementary Figure 4B**). On the basis of these results, we analyzed the exonic expression in 732 RNA-sequencing datasets of colorectal cancer samples from The Cancer Genome Atlas (TCGA) and observed a similar increase in expression of the kinase encoding exons in two datasets derived from colon adenocarcinomas, suggesting the presence of *NTRK3* fusion genes (**Supplementary Figure 4C**).[7]

Neurotrophin tyrosine kinase (NTRK) 1 and 3 are receptor kinases that are frequently activated by gene fusion in a variety of cancers.[10] The tyrosine kinase domain is always maintained in the chimeric proteins and fused to an oligomerization domain provided by the N-terminal fusion partner. To assess the molecular effects of the *EML4-NTRK3* fusion gene reported here, we expressed it in HEK293 cells together with *ERK1* and found that the *EML4-NTRK3* fusion activates MAPK/ERK signaling by phosphorylation of ERK1 (**Figure 2B**). A truncated version of the fusion gene was not active, suggesting that the EML4 coiled-coil domain (CCD) is supposed to promote receptor activation by dimerization, similar as for *EML4-ALK* fusions found in lung cancer.[29] We also tested the same fusion construct in the context of A14 cells cotransfected with AKT. Following serum starvation, we observed phosphorylation of AKT exceeding the levels of AKT phosphorylation by KRAS V12A under the same conditions (**Figure 2C**). Altogether, we conclude that the *EML4-NTRK3* fusion affects oncogenic signaling pathways and that *NTRK3* fusion genes are recurrent but low-frequent in colorectal cancer.

Another top candidate with a high expression Z-score involved an in-frame fusion with exon 1–9 of the integral endoplasmic reticulum membrane protein Ribosome-binding protein 1 (*RRBP1*) fused to exons 12–20 of the *RET* gene, harboring the complete N-terminal kinase domain (**Supplementary Figure 5A; Supplementary Figure 5B**). The N-terminal part of RRBP1 contains the ribosome receptor lysine/proline domain as well as a coiled coil domain (CCD). Previously reported fusions of *RET* to CCDs of 50 partners have been shown to initiate ligand-independent activation of the kinase domain, suggesting a similar mechanism in this fusion.[30,31] The *RET* gene is a known target for gene fusions in hereditary and sporadic papillary thyroid cancers and lung adenocarcinoma, and *RET* fusions have recently also been described in advanced colorectal cancer.[32-34]

In addition, we observed a fusion involving the kinase gene *PSKH2*, which is highly expressed in the sample with the fusion, but not at all in other tumor samples (**Supplementary Table 2**). However, this fusion was not pursued further because we observed several different splice variants with only partial open reading frames.

ERAS ACTIVATION THROUGH GENE FUSION IN COLON CANCER

One particularly interesting novel fusion gene that resulted from our outlier expression analysis contained the *ERAS* gene (**Figure 3A and B**). *ERAS* is a single-exon *RAS*-family member that is expressed only in embryonic stem cells.[35] The ERAS protein is constitutively active and leads to enhanced PI3K signaling and cellular transformation. Elevated *ERAS* expression has been described in some gastric cancer samples and a role in tumorigenesis has been implied.[36] We observed that *ERAS* was

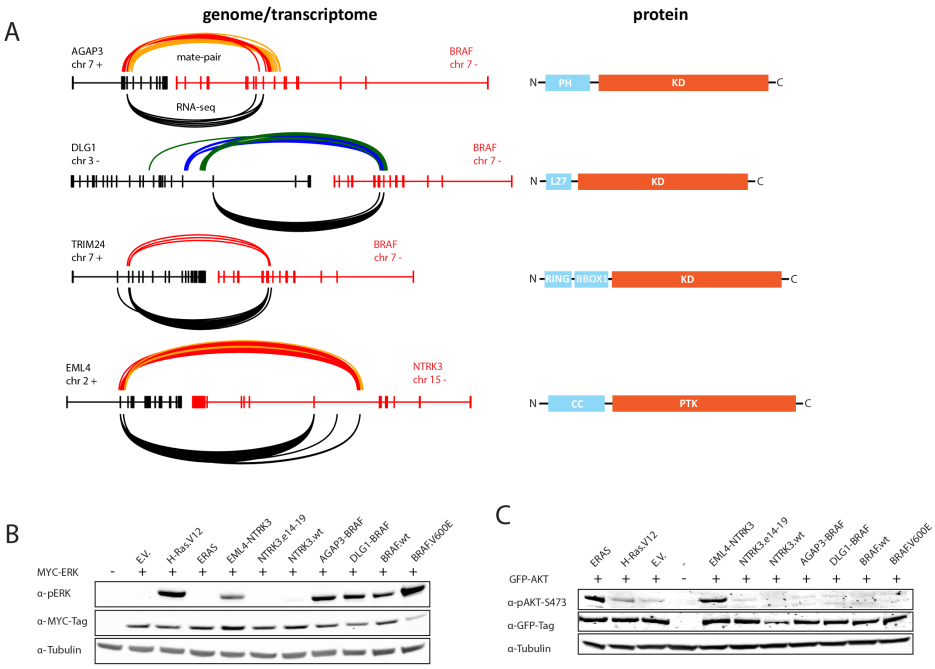


FIGURE 2. STRUCTURE AND CHARACTERIZATION OF FUSION GENES.
(**A**) Exon and protein structure of the TRIM24-BRAF, AGAP3-BRAF, DLG1-BRAF, and EML4-NTRK3 fusions. On top of the exonic structures, we plotted arcs indicating breakpoint junction sequence reads detected by mate-pair sequencing of tumor genomicDNA. Coloring of the arcs indicates orientation of the breakpoint junction as indicated by the respective mate-pair reads: red, head-to-head inverted; yellow, tail-to-tail inverted; blue, tail-to-head; green, head-to-tail. Below the exonic structures, chimeric RNA-seq reads are plotted (black arcs) indicating which exon–exon connections were observed from the sequence data. KD, kinase domain; L27, L27 protein interaction module; PH, pleckstrin homology; RING, zinc finger domain ring type; BBOX1, B-box-type zinc finger domain; PTK, protein tyrosine kinase domain; CC, coiled-coil domain. (**B**) Western blot results depicting the effects of fusion genes overexpression in HEK293 cells on ERK1 phosphorylation. (**C**) Western blot results displaying the effects of fusion gene overexpression on AKT phosphorylation in NIH-3T3-A14 cells.

highly expressed in one tumor sample in our dataset and no detectable expression was observed in the other tumor samples (**Figure 3B**). The high expression was driven by the fusion of *ERAS* with *USP9X*, a highly expressed housekeeping gene (**Figure 3A; Supplementary Figure 6A**). As opposed to canonical fusion genes, which often involve formation of a novel chimeric protein sequence, the *USP9X-ERAS* fusion was formed by fusion of 5'UTR sequences, which leads to an exchange of the *ERAS* promoter with the *USP9X* promoter and not the formation of a novel chimeric protein sequence.

To gain insight in the formation of the *USP9X-ERAS* fusion, we analyzed structural variations using mate-pair sequencing. This revealed that the fusion gene was caused

at the genomic level by a highly local chromothripsis event on chromosome X spanning solely the region covered by *USP9X* and *ERAS* (**Figure 3C**). The chromothripsis involved at least 18 genomic breakpoint junctions and led to multiple copy number changes. We cloned the *USP9X-ERAS* fusion gene and expressed it in NIH-3T3 A14 cells. Analysis of phosphorylated AKT showed that the *USPX9-ERAS* fusion can activate AKT signaling (**Figure 2C**). To get further support for a potential role of ERAS expression in cancer development, we assessed 521 RNA-seq datasets from colon cancer from the TCGA consortium (**Supplementary Table 3**). This revealed several colon cancer datasets showing detectable mRNA expression levels of *ERAS* (**Supplementary Figure 6B**), albeit not as high as the sample with the *USP9X-ERAS* fusion described here. We also assessed stomach cancer RNA-seq datasets from TCGA and observed one sample with high expression of *ERAS* (**Supplementary Figure 6C**). Altogether, our data suggest that induction of *ERAS* expression could be an alternative mode of promoting oncogenesis through the AKT pathway in colon cancer.

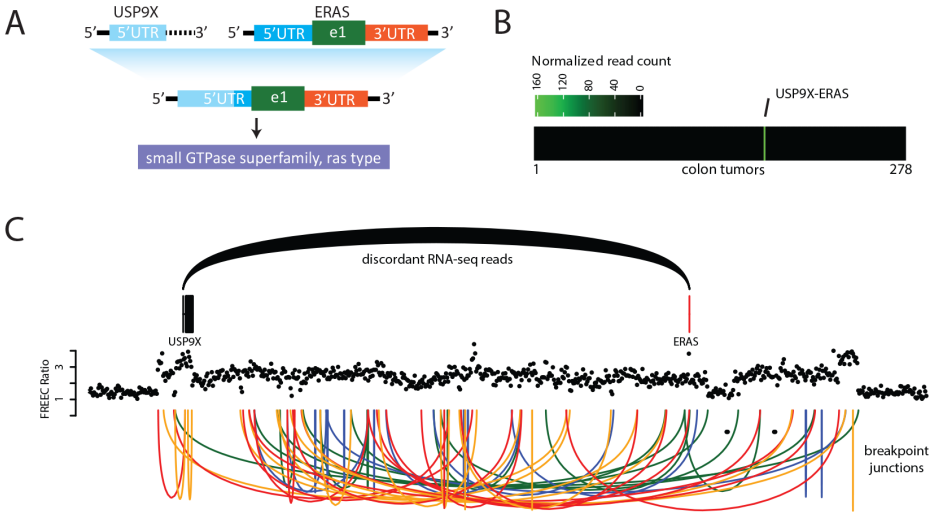


FIGURE 3. GENOMIC ORIGIN, STRUCTURE, AND EXPRESSION OF A NOVEL *USP9X-ERAS* FUSION GENE

(**A**) Schematic drawing indicating the transcript structure of the *USP9X-ERAS* fusion. The fusion was caused by a breakpoint junction in the 5′UTR of *USP9X* and *ERAS*, resulting in control of *ERAS* by the *USP9X* promoter. (**B**) *ERAS* expression levels across the entire cohort of 278 colon tumors included in this study. (**C**) RNA-seq and mate-pair sequencing data across chromosomal regions involving the *ERAS* and *USP9X* genes. Individual chimeric RNA-seq reads are depicted as black arcs. Genomic breakpoint junctions (not individual sequence reads) are shown as colored arcs (red, head-to-head inverted; yellow, tail-to-tail inverted; blue, tail-to-head; green, head-to-tail). The genomic copy number profile is displayed using black dots, which each represent the copy number of a genomic interval as determined based on analysis of mate-pair data using FREEC.

LOW FREQUENCY OF KNOWN R-SPONDIN FUSIONS

Previous work reported a number of different fusion genes in colon cancer most prominently those that involve genes that interact with the WNT signaling pathway.[11,13] Fusions involving the R-spondin genes *RSPO2* and *RSPO3* have been reported in up to 10% of microsatellite stable (MSS) colorectal cancers in one study and appear mutually exclusive with mutations in *APC*.[11] To achieve maximal sensitivity for picking up gene fusions, we evaluated our raw fusion gene calls for the presence of both types of RSPO fusion genes, but could only detect one *EIF3E-RSPO2* fusion in an MSS sample (**Supplementary Figure 7A**). To verify the sensitivity of our pipeline for picking up *RSPO2* and *RSPO3* fusion genes, we reanalyzed the raw RNA-seq FASTQ files as published recently using our STAR-based pipeline.[11] Our bioinformatics pipeline could detect all seven published fusions. In addition, we measured normalized read depth across the *RSPO2* and *RSPO3* genes, revealing a strong upregulation of expression for samples with the corresponding R-spondin fusion in the published tumor samples (**Supplementary Figure 7B**). We only observed elevated *RSPO2* expression for the one tumor sample in our cohort that showed the presence of an *EIF3E-RSPO2* fusion (**Supplementary Figure 7C**), further supporting the low frequency of *RSPO* fusion genes in our dataset. We conclude that R-spondin fusions may not be as frequently present as previously indicated or that sampling bias, selection bias or treatment regime may explain the observed discrepancies.

ONCOGENIC FUSIONS ARE MUTUALLY EXCLUSIVE WITH ACTIVATING MUTATIONS IN *KRAS*, *BRAF*, AND *NRAS* AND RESTRICTED TO STAGE I AND II TUMORS

We used the GATK-RNAseq mutation calling pipeline to detect indels and single-nucleotide changes in the RNA-seq data from all 278 tumor samples (stage I–III). The analysis was focused on cancer genes that are of major relevance for colon cancer development, including *BRAF*, *KRAS*, *HRAS*, *NRAS*, *SMAD4*, *TP53*, *APC*, and *PTEN*. Passed variant calls in *BRAF*, *KRAS*, *HRAS*, and *NRAS* were overlapped with known hotspot mutations from the COSMIC database.[37] For mutations in tumor suppressor genes, we filtered the variants against existing databases of germline variants to enrich for somatic variants. To estimate the reliability of the RNA-based variant calls, we compared them against paired tumor–normal exome sequencing data that were generated for a subset of 44 samples. For *BRAF*, *KRAS*, *HRAS*, and *NRAS*, all variants identified in the RNA-seq data were also found in the exome data and false negatives were not observed. On the basis of the RNA-seq variant calls, we identified 27 *BRAF* V600E mutations in the entire cohort of 278 tumors, which is in line with the estimated frequency (10%) of this mutation type in colorectal cancer.[7] Our analysis of fusion

genes showed that activation of *BRAF* may additionally be caused by fusion gene formation in an additional 1.1% of colon cancers (**Figure 4**).

Besides mutations in *BRAF*, we also found 103 tumors with a hotspot mutation in *KRAS* (n = 99.36%) and *NRAS* (n = 4, 1.4%). In line with previous observations in other cancer types, we observed that the presence of MAPK/ERK and PI3K/AKT activating hotspot mutations in *BRAF*, *KRAS*, and *NRAS* are mutually exclusive with the presence of oncogenic fusion genes in colon cancer (P = 0.018).[15,16] Finally, we noted that all oncogenic fusions (including *EIF3E-RSPO2*) were found in samples with lymph node–negative stage I and II tumors (P = 0.047) and none of the samples showed a relapse in subsequent years (median follow up 50.9 months). However, the latter results should be interpreted with caution due to the small numbers.

DISCUSSION

Our comprehensive analysis of RNA sequencing data from 278 well-characterized stage I to III colon cancers yielded a number of known and novel fusion genes, which may have clinical implications. In the era of personalized medicine, tumors are increasingly molecularly profiled, leading to better identification of patients for specific treatments.[38] For colorectal cancer, small gene panels including *BRAF*, *KRAS*, and *NRAS* are most often used since mutations in these genes are of clinical relevance.[39] Our analyses show that beyond these single gene tests, fusion genes may also be important.

Three of the fusion genes identified in our cohort involve the *BRAF* oncogene, which has previously been found in 4 (0.2%) colon cancer samples out of 2,154 colorectal cancer samples.[15] Here we show that *BRAF* fusions occur in 1.1% of stage I–III colon cancers. Two of the *BRAF* fusions (*AGAP3-BRAF* and *TRIM24-BRAF*) consist of know fusion configurations, while the *DLG1-BRAF* fusion is novel.[15] The *BRAF* fusions activate oncogenic signaling pathways in cells lines, indicating that they form genuine oncogenes in colon cancer, in addition to known oncogenic mutations in *BRAF* and *KRAS*. Although *BRAF* fusions are relatively rare, they may be highly relevant drug targets for the individual patient, similar as mutations in *BRAF*.[40-42]

An expression outlier analysis involving samples with and without fusion genes, revealed *EML4-NTRK3*, *RRBP1-RET*, and *USPX9-ERAS* fusion genes. The *EML4-NTRK3* fusion gene has not been described in colon cancer, but was reported in a single case of glioma.[28] However separately, both the *EML4* and *NTRK3* gene have been described as part of gene fusions in various types of cancers.[10] *EML4* has mainly been described in conjunction with the ALK kinase gene in non–small cell lung cancer

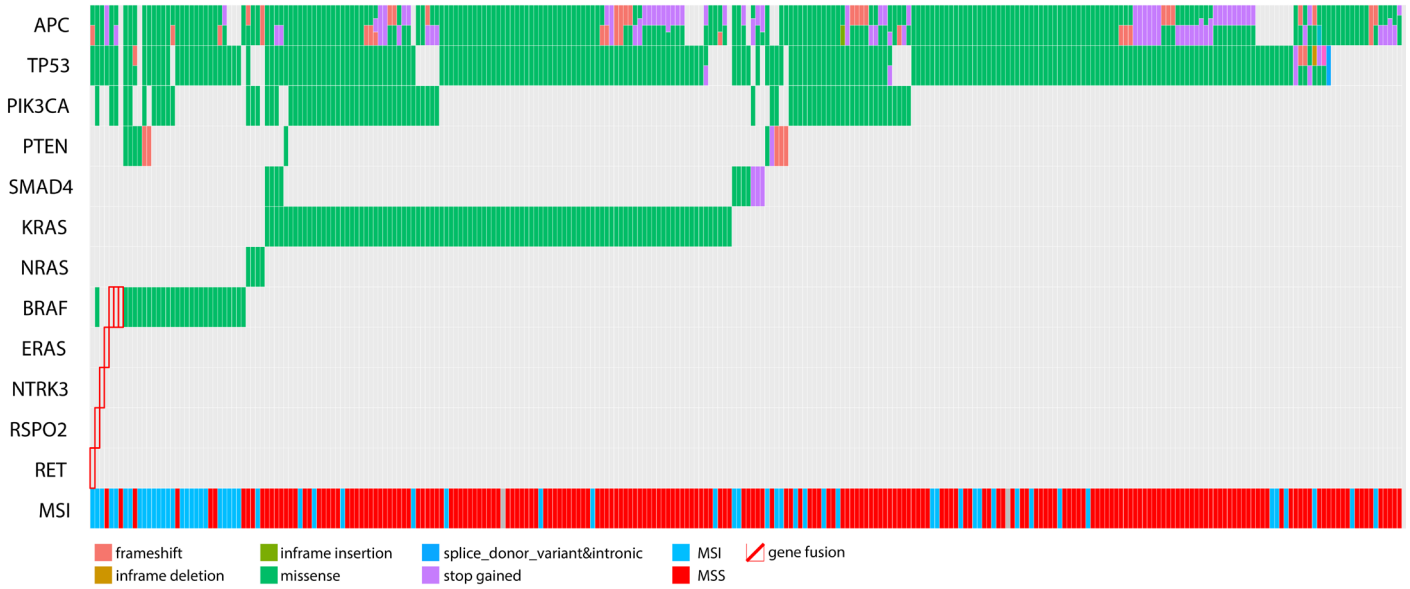


FIGURE 4. SCHEMATIC OVERVIEW OF MUTATION STATUS, MICROSATELLITE INSTABILITY (MSI) STATUS, AND PRESENCE OF FUSION GENES IN COLON TUMORS
Mutations were called for a selected set of known colon cancer genes based on RNA-seq data and intersection with COSMIC mutations. Mutation types are indicated with different colors according to their predicted effect.

occurring in five different variants.[43] All of these contain a CCD, which is responsible for the dimerization and constitutive activation of its acceptor gene product. This is consistent with our findings that the *EML4-NTRK3* fusion induces ERK1 and AKT phosphorylation, while expression of a truncated version of *NTRK3* or the entire *NTRK3* gene did not reveal such activity.

RET fusions have been described in up to one-third of papillary thyroid cancers, in 2% of lung adenocarcinoma and recently in 0.2% of 3,117 advanced colorectal tumors.[32-34] Tumors carrying a *RET* fusion in that colorectal cancer cohort were pan negative for known driver mutations such as *KRAS*, *BRAF*, *PIK3CA*, and *EGFR*, which was also true for the tumor carrying the *RRBP1-RET* fusion in our cohort. RET kinase inhibitors might form a promising treatment for colorectal cancers containing oncogenic *RET* fusions.[33]

An entirely novel fusion gene described in this work, comprises the *USP9X* and *ERAS* genes. Although this fusion has only been found in a single colon cancer sample in our study, its high expression and *in vitro* activity demonstrate that expression of ERAS has strong oncogenic capacity. We observed *ERAS* expression in colon cancer RNA-seq datasets from TCGA, suggesting that *ERAS* expression could be a recurrent oncogenic mechanism in colon cancer, similarly as has been proposed for stomach cancer.[36]

R-spondin fusions were described as a recurrent genomic aberration in colon cancer patients by Seshagiri and colleagues, whom identified seven R-spondin fusions in a cohort of 74 colon cancer patients (9.5%).[11] In their cohort, tumors with an R-spondin fusion did not contain a loss of function mutation in *APC* or copy loss, except for one tumor, which contained a single *APC* allele. Five out of seven R-spondin fusions occurred in a tumor with a *KRAS* mutation (13.5% of all KRAS mutant tumors) and two in a tumor carrying a *BRAF* mutation (40% of all *BRAF* mutant tumors). In our cohort of 278 patients we observed only one R-spondin fusion, which was present in a tumor sample carrying a *BRAF* mutation (3.7% of all *BRAF* mutated tumors). However, the percentage of *KRAS* mutated tumors differed substantially between the cohort of Seshagiri and our cohort (*KRAS* 50% vs. 35.6% P = 0.024 and *BRAF* 6.8% vs. 9.7% P = 0.43, respectively). The presence of R-spondin fusions in a subset of colorectal adenomas (traditional serrated adenoma) with frequent *KRAS* mutations has been recently shown.[44] These data suggest that differences between tumor cohorts may explain the differences in the total number of identified R-spondin fusions.

Our findings are in line with new insights that broader and systematic use of genetic profiling including DNA and RNA sequencing is needed to maximize identification of patients that could potentially benefit from targeted treatment.[45] Sharing of datasets including clinical characteristics and treatment outcome, such as our dataset, may help to overcome sample size limitations of individual studies and improve insight

into the clinical merit of specific infrequent genetic aberrations and fusion genes.[46] We found that oncogenic fusion genes were present in lymph node– negative tumors, although this finding needs to be substantiated in larger studies. Most of the previous studies reporting fusion genes in colorectal cancer did not include clinical or histopathologic characteristics, especially not stage.

In conclusion, we have created a large and comprehensive catalog of fusion genes in a unique clinically well-defined prospectively collected cohort of stage I to III primary colorectal cancers and identified several known and novel fusion genes with biological activity and possible prognostic value. We anticipate that incorporating *in vitro* platforms such as (tumor)organoids may facilitate testing of fusion genes for functional relevance, differences in oncogenic capacity and response to antitumor drugs.[5]

REFERENCES

1. GLOBOCAN. (Accessed 1st November 2016, at http://globocan.iarc.fr/Default.aspx.)

2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65:87-108.

3. Tamas K, Walenkamp AM, de Vries EG, et al. Rectal and colon cancer: Not just a different anatomic site. Cancer Treat Rev 2015;41:671-9.

4. Fearon ER. Molecular genetics of colorectal cancer. Annu Rev Pathol 2011;6:479-507.

5. Drost J, van Jaarsveld RH, Ponsioen B, et al. Sequential cancer mutations in cultured human intestinal stem cells. Nature 2015;521:43-7.

6. 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.

7. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-7.

8. Kumar-Sinha C, Kalyana-Sundaram S, Chinnaiyan AM. Landscape of gene fusions in epithelial cancers: seq and ye shall find. Genome Med 2015;7:129.

9. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 2015;15:371-81.

10. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. Nat Commun 2014;5:4846.

11. Seshagiri S, Stawiski EW, Durinck S, et al. Recurrent R-spondin fusions in colon cancer. Nature 2012;488:660-4.

12. Storm EE, Durinck S, de Sousa e Melo F, et al. Targeting PTPRK-RSPO3 colon tumours promotes differentiation and loss of stem-cell function. Nature 2016;529:97-100.

13. Bass AJ, Lawrence MS, Brace LE, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet 2011;43:964-8.

14. Nome T, Hoff AM, Bakken AC, Rognum TO, Nesbakken A, Skotheim RI. High frequency of fusion transcripts involving TCF7L2 in colorectal cancer: novel fusion partner and splice variants. PLoS One 2014;9:e91264.

15. Ross JS, Wang K, Chmielecki J, et al. The distribution of BRAF gene fusions in solid tumors and response to targeted therapy. Int J Cancer 2016;138:881-90.

16. Yoshihara K, Wang Q, Torres-Garcia W, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene 2015;34:4845-54.

17. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.

18. Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, et al. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. Cell Rep 2012;1:648-55.

19. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28:423-5.

20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

21. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.

22. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.

23. Burgering BM, Medema RH, Maassen JA, et al. Insulin stimulation of gene expression mediated by p21ras activation. EMBO J 1991;10:1103-9.

24. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499:43-9.

25. Chase A, Ernst T, Fiebig A, et al. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. Haematologica 2010;95:20-6.

26. Ciampi R, Knauf JA, Kerler R, et al. Oncogenic AKAP9-BRAF fusion is a novel mechanism of MAPK pathway activation in thyroid cancer. J Clin Invest 2005;115:94-101.

27. Giacomini CP, Sun S, Varma S, et al. Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. PLoS Genet 2013;9:e1003464.

28. Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol 2015;33:306-12.

29. Choi YL, Takeuchi K, Soda M, et al. Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer. Cancer Res 2008;68:4971-6.

30. Mulligan LM. RET revisited: expanding the oncogenic portfolio. Nat Rev Cancer 2014;14:173-86.

31. Takahashi M, Ritz J, Cooper GM. Activation of a novel human transforming gene, ret, by DNA rearrangement. Cell 1985;42:581-8.

32. Ju YS, Lee WC, Shin JY, et al. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res 2012;22:436-45.

33. Le Rolle AF, Klempner SJ, Garrett CR, et al. Identification and characterization of RET fusions in advanced colorectal cancer. Oncotarget 2015;6:28929-37.

34. Takeuchi K, Soda M, Togashi Y, et al. RET, ROS1 and ALK fusions in lung cancer. Nat Med 2012;18:378-81.

35. Takahashi K, Mitsui K, Yamanaka S. Role of ERas in promoting tumour-like properties in mouse embryonic stem cells. Nature 2003;423:541-5.

36. Kubota E, Kataoka H, Aoyama M, et al. Role of ES cell-expressed Ras (ERas) in tumorigenicity of gastric cancer. Am J Pathol 2010;177:955-63.

37. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805-11.

38. Moriarity A, O'Sullivan J, Kennedy J, Mehigan B, McCormick P. Current targeted therapies in the treatment of advanced colorectal cancer: a review. Ther Adv Med Oncol 2016;8:276-93.

39. Han SW, Kim HP, Shin JY, et al. Targeted sequencing of cancer-related genes in colorectal cancer using next-generation sequencing. PLoS One 2013;8:e64271.

40. Kopetz S, Desai J, Chan E, et al. Phase II Pilot Study of Vemurafenib in Patients With Metastatic BRAF-Mutated Colorectal Cancer. J Clin Oncol 2015;33:4032-8.

41. Prahallad A, Sun C, Huang S, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature 2012;483:100-3.

42. Sievert AJ, Lang SS, Boucher KL, et al. Paradoxical activation and RAF inhibitor resistance of BRAF protein kinase fusions characterizing pediatric astrocytomas. Proc Natl Acad Sci U S A 2013;110:5957-62.

43. Takeuchi K, Choi YL, Soda M, et al. Multiplex reverse transcription-PCR screening for EML4-ALK fusion transcripts. Clin Cancer Res 2008;14:6618-24.

44. Sekine S, Yamashita S, Tanabe T, et al. Frequent PTPRK-RSPO3 fusions and RNF43 mutations in colorectal traditional serrated adenoma. J Pathol 2016;239:133-8.

45. Voest EE, Bernards R. DNA-Guided Precision Medicine for Cancer: A Case of Irrational Exuberance? Cancer Discov 2016;6:130-2.

46. Siu LL, Lawler M, Haussler D, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med 2016;22:464-71.

5

SUPPLEMENTARY DATA

SUPPLEMENTARY MATERIALS AND METHODS

SAMPLE COLLECTION

Patients were selected from the MATCH-study, a prospective multicenter cohort study from 2007 onwards including adult patients undergoing curative surgery in one of seven hospitals in the Rotterdam region, the Netherlands (MEC-2007-088). All patients gave written informed consent for the storage and use of tissue samples for research purposes, and the collection of clinical data.

The selection criteria for this study were: 55-85 years old, stage I-III colon cancer, treatment was either curative surgery only (stage I-II) or curative surgery combined with adjuvant systemic therapy (stage III), and either disease free follow-up of at least 30 months or recurrence of disease at any point in time.

Directly following resection of the bowel segment containing the tumor, the specimen was transported to the pathology lab. Then, 2 to 5 biopsies of both central and peripheral regions of the tumor were sampled as well as 1 to 2 nontumor colon tissue sampled as distant from the tumor as possible. The samples were then fresh frozen using liquid nitrogen or isopentane with a maximum cold ischemia time of 2 hours. All samples were temporarily kept at -80˚C at the local centers, transported to the central tissue bank on dry ice and stored in liquid nitrogen.

HISTOLOGICAL EXAMINATION OF TUMOR SAMPLES

Sections of all samples were produced using a cryostat microtome (Thermo Scientific Microm HM 560, Thermo Fisher Scientific, inc.) set at -20˚C. For each sample, a Haematoxylin-eosin (HE) stained 5 µm section was reviewed by two pathologists for determination of final percentage of invasive tumor cells, necrosis, infiltrative cells, normal cells, tumor type (adenocarcinoma or other) and grade if possible. If needed, the paraffin slides of the primary tumor were used to aid in the assessment of the fresh frozen slides. Only samples with at least 40% invasive tumor cells were included in the final analysis, with the number of samples per sequencing run depending upon this percentage (see also below).

ISOLATION OF DNA AND RNA FROM TUMOR SAMPLES

DNA and RNA was isolated from 30 µm sections taken from the frozen tumor tissue obtained at primary surgery, preceded and followed by a 5 µm section for HE-staining and subsequent evaluation by the pathologist.

Genomic DNA was isolated using the NucleoSpin DNA Tissue kit (Macherey- Nagel; Bioké, Leiden, The Netherlands) according to the manufacturer's protocol. RNA isolation was done using RNA-Bee® (Tel-Test inc., Bio-Connect BV, Huissen, The

Netherlands), chloroform, isopropanol and ethanol according to the manufacturer's protocol and as described before.[1] Prior to sequencing, 20 µg of the isolated RNA was DNase treated and cleaned using the NucleoSpin® RNA II kit (Macherey-Nagel; Bioké, Leiden, The Netherlands) according to the manufacturer's protocol.
The quality and quantity of the DNA and RNA before and after clean-up was assessed with the Nanodrop ND-1000 (Thermo Scientific, Bleiswijk, The Netherlands) for measurement of A260/280 ratio, A260/230 ratio and total nucleic acid concentration. In addition, DNA quality and quantity was assessed using agarose gel electrophoresis and the PicoGreen® dsDNA quantitation assay (Thermo Scientific), respectively. The MultiNA Microchip Electrophoresis system (Shimadzu, Kyoto, Japan) was used for assessment of RNA quality, which was reported as an in-house adapted RNA integrity value ranging from 2 to 10. Only samples with an RNA integrity value of at least 7.0 were included in the final analysis.

RNA-SEQUENCING

500 ng of total RNA from tumor samples was used as input for the Illumina TruSeq stranded RNA-seq protocol. Libraries were pooled and sequenced on Illumina HiSeq2500 or NextSeq instruments. Pool sizes and the amount of samples per run were determined based on the percentage of tumor cells estimated from histological examination. **Supplementary Table S1** provides an overview of the number of sequence reads generated per sample.

BIOINFORMATIC DETECTION OF FUSION TRANSCRIPTS

We used the STAR fusion gene detection pipeline (version STAR-2.4.1) for analysis of RNA-seq data.[2] In a first step, fastq files were merged in case tumor samples (libraries) had been sequenced on separate sequencing runs. A total of 348 tumor datasets (70 published[3] and 278 generated in our work) and 74 control datasets (69 previously published and 5 generated in our work) were used as input. Subsequently, fastq files were used as input for STAR mapping and fusion detection with settings: -chimSegmentMin 15, chimJunctionOverhangMin 15, outSJfilterIntronMaxVsReadN 10,000,000. The resulting splice junction (SJ) and chimeric junction files were annotated with ensembl gene IDs (Ensembl version 72) and the annotated junctions were clustered per sample based on their overlapping gene IDs, i.e. junctions with the same overlapping donor and acceptor genes were merged. Subsequently, the clustered junctions were annotated based on the longest overlapping Ensembl CCDS transcript. The last exon of the donor gene and the first exon of the acceptor gene are reported. In addition, the type of junction was indicated as intronic (if junction reads fall in an intron), exonic (if junction reads fall in an intron) or boundary (if the

junction matches an intron-exon boundary). Furthermore, the exon rank within the transcript and the exon phase were assigned for both the donor and acceptor gene. Using this analysis and annotation pipeline we obtained 3 million raw junction calls. We applied a series of filtering steps to extract high-quality junctions from these raw data (**Figure 1**). This list of remaining junctions was annotated by adding fusion gene counts, donor gene counts, acceptor gene counts, overlap with protein domains and calculation of expression z-scores for the donor and acceptor genes relative to samples without the fusion. For calculation of z-scores, read counts per gene were calculated using HTSeq[4] using settings –m union -s yes. Read counts were normalized for library size (total read counts per sample). Normalized read counts were used as input for calculating z-scores according to:

$z - score = (X - \mu)/\sigma$

where X is the normalized read count for the sample with the fusion gene, μ is the mean read count for all samples without fusion and σ is the standard deviation of the normalized read counts across all samples without the fusion gene.

BIOINFORMATIC DETECTION OF SOMATIC VARIATIONS IN RNA-SEQ DATA

Somatic genetic variations in *BRAF*, *HRAS*, *NRAS*, *KRAS*, *PIK3CA*, *TP53*, *APC*, *SMAD4*, *PTEN* were detected in RNA-seq data using the GATK RNA-seq variant calling best practices. In brief, RNA-seq reads were mapped with STAR (version STAR-2.4.1) followed by a 2-pass STAR mapping using a new index of the genome reference based on the splice junctions identified by STAR (SJ.out.tab). The resulting BAM files were processed by Picard[5] to mark duplicates, sorting, indexing and adding read group information. Subsequently, we used GATK SplitNCigarReads to split reads in exon segments and hard-clip parts of reads that overlap intronic regions. Following indel realignment and base recalibration, variant calling and filtering was done using GATK Haplotypecaller, which resulted in a vcf file with indels and single-nucleotide variants. The vcf file was overlapped with COSMIC hotspot mutations in *BRAF*, *KRAS*, *NRAS* and *HRAS* genes. For variations in the remaining genes we checked for overlap with variations in 1000G (phase 3), TWINSUK, ALSPAC and ExAC. The non-overlapping variations were annotated based on variant effect and only non-synonymous, stop-gained/lost and frameshift variants that were flagged with PASS and SnpCluster were retained.

MATE-PAIR SEQUENCING AND ANALYSIS

Mate-pair library preparation was done using the Illumina Nextera Mate Pair library kit. Libraries were barcoded, pooled and sequenced on NextSeq using 2*75bp configuration. Mate-pair sequencing reads were mapped to the human reference genome (GRCh37) using BWA.[6] Subsequently, discordant read pairs were detected

from BAM files using a custom analysis pipeline as described previously.[7] To enrich our data for somatic rearrangements, clusters of discordant reads (rearrangement calls) were filtered against an in-house database of structural variation calls generated from mate-pair sequencing data using the same analysis pipeline. Previous work has shown that filtering against a large control dataset is a highly efficient approach to select for somatic variants.[8] BAM files containing mate-pair sequencing reads were also used as input for calling copy number changes with the FREEC tool.[9]

WHOLE EXOME SEQUENCING

DNA was was sequenced by GATC Biotech (Constance, Germany) using Illumina's HiSeq protocol (paired-end 100 bp, captured regions according to SureSelect v5), with a minimum of 60x coverage for normal DNA, and 90x or 120x for tumor DNA of tissues with >70% or 40-70% tumor cells, respectively. Raw sequence data for both normal and tumor DNA were mapped to reference genome GRCh37 using BWA (v 0.7.5a-r405)[6] after which data was sorted, indexed and duplicates marked using Picard (http://broadinstitute.github.io/picard/). Resulting BAM files were used for variant calling using two different software-packages: GATK (v3.3.0)[10] and Mutect (v1.1.6).[11] Mutations were filtered for high-quality calls and overlapping results from both variant callers. Next, the resulting mutations were annotated using SnpEff[12] to judge the effect of the mutation on the coding sequence (i.e. to check e.g. if the mutation results in an amino-acid change) and to check if the mutation is a known SNP in dbSNP (138_v37). Next, these data are combined per patient; only mutations found in the tumor but not in the normal and unknown in dbSNP were considered as somatic. Finally, the list was filtered for any variant with a minimal number of 20 reads per position and at least 15 reads having the variant.

FUSION GENE VALIDATION BY RT-PCR

We used RT-PCR as an assay to validate fusion genes. For this, 0.5µg of total RNA was taken as input for cDNA preparation (20 µL total reaction volume) using the High Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific). Junction PCR specific primers were designed based on the RNA-seq chimeric junction reads using primer3 software (**Supplementary Table 2**). PCR was performed in a final volume of 10 µL with 0.5 µL cDNA reaction using 50 nM junction specific PCR primers and 0.1 µL AmpliTaq Gold polymerase (Thermo Fisher Scientific) with standard PCR cycling conditions. cDNA samples from matching normal tissue were used as a control to be able to confirm that fusion genes were tumor-specific.

FUSION GENE CLONING

Primers specific for the entire fusion transcripts were designed using primer3 software. PCR products containing the entire fusion genes were cloned into the pBABE-puro retroviral mammalian expression vector. pBABE-puro was obtained through Addgene (Addgene plasmid # 1764).

FUSION GENE EXPRESSION AND WESTERN BLOTTING

HEK293T cells or NIH-3T3-A14 cells were transfected with 2μg of pBABE constructs containing fusion genes by the calcium phosphate method. Cells were co-transfected with constructs containing either myc-tagged ERK1 and GFPtagged AKT. After 48 hours cells were serum starved for 12 hours before cells were lysed with NP40 lysis buffer (150 mM NaCl, 50 mM Tris pH 8, 1 mM EDTA pH 8, 5 mM NaF, 1% NP-40, 0.25%, Na deoxycholate, 2mM NaVO3, protease inhibitors (cOmplete, Roche)). Lysates were clarified by centrifugation (12,000 × *g* for 10 minutes at 4˚C), and the concentration of total protein in the supernatant fraction was quantified by the Qubit protein assay kit (Thermo Fisher Scientific). Samples were denatured in NuPage® LDS Sample buffer 4X (Invitrogen, UK) at 70˚C for 10 minutes and 30 μg of protein was loaded and run on commercially produced pre-cast 4–12% Bis-Tris gels (Life Technologies). Proteins were transferred to a polyvinylidene fluoride (PVDF) membrane (Immobilon-FL, Merck-Millipore). Membranes were blocked with 2% BSA (Sigma) for 1 hour in TBS-T prior to incubation with rabbit polyclonal antibodies directed against phospho-AKT (Ser473) (D9E, Cell Signaling), phospho-p44/42 MAPK (ERK1/2, Thr202/Tyr204, Cell Signaling), c-MYC (910E, Santa Cruz Biotechnology) and mouse monoclonal anti-α-Tubulin (CP06, Calbiochem, San Diego, CA) overnight at 4˚C and diluted 1:5000 in TBS-T. Fluorescently labeled secondary antibodies (goat-anti-rabbit IgG (H+L) 800CW and donkey-antimouse (680 RD) from LI-COR) were applied for 90 minutes at room temperature (1:5,000 in TBS-T) prior to washing with TBS-T. Blots were imaged using a LICOR Odyssey® Infrared Imaging System and software and scanned at a resolution of 169 μm.

TISSUE CULTURE

Human HEK-293T cells were maintained in DMEM supplemented with 10% (v/v) fetal bovine serum, 100 U/ml penicillin, 100 μg/ml streptomycin and 2mM UltraGlutamine. For analysis of the AKT/PI3K pathway, NIH-3T3-A14 cells were used.[13] These cells overexpress the insulin receptor and were cultured in DMEM with low glucose (1000 mg/L). All cell lines were maintained at 37˚C in a humidified atmosphere consisting of 5% CO2/95% air.

SUPPLEMENTARY REFERENCES

1. Sieuwerts AM, Meijer-van Gelder ME, Timmermans M, et al. How ADAM-9 and ADAM-11 differentially from estrogen receptor predict response to tamoxifen treatment in patients with recurrent breast cancer: a retrospective study. Clin Cancer Res 2005;11:7311-21.

2. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.

3. Seshagiri S, Stawiski EW, Durinck S, et al. Recurrent R-spondin fusions in colon cancer. Nature 2012;488:660-4.

4. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-9.

5. Picard Tools - By Broad Institute. at http://broadinstitute.github.io/picard/.)

6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

7. Kloosterman WP, Guryev V, van Roosmalen M, et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet 2011;20:1916-24.

8. Hiltemann S, Jenster G, Trapman J, van der Spek P, Stubbs A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. Genome Res 2015;25:1382-90.

9. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28:423-5.

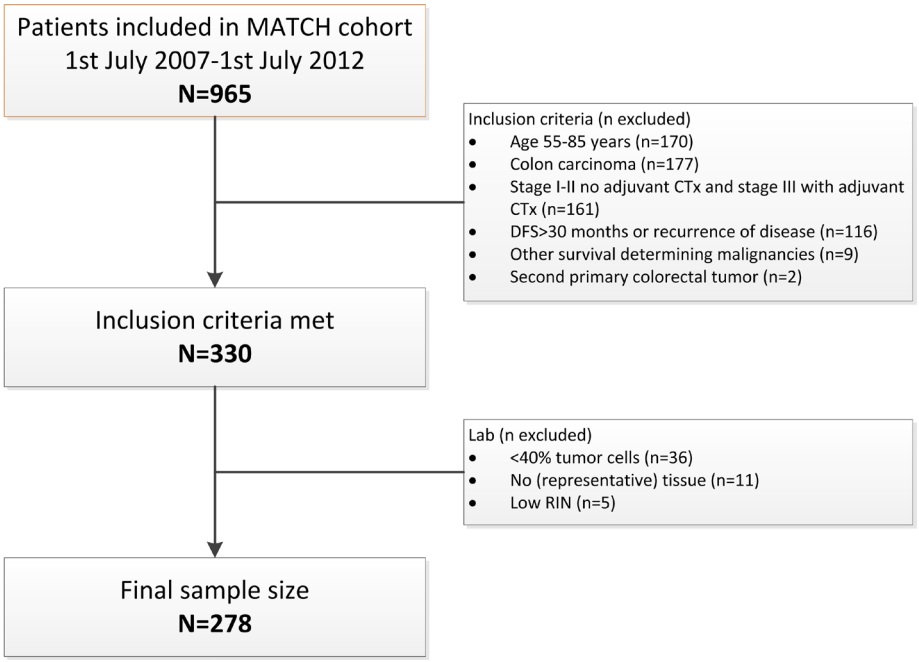
10. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.

11. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.

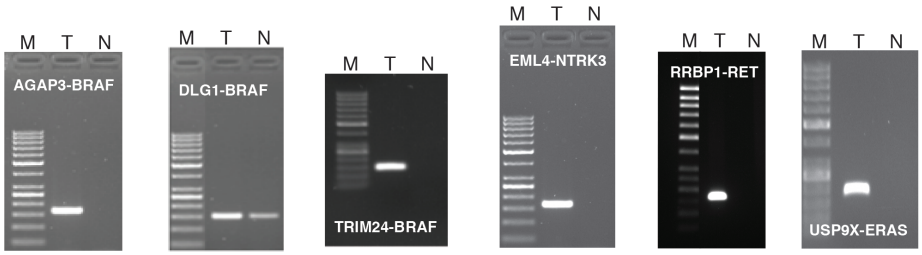
12. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80-92.

13. Burgering BM, Pronk GJ, van Weeren PC, Chardin P, Bos JL. cAMP antagonizes p21ras-directed activation of extracellular signal-regulated kinase 2 and phosphorylation of mSos nucleotide exchange factor. EMBO J 1993;12:4211-20.

