Oncogene (2016) **35,** 6026–6037
© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved 0950-9232/16
www.nature.com/onc

OPEN

ORIGINAL ARTICLE

A multidimensional network approach reveals microRNAs as determinants of the mesenchymal colorectal cancer subtype

E Fessler[1,2], M Jansen[3,9], F De Sousa E Melo[1,2,10], L Zhao[4], PR Prasetyanti[1,2], H Rodermond[1,2], R Kandimalla[1,2], JF Linnekamp[1,2], M Franitza[5,6], SR van Hooff[1,2], JH de Jong[1,2], SC Oppeneer[7], CJM van Noesel[3], E Dekker[7], G Stassi[8], X Wang[4,11], JP Medema[1,2,11] and L Vermeulen[1,11]

Colorectal cancer (CRC) is a heterogeneous disease posing a challenge for accurate classification and treatment of this malignancy. There is no common genetic molecular feature that would allow for the identification of patients at risk for developing recurrences and thus selecting patients who would benefit from more stringent therapies still poses a major clinical challenge. Recently, an international multicenter consortium (CRC Subtyping Consortium) was established aiming at the classification of CRC patients in biologically homogeneous CRC subtypes. Four consensus molecular subtypes (CMSs) were identified, of which the mesenchymal CMS4 presented with worse prognosis signifying the importance of identifying these patients. Despite the large number of samples analyzed and their clear association with unifying biological programs and clinical features, single-driver mutations could not be identified and patients are heterogeneous with regard to currently used clinical markers. We therefore set out to define the regulatory mechanisms underlying the distinct gene expression profiles using a network-based approach involving multiple molecular modalities such as gene expression, methylation levels and microRNA (miR) expression. The miR-200 family presented as the most powerful determinant of CMS4-specific gene expression, tuning the majority of genes differentially expressed in the poor prognosis subtype, including genes associated with the epithelial–mesenchymal transition program. Furthermore, our data show that two epigenetic marks, namely the methylation of the two miR-200 promoter regions, can identify tumors belonging to the mesenchymal subtype and is predictive of disease-free survival in CRC patients. Importantly, epigenetic silencing of the miR-200 family is also detected in epithelial CRC cell lines that belong to the mesenchymal CMS. We thus show that determining regulatory networks is a powerful strategy to define drivers of distinct cancer subtypes, which possess the ability to identify subtype affiliation and to shed light on biological behavior.

Oncogene (2016) **35,** 6026–6037; doi:10.1038/onc.2016.134; published online 9 May 2016

INTRODUCTION

Colorectal cancer (CRC) is a heterogeneous disease, which until recently has only been reflected by (i) histopathological classification, (ii) single molecular conditions such as microsatellite instability (MSI) and (iii) mutation status of major cancer genes. However, the diversity of the disease has made it difficult to accurately classify and treat CRC patients. Only lately, CRC patients have been categorized using unsupervised classification of gene expression profiling which resulted in the identification of distinct CRC subtypes.[1–6] In an effort to generate a unified classification system, the CRC Subtyping Consortium (CRCSC) was established, which identified the existence of four consensus molecular subtypes (CMSs) in a large panel of CRC patients ($n = 4151$).[7] Different CMSs associate with specific molecular, biological and clinical features and thus present distinct entities. Linking the subtypes to disease outcome revealed that the mesenchymal CMS4 displayed worse prognosis, highlighting the clinical relevance of this approach.[7]

To date subtyping of CRC based on gene expression data has not taken regulatory networks into account, thus the underlying determinants remain elusive.[8] In glioblastoma such a transcriptional network driving the expression of mesenchymal genes of a poor prognosis subtype was identified.[9] Also in ovarian and gastric carcinoma a poor prognosis mesenchymal subtype was detected, driven by a microRNA (miR) regulatory network.[10,11] miRs are short noncoding RNAs that regulate translation of mRNAs into proteins.[12] Individual miRs can affect an extensive number of genes simultaneously,[13,14] explaining their ability to influence expression of complete networks defining a cancer subtype. Of special interest with regard to driving poor prognosis subtypes is