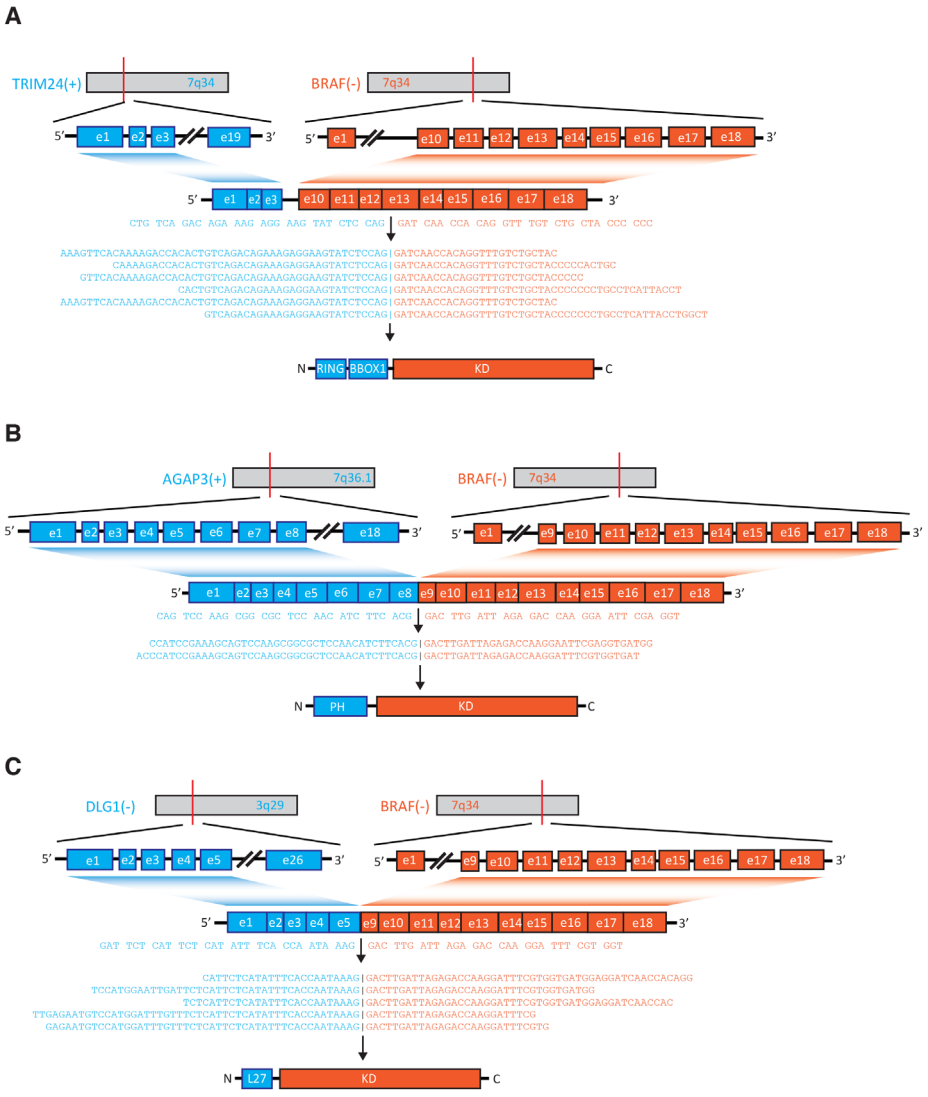
SUPPLEMENTARY FIGURES



SUPPLEMENTARY FIGURE 1.
Flow diagram showing patient inclusion in this study. Patient samples were gathered from July 2007 to July 2012 as part of the Rotterdam MATCH study.

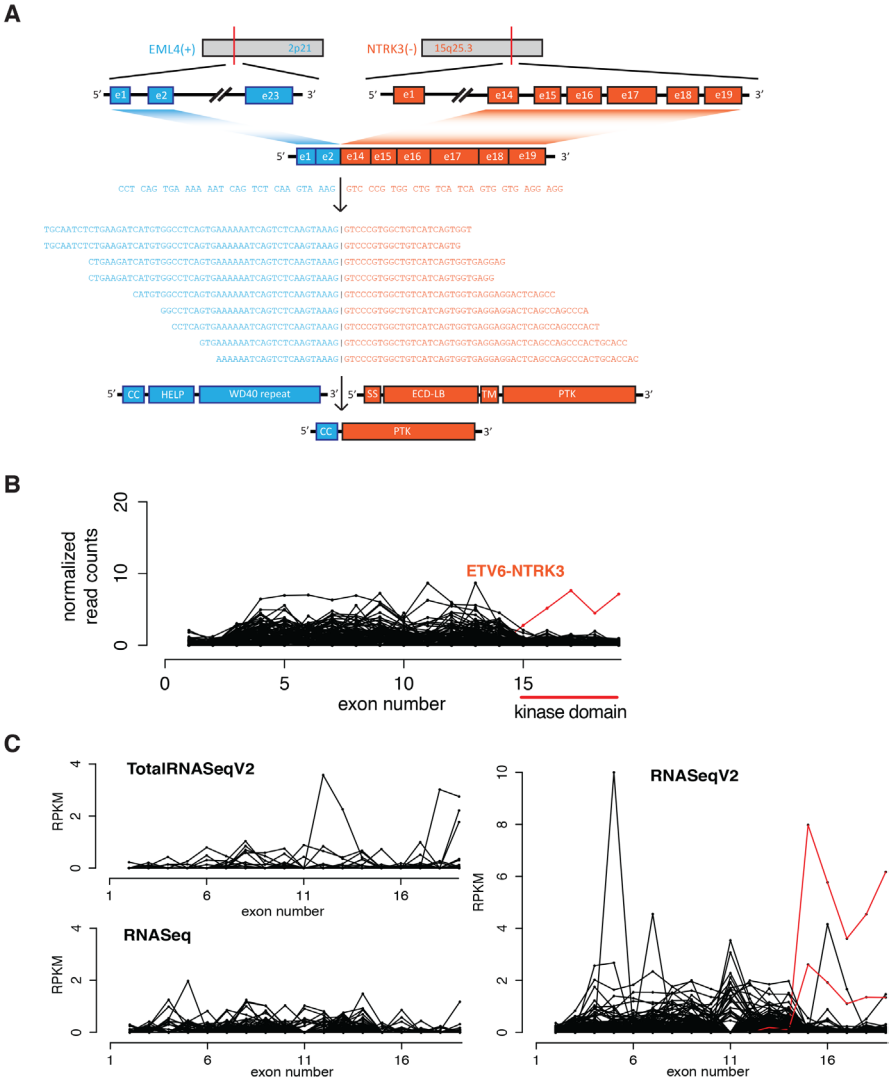


SUPPLEMENTARY FIGURE 2.
RT-PCR results of fusion gene verification. Primers for verification of cancer fusion genes were designed based on chimeric RNA-seq reads. M = marker, T = primary colon tumor with indicated fusion, N = matching normal colon tissue.



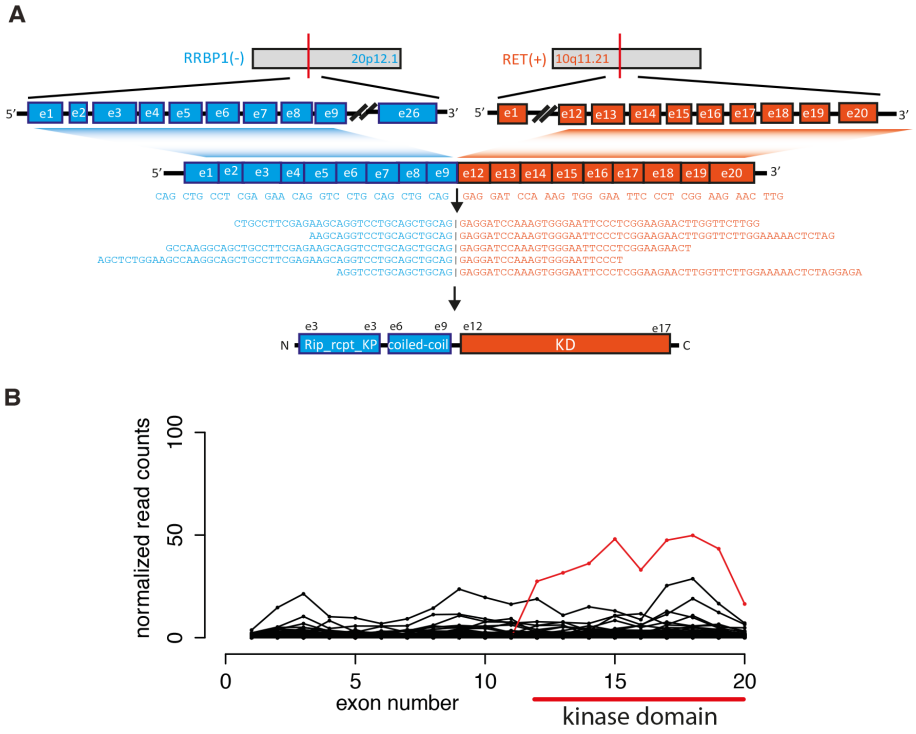
SUPPLEMENTARY FIGURE 3.

Exonic and protein structure of three BRAF fusions. Exonic structure, selected split RNA-seq reads and protein structure of *TRIM24-BRAF* (**A**), *AGAP3-BRAF* (**B**), *DLG1-BRAF* (**C**). KD = kinase domain, L27 = L27 protein interaction module, PH = pleckstrin homology, RING = zinc finger domain, ring type , BBOX1 = B-box-type zinc finger domain.



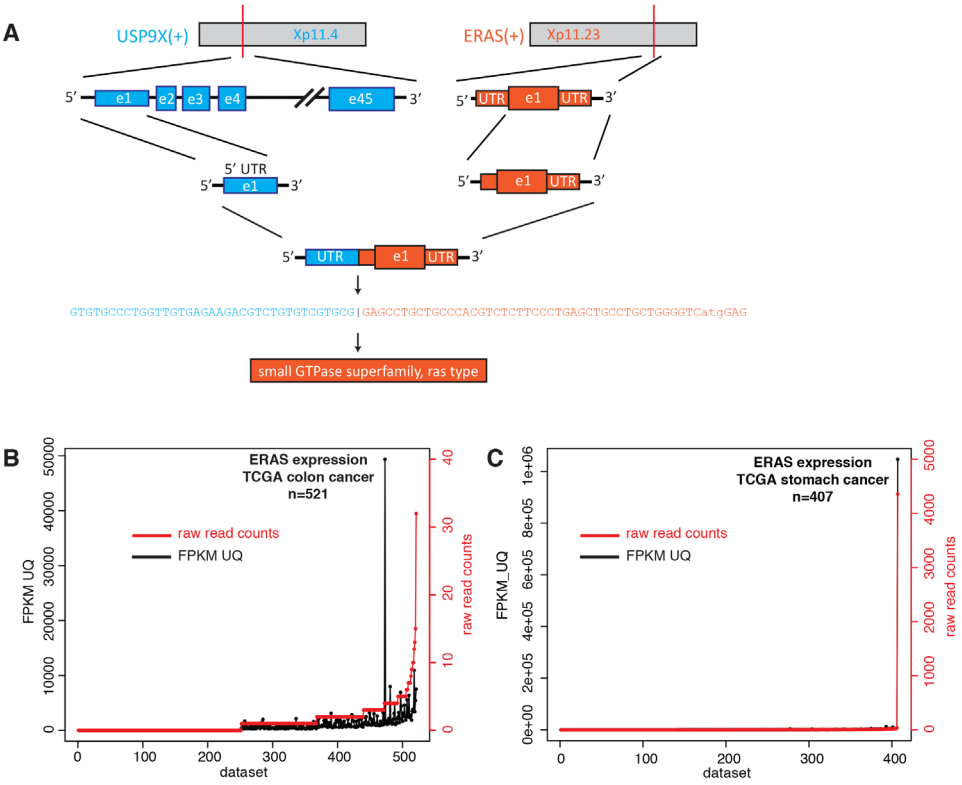
SUPPLEMENTARY FIGURE 4.

Structure of an *EML4-NTRK3* fusion gene and exonic expression of *NTRK3*. (**A**) Exonic structure, split-reads and protein structure of *EML4-NTRK3* fusion gene. (**B**) Exonic expression of NTRK3 in 70 colorectal cancers published previously.[1] The data represented by a red line are derived from a colorectal tumor expressing an *ETV6-NTRK3* fusion gene. (**C**) Exonic expression of *NTRK3* in 732 colorectal cancer RNA-seq datasets from TCGA data (https://tcga-data.nci.nih.gov/tcga/). Red lines indicate samples with increased expression of the entire *NTRK3* kinase domain, possibly indicating the presence of an *NTRK3* fusion gene. Both RNA-seq datasets were obtained from colon adenocarcinoma tissue. Data file identifiers and exonic expression values are shown in Supplementary Table 3.



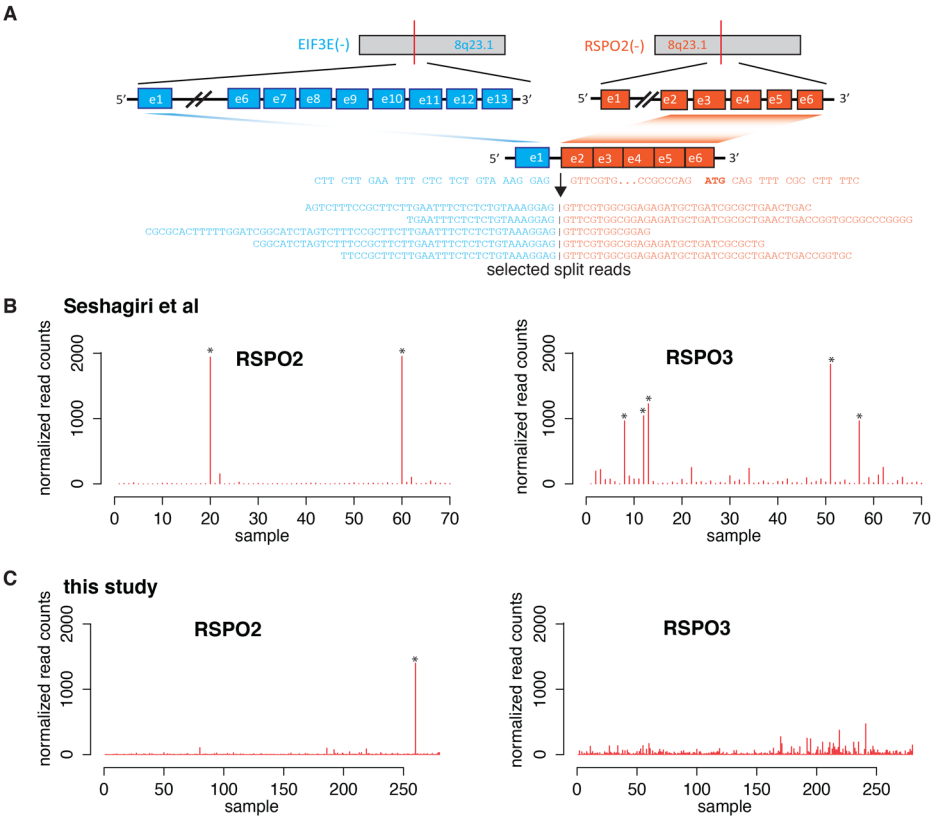
SUPPLEMENTARY FIGURE 5.

Structure and expression of a novel *RRBP1-RET* fusion. (**A**) Exonic structure, as subset of split-reads and protein structure of the *RRBP1-RET* fusion gene. The entire kinase domain of *RET* is fused to an N-terminal Rip-rcpt-KP and coiled-coil domain of *RRBP1*. (**B**) Expression of exons in the *RET* gene in 278 primary colon cancers from this study. The red curve indicates the sample with the *RRBP1-RET* fusion gene.



SUPPLEMENTARY FIGURE 6.

Structure and expression of a novel *USP9X-ERAS* fusion gene. (**A**) Exonic structure and protein domains of the *USP9X-ERAS* fusion gene. (**B**) Expression of the *ERAS* gene across 521 TCGA RNA-seq data sets from colon cancer samples. (**C**) Expression of the *ERAS* gene across 407 TCGA RNA-seq datasets from stomach cancer samples. For (B) and (C), both raw read counts and upper quartile normalized FPKM values (FPKM UQ) are plotted. Datasets (x-axis) were ranked from low to high raw read count. Data were obtained from the Genomic Data Commons Data Portal (NIH). Plotted values and dataset identifiers are provided in Supplementary Table 3.



SUPPLEMENTARY FIGURE 7.

Rspondin fusion gene detection and expression. (**A**) Structure of an *EIF3E-RSPO2* fusion gene detected in a colon cancer sample in the cohort described in this paper. Note that the fusion does not lead to a fusion protein, since the *RSPO2* start codon is present after the fusion point in exon 2. (**B**) Expression of *RSPO2* and *RSPO3* genes in colorectal cancer datasets from Seshagiri *et al*.[3] (**C**) Expression of *RSPO2* and *RSPO3* genes in 278 colon cancer datasets from this study.

SUPPLEMENTARY TABLES

SUPPLEMENTARY TABLE S1. COLON TUMOR SAMPLE OVERVIEW, CLINICAL CHARACTERISTICS, FUSION GENES AND *KRAS*, *NRAS*, *BRAF* MUTATION STATUS AND SEQUENCING STATISTICS.

This supplementary table can be found online.

SUPPLEMENTARY TABLE S2. ANNOTATED LIST OF PREDICTED GENE FUSIONS THAT PASSED ALL FILTERING STEPS.

This supplementary table can be found online.

SUPPLEMENTARY TABLE S3. LIST OF RNA-SEQ DATASET IDENTIFIERS FROM TCGA USED FOR ANALYZING EXPRESSION OF *ERAS* AND *NTRK3* IN COLON AND STOMACH CANCER SAMPLES.

This supplementary table can be found online.

R.R.J. Coebergh van den Braak, A.M. Sieuwerts, R. Kandimalla, Z.S. Lalmahomed, S.I. Bril, A. van Galen, M. Smid, K. Biermann, J.H. van Krieken, W.P. Kloosterman, J.A. Foekens, A. Goel, J.W.M. Martens, J.N.M. IJzermans, on behalf of the MATCH study group.

6

HIGH MRNA EXPRESSION OF SPLICE VARIANT SYK SHORT CORRELATES WITH HEPATIC DISEASE PROGRESSION IN CHEMONAIVE LYMPH NODE NEGATIVE COLON CANCER PATIENTS

PLOS ONE 2017

ABSTRACT

OBJECTIVE

Overall and splice specific expression of Spleen Tyrosine Kinase (*SYK*) has been posed as a marker predicting both poor and favorable outcome in various epithelial malignancies. However, its role in colorectal cancer is largely unknown. The aim of this study was to explore the prognostic role of *SYK* in three cohorts of colon cancer patients.

METHODS

Total messenger RNA (mRNA) expression of *SYK, SYK(T),* and mRNA expression of its two splice variants *SYK* short *(S)* and *SYK* long *(L)* were measured using quantitative reverse transcriptase (RT-qPCR) in 240 primary colon cancer patients (n=160 patients with chemonaive lymph node negative [LNN] and n=80 patients with adjuvant treated lymph node positive [LNP] colon cancer) and related to microsatellite instability (MSI), known colorectal cancer mutations, and disease-free (DFS), hepatic metastasis-free (HFS) and overall survival (OS). Two independent cohorts of patients with respectively 48 and 118 chemonaive LNN colon cancer were used for validation.

RESULTS

Expression of *SYK* and its splice variants was significantly lower in tumors with MSI, and in *KRAS* wild type, *BRAF* mutant and *PTEN* mutant tumors. In a multivariate Cox regression analysis, as a continuous variable, increasing *SYK(S)* mRNA expression was associated with worse HFS (Hazard Ratio[HR]=1.83; 95% Confidence Interval[CI]=1.08-3.12; p=0.026) in the LNN group, indicating a prognostic role for *SYK(S)* mRNA in patients with chemonaive LNN colon cancer. However, only a non-significant trend between *SYK(S)* and HFS in one of the two validation cohorts was observed (HR=4.68; 95%CI=0.75-29.15; p=0.098).

CONCLUSION

In our cohort, we discovered *SYK(S)* as a significant prognostic marker for HFS for patients with untreated LNN colon cancer. This association could however not be confirmed in two independent smaller cohorts, suggesting that further extensive validation is needed to confirm the prognostic value of *SYK(S)* expression in chemonaive LNN colon cancer.

INTRODUCTION

Colon cancer is the second most common malignancy in the Western World with close to 450,000 new cases in Europe in 2012.[1] As in most solid cancers, histological tumor staging (TNM) is the best determinant of prognosis and as a result provides recommendations for treatment decisions. The current treatment for stage I-III colon cancer is surgery alone for stages I and II, and surgery combined with adjuvant chemotherapy for stage III. However, up to 21% of the patients with stage I-II and up to 40% of the patients with stage III colon cancer will develop metastatic disease after curative surgery.[2,3] Therefore, prognostic biomarkers complementing the TNM classification are urgently needed.[4,5]

Tyrosine-protein kinases are key regulators of cell proliferation associated with poor survival and tumorigenesis, and are therefore extensively studied in the field of oncological biomarker research.[6,7] Spleen tyrosine kinase (SYK) has been posed as marker predicting both poor and favorable outcome in various epithelial malignancies including colorectal cancer.[8-11] However, most of these studies have focused on functional outcome in cell lines or associated tumor characteristics to the total mRNA or protein expression of SYK instead of linking mRNA and/or protein expression of SYK to long term clinical outcome. Furthermore, evidence suggesting different biological effects for the two known splice variants of SYK on growth properties of cancer cells is accumulating. In aggregate, the long isoform SYK(L) appears to be associated with tumor suppressive activities while the short isoform SYK(S) appears to be associated with tumor promoting activities. For instance, in patients with hepatocellular cancer, the expression of SYK(S) has been reported to be a significant indicator of poor prognosis.[12]

The significance of SYK and its isoforms in colorectal cancer is largely unknown. Yang *et al*. showed that hypermethylation of the *SYK* promoter region resulted in loss of overall *SYK* mRNA expression, which was associated with a higher tumor stage and reduced five-year overall survival in a heterogeneous group of stage I-IV colon and rectum carcinoma.[13] In a second study by Ni *et al*. SYK(L) but not SYK(S) was downregulated in the majority of cancer and adjacent non-cancerous colon tissues.[14] Lastly, *SYK* is part of various prognostic gene signatures and the gene set used to define the consensus molecular subtypes of colorectal cancer.[15,16]

We aimed to assess the association of mRNA expression of overall *SYK* (*SYK(T)*) and its splice variants *SYK(L)* and *SYK(S)* with disease outcome in a well-defined homogeneous prospectively collected set of primary tumor tissues of patients with stage I-III colon cancer. Patients with lymph node negative (LNN) colon cancer who did not receive systemic adjuvant chemotherapy (chemonaive) and patients with

lymph node positive (LNP) colon cancer who did receive adjuvant chemotherapy were analyzed separately to distinguish between pure disease prognosis and prognosis after adjuvant chemotherapy.

MATERIAL AND METHODS

Where possible, the guidelines for Reporting recommendations for tumour MARKer (REMARK) prognostic studies were followed, and the paper was written accordingly.[17]

PATIENT SELECTION

Patients were selected from the MATCH study, an ongoing prospective multicenter observational cohort study from 2007 onwards including adult patients who undergo curative surgery in one of seven participating hospitals in the Rotterdam region, the Netherlands. Patients received treatment according to the current national guideline.[18] Patients were verbally informed about the storage and use of tissue samples, and the collection of clinical data for research purposes. The institutional review board of the Erasmus MC University Medical Center approved the MATCH study and specifically approved studies on (epi)genetic biomarkers to predict recurrence of diseases including the current study (Institutional Review Board number MEC 2007-088). Written informed consent was obtained from all patients.

Inclusion criteria for this study were: informed consent available, inclusion date between 1st July 2007 and 1st July 2012 to ensure sufficient follow up, age > 55 years, stage I-II without adjuvant chemotherapy or stage III with adjuvant chemotherapy, radical surgery, fresh frozen tissue with at least 40% invasive tumor cells available, and either recurrence of disease or at least 30 months of disease-free follow-up. A diagram of the analysis workflow is shown in **Figure 1**.

The two independent validation cohorts consisted of 84 and 196 fresh frozen samples of primary colorectal cancers obtained through the Baylor Scott and White Research Institute and Charles A Sammons Cancer Center (Dallas, TX, USA) (cohort A and cohort B). Details on samples collection, processing and RNA isolation have been described previously.[19] For 82 and 185 patients of these cohorts respectively, RNA was sent to our lab to perform the cDNA synthesis and mRNA transcript level quantifications using the methodology as was used for the discovery study (see below). In cohort A, 34 patients were excluded (failed RNA/cDNA quality control [n=10], rectal carcinoma [n=23] and irradical resection [n=1]) leaving a total of 48 patients for analysis

(**Supplementary Figure 1**). In cohort B, 80 patients were excluded (failed RNA/cDNA quality control [n=7], rectal carcinoma [n=51] and age < 50 years [n=9] and incomplete survival data [n=2]) leaving a cohort of 116 patients for analysis (**Supplementary Figure 2**).

SAMPLE COLLECTION AND PROCESSING

Immediately following removal of the resection specimen during surgery, the specimen was transported to the pathology lab at room temperature and without any conservation fluids. In the pathology lab, two to four biopsies of both central and peripheral regions of the tumor as well as one or two adjacent non-tumor colon tissue samples were taken and fresh frozen with a maximum cold ischemia time of two hours. All samples were stored in liquid nitrogen.

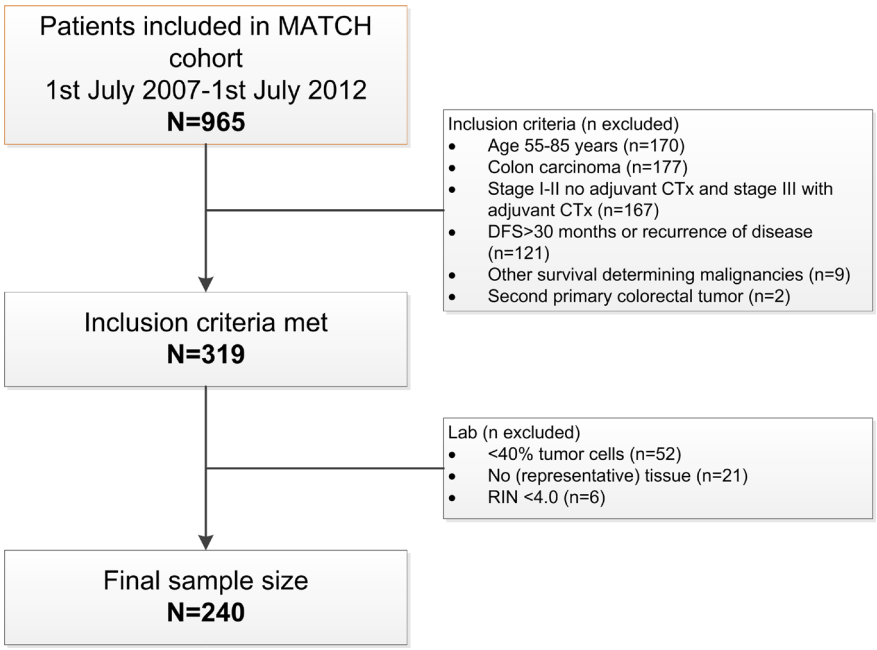


FIGURE 1. DIAGRAM OF ANALYSIS WORKFLOW OF THE MATCH COHORT

RNA ISOLATION, CDNA SYNTHESIS AND MRNA TRANSCRIPT LEVEL QUANTIFICATION

Sectioning of fresh frozen colon cancer and normal colon tissue was done using a cryostat microtome (Thermo Scientific Microm HM 560, Thermo Fisher Scientific, inc.) set at -20°C. Before, during and after sectioning for RNA isolation, 3 x 5 µm sections were cut and after hematoxylin-eosin (HE) staining reviewed by two pathologists independently. For the MATCH cohort, the percentage of tumor cells, necrosis, infiltrate and normal cells were estimated relative to other cells (e.g. stromal cells, inflammatory infiltrate and pre-existing epithelial cells). The estimates were scored in categories of 0–5%, 6–10%, 11–20%, 21–30%, 31–40%, 41–50%, 51–60%, 61–70%, 71–80%, 81–90%, and 91–100% tumor cells. Differentiation grade of the tumor was estimated according to the WHO 2010 classification for the carcinoma of the colon and rectum (WHO Press, World Health Organization, 20 Avenue Appia, Geneva, Switzerland). For the validation cohorts, no HE slides were available for evaluation.

For the discovery cohort, RNA was isolated from 30 µm sections using RNA-Bee® according to the manufacturer's instructions (Tel-Test inc., USA). For the validation cohorts, RNA was isolated with the RNeasy tissue kit (Qiagen, Germany). Quality and quantity of RNA was assessed with the Nanodrop ND-1000 (Thermo Scientific, Wilmington, USA) and the MultiNA Microchip Electrophoresis system (Shimadzu, Kyoto, Japan). Next, cDNA was generated from 2 µg of the isolated total RNA for the discovery cohort and from 0.1-1 µg of the isolated total RNA for the validation cohort using Reverse Transcriptase (RT) with the Thermo Scientific RevertAid H Minus First Strand cDNA Synthesis Kit (Fermentas, Thermo Scientific, USA) using the protocol supplied by the manufacturer, followed by an RNAse H step (Ambion, Life Technologies, USA) to digest any remaining RNA. Quantitative real-time PCR (qPCR) was performed with the Mx3000P QPCR machine (Agilent Technologies, NL) using ABgene Absolute Universal or Absolute SYBR Green with ROX PCR reaction mixtures (Thermo Scientific, USA) according the manufacturer's instructions.[20]

SYK mRNA expression levels were quantified with commercially available and validated TaqMan assays (Applied Biosystems, Thermo Scientific, USA) for the total expression of *SYK* (*SYK(T)); Hs00374292_m1*, and for its two alternative splicing variants, full-length *SYK* (*SYK(L)); Hs00895384_m1* and the short gene product lacking a 23-amino acid insert within the "linker" region located between the second Src homology 2 and the catalytic domain (*SYK(S)); Hs00177369_m1*. *SYK* mRNA expression levels were normalized using the average of three reference genes (*HMBS*, *HPRT1* and *TBP*) using the 2-ΔΔCq method as described in detail before by Livak and Schmittgen [21] and Sieuwerts *et al* [22], using a serially diluted pooled tumor cDNA sample as calibrator in

every run to allow comparisons between runs. Only cDNA samples that were at a 100-fold final dilution in the qPCR able to generate a Cq value for the average of the reference genes within 28 cycles were considered of sufficient quality and quantity to be included in the study. Specifics of the gene assays used are provided in **Supplementary Table 1**.

MESENCHYMAL AND INFILTRATE MARKERS

To capture epithelial to mesenchymal transition (EMT), the mRNA expression levels of one epithelial marker (*EPCAM*) and the three mesenchymal markers from the Oncotype Dx (*BGN*, *FAP*, *INHBA*) were measured using RT-qPCR.[23]

PTPRC mRNA levels (a measure for CD45), which is present on all differentiated hematopoietic cells except erythrocytes and plasma cells, were used to estimate the contribution of infiltrate. Specifics of the gene assays to generate these indices are provided in **Supplementary Table 1**.

MUTATION CALLING

For n=238 patients, RNA sequencing data was available.[24] In short, somatic genetic variations were detected in RNA-seq data using the GATK RNA-seq variant calling tool.[25] From the variant call list produced by the GATK workflow, we only retained calls that overlapped known cancer mutations present in the COSMIC database.[26]

MICROSATELLITE INSTABILITY (MSI)

MSI was analyzed with the MSI Analysis System from Promega, a fluorescent PCR-based assay for the detection of MSI in 5 mononucleotide repeat markers (BAT-25, BAT-26, NR-21, NR-24 and MONO-27) and two pentanucleotide repeat markers (Penta C and Penta D). The mononucleotide markers were used for MSI determination, and the pentanucleotide markers to detect potential sample mix ups and/or contamination using the protocol supplied by the manufacturer. In brief, genomic DNA was extracted with the NucleoSpin Tissue kit (Macherey-Nagel, BIOKE, Leiden, NL) from 2 to 5 x 30 µm sections cut in between the sections used for the RNA isolation. Quality and quantity were assessed by both Nanodrop, the Quant-iT PicoGreen dsDNA kit (Life Technnologies) and agarose gel electrophoresis. Next, 2 ng of PicoGreen measured DNA was used in the analysis for MSI.

The technical personnel performed all the above-mentioned analyses blinded from clinical outcome since they received the samples with according sample numbers and had no access to the patient identifying data nor the clinical data.

SURVIVAL DATA

Disease free survival (DFS) was defined as the time elapsed between the date of surgery, and either the date of any recurrence of disease or the date of the last follow-up visit at which a patient was considered to have no recurrence. Hepatic metastasis free survival (HFS) was defined as the time elapsed between the date of surgery, and either the date of the appearance of liver metastasis or the date of the last follow-up visit at which a patient was considered to have no liver metastases. Overall survival (OS) was defined as the time elapsed between the date of surgery, and either the date of death or the date of the last check in the Municipal Personal Records Database.

STATISTICAL ANALYSES

Statistical analyses were performed using the SPSS statistical package version 21. mRNA expression levels of *SYK(T)*, *SYK(S)* and *SYK(L)* were correlated with each other, the epithelial, mesenchymal and infiltrate markers, the clinicopathological characteristics and assessed CRC mutations using the Spearman Rank correlation test, Mann-Whitney U test, Kruskal-Wallis test and Jonckheere-Terpstra test where appropriate. Univariate Cox regression analysis was used to assess the association of the mRNA expression levels of *SYK(T)*, *SYK(S)* and *SYK(L)* as a continuous variable and clinicopathological characteristics with the clinical endpoints. Kaplan Meier estimates were used to visualize the association between mRNA expression of SYK and its splice variants with the relevant clinical endpoints. To this end, mRNA expression levels were split at the median level. Multivariate Cox regression analysis was used to assess the association between mRNA expression and clinical outcome while correcting for other clinicopathological factors associated with the clinical endpoint of interest. All analyses were two-sided and $P<0.05$ was considered significant.

RESULTS

CORRELATION OF MRNA EXPRESSION LEVELS *SYK(T)* AND ITS SPLICE VARIANTS

First, we assessed the correlation between *SYK(T), SYK(S)* and *SYK(L)*. *SYK(T)* showed a good correlation with both *SYK(S)* and *SYK(L)* (Spearman's Rho (r_s)=0.74, p<0.001 and r_s=0.86 p<0.001, respectively) while *SYK(S)* and *SYK(L)* expression levels showed only a moderate association (r_s=0.48 p<0.001) (**Supplementary Figure 3**). The worse correlation between the two splice variants suggested that a separate analysis of the splice variants may be of added value.

TABLE 1. CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS OF THE TOTAL MATCH COHORT

			SYK(T)		*SYK(S)*		*SYK(L)*		
			MEDIAN (IQR)	P VALUE	MEDIAN (IQR)	P VALUE	MEDIAN (IQR)	P VALUE	PERFORMED TEST
	n	%							
Gender									
Female	112	46.7%	-4.24 (-4.60 • -3.72)	0.11	-4.81 (-5.47 • -4.14)	0.18	-4.76 (-5.16 • -4.17)	0.40	Mann-Whitney U
Male	128	53.3%	-4.05 (-4.58 • -3.50)		-4.63 (-5.27 • -3.93)		-4.66 (-5.19 • -4.03)		
Age									
	240	100%	-0.08	0.21	-0.009	0.89	-0.011	0.86	Spearman's Rho
Tumor stage									
Stage I	60	25.0%	-4.31 (-4.71 • -3.67)	0.31	-4.62 (-5.21 • -3.96)	0.037	-4.73 (-5.26 • -4.28)	0.45	Jonckheere-Terpstra
Stage II	100	41.7%	-3.96 (-4.55 • -3.47)		-4.61 (-5.36 • -4.07)		-4.54 (-5.04 • -3.95)		
Stage III	80	33.3%	-4.15 (-4.58 • -3.68)		-4.98 (-5.70 • -4.19)		-4.81 (-5.30 • -4.33)		
T status									
T2	71	29.6%	-4.32 (-4.69 • -3.76)	0.03	-4.69 (-5.26 • -3.97)	0.37	-4.77 (-5.33 • -4.46)	0.021	Mann-Whitney U
T3	169	70.4%	-4.05 (-4.57 • -3.56)		-4.70 (-5.43 • -4.08)		-4.65 (-5.11 • -4.04)		
Nodal status									
N0 ≥ 10 nodes assessed	131	54.6%	-4.08 (-4.60 • -3.62)	0.97	-4.63 (-5.32 • -4.08)	0.07	-4.62 (-5.13 • -4.04)	0.037	Jonckheere-Terpstra
N0 < 10 nodes assessed	29	12.1%	-3.81 (-4.66 • -3.43)		-4.30 (-5.28 • -3.74)		4.68 (-5.17 • -3.90)		
N1	53	22.1%	-4.09 (-4.48 • -3.64)		-4.89 (-5.68 • -4.21)		-4.77 (-5.34 • -4.38)		
N2	27	11.3%	-4.30 (-4.60 • -3.69)		-5.08 (-5.74 • -4.03)		-4.96 (-5.21 • -4.31)		
Tumor grade									
Good	20	8.3%	-3.95 (-4.43 • -3.46)	0.57	-4.61 (-5.30 • -3.88)	0.28	-4.59 (-4.81 • -3.97)	0.61	Jonckheere-Terpstra
Moderate	192	80.0%	-4.14 (-4.60 • -3.61)		-4.69 (-5.33 • -4.08)		-4.72 (-5.20 • -4.15)		
Poor	20	8.3%	-4.10 (-4.60 • -3.74)		-4.95 (-5.87 • -4.35)		-4.78 (-5.24 • -4.08)		
Other	8	3.3%	-3.94 (-4.62 • -3.53)		-4.30 (-5.79 • -3.79)		-4.48 (-4.86 • -4.03)		
Location									
Right	121	50.4%	-4.24 (-4.71 • -3.70)	0.015	-4.89 (-5.58 • -4.15)	0.004	-4.84 (-5.28 • -4.28)	0.008	Mann-Whitney U
Left	119	49.6%	-4.01 (-4.47 • -3.53)		-4.52 (-5.21 • -3.85)		-4.62 (-4.96 • -4.13)		
MSI status[a]									
MSI	49	20.4%	-4.59 (-5.02 • -4.25)	<0.001	-5.34 (-5.77 • -4.86)	<0.001	-5.05 (-5.49 • -4.54)	<0.001	Mann-Whitney U
MSS	190	79.2%	-3.96 (-4.47 • -3.49)		-4.46 (-5.21 • -3.92)		-4.63 (-5.08 • -4.03)		

SYK mRNA expression levels were normalized using the average of three reference genes (*HMBS*, *HPRT1* and *TBP*) using the 2-ΔΔcq method as described in detail before by Livak and Schmittgen [21] and Sieuwerts et al [22]. [a] n=1 missing

ASSOCIATION OF *SYK* MRNA EXPRESSION LEVELS WITH CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS

In total, 240 patients were included in the discovery cohort. Clinical and histopathological characteristics, and median *SYK(T), SYK(S)* and *SYK(L)* mRNA expression levels and their associations for the entire group are shown in **Table 1**, for the 160 patients with lymph node negative (LNN) disease in **Supplementary Table 2a** and for the 80 patients with lymph node positive (LNP) disease in **Supplementary Table 2b**.

A significantly lower expression of *SYK(T)*, *SYK(S)* and *SYK(L)* was found in MSI tumors as compared to MicroSatellite Stable (MSS) tumors. This finding was observed in the total group as well as in both subgroups, except for *SYK(L)* in the LNP group. *SYK* expression was also significantly associated with tumor stage and location, but significance was dependent on the type of variant analyzed. While expression of *SYK(S)* was higher in stage I and II than in stage III, expression of *SYK(T)* and *SYK(L)* was not found to correlate in an unambiguous way with tumor stage. Independent of stage, a higher expression of *SYK(T)*, *SYK(S)* and *SYK(L)* was found in left sided tumors, which was also observed for *SYK(S)* in the LNN group and for *SYK(T)* and *SYK(L)* in the LNP group.

These data indicated a differential expression of *SYK* splice variants as compared to total *SYK* expression, with significant differences in mRNA expression of *SYK(T), SYK(S)* and/or *SYK(L)* with MSI status, stage and tumor location.

ASSOCIATION OF *SYK* MRNA EXPRESSION LEVELS AND MESENCHYMAL MARKERS

To explore the association between the *SYK* isoform variants and features of EMT in our cohort, mRNA expression levels of one epithelial marker (*EPCAM*) and the three mesenchymal markers from the Oncotype Dx[23] (*BGN, FAP, INHBA*) were measured using RT-qPCR (**Supplementary Figure 3**). mRNA expression levels of *SYK(T)*, *SYK(S)* or *SYK(L)* all showed a moderate positive correlation with mRNA expression of *EPCAM* (r_s=0.47 p<0.001, r_s=0.58 p=0.001 and r_s=0.41 p<0.001, respectively). For the stromal markers, only *FAP* showed a significant but less striking negative association with *SYK(S)* in the total group (r_s=-0.13 p=0.046) and LNP group (r_s=-0.24 p=0.031).

ASSOCIATION OF SYK MRNA EXPRESSION LEVELS AND INFILTRATE

As *SYK* is a known infiltrate marker[27], we next explored the association between mRNA and protein expression levels of *SYK* and its isoform variants, and the extent of possible infiltrate contribution. We measured mRNA expression levels of an infiltrate marker (*PTPRC*/CD45) using RT-qPCR and scored the percentage of infiltrate on H&E slides. Although mRNA expression levels of *SYK(S)* correlated moderately

negatively with the percentage of infiltrate as scored by a pathologist in the total group (r_s=-0.14 p=0.043), we did not observe a significant association between *PTPRC*/CD45 and mRNA expression of *SYK* or its splice variants (**Supplementary Table 4**).

ASSOCIATION OF *SYK* MRNA EXPRESSION LEVELS WITH KNOWN CRC MUTATIONS

Because of the correlation of *SYK* mRNA expression with MSI and a previous study which showed that *SYK* is differentially expressed in *KRAS*-dependent and *KRAS*-independent cancer cell lines[28], we explored the association between known CRC mutations and *SYK* expression in our MATCH cohort and TCGA (**Figure 2**). The mutation rates were: *APC* 90.4%, *TP53* 83.3%, *KRAS* 35.4%, *BRAF* 7.9%, *PTEN* 3.8%, *SMAD4* 3.3% and *NRAS* 1.7% (**Figure 2a**). mRNA expression of *SYK(T)* was significantly higher in *KRAS* mutant (mt), and lower in *BRAF* mt and *PTEN* mt tumors compared to wild type (wt) tumors (p=0.021, p=0.01 and p=0.031, respectively) (**Figure 2b**). A similar association was observed for *SYK(S)* (*BRAF* p<0.001 and *PTEN* p=0.002, respectively), while no significant associations were found for *SYK(L)* (**Figure 2c-d**). In line with literature[29], these mutations were more prevalent in MSI tumors than in MSS tumors (*BRAF* 30.6% vs 2.1%, p<0.001 and *PTEN* 10.2% vs 2.1% p=0.008). No significant differences in mRNA expression for *SYK(T)*, *SYK(S)* and *SYK(L)* were observed. Next, we analyzed all cases of stage I-III colon cancer in the TCGA for which both the known CRC mutations and *SYK(T)* expression levels were available (n=108) (**Figure 2e**). In this cohort, *SYK(T)* expression was significantly lower in *BRAF* mt tumors compared to wild type tumors (p=0.0018) and significantly lower in *APC* mt tumors compared to wild type tumors (p=0.009) (**Figure 2f**).

ASSOCIATION OF *SYK* MRNA EXPRESSION LEVELS WITH SURVIVAL

First, associations between basic patient characteristics and survival outcome were assessed using Cox regression analysis. In the total MATCH cohort, having a stage III tumor or more than three positive lymph nodes (N2 versus N0) was significantly associated with an adverse DFS. Age, gender and more than three positive lymph nodes were significantly associated with poor OS (**Supplementary Table 5a**). In the LNN subgroup, less than ten lymph nodes assessed in total was associated with an adverse HFS. In this sub group, only age at time of surgery was significantly associated with OS (**Supplementary Table 5b**). In the LNP group, more than three positive lymph nodes was significantly associated with an adverse DFS. Also, the presence of more than three positive lymph nodes and increasing age were significantly associated with poor OS in the LNP subgroup of the MATCH cohort (**Supplementary Table 5c**).

Subsequently, the associations between mRNA expression levels of *SYK* and its splice variants with DFS, HFS and OS were assessed using Cox regression analysis. For the whole MATCH cohort (n=240), no significant associations were found between mRNA expression of *SYK(T)*, *SYK(S)* and *SYK(L),* and the clinical endpoints (**Supplementary Table 5a**). Next, the LNN chemonaive group (n=160) and the LNP group who had received adjuvant therapy (n=80) were analyzed separately.

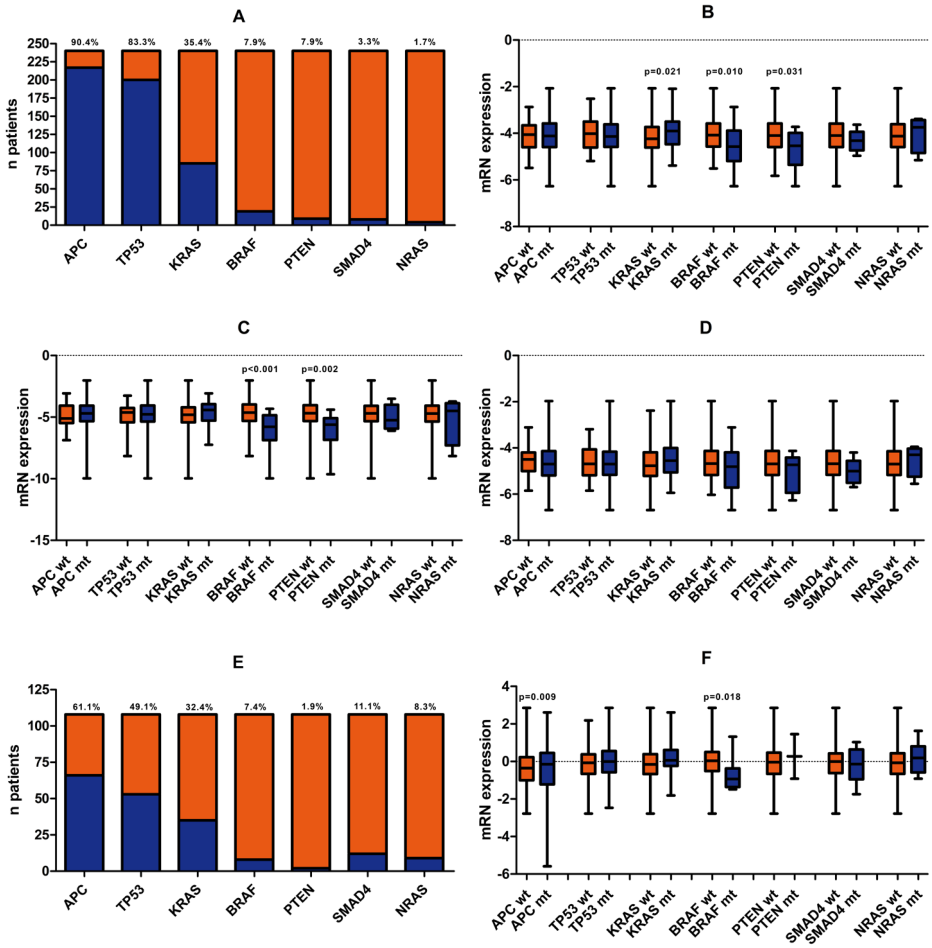


FIGURE 2. THE ASSOCIATION BETWEEN SYK MRNA EXPRESSION AND KNOWN CRC MUTATIONS
Mutation rates in the MATCH cohort (n=240)(**a**); differences in mRNA expression of *SYK(T)* (**b**), *SYK(S)* (**c**) and *SYK(L)* (**d**) in the MATCH cohort; mutation rates in the TCGA (n=108)(**e**); differences in mRNA expression of *SYK(T)* in the TCGA cohort (**f**).

In the LNN group, higher mRNA expression levels of *SYK(T)* and *SYK(S)* (continuous variables) were significantly associated with poor HFS (Hazard Ratio [HR]=2.05; 95% Confidence Interval [CI]=1.01-4.17; p=0.047 and HR=2.14; 95% CI=1.14-4.01; p=0.018, respectively) (**Supplementary Table 5b**). The association of mRNA expression of *SYK(T)* and *SYK(S)* split into four quartiles (Q1 with lowest mRNA expression levels through Q4 with the highest mRNA expression levels) with HFS was visualized by Kaplan-Meier curves (**Figures 3a** and **3b**), which suggested an impaired HFS particularly for patients with *SYK(S)* mRNA expression levels of the tumor in Q4. These findings were confirmed in an exploratory analysis with Cox regression analysis showing a significantly worse HFS for Q4 versus Q1-Q3 (HR=3.83; 95%CI=1.23-11.86; p=0.02). To explore the prognostic role of *SYK(S)* for HFS independent of other significantly associated factors in the LNN group, we performed a multivariate Cox regression model including N-status, the only other factor significantly related to HFS in the LNN group, and *SYK(S)* mRNA expression level. In this analysis, both continuous mRNA expression levels of *SYK(S)* and nodal status remained significantly associated with HFS (HR=1.83; 95% CI=1.08-3.12; p=0.026 and HR=1.27; 95%CI=1.01-1.60; p=0.042) (**Table 2**). However, since the total number of events in this low-risk group was only 12, these results should be interpreted with caution.
In the LNP group, no significant associations between any of the *SYK* mRNA expression levels and clinical endpoints were observed (**Supplementary Table 5c**).

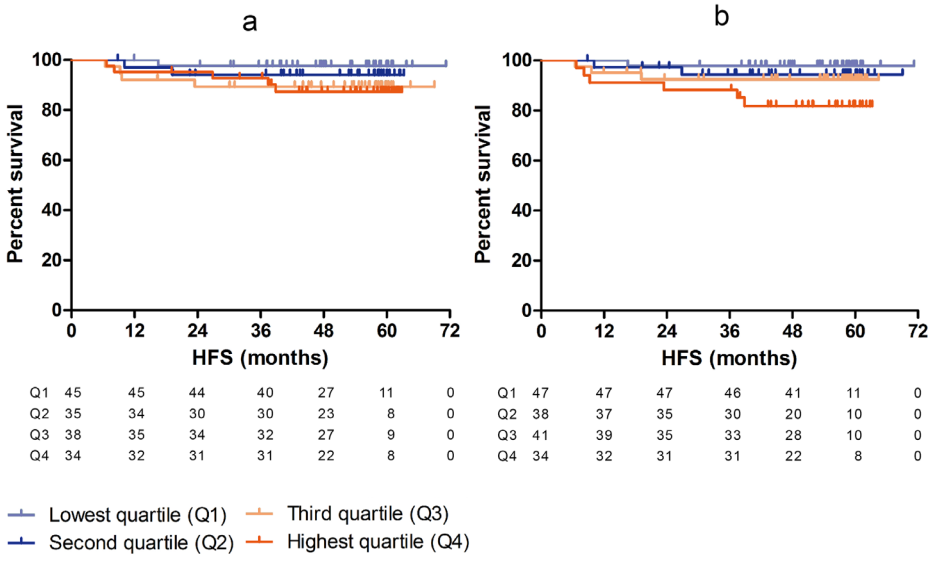


FIGURE 3. SURVIVAL CURVES FOR HFS IN THE LNN SUBGROUP OF THE MATCH COHORT FOR *SYK(T)* SPLIT IN QUARTILES (A) AND FOR *SYK(S)* SPLIT IN QUARTILES (B)

VALIDATION COHORTS

Details on patient and tumor characteristics for both cohorts can be found in **Table 3**. More patients in cohort A and B had a T1/T4 tumor compared to patients in the LNN subgroup of the MATCH cohort (14.6% and 8.6% vs 0%, p<0.001 respectively). The total number of assessed lymph nodes was less often below the cut-off of 10 lymph nodes (cohort A 4.2% and cohort B 12.9% vs 18.1% in the LNN subgroup of the MATCH cohort, p<0.001). Both cohort A and B contained more well differentiated tumors compared to the LNN MATCH cohort (83.3% and 93.1% vs 8.1%, p<0.001

TABLE 2. UNIVARIATE AND MULTIVARIATE COX REGRESSION ANALYSIS FOR THE LNN MATCH COHORT

			HFS (EVENTS=12)			
			HR (95%CI)	P VALUE	HR (95%CI)	P VALUE
	n	%				
mRNA expression						
SYK(S)	160	100%	2.14 (1.14 • 4.01)	0.018	1.83 (1.08 • 3.12)	0.026
Gender						
Female	78	48.8%	1			
Male	82	51.3%	1.97 (0.59 • 6.53)	0.27		
Age	160	100%	1.01 (0.94 • 1.08)	0.79		
Tumor stage						
Stage I	60	37.5%	1			
Stage II	100	62.5%	1.21 (0.36 • 4.02)	0.76		
Stage III	-					
T status						
T2	60	37.5%	1			
T3	100	62.5%	1.21 (0.36 • 4.02)	0.76		
Nodal status						
N0 ≥ 10 nodes assessed	131	81.9%	1			
N0 < 10 nodes assessed	29	18.1%	3.42 (1.09 • 10.78)	0.04	1.27 (1.01 • 1.60)	0.042
N1	-					
N2	-					
Tumor grade						
Good	13	8.1%	1			
Moderate	135	84.4%	0.88 (0.11 • 6.91)	0.27		
Poor	9	5.6%	2.85 (0.26 • 31.39)	0.39		
Other[a]	3	1.9%	-	-		
Location						
Right	82	51.3%	1			
Left	78	48.8%	1.09 (0.35 • 3.38)	0.88		
MSI status[b]						
MSI	37	23.1%	1			
MSS	122	76.1%	31.28 (0.12 • 8430.24)	0.23		

[a] there were no events in this subgroup; [b] n=1 missing

respectively). In cohort A, more tumors were right-sided compared to the LNN subgroup of the MATCH cohort (79.2% vs 51.3%, p>0.001). In cohort B, less tumors for which MSI status was known were MSI compared to the LNN subgroup of the MATCH cohort (9.5% vs 23.3%, p=0.019). No differences for the distribution of gender, age, tumor stage, or location of recurrence between the validation cohorts and the LNN subgroup of the MATCH cohort were observed (**Table 3**). No significant association between mRNA expression of *SYK(T)* or its splice variants with any of these characteristics were observed.

TABLE 3. BASIC CHARACTERISTICS OF THE LNN MATCH COHORT, AND VALIDATION COHORTS A AND B

	MATCH COHORT	COHORT A	COHORT B	P VALUE
Gender				
Female	78 (48.8%)	24 (50.0%)	51 (44.0%)	0.67
Male	82 (51.3%)	24 (50.0%)	65 (56.0%)	
Age	68 (62-75)		69 (61-78)	0.89
Tumor stage				
Stage I	60 (37.5%)	16 (33.3%)	37 (31.9%)	0.61
Stage II	100 (62.5%)	32 (66.7%)	79 (68.1%)	
T status				
T1	0 (0.0%)	3 (6.3%)	10 (8.6%)	<0.001
T2	60 (37.5%)	13 (27.1%)	27 (23.3%)	
T3	100 (62.5%)	28 (58.3%)	79 (68.1%)	
T4	0 (0.0%)	4 (8.3%)	0 (0.0%)	
Nodal status				
N0 ≥ 10 nodes assessed	131 (81.9%)	46 (95.8%)	101 (87.1%)	0.046
N0 < 10 nodes assessed	29 (18.1%)	2 (4.2%)	15 (12.9%)	
Tumor grade				
Good	13 (8.1%)	40 (83.3%)	108 (93.1%)	<0.001
Moderate	135 (84.4%)	5 (10.4%)	0 (0.0%)	
Poor	9 (5.6%)	2 (4.2%)	8 (6.9%)	
Other	3 (1.9%)	1 (2.1%)	0 (0.0%)	
Location				
Right	82 (51.3%)	38 (79.2%)	53 (45.7%)	<0.001
Left	78 (48.8%)	10 (20.8%)	63 (54.3%)	
MSI status[a]				
MSI	37 (23.1%)	-	6 (9.5%)	0.019
MSS	122 (76.3%)	-	57 (90.5%)	
Location of recurrence				
No recurrence	133 (83.1%)	41 (85.4%)	105 (90.5%)	0.65
Local	2 (1.3%)	1 (2.1%)	2 (17%)	
Hepatic	11 (6.9%)	2 (4.2%)	5 (4.3%)	
Non hepatic	11 (6.9%)	3 (6.3%)	4 (3.4%)	
Combined	3 (1.9%)	1 (2.1%)	0 (0.0%)	

[a] n=1 missing in the MATCH cohort and n=53 missing in cohort B

In both cohorts, no significant associations were observed between mRNA expression of *SYK(T)* nor the splice variants with DFS or HFS, although a non-significant trend between mRNA expression levels *SYK(S)* and HFS was observed in cohort A (HR=4.68; 95%CI=0.75-29.15; p=0.098).

DISCUSSION

In epithelial malignancies, both tumor promoting and tumor suppressing roles have been ascribed to *SYK*. Evidence suggesting different effects of the *SYK* splice variants on growth properties of cancer cells is accumulating.[10] The dual role of *SYK* in epithelial cancers combined with the scarce literature on the role of *SYK* and its splice variants in colorectal cancer provided a rationale to assess their prognostic value in primary tumors of colon cancer patients. This study showed that high mRNA expression level of *SYK(S)* is associated with short HFS in our MATCH cohort of chemonaive LNN colon cancer patients, although these findings could not be validated in two independent clinically less well-defined and smaller cohorts of patients with chemonaive LNN colon cancer.

Three major mechanisms through which *SYK* may affect cancer cell properties have been identified: *SYK* promoting cell survival through anti-apoptotic factors, *SYK* altering cellular differentiation programs regulating EMT and *SYK* altering cell motility. Importantly, SYK has two alternatively spliced variants, SYK(L) and SYK(S). In the short splice variant which a stretch of 23 amino acids in linker B (Exon 7) is spliced out. In normal hematopoietic cells, SYK(S) is intrinsically less active compared to SYK(L). In most epithelial cancers, overall *SYK* mRNA levels are higher in cancerous cells compared to normal cells of the same organ, including colon, suggesting a tumor promotor role of SYK in tumorigenesis.[10]

However, *SYK* mRNA or SYK protein expression have been both positively and negatively associated with tumor characteristics such as tumor grade and tumor stage. This paradoxical association may be explained by the accumulating observations that SYK(S) and SYK(L) both have an active but opposing role in solid cancers.[30,31] These opposing effects are generally attributed to a different location within the cell with SYK(L) being present in both the nucleus and cytoplasm, and SYK(S) being confined to the cytoplasm.[12,30,32] Wang and co-workers showed that SYK(L) was present in both normal and cancerous cells, and suppressed cell invasiveness in breast cancer cell lines. In contrast, SYK(S) was present only present in cancerous cells, but did not affect cell invasiveness.[30] Hong *et al*. observed similar differential expression patterns in hepatocellular carcinoma (HCC) as *SYK(L)* mRNA

expression was downregulated in 38% of the tumor samples while *SYK(S)* mRNA expression was detectable in 40% of the tumor samples and none of the normal liver tissue samples. Furthermore, *SYK(S)* mRNA expression levels were higher in poorly differentiated tumors compared to well differentiated tumors, while *SYK(L)* was expressed vice versa.[12] Ni *et al*. showed that overexpression of SYK(L) significantly reduced cell proliferation *in vitro* while SYK(S) overexpression did not in the human colorectal cancer HCT116 cell line. They also observed downregulation of SYK(L) but not SYK(S) 69% of tumor tissue samples compared to adjacent non-cancerous tissues.[14] In the current study we observed an decreased mRNA expression of *SYK(S)* in stage III compared to stage I-II colon cancers, but no association between mRNA expression of the splice variants with tumor grade. The latter may be explained by the large portion (88.3%) of well to moderately differentiated tumors in our cohort. Overall, the findings in literature and the current study suggest that SYK(S) is associated with tumor promoting activities while SYK(L) is associated with tumor suppressing activities.

We also observed differential expression between left and right-sided tumors, MSI and MSS tumors, and between tumors with and without known CRC mutations. Right-sided tumors, MSI tumors, *BRAF* mt tumors and *PTEN* mt tumors expressed *SYK(T)* and *SYK(S)* at a significantly lower level compared to left-sided tumors, MSS tumors, and wild type tumors in both the total and LNN subgroup, respectively. The lower expression of *SYK(T)* and *SYK(S)* in tumors harboring a *PTEN* mutation supports the findings of a previous study on diffuse large B-cell lymphomas in which a subset of samples exhibited an increase in the *SYK* gene copy number variation while a different subset exhibited loss of *PTEN* suggesting two independent mechanisms to promote cell survival.[33] The association between high mRNA expression of *SYK(T)* and both splice variants and microsatellite stability is interesting, as microsatellite stability is considered to be a phenotype associated with poor prognosis.[34] In aggregate, these findings may suggest a different role for *SYK* in hypermutated versus non-hypermutated tumors, although these findings should be verified in independent cohorts. We also observed a higher expression of *SYK(T)* in *KRAS* mt compared to *KRAS* wt tumors in our own cohort. These findings were in line with a previous study reporting higher expression *KRAS*-dependent compared to *KRAS*-independent pancreatic and lung cancer cell lines.[28] Thus, *SYK* may play a different role in *KRAS* mt and *KRAS* wt tumors. Functional studies should be conducted in colorectal cancer cell lines and/or samples to confirm this assumption.

Next to the associations with tumor characteristics, we showed that high *SYK(S)* mRNA expression is associated with short HFS in our MATCH cohort of chemonaive LNN colon cancer patients. To our knowledge, one previous study of colorectal

cancer patients explored the prognostic role of *SYK*. Yang *et al*. showed that methylation of the *SYK* gene promoter region was associated with decreased *SYK* mRNA and SYK protein expression, and subsequently showed a significantly worse five-year OS in the group with methylated *SYK* gene promoter region compared to the group with unmethylated *SYK* gene promoter region (5-year overall survival 59% vs 80% p<0.001, respectively.[13] However, the cohort consisted of stage I to IV colon and rectum carcinoma, and no details regarding neoadjuvant or adjuvant therapy and DFS were provided. Furthermore, only total expression of *SYK* was measured leaving questions regarding the prognostic value of the splice variants in their cohort unanswered. Interestingly, the prognostic role of the splice variants of *SYK* was investigated by Hong *et al*., who showed that patients with a *SYK(S)*-positive HCC were more likely to develop early and late recurrence (80.3% vs 53.8% P=0.001 and 66.7% vs 16.7%; P=0.002 respectively) compared to patients with a *SYK(S)*-negative HCC, which supports the findings in the MATCH cohort. Hong *et al*. also showed that patients with a *SYK(S)*-positive HCC had a worse OS compared to patients with a *SYK(S)*-negative HCC.[12] We did not observe an association between *SYK* mRNA expression and OS in our cohort. Furthermore, we did not find evidence supporting a tumor suppressor role for *SYK(L)*.

Unfortunately, the findings in the MATCH cohort could not be confirmed in two independent cohorts of patients with chemonaive LNN colon cancer and therefore warrant further investigation. The different observations in the MATCH cohort and the two validation cohorts with regard to clinical outcome may be explained by the limited number of patients and events (especially in cohort A with 48 patients and only 3 events for HFS). Second, the observed differences may be explained by differences in tumor biology. The large majority of tumors in both validation cohorts were well-differentiated compared to a large majority of moderately differentiated tumors in the MATCH cohort. Furthermore, Cohort A contained significantly more right-sided tumors while cohort B contained significantly less MSI tumors compared to the LNN subgroup of the MATCH cohort. Beside the biological differences associated with these tumor characteristics, we showed that expression of *SYK(T)* and its splice variants was significantly different for left- vs right-sided tumors and for MSS vs MSI tumors in the MATCH cohort. Lastly, the two validation cohorts originated from Japan, which may account for some of the observed differences as worldwide variations in clinical outcome in colorectal cancer patients have been shown.[35]

In conclusion, the differential expression of *SYK(T)* and its splice variants between left and right-sided tumors, MSI and MSS tumors, and tumors with and without a *BRAF* and/or *PTEN* mutation suggest a different role for *SYK* in hypermutated and

non-hypermutated tumors. Furthermore, high *SYK(S)* was associated with poor HFS in the prospectively collected MATCH cohort of patients with chemonaive LNN colon cancer. However, the association was not confirmed in two independent, clinically less well-defined and smaller cohorts of patients with chemonaive LNN colon cancer. Further research is warranted to elucidate the role of *SYK* and its splice variants in colorectal cancer.

REFERENCES

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer 2013;49:1374-403.

2. Sargent DJ, Patiyil S, Yothers G, et al. End points for colon cancer adjuvant trials: observations and recommendations based on individual patient data from 20,898 patients enrolled onto 18 randomized trials from the ACCENT Group. J Clin Oncol 2007;25:4569-74.

3. Elferink MA, de Jong KP, Klaase JM, Siemerink EJ, de Wilt JH. Metachronous metastases from colorectal cancer: a population-based study in North-East Netherlands. Int J Colorectal Dis 2015;30:205-12.

4. Lochhead P, Kuchiba A, Imamura Y, et al. Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication. J Natl Cancer Inst 2013;105:1151-6.

5. Roth AD, Delorenzi M, Tejpar S, et al. Integrated analysis of molecular and clinical prognostic factors in stage II/III colon cancer. J Natl Cancer Inst 2012;104:1635-46.

6. Krause DS, Van Etten RA. Tyrosine kinases as targets for cancer therapy. N Engl J Med 2005;353:172-87.

7. Hunter T. Tyrosine phosphorylation: thirty years and counting. Curr Opin Cell Biol 2009;21:140-6.

8. Coopman PJ, Mueller SC. The Syk tyrosine kinase: a new negative regulator in tumor growth and progression. Cancer Lett 2006;241:159-73.

9. Yanagi S, Inatome R, Takano T, Yamamura H. Syk expression and novel function in a wide variety of tissues. Biochem Biophys Res Commun 2001;288:495-8.

10. Krisenko MO, Geahlen RL. Calling in SYK: SYK's dual role as a tumor promoter and tumor suppressor in cancer. Biochim Biophys Acta 2015;1853:254-63.

11. Shin G, Kang TW, Yang S, Baek SJ, Jeong YS, Kim SY. GENT: gene expression database of normal and tumor tissues. Cancer Inform 2011;10:149-57.

12. Hong J, Yuan Y, Wang J, et al. Expression of variant isoforms of the tyrosine kinase SYK determines the prognosis of hepatocellular carcinoma. Cancer Res 2014;74:1845-56.

13. Yang Z, Huo L, Chen H, et al. Hypermethylation and prognostic implication of Syk gene in human colorectal cancer. Med Oncol 2013;30:586.

14. Ni B, Hu J, Chen D, et al. Alternative splicing of spleen tyrosine kinase differentially regulates colorectal cancer progression. Oncol Lett 2016;12:1737-44.

15. Sanz-Pamplona R, Berenguer A, Cordero D, et al. Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. PLoS One 2012;7:e48877.

16. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

17. McShane LM, Altman DG, Sauerbrei W, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). Eur J Cancer 2005;41:1690-6.

18. Oncoline. Netherlands Comprehensive Cancer Organisation, 2017. at www.oncoline.nl.)

19. Ozawa T, Matsuyama T, Toiyama Y, et al. CCAT1 and CCAT2 long noncoding RNAs, located within the 8q.24.21 'gene desert', serve as important prognostic biomarkers in colorectal cancer. Ann Oncol 2017:Forthcoming.

20. Sieuwerts AM, Meijer-van Gelder ME, Timmermans M, et al. How ADAM-9 and ADAM-11 differentially from estrogen receptor predict response to tamoxifen treatment in patients with recurrent breast cancer: a retrospective study. Clin Cancer Res 2005;11:7311-21.

21. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001;25:402-8.

22. Sieuwerts AM, Lyng MB, Meijer-van Gelder ME, et al. Evaluation of the ability of adjuvant tamoxifen-benefit gene signatures to predict outcome of hormone-naive estrogen receptor-positive breast cancer patients treated with tamoxifen in the advanced setting. Molecular oncology 2014;8:1679-89.

23. Kennedy RD, Bylesjo M, Kerr P, et al. Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue. J Clin Oncol 2011;29:4620-6.

24. Kloosterman WP, Coebergh van den Braak RRJ, Pieterse M, et al. A Systematic Analysis of Oncogenic Gene Fusions in Primary Colon Cancer. Cancer Res 2017;77:3814-22.

25. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.

26. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 2017;45:D777-D83.

27. Blancato J, Graves A, Rashidi B, et al. SYK allelic loss and the role of Syk-regulated genes in breast cancer survival. PLoS One 2014;9:e87610.

28. Singh A, Greninger P, Rhodes D, et al. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. Cancer Cell 2009;15:489-500.

29. Lin EI, Tseng LH, Gocke CD, et al. Mutational profiling of colorectal cancers with microsatellite instability. Oncotarget 2015;6:42334-44.

30. Wang L, Duke L, Zhang PS, et al. Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. Cancer Res 2003;63:4724-30.

31. Latour S, Chow LM, Veillette A. Differential intrinsic enzymatic activity of Syk and Zap-70 protein-tyrosine kinases. J Biol Chem 1996;271:22782-90.

32. Prinos P, Garneau D, Lucier JF, et al. Alternative splicing of SYK regulates mitosis and cell survival. Nat Struct Mol Biol 2011;18:673-9.

33. Chen L, Monti S, Juszczynski P, et al. SYK inhibition modulates distinct PI3K/AKT- dependent survival pathways and cholesterol biosynthesis in diffuse large B cell lymphomas. Cancer Cell 2013;23:826-38.

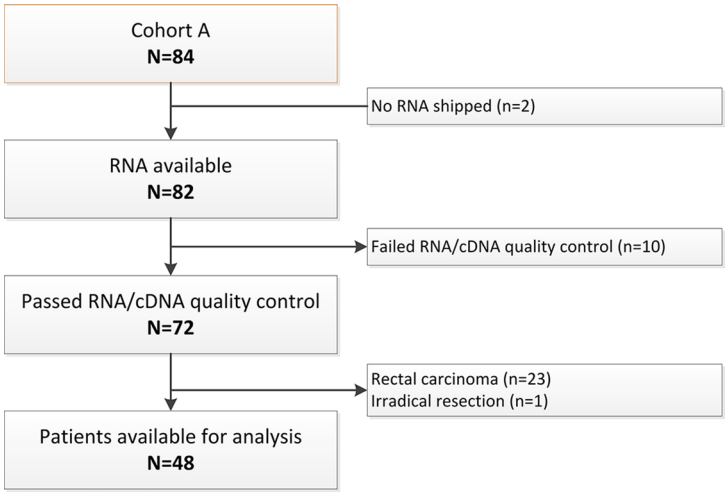
34. Gelsomino F, Barbolini M, Spallanzani A, Pugliese G, Cascinu S. The evolving role of microsatellite instability in colorectal cancer: A review. Cancer Treat Rev 2016;51:19-26.

35. Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin 2017;67:177-93.

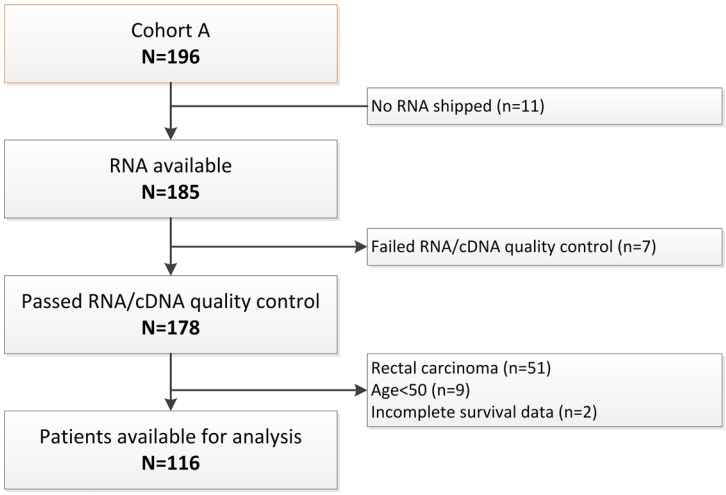
6

SUPPLEMENTARY DATA

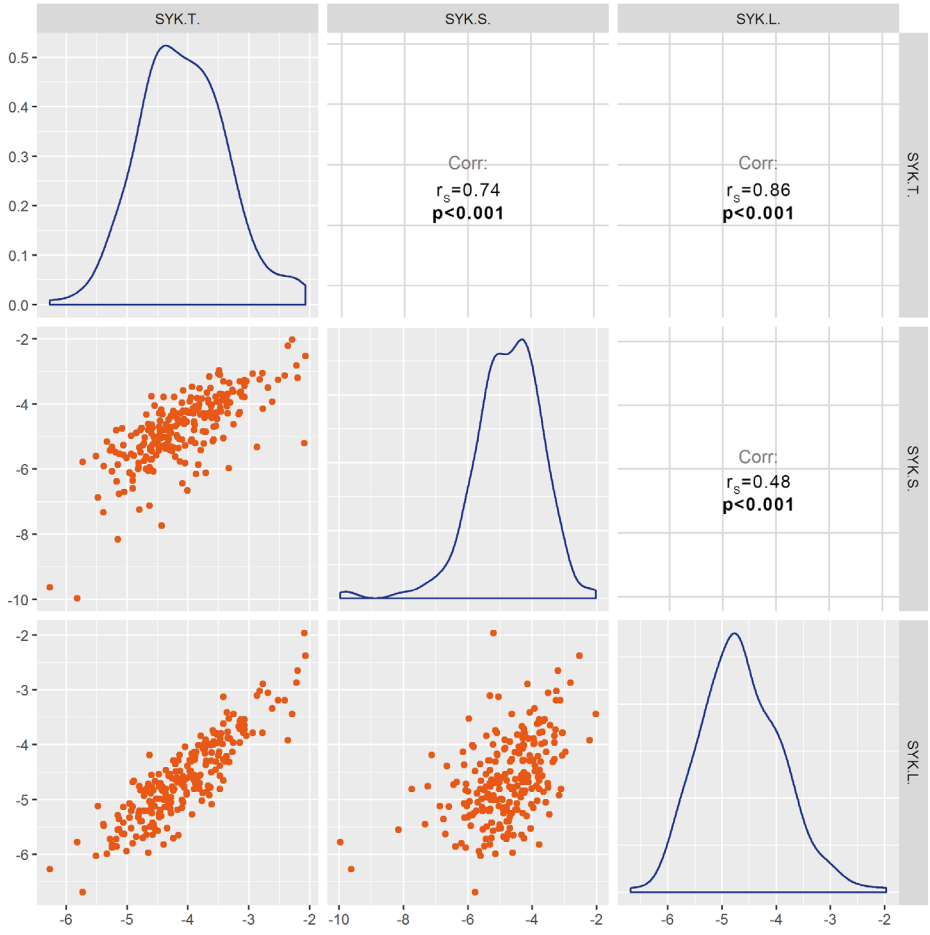
SUPPLEMENTARY FIGURES



SUPPLEMENTARY FIGURE 1. DIAGRAM OF ANALYSIS WORKFLOW OF VALIDATION COHORT A



SUPPLEMENTARY FIGURE 2. DIAGRAM OF ANALYSIS WORKFLOW OF VALIDATION COHORT B



SUPPLEMENTARY FIGURE 3. CORRELATION PLOTS AND PEARSON CORRELATION COEFFICIENTS OF THE CORRELATION BETWEEN *SYK(T)*, *SYK(S)* AND *SYK(L)*

SUPPLEMENTARY TABLE 1.

Gene assays used to measure mRNA expression of *SYK*, *SYK* splice variants and reference genes, and generate EMT, infiltrate and GGI indices.

INDEX	GENE SYMBOL	GENE NAME	QPCR DETECTION METHOD	ASSAY ID THERMOFISHER SCIENTIFIC	F SEQUENCE	R SEQUENCE
SYK total	*SYK (T)*	Spleen Tyrosine Kinase	Taqman	Hs00374292_m1		
SYK short	*SYK (S)*	Spleen Tyrosine Kinase	Taqman	Hs00177369_m1		
SYK long	*SYK (L)*	Spleen Tyrosine Kinase	Taqman	Hs00895384_m1		
EMT	*BGN*	biglycan	Taqman	Hs00959141_g1		
EMT	*CDH2*	cadherin 2, type 1, N-cadherin (neuronal)	Taqman	Hs00983062_m1		
EMT	*FAP*	fibroblast activation protein, alpha	Taqman	Hs00990806_m1		
EMT	*FN1*	fibronectin 1	Taqman	Hs00277509_m1		
EMT	*INHBA*	inhibin, beta A	Taqman	Hs01081598_m1		
EMT	*EPCAM*	tumor-associated calcium signal transducer 1	SYBR		AGTTTGCGGACTGCACTTCA	AATACTCGTGATAAATTTTGGATCCA
EMT	*ESR1*	estrogen receptor 1	SYBR		ATCCTACCAGACCCTTCAGTG	GCCAGACGAGACCAATCATC
EMT	*ESR2*	estrogen receptor 2 (ER beta)	SYBR		CATGCTCCTGGCAACTACTTC	GCTCTTGGCAATCACCCAAAC
EMT	*IGF1*	insulin-like growth factor 1 (somatomedin C)	SYBR		TGGTGGATGCTCTTCAGTTC	GACAGAGCGAGCTGACTTG
EMT	*IGF2*	insulin-like growth factor 2 (somatomedin A)	SYBR		GCGGCTTCTACTTCAGCAG	CAGGTGTCATATTGGAAGAAC
EMT	*TGFB1*	transforming growth factor, beta 1	SYBR		GCCCTGGACACCAACTATTG	CGTGTCCAGGCTCCAAATG
EMT	*VIM*	vimentin	SYBR		CAGATTCAGGAACAGCATGTC	TCAGAGAGGTCAGCAAACTTG
EMT & Infiltrate	*VEGFA*	vascular endothelial growth factor A	Taqman	Hs00900055_m1		
Infiltrate	*PTPRC*	protein tyrosine phosphatase, receptor type, C, CD45	Taqman	Hs00236304_m1		
Reference gene	*HMBS*	hydroxymethylbilane synthase	SYBR		CATGTCTGGTAACGGCAATG	GTACGAGGCTTTCAATGTTG
Reference gene	*HPRT1*	hypoxanthine phosphoribosyltransferase 1	SYBR		TATTGTAATGACCAGTCAACAG	GGTCCTTTTCACCAGCAAG
Reference gene	*TBP*	TATA-box binding protein	SYBR		TTCGGAGAGTTCTGGGATTG	ACGAAGTGCAATGGTCTTTAG

SUPPLEMENTARY TABLE 2A. CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS OF THE LNN SUBGROUP OF THE MATCH COHORT

			SYK(T)		*SYK(S)*		*SYK(L)*		
			MEDIAN (IQR)	P VALUE	MEDIAN (IQR)	P VALUE	MEDIAN (IQR)	P VALUE	PERFORMED TEST
	n	%							
Gender									
Female	78	48.8%	-4.24 (-4.60 • -3.75)	0.07	-4.71 (-5.43 • -4.13)	0.09	-4.70 (-5.13 • -4.13)	0.17	Mann-Whitney U
Male	82	51.3%	-3.97 (-4.61 • -3.42)		-4.48 (-5.14 • -3.82)		-4.56 (-5.14 • -3.91)		
Age	160	100%	-1.12	0.14	0.11	0.17	0.11	0.17	Spearman's Rho
Tumor stage									
Stage I	60	37.5%	-4.31 (-4.71 • -3.67)	0.036	-4.62 (-5.21 • -3.96)	0.92	-4.73 (-5.26 • -4.28)	0.012	Jonckheere-Terpstra
Stage II	100	62.5%	-3.96 (-4.55 • -3.47)		-4.61 (-5.36 • -4.07)		-4.54 (-5.04 • -3.95)		
Stage III	-	-	-		-		-		
T status									
T2	60	37.5%	-4.31 (-4.71 • -3.67)	0.036	-4.62 (-5.21 • -3.96)	0.92	-4.73 (-5.26 • -4.28)	0.012	Mann-Whitney U
T3	100	62.5%	-3.96 (-4.55 • -3.47)		-4.61 (-5.36 • -4.07)		-4.54 (-5.04 • -3.95)		
Nodal status									
N0	131	81.9%	-4.08 (-4.60 • -3.62)	0.60	-4.63 (-5.32 • -4.08)	0.39	-4.62 (-5.13 • -4.04)	0.98	Jonckheere-Terpstra
Nx	29	18.1%	-3.81 (-4.66 • -3.43)		-4.30 (-5.28 • -3.74)		4.68 (-5.17 • -3.90)		
N1	-	-	-		-		-		
N2	-	-	-		-		-		
Tumor grade									
Good	13	8.1%	-3.85 (-4.52 • -3.40)	0.55	-4.70 (-5.45 • -3.77)	0.42	-4.58 (-4.80 • -3.68)	0.51	Jonckheere-Terpstra
Moderate	135	84.4%	-4.12 (-4.60 • -3.58)		-4.54 (-5.20 • -4.05)		-4.66 (-5.14 • -4.01)		
Poor	9	5.6%	-4.07 (-4.71 • -3.69)		-5.37 (-5.79 • -4.09)		-4.75 (-5.16 • -3.94)		
Other[a]	3	1.9%	-		-		-		
Location									
Right	82	51.3%	-4.16 (-4.80 • -3.58)	0.16	-4.79 (-5.45 • -4.15)	0.015	-4.77 (-5.18 • -4.00)	0.14	Mann-Whitney U
Left	78	48.8%	-4.04 (-4.50 • -3.52)		-4.37 (-5.10 • -3.79)		-4.58 (-4.94 • -4.03)		
MSI status[b]									
MSI	37	23.3%	-4.59 (-5.11 • -4.21)	<0.001	-5.31 (-5.58 • -4.77)	<0.001	-4.99 (-5.58 • -4.54)	<0.001	Mann-Whitney U
MSS	122	76.7%	-3.94 (-4.47 • -3.45)		-4.29 (-5.06 • -3.82)		-4.56 (-4.99 • -3.96)		

[a] there were no events in this subgroup. [b] n=1 missing

SUPPLEMENTARY TABLE 2B. CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS OF THE LNP SUBGROUP OF THE MATCH COHORT

			SYK(T)		*SYK(S)*		*SYK(L)*		
			MEDIAN (IQR)	P VALUE	MEDIAN (IQR)	P VALUE	MEDIAN (IQR)	P VALUE	PERFORMED TEST
	n	%							
Gender									
Female	34	42.5%	-4.20 (-4.60 • -3.59)	0.79	-5.07 (-5.75 • -4.14)	0.83	-4.79 (-5.37 • -4.30)	0.97	Mann-Whitney U
Male	46	57.5%	-4.12 (-4.52 • -3.70)		-4.92 (-5.69 • -4.22)		-4.84 (-5.27 • -4.38)		
Age	80	100%	0.05	0.65	0.11	0.31	0.11	0.31	Spearman's Rho
Tumor stage									
Stage I	-	-	-		-		-		Jonckheere-Terpstra
Stage II	-	-	-		-		-		
Stage III	80	100%	-4.15 (-4.58 • -3.68)	-	-4.98 (-5.70 • -4.19)	-	-4.81 (-5.30 • -4.33)	-	
T status									
T2	11	13.8%	-4.34 (-4.63 • -3.76)	0.42	-4.69 (-5.30 • -4.11)	0.56	-5.04 (-5.70 • -4.68)	0.08	Mann-Whitney U
T3	69	86.3%	-4.13 (-4.58 • -3.64)		-5.02 (-5.72 • -4.21)		-4.77 (-5.20 • -4.30)		
Nodal status									
N0	-	-	-		-		-		
Nx	-	-	-		-		-		
N1	53	66.3%	-4.09 (-4.48 • -3.64)	0.38	-4.89 (-5.68 • -4.21)	0.67	-4.77 (-5.34 • -4.38)	0.86	Jonckheere-Terpstra
N2	27	33.8%	-4.30 (-4.60 • -3.69)		-5.08 (-5.74 • -4.03)		-4.96 (-5.21 • -4.31)		
Tumor grade									
Good	7	8.8%	-4.16 (-4.34 • -3.85)	0.98	-4.52 (-5.23 • -4.11)	0.68	-4.68 (-5.01 • -4.29)	0.71	Jonckheere-Terpstra
Moderate	57	71.3%	-4.16 (-4.59 • -3.65)		-5.09 (-5.75 • -4.26)		-4.91 (-5.44 • -4.37)		
Poor	11	13.8%	-4.14 (-4.58 • -3.76)		-4.89 (-6.65 • -4.33)		-4.80 (-5.36 • -4.19)		
Other	5	6.3%	-3.67 (-4.62 • -3.35)		-3.87 (-6.41 • -3.71)		-4.46 (-4.82 • -3.88)		
Location									
Right	39	48.8%	-4.29 (-4.71 • -3.72)	0.020	-5.25 (-5.99 • -4.46)	0.07	-5.09 (-5.53 • -4.48)	0.01	Mann-Whitney U
Left	41	51.3%	-3.95 (-4.38 • -3.51)		-4.70 (-5.35 • -4.09)		-4.65 (-5.00 • -4.24)		
MSI status									
MSI	12	15.0%	-4.59 (-4.80 • -4.28)	0.002	-5.72 (-6.57 • -5.29)	0.00	-5.19 (-5.34 • -4.59)	0.15	Mann-Whitney U
MSS	68	85.0%	-4.02 (-4.46 • -3.56)		-4.71 (-5.41 • -4.02)		-4.76 (-5.19 • -4.31)		

SUPPLEMENTARY TABLE 3.

The association between epithelial and mesenchyal markers and *SYK(T)*, *SYK(S)* and *SYK(L)* for the total MATCH cohort, and the LNN and LNP subgroups of the MATCH cohort.

	EPCAM		BGN		FAP		INHBA	
	r_s	P value	r_s	P value	r_s	P value	r_s	P value
Total MATCH cohort								
SYK(T)	0.47	<0.001	0.00	0.97	-0.08	0.22	-0.01	0.86
SYK(S)	0.58	<0.001	-0.06	0.40	-0.13	0.046	-0.07	0.31
SYK(L)	0.41	<0.001	0.12	0.07	0.03	0.65	0.08	0.24
LNN cohort								
SYK(T)	0.47	<0.001	0.02	0.83	-0.06	0.46	-0.02	0.80
SYK(S)	0.55	<0.001	0.02	0.80	-0.06	0.43	-0.05	0.53
SYK(L)	0.36	<0.001	0.10	0.20	0.03	0.75	0.04	0.60
LNP cohort								
SYK(T)	0.51	<0.001	-0.01	0.91	-0.13	0.27	-0.00	0.99
SYK(S)	0.59	<0.001	-0.21	0.07	-0.24	0.031	-0.17	0.14
SYK(L)	0.39	<0.001	0.16	0.16	0.06	0.58	0.12	0.29

SUPPLEMENTARY TABLE 4.

The association between infiltrate markers and *SYK(T)*, *SYK(S)* and *SYK(L)* for the total MATCH cohort, and the LNN and LNP subgroups of the MATCH cohort.