[1]Laboratory for Experimental Oncology and Radiobiology (LEXOR), Center for Experimental Molecular Medicine (CEMM), Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands; [2]Cancer Genomics Center, Utrecht, The Netherlands; [3]Department of Pathology, AMC, University of Amsterdam, Amsterdam, The Netherlands; [4]Department of Biomedical Sciences, City University of Hong Kong, Kowloon Tong, Hong Kong; [5]Cologne Center for Genomics (CCG), University of Cologne, Cologne, Germany; [6]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany; [7]Department of Gastroenterology, AMC, University of Amsterdam, Amsterdam, The Netherlands and [8]Cellular and Molecular Pathophysiology Laboratory, Department of Surgical and Oncological Sciences, University of Palermo, Palermo, Italy. Correspondence: Dr X Wang, Department of Biomedical Sciences, City University of Hong Kong, Kowloon Tong, Hong Kong or Professor Dr JP Medema or Dr L Vermeulen, Laboratory for Experimental Oncology and Radiobiology (LEXOR), Center for Experimental Molecular Medicine (CEMM), Academic Medical Center (AMC), University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands.
E-mail: xin.wang@cityu.edu.hk or j.p.medema@amc.nl or l.vermeulen@amc.nl
[9]Current Address: Centre for Tumour Biology, Barts Cancer Institute, University of London, London, UK.
[10]Current Address: Department of Molecular Oncology, Genentech Inc., San Francisco, CA, USA.
[11]These three authors contributed equally to this work.
Received 13 October 2015; revised 17 March 2016; accepted 22 March 2016; published online 9 May 2016

the ability of miRs to regulate the expression of genes associated with aggressive features of malignancies, such as the epithelial–mesenchymal transition (EMT).[15] Cancer cells undergoing EMT lose epithelial characteristics, such as polarity and cell–cell adhesion, and gain mesenchymal features that allow them to disseminate to distant organs.[16] A direct link between miRs and EMT was firmly established and specifically the miR-200 family has been implicated in repressing the EMT-inducing zinc-finger E-box binding homeobox transcription factors ZEB1 and ZEB2,[17,18] thereby inhibiting the EMT program. Also in CRC miR-200 family members have been shown to regulate the transition between distinct cellular states,[19] and to play a role in development of metastatic disease. Cells at the invasive front of tumors down-regulate miR-200 expression, allowing them to undergo EMT and leave the primary tumor. This process is reversed in metastatic lesions, leading to the re-acquisition of an epithelial phenotype.[20,21] The expression of miR-200 family members, and thus the orchestration of the EMT program, has been shown to be regulated by miR-200 promoter methylation.[21–25]

Herein, we describe a network-based approach enabling the identification of subtype-specific drivers to gain insight into the molecular determinants of distinct CRC subtypes. The dominant regulatory network identified in the CMS4 group consists of the miR-200 family. Methylation of the miR-200 promoter regions controls the activity of this network and is responsible for tuning the gene expression of the poor prognosis CMS. We show that in CRC the miR-200 family cannot only be used to define distinct cellular states, such as associated with the EMT program, but regulates subtype-specific gene expression, thus being a determinant for the subset of poor prognosis, mesenchymal CRCs. Defining the underlying regulatory network of CMS4 tumors allowed us to identify miR-200 promoter methylation as a molecular marker that can be used to determine subtype affiliation at early stages in tumor development and that presents an independent prognostic factor in stage II CRC.

RESULTS

Gene expression profiling identifies a heterogeneous group of poor prognosis CRCs

Based on gene expression profiling, CRCs can be divided into four main CMSs using a random forest classifier.[7] To illustrate this finding, the classifier was applied to a large publicly available data set of 566 colon cancer patients (Cartes d'Identité des Tumeurs, CIT) (Supplementary Figure S1a).[6] A total of 557 patients were classified in one of the four CMSs (Figures 1a and b). This gene expression-based classification allowed for the identification of distinct subsets of CRC. The largest CMS2 subgroup, for instance, associated with the well-characterized chromosomal-instable (CIN) type of CRC, as 95.5% of the tumors identified as CMS2 were CIN+ and mainly anatomically left-sided, displaying *KRAS* and/or *TP53* mutations. The CMS1 group contained a high level of MSI and CIMP+ right-sided tumors, therefore associating with the reported MSI/CIMP+ type of CRC (Figure 1b).[26–28] However, none of these molecular markers was able to unequivocally point to one subtype and especially tumors belonging to the CMS3 and CMS4 subtypes represented heterogeneous groups of CRC patients with respect to molecular aberrations commonly occurring in this disease (Figure 1b). There was however a strong association of clinical characteristics, as the CMS1–3 subgroups were enriched in stage I +II tumors, whereas more than 55% of CMS4 tumors were classified as stage III+IV (Figure 1b). Although representing a mixed group based on molecular markers, we observed a significantly poorer prognosis for patients with CMS4 tumors (Figure 1c). Therefore the clinical relevance of this approach was confirmed as it allows the identification of a subtype with significantly decreased disease-free survival (DFS).