	PTPRC/CD45		VEGFA		% OF INFILTRATING CELLS		% OF INVASIVE TUMOR CELLS	
	r_s	P value	r_s	P value	r_s	P value	r_s	P value
Total MATCH cohort								
SYK(T)	0.09	0.18	0.16	0.012	-0.04	0.56	-0.10	0.14
SYK(S)	0.12	0.07	0.16	0.015	-0.14	0.043	0.03	0.64
SYK(L)	0.08	0.21	0.12	0.06	0.04	0.56	-0.09	0.18
LNN cohort								
SYK(T)	0.02	0.81	0.16	0.05	-0.04	0.63	-0.10	0.19
SYK(S)	0.09	0.26	0.19	0.017	-0.24	0.007	0.04	0.59
SYK(L)	0.01	0.93	0.13	0.09	0.05	0.62	-0.14	0.07
LNP cohort								
SYK(T)	0.18	0.11	0.18	0.11	-0.03	0.80	-0.10	0.38
SYK(S)	0.10	0.38	0.09	0.42	0.02	0.87	-0.14	0.21
SYK(L)	0.08	0.48	0.10	0.39	0.07	0.57	-0.13	0.24

SUPPLEMENTARY TABLE 5A. UNIVARIATE COX REGRESSION ANALYSIS FOR THE TOTAL MATCH COHORT

			DFS (EVENTS=48)		HFS (EVENTS=19)		OS (EVENTS=42)	
			HR (95%CI)	P value	HR (95%CI)	P value	HR (95%CI)	P value
	n	%						
mRNA expression								
SYK(T)	240	100%	1.10 (0.74 • 1.64)	0.62	1.57 (0.86 • 2.88)	0.14	0.78 (0.50 • 1.22)	0.28
SYK(S)	240	100%	1.07 (0.82 • 1.41)	0.61	1.57 (0.97 • 2.54)	0.07	0.87 (0.67 • 1.14)	0.32
SYK(L)	240	100%	0.98 (0.66 • 1.43)	0.90	1.26 (0.70 • 2.26)	0.43	0.80 (0.52 • 1.24)	0.31
Gender								
Female	112	46.7%	1		1		1	
Male	128	53.3%	1.25 (0.70 • 2.21)	0.45	1.22 (0.49 • 3.04)	0.66	2.26 (1.15 • 4.41)	0.017
Age	240	100%	1.0002 (0.96 • 1.04)	0.99	1.001 (0.95 • 1.06)	0.96	1.06 (1.02 • 1.10)	0.004
Tumor stage								
Stage I	60	25.0%	1		1		1	
Stage II	100	41.7%	1.36 (0.59 • 3.13)	0.47	1.21 (0.36 • 4.02)	0.76	1.11 (0.47 • 2.61)	0.82
Stage III	80	33.3%	2.25 (1.002 • 5.06)	0.049	1.42 (0.42 • 4.87)	0.57	1.90 (0.83 • 4.34)	0.13
T status								
T2	71	29.6%	1		1		1	
T3	169	70.4%	1.93 (0.93 • 3.98)	0.08	1.23 (0.45 • 3.43)	0.69	1.94 (0.89 • 4.19)	0.09
Nodal status								
N0	131	54.6%	1		1		1	
Nx	29	12.1%	1.15 (0.43 • 3.05)	0.78	3.41 (1.08 • 10.74)	0.04	1.53 (0.60 • 3.89)	0.37
N1	53	22.1%	1.20 (0.57 • 2.55)	0.64	1.82 (0.58 • 5.72)	0.31	1.003 (0.42 • 2.42)	0.99
N2	27	11.3%	3.58 (1.76 • 7.29)	<0.001	1.71 (0.36 • 8.26)	0.50	4.42 (2.10 • 9.27)	<0.001
Tumor grade								
Good	20	8.3%	1		1		1	
Moderate	192	80.0%	0.71 (0.28 • 1.80)	0.47	0.71 (0.16 • 3.12)	0.65	1.67 (0.40 • 6.97)	0.48
Poor	20	8.3%	1.37 (0.43 • 4.31)	0.59	1.50 (0.25 • 8.97)	0.66	4.18 (0.89 • 19.70)	0.07
Other[a]	8	3.3%	-	-	-	-	-	-
Location								
Right	121	50.4%	1		1		1	
Left	119	49.6%	1.19 (0.68 • 2.11)	0.54	0.91 (0.37 • 2.24)	0.84	0.58 (0.31 • 1.08)	0.09
MSI status[b]								
MSI	49	20.5%	1		1		1	
MSS	190	79.5%	2.32 (0.92 • 5.85)	0.08	29.05 (0.26 • 3292.13)	0.16	0.67 (0.34 • 1.32)	0.25

[a] there were no events in this subgroup. [b] n=1 missing

SUPPLEMENTARY TABLE 5B. UNIVARIATE COX REGRESSION ANALYSIS FOR THE LNN SUBGROUP OF THE MATCH COHORT

			DFS (EVENTS=26)		HFS (EVENTS=12)		OS (EVENTS=23)	
			HR (95%CI)	P value	HR (95%CI)	P value	HR (95%CI)	P value
	n	%						
mRNA expression								
SYK(T)	160	100%	1.33 (0.81 • 2.19)	0.26	2.05 (1.01 • 4.17)	0.047	0.80 (0.45 • 1.42)	0.44
SYK(S)	160	100%	1.35 (0.89 • 2.04)	0.16	2.14 (1.14 • 4.01)	0.018	0.91 (0.62 • 1.32)	0.60
SYK(L)	160	100%	1.02 (0.62 • 1.68)	0.95	1.44 (0.72 • 2.87)	0.31	0.78 (0.44 • 1.39)	0.39
Gender								
Female	78	48.8%	1		1		1	
Male	82	51.3%	1.62 (0.73 • 3.56)	0.24	1.97 (0.59 • 6.53)	0.27	2.33 (0.96 • 5.68)	0.06
Age	160	100%	1.002 (0.96 • 1.05)	0.94	1.01 (0.94 • 1.08)	0.79	1.08 (1.02 • 1.14)	0.011
Tumor stage								
Stage I	60	37.5%	1		1		1	
Stage II	100	62.5%	1.36 (0.59 • 3.12)	0.47	1.21 (0.36 • 4.02)	0.76	1.10 (0.47 • 2.61)	0.82
Stage III	-	-	-	-	-	-	-	-
T status								
T2	60	37.5%	1		1		1	
T3	100	62.5%	1.36 (0.59 • 3.12)	0.47	1.21 (0.36 • 4.02)	0.76	1.10 (0.47 • 2.61)	0.82
Nodal status								
N0	131	81.9%	1		1		1	
Nx	29	18.1%	1.16 (0.44 • 3.07)	0.77	3.42 (1.09 • 10.78)	0.036	1.52 (0.60 • 3.87)	0.38
N1	-	-	-	-	-	-	-	-
N2	-	-	-	-	-	-	-	-
Tumor grade								
Good	13	8.1%	1		1		1	
Moderate	135	84.4%	1.02 (0.24 • 4.36)	0.98	0.88 (0.11 • 6.91)	0.27	1.92 (0.26 • 14.38)	0.52
Poor	9	5.6%	2.14 (0.36 • 12.82)	0.40	2.85 (0.26 • 31.39)	0.39	4.47 (0.46 • 43.15)	0.20
Other[a]	3	1.9%	-	-	-	-	-	-
Location								
Right	82	51.3%	1		1		1	
Left	78	48.8%	2.06 (0.92 • 4.62)	0.08	1.09 (0.35 • 3.38)	0.88	0.60 (0.26 • 1.42)	0.25
MSI status[b]								
MSI	37	23.3%	1		1		1	
MSS	122	76.7%	3.90 (0.92 • 16.49)	0.07	31.28 (0.12 • 8430.24)	0.23	0.51 (0.22 • 1.17)	0.11

[a] there were no events in this subgroup. [b] n=1 missing

SUPPLEMENTARY TABLE 5C. UNIVARIATE COX REGRESSION ANALYSIS FOR THE LNP SUBGROUP OF THE MATCH COHORT

			DFS (EVENTS=22)		HFS (EVENTS=7)		OS (EVENTS=19)	
			HR (95%CI)	P value	HR (95%CI)	P value	HR (95%CI)	P value
	n	%						
mRNA expression								
SYK(T)	80	100%	0.78 (0.39 • 1.56)	0.48	0.78 (0.23 • 2.67)	0.69	0.71 (0.33 • 1.52)	0.37
SYK(S)	80	100%	0.95 (0.65 • 1.37)	0.77	1.08 (0.54 • 2.16)	0.83	0.90 (0.60 • 1.34)	0.59
SYK(L)	80	100%	1.08 (0.56 • 2.07)	0.82	1.02 (0.32 • 3.27)	0.98	0.95 (0.46 • 1.94)	0.88
Gender								
Female	34	42.5%	1		1		1	
Male	46	57.5%	0.85 (0.37 • 1.96)	0.70	0.54 (0.12 • 2.41)	0.42	1.96 (0.70 • 5.48)	0.20
Age	80	100%	1.03 (0.96 • 1.10)	0.37	0.99 (0.88 • 1.12)	0.89	1.10 (1.02 • 1.18)	0.01
Tumor stage								
Stage I	-	-	-		-		-	
Stage II	-	-	-		-		-	
Stage III	80	100%	-		-		-	
T status								
T2	11	13.8%	1		1		1	
T3	69	86.3%	3.84 (0.52 • 28.57)	0.19	1.08 (0.13 • 8.97)	0.94	26.85 (0.14 • 5108.71)	0.22
Nodal status								
N0	-	-	-	-	-		-	
Nx	-	-	-	-	-		-	
N1	53	66.3%	1		1		1	
N2	27	33.8%	2.85 (1.23 • 6.60)	0.015	0.93 (0.18 • 4.81)	0.93	4.14 (1.63 • 10.57)	0.003
Tumor grade								
Good	7	8.8%	1		1		1	
Moderate	57	71.3%	0.56 (0.16 • 1.92)	0.35	0.58 (0.07 • 4.93)	0.62	1.43 (0.19 • 10.98)	0.73
Poor	11	13.8%	0.82 (0.18 • 3.66)	0.79	0.63 (0.04 • 10.01)	0.74	3.12 (0.36 • 26.75)	0.30
Other[a]	5	6.3%	-	-	-	-	-	-
Location								
Right	39	48.8%	1		1		1	
Left	41	51.3%	0.60 (0.26 • 1.40)	0.24	0.65 (0.15 • 2.91)	0.57	0.52 (0.20 • 1.31)	0.17
MSI status								
MSS	12	15.0%	1		1		1	
MSI	68	85.0%	0.995 (0.29 • 3.36)	0.99	25.07 (0.002 • 332856.52)	0.51	0.86 (0.25 • 2.98)	0.81

[a] there were no events in this subgroup

R.R.J. Coebergh van den Braak, A.M. Sieuwerts, Z.S. Lalmahomed, M. Smid, S.M. Wilting, S.I. Bril, S. Xiang, M. Daane – van der Vlugt, V. de Weerd, A. van Galen, K. Biermann, J.H. van Krieken, W.P. Kloosterman, J.A. Foekens, J.W.M. Martens, J.N.M. IJzermans, on behalf of the MATCH study group.

7

CONFIRMATION OF A METASTASIS-SPECIFIC MICRORNA SIGNATURE IN PRIMARY COLON CANCER

SUBMITTED

ABSTRACT

The identification of patients with high-risk stage II colon cancer who may benefit from adjuvant therapy remains an unmet need. Understanding of tumour biology may be used to tailor the clinical approach for these patients. MicroRNAs have been proposed as markers for prognosis or treatment response in colorectal cancer. Recently, a 2-microRNA signature (*let-7i* and *miR-10b*) was proposed to identify colorectal cancer patients at risk of developing distant metastasis. We assessed the prognostic value of this signature and additional candidate microRNAs in an independent clinically well-defined prospectively collected cohort of primary colon cancers including stage I-II colon cancer without and stage III colon cancer with adjuvant treatment. The 2-microRNA signature predicted specifically hepatic recurrence in the stage I-II group, but not overall the ability to develop distant metastasis. Addition of *miR-30b* to the 2-microRNA signature allowed prediction of both distant metastasis and hepatic recurrence in patients with stage I-II colon cancer who did not receive adjuvant chemotherapy. Available gene expression data allowed us to associate *miR-30b* expression with axon guidance and *let-7i* expression with cell adhesion, migration and motility.

INTRODUCTION

Colorectal cancer (CRC) is the second most common malignancy in the Western World with close to 450,000 new cases in Europe in 2012.[1] As in most solid cancers, histologic tumour staging (TNM) is to date the best determinant for prognosis and treatment. The current treatment for stage I-III colon cancer is surgery alone for stages I and II, and surgery followed by adjuvant chemotherapy for stage III. Despite treatment, up to 21% of the patients with stage I-II and up to 40% of the patients with stage III colon cancer will develop metastatic disease after curative surgery.[2,3] Therefore, prognostic biomarkers complementing the TNM classification are urgently needed.[4,5] The only biomarker that is currently used to predict prognosis and response to therapy in resectable colon cancer is microsatellite instability (MSI), a phenotype associated with a favourable prognosis compared to microsatellite stable (MSS) tumours.[6]

MicroRNAs (miRNAs) are a group of short noncoding RNAs that regulate gene expression at the post-transcriptional level.[7] In cancer, miRNAs are known to play a central role in key pathways. In CRC, a growing number of miRNAs has been connected to different steps of tumorigenesis and has been proposed as markers for prognosis or treatment response.[8] Recently, Hur et al. identified six miRNAs as potential markers for the development of metastases in CRC patients (*miR-320*, *miR-221*, *miR-30b*, *miR-10b*, *miR-885-5p*, *let-7i*) through a metastasis-specific miRNA biomarker discovery approach showing differential expression of these six miRNAs between primary CRC and paired liver metastases tissues.[9] Two of these miRNAs (*miR-10b* and *let-7i*), measured in primary tumours, were associated with the development of distant metastases. The combination of the expression of these two miRNAs identified a group of patients which remained entirely free of distant metastases.

In this study, we aimed to assess the prognostic value of the above-mentioned miRNAs and the 2-miRNA metastasis-specific signature in a clinically well-defined independent prospectively collected cohort of primary colon cancers. Patients with lymph node negative (LNN) colon cancer who did not receive systemic adjuvant chemotherapy (untreated) and patients with lymph node positive (LNP) colon cancer who did receive adjuvant chemotherapy were analysed separately to distinguish between the natural course of the disease (pure prognosis) and prognosis while receiving adjuvant chemotherapy.

METHODS

All aspects of the guidelines for REporting recommendations for tumour MARKer (REMARK) prognostic studies were followed, and the paper was written accordingly.[10] The study was conducted according to Declaration of Helsinki. All procedures involving human subjects were approved by the Erasmus MC University Medical Centre Institutional Review Board (MEC 2007-088).

PATIENT SELECTION

Patients were selected from the MATCH-cohort, an observational prospective multicentre cohort study from 2007 onwards including patients who undergo curative surgery for CRC in one of seven participating hospitals in the Rotterdam region, the Netherlands. Patients gave written informed consent for the storage and use of biobank samples for research purposes, and the collection of clinical data (Institutional Review Board number MEC 2007-088).

Inclusion criteria for this study were: informed consent available, inclusion date between 1st July 2007 and 1st July 2012, age 55-85 years, stage I-II without adjuvant chemotherapy or stage III with adjuvant chemotherapy, fresh frozen tissue with at least 40% invasive tumour cells available, and either recurrence of disease or at least 30 months of disease-free follow-up. A diagram of patient selection is shown in **Figure 1**.

SAMPLE COLLECTION AND PROCESSING

The resection specimens were transported to the pathology laboratory immediately following removal during surgery. In the pathology laboratory, two to four samples of both central and peripheral regions of the tumour, and one or two adjacent non-tumour colon tissue samples were taken. These samples were fresh frozen with a maximum cold ischemia time of two hours. All samples have been stored in liquid nitrogen.

RNA ISOLATION, CDNA SYNTHESIS AND MRNA TRANSCRIPT LEVEL QUANTIFICATION

A cryostat microtome set at -20°C was used to cut the fresh frozen colon cancer and normal colon tissues (Thermo Scientific Microm HM 560, Thermo Fisher Scientific, inc.). Before, in the middle and after sectioning for RNA isolation, a 5 µm section was cut. After hematoxylin-eosin (HE) staining the sections were reviewed by two pathologists independently. The percentage of neoplastic cells, infiltrating immune cells, necrosis and normal mucosa was scored in categories of 0–5%, 6–10%, 11–20%, 21–30%, 31–40%, 41–50%, 51–60%, 61–70%, 71–80%, 81–90%, and 91–100% tumour cells, relative to other cells. Classification and grading of the tumour was determined

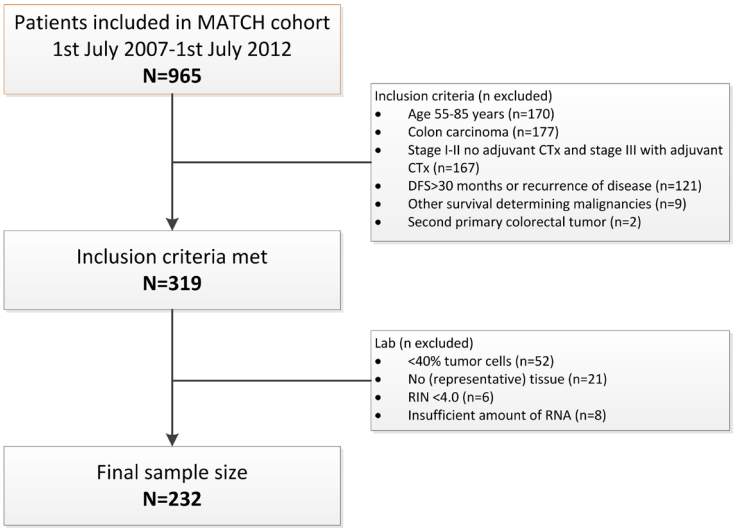


FIGURE 1. DIAGRAM OF THE PATIENT SELECTION

using the WHO 2010 classification for the carcinoma of the colon and rectum (WHO Press, World Health Organization, 20 Avenue Appia, Geneva, Switzerland).

Total RNA was isolated from 30 µm sections using RNA-Bee® according the manufacturer's instructions (Tel-Test inc., USA). Quality and quantity of RNA was assessed with the MultiNA Microchip Electrophoresis system (Shimadzu, Kyoto, Japan) and the Nanodrop ND-1000 (Thermo Scientific, Wilmington, USA), respectively.

Expression levels of the six miRNAs were quantified in a 9-plex miRNA Assay Protocol utilizing commercially available and validated Taqman miRNA assays (Applied Biosystems, Thermo Scientific, USA) relative to the average level of reference genes *miRNA-16*, *RNU6B* and *RNU44* measured in the same total RNA sample.

In brief, 5 µL sample containing 20 ng/µL total RNA was reverse transcribed for 30 min at 16°C, 30 min at 42°C, 5 min at 85°C and stopped at 4°C in the presence of 18 nM final of each of the 9 Taqman RT-primers in 1x RT buffer [Fermentas] supplemented with 0.65 mM each dNTP [Fermentas], 0.25 U/µL RNAseout [Fermentas], 3.8 mM MgCl2 [Invitrogen) in the absence (negative control) or presence of 15 U/µL revertAid MLV H-minus RT [Fermentas]. After the RT-reaction samples were diluted 20-fold in LoTE (3 mM Tris-HCl/0.2 mM EDTA, pH 8.0) prior to performing 40 cycles of individual qPCR reactions for each of the 9 miRNA assays in the presence of Taqman Master mix without UNG as advised by the manufacturer (Applied Biosystems, Thermo Scientific, USA). Specifics of the 6 target miRNA assays and 3 reference miRNA assays used to normalize the data are provided in **Supplementary Table 1**.

MICROSATELLITE INSTABILITY (MSI)

In short, genomic DNA was extracted from 2 to 5 x 30 µm sections cut in between the sections used for RNA isolation (NucleoSpin Tissue kit, Macherey-Nagel, BIOKE, Leiden, the Netherlands). MSI status was determined with a fluorescent PCR-based assay in five mononucleotide repeat markers (BAT-25, BAT-26, NR-21, NR-24 and MONO-27 (Promega MSI Analysis System) using 2 ng of PicoGreen measured DNA. Quality and quantity were assessed by both agarose gel electrophoresis, Nanodrop and the Quant-iT PicoGreen dsDNA kit (Life Technnologies). Two pentanucleotide repeat markers (Penta C and Penta D) were also included to detect potential sample mix ups and/or contamination using the protocol as supplied by the manufacturer.

PATHWAY ANALYSIS

For n=231 patients, RNA sequencing data were available.[11] In short, we used the STAR[12] algorithm (version 2.4.2a) to align the RNA-seq data with the GRCh38 reference using the '--quantmode Genecounts' option to obtain the raw readcounts for each gene. Gene annotation was derived from gencode v23 (https://www.gencodegenes.org/). Next, the Trimmed Mean of M-values normalization[13] - as implemented in EdgeR[14] - was used to normalize the raw read count data. These data were used as input for the pathway analysis.

For the miRNAs associated with long-term clinical outcome in our cohort, the 50 tumours with the highest expression and 50 tumours with the lowest expression per miRNA were grouped and used as input. Pathway analyses were performed using the R-package 'global test' using KEGG.[15] Importantly, only genes for which expression data was available were used as input. The Bonferroni-Holm method was used to correct all p values for multiple testing. Re-sampling (n=1,000) was performed to determine the number of times a randomly selected group of genes of equal size was at least as significant as the true set of genes assigned to a pathway. The gene plots of pathways with a corrected p value < 0.05 and a re-sampling probability <0.05 were reviewed. Three target gene prediction databases were used to identify which genes within the selected pathways were predicted to be potential targets of the respective miRNAs (Targetscan version 7.1, http://www.targetscan.org[16]; MicroRNA Target Prediction and Functional Study Database [miRDB], http://mirdb.org/miRDB[17]; RNA22 version 2.0, https://cm.jefferson.edu/rna22/[18]).We considered genes to be a potential target of a miRNA when predicted by all three databases and when the binding site was a conserved site within the 3' UTR region.

SURVIVAL DATA

Metastasis free survival (MFS) was defined as the time elapsed between date of surgery, and either date of the appearance of distant metastasis or date of the last follow-up visit at which a patient was considered to have no recurrence.
Hepatic metastasis free survival (HFS) was defined as the time elapsed between date of surgery, and either date of the appearance of liver metastasis or date of the last follow-up visit at which a patient was considered to have no liver metastases.
Overall survival (OS) was defined as the time elapsed between date of surgery, and either date of death or date of the last check in the Municipal Personal Records Database.

STATISTICAL ANALYSES

Statistical analyses were performed using the SPSS statistical package version 21. Associations between the expression of the six miRNAs as a continuous variable, and clinical and histopathological characteristics were assessed using the Mann-Whitney U test, Spearman Rank correlation test, Kruskal-Wallis test and Jonckheere-Terpstra test where appropriate. Cox regression analysis was used to assess the association between the expression of the six miRNAs as a continuous variable, and MFS, HFS and OS. A one-way ANOVA was used to assess the difference between the two log likelihood estimates of the Cox regression models when adding a variable to the model. Kaplan Meier estimates were used to visualize the relevant associations between miRNAs and long-term clinical outcome. The Youden's index was calculated as previously described.[19] All analyses were two-sided and $p<0.05$ was considered significant.

DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article (and its Supplementary Information files). The RNA sequencing data have already been published and made available elsewhere.[11]

RESULTS

ASSOCIATION OF MIRNA EXPRESSION LEVELS WITH CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS

The total cohort consisted of 232 patients, 155 patients with LNN primary colon cancer and 77 patients with LNP primary colon cancer.

First, the distribution of the mRNA expression levels for each miRNA, and the correlation between the expression levels of the miRNAs were assessed (**Supplementary Figure 1**). The only miRNA for which the expression levels did not follow a normal distribution was miR-885-5p ($p<0.001$). Furthermore, the Spearman correlation between the six miRNAs was assessed. All miRNAs were significantly and positively correlated with three or more of the six assessed miRNAs. However, only poor to moderate associations were observed (spearman's rho 0.13 - 0.33, $p<0.001$ - $p=0.044$) (**Supplementary Figure 1**). Then, possible associations between miRNA expression levels and clinical and histopathological characteristics were assessed. Importantly, most of the significant differences in miRNA expression observed in the total group derived from differences in the LNN group and not the smaller LNP group. Since we were primarily interested in the association with pure disease prognosis, the LNN subgroup was the main focus for further analyses. The associations between clinical and histopathological features for the LNN group are shown in **Table 1**, for the total group in **Supplementary Table 2A** and for the LNP group in **Supplementary Table 2B**.

Expression of *miR-221*, *miR-30b* and *miR-885-5p* was significantly lower while expression of *miR-10b* was significantly higher in MSI tumours compared to Microsatellite Stable (MSS) tumours (**Table 1**). Expression of *miR-221* and *miR-30b* was significantly lower and *miR-10b* significantly higher in left-sided tumours compared to right-sided tumours. Besides the associations of the miRNAs with these two clinically important characteristics, expression of *let-7i* was significantly lower in stage II compared to stage I tumours.

LET-7I, *MIR-30B* AND MIRNA SIGNATURE AS PROGNOSTIC MARKERS FOR CLINICAL OUTCOME

First the prognostic value of the six miRNAs was assessed using the expression levels as a continuous variable in a univariate Cox regression model (**Table 2**). *Let-7i* expression levels were significantly associated with HFS (Hazard Ratio [HR]=0.32, 95% Confidence Interval [CI]=0.17-0.60, $p<0.001$). In contrast to the findings of Hur et al., *miR-10b* was not significantly associated with any of the clinical endpoints. The expression of m*iR-30b* however was significantly associated with MFS and HFS (HR=2.13, 95%CI=1.22-3.72, $p=0.008$ and HR=2.77, 95%CI=1.24-6.18, $p=0.013$, respectively). None of the other miRNAs were associated with MFS and HFS, and none of the miRNAs were significantly associated with OS.

TABLE 1. CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS OF THE LYMPH NODE NEGATIVE PATIENTS

	N	%	*MIR-320*	IQR	P VALUE	*MIR-221*	IQR	P VALUE	*MIR-30B*	IQR	P VALUE	*MIR-10B*	IQR	P VALUE	*MIR-885-5P*	IQR	P VALUE	*LET-7I*	IQR	P VALUE
Gender																				
Female	73	47.1%	-4.50	0.92	0.94	-1.78	1.37	0.42	-1.79	1.10	0.025	-2.55	1.33	0.09	-11.20	2.69	0.96	-2.97	1.17	0.14
Male	82	52.9%	-4.55	0.98		-1.77	1.43		-1.59	0.86		-2.68	1.04		-11.39	2.26		-3.16	1.39	
Age	155	100.0%	-0.05		0.54	-0.13		0.11	0.02		0.82	0.09		0.28	0.02		0.79	0.07		0.42
Tumour stage																				
Stage I	57	36.8%	-4.61	0.97	1.00	-1.78	1.57	0.65	-1.55	0.90	0.13	-2.62	1.39	0.96	-11.21	2.79	0.16	-2.59	1.46	0.037
Stage II	98	63.2%	-4.49	0.91		-1.77	1.36		-1.75	0.81		-2.66	1.11		-11.39	2.48		-3.15	0.95	
Stage III			-	-		-	-		-	-		-	-		-	-		-	-	
T status																				
T2	57	36.8%	-4.61	0.97	1.00	-1.78	1.57	0.65	-1.55	0.90	0.13	-2.62	1.39	0.96	-11.21	2.79	0.16	-2.59	1.46	0.037
T3	98	63.2%	-4.49	0.91		-1.77	1.36		-1.75	0.81		-2.66	1.11		-11.39	2.48		-3.15	0.95	
Nodal status																				
N0	127	81.9%	-4.47	0.94	0.38	-1.77	1.38	0.85	-1.75	0.81	0.08	-2.58	1.31	0.18	-11.12	2.78	0.64	-3.05	1.02	0.40
N0 <10 nodes	28	18.1%	-4.72	1.01		-1.71	1.80		-1.40	1.08		-2.65	1.26		-11.04	3.50		-3.14	1.44	
N1			-	-		-	-		-	-		-	-		-	-		-	-	
N2			-	-		-	-		-	-		-	-		-	-		-	-	
Tumour grade																				
Good	13	8.4%	-4.38	0.92	0.77	-1.59	0.91	0.21	-1.54	0.70	0.95	-2.74	1.40	0.98	-11.61	2.50	0.28	-3.01	1.47	0.51
Moderate	130	83.9%	-4.52	0.99		-1.82	1.37		-1.74	0.87		-2.65	1.18		-11.32	2.57		-3.12	1.21	
Poor	9	5.8%	-4.58	1.18		-1.98	2.37		-1.73	1.58		-2.87	0.96		-11.64	2.31		-2.87	1.21	
Other	3	1.9%	-4.27	-		-1.77	-		-1.08	-		-1.59	-		-9.15	-		-2.55	-	
Location																				
Right	79	51.0%	-4.49	0.79	0.76	-2.16	1.22	0.001	-1.82	0.70	0.029	-2.49	0.99	<0.001	-11.22	2.59	0.70	-3.16	1.22	0.55
Left	76	49.0%	-4.58	1.10		-1.48	1.39		-1.56	0.87		-2.94	1.12		-11.35	2.34		-3.03	1.11	
MSI-status[a]																				
MSS	120	77.9%	-4.51	1.03	0.44	-1.61	1.36	<0.001	-1.61	0.84	0.007	-2.73	1.12	<0.001	-11.21	2.39	0.011	-3.08	1.33	0.50
MSI	34	22.1%	-4.50	0.80		-2.53	0.96		-2.00	0.59		-2.12	1.18		-12.33	2.43		-3.01	0.88	

[a] n=1 missing

TABLE 2. UNIVARIATE AND MULTIVARIATE COX REGRESSION ANALYSIS FOR THE LNN GROUP

				MFS (EVENTS=25)				HFS (EVENTS=12)				OS (EVENTS=23)			
				HR	95% CI		p value	HR	95% CI		p value	HR	95% CI		p value
Univariate															
		n	%												
mRNA expression	*MiR-320*	155	100.0%	0.83	0.51	1.35	0.46	0.71	0.36	1.40	0.32	0.86	0.51	1.47	0.59
	MiR-221	155	100.0%	1.46	1.00	2.13	0.051	0.92	0.53	1.61	0.78	0.99	0.65	1.51	0.96
	MiR-30b	155	100.0%	2.07	1.21	3.53	0.008	2.94	1.38	6.25	0.005	1.28	0.69	2.38	0.44
	MiR-10b	155	100.0%	0.96	0.60	1.53	0.85	0.76	0.39	1.48	0.42	0.84	0.51	1.38	0.48
	MiR-885-5p	155	100.0%	0.89	0.69	1.13	0.33	0.90	0.63	1.28	0.56	0.77	0.57	1.04	0.08
	Let-7i	155	100.0%	0.66	0.41	1.06	0.087	0.28	0.14	0.56	<0.001	0.74	0.47	1.15	0.18
Gender		155	100.0%	1.41	0.63	3.14	0.40	1.84	0.55	6.11	0.32	2.23	0.92	5.42	0.08
Age		155	100.0%	1.00	0.96	1.05	0.88	1.01	0.94	1.09	0.75	1.08	1.02	1.14	0.009
Tumour stage		155	100.0%	1.51	0.63	3.61	0.36	1.17	0.35	3.90	0.79	1.07	0.45	2.53	0.88
T status		155	100.0%	1.51	0.63	3.61	0.36	1.17	0.35	3.90	0.79	1.07	0.45	2.53	0.88
Nodal status		155	100.0%	1.20	0.45	3.20	0.72	3.42	1.08	10.76	0.036	1.51	0.59	3.86	0.39
Tumour grade		155	100.0%	1.08	0.49	2.39	0.84	1.31	0.44	3.85	0.63	1.29	0.60	2.76	0.51
Location		155	100.0%	1.91	0.84	4.32	0.12	1.08	0.35	3.33	0.90	0.60	0.25	1.40	0.24
MSI-status[a]		154	99.4%	0.29	0.07	1.23	0.09	0.03	0.00	9.97	0.033	2.08	0.90	4.82	0.09
Modified signature	low risk	79	51.0%	1.00				1.00				1.00			
	high risk	76	49.0%	1.65	1.07	2.56	0.024	3.35	1.20	9.31	0.021	1.14	0.76	1.73	0.52
Multivariate															
Nodal status		155	100.0%	1.08	0.40	2.88	0.88	2.91	0.92	9.19	0.07	1.44	0.56	3.71	0.46
Modified signature	low risk	79	51.0%	1.00				1.00				1.00			
	high risk	76	49.0%	2.72	1.13	6.53	0.025	10.30	1.33	79.99	0.026	1.36	0.59	3.11	0.47

[a] n=1 missing

The significant associations were visualised using Kaplan Meier analysis dividing the expression levels of *let-7i* and *miR-30b* into quartiles (**Figure 2**) showing a split course for Q1-2 (below median) and Q3-4 (above median). The median expression levels were used to assess the prognostic value of the original 2-miRNA signature in our cohort (*Let-7i* high and *miR-10b* low vs. *Let-7i* low and/or *miR-10b* high). The 2-miRNA signature was significantly associated with HFS (5-year survival 100% vs 89.3%, p=0.04)(**Figure 3A**), but did not show a significant difference for MFS (**Figure 3B**) or OS (**Figure 3C**).

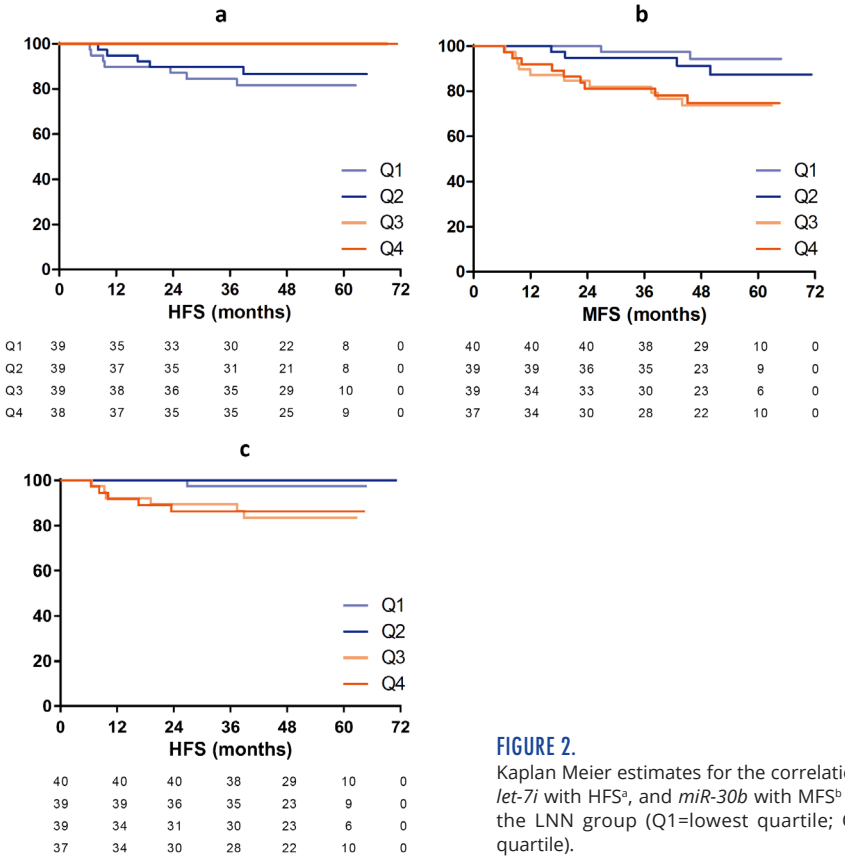


FIGURE 2.
Kaplan Meier estimates for the correlation between *let-7i* with HFS[a], and *miR-30b* with MFS[b] and HFS[c] in the LNN group (Q1=lowest quartile; Q4=highest quartile).

Since *miR-30b* was significantly associated with MFS and HFS (**Figure 2B** and **Figure 2C**, respectively) in our cohort, we explored whether *miR-30b* could contribute to the discriminating value of the signature. In a multivariate Cox regression model including the 2-miRNA signature and *miR-30b* expression split at the median level, *miR-30b* was significantly associated with MFS (HR=3.65, 95%CI=1.44-9.27, p=0.007) and HFS

(HR=10.04, 95%CI=1.30-77.78, p=0.027) independent of the original signature. For both models, the log likelihood significantly increased when adding *miR-30b* expression split at the median level (Δ log likelihood=4.36, p=0.003 and Δ log likelihood=4.43, p=0.003). We therefore added *miR-30b* split at the median expression level to the original metastasis-specific miRNA signature (modified 3-miRNA signature), in which patients were categorized as low risk when having a tumour with two or three of the following: *let-7i* high, *miR-10b* low and *miR-30b* low. All other patients were categorized as high risk. The modified 3-miRNA signature showed a significant difference in MFS (5-year survival 90.2% vs 75.5%, p=0.019) and HFS (5-year survival 98.6% vs 85.6%, p=0.004) (**Figure 3A** and **Figure 3B**, respectively). In the total group, this modified signature also showed a significant difference between patients in the 'low risk' vs 'high risk' group for both MFS (5-year survival 86.8% vs 73.8%, p=0.018) and HFS (5-year survival 97.0% vs 86.2%, p=0.004). Still, the modified 3-miRNA signature did not correlate with OS in these groups, or with any of the clinical endpoints in the LNP group.

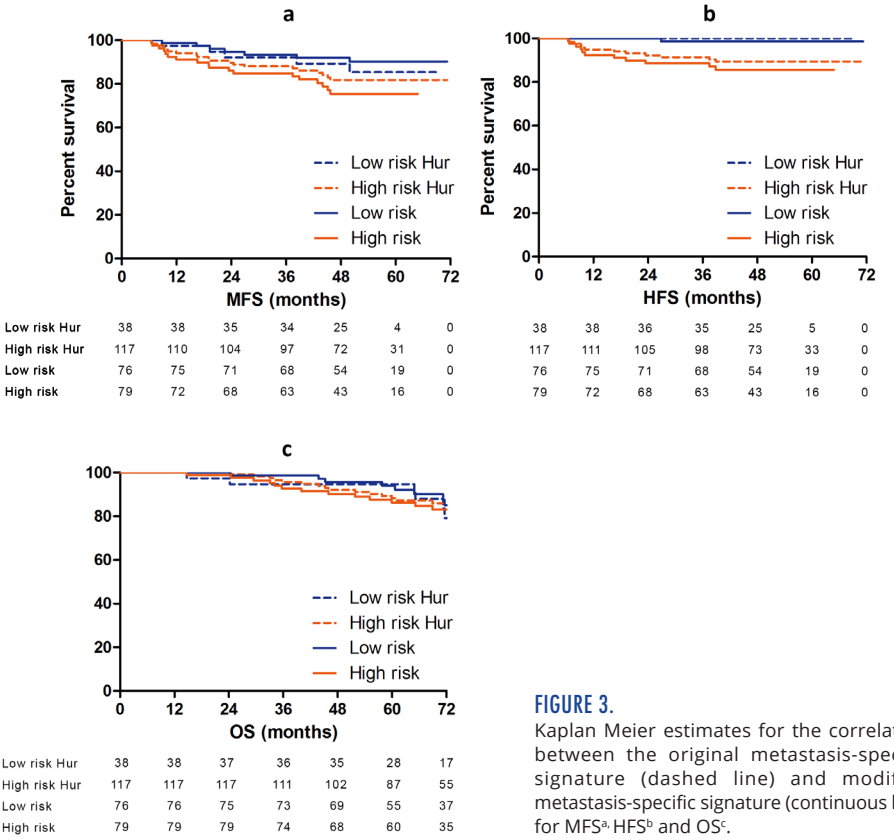


FIGURE 3.
Kaplan Meier estimates for the correlation between the original metastasis-specific signature (dashed line) and modified metastasis-specific signature (continuous line) for MFS[a], HFS[b] and OS[c].

Nodal status was the only traditional clinical factor significantly associated with disease outcome in univariate Cox regression analysis in both the LNN and total group (**Table 2** and **Supplementary Table 3**, respectively). When adjusting for lymph node status in a multivariate Cox regression model (which in the LNN group stratifies in two groups based on the total number of assessed lymph nodes with a cut-off of 10), the modified metastasis-specific 3-miRNA signature was still significantly associated with MFS (HR=2.72, 95%CI=1.13-6.53, p=0.025) and HFS (HR=10.30, 95%CI=1.33-79.99, p=0.026) in the LNN group, and HFS in the total group (HR=5.07, 95%CI=1.46-17.59, p=0.011) (**Table 2** and **Supplementary Table 3**, respectively).

PATHWAY ANALYSIS

Pathway analysis showed that *let-7i* expression was associated with axon guidance, glycosphingolipid and glycosaminoglycan biosynthesis, focal adhesion, extracellular matrix (ECM) receptor interaction and regulation of the actin cytoskeleton which are all related to cell adhesion, migration and motility. Furthermore, we observed an association with the hedgehog, WNT and Transforming Growth Factor (TGF)-β signalling pathways (**Supplementary Table 4** and **5**). Combined use of 3 independent target prediction algorithms did not reveal any overlapping *let-7i* target genes involved in glycosphingolipid biosynthesis, glycosaminoglycan biosynthesis, or the hedgehog signalling pathway. However, the axon guidance, focal adhesion, ECM receptor interaction, regulation of actin cytoskeleton TGF beta signalling and WNT signalling pathways do contain a number of genes predicted to be *let-7i* targets by all 3 algorithms (**Supplementary Table 6**). mRNA Expression of *COL4A6* and *FNDC3A*, involved in focal adhesion and ECM receptor interactions, was significantly negatively correlated to *let-7i* expression in our samples (both in LNN and LNP samples). Expression of *ACVR1C,* part of the TGF-β signalling pathway, showed a borderline significant negative correlation with *let-7i* only in LNP samples (p=0.056).

Expression of *miR-30b* was associated with axon guidance and showed a significant negative correlation with expression of *PPP3R1, NFAT5,* and *SEMA6B* in this pathway (**Supplementary Table 4-6**).

Noteworthy, we observed more significant positive correlations between *let-7i* expression and its predicted targets in all significantly associated pathways, suggesting the effect of *let-7i* on genes in these pathways are mostly indirect.

DISCUSSION

Our study confirmed the clinical significance of measuring *let-7i* and the miRNA-signature as suggested by Hur et al, and have extended these findings to a well-defined independent cohort of patients with colon cancer. We showed that the expression of most of these miRNAs was different for MSI and MSS tumours, and for left and right-sided tumours. Furthermore, let-7i was expressed at a lower level in stage II compared to stage I colon cancers. We validated *let-7i* as a prognostic marker, but could not confirm *miR-10b* as a prognostic factor. In contrast, *miR-30b* was prognostic with regard to MFS and HFS. The original metastasis-specific 2-miRNA signature was significantly associated with HFS but not MFS. We therefore propose a modified metastasis-specific 3-miRNA signature combining *miR-10b, miR-30b* and *let-7i* which identified a group with low and high risk in terms of MFS and HFS. Subsequent pathway analysis suggested an association between *let-7i* expression and cell adhesion, migration and motility, and the hedgehog, WNT and Transforming Growth Factor (TGF)-β signalling pathways. *miR-30b* expression was associated to axon guidance.

The findings with regard to the differential expression of the miRNAs in MSI versus MSS and left-sided versus right-sided tumours suggest a different role for these miRNAs in hyper and non-hypermutated tumours, and add to the understanding of these biologically different entities. Although these characteristics were not included in the differential miRNA expression analysis by Hur et al., the findings are in line with accumulating evidence that supports the role of miRNAs in the pathogenesis of MSI tumours, such as the involvement *miR-21* and *miR-155* in the regulation of mismatch repair gene and protein expression.[20] Similarly, left- and right-sided tumours are reported to express miRNAs at different levels.[21,22] Our findings add to the fast expanding knowledge on the different roles of miRNAs in these biologically different entities. Furthermore, we found that *let-7i* expression was lower in stage II compared to stage I tumours, which is line with the findings by Hur et al. and the tumour suppressing role of the *let-7* family as described in literature.[23]

Furthermore, low expression of *let-7i* was associated with poor HFS in our cohort of patients with lymph node negative colon cancer who did not receive adjuvant chemotherapy which confirmed the findings of Hur et al. We also found that high expression of *miR-30b* was associated with poor MFS and HFS, which was not observed by Hur et al., since *miR-30b* was not validated in the comparative analysis of primary CRC and matched liver metastases (LM) tissues used for the identification of miRNAs that may hold prognostic potential in primary tumours. In our cohort, all patients had colon cancer which may explain the different observations in our cohort.

In contrast to Hur et al., we did not find a significant association or even trend between expression of *miR-10b* and any of the long term clinical endpoints. Interestingly, Hur et al. showed that *miR-10b* was downregulated in LM tissue compared to primary CRC tissue while it was upregulated in primary CRC versus normal tissue, and high expression of *miR-10b* was associated with poor MFS. This suggests that differential expression between primary and metastatic lesions does not always reflect prognostic and/or predictive value in primary tumour lesions.[24-29] This may also explain the discrepancy for *miR-30b* between absence of differential expression in their comparative analysis and the association with MFS/HFS in our cohort, although further studies should be conducted to confirm the latter.

Hur et al. also showed that their signature (consisting of *miR-10b* and *let-7i*) was associated with MFS in one cohort of primary colorectal cancers. Although this signature did segregate our cohort of patients with LNN colon cancer in two groups with significantly different outcome in terms of HFS, we did not find a significant difference in terms of MFS. Incorporation of *miR-30b* (modified signature) did identify a low and high risk group were clearly distinctive with regard to MFS and HFS. These results remained significant when correcting for nodal status. Next to nodal status, MSI was significantly associated with HFS in the LNN group. Still, when correcting for MSI alone or together with nodal status in the LNN group, the modified metastasis-specific signature remained significantly associated with MFS and HFS (data not shown). However, due to the low number of events, these multivariable analyses should be interpreted with caution. In conclusion, our results underline the prognostic value of *let-7i* and provide support for the prognostic value of the modified miRNA signature.

The different observations may be explained by differences in the patient cohorts used in the study by Hur et al and the current study. The cohorts used by Hur et al were heterogeneous with regard to tumour stage and tumour location (both colon and rectum were included), and no information on pre and postoperative local and systemic treatment was provided.[9] This heterogeneity may for instance explain the failure of validating miR-30b in the comparative analysis since expression levels of *miR-30b* are different between colon and rectal cancers.[22] In the current study, a well-defined prospectively included cohort of patients with colon cancer was used. Another explanation for the observed differences may be the cut-off used to dichotomize expression levels. Hur et al. used the Youden's index (i.e. the most optimal combination of sensitivity and specificity) to determine the cut-off value for which the miRNAs had the maximum potential effectiveness based on a specific outcome parameter.[19] Inherent to the dependency of the cut-off to the outcome parameter, patients may change from the low risk to the high risk group depending

on the outcome of interest. In our cohort, the cut-off value based on the Youden's index was nearly identical to the median expression level, which is independent of the outcome parameter of interest and per definition gives two equally sized subgroups (data not shown). In their cohort, the Youden's index stratified the patients in two unequally sized groups (4.8% low *let-7i* expression and 21% low *miR-10b* expression) meaning the cut-off value based on the Youden's index was quite different from the median expression level. Therefore, future studies may explore the median expression level to assign patients to a low or high risk group. Lastly, differences in the used material (formalin-fixed paraffin embedded versus fresh frozen), RNA isolation and/or the assays used to measure the miRNA expression may also account for part of the observed differences.

The pathway analysis for *let-7i* and *miR-30b* revealed several interesting associations. *Let-7i* expressionwas associated with several pathways related to cell adhesion, migration and motility. This is in line with literature that associates the expression of this miRNA with the expression of genes involved in these pathways that are to known to be altered in carcinogenesis and involved in progression.[30,31] Activation of either LIN28A or LIN28B is thought to responsible for global post-transcriptional downregulation of the *let-7* family in cancers.[32] Furthermore, *let-7i* expression was associated with the among others hedgehog, WNT and Transforming Growth Factor (TGF)-β signalling pathways. Interestingly, the *let-7* family was shown to be involved in hedgehog-mediated drug resistance in lung cancer which supports the observed association in our cohort.[33] Similarly, the *let-7* family was shown to be involved in both the Wnt signalling in breast cancer which again involved LIN28.[34] Lastly, high expression of *let-7i* was associated with increased TGF-β signalling. TGF-β is known to play a major role in the tumorigenesis of at least half of all CRCs in which inactivating mutations abolish the tumour suppressing effect of TGF-β signalling pathway. Furthermore, decreased SMAD4 expression, a downstream target of TGF-β, is associated with poor prognosis in colon cancer providing indirect evidence that inactivation of TGF-β signalling leads to invasive behaviour of colon cancer.[35] Interestingly, TGF-β appears to convert from a tumour suppressor to a tumour promotor in more advanced stages of cancers, which is known as the TGF-β paradox.[6] A few studies have specifically addressed the association between let-7 and TGF-β expression showing an inverse correlation.[37,38] However, these studies were performed in cell lines and a melanoma xenograft model, which may preclude extrapolating these results to CRC in a clinical setting. In our study, 6% of the genes in the 'TGF-β signalling pathway' were considered to be a potential target of *let-7i* using three prediction algorithms. The positive correlation between tumour suppressor *let-7i* and genes involved in the TGF-β signalling pathway in this study

suggests that TGF-β has yet to go through this conversion, and both TGF-β and let-7i act as a tumour suppressor in our homogeneous prospectively collected cohort of early stage colon cancers. Interestingly, expression of *ACVR1C* was negatively associated with *let-7i* expression in only the LNP group, suggesting these tumours may have gone through the conversion. The association with the above-mentioned pathways may also suggest that *let-7i* is linked to stromal content and has an indirect role in TGF-β signalling in stromal cells, which is known to play a role in metastasis initiation.[39]

The *miR-30* family has been extensively studied in the field of cancer research. *In vitro*, *miR-30* family members have been associated with several aspects of tumorigenesis such as cell migration, cell growth, cell invasiveness and apoptosis.[40-48] Interestingly, both negative and positive associations with tumorigenesis have been described for *miR-30* family members including *miR-30b* which precludes definitive conclusions. Contrasting associations have also been reported between the *miR-30* family and tumour characteristics such as tumour stage and tumour grade.[40,46,47,49-52] In terms of clinical outcome defined as MFS and/or OS, *miR-30* family members have been described as markers of poor outcome in melanoma[49], ovarian cancer[41], prostate cancer[44] and oesophageal cancer[53], and as a tumour suppressive miRNA in breast cancer[51], lung cancer[54], CRC[42], prostate cancer[40] and ovarian cancer.[55] In our study, high *miR-30b* was associated with poor MFS and HFS. The pathway analysis mainly revealed a positive association between *miR-30b* expression and axon guidance. The genes traditionally described for their roles in axon guidance are important regulators of neuronal migration and positioning during embryonic development. However, they have been implicated in cancer cell survival, growth, invasion and angiogenesis.[56,57] Very little is known about the direct association between *miR-30b* and axon guidance in current literature. However, the target gene prediction we performed subsequently to the pathway analysis showed that >11% of the genes listed in the 'axon guidance' pathway in KEGG were predicted to be direct targets of *miR-30b*. In support of this direct interaction, significant negative correlations between expression of *miR-30b* and axon-guidance genes *PPP3R1, NFAT5, SEMA6B* was observed in our cohort. Further research may be directed to investigate the possible role of *miR-30b* in axon guidance. These insights combined with the fact that miR-30b was upregulated in patients with a poor prognosis in our cohort, may provide a rationale to investigate the miRNA as a therapeutic target.

Although overlapping predicted targets were found for the majority of *let-7i* and *miR-30b* associated pathways in 3 prediction algorithms, we observed many significantly positive correlations between particularly *let-7i* and its predicted targets in our cohort. This suggests that effects of *let-7i* on the associated pathways is

(partially) indirect , although functional studies are needed to confirm this. All in all, the observed associations suggest that *let-7i* and *miR-30b* may be a relevant factor for cancer cells in their ability to move potentially also involving their stromal component together increasing their metastatic potential.

Our data suggest that *let-7i* and *miR-30b,* and a 3-miRNA signature hold prognostic value in lymph node negative colon cancers, although independent validation in a large cohort is needed. Further studies should ideally include an analysis of circulating serum levels of these miRNAs.

REFERENCES

1 Ferlay, J. *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer*. **49**, 1374-1403 (2013).

2 Sargent, D. J. *et al.* End points for colon cancer adjuvant trials: observations and recommendations based on individual patient data from 20,898 patients enrolled onto 18 randomized trials from the ACCENT Group. *J Clin Oncol*. **25**, 4569-4574 (2007).

3 Elferink, M. A., de Jong, K. P., Klaase, J. M., Siemerink, E. J. & de Wilt, J. H. Metachronous metastases from colorectal cancer: a population-based study in North-East Netherlands. *Int J Colorectal Dis*. **30**, 205-212 (2015).

4 Lochhead, P. *et al.* Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication. *J Natl Cancer Inst*. **105**, 1151-1156 (2013).

5 Roth, A. D. *et al.* Integrated analysis of molecular and clinical prognostic factors in stage II/III colon cancer. *J Natl Cancer Inst*. **104**, 1635-1646 (2012).

6 Popat, S., Hubner, R. & Houlston, R. S. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. **23**, 609-618 (2005).

7 Lin, S. & Gregory, R. I. MicroRNA biogenesis pathways in cancer. *Nat Rev Cancer*. **15**, 321-333 (2015).

8 Chi, Y. & Zhou, D. MicroRNAs in colorectal carcinoma--from pathogenesis to therapy. *J Exp Clin Cancer Res*. **35**, 43 (2016).

9 Hur, K. *et al.* Identification of a metastasis-specific MicroRNA signature in human colorectal cancer. *J Natl Cancer Inst*. **107** (2015).

10 McShane, L. M. *et al.* REporting recommendations for tumour MARKer prognostic studies (REMARK). *Eur J Cancer*. **41**, 1690-1696 (2005).

11 Kloosterman, W. P. *et al.* A Systematic Analysis of Oncogenic Gene Fusions in Primary Colon Cancer. *Cancer Res*, 3814-3822 (2017).

12 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15-21 (2013).

13 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11**, R25 (2010).

14 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139-140 (2010).

15 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28**, 27-30 (2000).

16 Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. **120**, 15-20 (2005).

17 Wong, N. & Wang, X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*. **43**, D146-152 (2015).

18 Loher, P. & Rigoutsos, I. Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics*. **28**, 3322-3323 (2012).

19 Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. & Schisterman, E. F. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*. **50**, 419-430 (2008).

20 Yamamoto, H. & Imai, K. Microsatellite instability: an update. *Arch Toxicol*. **89**, 899-921 (2015).

21 Schee, K. *et al.* Deep Sequencing the MicroRNA Transcriptome in Colorectal Cancer. *PLoS One*. **8**, e66165 (2013).

22 Slattery, M. L. *et al.* MicroRNAs and colon and rectal cancer: differential expression by tumor location and subtype. *Genes Chromosomes Cancer*. **50**, 196-206 (2011).

23 Boyerinas, B., Park, S. M., Hau, A., Murmann, A. E. & Peter, M. E. The role of let-7 in cell differentiation and cancer. *Endocr Relat Cancer*. **17**, F19-36 (2010).

24 Nishida, N. *et al.* MicroRNA-10b is a prognostic indicator in colorectal cancer and confers resistance to the chemotherapeutic agent 5-fluorouracil in colorectal cancer cells. *Ann Surg Oncol*. **19**, 3065-3071 (2012).

25 Pu, X. X. *et al.* Circulating miR-221 directly amplified from plasma is a potential diagnostic and prognostic marker of colorectal cancer and is correlated with p53 expression. *J Gastroenterol Hepatol*. **25**, 1674-1680 (2010).

26 Baffa, R. *et al.* MicroRNA expression profiling of human metastatic cancers identifies cancer gene targets. *J Pathol*. **219**, 214-221 (2009).

27 Lanza, G. *et al.* mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. *Mol Cancer*. **6**, 54 (2007).

28 Schimanski, C. C. *et al.* High miR-196a levels promote the oncogenic phenotype of colorectal cancer cells. *World J Gastroenterol*. **15**, 2089-2096 (2009).

29 Bitarte, N. *et al.* MicroRNA-451 is involved in the self-renewal, tumorigenicity, and chemoresistance of colorectal cancer stem cells. *Stem Cells*. **29**, 1661-1671 (2011).

30 Paschos, K. A., Canovas, D. & Bird, N. C. The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. *Cell Signal*. **21**, 665-674 (2009).

31 Ono, M. & Hakomori, S. Glycosylation defining cancer cell motility and invasiveness. *Glycoconj J*. **20**, 71-78 (2004).

32 Balzeau, J., Menezes, M. R., Cao, S. & Hagan, J. P. The LIN28/let-7 Pathway in Cancer. *Front Genet*. **8**, 31 (2017).

33 Ahmad, A. *et al.* Inhibition of Hedgehog signaling sensitizes NSCLC cells to standard therapies through modulation of EMT-regulating miRNAs. *J Hematol Oncol*. **6**, 77 (2013).

34 Cai, W. Y. *et al.* The Wnt-beta-catenin pathway represses let-7 microRNA expression through transactivation of Lin28 to augment breast cancer stem cell expansion. *J Cell Sci*. **126**, 2877-2889 (2013).

35 Alazzouzi, H. *et al.* SMAD4 as a prognostic marker in colorectal cancer. *Clin Cancer Res*. **11**, 2606-2611 (2005).

36 Morrison, C. D., Parvani, J. G. & Schiemann, W. P. The relevance of the TGF-beta Paradox to EMT-MET programs. *Cancer Lett*. **341**, 30-40 (2013).

37 Dangi-Garimella, S., Strouch, M. J., Grippo, P. J., Bentrem, D. J. & Munshi, H. G. Collagen regulation of let-7 in pancreatic cancer involves TGF-beta1-mediated membrane type 1-matrix metalloproteinase expression. *Oncogene*. **30**, 1002-1008 (2011).

38 Zhang, Z. *et al.* Lin28B promotes melanoma growth by mediating a microRNA regulatory circuit. *Carcinogenesis*. **36**, 937-945 (2015).

39 Calon, A. *et al.* Dependency of colorectal cancer on a TGF-beta-driven program in stromal cells for metastasis initiation. *Cancer Cell*. **22**, 571-584 (2012).

40 Ling, X. H. *et al.* MicroRNA-30c serves as an independent biochemical recurrence predictor and potential tumor suppressor for prostate cancer. *Mol Biol Rep*. **41**, 2779-2788 (2014).

41 Li, N. *et al.* A combined array-based comparative genomic hybridization and functional library screening approach identifies mir-30d as an oncomir in cancer. *Cancer Res*. **72**, 154-164 (2012).

42 Moreno-Mateos, M. A. *et al.* Novel small RNA expression libraries uncover hsa-miR-30b and hsa-miR-30c as important factors in anoikis resistance. *RNA*. **19**, 1711-1725 (2013).

43 Jiang, L. *et al.* MicroRNA-30e* promotes human glioma cell invasiveness in an orthotopic xenotransplantation model by disrupting the NF-kappaB/IkappaBalpha negative feedback loop. *J Clin Invest*. **122**, 33-47 (2012).

44 Kobayashi, N. *et al.* Identification of miR-30d as a novel prognostic maker of prostate cancer. *Oncotarget*. **3**, 1455-1471 (2012).

45 Liu, M. *et al.* Heterochromatin protein HP1gamma promotes colorectal cancer progression and is regulated by miR-30a. *Cancer Res*. **75**, 4593-4604 (2015).

46 Wang, W. *et al.* MicroRNA-30a-3p inhibits tumor proliferation, invasiveness and metastasis and is downregulated in hepatocellular carcinoma. *Eur J Surg Oncol*. **40**, 1586-1594 (2014).

47 Yao, J. *et al.* MicroRNA-30d promotes tumor invasion and metastasis by targeting Galphai2 in hepatocellular carcinoma. *Hepatology*. **51**, 846-856 (2010).

48 Liu, X. *et al.* miR-30c regulates proliferation, apoptosis and differentiation via the Shh signaling pathway in P19 cells. *Exp Mol Med*. **48**, e248 (2016).

49 Gaziel-Sovran, A. *et al.* miR-30b/30d regulation of GalNAc transferases enhances invasion and immunosuppression during metastasis. *Cancer Cell*. **20**, 104-118 (2011).

50 Liao, W. T. *et al.* MicroRNA-30b functions as a tumour suppressor in human colorectal cancer by targeting KRAS, PIK3CD and BCL2. *J Pathol*. **232**, 415-427 (2014).

51 Cheng, C. W. *et al.* MicroRNA-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer. *Breast Cancer Res Treat*. **134**, 1081-1093 (2012).

52 Yu, F. *et al.* Mir-30 reduction maintains self-renewal and inhibits apoptosis in breast tumor-initiating cells. *Oncogene*. **29**, 4194-4204 (2010).

53 Hu, Y. *et al.* Prognostic significance of differentially expressed miRNAs in esophageal cancer. *Int J Cancer*. **128**, 132-143 (2011).

54 Suh, S. S. *et al.* FHIT suppresses epithelial-mesenchymal transition (EMT) and metastasis in lung cancer through modulation of microRNAs. *PLoS Genet*. **10**, e1004652 (2014).

55 Wang, Y. *et al.* The expression of miR-30a* and miR-30e* is associated with a dualistic model for grading ovarian papillary serious carcinoma. *Int J Oncol*. **44**, 1904-1914 (2014).

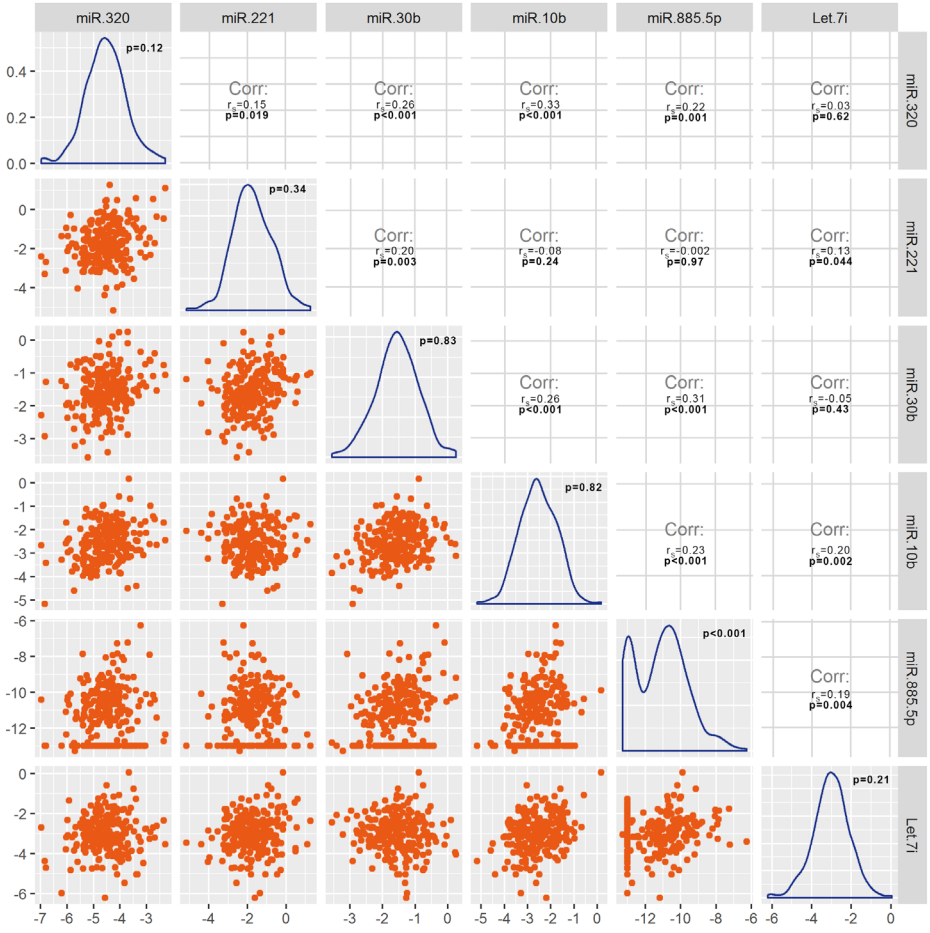
56 Chedotal, A., Kerjan, G. & Moreau-Fauvarque, C. The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death Differ*. **12**, 1044-1056 (2005).

57 Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*. **491**, 399-405 (2012).

7

SUPPLEMENTARY DATA

SUPPLEMENTARY FIGURES



SUPPLEMENTARY FIGURE 1. THIS FIGURE DISPLAYS THE CORRELATION BETWEEN THE MIRNAS (SCATTER PLOTS) WITH THEIR RESPECTIVE SPEARMAN'S CORRELATION COEFFICIENT AND P VALUES

The distribution of the expression levels of the six miRNAs are displayed on the diagonal with their respective p values (Shapiro-Wilk test).

SUPPLEMENTARY TABLES

SUPPLEMENTARY TABLE 1. GENE ASSAYS USED TO MEASURE MRNA EXPRESSION OF THE SIX MIRNAS AND 3 REFERENCE GENES

INDEX	GENE SYMBOL	GENE NAME	QPCR DETECTION METHOD	ASSAY ID THERMOFISHER SCIENTIFIC
Candidate miR	*MIR320*	microRNA 320	TaqMan® MicroRNA Assay	384
Candidate miR	*MIR221*	microRNA 221	TaqMan® MicroRNA Assay	2277
Candidate miR	*MIR30B*	microRNA 30b	TaqMan® MicroRNA Assay	2218
Candidate miR	*MIR10B*	microRNA 10b	TaqMan® MicroRNA Assay	524
Candidate miR	*MIR885*	microRNA 885	TaqMan® MicroRNA Assay	602
Candidate miR	*MIRLET7I*	microRNA let-7i	TaqMan® MicroRNA Assay	2296
Reference miR	*MIR16-1*	microRNA 16-1	TaqMan® MicroRNA Assay	391
Reference miR	*RNU6B*	RNA, U6 small nuclear 6, pseudogene	TaqMan® MicroRNA Assay	1093
Reference miR	*SNORD44*	small nucleolar RNA, C/D box 44	TaqMan® MicroRNA Assay	1094

SUPPLEMENTARY TABLE 2A. CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS OF THE TOTAL GROUP

			MIR-320		*MIR-221*		*MIR-30B*	
	n	**%**	**median (IQR)**	**P value**	**median (IQR)**	**P value**	**median (IQR)**	**P value**
Gender								
Female	104	44.8%	-4.52 (-5.07 • -4.10)	0.82	-1.84 (-2.49 • -1.05)	0.64	-1.67 (-2.12 • -1.19)	0.008
Male	128	55.2%	-4.56 (-4.94 • -4.03)		-1.77 (-2.35 • -1.08)		-1.47 (-1.92 • -1.03)	
Age	232	100.0%	-0.04	0.52	-0.09	0.20	-0.07	0.26
Tumor stage								
Stage I	57	24.6%	-4.61 (-5.01 • -4.03)	0.62	-1.78 (-2.45 • -0.88)	0.92	-1.55 (-1.97 • -1.07)	0.002
Stage II	98	42.2%	-4.49 (-4.99 • -4.08)		-1.77 (-2.51 • -1.16)		-1.75 (-2.17 • -1.35)	
Stage III	77	33.2%	-4.61 (-5.08 • -4.11)		-1.82 (-2.33 • -1.01)		-1.33 (-1.62 • -1.01)	
T status								
T2	68	29.3%	-4.58 (-4.93 • -4.03)	0.61	-1.78 (-2.39 • -0.99)	0.52	-1.48 (-1.92 • -0.99)	0.33
T3	164	70.7%	-4.51 (-5.05 • -4.10)		-1.80 (-2.37 • -1.13)		-1.59 (-1.97 • -1.14)	
Nodal status								
N0	127	54.7%	-4.47 (-4.95 • -4.01)	0.38	-1.77 (-2.51 • -1.13)	0.48	-1.75 (-2.13 • -1.32)	<0.001
N0 <10 nodes	28	12.1%	-4.72 (-5.14 • -4.13)		-1.71 (-2.46 • -0.67)		-1.40 (-1.99 • -0.91)	
N1	52	22.4%	-4.69 (-5.06 • -4.16)		-1.89 (-2.32 • -1.15)		-1.32 (-1.64 • -0.92)	
N2	25	10.8%	-4.41 (-5.19 • -3.83)		-1.42 (-2.35 • -0.52)		-1.38 (1.53 • -1.01)	
Tumor grade								
Good	20	8.6%	-4.61 (-5.12 • -4.11)	0.41	-1.61 (-2.24 • -1.12)	0.72	-1.58 (-1.98 • -1.22)	0.056
Moderate	184	79.3%	-4.55 (-5.05 • -4.04)		-1.81 (-2.40 • -1.06)		-1.58 (-1.99 • -1.16)	
Poor	20	8.6%	-4.38 (-4.84 • -4.14)		-1.58 (-2.50 • -0.83)		-1.24 (-1.71 • -0.92)	
Other	8	3.4%	-4.36 (-5.14 • -3.68)		-2.21 (-2.35 • -1.85)		-1.49 (-1.61 • -0.80)	
Location								
Right	115	49.6%	-4.41 (-4.84 • -4.03)	0.036	-2.11 (-2.56 • -1.17)	0.006	-1.62 (-2.01 • -1.23)	0.016
Left	117	50.4%	-4.67 (-5.15 • -4.11)		-1.60 (-2.23 • -0.97)		-1.41 (-1.91 • -1.04)	
MSI-status[a]								
MSI	46	19.8%	-4.38 (-4.86 • -3.90)	0.18	-2.45 (-3.05 • -2.10)	<0.001	-1.88 (-2.14 • -1.52)	<0.001
MSS	185	79.7%	-4.55 (-5.08 • -4.09)		-1.65 (-2.24 • -0.96)		-1.48 (-1.91 • -1.06)	

[a] n=1 missing

MIR-10B		*MIR-885-5P*		*LET-7I*	
median (IQR)	**P value**	**median (IQR)**	**P value**	**median (IQR)**	**P value**
-2.49 (-3.07 • -1.90)	0.26	-11.10 (-13.00 • -10.16)	0.30	-2.88 (-3.44 • -2.37)	0.14
-2.64 (-3.15 • -2.03)		-10.93 (-12.29 • -10.20)		-3.10 (-3.71 • -2.50)	
.01	0.94	-0.03	0.70	0.03	0.61
-2.62 (-3.29 • -1.90)	0.27	-11.21 (-12.88 • -10.09)	0.041	-2.59 (-3.55 • -2.09)	0.47
-2.66 (-3.13 • -2.02)		-11.39 (-13.00 • -10.52)		-3.15 (-3.61 • -2.66)	
-2.45 (-2.97 • -1.98)		-10.52 (-11.46 • -9.86)			
-2.58 (-3.17 • -1.89)	0.93	-10.79 (-12.58 • -9.99)	0.20	-2.65 (-3.54 • -2.12)	0.054
-2.59 (-3.08 • -2.01)		-11.11 (-13.00 • -10.26)		-3.09 (-3.56 • -2.62)	
-2.61 (-3.11 • -1.91)	0.54	-11.34 (-13.00 • -10.48)	0.002	-3.05 (-3.54 • -2.38)	0.87
-2.73 (-3.37 • -2.06)		-11.28 (-12.94 • -9.50)		-3.12 (-3.96 • -2.36)	
-2.46 (-3.13 • -2.03)		-10.62 (-11.49 • -9.89)		-2.86 (-3.39 • -2.46)	
-2.36 (-2.84 • -1.95)		-10.41 (-12.12 • -9.66)		-3.08 (-3.73 • -2.65)	
-2.73 (-3.13 • -2.31)	0.56	-11.36 (-11.85 • -10.26)	0.14	-3.08 (-3.72 • -2.68)	0.34
-2.58 (-3.11 • -1.95)		-11.11 (-13.00 • -10.24)		-3.06 (-3.56 • -2.44)	
-2.84 (-3.28 • -2.28)		-10.51 (-12.52 • -9.94)		-2.95 (-3.72 • -2.44)	
-1.79 (-2.55 • -1.17)		-9.98 (-10.94 • -9.19)		-2.86 (-3.04 • -2.24)	
-2.25 (-2.72 • -1.71)	<0.001	-11.10 (-13.00 • -10.10)	0.98	-2.99 (-3.55 • -2.38)	0.48
-2.88 (-3.39 • -2.34)		-11.01 (-12.62 • -10.21)		-3.09 (-3.59 • -2.50)	
-2.12 (-2.59 • -1.64)	<0.001	-11.65 (-13.00 • -10.53)	0.001	-2.79 (-3.26 • -2.38)	0.09
-2.65 (-3.19 • -2.13)		-10.92 (-12.15 • -10.10)		-3.08 (-3.65 • -2.46)	

SUPPLEMENTARY TABLE 2B. CLINICAL AND HISTOPATHOLOGICAL CHARACTERISTICS OF THE LYMPH NODE POSITIVE GROUP

			MIR-320		*MIR-221*		*MIR-30B*	
	n	**%**	**median (IQR)**	**P value**	**median (IQR)**	**P value**	**median (IQR)**	**P value**
Gender								
Female	31	40.3%	-4.73 (-5.18 • -4.11)	0.61	-2.10 (-2.37 • -0.79)	0.84	-1.47 (-1.77 • -1.06)	0.22
Male	46	59.7%	-4.59 (-4.93 • -4.11)		-1.81 (-2.25 • -1.19)		-1.32 (-1.57 • -0.94)	
Age	77	100%	-0.06	0.60	0.01	0.91	-0.09	0.43
Tumor stage								
Stage I	0	0.0%	-	-	-		-	
Stage II	0	0.0%	-	-	-		-	
Stage III	77	100%	-4.61 (-5.08 • -4.11)		-1.82 (-2.33 • -1.01)		-1.33 (-1.62 • -1.01)	
T status								
T2	11	14.3%	-4.52 (-4.83 • -3.90)	0.38	-1.71 (-2.20 • -1.02)	0.44	-1.20 (-1.37 • -0.74)	0.128
T3	66	85.7%	-4.65 (-5.09 • -4.13)		-1.84 (-2.36 • -0.99)		-1.38 (-1.66 • -1.02)	
Nodal status								
N0	0	0.0%	-	-	-		-	
N0 <10 nodes	0	0.0%	-	-	-		-	
N1	52	67.5%	-4.69 (-5.06 • -4.16)	0.38	-1.89 (-2.32 • -1.15)	0.44	-1.32 (-1.64 • -0.92)	0.13
N2	25	32.5%	-4.41 (-5.19 • -3.83)		-1.42 (-2.35 • -0.52)		-1.38 (1.53 • -1.01)	
Tumor grade								
Good	7	9.1%	-4.88 (-5.44 • -4.25)	0.11	-2.20 (-2.34 • -1.36)	0.57	-1.79 (-2.09 • -1.17)	0.11
Moderate	54	70.1%	-4.69 (-5.07 • -4.09)		-1.80 (-2.36 • -0.99)		-1.32 (-1.60 • -0.99)	
Poor	11	14.3%	-4.28 (-4.61 • -4.11)		-1.11 (-1.82 • -0.78)		-1.09 (-1.38 • -0.78)	
Other	5	6.5%	-4.39 (-5.33 • -3.59)		-2.32 (-2.57 • -2.18)		-1.60 (-1.68 • -1.05)	
Location								
Right	36	46.8%	-4.31 (-4.80 • -3.80)	0.00	-1.94 (-2.36 • -0.90)	0.96	-1.47 (-1.65 • -1.03)	0.36
Left	41	53.2%	-4.82 (-5.24 • -4.45)		-1.80 (-2.27 • -1.27)		-1.27 (-1.61 • -0.91)	
MSI-status								
MSI	12	15.6%	-4.31 (-5.23 • -3.55)	0.33	-2.37 (-2.86 • -1.68)	0.015	-1.58 (-1.89 • -1.47)	0.009
MSS	65	84.4%	-4.65 (-5.07 • -4.12)		-1.77 (-2.21 • -0.98)		-1.20 (-1.60 • -0.94)	

MIR-10B		*MIR-885-5P*		*LET-7I*	
median (IQR)	**P value**	**median (IQR)**	**P value**	**median (IQR)**	**P value**
-2.46 (-2.97 • -2.13)	0.57	-11.03 (-13.00 • -10.06)	0.09	-2.80 (-3.39 • -2.56)	0.57
-2.39 (-2.98 • -1.92)		-10.47 (-11.13 • -9.62)		-3.06 (-3.57 • -2.51)	
-0.15	0.19	-0.02	0.87	-0.05	0.70
-		-		-	
-		-		-	
-2.45 (-2.97 • -1.98)		-10.52 (-11.46 • -9.86)		-2.99 (-3.52 • -2.52)	
-2.41 (-2.97 • -1.66)	0.71	-10.23 (-10.52 • -9.47)	0.10	-3.04 (-3.20 • -2.45)	0.81
-2.46 (-2.98 • -1.98)		-10.65 (-11.64 • -9.90)		-2.97 (-3.53 • -2.53)	
-		-		-	
-		-		-	
-2.46 (-3.13 • -2.03)	0.71	-10.62 (-11.49 • -9.89)	0.10	-2.86 (-3.39 • -2.46)	0.81
-2.36 (-2.84 • -1.95)		-10.41 (-12.12 • -9.66)		-3.08 (-3.73 • -2.65)	
-2.72 (-2.96 • -2.47)	0.52	-10.26 (-11.52 • -9.82)	0.61	-3.08 (-3.72 • -2.80)	0.50
-2.35 (-2.94 • -1.89)		-10.74 (-11.79 • -10.01)		-2.92 (-3.53 • -2.49)	
-2.78 (-3.29 • -2.17)		-10.30 (-10.52 • -9.73)		-3.11 (-3.80 • -2.38)	
-1.99 (-2.81 • -1.06)		-10.63 (-12.02 • -9.39)		-2.93 (-3.22 • -2.47)	
-2.06 (-2.53 • -1.57)	<0.001	-10.54 (-11.39 • -9.62)	0.36	-2.78 (-3.11 • -2.39)	0.028
-2.74 (-3.31 • -2.34)		-10.52 (-11.67 • -10.11)		-3.12 (-3.68 • -2.68)	
-2.06 (-2.52 • -1.82)	0.049	-11.45 (-13.00 • -9.79)	0.11	-2.71 (-2.87 • -2.20)	0.018
-2.47 (-3.06 • -2.10)		-10.45 (-11.21 • -9.86)		-3.08 (-3.62 • -2.55)	

SUPPLEMENTARY TABLE 3. UNIVARIATE AND MULTIVARIATE COX REGRESSION ANALYSIS FOR THE TOTAL GROUP

			UNIVARIATE					
			MFS (events=44)		HFS (events=12)		OS (events=23)	
	n	%	HR (95%CI)	P value	HR (95%CI)	P value	HR (95%CI)	P value
mRNA expression								
MiR-320	232	100%	0.88 (0.60 • 1.30)	0.53	0.95 (0.53 • 1.71)	0.87	0.96 (0.64 • 1.45)	0.86
MiR-221	232	100%	1.36 (1.01 • 1.82)	0.041	1.24 (0.79 • 1.94)	0.35	1.02 (0.74 • 1.41)	0.89
MiR-30b	232	100%	1.93 (1.24 • 3.01)	0.004	2.37 (1.20 • 4.66)	0.013	1.35 (0.84 • 2.17)	0.22
MiR-10b	232	100%	1.03 (0.71 • 1.49)	0.87	0.65 (0.37 • 1.13)	0.13	1.02 (0.70 • 1.50)	0.92
MiR-885-5p	232	100%	0.98 (0.80 • 1.20)	0.85	0.89 (0.64 • 1.22)	0.45	0.94 (0.76 • 1.16)	0.57
Let-7i	232	100%	0.86 (0.62 • 1.19)	0.36	0.42 (0.27 • 0.64)	<0.001	0.92 (0.66 • 1.27)	0.60
Gender								
Female	73	47.1%	1		1		1	
Male	82	52.9%	1.19 (0.65 • 2.18)	0.57	1.14 (0.46 • 2.82)	0.78	2.23 (1.15 • 4.59)	0.018
Age	232	100%	1.0002 (0.96 • 1.04)	0.99	1.003 (0.95 • 1.06)	0.92	1.06 (1.02 • 1.11)	0.003
Tumor stage								
Stage I	57	36.8%	1		1		1	
Stage II	98	63.2%	1.51 (.63 • 3.61)	0.36	1.17 (0.35 • 3.90)	0.79	1.08 (0.46 • 2.54)	0.87
Stage III			2.18 (0.92 • 5.18)	0.08	1.40 (0.41 • 4.79)	0.59	1.80 (0.78 • 4.15)	0.17
T status								
T2	57	36.8%	1		1		1	
T3	98	63.2%	1.97 (0.92 • 4.24)	0.08	1.22 (0.44 • 3.38)	0.71	1.87 (0.86 • 4.07)	0.11
Nodal status								
N0	127	81.9%	1		1		1	
N0 <10 nodes	28	18.1%	1.20 (0.45 • 3.20)	0.72	3.41 (1.08 • 10.75)	0.036	1.53 (0.6 • 3.88)	0.38
N1			0.99 (0.44 • 2.26)	0.99	1.79 (0.57 • 5.65)	0.32	1.01 (0.42 • 2.45)	0.98
N2			3.57 (1.71 • 7.47)	0.001	1.78 (0.37 • 8.59)	0.47	4.20 (1.96 • 9.00)	<0.001
Tumor grade								
Good	13	8.4%	1		1		1	
Moderate	130	83.9%	0.85 (0.30 • 2.39)	0.85	0.74 (0.17 • 3.26)	0.69	1.68 (0.40 • 7.04)	0.48
Poor	9	5.8%	1.70 (0.50 • 5.81)	0.40	1.50 (0.25 • 8.97)	0.66	4.17 (0.89 • 19.66)	0.07
Other^	3	1.9%	-		-		-	
Location					-		-	
Right	79	51.0%	1		1		1	
Left	76	49.0%	1.17 (0.65 • 2.12)	0.60	0.88 (0.36 • 2.17)	0.78	0.58 (0.31 • 1.09)	0.09
MSI-status[a]								
MSI	34	21.9%	1		1		1	
MSS	120	77.4%	2.00 (0.79 • 5.08)	0.14	28.64 (0.24 • 3466.30)	0.17	0.64 (0.32 • 1.25)	0.19
Modified signature								
low risk	79	51.0%	1		1		1	
high risk	76	49.0%	2.12 (1.12 • 4.00)	0.02	5.17 (1.51 • 17.76)	0.009	2.12 (1.12 • 4.00)	0.02

^ there were no events in this subgroup; [a] n=1 missing

MULTIVARIATE					
MFS (events=25)		HFS (events=12)		OS (events=23)	
HR (95%CI)	**P value**	**HR (95%CI)**	**P value**	**HR (95%CI)**	**P value**
1		1		1	
1.16 (0.43 • 3.08)	0.77	3.18 (1.01 • 10.02)	0.049	1.44 (0.56 • 3.68)	0.45
0.90 (0.39 • 2.05)	0.80	1.43 (0.45 • 4.54)	0.54	0.94 (0.39 • 2.29)	0.89
2.93 (1.37 • 6.28)	0.006	1.19 (0.24 • 5.78)	0.83	3.60 (1.62 • 8.00)	0.002
1		1		1	
1.81 (0.94 • 3.50)	0.08	5.07 (1.46 • 17.59)	0.011	1.50 (0.76 • 2.96)	0.24

SUPPLEMENTARY TABLE 4. PATHWAY ANALYSIS RESULTS

	LET-7I		*MIR-30B*	
	P value	**% of predicted genes**	**P value**	**% of predicted genes**
GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIO_SERIES	1.35E-11	-		
GLYCOSAMINOGLYCAN_BIOSYNTHESIS_CHONDROITIN_SULFATE	1.95E-10	-		
AXON_GUIDANCE	3.41E-10	3.9%	3.59E-4	11.6%
FOCAL_ADHESION	7.61E-10	3.6%		
ECM_RECEPTOR_INTERACTION	1.33E-9	7.0%		
REGULATION_OF_ACTIN_CYTOSKELETON	1.80E-9	2.0%		
PATHWAYS_IN_CANCER	2.22E-9	-		
HEDGEHOG_SIGNALING_PATHWAY	2.44E-9	0.0%		
RENAL_CELL_CARCINOMA	2.72E-9	-		
DILATED_CARDIOMYOPATHY	4.69E-9	-		
HYPERTROPHIC_CARDIOMYOPATHY_HCM	4.83E-9	-		
WNT_SIGNALING_PATHWAY	9.16E-9	2.7%		
ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	1.05E-7	-		
TGF_BETA_SIGNALING_PATHWAY	2.73E-7	6.0%		
MELANOMA	4.10E-7	-		
GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES	3.05E-6	-		
UBIQUITIN_MEDIATED_PROTEOLYSIS			3.90E-4	-
MELANOGENESIS			1.35E-3	-
RNA_POLYMERASE			2.58E-3	-
HOMOLOGOUS_RECOMBINATION			2.69E-3	-
SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT			2.95E-3	-
PRION_DISEASES			4.39E-3	-
EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION			8.69E-3	-
TYROSINE_METABOLISM			1.09E-2	-
FATTY_ACID_METABOLISM			2.08E-2	-

SUPPLEMENTARY TABLE 5. GENES PER PATHWAY FOR WHICH EXPRESSION DATA WAS AVAILABLE

GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIO_SERIES							
B3GALT4	GLB1	HEXB	SLC33A1	ST3GAL2	ST6GALNAC3	ST6GALNAC5	ST8SIA1
B4GALNT1	HEXA	LCT	ST3GAL1	ST3GAL5	ST6GALNAC4	ST6GALNAC6	ST8SIA5

GLYCOSAMINOGLYCAN_BIOSYNTHESIS_CHONDROITIN_SULFATE						
B3GALT6	B3GAT3	CHPF2	CHST13	CHST3	CHSY3	UST
B3GAT1	B4GALT7	CHST11	CHST14	CHST7	CSGALNACT1	XYLT1
B3GAT2	CHPF	CHST12	CHST15	CHSY1	CSGALNACT2	XYLT2

AXON_GUIDANCE								
ABL1	EFNA1	EPHB1	LIMK1	NRAS	PLXNB2	RHOD	SEMA3G	SLIT2
ABLIM1	EFNA2	EPHB2	LIMK2	NRP1	PLXNB3	RND1	SEMA4A	SLIT3
ABLIM2	EFNA3	EPHB3	LRRC4C	NTN1	PLXNC1	ROBO1	SEMA4B	SRGAP1
ABLIM3	EFNA4	EPHB4	MAPK1	NTN3	PPP3CA	ROBO2	SEMA4C	SRGAP2
ARHGEF12	EFNA5	EPHB6	MAPK3	NTN4	PPP3CB	ROBO3	SEMA4D	SRGAP3
CDC42	EFNB1	FES	MET	NTNG1	PPP3CC	ROCK1	SEMA4F	UNC5A
CDK5	EFNB2	FYN	MRAS	PAK1	PPP3R1	ROCK2	SEMA4G	UNC5B
CFL1	EFNB3	GNAI1	NCK1	PAK2	PPP3R2	RRAS	SEMA5A	UNC5C
CFL2	EPHA1	GNAI2	NCK2	PAK3	PTK2	RRAS2	SEMA5B	UNC5D
CHP1	EPHA2	GNAI3	NFAT5	PAK4	RAC1	SEMA3A	SEMA6A	
CXCL12	EPHA3	GSK3B	NFATC1	PAK7	RAC2	SEMA3B	SEMA6B	
CXCR4	EPHA4	HRAS	NFATC2	PLXNA1	RAC3	SEMA3C	SEMA6C	
DCC	EPHA5	ITGB1	NFATC3	PLXNA2	RASA1	SEMA3D	SEMA6D	
DPYSL2	EPHA7	KRAS	NFATC4	PLXNA3	RGS3	SEMA3E	SEMA7A	
DPYSL5	EPHA8	L1CAM	NGEF	PLXNB1	RHOA	SEMA3F	SLIT1	

FOCAL_ADHESION								
ACTB	CCND1	COMP	IBSP	ITGB8	MYL6	PIK3CD	RAP1A	TNN
ACTC1	CCND2	CRK	IGF1	JUN	MYLK	PIK3CG	RAP1B	TNR
ACTG1	CCND3	CRKL	IGF1R	KDR	MYLK2	PIK3R1	RAPGEF1	TNXB
ACTN1	CDC42	CTNNB1	ILK	LAMA1	MYLK3	PIK3R2	RELN	TTN
ACTN2	CHAD	DIAPH1	ITGA10	LAMA2	MYLK4	PIK3R3	RHOA	VASP
ACTN3	COL11A1	DOCK1	ITGA11	LAMA3	PAK1	PIK3R5	ROCK1	VAV1
ACTN4	COL11A2	EGF	ITGA2	LAMA4	PAK2	PIP5K1C	ROCK2	VAV2
AKT1	COL1A1	EGFR	ITGA2B	LAMA5	PAK3	POTEKP	SHC1	VAV3
AKT2	COL1A2	ELK1	ITGA3	LAMB1	PAK4	PPP1CA	SHC2	VCL
AKT3	COL2A1	ERBB2	ITGA4	LAMB2	PAK7	PPP1CB	SHC3	VEGFA
ARHGAP35	COL3A1	FARP2	ITGA5	LAMB3	PARVA	PPP1CC	SHC4	VEGFB
ARHGAP5	COL4A1	FIGF	ITGA6	LAMB4	PARVB	PPP1R12A	SOS1	VEGFC
BAD	COL4A2	FLNA	ITGA7	LAMC1	PARVG	PRKCA	SOS2	VTN
BCAR1	COL4A4	FLNB	ITGA8	LAMC2	PDGFB	PRKCB	SPP1	VWF
BCL2	COL4A6	FLNC	ITGA9	LAMC3	PDGFC	PRKCG	SRC	XIAP
BIRC2	COL5A1	FLT1	ITGAV	MAP2K1	PDGFD	PTEN	THBS1	ZYX
BIRC3	COL5A2	FN1	ITGB1	MAPK1	PDGFRA	PTK2	THBS2	
BRAF	COL5A3	FYN	ITGB3	MAPK10	PDGFRB	PXN	THBS3	
CAPN2	COL6A1	GRB2	ITGB4	MAPK3	PDPK1	RAC1	THBS4	
CAV1	COL6A2	GSK3B	ITGB5	MAPK8	PGF	RAC2	TLN1	
CAV2	COL6A3	HGF	ITGB6	MAPK9	PIK3CA	RAC3	TLN2	
CAV3	COL6A6	HRAS	ITGB7	MET	PIK3CB	RAF1	TNC	

ECM_RECEPTOR_INTERACTION								
AGRN	COL3A1	COL6A3	GP5	ITGA3	ITGB4	LAMB1	SDC3	TNC
CD36	COL4A1	COL6A6	GP6	ITGA4	ITGB5	LAMB2	SDC4	TNN
CD44	COL4A2	DAG1	GP9	ITGA5	ITGB6	LAMB3	SPP1	TNR
CD47	COL4A4	FN1	HMMR	ITGA6	ITGB7	LAMB4	SV2A	TNXB
CHAD	COL4A6	FNDC1	HSPG2	ITGA7	ITGB8	LAMC1	SV2B	VTN
COL11A1	COL5A1	FNDC3A	IBSP	ITGA8	LAMA1	LAMC2	SV2C	VWF
COL11A2	COL5A2	FNDC4	ITGA10	ITGA9	LAMA2	LAMC3	THBS1	
COL1A1	COL5A3	FNDC5	ITGA11	ITGAV	LAMA3	RELN	THBS2	
COL1A2	COL6A1	GP1BA	ITGA2	ITGB1	LAMA4	SDC1	THBS3	
COL2A1	COL6A2	GP1BB	ITGA2B	ITGB3	LAMA5	SDC2	THBS4	

SUPPLEMENTARY TABLE 5. CONTINUED

REGULATION_OF_ACTIN_CYTOSKELETON								
ABI2	BAIAP2	DOCK1	FGF3	ITGA2	KRAS	PAK1	PIP4K2C	SOS1
ACTB	BCAR1	EGF	FGF4	ITGA2B	LIMK1	PAK2	PIP5K1A	SOS2
ACTC1	BDKRB1	EGFR	FGF5	ITGA3	LIMK2	PAK3	PIP5K1B	SSH1
ACTG1	BDKRB2	EZR	FGF6	ITGA4	MAP2K1	PAK4	PIP5K1C	SSH2
ACTN1	BRAF	F2	FGF7	ITGA5	MAP2K2	PAK7	POTEKP	SSH3
ACTN2	BRK1	F2R	FGF8	ITGA6	MAPK1	PDGFB	PPP1CA	TIAM1
ACTN3	CD14	FGD1	FGF9	ITGA7	MAPK3	PDGFRA	PPP1CB	TIAM2
ACTN4	CDC42	FGD3	FGFR1	ITGA8	MATK	PDGFRB	PPP1CC	TMSB4X
APC	CFL1	FGF1	FGFR2	ITGA9	MOS	PFN1	PPP1R12A	TMSB4XP8
APC2	CFL2	FGF10	FGFR3	ITGAD	MRAS	PFN2	PPP1R12B	TMSB4Y
ARHGAP35	CHRM1	FGF11	FGFR4	ITGAE	MSN	PFN3	PTK2	TTN
ARHGEF1	CHRM2	FGF12	FN1	ITGAL	MYH10	PFN4	PXN	VAV1
ARHGEF12	CHRM3	FGF13	GIT1	ITGAM	MYH14	PIK3CA	RAC1	VAV2
ARHGEF4	CHRM4	FGF14	GNA12	ITGAV	MYH9	PIK3CB	RAC2	VAV3
ARHGEF6	CHRM5	FGF16	GNA13	ITGAX	MYL1	PIK3CD	RAC3	VCL
ARHGEF7	CRK	FGF17	GNG12	ITGB1	MYL3	PIK3CG	RAF1	WAS
ARPC1A	CRKL	FGF18	GSN	ITGB2	MYLK	PIK3R1	RDX	WASF1
ARPC1B	CSK	FGF19	HRAS	ITGB3	MYLK2	PIK3R2	RHOA	WASF2
ARPC2	CYFIP1	FGF2	IQGAP1	ITGB4	MYLK3	PIK3R3	ROCK1	WASL
ARPC3	CYFIP2	FGF20	IQGAP2	ITGB5	MYLK4	PIK3R5	ROCK2	
ARPC4	DIAPH1	FGF21	IQGAP3	ITGB6	NCKAP1	PIKFYVE	RRAS	
ARPC5	DIAPH2	FGF22	ITGA10	ITGB7	NCKAP1L	PIP4K2A	RRAS2	
ARPC5L	DIAPH3	FGF23	ITGA11	ITGB8	NRAS	PIP4K2B	SLC9A1	