A microRNA regulatory network accounts for differential gene expression between CMS1–3 and CMS4 tumors

Mutation status or molecular characteristics such as MSI do not allow the distinction between the different CMSs, indicating that the driving force for specific cancer subtypes is more complex. To investigate which regulatory network is predominantly responsible for differences in gene expression between the subtypes we generated three regulatory networks comprising various dimensions of molecular regulation. In particular we studied (i) transcription factors (TFs) (Supplementary Figure S2), (ii) methylation marks (Supplementary Figure S3) and (iii) microRNAs (miRs) (Figure 2) as putative subtype-specific regulators. Network analyses were performed to identify which level of regulation was most stringently associated with the subtype-specific gene expression distinction in the TCGA data set (Supplementary Figure S1a). We chose to investigate the differences between CMS1–3 and CMS4 tumors, as this mesenchymal subtype is associated with worse clinical outcome. The miR network, consisting of miRs significantly lower expressed in CMS4 compared with CMS1–3 tumors (Supplementary Figure S4a), was found to be the most powerful determinant of subtype-specific regulation of gene expression. The largest regulatory element was predicted to be responsible for 74.8% of the variation in gene expression (Supplementary Figure S4b; based on reference networks; for comparisons with the other networks the percentage was calculated based on the transcriptional network with redundant edges removed based on ARACNE which resulted in a percentage of 61.9%). The largest elements in the TF- and methylation mark-based regulatory networks account for approximately 10.3 and 10.4%, respectively. The dominant regulatory element in the miR network was generated by the miR-200 family, namely miR-200a, miR-200b, miR-200c, miR-141 and miR-429, representing miRs whose expression levels differ significantly between CMS1–3 and CMS4 tumors (Supplementary Figure S5a). The network analysis predicted that these miRs target similar genes, suggesting that the miR-200 family members are functionally related, which was confirmed by analyzing the overlap of genes most strongly predicted ($P < 0.01$) to be regulated by one of the miR-200 family members based on the network analysis (Supplementary Figure S5b).

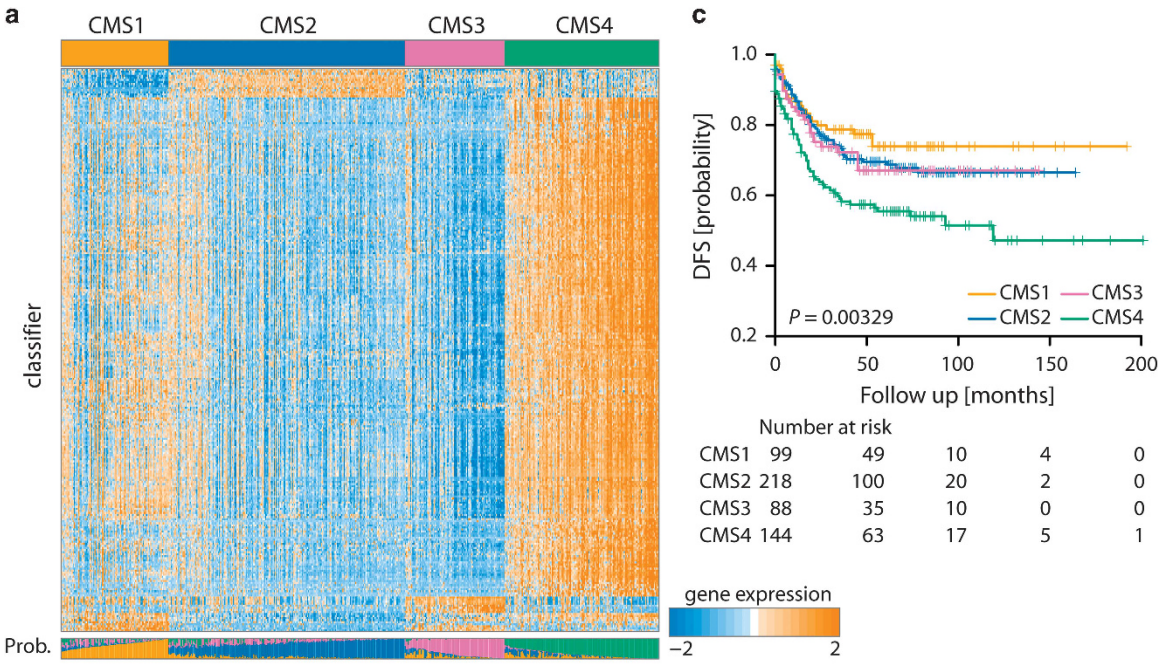
These findings point to a critical role of the miR-200 regulatory network in orchestrating subtype-specific gene expression. As reported before, expression of the miR-200 family associated with low expression of mesenchymal genes and inhibition of the EMT program by modulating ZEB1 levels,[17,18,29,30] thus the low abundance of mi-200 family members in CMS4 tumors could explain their mesenchymal phenotype.