HEDGEHOG_SIGNALING_PATHWAY								
BMP2	BTRC	CSNK1G3	GSK3B	PRKX	STK36	WNT2	WNT6	ZIC2
BMP4	CSNK1A1	DHH	HHIP	PRKY	SUFU	WNT2B	WNT7A	
BMP5	CSNK1A1L	FBXW11	IHH	PTCH1	WNT1	WNT3	WNT7B	
BMP6	CSNK1D	GAS1	LRP2	PTCH2	WNT10A	WNT3A	WNT8A	
BMP7	CSNK1E	GLI1	PRKACA	RAB23	WNT10B	WNT4	WNT8B	
BMP8A	CSNK1G1	GLI2	PRKACB	SHH	WNT11	WNT5A	WNT9A	
BMP8B	CSNK1G2	GLI3	PRKACG	SMO	WNT16	WNT5B	WNT9B	

WNT_SIGNALING_PATHWAY								
APC	CSNK1A1L	DVL2	GSK3B	NKD2	PPP3CC	RBX1	TBL1XR1	WNT3A
APC2	CSNK1E	DVL3	JUN	NLK	PPP3R1	RHOA	TBL1Y	WNT4
AXIN1	CSNK2A1	EP300	LEF1	PLCB1	PPP3R2	ROCK1	TCF7	WNT5A
AXIN2	CSNK2A2	FBXW11	LRP5	PLCB2	PRICKLE1	ROCK2	TCF7L1	WNT5B
BTRC	CSNK2B	FOSL1	LRP6	PLCB3	PRICKLE2	RUVBL1	TCF7L2	WNT6
CACYBP	CTBP1	FRAT1	MAP3K7	PLCB4	PRKACA	SENP2	TP53	WNT7A
CAMK2A	CTBP2	FRAT2	MAPK10	PORCN	PRKACB	SFRP1	VANGL1	WNT7B
CAMK2B	CTNNB1	FZD1	MAPK8	PPARD	PRKACG	SFRP2	VANGL2	WNT8A
CAMK2D	CTNNBIP1	FZD10	MAPK9	PPP2CA	PRKCA	SFRP4	WIF1	WNT8B
CAMK2G	CUL1	FZD2	MMP7	PPP2CB	PRKCB	SFRP5	WNT1	WNT9A
CCND1	CXXC4	FZD3	MYC	PPP2R1A	PRKCG	SIAH1	WNT10A	WNT9B
CCND2	DAAM1	FZD4	NFAT5	PPP2R1B	PRKX	SKP1	WNT10B	
CCND3	DAAM2	FZD5	NFATC1	PPP2R2A	PRKY	SMAD2	WNT11	
CER1	DKK1	FZD6	NFATC2	PPP2R2B	PSEN1	SMAD3	WNT16	
CHP1	DKK2	FZD7	NFATC3	PPP2R2C	RAC1	SMAD4	WNT2	
CREBBP	DKK4	FZD8	NFATC4	PPP3CA	RAC2	SOX17	WNT2B	
CSNK1A1	DVL1	FZD9	NKD1	PPP3CB	RAC3	TBL1X	WNT3	

TGF_BETA_SIGNALING_PATHWAY								
ACVR1	BMP5	COMP	GDF7	LEFTY1	PPP2CB	SMAD1	SP1	THBS4
ACVR1B	BMP6	CREBBP	ID1	LEFTY2	RBL1	SMAD2	TFDP1	TNF
ACVR1C	BMP7	CUL1	ID2	LTBP1	RBL2	SMAD3	TGFB1	ZFYVE16
ACVR2A	BMP8A	DCN	ID3	MAPK1	RBX1	SMAD4	TGFB2	ZFYVE9
ACVR2B	BMP8B	E2F4	ID4	MAPK3	RHOA	SMAD5	TGFB3	
ACVRL1	BMPR1A	E2F5	IFNG	MYC	ROCK1	SMAD6	TGFBR1	
AMH	BMPR1B	EP300	INHBA	NODAL	ROCK2	SMAD7	TGFBR2	
AMHR2	BMPR2	FST	INHBB	NOG	RPS6KB1	SMAD9	THBS1	
BMP2	CDKN2B	GDF5	INHBC	PITX2	RPS6KB2	SMURF1	THBS2	
BMP4	CHRD	GDF6	INHBE	PPP2CA	SKP1	SMURF2	THBS3	

SUPPLEMENTARY TABLE 6. LIST OF GENES THAT WERE PREDICTED AS DIRECT TARGETS BY THE COMBINED USE OF 3 INDEPENDENT TARGET PREDICTION ALGORITHMS

MIRNA	GENE SYMBOL	GENE NAME	SPEARMAN'S RHO	P VALUE
miR-30b	*ABL1*	ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase	-0.032	0.63
	CFL2	Cofilin 2	0.107	0.11
	DPYSL2	Dihydropyrimidinase Like 2	0.072	0.28
	EFNA3	Ephrin A3	0.121	0.07
	GNAI2	G Protein Subunit Alpha I2	-0.030	0.65
	NFAT5	Nuclear Factor Of Activated T-Cells 5	-0.286	<0.0001
	NFATC2	Nuclear Factor Of Activated T-Cells 2	-0.009	0.90
	NFATC3	Nuclear Factor Of Activated T-Cells 3	0.067	0.32
	PLXNA2	Plexin A2	-0.077	0.25
	PLXNC1	Plexin C1	-0.007	0.92
	PPP3R1	Protein Phosphatase 3 Regulatory Subunit B, Alpha	-0.331	<0.001
	RASA1	RAS P21 Protein Activator 1	-0.083	0.22
	SEMA6B	Semaphorin 6B	-0.198	0.003
	SEMA6D	Semaphorin 6D	0.178	0.007
	SRGAP3	SLIT-ROBO Rho GTPase Activating Protein 3	-0.018	0.79
	UNC5C	Unc-5 Netrin Receptor C	0.295	<0.0001
let-7i	*ACVR1C*	Activin A Receptor Type 1C	-0.113	0.09
	CHRD	Chordin	0.422	<0.0001
	COL1A2	Collagen Type I Alpha 2 Chain	0.561	<0.0001
	COL3A1	Collagen Type III Alpha 1 Chain	0.576	<0.0001
	COL4A6	Collagen Type IV Alpha 6 Chain	-0.258	0.000
	COL5A2	Collagen Type V Alpha 2 Chain	0.512	<0.0001
	E2F5	E2F Transcription Factor 5	0.076	0.25
	FNDC3A	Fibronectin Type III Domain Containing 3A	-0.295	<0.0001
	GDF6	Growth Differentiation Factor 6	0.053	0.54
	ITGB3	Integrin Subunit Beta 3	0.196	0.003
	TGFBR1	Transforming Growth Factor Beta Receptor 1	0.346	<0.0001

J.P.M. Burbach*, S.A. Kurk*, R.R.J. Coebergh van den Braak, V.K. Dik, A.M. May, G.A. Meijer, C.J.A. Punt, G.R. Vink, M.Los, N. Hoogerbrugge, P.C. Huijgens, J.N.M. IJzermans, E.J. Kuipers, M.E. de Noo, J.P. Pennings, A.M.T. van der Velden, C. Verhoef, P.D. Siersema, M.G.H. van Oijen, H.M. Verkooijen, M. Koopman.

* These authors contributed equally

8

PROSPECTIVE DUTCH COLORECTAL CANCER COHORT: AN INFRASTRUCTURE FOR LONG-TERM OBSERVATIONAL, PROGNOSTIC, PREDICTIVE AND (RANDOMIZED) INTERVENTION RESEARCH

ACTA ONCOLOGICA 2016

ABSTRACT

BACKGROUND

Systematic evaluation and validation of new prognostic and predictive markers, technologies and interventions for colorectal cancer (CRC) is crucial for optimizing patients' outcomes. With only 5-15% of patients participating in clinical trials, generalizability of results is poor. Moreover, current trials often lack the capacity for post-hoc subgroup analyses. For this purpose, a large observational cohort study, serving as a multiple trial and biobanking facility, was set up by the Dutch Colorectal Cancer Group (DCCG).

METHODS / DESIGN

The Prospective Dutch ColoRectal Cancer cohort is a prospective multidisciplinary nation-wide observational cohort study in The Netherlands (yearly CRC incidence of 15,500). All CRC patients (stage I-IV) are eligible for inclusion, and longitudinal clinical data are registered. Patients give separate consent for the collection of blood and tumor tissue, filling out questionnaires, and broad randomization for studies according to the innovative cohort multiple randomized controlled trial design (cmRCT), serving as an alternative study design for the classic randomized controlled trial.

Objectives of the study include 1) systematically collected long-term clinical data, patient-reported outcomes and biomaterials from daily CRC practice and 2) to facilitate future basic, translational and clinical research including interventional and cost-effectiveness studies for both national and international research groups with short inclusion periods, even for studies with stringent inclusion criteria.

RESULTS

Seven months after initiation 650 patients have been enrolled, 8 centers participate, 15 centers await IRB approval and 9 embedded cohort- or cmRCT-designed studies are currently recruiting patients.

CONCLUSION

This cohort provides a unique multidisciplinary data, biobank, and patient reported outcomes collection initiative, serving as an infrastructure for various kinds of research aiming to improve treatment outcomes in CRC patients. This comprehensive design may serve as an example for other tumor types.

BACKGROUND

Worldwide, colorectal cancer (CRC) is the third most common malignancy in men and second in women.[1] With a continuously rising incidence, an estimated 1.35 million new cases are diagnosed yearly, associated with 694,000 annual deaths. In the past decades, substantial progress has been made in diagnosis and treatment of CRC, resulting in an increasing number of CRC survivors.[2-8] The implementation of national CRC screening programmes is expected to increase the incidence of CRC.[9] The increasing incidence of CRC, in combination with improved survival rates, has led to high numbers of people living with (the consequences of) CRC. In addition to treatment parameters and outcomes, also quality of life, workability, and daily functioning during and after CRC treatment are becoming increasingly important parameters in research.

There is no consensus on the use of prognostic parameters in CRC, and predictive factors for treatment are scarce. Also there is a growing availability of new molecular markers[10-13] and innovative treatment options. This puts increasing pressure on the current research system, since large numbers of patients are required to assess relevance or superiority before their implementation into clinical practice. This warranted large number greatly exceeds the amount of patients that currently participate in clinical trials (5%-15%).[14-16] Low recruitment-rates may also imply selective inclusion of patients in trials rather than representative population samples[17], which may result in limited external validity of outcomes. The danger of the extrapolation of study results to the general population was recently shown. Survival outcomes of patients with metastatic CRC (mCRC) treated within the scope of a randomized study were significantly better than the survival outcomes in patients not fulfilling the study eligibility criteria and treated outside the trial with the same drugs during the same period.[17,18] Moreover, study designs classically used for comparative research often lack the ability to provide sufficient data for subgroup treatment effects or post-hoc evaluation. For example, immunotherapy showed to be effective in mCRC patients with microsatellite instability (MSI). As MSI is only observed in 3-5% of the mCRC patients, the conduction of a large randomized phase 3 trial will be challenging.[19] It is therefore desirable to include all these patients in a large representative cohort of CRC patients who are prospectively followed for relevant outcomes that enables to study the value of novel prognostic and predictive biomarkers in large, but also small subgroups of patients. It would be ideal to use data from routine sources such as hospital systems or (cancer) registries, but these sources often lack the required detail about (changes in) treatments, doses, toxicity, and response, which is paramount for this purpose. As an alternative, a large

observational cohort has the advantage that it can provide a standardized data-collection, dedicated data-monitoring and intensive follow-up, all of which are especially valuable for long term research in prognostic or predictive determinants. Ideally, all new interventions should be evaluated in Randomized Controlled Trials (RCTs) since this is considered the gold standard to prove effectiveness. However, this design in itself is often not only complicated by slow recruitment rates and limited generalizability[15], it is also subject to a considerable delay between conceptualization and start, limited long-term follow-up, inadequate collection of patient-reported outcomes (PROMs), high non-completion rates and high costs.[16] An innovative alternative proposed for the classic RCT is the 'cohort multiple Randomized Controlled Trial' (cmRCT).[20] This design was originally developed as an alternative for classic pragmatic RCTs, and combines useful features from both classic RCTs (randomization) and prospective observational cohort studies. The design is characterized by three features: 1) patient-centred informed-consent approach; 2) framework to systematically collect long-term clinical follow-up as well as PROMs; and 3) efficient recruitment for trials by asking patients to give 'broad consent for randomization' in future trials. Unique features of the cmRCT design are that it allows to conduct multiple randomized trials simultaneously and that patients can participate in multiple non-conflicting trials at the same time.[20] The design itself and its implementation in this study are explained in more detail in **Box 2**.

We believe that a prospective observational cohort study can provide a standardized and validated collection of long term clinical data, tissue and blood samples and PROMs to establish a continuous source for a variety of research purposes. This research database can, among others, be used to investigate what (intrinsic and environmental) factors are associated with survival and PROMs, to find new predictive markers for treatment outcomes and side-effects, and to develop more accurate diagnostic tests and efficient follow-up surveillance strategies.

METHODS / DESIGN

DESIGN AND OBJECTIVES

This is a project of the Dutch Colorectal Cancer Group (DCCG) and was launched as the prospective Dutch colorectal cancer cohort (Dutch: 'Prospectief Landelijk ColoRectaal kanker Cohort' (PLCRC)). The cohort is designed in accordance with the 'Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement' guidelines.[21] The project aims to collect high quality clinical data, biomaterials and PROMs of a large cohort of CRC patients that are prospectively followed from

primary diagnosis until death. All data are collected under a broad informed consent to facilitate future basic, translational and clinical research (**Box 1**). Furthermore, the cohort aims to serve as an infrastructure to conduct multiple simultaneous (randomized controlled) trials (according to the cmRCT design (**Box 2**)).

BOX 1: THE MAIN OBJECTIVES OF THE PROSPECTIVE DUTCH COLORECTAL CANCER COHORT (PLCRC) ARE:

- To execute a prospective observational cohort study aiming to include all Dutch CRC patients and follow them until death.
- To prospectively collect high quality data on medical history, comorbidities, clinical characteristics, imaging, pathology results, tumor characteristics, treatment, survival, recurrence, hospitalization, adverse events, toxicity and (long-term) outcomes of experimental interventions (table 1).
- To collect, store and make available blood and tumor tissue samples.
- To systematically collect patient-reported outcomes on quality of life, workability and functional outcomes.
- To provide detailed data on daily clinical care in the Netherlands.
- To create an infrastructure to facilitate studies of different nature, including:

  A. Prognostic and predictive research

  B. Biological research and (epi-)genetic research

  C. Studies that compare novel therapies or interventions in a target population according to the innovative cohort multiple randomized controlled trial (cmRCT) design serving as a pragmatic alternative for classic RCTs.

  D. Cost-effectiveness studies

STUDY POPULATION

Patients with histologically proven CRC are eligible for participation if they are 18 years or older and have given written informed consent. Only mentally incompetent and non-Dutch speaking patients are withheld from participation. The aim of the PLCRC project is to include all eligible patients in The Netherlands, a country with a yearly incidence of 15,500.

INFORMED CONSENT

Study information is given by researchers, research assistants, nonphysician clinicians and/or physicians during routine hospital visits after initial diagnosis, preferably before start of treatment. 'General' informed consent is mandatory for participation in this study and allows the collection of long term clinical and survival data. Subsequently, patients are given the option to consent to 1) filling out questionnaires on health-related quality of life, functional outcomes and workability, 2) biobanking

of tumor and normal tissue, 3) collection of blood samples, and 4) to be offered studies conducted within the infrastructure of the cohort, either in accordance with the cmRCT design or not. When participants are offered to participate in a trial or intervention, an additional informed consent needs to be signed before patients can be enrolled in that trial (**Box 2**). The PLCRC informed consent procedure is a dynamic process since patients can withdraw or alter their consent preferences at any time during the study.

BOX 2: THE COHORT MULTIPLE RANDOMIZED CONTROLLED TRIAL DESIGN

The basis of the cohort multiple Randomized Controlled Trial (cmRCT) design is a prospective observational cohort of patients with a certain condition [20], in our case CRC, in which all patients in principle undergo standard care. Within this cohort, clinical characteristics and standardized outcome measures are collected at baseline and regularly during follow-up. Clinical and self-reported data are used to compare effectiveness and safety of trialled interventions.

Practically, when an RCT is conducted within the cmRCT cohort, the first step is to identify a subcohort of all patients eligible for the intervention. Some of these patients are randomly selected and offered the experimental intervention (intervention group). If patients accept the offer, they are sure to undergo the experimental intervention. If they refuse they will undergo standard care. Eligible patients in the subcohort not randomly selected (control group), undergo standard treatment and do not receive any information on the trial. Outcomes are compared between randomly selected and non-selected patients. This process can be repeated for multiple (experimental) interventions simultaneously, offering (more) reliable direct comparisons between interventions and standard care.

In the cmRCT design a patient-centered informed consent procedure is obtained [42] by asking all patients to give 'broad consent for randomization' after enrolment [42, 43] This allows researchers to randomly selected patients from the cohort, and offer them experimental interventions, while patients who are not randomly selected serve as controls and undergo standard care without further notification. When informing patients about broad consent for randomization, we explicitly state that not all patients that consent will be offered an experimental intervention since offers are based on random selection. When they got offered an experimental intervention they can either accept the intervention or they can refuse and undergo standard care. Also they are told that they may become (temporarily) ineligible for future trials if they already participate in a trial which measures interfering endpoints. We ensure that patients will never be withheld proven effective care.

With this consent procedure we aim to mimic clinical practice, where people are usually not told about treatments they will not / cannot receive. The patient centered informed consent is expected to prevent cross-over and disappointment bias, especially in situations where, regardless of clinical equipoise, a new intervention is highly preferred by doctors and patients. Asking broad consent for randomisation also deals with the controversial ethical aspect of pre-randomization (as introduced by Zelen [44]) by obtaining upfront consent from all patients for randomization and data use in future comparative research, thereby not randomizing patients without prior notification and their consent.

After inclusion, participants are assigned a unique study identification(ID), which remains the only patient identifier throughout all further processes in the cohort's infrastructure. Cohort inclusion does not limit participation in other observational studies. However, patients may become temporarily ineligible to participate in clinical trials outside the cohort in case they already participate in a cohort-embedded trial that has interfering endpoints.

ETHICS

The study was originally initiated as a monocenter study for which it received approval of the medical ethical review committee of the University Medical Center Utrecht (The Netherlands) in June 2013 (METC 12-510). Subsequently, approval was extended by the same IRB for a multicenter set-up, which was implemented in September 2015. All new intervention studies trialled within the cohort require separate approval from a medical ethical review committee. Study protocols and final results of PLCRC trials are available on the website: www.plcrc.nl excluding study protocols of cmRCT trials, since this design does not allow patients enrolled in the control arm to be informed on these studies (**Box 2**). PLCRC is registered at Clinicaltrials.gov under NCT02070146.

DATA COLLECTION AND ENDPOINTS

OBSERVATIONAL CLINICAL AND SURVIVAL DATA

Extensive observational clinical data (**Table 1**) are collected from medical charts by trained data-managers of the "Netherlands Comprehensive Cancer Organisation" (Dutch: Integraal Kankercentrum Nederland (IKNL)[22]) and does not require additional effort from participating hospitals or patients. The collected data is stored in the "Netherlands Cancer Registry" (Dutch: Nederlandse Kankerregistratie (NKR) [23]) Study spe\cific data, not standardly collected in the NKR, is gathered separately by IKNL data managers, or by study-personal or researchers.

BIOBANKING OF BLOOD AND TUMOR TISSUE MATERIALS

Tumor tissues are collected after routine surgery and stored as five snap frozen tissue samples, two Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples and two tissue sample cores for Tissue Micro Arrays. Blood samples (10ml serum and 10ml EDTA) are collected during routine blood withdrawal before treatment. Serum is divided over six 0.5ml samples and the EDTA sample is divided over six 0.5ml plasma samples and three 0.9ml pellet samples before being frozen and stored. Snap freezing of tumor tissue, FFPE processing and blood sample processing are performed locally in

TABLE 1. CLINICAL DATA COLLECTION WITHIN THE PROSPECTIVE DUTCH COLORECTAL CANCER COHORT

PART A: PATIENT ID AND DATA SOURCES

Patient identification & demographics
- Patient-ID code
- Birth information (date and city)
- Gender

Data source
- Hospital
- Patient number within hospital

Data capture
- ID of person who captures data

Study participation within the cohort
- Number and name of studies/trials
- Date(s) of inclusion
- Date(s) of completed study/trial follow-up

PART B: PRE-TREATMENT RECORD

Medical history
Cancer specific
- Date, location, type, treatment, treatment outcome

Comorbidity
- Cardiac, pulmonal, vascular, gastro-intestinal, neurological, gynecological, urological, muscle/bones, endocrine.

Intoxication
- Smoking at diagnosis
- Alcohol use at diagnosis

Physical examination
- BMI (length & weight)
- WHO performance status

Diagnosis & tumor information
- Sequential tumor number
- Date of diagnosis
- Source/procedure of diagnosis

Laboratory investigations
- CEA

Diagnostic work-up
Endoscopy
- Date, hospital, procedure, procedure complete?
- Number of tumors/polyps, distance from anal verge
- Endoscopic treatment

Pathology
- Type, differentiation, T-stage

Imaging
- Modalities, date(s), hospital
- cTNM, MRF involvement, distance from anal verge

Multidisciplinary Tumor Board
- Date & final staging

PART C: TREATMENT RECORD

Radiotherapy*
Setting (neo-adjuvant/adjuvant)
Indication for radiotherapy
Treatment
- Start date first fraction
- Standard: # fractions, fraction dose, total dose
- Boost: # fractions, fraction dose, total boost dose
- Total received dose and fractions
- Stop/completion date
- Response (TRG, ycTNM)
- Adverse events (date, cause, management)

Medical oncology*
Setting (neo-adjuvant/adjuvant)
Indication for systemic therapy
Treatment
- Start date first cycle
- Agent, dose, number of cycles
- Total received dose and cycles
- Stop/completion date
- Response (TRG, ycTNM)
- Adverse events (date, cause, management)

Surgery*
ASA classification
Procedure
- Date, hospital
- Setting (elective/acute)
- Approach (open/laparoscopic/robot
- Type ((Sub)Total Colectomy, LAR, APR,Hartmann)
- Anastomosis (type, stapled/sewn)
- Date of discharge

Stoma
- Date, hospital
- Setting (elective/acute)
- Temporary/definitive
- Type (ileostoma, colostoma)
- Date of stoma reversal
- Peri-operative complications (anastomic leakage, abcess, ileus)
- Post operative complications (incl. wound complications)

Pathology*
- pTNM
- Tumor regression grade
- Radicality of resection
- Circumferential resection margin (CRM)
- # lymph nodes & # positive lymph nodes in specimen
- Angio- and lymphatic invasion
- Perforation of the bowel
- Molecular markers (BRAF, RAS, MSI status)

TABLE 1. CONTINUED

PART D: POST-TREATMENT / FOLLOW-UP RECORD

Oncological follow-up
Recurrence*
- Recurrence (date, number, location(s),
- Treatment of recurrence (new PART C entry)
- Setting (curative/palliative)

Metastases*
- Metachronic metastases (date, number, location(s))
- Treatment of metastases (new PART C entry)
- Setting (curative/palliative)

Complications*
- Grade 3/4 adverse events or complications

Survival
- Date of last hospital visit
- Death (including date and cause)

* Multiple entries are allowed within each tumor episode

participating hospitals and transported to regional biobank facilities for long-term storage. To provide a sustainable infrastructure for biobanking, we established close collaborations with existing national organisations for use of their expertise, and to prevent duplicate data entry and unnecessary costs. These initiatives include the Dutch Biobanking and BioMolecular resources Research Infrastructure (BBMRI-NL; www.bbmri.nl) and the CTMM Translational Research IT (TraIT, www.ctmm-trait.nl).

LONGITUDINAL ASSESSMENT OF PROMS

Nationally and internationally accepted and validated questionnaires are used to measure PROMs, which include EORTC QLQ-C30[24], -CR29[25] and –CIPN20[26], Euroqol- 5 dimensional (EQ-5D)[27], Work Ability Index (WAI)[28], Low Anterior Resection Syndrome (LARS)[29], Stoma quality of life scale (SQOLS)[30], Short Questionnaire to assess Health-enhancing physical activity (SQUASH)[31], Hospital Anxiety and Depression Score (HADS)[32], multidimensional fatigue score (MFI-20)[33] and the Self-administered Comorbidity Questionnaire (SCQ)[34]. Patients have the option to fill out paper questionnaires, or use the digital patient tracking system PROFILES (Patient Reported Outcomes Following Initial Long term treatment and Survivor Ship).[35] Questionnaires are provided at enrolment (baseline) and 3, 6, 12, 18 and 24 months thereafter, followed by an annual questionnaire for the remainder of their participation or until death. The comprehensive selection of PROM questionnaires which are administered frequently at pre-defined time points enable the use of PROM outcomes as consistent endpoints in research. Within PROFILES, the option exists to compare PROMs of the PLCRC patient population to those of large population-based samples of cancer patients and a normative Dutch cohort.

DATA FOR FUTURE STUDIES

Data collected and stored in the NKR is at all times available to centres where the data were originally captured. Additional data required for future research, including study specific data not standardly collected in the NKR, PROMs and biomaterials, is available upon request.

SAFETY

The observational nature of this study eliminates the appearance of adverse events (AEs) and serious adverse events (SAEs) as a result of participation in this study. However, grade 3/4 incidents according to the Common Terminology Criteria for Adverse Events (CTCAE) are important outcome parameters in research, and are therefore systematically collected. In cohort-embedded trials, reporting of SAEs occurs as specified in the separate trial protocols.

PROCEEDINGS

Recruitment of patients initially started in one center in February 2013. At this first site, a highly dedicated patient-routine was introduced in which almost all CRC patients visiting the radiotherapy department were approached for participation.[36] During the observed period, 90% of all approached patients consented to inclusion, of whom 90% additionally consented to receive questionnaires, 83% to the storage of biomaterials and 85% to 'broad consent for randomization' in future trials. From September 2015 onwards, recruitment has been extended to multiple centers throughout The Netherlands and more centers expect to start recruitment in the (near) future. Currently, 8 hospitals are open for inclusion, 15 hospitals are preparing or awaiting IRB approval and 650 patients have been enrolled of which 160 patients were included over the last three months. In addition 9 cohort studies that are currently recruiting patients have been embedded within the PLCRC infrastructure, including 2 RCTs that are designed according to the cmRCT design (**Table 2**). For both RCTs inclusion rates have looked promising so far, with numbers greatly exceeding those of other RCTs.[16,17,36] PLCRC patients may be eligible for both trials; therefore patients that participate in both trials are stratified according to their received neo-adjuvant treatment as a first step to investigate the feasibility of overlapping trials within the cmRCT infrastructure.

TABLE 2. ONGOING STUDIES EMBEDDED IN THE PLCRC COHORT CURRENTLY RECRUITING PATIENTS

NAME STUDY	DESIGN	DESCRIPTION OF STUDY	STUDY POPULATION	SAMPLE SIZE (N)
Rectal Boost	RCT*	Pre-operative dose-escalation BOOST versus standard chemoradiation on pathologic response rates in locally advanced rectal cancer [42]	T3 with threatened mesorectal fascia (<1 mm), T4 or N2M0 rectal cancer	60 versus 60
Sponge	RCT*	Effect of a retractor SPONGE during laparoscopic rectal cancer surgery versus Trendelenburg positioning on perioperative complications and hospital stay [43]	Stage I-IV CRC undergoing laparoscopic surgery	94 versus 94
PROTECT	Prospective cohort	PlcRc cOhorT: diEtary intake after diagnosis and ColorecTal cancer outcomes	Stage I-IV CRC	1000
CONNECTION	Validation study with different workpackages including a prospective cohort study	A nation-wide COloN CaNcer rEgistry and stratifiCaTION effort for the development and validation of genetic, proteomic and histopathological assays to stratify patients for adjuvant therapy	Stage II & III colon cancer	NA
MEDOCC	Prospective cohort	Molecular Early Detection of Colorectal Cancer study to investigate the prognostic value of circulating tumor DNA [44]	Stage II colon cancer	846
SPECTRE	Pilot study	Ultra-high field 7.0 Tesla MR SPECTRoscopy to monitor capEcitabine metabolism in liver metastases	metastatic CRC	26
Recap	Case-control	Case match control study investigating the benefit of last line regorafenib treatment	Metastatic colon cancer and metastatic RAS wildtype rectal cancer	125 versus 125
Quality of life study 1	Prospective cohort	Impact of short-course radiation versus long-course chemoradiation for rectal cancer on quality of life	Stage II-IV rectal cancer	>60 versus >60
Quality of life study 2	Prospective cohort	Quality of life comparison between patients undergoing radiation followed by low anterior resection versus abdomino-perineal excision	Stage II-IV rectal cancer	>100 versus >100

*Randomized controlled trial according to the cohort multiple randomized trial design (box 2)

DISCUSSION

This multidisciplinary prospective observational cohort study provides a validated and standardized collection of high quality clinical data, PROMs and biomaterials of a large cohort of CRC patients to facilitate future basic, translational and clinical research. By making this collection available to researchers upon request, the cohort foresees in the growing need for comprehensive data collection and sharing.[37] Through its broad eligibility the cohort is likely to reach high recruitment rates, thereby allowing to conduct highly powered analyses, improve recruitment rates to trials and reduce long inclusion periods for studies that use stringent inclusion criteria, i.e. aim to include specific subgroups of patients.

Over the past decades several other cancer registries and prospective observational cohort studies have been initiated in The Netherlands.[38-41] These initiatives serve different purposes, such as providing insight in incidence and prevalence, in the effects of nutrition, lifestyle or treatments in current daily practice, or to serve as a platform for monitoring quality of care. Often these databases or registries are used for various types of research, even though they were originally not intended for this (specific type of research) purpose. In addition, most of these existing cohorts are closed cohorts, or maintain restricted inclusion criteria that limit the inclusion to patients with certain cancer subtypes or to patients that received certain treatment(s). The PLCRC initiative differs in respect to these limitations by its dynamic design, its unlimited accrual potential, and by allowing the inclusion of CRC patients of all stages, independent of their received treatments. Furthermore, some of the registries contain data that are provided by healthcare professionals themselves. Therefore, these registries may lack adequate validation and monitoring of the included data, which likely increases the risk of misclassification and/or underreporting of (adverse) outcomes. By harboring independent data managers and monitors for the PLCRC cohort, we attempt to limit incorrect data registration, which should improve the robustness of outcomes and trial results from our cohort. Finally, the PLCRC cohort is unique in the sense that it provides a comprehensive dataset, which includes aggregated high quality multidisciplinary clinical information, biomaterials and PROMs, and with the possibility of performing studies according to the cmRCT design.

We acknowledge potential challenges and limitations arise from our cohort's infrastructure. First, by asking informed consent we introduce the risk of selection at a patient level (if specific subgroups do not provide consent as much as other subgroups), or, in case hospitals decide not to participate, at a hospital level. However, since we parallel our data to data from The Netherlands cancer registry (recording baseline and clinical data from all histologically confirmed CRC patients in The

Netherlands), we are able to obtain insight in the selection that exists in our cohort both within and between participating and non-participating centers. Secondly, the cmRCT infrastructure is not appropriate for all types of research. Since experimental interventions are compared against standard care, the design does not allow placebo-controlled settings or the use of non-standardly measured outcomes. Nevertheless, such trials can still be embedded in the cohort as classic RCTs for which the cohort can be used as a recruitment pool. The high participation rates, high levels of consent to the additional consent options and the willingness of hospitals to participate in PLCRC indicate that this innovative design is feasible in the oncology practice, acceptable for patients and healthcare professionals, facilitate research projects and is likely to provide generalizable results. Future results are needed to confirm whether the cmRCT design indeed provides an acceptable alternative for classic pragmatic RCTs.

In summary, this cohort provides a unique high-quality multidisciplinary data collection initiative, including biobanking and PROMs, which serves as an infrastructure to perform various kinds of research in the field of CRC. The set-up allows evaluation of long-term clinical and PROMs of patients treated in current routine care, and that of patients treated by experimental interventions in a randomized controlled setting. This comprehensive design may serve as an example for research in other tumor types.

REFERENCES

1. GLOBOCAN. 2012. at http://globocan.iarc.fr/Default.aspx.)

2. Baca B, Beart RW, Jr., Etzioni DA. Surveillance after colorectal cancer resection: a systematic review. Diseases of the colon and rectum 2011;54:1036-48.

3. Faivre J, Dancourt V, Lejeune C, et al. Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. Gastroenterology 2004;126:1674-80.

4. Kapiteijn E, Marijnen CA, Nagtegaal ID, et al. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer. The New England journal of medicine 2001;345:638-46.

5. Lemmens VE, de Haan N, Rutten HJ, et al. Improvements in population-based survival of patients presenting with metastatic rectal cancer in the south of the Netherlands, 1992-2008. Clinical & experimental metastasis 2011;28:283-90.

6. Ragnhammar P, Hafstrom L, Nygren P, Glimelius B, Care SB-gSCoTAiH. A systematic overview of chemotherapy effects in colorectal cancer. Acta oncologica 2001;40:282-308.

7. van Steenbergen LN, Elferink MA, Krijnen P, et al. Improved survival of colon cancer due to improved treatment and detection: a nationwide population-based study in The Netherlands 1989-2006. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2010;21:2206-12.

8. Gustavsson B, Carlsson G, Machover D, et al. A review of the evolution of systemic chemotherapy in the management of colorectal cancer. Clinical colorectal cancer 2015;14:1-10.

9. Schreuders EH, Ruco A, Rabeneck L, et al. Colorectal cancer screening: a global overview of existing programmes. Gut 2015;64:1637-49.

10. 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. Nature medicine 2013;19:614-8.

11. Locker GY, Hamilton S, Harris J, et al. ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2006;24:5313-27.

12. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, Kerr D. Genetic prognostic and predictive markers in colorectal cancer. Nature reviews Cancer 2009;9:489-99.

13. Sinicrope FA, Okamoto K, Kasi PM, Kawakami H. Molecular Biomarkers in the Personalized Treatment of Colorectal Cancer. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association 2016.

14. Wilt TJ, Brawer MK, Jones KM, et al. Radical prostatectomy versus observation for localized prostate cancer. The New England journal of medicine 2012;367:203-13.

15. Bennette CS, Ramsey SD, McDermott CL, Carlson JJ, Basu A, Veenstra DL. Predicting Low Accrual in the National Cancer Institute's Cooperative Group Clinical Trials. Journal of the National Cancer Institute 2016;108.

16. Young RC. Cancer clinical trials--a chronic but curable crisis. The New England journal of medicine 2010;363:306-9.

17. Mol L, Koopman M, van Gils CW, Ottevanger PB, Punt CJ. Comparison of treatment outcome in metastatic colorectal cancer patients included in a clinical trial versus daily practice in The Netherlands. Acta oncologica 2013;52:950-5.

18. Sorbye H, Pfeiffer P, Cavalli-Bjorkman N, et al. Clinical trial enrollment, patient characteristics, and survival differences in prospectively registered metastatic colorectal cancer patients. Cancer 2009;115:4679-87.

19. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. The New England journal of medicine 2015;372:2509-20.

20. Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. Bmj 2010;340:c1066.

21. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370:1453-7.

22. Netherlands Cancer Registry. 2015. at http://www.iknl.nl/.)

23. Dutch Cancer Registry. 2015. at http://www.cijfersoverkanker.nl/.)

24. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 1993;85:365-76.

25. Whistance RN, Conroy T, Chie W, et al. Clinical and psychometric validation of the EORTC QLQ-CR29 questionnaire module to assess health-related quality of life in patients with colorectal cancer. European journal of cancer 2009;45:3017-26.

26. Postma TJ, Aaronson NK, Heimans JJ, et al. The development of an EORTC quality of life questionnaire to assess chemotherapy-induced peripheral neuropathy: the QLQ-CIPN20. European journal of cancer 2005;41:1135-9.

27. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Annals of medicine 2001;33:337-43.

28. de Zwart BC, Frings-Dresen MH, van Duivenbooden JC. Test-retest reliability of the Work Ability Index questionnaire. Occupational medicine 2002;52:177-81.

29. Emmertsen KJ, Laurberg S. Low anterior resection syndrome score: development and validation of a symptom-based scoring system for bowel dysfunction after low anterior resection for rectal cancer. Annals of surgery 2012;255:922-8.

30. Baxter NN, Novotny PJ, Jacobson T, Maidl LJ, Sloan J, Young-Fadok TM. A stoma quality of life scale. Diseases of the colon and rectum 2006;49:205-12.

31. Wendel-Vos GC, Schuit AJ, Saris WH, Kromhout D. Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. Journal of clinical epidemiology 2003;56:1163-9.

32. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta psychiatrica Scandinavica 1983;67:361-70.

33. Smets EM, Garssen B, Bonke B, De Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. Journal of psychosomatic research 1995;39:315-25.

34. Sangha O, Stucki G, Liang MH, Fossel AH, Katz JN. The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research. Arthritis and rheumatism 2003;49:156-63.

35. van de Poll-Franse LV, Horevoorts N, van Eenbergen M, et al. The Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship registry: scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. European journal of cancer 2011;47:2188-94.

36. Burbach JPM YA, van der Velden, van den Bongard, Intven, Reerink, Relton, van Gils, van Vulpen, Verkooijen. Implementation of the 'cohort multiple randomized controlled trial' design for systematic randomized evaluation of multiple interventions in oncology. Int J Clin Epi (submitted) 2015.

37. Taichman DB, Backus J, Baethge C, et al. Sharing Clinical Trial Data--A Proposal from the International Committee of Medical Journal Editors. The New England journal of medicine 2016;374:384-6.

38. Dutch Surgical Colorectal Audit. (Accessed 03-02-2016, at http://dsca.clinicalaudit.nl/.)

39. ENCORE study. (Accessed 3-02-2016, at www.encorestudie.nl.)

40. Netherlands Cancer Registry - Colorectal cancer incidence. 2014. at http://www.cijfersoverkanker.nl/selecties/Incidentie_darmkanker/img54d0ec8382b2f.)

41. Winkels RM, Heine-Broring RC, van Zutphen M, et al. The COLON study: Colorectal cancer: Longitudinal, Observational study on Nutritional and lifestyle factors that may influence colorectal tumour recurrence, survival and quality of life. BMC cancer 2014;14:374.

R.R.J. Coebergh van den Braak*, L.B. van Rijssen*, J.J. van Kleef*, G.R. Vink, M. Berbee, M.I. van Berge Henegouwen, H.J. Bloemendal, M.J. Bruno, M.C. Burgmans, O.R.C. Busch, P.P.L.O. Coene, V.M.H. Coupé, J.W.T. Dekker, C.H.J. van Eijck, M.A.G. Elferink, F.L.G. Erdkamp, W.M.U. van Grevenstein, J.W.B. de Groot, N.C.T. van Grieken, I.H.J.T. de Hingh, M.C.C.M. Hulshof, J.N.M. IJzermans, L. Kwakkenbos, V.E.P.P. Lemmens, M. Los, G.A. Meijer, I.Q. Molenaar, G.A.P. Nieuwenhuijzen, M.E. de Noo, L.V. van de Poll-Franse, C.J.A. Punt, R.C. Rietbroek, W.W.H. Roeloffzen, T. Rozema, J.P. Ruurda, J.W. van Sandick, A.H.W. Schiphorst, H. Schippers, P.D. Siersema, M. Slingerland, D.W. Sommeijer, M.C.W. Spaander, M.A.G. Sprangers, H.B.A.C. Stockmann, M. Strijker, G. van Tienhoven, L.M. Timmermans, M.L.R. Tjin-a-Ton, A.M.T. van der Velden, M.J. Verhaar, H.M. Verkooijen, W.J. Vles, J. de Vos-Geelen, J.W. Wilmink, D.D.E. Zimmerman, M.G.H. van Oijen†, M. Koopman†, M.G.H. Besselink†, and H.W.M. van Laarhoven†, on behalf of the Dutch Pancreatic Cancer Group, Dutch Upper GI Cancer Group and PLCRC working group.

* These authors contributed equally; † These authors share senior co-authorship.

9

THE 3P INITIATIVE: THREE COMPREHENSIVE NATIONWIDE POPULATION-BASED CANCER COHORT STUDIES

ACTA ONCOLOGICA 2017

ABSTRACT

BACKGROUND

The increasing sub-classification of cancer patients due to more detailed molecular classification of tumors, and limitations of current trial designs, require innovative research designs. We present the design, governance and current standing of three comprehensive nationwide cohorts including pancreatic, esophageal/gastric, and colorectal cancer patients (NCT02070146). Multidisciplinary collection of clinical data, tumor tissue, blood samples, and patient-reported outcome (PRO) measures with a nationwide coverage, provides the infrastructure for future and novel trial designs and facilitates research to improve outcomes of gastrointestinal cancer patients.

MATERIAL AND METHODS

All patients aged ≥18 years with pancreatic, esophageal/gastric or colorectal cancer are eligible. Patients provide informed consent for: (1) reuse of clinical data; (2) biobanking of primary tumor tissue; (3) collection of blood samples; (4) to be informed about relevant newly identified genomic aberrations; (5) collection of longitudinal PROs; and (6) to receive information on new interventional studies and possible participation in cohort multiple randomized controlled trials (cmRCT) in the future.

RESULTS

In 2015, clinical data of 21,758 newly diagnosed patients were collected in the Netherlands Cancer Registry. Additional clinical data on the surgical procedures were registered in surgical audits for 13,845 patients. Within the first two years, tumor tissue and blood samples were obtained from 1507 patients; during this period, 1180 patients were included in the PRO registry. Response rate for PROs was 90%. The consent rate to receive information on new interventional studies and possible participation in cmRCTs in the future was >85%. The number of hospitals participating in the cohorts is steadily increasing.

CONCLUSION

A comprehensive nationwide multidisciplinary gastrointestinal cancer cohort is feasible and surpasses the limitations of classical study designs. With this initiative, novel and innovative studies can be performed in an efficient, safe, and comprehensive setting.

BACKGROUND

Patients with gastrointestinal cancer are traditionally treated according to several clinical and histopathological characteristics (e.g., tumor location, TNM stage, tumor grade). However, patients with similar traditional features, undergoing similar treatment, may show important differences in clinical outcome.[1-3] The underlying biological differences result in an increasing number of disease sub-classifications, based on efforts towards more individualized (or tailored) patient treatment.

However, due to the increasing sub-classifications of patients, there is a need for novel clinical trial designs and methods for data acquisition and patient recruitment. Current clinical trials have important limitations. First, recruitment is extremely restricted: only 5–10% of all patients are enrolled in a clinical trial.[4] Second, clinical trials include only highly selected patient populations, which leads to low inclusion rates and further limits their external validity. Third, data collection may be inadequate, due to insufficient follow-up or the absence of patient-reported outcomes (PROs); this, in turn, results in high costs or premature termination of the study. Finally, clinical trials are often underpowered for post hoc subgroup analyses.[4,5] Consequently, the current clinical trial system has been described as 'broken', 'in crisis', and 'not fit for purpose'.[6] However, there is a paucity of data on a population level. Current nationwide data initiatives such as the Surveillance, Epidemiology, and End Results (SEER) and the Medicare database, contain selected populations and do not include biobanking or data on quality of life.[7]

In an effort to address these issues, three comprehensive nationwide cohorts of pancreatic, esophageal/gastric, and colorectal cancer patients were started in the Netherlands (the Dutch PAncreatic CAncer Project 'PACAP'; the Prospective Observational Cohort study of Oesophageal-gastric cancer Patients 'POCOP'; and the Prospective Dutch ColoRectal Cancer cohort 'PLCRC').

Collaborating as the 3P initiative, these three cohorts collect clinical data, tumor tissue, blood samples, and PROs of gastrointestinal cancer patients. The goal is to facilitate research by (inter)national research groups to improve the survival and quality of life of patients with one of these three cancers. The protocol of PLCRC (describing the practical procedures and considerations of the cohort) was previously published.[8]

We present the design, proceedings, governance, opportunities, and pitfalls of the three collaborating comprehensive prospective nationwide gastrointestinal cancer cohorts.

MATERIAL AND METHODS

INCLUSION AND INFORMED CONSENT

All patients aged ≥18 years with pancreatic, esophageal/gastric, or colorectal cancer are eligible. Excluded from participation are patients with mental incompetence or insufficient understanding of the Dutch language to provide informed consent. Patients are asked to sign a multi-source informed consent including the following components: (1) reuse of clinical data from all medical files; (2) tissue sampling; (3) blood sampling; (4) to be informed about relevant newly identified genomic aberrations; (5) PROs; and (6) receiving information on new interventional studies and possible participation in cohort multiple randomized controlled trials (cmRCT) in the future. Patients can provide written informed consent for each component separately, and may alter or retract consent for each component at any point in time. As clinical data are crucial to the cohorts for obvious reasons, patients that do not give informed consent for this part of the study are considered ineligible.

Patients who want to consent to receiving information on new interventional studies in the future are informed in detail about the cmRCT design. They are informed: (i) that their data may be (re)used for the evaluation of new interventions offered to patients within the cohort; (ii) that they may in future be randomly selected for an experimental intervention, which they may accept or refuse at a later stage; and (iii) that when enrolled in the cmRCT, they cannot participate in other studies investigating the same intervention or outcome. This procedure is identical to current practice for classical RCTs. However, patients participating in one of the 3P cohorts, but who are not enrolled in a cmRCT, may participate in other studies (e.g., classical RCTs) outside of the cohorts.

CLINICAL DATA

Clinical data are obtained from the Netherlands Cancer Registry (NCR), hosted by the Netherlands Comprehensive Cancer Organization. The NCR contains clinical data from all relevant medical charts registered by trained data managers for every patient diagnosed with cancer in the Netherlands.[9] In 2015, the item set of the NCR was renewed and expanded to meet the requirements of the gastrointestinal cancer cohorts and to facilitate research. Items focus on patient, tumor and treatment characteristics, adverse events, and survival. Importantly, medical files are revisited multiple times to ensure the registration of clinical items from diagnosis until death. For every new cancer patient, 200-400 clinical data items are stored in an online secured database using Snowmed ontologies. A collaboration between the national tumor working groups, research groups, and the NCR has resulted in data sharing

initiatives allowing data from the NCR to be merged with other databases, e.g., for surgical audits. In these audits, oncologic surgeons collect data for a nationwide auditing initiative, supervised by the Dutch Institute for Clinical Auditing (DICA).[10] Participation in these audits is mandatory for each hospital. For each surgical patient, an additional set of 100–150 surgical clinical data items is collected.

Importantly, according to Dutch law, collection of clinical data in the NCR and surgical audits does not require informed consent; patients sign informed consent for reuse of these data (as described above).

TISSUE AND BLOOD SAMPLES

For POCOP and PACAP, tissue and blood sampling is organized in close collaboration with the Parelsnoer Institute, an existing national initiative facilitating biobanking for 17 different diseases, including esophageal/gastric and pancreatic cancer.[11,12] Fresh frozen tumor and normal tissue samples are taken from the surgical resection specimen of the primary tumor. Blood samples are withdrawn before and after surgery.

For colorectal cancer, the Parelsnoer Institute facilitates biobanking for hereditary cases. Therefore, patients enrolling in PLCRC cohort can consent to tumor and tissue collection and biobanking separately, in order to collect biomaterial of all cases of colorectal cancer.[8] Both fresh frozen and formalin- fixed paraffin embedded tumor and normal tissue samples are obtained from the surgical resection specimen of the primary tumor. Furthermore, blood samples may be collected as needed for specific study protocols. In PLCRC, the informed consent allows for the withdrawal of blood samples for future research questions without precisely specifying the time point of withdrawal, the patient population studied, and/or the specific tests that will be performed. There is a limit of 10 tubes per patient per year, collected only at the time of regular blood withdrawals. Details of the biobanking standard operating protocols are available in the **Supplementary Materials**.

PATIENT-REPORTED OUTCOMES

PROs, including health-related quality of life (HRQoL), are increasingly important outcomes for patients and physicians, and are also of growing interest to other healthcare partners. The PROs that are administered longitudinally were selected in close collaboration with national experts, international advisors, patients, and patient advocates. A core set of validated questionnaires is used to measure generic and disease-specific HRQoL (e.g., the EuroQol and European Organization for Research and Treatment of Cancer (EORTC) questionnaires).[13,14] In addition, a cohort-specific set of questionnaires is used to measure, e.g., self-reported adverse events and work

productivity. The composition of questionnaires is flexible and may be altered depending on the inclusion of new studies.

To increase patient participation and response rates, patients may complete questionnaires on paper, or online (computer, tablet or smartphone). Questionnaires are provided by the digital tracking system Patient-Reported Outcomes Following Initial treatment and Long-term Evaluation of Survivorship (PROFILES), a noncommercial initiative with an online patient management system used to send online or paper PROs and automatic reminders.[15]

DATA ACCESS AND INTEGRATION

As mentioned, the main goal of the 3P initiative is to facilitate (inter)national research by collecting and sharing high quality data. Every researcher (national and international) can use the data and biomaterial gathered to improve the outcome for patients with pancreatic, esophageal/gastric, and colorectal cancer. To ensure a sustainable and secure use of the data, a procedure to evaluate requests to access the data is in place (see: 'Governance' below). Furthermore, participating centers may at any time request data of patients enrolled at their own center. Besides the evaluation of the request, a generic and easy-to-use information technology (IT) infrastructure to facilitate the actual use of data is essential. The IT backbone of the three cohorts is based on the FAIR (Findable, Accessible, Interoperable, and Reusable) principles[16] and has been created in close collaboration with national and international research initiatives, such as the Dutch national node of the Biobanking and BioMolecular resources Research Infrastructure (BBMRI), Dutch Translational Research IT (TraIT), and the AACR (American Association for Cancer Research) project 'Genomics, Evidence, Neoplasia, Information, Exchange' (GENIE). As mentioned, the different data types are gathered through existing best practices (e.g., NCR, PROFILES). These data are combined using an IT solution in which the data types are matched through a unique study registration (USR) number which is assigned to each patient at enrollment. A separate enrollment log, only containing USR numbers with corresponding patient identifiers (name, date of birth, gender and date of inclusion), is stored on a different secured server to secure patients' identity. Data from the different sources are regularly added using data dumps. In the future, the databases can and will be enriched with data from other studies. Eventually all clinical, biological, and PROs data are integrated and made accessible in a secure way. This allows (among other tasks) to scrutinize the data for selection bias of the informed consent components, and attrition or responder bias in patients that did respond compared with those who did not respond to the PROs.

STATISTICAL ANALYSIS

Data were analyzed using IBM SPSS statistics version 21 (IBM, Armonk, NY, USA). Frequency tables were provided, and categorical data were presented as frequencies with percentages. No comparative analyses were performed.

PROSPECTIVE STUDIES

The 3P initiative provides the infrastructure for efficient, safe and comprehensive clinical evaluation of new interventions for patients with pancreatic, esophageal/gastric, and colorectal cancer based on classical observational and interventional clinical study designs, or on the cmRCT design (**Figure 1**).[17] Studies based on the latter design can be performed because clinical data are collected for all patients enrolled in the cohorts, including patients who will be randomized to the standard of care arm and do not need to be approached for informed consent at the time of a new cmRCT study.

Both cmRCT and multiple simultaneous prospective observational studies can be performed within the cohorts, as many variables and endpoints are collected in a standardized way. If required, the composition of clinical data and PROs can be altered to accommodate prospective studies. Additionally, data from the cohorts may be used for studies performed outside the 3P initiative.

Because clinical data are collected for all patients enrolled in the cohorts, the cohorts are well suited to serve as the basis for cmRCTs.[17] To enable cmRCT studies within the cohorts, patients are not only asked for informed consent for data collection and to be approached for future clinical trials, but are also specifically asked for future randomization according to the cmRCT design. To use the database for cmRCT to evaluate a new intervention, eligible patients within the cohort can be identified. A randomly selected subgroup will be offered the experimental intervention. The outcomes of these patients are compared to the routinely collected outcomes of eligible patients who were not randomly selected to receive the intervention (i.e., the control group). Patients not receiving the intervention will not be informed that they are serving as controls, as is explained at initial enrollment in the cohort. The intervention is described in a separate protocol that requires approval of the institutional review board. To avoid overlap with other studies, each center can decide if it wants to participate in a specific cmRCT. Patients who accept the intervention have to sign a separate informed consent.

GOVERNANCE

In order to obtain data, (inter)national researchers can file a study proposal using a pre-specified format which will be assessed by the appropriate tumor-specific

scientific committee (www.dpcg.nl, www.ducg.nl, and www.plcrc.nl). Submitted research proposals are then reviewed to determine feasibility and quality, and to ascertain possible duplicate studies. For POCOP and PACAP, study proposals that are approved by the scientific committee are subsequently presented to the Dutch Upper GI Cancer Group and the Dutch Pancreatic Cancer Group, respectively. Participating centers requesting data of patients enrolled at their own center can obtain data without following this procedure.

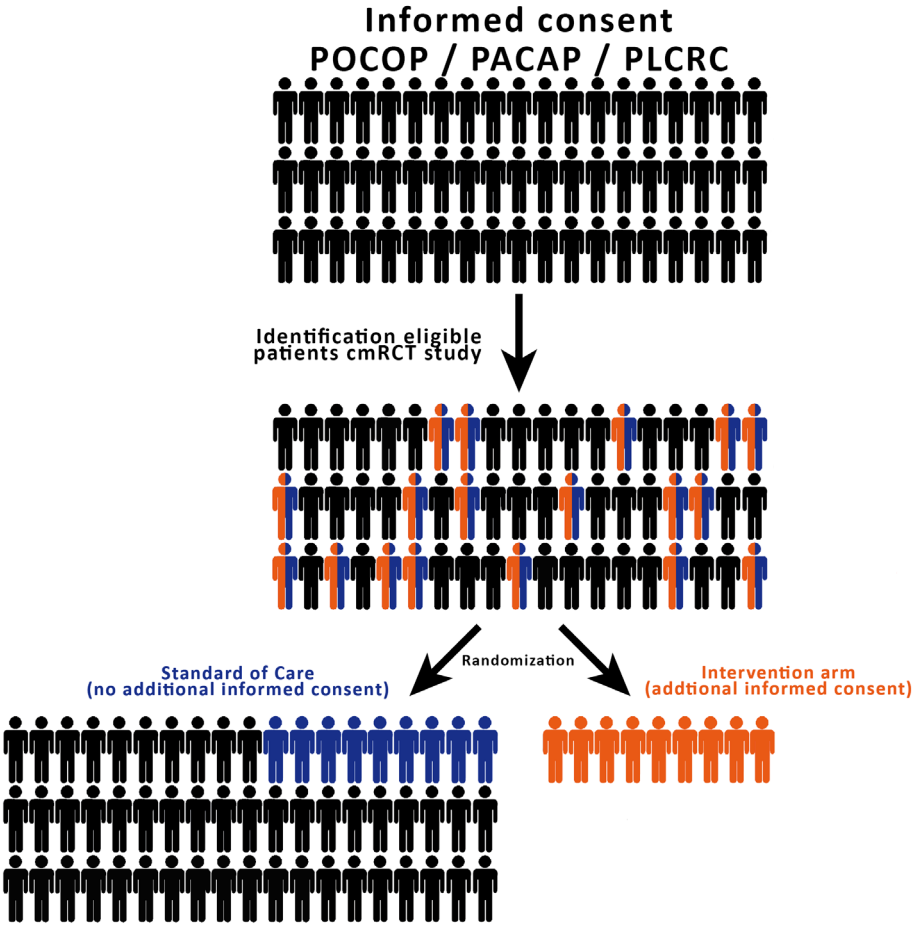


FIGURE 1. FLOWCHART OF THE COHORT MULTIPLE RANDOMIZED CLINICAL TRIAL (CMRCT) DESIGN
At enrollment in one of the three cohorts, patients can consent to be selected and randomized according to the cmRCT design. When a patient is eligible to enter a cmRCT trial, the patient is randomized. When a patient is randomized to the standard of care arm, no additional steps are undertaken. When a patient is randomized to the intervention arm, the patient is approached and offered the intervention, and signs an additional informed consent when participating in the trial.

The scientific committees meet 3–4 times a year and are composed of a multidisciplinary team (including basic scientists, social scientists, clinicians, and patient advocates), representatives from all participating centers, and representatives of the boards of the respective research groups. The scientific committees control the release of clinical data, PROs, and blood or tissue samples, to various healthcare partners such as government and industry (**Supplementary Figure 1**).

FUNDING

Financial support of the 3P initiative is based on ad hoc funding and structural funding. Ad hoc funding consists of public funds (e.g., the Dutch Cancer Society; The Netherlands Organization for Health Research, and Development) and public–private partnerships with pharmaceutical companies. These partnerships are increasingly popular to create the critical mass of partners in specific areas and allow to combine resources, expertise and complementary skills to advance the understanding of the factors underlying differences in the clinical outcome. Furthermore, these collaborations allow to develop drugs at a (possibly) lower cost and faster rate, and to evaluate the cost-effectiveness of new and costly medication.[18] Besides the financial support, the 3P initiative is also supported through data and knowledge-sharing, and access to information technology (IT) tools.

ETHICAL CONSIDERATIONS AND PRIVACY

Studies on the three cohorts are conducted in accordance with the principles of the Declaration of Helsinki [64th WMA (World Medical Association) General Assembly, Fortaleza, Brazil, October 2013] and in accordance with the Dutch Medical Research Involving Human Subjects Act.

RESULTS

In 2015, clinical data of all newly diagnosed patients with gastrointestinal cancer were collected within the NCR including 2,284 pancreatic, 3,925 esophageal/gastric, and 15,549 colorectal cancer patients. Additional clinical data regarding the surgical procedure (registered in the surgical audit) were available for patients who underwent surgery: 881 (39%) pancreatic, 1244 (32%) esophageal/gastric, and 11,720 (76%) colorectal cancer patients. Extensive data on clinical characteristics and data completeness are reported elsewhere.[8,19]

In an increasing number of participating hospitals, the informed consent procedure for PACAP, POCOP, and/or PLCRC has been implemented (n = 22, n = 16, and n = 12

centers, respectively, as at 1 December 2016). At time of manuscript acceptance, informed consent to collect tumor tissue and blood samples was obtained from 538 pancreatic, 199 esophageal/gastric, and 1,313 colorectal cancer patients. During this period, 309 pancreatic, 416 esophageal/gastric, and 1145 colorectal cancer patients were included in the PRO registry. Analysis showed that >90% of the patients who were informed, provided informed consent for one or more components. Of all patients that consented to receiving questionnaires, at baseline the response rate was 91%, whereas this decreased to 64% after 3 months and to 31% after 6 months. **Table 1** and **Figure 2** provide overviews of patients in the clinical data collection, the PROs, and the biobanking initiatives per tumor type. The decrease in response rate over time is partly because some patients died or dropped out between the two time points, or had not yet reached the 3-month or 6-month time point. Of all patients who completed at least one questionnaire, 54% completed the questionnaires online. In the first 67 PACAP patients, the median completion time was 40 (IQR 30, range 15–350) min, which was acceptable to most (70%) of the patients. Over 80% of patients were satisfied with the questionnaire and would participate in the cohort on a regular basis. In addition, most patients (80%) felt that physicians should pay more attention to HRQoL.

Consent to receive information about intervention studies and to participate in cmRCT studies in the future were provided by 94% of PACAP, 84% of POCOP, and by 85% of PLCRC patients.

TABLE 1. NUMBER OF PATIENTS IN THE NETHERLANDS CANCER REGISTRY, SURGICAL AUDITS, PATIENT-REPORTED OUTCOMES REGISTRIES, AND BIOBANKING

	PACAP PANCREAS	POCOP ESOPHAGEAL/ GASTRIC	PLCRC COLORECTAL	TOTAL
Netherlands Cancer Registry[a]	2,284	3,925	15,549	21,758
Surgical Audit[a,b]	881	1,244	11,720	13,845
Biobank	538	199	1313	2,050
PROs (eligible)	309 (506)	416 (675)	1145 (1575)	2,580 (2,756)
PROs response rate, t= 0 months	98%	95%	79%[c]	91%
PROs response rate, t= 3 months	63%	73%	63%[c]	64%

Informed consent was obtained from all patients in the patient-reported outcomes registry and biobanking, as collection of clinical data does not require informed consent.
[a] 2015 only.
[b] Number of registered resections.
[c] Date of inclusion until 31 August 2016.
PROs: patient-reported outcomes; PACAP: Dutch PAncreatic CAncer Project; POCOP: Prospective Observational Cohort study of esophageal-gastric cancer Patients; PLCRC: Prospective Dutch ColoRectal Cancer cohort; N/A: not applicable.

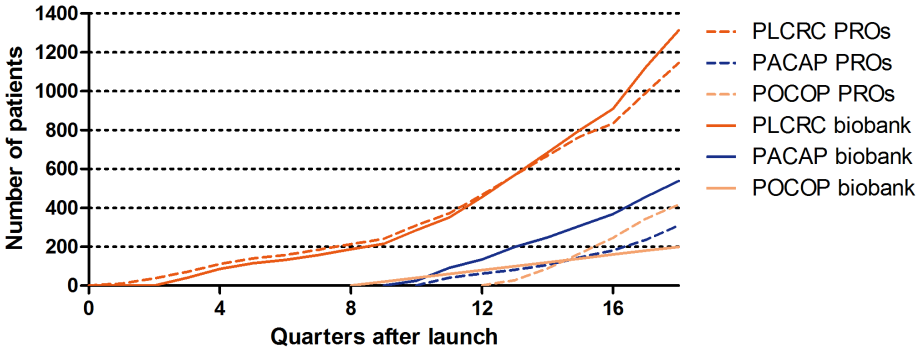


FIGURE 2. NUMBER OF SIGNED INFORMED CONSENTS OBTAINED, FOR THE COLLECTION OF BIOMATERIALS AND PROMS IN PATIENTS WITH COLORECTAL, ESOPHAGEAL/GASTRIC, AND PANCREATIC CANCER

The number of signed informed consents obtained for collection of biomaterials in patients with esophageal/gastric cancer was not available per quarter, and is therefore depicted as a total only.

DISCUSSION

The 3P initiative provides a comprehensive, nationwide, multidisciplinary research infrastructure that accommodates studies on a national level, providing population-based data. Extensive and accurate clinical data, tissue samples, blood samples, and PROs are collected from diagnosis until the death of patients with pancreatic, esophageal/gastric, and colorectal cancer after a broad-based informed consent has been given. The participation rate for each informed consent item was >80%, including consent to be informed about interventional studies and to participate in cmRCT studies in the future. The cohorts overcome many limitations of classical study designs and allow the performance of multiple concurrent studies. The collaborative nature of the 3P initiative combined with involvement of all relevant disciplines and mandated representatives of professional associations ensures a broad nationwide support.

Although many other clinical registries and biobank initiatives are available, only a few initiatives manage to combine both. The 3P initiative not only contains detailed longitudinal clinical data and biomaterial of patients, the PROs are collected and patients can easily be approached for future clinical trials. Based on collaboration with the NCR, which contains clinical data of all Dutch patients diagnosed with (pancreatic, esophageal/gastric, and colorectal) cancer, a nationwide coverage is ensured. Completeness of the NCR is reported to be at least 95%.[20] For comparison: although the SEER program in the USA has greater absolute numbers, coverage is

only 28% of the total US population.[7] Regarding the Nordic cancer registries (Denmark, Finland, Iceland, Norway, Sweden, and Faroe Islands), their national coverage is comparable to the NCR and is reported to be close to 100%.[21] Recognizing the importance of PROs, Sweden has also started to collect PROs prospectively over time and organized by tumor type.[22]

In the Netherlands, considerable experience has been obtained with surgical auditing, leading to case ascertainments of 95%, data completeness of almost 100%, and data accuracy of 95–99%.[10] The main equivalent for these audits is the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP)[23]; although the reported completeness and accuracy is lower compared to the Dutch audits[24], the absence of universal definitions or scoring systems hampers proper comparison. Therefore, a generic and easy-to-use IT infrastructure based on the FAIR (Findable, Accessible, Interoperable, and Reusable) principles is essential, but not easy to achieve.[16] Until now, many integration/sharing initiatives in (translational) research have failed to achieve full potential. This may be due to either focusing on technology push without sufficient user buy-in and content, or on supplying only technical solutions for one individual dataset thereby creating information silos instead of accessible data. The Handbook for Adequate Natural Data Stewardship (HANDS) published by the Netherlands Federation of University Medical Centers, illustrates the active attitude to break down these information silos and converge the current (inter)national ongoing efforts.[25]

Continuously evolving ethical and legal changes lead to more stringent criteria, lengthier protocols/patient information leaflets, and more informed consents for the use and even (retrospective) reuse of patient data and biomaterials. Between 1987 and 2007, the length of the informed consent documents has doubled, mostly due to formal components that aim to inform patients as fully as possible.[26] However, patients who receive brief/simple documents remember the information provided better than those who receive detailed/ lengthy information.[27] To maximize information retention of the informed consent procedure of the 3P initiative, information is provided through multiple sources including: the treating physician, a research nurse or physician assistant, study websites, an online patient movie, brochures, and small executive summary folders. Cervo et al. studied a similar multisource informed consent procedure and showed that these patients retain much more information (≥95% of the questions about the informed consent answered correctly)[28] compared to 56-88% without the provision of multisource information.[27]

In the Netherlands, the cmRCT design complies with the laws on human medical research and is becoming increasingly accepted in clinical practice. Nevertheless, the

design and the possible future impact on patients have provoked resistance in other countries. Patients consenting to randomization following the cmRCT design, may be included in an intervention study in the future.[17] Patients who consent to be approached for future investigations and are later selected for an cmRCT intervention arm, will receive additional detailed information and will be asked to sign a separate informed consent for the intervention. Patients in the control group previously consented at baseline and are fully aware that, when selected for the control group, they will not be informed about that particular cmRCT.

Although it remains debatable whether it is ethical not to offer the intervention to these (control) patients, this is no different from a classical RCT. A progressive, practice-changing agreement in the PLCRC cohort is that a maximum of 10 tubes per year may be withdrawn at regular blood withdrawals without the need to amend the study protocol to specify the timing, type of tube, and processing steps. This avoids multiple informed consents, while the assessment of study proposals from researchers or research groups by the tumor-specific scientific committees ensures scientific and ethical integrity. Although the 3P initiative has a nationwide coverage, intrinsic features of the study population (e.g., the distribution of age/gender/race, and what is considered the 'standard of care' in the Netherlands) may limit external international validity. However, the large number of included patients allows the selection of sufficiently large subgroups. Also, the standardized collection of data and biomaterials using international guidelines allows researchers to integrate data from multiple population-based registries, and to analyze differences in the standard of practice and subsequent clinical outcome.

A second limitation is that the clinical data in the NCR are only as accurate as the information provided in the relevant medical files. Therefore, synoptic and standardized reporting initiatives are ongoing.[29] A third limitation (or challenge) is the current dependency on ad hoc funding. This may result in additional costs for researchers if funding is insufficient to maintain the initiative, which might raise the threshold for researchers to make use of the cohorts. However, since the data and biomaterials are shared with multiple researchers, the financial contribution per research protocol will (if introduced) be lower than the costs for conducting each protocol separately. Importantly, retrospective observational studies using the available data can be performed without making a financial contribution. Nevertheless, structural financial support is preferable to ad hoc funding, which has been (in part) realized through public–private partnerships. Ideally, public funds may redirect part of their funding towards structural funding of longitudinal research initiatives, or the maintenance of longitudinal research initiatives may be considered part of daily clinical practice and be reimbursed as such.

CONCLUSIONS

The 3P initiative provides a comprehensive nationwide multidisciplinary research infrastructure to accommodate studies on a national level and complement the paucity of population based data. Three nationwide comprehensive cohorts for gastrointestinal cancer combine long-term clinical data, biobank material (including tissue and blood), and PROs. These are implemented using available best practices, internationally accepted standards for data collection, and a broad multi-step informed consent. Funding remains a challenge. Data from this initiative are accessible for further (inter)national research that aims to improve health outcomes for pancreatic, esophageal/gastric, and colorectal cancer patients.

REFERENCES

1. Bailey P, Chang DK, Nones K, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016;531:47-52.

2. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

3. Secrier M, Li X, de Silva N, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat Genet 2016;48:1131-41.

4. Lara PN, Jr., Higdon R, Lim N, et al. Prospective evaluation of cancer clinical trial accrual patterns: identifying potential barriers to enrollment. J Clin Oncol 2001;19:1728-33.

5. Bennette CS, Ramsey SD, McDermott CL, Carlson JJ, Basu A, Veenstra DL. Predicting Low Accrual in the National Cancer Institute's Cooperative Group Clinical Trials. J Natl Cancer Inst 2016;108.

6. DeVita VT, Jr. The clinical trials system is broken. Nat Clin Pract Oncol 2008;5:683.

7. Surveillance, Epidemiology, and End Results (SEER) Program. (Accessed 50 October 2016, at seer.cancer.gov.)

8. Burbach JP, Kurk SA, Coebergh van den Braak RR, et al. Prospective Dutch colorectal cancer cohort: an infrastructure for long-term observational, prognostic, predictive and (randomized) intervention research. Acta Oncol 2016;55:1273-80.

9. Cijfers over kanker. 2016. (Accessed 1th November at www.cijfersoverkanker.nl.)

10. Van Leersum NJ, Snijders HS, Henneman D, et al. The Dutch surgical colorectal audit. Eur J Surg Oncol 2013;39:1063-70.

11. Haverkamp L, Parry K, van Berge Henegouwen MI, et al. Esophageal and Gastric Cancer Pearl: a nationwide clinical biobanking project in the Netherlands. Dis Esophagus 2016;29:435-41.

12. Manniën J, Ledderhof T, Verspaget HW, et al. The Parelsnoer Institute: A National Network of Standardized Clinical Biobanks in the Netherlands. Open Journal of Bioresources 2017;4.

13. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.

14. EuroQol G. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.

15. van de Poll-Franse LV, Horevoorts N, van Eenbergen M, et al. The Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship registry: scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. Eur J Cancer 2011;47:2188-94.

16. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.

17. Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. BMJ 2010;340:c1066.

18. Mittleman B, Neil G, Cutcher-Gershenfeld J. Precompetitive consortia in biomedicine--how are we doing? Nat Biotechnol 2013;31:979-85.

19. van Rijssen LB, Koerkamp BG, Zwart MJ, et al. Nationwide prospective audit of pancreatic surgery: design, accuracy, and outcomes of the Dutch Pancreatic Cancer Audit. HPB (Oxford) 2017;19:919-26.

20. Schouten LJ, Jager JJ, van den Brandt PA. Quality of cancer registry data: a comparison of data provided by clinicians with those of registration personnel. Br J Cancer 1993;68:974-7.

21. Engholm G, Ferlay J, Christensen N, et al. NORDCAN--a Nordic tool for cancer information, planning, quality control and research. Acta Oncol 2010;49:725-36.

22. Nationella Kvalitetregister. (Accessed 1st June, 2017, at http://www.kvalitetsregister.se/englishpages/.)

23. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. Ann Surg 2009;250:363-76.

24. Epelboym I, Gawlas I, Lee JA, Schrope B, Chabot JA, Allendorf JD. Limitations of ACS-NSQIP in reporting complications for patients undergoing pancreatectomy: underscoring the need for a pancreas-specific module. World J Surg 2014;38:1461-7.

25. Handbook for Adequate Natural Data Stewardship (HANDS). (Accessed 1st December 2016, at http://data4lifesciences.nl/hands/handbook-for-adequate-natural-data-stewardship/.)

26. Berger O, Gronberg BH, Sand K, Kaasa S, Loge JH. The length of consent documents in oncological trials is doubled in twenty years. Ann Oncol 2009;20:379-85.

27. Davis TC, Holcombe RF, Berkel HJ, Pramanik S, Divers SG. Informed consent for clinical trials: a comparative study of standard versus simplified forms. J Natl Cancer Inst 1998;90:668-74.

28. Cervo S, Rovina J, Talamini R, et al. An effective multisource informed consent procedure for research and clinical practice: an observational study of patient understanding and awareness of their roles as research stakeholders in a cancer biobank. BMC Med Ethics 2013;14:30.

29. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Arch 2016;468:639-49.

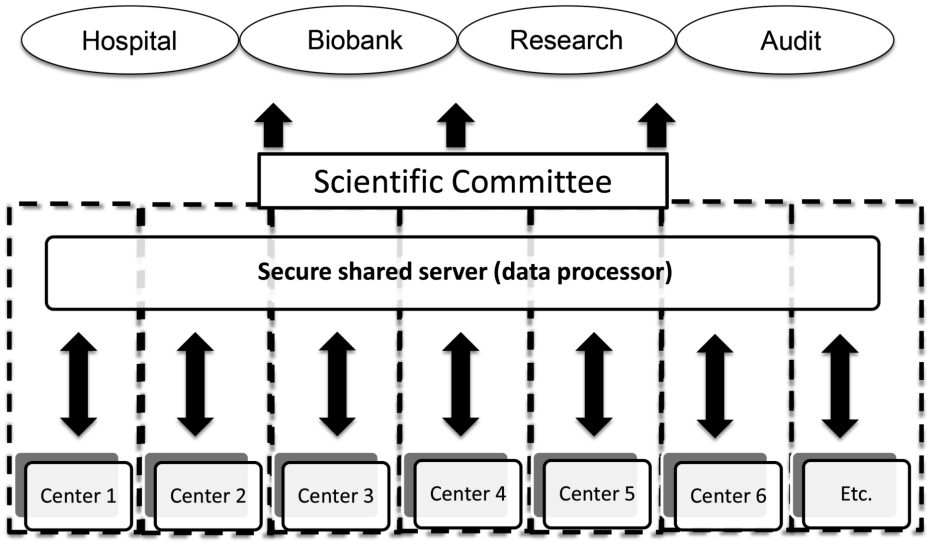
9

SUPPLEMENTARY DATA

SUPPLEMENTARY MATERIAL

SAMPLES COLLECTION WITHIN BIOBANKING INITIATIVES

In the Dutch Pancreas Biobank within the Parelsnoer Institute all patients with an indication for pancreatic surgery are included. Preoperatively, one 10 ml serum clot tube and one 10 EDTA plasma tube are collected. Blood is centrifuged, and serum and plasma are stored at -80°C in 0.5 ml aliquots. Moreover, for the purpose of DNA isolation, a second EDTA tube is collected preoperatively or the pellet of the first EDTA tube is used. At time of surgery the following samples are collected from the resection specimen: 2 samples of tumor tissue, 1 sample of normal pancreatic tissue and one sample of duodenum or spleen. If possible, also pancreatic (cyst) fluid is obtained. Postoperatively, one 10 ml serum clot tube and one 10 EDTA plasma tube are collected during the first postoperative visit to the out patients clinic, at 6 months and 12 months after resection, and in case of recurrence. All procedures within the Dutch Pancreatic and also Esophageal/Gastric Biobank are performed according to the Parelsnoer Institute Standard Operating Procedures (SOPs).[1]



SUPPLEMENTARY FIGURE 1. GOVERNANCE OF CLINICAL DATA, PATIENT-REPORTED OUTCOMES (PROS), AND TISSUE AND BLOOD SAMPLES

Ownership of the data remains with each individual center. All data are available to relevant stakeholders via an official request filed with the relevant tumor-specific scientific committee.

In the Dutch Esophageal/Gastric Biobank within the Parelsnoer Institute patients are included who undergo an upper gastrointestinal endoscopy with the taking of biopsies to confirm diagnosis and/or an esophagectomy or gastrectomy. During endoscopy, six additional biopsies are taken from the tumor for study purposes. Three samples are embedded in paraffin and three samples are snap frozen in isopentane on dry ice. Similarly, one paraffin and one snap frozen biopsy of normal mucosa are collected if possible. Paraffin biopsies are stored at room temperature and snap-frozen biopsies at −80°C. Blood samples (serum, EDTA plasma and EDTA blood) are preoperatively and postoperatively collected. 10 mL blood is obtained in order to take serum and EDTA plasma samples which are stored in 0.5 mL aliquots at −80°C. EDTA blood samples (10 mL) for DNA extraction, is stored after quality control at 4°C or lower. During surgery, six resection specimens are collected (three paraffin and three snap frozen in isopentane). In addition, normal tissue will be taken for storage. During follow-up, when biopsies are taken to histopathologically confirm recurrent or metastasized disease, two additional biopsies are taken for storage in the biobank.[2]

For PLCRC, the informed consent allows for the withdrawal of blood samples for future research questions without specifying exactly the time point of withdrawal, the studied patient population, and the specific tests that will be performed. There is a limit of ten tubes per patient per year, collected only at the time of regular blood withdrawals. Details on the alliquoting, quality assurance and sampling timing is defined per project.

REFERENCES

1 Manniën J, Ledderhof T, Verspaget HW, et al.. The Parelsnoer Institute: A National Network of Standardized Clinical Biobanks in the Netherlands. Open Journal of Bioresources. 2017;4(1):3. DOI: http://doi.org/10.5334/ojb.23

2 Haverkamp L, Parry K, van Berge Henegouwen MI, et al. Esophageal and Gastric Cancer Pearl: a nationwide clinical biobanking project in the Netherlands. Dis Esophagus. 2016 Jul;29(5):435-41.

10

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

GENERAL DISCUSSION

PREVENTION AND THE DETECTION OF EARLY STAGE COLORECTAL CANCER

Over the past decades, the knowledge of the genetic basis of cancer and tumour biology has increased exponentially.[1] In the metastatic setting, this data has translated to advances in patient stratification and the development of targeted therapies.[2-4] However, so far these efforts have not resulted in an effective adjuvant treatment of early stage colon cancer. Most recently published adjuvant trials have shown no benefit of targeted agents added to standard chemotherapy regimens.[5-8] To date, the only predictive marker in early stage colon cancer is MSI, which is used to select patients who do not benefit from adjuvant chemotherapy.[9, 10] The vast majority of colon cancer research is focused on metastasized disease, while only a minority is directed to early stage cancer. This uneven distribution of efforts may be an important factor in the lack of our understanding of early cancers at risk for disease recurrence. As an analogy, Bert Vogelstein used the focus of research in heart disease: "In cardiology, the focus over the past half century has been on prevention, not treating massive infarcts or strokes. In cancer, we've taken the opposite approach".[11]

Therefore, beside the current investment to optimize treatment of metastasized cancer disease, resources and intellectual energy should be aimed at prevention and early detection to detect disease when it is still at a curable stage, and optimization of curative treatment of early stage disease. In the Netherlands, the introduction of the Dutch bowel cancer screening program in 2014 is an important step towards early detection of (pre)malignant colorectal lesions. In the first year of the screening program, over 20,000 premalignant lesions and almost 2,500 cancers were detected in a cohort of roughly 865,000 individuals.[12] The early detection of (pre)malignant lesions will increase the proportion of patients who can be treated with curative intent using minimally invasive techniques such as endoscopic mucosal resection, laparoscopic bowel resection and transanal resection by endoscopic microsurgery.

PATIENT STRATIFICATION IN EARLY STAGE COLORECTAL CANCER

The expected shift towards earlier stages of colorectal cancer adds to the need for reliable criteria to identify which patients are at risk to develop recurrence of disease and which patients will benefit from (neo)adjuvant therapy. A deeper understanding of the key features of a tumour that define the clinical course of the disease is an important step towards the optimization of patient selection. However, the interpretation of the numerous biomarkers that are published must be done carefully keeping two important things in mind. First, associations between biomarkers and endpoints such as the proliferation or apoptotic rate and differentiation grade used

in basic research do not directly implicate a clinically relevant difference or effect in patients. Second, the use of heterogeneous patient cohorts in terms of clinically relevant features such as tumour stage and type of treatment makes it difficult to segregate the prognostic value of a biomarker from the prognostic value of known clinical/pathological features. Therefore, these kinds of studies should always be validated in homogeneous and clinically relevant patient cohorts. The findings in Chapter 4 illustrate this important issue, as the prognostic value of the CMS groups was less apparent in the homogeneous cohorts of lymph node negative colon cancer patients treated with surgery alone than in the heterogeneous cohorts used to posit the prognostic value of this gene signature.

Biomarkers should improve the clinical decision model to be clinically relevant. Therefore, comprehensive efforts are made to integrate well-defined and frequently recurring markers to quantify the added value to currently used decision tools. Recently, a multi-institutional effort to optimize prognostic stratification of stage II/III colon cancer by integrating molecular biomarkers (MSI status and KRAS/BRAF mutation) with clinical and pathological features was published.[13] Adding molecular biomarkers to a model with clinical and pathological features only marginally increased the performance of the model to predict overall survival in patients treated with adjuvant chemotherapy, while a significant increase in performance was observed in patients who were treated with surgery alone. Although these results should be interpreted with caution given the retrospective origin of most patients in the validation cohorts, this stratification effort has set the stage for similar efforts to validate these findings and come to a tailor-made treatment of patients with early stage colon cancer. It should be stressed that the differences in outcome underline the importance of validation in clinically relevant subgroups.

It is important to note that biological features occurring even at a low frequency may determine the metastatic potential of cancer cells, the response to therapy or the prognosis of a single patient, while no effects on population level may be measured. For instance, RET kinase inhibitors may be very effective in patients with a tumour carrying a RET fusion as these tumours are pan-negative for all other known driver mutations (Chapter 5).[14] Considerable hurdles must be overcome to implement technologies such as RNA sequencing to identify these low frequently occurring features in daily clinical practice. These hurdles include lowering the costs to sequence a low number of tissues and technical aspects such as sample preparation, benchmark standards and reproducibility.[15]

CONSENSUS MOLECULAR SUBTYPES

The CMS classification is currently the most robust and complete molecular

classification system for colorectal cancers with clear biological interpretability.[16] However, before this stratification system can be applied in a clinical setting additional information is needed. An important issue that should be addressed is the prognostic and predictive value of the CMS classification in clinically relevant subgroups, such as stage II and stage III colon cancer. In Chapter 4, we showed that CMS4 (the mesenchymal subtype) was more prevalent in advanced stages of colorectal cancer, and that it was associated with inadequate lymph node assessment in patients with stage II colon cancer. Subsequently, CMS4 only had a worse disease-free survival in the subgroup of patients with inadequate lymph node assessment, but did not hold significant prognostic value in patients in whom the number of assessed lymph nodes was considered adequate. Focusing on the predictive value of CMS, literature provides some hints. Patients with a CMS1 tumour (generally MSI and hypermutated) are expected not to respond to chemotherapy but may very well benefit from immunotherapy, as MSI predicted response to immunotherapy in patients with heavily pre-treated metastasised colorectal cancer. [2, 9, 10] Furthermore, patients with a CMS2 tumour were shown to be responsive to Oxaliplatin in a multi-agent regimen while mesenchymal tumours (CMS4) did not response to chemotherapy.[17, 18] However, future studies should be conducted to show a solid response monitoring per subtype given the retrospective nature of these studies and the fact that differences in prognosis and response to therapy are difficult to segregate in these patients. The CONNECTION-consortium (a nation-wide Colon Cancer Registry and Stratification effort) will initiate such a study on the role of subtypes on therapy response in a novel neoadjuvant setting to determine therapy efficacy on the individual subtypes. A second essential step towards the clinical application of CMS is the development of a 'CMS test', which can reliably distinguish the four different subtypes. The current gold standard to determine CMS is Affymetrix Array on fresh frozen samples, which is impractical, costly and labour intensive. The transition from fresh frozen to formalin-fixed paraffin embedded (FFPE) samples poses a real challenge as significant differences in Affymetrix array data between the two sample types can be expected and thus validation of a FFPE-based test (preferably using Affymetrix array expression data generated with the same FFPE sample) can prove to be difficult. Importantly, a 'CMS test' should distinguish each subtype and not one subtype, as each CMS group is likely to require a different therapeutic approach. Two techniques that have been explored to develop such as test are immunohistochemical staining and polymerase chain reactions (PCRs), although these attempts have not yet resulted in a validated paraffin-based 'CMS test' .[19, 20] Still, a 'CMS test' based on these techniques is most likely be implemented in a clinical setting as these techniques are already used by pathologists in daily clinical practice.

FUTURE PERSPECTIVES

TUMOUR HETEROGENEITY

As in most translational research on tumour tissues, this thesis focuses on inter tumour heterogeneity using one sample per tumour. However, intratumoral heterogeneity caused by the ongoing accumulation of genetic aberrations and selective pressure has become increasingly apparent.[21] Spatial intratumoral heterogeneity (i.e. the existence of multiple genetically distinct clones within one tumour) implies that the presence or absence of a biological feature in one sample may not reflect the presence or absence of that feature in all clones.[21] For instance, a recent study showed that lymphatic and distant metastases can arise from independent clones within the primary tumour, which may have implication for the prediction of the metastatic potential of a tumour.[22] In a clinical setting, patients may wrongfully be offered or withheld therapy if the decision is based on a single biopsy. Furthermore, the extent of intratumoral heterogeneity itself may be a relevant feature to take into account when treating and informing a patient.[23] In the future, multiple biopsies may be needed to determine biological features with acceptable accuracy. The evolutionary nature of cancer (temporal intratumoral heterogeneity) and the subsequent emerging therapy resistance as a result of selective pressure has major consequences for therapy efficacy and strategy.[24] The current dogma for the systemic treatment of cancer is to maximize cell death. However, this very strategy enables the outgrowth of resistant clones after eradication of treatment-sensitive clones.[25] Therefore, alternative strategies should be explored to control this mechanism of resistance. A possible alternative strategy is 'adaptive therapy', which aims to control the disease by allowing treatment-sensitive clones to exist at a stable level which in turn keep treatment-resistant clones stable as well.[26] Other possibilities to overcome therapy-induced resistance may lie in the inhibition of clonal dispersion to defer the emergence of resistance and in 'temporal collateral sensitivity', a transient state of the tumour during the development of resistance in which the tumour is (more) vulnerable to therapy.[27]

LIQUID BIOPSIES

The above-mentioned tumour heterogeneity encountered when using tumour biopsies has pushed research to develop new strategies to gain insight in the molecular tumour profile and the concurrent course of the disease. A promising source of biomarkers which is thought to be less sensitive to intratumoral heterogeneity is blood withdrawal or 'liquid biopsies'. Since the discovery of cell-free DNA in blood in 1948, numerous efforts have tried to harness liquid biopsies focusing

on circulating tumour cells, exosomes and circulating tumour DNA (ctDNA).[28] An important study of stage II colon cancer patients demonstrated that ctDNA analysis of blood samples taken after radical surgery can be used to define a population at very high risk of recurrence.[29] The MEDOCC-project, a collaboration between Dutch institutes with the John's Hopkins hospital aims to validate these findings in a large cohort of stage II colon cancer patients. In the future, patients with ctDNA after radical surgery for non-metastasised colon cancer may be randomized between follow-up or adjuvant chemotherapy to assess the predictive value of ctDNA. ctDNA analysis can also be used to detect recurrence of disease (possibly before a lesion is detectable with imaging), to evaluate a patient's response to therapy, and to guide adaptive therapy approaches is also being explored.[30, 31]

LATERALITY

Laterality as a factor to predict prognosis and response to therapy has gained considerable interest. Laterality is defined as the location of the primary tumour in the left or right colon with the splenic flexure as the demarcation between left and right. Both patients with and without distant metastases and a right sided tumour have worse prognosis compared to patients with a left sided tumour, even when the primary tumour is resected. Laterality was reported to be predictive in RAS and BRAF wild type metastatic colorectal cancer for the response to anti-EGFR therapy.[13, 32] Furthermore, factors such as MSI, BRAF and CMS are unequally distributed between left and right sided tumours.[16] This suggests that tumours located in the right and left hemi colon are two biologically separate entities, although none of the known biological features is observed in either left or right sided tumours. An important advantage of laterality to most biomarkers is the fact that the location of the primary tumour is known based on the preoperative colonoscopy and/or imaging. Laterality has been included in the most recent version of treatment guidelines, which will likely change the field of biomarker development. This is illustrated by the comprehensive analysis mentioned earlier showing that a model including amongst others laterality did not improve by the incorporation of MSI, KRAS and BRAF.

(INTER)NATIONAL COHORTS TO FACILITATE RESEARCH

The progressive understanding of tumour biology will lead to increasingly complex sub-classifications of cancer. Novel study designs and methods for data acquisition as well as changes in patient recruitment are needed to facilitate basic, translational and clinical research in the era of personalized medicine. A single centre approach will not solve the problem. Larger well-organized consortia are pivotal to improve our insight in tumour biology and improve cancer care. In the Netherlands, the Dutch

Prospective ColoRectal Cancer cohort ('PLCRC') was initiated to answer to these needs in the field of colorectal cancer by gathering clinical data, biomaterial and patient reported outcome measures in standardized way under a broad informed consent.[33, 34] A well-built IT infrastructure to store and unlock data according to the FAIR principles (Findable, Accessible, Interoperable and Reusable) is key to make these kind of initiatives successful.[35] Also, rewarding and recognizing the importance of data collecting and data sharing itself is important to create an incentive for researchers to share data. For instance, contributing 'data authors' could be listed in the primary publication, a status that should be recognized by all stakeholders and should be searchable in relevant sources such as Medline.[36] Such an incentive is needed to make current efforts on data sharing successful. In the Netherlands, the Handbook for Adequate Natural Data Stewardship published by the Netherlands Federation of University Medical Centres illustrates the active attitude to break down information silos and converge current (inter)national ongoing efforts such as PLCRC.[37] In the (near) future, intensified collaboration between research initiatives such as PLCRC, clinical audits such as the Dutch ColoRectal Audit and government driven programs such as the bowel cancer screening programme is key to successfully meet the problem of colon cancer. Furthermore, we envision that research and clinical practice will become an inseparable unity. Ultimately, research could be funded as part of the health care reimbursement system to secure structural research funding. With structural funding, hospitals can establish a research agency to lighten the load for clinicians.

To conclude, comprehensive multi-institutional efforts are needed to determine the added value of the available prognostic and predictive biomarkers in light of the current decision models, and to identify new clinically relevant biomarkers. These large-scale collaborations require an open and collaborative research community, which in turn requires change in attitude and recognition of currently underappreciated efforts such as data collection across all stakeholders. The role of tumour heterogeneity in tumour biology and its effect on research involving single biopsies will become increasingly important. Biomarkers such as ctDNA analysis in liquid biopsies will change the landscape of the detection of early and recurrent disease. Finally, molecular diagnostics will help to monitor actual disease burden and temporal heterogeneity, leading to a better understanding of the changing face of cancer and enabling a more successful approach in personalized cancer treatment.

REFERENCES

1. Ward RL, Hawkins NJ. Checking the scoreboard: the impact of cancer genetics on clinical practice. Intern Med J 2001;31:249-53.

2. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med 2015;372:2509-20.

3. Tol J, Koopman M, Cats A, et al. Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer. N Engl J Med 2009;360:563-72.

4. Battaglin F, Dadduzio V, Bergamo F, et al. Anti-EGFR monoclonal antibody panitumumab for the treatment of patients with metastatic colorectal cancer: an overview of current practice and future perspectives. Expert Opin Biol Ther 2017:1-12.

5. Alberts SR, Sargent DJ, Nair S, et al. Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: a randomized trial. JAMA 2012;307:1383-93.

6. Taieb J, Tabernero J, Mini E, et al. Oxaliplatin, fluorouracil, and leucovorin with or without cetuximab in patients with resected stage III colon cancer (PETACC-8): an open-label, randomised phase 3 trial. Lancet Oncol 2014;15:862-73.

7. Allegra CJ, Yothers G, O'Connell MJ, et al. Bevacizumab in stage II-III colon cancer: 5-year update of the National Surgical Adjuvant Breast and Bowel Project C-08 trial. J Clin Oncol 2013;31:359-64.

8. de Gramont A, Van Cutsem E, Schmoll HJ, et al. Bevacizumab plus oxaliplatin-based chemotherapy as adjuvant treatment for colon cancer (AVANT): a phase 3 randomised controlled trial. Lancet Oncol 2012;13:1225-33.

9. Des Guetz G, Schischmanoff O, Nicolas P, et al. Does microsatellite instability predict the efficacy of adjuvant chemotherapy in colorectal cancer? A systematic review with meta-analysis. Eur J Cancer 2009;45:1890-6.

10. Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol 2010;28:3219-26.

11. The ASCO Post. (Accessed 2017, at www.ascopost.com.)

12. Toes-Zoutendijk E, van Leerdam ME, Dekker E, et al. Real-Time Monitoring of Results During First Year of Dutch Colorectal Cancer Screening Program and Optimization by Altering Fecal Immunochemical Test Cut-Off Levels. Gastroenterology 2017;152:767-75 e2.

13. Dienstmann R, Mason MJ, Sinicrope FA, et al. Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. Ann Oncol 2017;28:1023-31.

14. Le Rolle AF, Klempner SJ, Garrett CR, et al. Identification and characterization of RET fusions in advanced colorectal cancer. Oncotarget 2015;6:28929-37.

15. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet 2016;17:257-71.

16. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.

17. Song N, Pogue-Geile KL, Gavin PG, et al. Clinical Outcome From Oxaliplatin Treatment in Stage II/III Colon Cancer According to Intrinsic Subtypes: Secondary Analysis of NSABP C-07/NRG Oncology Randomized Clinical Trial. JAMA Oncol 2016;2:1162-9.

18. 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.

19. Ubink I, Elias SG, Moelans CB, et al. A Novel Diagnostic Tool for Selecting Patients With Mesenchymal-Type Colon Cancer Reveals Intratumor Subtype Heterogeneity. J Natl Cancer Inst 2017;109.

20. Trinh A, Trumpi K, De Sousa EMF, et al. Practical and Robust Identification of Molecular Subtypes in Colorectal Cancer by Immunohistochemistry. Clin Cancer Res 2017;23:387-98.

21. Punt CJ, Koopman M, Vermeulen L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. Nat Rev Clin Oncol 2017;14:235-46.

22. Naxerova K, Reiter JG, Brachtel E, et al. Origins of lymphatic and distant metastases in human colorectal cancer. Science 2017;357:55-60.

23. Andor M, Tudor A, Paralescu S, et al. Methods for Sonic Representation of ST Depression During Exercise. Stud Health Technol Inform 2015;216:1041.

24. Enriquez-Navas PM, Wojtkowiak JW, Gatenby RA. Application of Evolutionary Principles to Cancer Therapy. Cancer Res 2015;75:4675-80.

25. Molinari F, Felicioni L, Buscarino M, et al. Increased detection sensitivity for KRAS mutations enhances the prediction of anti-EGFR monoclonal antibody resistance in metastatic colorectal cancer. Clin Cancer Res 2011;17:4901-14.

26. Gatenby RA, Silva AS, Gillies RJ, et al. Adaptive therapy. Cancer Res 2009;69:4894-903.

27. Zhao B, Sedlak JC, Srinivas R, et al. Exploiting Temporal Collateral Sensitivity in Tumor Clonal Evolution. Cell 2016;165:234-46.

28. Jia S, Zhang R, Li Z, et al. Clinical and biological significance of circulating tumor cells, circulating tumor DNA, and exosomes as biomarkers in colorectal cancer. Oncotarget 2017.

29. Tie J, Wang Y, Tomasetti C, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci Transl Med 2016;8:346ra92.

30. Scholer LV, Reinert T, Orntoft MW, et al. Clinical implications of monitoring circulating tumor DNA in patients with colorectal cancer. Clin Cancer Res 2017.

31. Garlan F, Laurent-Puig P, Sefrioui D, et al. Early evaluation of circulating tumor DNA as marker of therapeutic efficacy in metastatic colorectal cancer patients (PLACOL study). Clin Cancer Res 2017.

32. Boeckx N, Koukakis R, Op de Beeck K, et al. Primary tumor sidedness has an impact on prognosis and treatment outcome in metastatic colorectal cancer: results from two randomized first-line panitumumab studies. Ann Oncol 2017.

33. Burbach JP, Kurk SA, Coebergh van den Braak RR, et al. Prospective Dutch colorectal cancer cohort: an infrastructure for long-term observational, prognostic, predictive and (randomized) intervention research. Acta Oncol 2016;55:1273-80.

34. Coebergh van den Braak RRJ, van Rijssen LB, van Kleef JJ, et al. Nationwide comprehensive gastro-intestinal cancer cohorts: the 3P initiative. Acta Oncol 2017:1-8.

35. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.

36. Bierer BE, Crosas M, Pierce HH. Data Authorship as an Incentive to Data Sharing. N Engl J Med 2017;377:402.

37. Handbook for Adequate Natural Data Stewardship (HANDS). (Accessed 1st December 2016, at http://data4lifesciences.nl/hands/handbook-for-adequate-natural-data-stewardship/.)

11

SUMMARY

The research described in this thesis focuses on the identification and validation of molecular biomarkers to identify patients at risk of recurrence after curative surgery for early stage colon cancer. If we could realize to identify these patients at risk, a personalized approach would become more successful.

As many hospitals contribute to the daily care of patients with colorectal cancer it is important to develop standard operating procedures that are applicable in all these institutes when studying colorectal cancer. In **Chapter 2** we show that collecting high quality fresh frozen tissue samples of primary colorectal cancer is feasible in a multicentre setting. Over 90% of the tissue samples randomly selected from the MATCH study could be used for highly demanding techniques such as high-throughput RNA sequencing.

RNA sequencing data must be processed by several steps including a normalization step before any downstream analysis can be performed. Importantly, current normalization methods allow either the comparison within samples or between samples depending on whether gene length correction is or is not applied, respectively. In **Chapter 3**, we introduce GeTMM (Gene length corrected Trimmed Mean of M-values), a new normalization method that combines the individual advantages of two normalization methods to enable the execution of comparisons within and between samples using the same normalized data set. The analysis showed that the performance of GeTMM is not affected by poor RNA quality or GC-content. GeTMM outperforms the most commonly used normalization methods (EgdeR, DESeq2 and TPM) when comparing the normalized data with RT-qPCR mRNA expression data. Importantly, gene length correction impacted the prediction of the consensus molecular subtypes (CMS), an observation that should be considered when predicting CMS based on RNA sequencing data.

In **Chapter 4** we showed that CMS4 (the mesenchymal subtype) was found to be more prevalent in advanced stages of colorectal cancer, and that the number of assessed lymph nodes was often considered too low to reliably distinguish stage II from stage III disease. Lymph node assessment impacted the prognostic value of the CMS classification in aggregated cohort of patients with stage II colon cancer who did not receive adjuvant chemotherapy. CMS4 had a worse DFS in patients with inadequate lymph node assessment, but did not hold significant prognostic value in patients in whom the number of assessed lymph nodes was considered adequate. Combined these observations suggest substantial interconnectivity between tumour stage and tumour biology, and suggest that the prognostic value of the CMS classification is dependent on the heterogeneity of tumour stage in a patient cohort.

In **Chapter 5**, we present a systematic analysis of pathogenically relevant oncogenic fusions. The analysis yielded two known *BRAF* fusions and one novel *BRAF* fusion,

which were mutually exclusive with *BRAF* mutations. The analysis yielded four other relevant fusions: an *RRBP1-RET* fusion, an *EML4-NTRK3* fusion, an *USP9X-ERAS* fusion, and an *EIF3E-RSPO2* fusion. The tumour that harboured the *RET* fusion was pan-negative for the known driver mutations in colorectal cancer, suggesting that targeting the *RET* fusion may be very effective. The *ERAS* fusion was particularly interesting because the fusion was formed through a highly local chromothripsis event and because the ERAS protein normally is only expressed in embryogenic cells. The low frequency of R-spondin fusions in this well-defined cohort nuances the previously reported 10% recurrent R-spondin fusions in a smaller cohort of colorectal cancers. All fusions led to increased mRNA expression, and the fusions that were introduced in cell lines showed increased oncogenic activity, which both underlined the oncogenic potential of these fusions.

Chapter 6 describes our findings on the prognostic value of *SYK* and more specifically its splice variants *SYK(S)* and *SYK(L)*. The differential expression of *SYK total [SYK(T)]* and its splice variants between MSI and MSS tumours as well as between tumours with or without a *BRAF* and/or *PTEN* mutation suggested a different role for *SYK* in hypermutated and non-hypermutated tumours. High *SYK(S)* was associated with poor hepatic metastasis free survival in the patients with stage I-II colon cancer who were treated with surgery alone. Notably, the association was not confirmed in two independent, clinically less well-defined and smaller cohorts. Further research is warranted to elucidate the role of *SYK* and its splice variants in colorectal cancer.

In **Chapter 7** we validated a metastasis-specific microRNA signature, a combination of *let-7i* and *miR-10b* expression, in a cohort largely identical to the cohort used in Chapter 7. Furthermore, we identified *miR-30b* as an additional prognostic marker and showed that adding this microRNA to the signature (modified signature) improved the discriminative performance in this cohort to predict metastasis-free survival. Additional analysis showed *let-7i* expression to be mainly associated with cell adhesion, migration and motility, and the hedgehog, Wnt and TGF-β signalling pathways. *MiR-30b* expression is associated to axon guidance. Future studies should be conducted to further validate the modified signature.

In **Chapter 8** the outline of the Prospective Dutch ColoRectal Cancer cohort (PLCRC) is described. This national cohort study provides a multidisciplinary data, biobank, and patient-reported outcomes collection initiative to serve as an infrastructure for basic, translational and clinical research aiming to improve the outcome of patients with colorectal cancer. Clinical data are obtained from the Netherlands Cancer Registry, which was revisited and expanded to facilitate research. Tissue and blood samples are gathered in collaboration with the local pathology and chemistry labs, while the questionnaires are sent out and received by PROFILES, a researcher driven

national initiative on quality of life. Patients can optionally provide consent to tissue sampling and storage, blood withdrawals, receiving questionnaires on quality of life and being contacted for interventional studies in the future. This unique initiative will allow the selection of homogeneous cohorts of patients to execute studies focusing amongst others on the identification and validation of molecular biomarkers to identify patients at risk of recurrence after curative surgery for early stage colon cancer.

Chapter 9 describes the collaboration between PLCRC and two similar initiatives in the field of pancreatic (PACAP) and oesophageal/gastric cancer (POCOP): the 3P initiative. We underlined the importance and inevitability of intensive collaboration between best practices for data registration, quality of life questionnaires and IT research infrastructure to advance cancer research. We discussed the opportunities that arise from these cohorts such as the execution of studies according to innovative study designs like the cohort multiple randomized controlled trial. We also displayed the challenges we encountered when building and maintaining these initiatives such as structural funding and the design of a broadly supported governance structure that is both agile and robust. Lastly, it highlights some of the progress made in the first years of PLCRC such as the agreement with the institutional review board on the withdrawal of a maximum of 10 blood tubes for future research questions without further specifications to better facilitate observational studies involving, for instance, liquid biopsies. By realizing a shorter inclusion period of a large number of patients and the collection of standardized data and tissue, studies will be more successfully and accelerate the progress of cancer research.

12

SUMMARY IN DUTCH / NEDERLANDSE SAMENVATTING

De studies in dit proefschrift richten zich op de identificatie en validatie van moleculaire biomarkers waarmee identificatie mogelijk is van patiënten met een vroeg stadium coloncarcinoom die na een curatieve resectie een hoog risico lopen op het ontwikkelen van recidief ziekte. Dergelijke biomarkers zullen het mogelijk maken om therapie op maat succesvoller toe te passen.

Darmkanker wordt in veel ziekenhuizen behandeld. Bij het doen van onderzoek naar darmkanker is het belangrijk om standaardprocedures te ontwikkelen die in ieder ziekenhuis kunnen worden toegepast. In **Hoofdstuk 2** laten wij zien dat het snel invriezen van vers tumormateriaal in veel ziekenhuizen goed mogelijk is met behoud van hoogwaardige kwaliteit van het materiaal. Ruim 90% van de willekeurig onderzochte biopten uit het MATCH-onderzoek waren van zeer hoge kwaliteit waardoor het toepassen van onderzoekstechnieken zoals high throughput RNA sequencing mogelijk zijn.

RNA sequencing data moet volgens een aantal stappen inclusief een normalisatiestap worden verwerkt voordat verdere analyses kunnen worden verricht. De huidige normalisatiemethoden maken het mogelijk om verschillen in genexpressie binnen één biopt óf verschillen in genexpressie tussen biopten te onderzoeken afhankelijk van het wel of niet corrigeren voor genlengte. In **Hoofdstuk 3** introduceren wij GeTMM (Gene length corrected Trimmed Mean of M-values), een nieuwe normalisatiemethode waarmee beide type analyses kunnen worden uitgevoerd door het combineren van de voordelen van twee normalisatiemethoden. De analyses lieten zien dat de uitkomsten van GeTMM niet worden beïnvloed door RNA kwaliteit of GC-inhoud. Bovendien zijn de uitkomsten beter dan de meest gebruikte normalisatiemethoden (EdgeR, DESeq2 en TPM), wanneer de genormaliseerde data vergekeken wordt met RT-qPCR expressiedata. Genlengte correctie bleek een belangrijke invloed te hebben op het voorspellen van de CMS classificatie, een belangrijke bevinding die meegewogen moet worden als CMS groepen worden voorspeld op basis van RNA sequencing data.

In **Hoofdstuk 4** laten we zien dat CMS4 (het mesenchymale subtype) vaker voor komt in meer gevorderde stadia van het colorectaal carcinoom, en dat het aantal onderzochte klieren in de onderzochte patiënten vaak te laag was om betrouwbaar stadium II van stadium III te onderscheiden. Het aantal onderzochte lymfklieren had invloed op de prognostische waarde van de CMS classificatie in het samengesteld cohort van patiënten met stadium II coloncarcinoom die geen adjuvante chemotherapie hebben ontvangen. CMS4 was alleen geassocieerd met een slechtere overleving in die patiënten waarbij te weinig klieren waren onderzocht, en niet in de patiënten waarbij voldoende lymfklieren onderzocht waren. De resultaten van dit hoofdstuk suggereren aanzienlijke verwevenheid van

tumorstadium en tumorbiologie, en suggereren dat de prognostische waarde van CMS afhankelijk is van de heterogeniteit in een patiëntencohort ten aanzien van tumorstadium.

In **Hoofdstuk 5** presenteren wij een systematische analyse naar pathogenetisch relevante oncogene fusiegenen. Hierbij werden twee bekende *BRAF* fusies en één nieuwe *BRAF* fusie gevonden in tumoren die geen *BRAF*-mutatie hadden. Daarnaast werden er vier andere relevante fusies gevonden: een *RRBP1-RET* fusie, een *EML4-NTRK3* fusie, een *USP9X-ERAS* fusie en een *EIF3E-RSPO2* fusie. De tumor die de *RET*-fusie bevatte, had geen andere 'driver' mutaties waardoor de fusie een aantrekkelijk doelwit voor gerichte therapie is. De *ERAS*-fusie was interessant, omdat de fusie tot stand was gekomen door chromothripsis en omdat het *ERAS* eiwit normaal alleen tot expressie komt in embryonale cellen. De lage frequentie van R-spondin fusies in dit goed gedefinieerde patiënten cohort nuanceert de eerder gerapporteerde incidentie van 10% in een ander, kleiner cohort van patiënten met darmkanker. Alle fusies resulteerden in een verhoogde mRNA expressie, en fusies die werden geïntroduceerd in cellijnen lieten een verhoogde oncogene activiteit zien wat het oncogene potentieel van de fusies onderstreepte.

Hoofdstuk 6 beschrijft de bevindingen betreffende de prognostische waarde van *SYK* en de *SYK* splice varianten *SYK(S)* en *SYK(L)*. De expressieverschillen tussen MSI en MSS tumoren, en tussen tumoren met en zonder een BRAF en/of PTEN mutatie suggereerde een verschillende rol voor totaal *SYK [SYK(T)]* en de *SYK* splice varianten in hypergemuteerde en niet-hypergemuteerde tumoren. Verder was hoge expressie van *SYK(S)* geassocieerd met een slechte levermetastasevrije overleving bij patiënten met stadium I-II coloncarcinoom die alleen met een operatie werden behandeld. Deze associatie werd niet bevestigd in twee kleinere, onafhankelijke, minder goed gedefinieerde cohorten. Verder onderzoek naar de rol van *SYK* en de splice varianten bij het colorectaal carcinoom is nodig.

In **Hoofdstuk 7** beschrijven wij de validatie van een metastase-specifieke signatuur, een combinatie van *Let-7i* en *miR-10b* expressie, in een cohort dat vrijwel identiek was aan het cohort dat gebruikt werd in hoofdstuk 7. Daarnaast werd *miR-30b* geïdentificeerd als additionele prognostische marker. Het toevoegen van deze marker aan de bestaande signatuur (gemodificeerde signatuur) vergrootte het onderscheidend vermogen van de signatuur wat betreft de metastasevrije overleving in dit cohort. Een additionele analyse liet zien dat *Let-7i* expressie geassocieerd was met celadhesie, celmigratie en celmotiliteit, en met de hedgehog, Wnt en TGF-b signaaltransductieroutes. *MiR-30b* expressie was geassocieerd met axon geleiding. Vervolgstudies zullen moeten worden uitgevoerd om de prognostische waarde van de gemodificeerde signatuur te valideren.

In **hoofdstuk 8** wordt het studieprotocol van het Prospectief Landelijk ColoRectaal Carcinoom cohort (PLCRC) samengevat. Dit landelijke cohortonderzoek is een onderzoeksinitiatief dat gestructureerd data, weefsel, en patiëntrapporteerde uitkomsten verzameld van patiënten met een colorectaal carcinoom om de uitkomsten van patiënten met deze ziekte te verbeteren. De klinische data worden opgevraagd bij de Nederlandse kankerregistratie, die werd gereviseerd en uitgebreid om onderzoek te kunnen faciliteren. Het verzamelen van weefsel en bloed gebeurd in samenwerking met de lokale pathologie- en chemielaboratoria. De vragenlijsten worden uitgestuurd en ingenomen door PROFIEL, een nationaal, vanuit onderzoek gedreven initiatief op het gebied van kwaliteit van leven. Patiënten kunnen optioneel toestemming geven voor de verzameling en opslag van weefsel, bloedafnames, het ontvangen van vragenlijsten over kwaliteit van leven en het in de toekomst uitgenodigd worden voor interventiestudies. Dit unieke initiatief maakt het mogelijk homogene patiëntcohorten te selecteren voor bijvoorbeeld de identificatie en validatie van moleculaire biomarkers waarmee patiënten met een vroeg stadium coloncarcinoom die recidief ziekte zullen ontwikkelen na curatieve chirurgie kunnen worden geïdentificeerd.

Hoofdstuk 9 beschrijft de samenwerking tussen PLCRC en twee vergelijkbare initiatieven op het gebied van het pancreascarcinoom (PACAP), en het slokdarm- en maagcarcinoom (POCOP): het 3P initiatief. In dit hoofdstuk gaven wij het belang en de onvermijdelijkheid aan van intensieve samenwerking tussen organisaties op het gebied van dataregistratie, kwaliteit van leven vragenlijsten en IT onderzoeks-infrastructuur om kankeronderzoek te bevorderen. We bespraken de mogelijkheden die deze cohorten bieden zoals het uitvoeren van een onderzoek volgens innovatieve onderzoeksmethoden zoals het 'cohort multiple randomized controlled trial'. Daarnaast bespraken we de uitdagingen zoals het verkrijgen van structurele financiering en het ontwerpen van een robuuste, breed ondersteunde bestuurs-structuur die mee kan evolueren met de zich ontwikkelde platformen. Daarnaast werd de progressie besproken die gemaakt is in de eerste jaren van PLCRC zoals de overeenkomst met de METC over het afnemen van bloed om observationeel onderzoek waarbij bloedafnames worden gebruikt beter te faciliteren. Na het geven van toestemming voor PCLRC mogen er per jaar 10 buizen worden afgenomen in het kader van onderzoek zonder dat hiervoor opnieuw toestemming moet worden gevraagd bij de METC. Het 3P initiatief zal naar verwachting de inclusieperiode voor onderzoeken verkorten waardoor snellere resultaten worden behaald en de voortgang in het kankeronderzoek wordt versneld.

APPENDICES

LIST OF PUBLICATIONS

Coebergh van den Braak RR, van der Elst M, Scheffers J, Heitmann M. [Splenic rupture not always painful: diagnostics after blunt abdominal trauma]. *Ned Tijdschr Geneeskd 2013*.

Ebbink BJ, Brands MM, van den Hout JM, Lequin MH, **Coebergh van den Braak RR**, van de Weitgraven RL, Plug I, Aarsen FK, van der Ploeg AT. Long-term cognitive follow-up in children treated for Maroteaux-Lamy syndrome. *J Inherit Metab Dis 2016*.

Lalmahomed ZS, Bröker ME, van Huizen NA, **Coebergh van den Braak RR**, Dekker LJ, Rizopoulos D, Verhoef C, Steyerberg EW, Luider TM, IJzermans JN. Hydroxylated collagen peptide in urine as biomarker for detecting colorectal liver metastases. *Am J Cancer Res 2016*.

Coebergh van den Braak RR, Martens JW, IJzermans JN. CDX2 as a Prognostic Biomarker in Colon Cancer. *N Engl J Med 2016*.

Coebergh van den Braak RR, Hartholt KA, Dekker JW. Necrotic lesions of the caecum: a rare cause of right iliac fossa pain. *BMJ Case Rep 2016*.

Burbach JP*, Kurk SA*, **Coebergh van den Braak RR**, Dik VK, May AM, Meijer GA, Punt CJ, Vink GR, Los M, Hoogerbrugge N, Huijgens PC, Ijzermans JN, Kuipers EJ, de Noo ME, Pennings JP, van der Velden AM, Verhoef C, Siersema PD, van Oijen MG, Verkooijen HM, Koopman M. Prospective Dutch colorectal cancer cohort: an infrastructure for long-term observational, prognostic, predictive and (randomized) intervention research. *Acta Oncol 2016*.

Büttner S, Lalmahomed ZS, **Coebergh van den Braak RR**, Hansen BE, Coene PP, Dekker JW, Zimmerman DD, Tetteroo GW, Vles WJ, Vrijland WW, Fleischeuer RE, van der Wurff AA, Kliffen M, Torenbeek R, Meijers JH, Doukas M, IJzermans JNM. Completeness of Pathology Reports in Stage II Colorectal Cancer. *Acta Chir Belg 2017*.

Coebergh van den Braak RR, Hartholt KA, Pannekoek BJ, Smedts F, van der Elst M. Solitary Fibrous Tumors of the Pleura – current diagnostic tools. *Am Surg 2017*.

van der Vlugt JJ, van der Meulen JJ, **Coebergh van den Braak RR**, Vermeij-Keers C, Horstman EG, Hovius SE, Verhulst FC, Wierdsma AI, Lequin MH, Okkerse JM. Insight into the pathophysiological mechanisms behind cognitive dysfunction in trigonocephaly. *Plast Reconstr Surg 2017*.

van Vugt JLA, **Coebergh van den Braak RRJ**, Schippers HJW, Veen KM, Levolger S, de Bruin RWF, Koek M, Niessen WJ, IJzermans JNM, Willemsen FEJA. The effect of contrast-enhancement on skeletal muscle mass and skeletal muscle density measurements on computed tomography. *Clin Nutr 2017*.

van Dam PJ, van der Stok EP, Teuwen LA, Van den Eynden GG, Illemann M, Frentzas S, Majeed AW, Eefsen RL, **Coebergh van den Braak RRJ**, Lazaris A, Fernandez MC, Galjart B, Laerum OD, Rayes R, Grünhagen DJ, Van de Paer M, Sucaet Y, Mudhar HS, Schvimer M, Nyström H, Kockx M, Bird NC, Vidal-Vanaclocha F, Metrakos P, Simoneau E, Verhoef C, Dirix LY, Van Laere S, Gao ZH, Brodt P, Reynolds AR, Vermeulen PB. International consensus guidelines for scoring the histopathological growth patterns of human liver metastasis. *BJC 2017*.

van Vugt JLA, Buettner S, Levolger S, **Coebergh van den Braak RRJ**, Suker M, Gaspersz MP, de Bruin RWF, Verhoef C, van Eijck CHC, Bossche N, Groot Koerkamp B, IJzermans JNM. Sarcopenia is Associated with Increased Hospital Expenditure in Patients Undergoing Cancer Surgery of the Alimentary Tract. *Plos One 2017*.

Lalmahomed ZS, **Coebergh van den Braak RR**, Oomen MH, Arshad SP, Riegman PH, IJzermans JN; MATCH study working group. Multicenter fresh frozen tissue sampling in colorectal cancer: does the quality meet the standards for state of the art biomarker research? *Cell Tissue Bank 2017*.

Kloosterman WP*, **Coebergh van den Braak RRJ***, Pieterse M, van Roosmalen MJ*, Sieuwerts AM*, Stangl C, Brunekreef R, Lalmahomed ZS, Ooft S, van Galen A, Smid M, Lefebvre A, Zwartkruis F, Martens JWM, Foekens JA, Biermann K, Koudijs MJ, Ijzermans JNM†, Voest EE†. A Systematic Analysis of Oncogenic Gene Fusions in Primary Colon Cancer. *Cancer Res 2017*.

Coebergh van den Braak RRJ*, van Rijssen LB*, van Kleef JJ*, Vink GR, Berbee M, van Berge Henegouwen MI, Bloemendal HJ, Bruno MJ, Burgmans MC, Busch ORC, Coene PPLO, Coupé VMH, Dekker JWT, van Eijck CHJ, Elferink MAG, Erdkamp FLG, van Grevenstein WMU, de Groot JWB, van Grieken NCT, de Hingh IHJT, Hulshof MCCM, Ijzermans JNM, Kwakkenbos L, Lemmens VEPP, Los M, Meijer GA, Molenaar IQ, Nieuwenhuijzen GAP, de Noo ME, van de Poll-Franse LV, Punt CJA, Rietbroek RC, Roeloffzen WWH, Rozema T, Ruurda JP, van Sandick JW, Schiphorst AHW, Schipper H, Siersema PD, Slingerland M, Sommeijer DW, Spaander MCW, Sprangers MAG, Stockmann HBAC, Strijker M, van Tienhoven G, Timmermans LM, Tjin-A-Ton MLR, van der Velden AMT, Verhaar MJ, Verkooijen HM, Vles WJ, de Vos-Geelen JMPGM, Wilmink JW, Zimmerman DDE, van Oijen MGH†, Koopman M†, Besselink MGH†, van

Laarhoven HWM†; Dutch Pancreatic Cancer Group, Dutch Upper GI Cancer Group and PLCRC working group. Nationwide comprehensive gastro-intestinal cancer cohorts: the 3P initiative. *Acta Oncologica 2017.*

Coebergh van den Braak RRJ, Sieuwerts AM, Kandimalla R, Lalmahomed ZS, Bril SI, van Galen A, Smid M, Biermann K, van Krieken JHJM, Kloosterman WP, Foekens JA, Goel A, Martens JWM, IJzermans JNM; MATCH study group. High mRNA expression of splice variant SYK short correlates with hepatic disease progression in untreated lymph node negative colon cancer patients. *Plos One 2017*.

Coebergh van den Braak RRJ, Lalmahomed ZS, Büttner S, Hansen BE, Ijzermans JNM; MATCH Study Group. Nonphysician clinicians in the follow up of resected patients with colorectal cancer. *Dig Dis 2018*.

M. Smid*, **R.R.J. Coebergh van den Braak***, H. van de Werken, J. van Riet, A. van Galen, V. de Weerd, M. van der Vlugt-Daane, S.I. Bril, Z.S. Lalmahomed, W.P. Kloosterman, S.M. Wilting, J.A. Foekens, J.N.M. IJzermans, J.W.M. Martens†, A.M. Sieuwerts†. Gene length corrected Trimmed Mean of M-values (GeTMM) improves RNA-seq data processing for intra- and intersample comparisons. *Submitted.*

R.R.J. Coebergh van den Braak, A.M. Sieuwerts, Z.S. Lalmahomed, M. Smid, S.M. Wilting, S.I. Bril, X. Xiang, M. Daane, V. de Weerd, A. van Galen, K. Biermann, J.H. van Krieken, W.P. Kloosterman, J.A. Foekens, J.W.M. Martens, J.N.M. IJzermans. Confirmation of a metastasis-specific microRNA signature in primary colon cancer. *Submitted*.

Jeroen L.A. van Vugt, **Robert R.J. Coebergh van den Braak**, Zarina S. Lalmahomed, Peter P.L.O. Coene, Jan W.T. Dekker, David D.E. Zimmerman, Wouter J. Vles, Wietske W. Vrijland, Jan N.M. IJzermans. The Short and Long Term Impact of Low Skeletal Muscle Mass and Density after Resection of Stage I-III Colorectal Cancer: Results from a Prospective Multicenter Observational Cohort Study. *Submitted*.

Joris J.B. van der Vlugt, **Robert R.J. Coebergh van den Braak**, Jacques J.M.N. van der Meulen, Steven E.R. Hovius, Frank C. Verhulst, Jolanda M.E. Okkerse, A.I. Wierdsma, Steven A. Kushner, Markus Klimek. Prolonged surgical time in open craniofacial surgery: detrimental for cognitive functioning?. *Submitted*.

N.A. van Huizen, **R.R.J. Coebergh van den Braak**, M. Doukas, L.J.M. Dekker, J.N.M. IJzermans, T.M. Luider. Distribution of Collagen in Colorectal Liver Metastasis and Normal Liver tissue. *Submitted*.

PHD PORTFOLIO

Name PhD student:	Robertus Rudolphus Johannes Coebergh van den Braak
Erasmus MC department:	Surgery
PhD period:	Okt 2013 – Dec 2017
Promoter:	Prof. dr. J.N.M. IJzermans
Copromoter:	Prof. dr. J.W.M. Martens
Date thesis defence:	31 January 2018

COURSES (12.6 ECTS)

		ECTS
2013	Basic introduction course on SPSS	1.0
2014	Cursus wetenschappelijke integriteit	0.6
	Biostatistics for clinicians (NIHES Erasmus Winter Program)	0.7
	Regression analysis for clinicians (NIHES Erasmus Winter Program)	1.4
	Survival analysis for clinicians (NIHES Erasmus Winter Program)	1.9
	BROK-cursus	1.5
	An Introduction to the analysis of the next-generation sequencing data	1.4
2015	Basic and Translational Oncology	1.8
	Graphpad course	0.3
	Biomedical English Writing Course for MSc and PhD-students	2.0

ORAL PRESENTATIONS (11 ECTS)

		ECTS
2014	Wetenschapsdag Heelkunde Erasmus MC - '*De toegevoegde waarde van protocollaire histopathologische verslaglegging bij stadium II coloncarcinomen*'	1.0
2015	Najaarsvergadering Nederlande Vereniging voor Heelkunde - '*Het effect van intensieve follow-up na in opzet curatieve oncologische colonresectie*'	1.0
	Najaarsvergadering Nederlande Vereniging voor Heelkunde - '*De voorspellende waarde van histopathologische risicofactoren voor hoog risico stadium II coloncarcinoom*'	1.0
	Wetenschapsdag Heelkunde Erasmus MC - '*Een gehydroxyleerd collageen peptide in de urine als biomarker voor de detectie van colorectale levermetastasen*'.	1.0
	Nel Kreeft prijs; prijs voor beste voordracht.	

2016	Voorjaarsvergadering Nederlandse Vereniging voor Heelkunde - '*Een gehydroxyleerd collageen peptide in de urine als biomarker voor de detectie van colorectale levermetastasen*'	1.0
	Najaarsvergadering Nederlande Vereniging voor Gastro-Enterologie - '*Next generation complete cancer exome sequencing to discover genetic aberrations associated with poor outcome in untreated lymph node negative colon cancer*'	1.0
	Dutch Colorectal Cancer Group avonden - '*Het Prospectief Landelijk ColoRectaal Carcinoom cohort: een landelijke prospectieve observationele studie*'	1.0
2017	Dutch Digestive Disease voorjaarsdagen '*Validation and pathway analysis of a metastasis-specific microRNA signature in primary colon cancer*'	1.0
	Dutch Digestive Disease voorjaarsdagen '*High mRNA expression of splice variant SYK short correlates with hepatic disease progression in untreated lymph node negative colon cancer patients*'	1.0
	Voorjaarsvergadering Nederlande Vereniging voor Heelkunde '*Validatie van een metastase-specifiek prognostisch microRNA profiel in lymfklier negatieve primaire coloncarcinomen*'	1.0
	Dutch Colorectal Cancer Group dag – '*CONNECTION-II, Improving clinical management of colon cancer through CONNECTION, a nation-wide Colon Cancer Registry and Stratification effort*'	1.0

CONFERENCES AND SEMINARS (18.4 ECTS)

		ECTS
2013	Voorjaarsvergadering Nederlandse Vereniging voor Heelkunde	1.0
	Daniel den Hoed dag	0.3
2014	European Multidisciplinary Colorectal Cancer Congress (poster)	1.0
	Leversymposium AMC	0.3
	Voorjaarsvergadering Nederlandse Vereniging voor Heelkunde	1.0
	Najaarsvergadering Nederlandse Vereniging voor Heelkunde	0.3
	Wetenschapsdag Heelkunde Erasmus MC	1.0
2015	Liver Metastasis Research Network meeting	1.0
	Transatlantic conference on personalized medicine	0.3
	Nederlandse Vereniging voor Gastro-Enterologie voorjaarsvergadering	1.0
	Voorjaarsvergadering Nederlandse Vereniging voor Heelkunde	1.0
	Najaarsvergadering Nederlandse Vereniging voor Heelkunde	0.3
	Wetenschapsdag Heelkunde Erasmus MC	0.3

2016	American Association of Cancer Research Annual meeting (poster)	1.0
	European Multidisciplinary Colorectal Cancer Congress (3xposter)	1.0
	5D congres	0.3
	MolMed day	0.3
	Voorjaarsvergadering Nederlandse Vereniging voor Heelkunde	1.0
	Nederlandse Vereniging voor Gastro-Enterologie najaarsvergadering	1.0
	Empowering Personalized Medicine Health-RI meeting (poster)	1.0
2017	American Association of Cancer Research Annual meeting (3xposter)	1.0
	Liver Metastasis Research Network meeting	1.0
	Dutch Digestive Disease voorjaarsdagen	1.0
	Voorjaarsvergadering Nederlandse Vereniging voor Heelkunde	1.0

TEACHING (9 ECTS)

		ECTS
2014	EHBO courses	0.25
	Supervision master thesis	2.0
2015	EHBO courses	0.25
	Supervision master thesis (2x)	4.0
2016	Supervision master thesis	2.0
	EHBO courses	0.25
2017	EHBO courses	0.25

ACKNOWLEDGEMENTS

Geachte promotor, beste professor IJzermans, bedankt voor uw begeleiding tijdens dit promotietraject. Na een korte periode van inwerken liet u mij al snel zeer zelfstandig werken, waardoor ik veel hebben leren van mijn tijd als promovendus. Ik heb genoten van de steeds gezelligere halve uurtjes op vrijdagochtend waarin de meest uiteenlopende onderwerpen voorbij zijn gekomen. Het is me een waar genoegen bij u te mogen promoveren.

Geachte promotor, beste John, allereerst gefeliciteerd met je benoeming. Het is een eer de eerste promovendus te zijn die bij jou als professor mag promoveren. Bedankt voor de geduldige begeleiding tijdens mijn ontdekkingstocht in de wereld van de tumorbiologie. De deur stond altijd bij je open, en elke vraag kon worden gesteld.

Geachte professor Foekens, beste John, veel dank voor het beoordelen van dit proefschrift. Daarnaast bedankt voor je vlijmscherpe en nauwkeurige correcties op alles wat ik ter lezing heb aangeboden, ik heb er veel van geleerd.

Geachte professor Medema, beste JP, ook jij bedankt voor het beoordelen van dit proefschrift. Maar veel belangrijker, bedankt voor de fijne samenwerking in het CONNECTION-project, en de waardering die je impliciet en expliciet liet blijken voor mijn inzet.

Geachte professor Koopman, beste Miriam, bedankt voor het beoordelen van dit proefschrift. Ik weet dat je die aanhef nog gek vindt, maar wat heb je het verdiend. Ik heb ongelofelijk veel respect voor je gedrevenheid en aanstekelijke ambitie. Ik heb veel van je geleerd tijdens de vele uren die we samen in PLCRC hebben gestoken, en vind het heel bijzonder dat ik de eer heb het symposium ter ere van je oratie te mogen voorzitten. We gaan er een mooie dag van maken!

Geachte leden van de grote commissie, dank voor de bereidheid van gedachten te wisselen over dit proefschrift.

Een speciaal dankwoord aan alle patiënten die toestemming hebben gegeven voor de MATCH-studie. Zonder jullie bereidheid deel te nemen en jullie vertrouwen in ons onderzoeksteam was dit proefschrift niet tot stand gekomen.

Anieta, al snel had ik door dat jij de motor was van het lab waarin alles moest gaan gebeuren. Maar veel belangrijker, dat samenwerken met jou helemaal zou gaan lukken. Ik heb enorm genoten van de (soms vroege) besprekingen bij jou op de kamer, ik op een ladekastje en jij achter je bureau. Ik wil je bedanken voor al die uren begeleiding en alles wat je me geleerd hebt.

Marcel, ik weet niet hoe vaak heb ik gedacht "ik kan niet weer Marcel gaan bellen om iets uit te zoeken". Maar altijd als ik de moed toch weer had verzameld, was jij direct bereid me te helpen en uit te leggen wat je allemaal deed. Dank voor al je hulp, en voor de gezellige tijd in New Orleans.

Monique en Shazia, ontelbaar veel cupjes, stickers en rekjes heb ik bij jullie opgehaald en weer gevuld teruggebracht. Bedankt voor de fijne samenwerking, jullie vrolijkheid en gezelligheid.

Katharina, Carolien en Han, bedankt voor de eindeloze hoeveelheid slides die jullie hebben beoordeeld.

Saskia, ik ken je nog maar kort maar wat ben je een topper. Ik weet zeker dat jij en Inge een aantal mooie ontdekkingen gaan doen.

Vanja, Michelle, Anne en Sandra, dank voor de leuke samenwerking en alle uren die jullie in het analyseren en opwerken van al het aangeleverde vriesmateriaal hebben gestoken.

Geraldine, ik durf wel te stellen dat ik niemand vaker aan de lijn heb gehad tijdens mijn onderzoekstijd dan jij. Ik heb met enorm veel plezier met je samengewerkt, en hoop dat je nog een paar jaar de drijvende kracht van PLCRC zal blijven.

Gerrit, iedere week heb ik me tijdens onze vergaderingen weer verbaasd over de hoeveelheid ballen die jij in de lucht weet te houden, en hoeveel kruisbestuiving jij in onderzoeksland teweeg weet te brengen. Bedankt voor de samenwerking, en het vertrouwen dat je me vanaf het begin als directe collega hebt gegeven.

Carola, jij mag natuurlijk niet ontbreken in dit dankwoord. Je altijd opgewekte stemming en je bijzondere talent om immer het overzicht te bewaren maakte je een top collega om mee samen te werken.

Leden van de MATCH-studiegroep, verpleegkundig specialisten, nurse practitioners, pathologiemedewerkers en OK-personeel, jullie zijn de basis die de MATCH en daardoor alles in dit proefschrift mogelijk hebben gemaakt. Enorm bedankt voor de samenwerking en de mogelijkheden die jullie hebben gecreëerd. Jullie zijn een voorbeeld voor het darmkankerveld in Nederland!

Zarina, jouw inspanningen voor de MATCH-studie en de door jou binnengehaalde subsidies hebben dit boekwerk mogelijk gemaakt. Ik hoop dat je je realiseert dat je

iets heel bijzonders hebt neergezet waarvoor ik je enorm dankbaar ben. Je bent een topper!

Tessa, ik heb het al vaker tegen je gezegd, maar nu echt zwart op wit: bedankt voor al je hulp en advies bij het maken van een aantal belangrijke beslissingen de afgelopen jaren. Ik waardeer je als collega, maar nog belangrijker als persoon. Succes met de fantastische plek die je hebt weten te bemachtigen, wellicht kom ik nog een stage bij je doen!

Inge, aan jou de schone taak om het stokje van mij over te nemen. Ik heb er alle vertrouwen in dat het goed gaat komen, zet hem op!

Joël en Bo, jullie als paranimf vragen was de eenvoudigste beslissing uit mijn hele promotietraject. We zijn begonnen als collega's, maar inmiddels zijn jullie beiden dierbare vrienden van me geworden. Bo, ik ben diep onder de indruk van het boekwerk dat jij aan het schrijven bent, en de efficiëntie waarmee je het doet. Heel veel succes met het laatste stukje en knallen bij de sollicitatie! Ik ga onze tijd in de Z enorm missen, maar weet zeker dat we elkaar zullen blijven zien. Joël, de uitzonderlijke combinatie van intelligentie, gezelligheid en humor maakt je een bijzonder persoon. Bij jou kan ik volledig mezelf zijn, dat maakt onze vriendschap voor mij heel belangrijk.

Bruders, als oudere broer is het fantastisch te zien dat jullie goed terecht zijn gekomen. Wout, ik hoop dat je de rest van je leven de vrolijke, inventieve en nieuwsgierige aap blijft die je nu bent. Thoom, nog een paar jaar en dan is je RA-diploma binnen. Ik heb er geen twijfel over dat je er slim genoeg voor bent, de truc is volhouden. Mijn deur staat altijd voor jullie open.

Papa en mama, hoe ouder ik word, hoe meer bewondering ik voor jullie krijg. Thuis is de plek geweest waar ik me veilig voelde en alles kon worden gevraagd, vroeger en nog steeds. Ik ben trots jullie zoon te zijn, en hoop nog lang van jullie als (o)pa en (o)ma te mogen genieten. En hoe langer ik ook nadenk over wat nog te schrijven, alles komt neer op 'ik hou van jullie, bedankt voor alles!'.

Lieve Liz, door jou promoveerde ik van man naar vader. Het is bijzonder te ervaren wat onvoorwaardelijke liefde is, je bent m'n persoonlijke wondertje.

Eef, natuurlijk zijn de laatste woorden voor jou. Jouw onvoorwaardelijke steun maakte de avonden en weekenden in de Z-flat zoveel makkelijker. Je ten huwelijk vragen is de beste beslissing geweest uit mijn leven, samen gaan we een fantastisch avontuur tegemoet. Je bent mijn maatje, mijn topper, ik kan me geen leven meer zonder jou en onze Liz voorstellen. Ik hou van je.

ABOUT THE AUTHOR

Robert Coebergh van den Braak werd op 25 juli 1987 geboren in het Catharina Ziekenhuis in Eindhoven als de eerste van drie zonen. Op zijn 11e verhuisde hij naar Helmond, waar hij in 2005 aan het Carolus Borromeus College zijn VWO-diploma haalde. In datzelfde jaar begon hij aan zijn studie geneeskunde aan de Erasmus Universiteit Rotterdam.

Na het behalen van zijn artsexamen in september 2012 werkte hij met veel plezier als arts-assistent bij de afdeling heelkunde in het Reinier de Graaf Gasthuis te Delft (Dr. M. van der Elst), waarna hij in oktober 2013 begon als arts-onderzoeker in het Erasmus MC Universitair Medisch Centrum te Rotterdam (promotor: Prof. dr. J.N.M. IJzermans en copromotor: Prof. dr. J.W.M. Martens). De resultaten van het onderzoek naar moleculaire biomarkers bij het coloncarcinoom zijn gebundeld in dit proefschrift. Op 1 januari 2018 begon hij aan zijn opleiding tot chirurg in regio Rotterdam (opleiders: Dr. B.P.L. Wijnhoven en Dr. L. van der Laan).

Robert woont gelukkig samen met zijn vrouw Eva Julia Niers. Op 9 oktober 2017 werden zij trotse ouders van hun eerste dochter Liz.