Highlighting the genes belonging to an experimentally derived EMT signature in the inferred regulatory network revealed that many of the EMT-associated genes were indeed predicted to be regulated by the identified miRs (Figure 2).[31] Most of the EMT genes belonged to the miR-200 family cluster, and reversely, further analysis revealed that 12 miRs, including all miR-200 family members, function as master regulators of the EMT signature (Supplementary Figure S4c). Consistent with the idea that miR-200 members are instrumental in orchestrating an EMT signature in CMS4 tumors, the expression of genes present in an experimentally derived EMT signature was highly enriched in CMS4 tumors of the CIT data set (Supplementary Figure S6).[6,31]

Methylation of promoter regions underlies differential expression of miR-200 family members

The members of the miR-200 family are organized as polycistronic transcripts on chromosome 1 (miR-200ba429) and chromosome 12 (miR-200c141).[32] The methylation mark regulator network identified the methylation of the miR-200 loci as regulator of subtype-specific gene expression (Supplementary Figure S3) and

© 2016 Macmillan Publishers Limited, part of Springer Nature.



b

	CMS1	CMS2	CMS3	CMS4
number of patients	100/557 (18)	220/557 (39.5)	93/557 (16.7)	144/557 (25.9)
***BRAF*-mutant**	36/95 (37.9)***	0/194 (0)	9/86 (10.5)	6/129 (4.7)
***KRAS*-mutant**	29/98 (29.6)	64/210 (30.5)	62/89 (69.7)***	57/140 (40.7)
***TP53*-mutant**	19/55 (34.5)	106/151 (70.2)***	18/60 (30)	46/84 (54.8)
MSI	54/87 (62.1)***	3/209 (1.4)	14/92 (15.2)	4/123 (3.3)
CIN+	39/87 (44.8)	169/177 (95.5)***	43/80 (53.8)	98/114 (86)**
CIMP+	54/92 (58.7)***	8/188 (4.3)	18/84 (21.4)	11/124 (8.9)
proximal location	74/100 (74)***	44/220 (20)	51/93 (54.8)**	49/144 (34)
distal location	26/100 (26)	176/220 (80)***	42/93 (45.2)	95/144 (66)
AJCC stage I+II	62/100 (62)	123/220 (55.9)	46/93 (49.5)	61/144 (42.4)
AJCC stage III+IV	38/100 (38)	97/220 (44.1)	44/93 (47.3)	83/144 (57.6)**

Figure 1. Classification of CRC samples into four distinct subtypes based on gene expression. (**a**) In total, 557 patients of the CIT data set are classified in four CMSs. Columns indicate patients, whose subtypes are indicated in the top bar. The posterior probability of belonging to the subtype is shown in the bar below the heatmap. Rows represent genes of a random forest classifier. The heatmap displays the median centered log2 gene expression levels (high expression: orange, low expression: blue). (**b**) Association of each subtype with molecular characteristics as well as stage and tumor location. The number of patients displaying the respective characteristic and the total number of patients analyzed are shown. The percentage of positivity is shown in brackets. Asterisks indicate the significance of association of one subtype with the respective feature as determined by hypergeometric tests (**$P < 0.01$, ***$P < 0.001$). (**c**) KM curve showing DFS of patients of the CIT set according to CMS. *P*-value is based on log-rank test.

promoter methylation has been described to regulate expression of miR-200 family members.[22–25] Indeed, both miR-200 loci displayed significantly higher methylation levels in CMS4 CRC samples from the TCGA set (Figure 3a). The methylation of the two loci was highly correlated, confirming the role of all miR-200 family members in regulating a common gene expression program (Figure 3b). Furthermore, methylation of each locus significantly correlated with expression of the respective miR-200 family member (Figure 3c and Supplementary Figures S5c–e), suggesting it to be decisive for the differential miR-200 expression. Additionally the EMT-inducing TF ZEB1, which is a direct target of the miR-200 family,[30] was significantly higher expressed in tumors belonging to the CMS4 group (Figure 3d).

Recently, two independent studies reported that not the epithelium, but the abundance or activation of the stroma accounts for the mesenchymal status of this CRC subtype.[33,34] As miR-200 family expression is suggested to be regulated by methylation in stromal fibroblasts as well,[25] the elevated levels of methylation could therefore simply reflect the amount of stroma in these tumors. However, our data point to the expression of ZEB1 in the epithelium of mesenchymal tumors,[2] suggesting that this mechanism of gene regulation is active in the malignant, epithelial part of CMS4 tumors. Indeed, the analysis of primary tumors and cell lines or xenografts derived from these cancers confirmed this observation (Supplementary Figure S8). In primary tumors, whose corresponding cell lines or xenografts do not

© 2016 Macmillan Publishers Limited, part of Springer Nature.

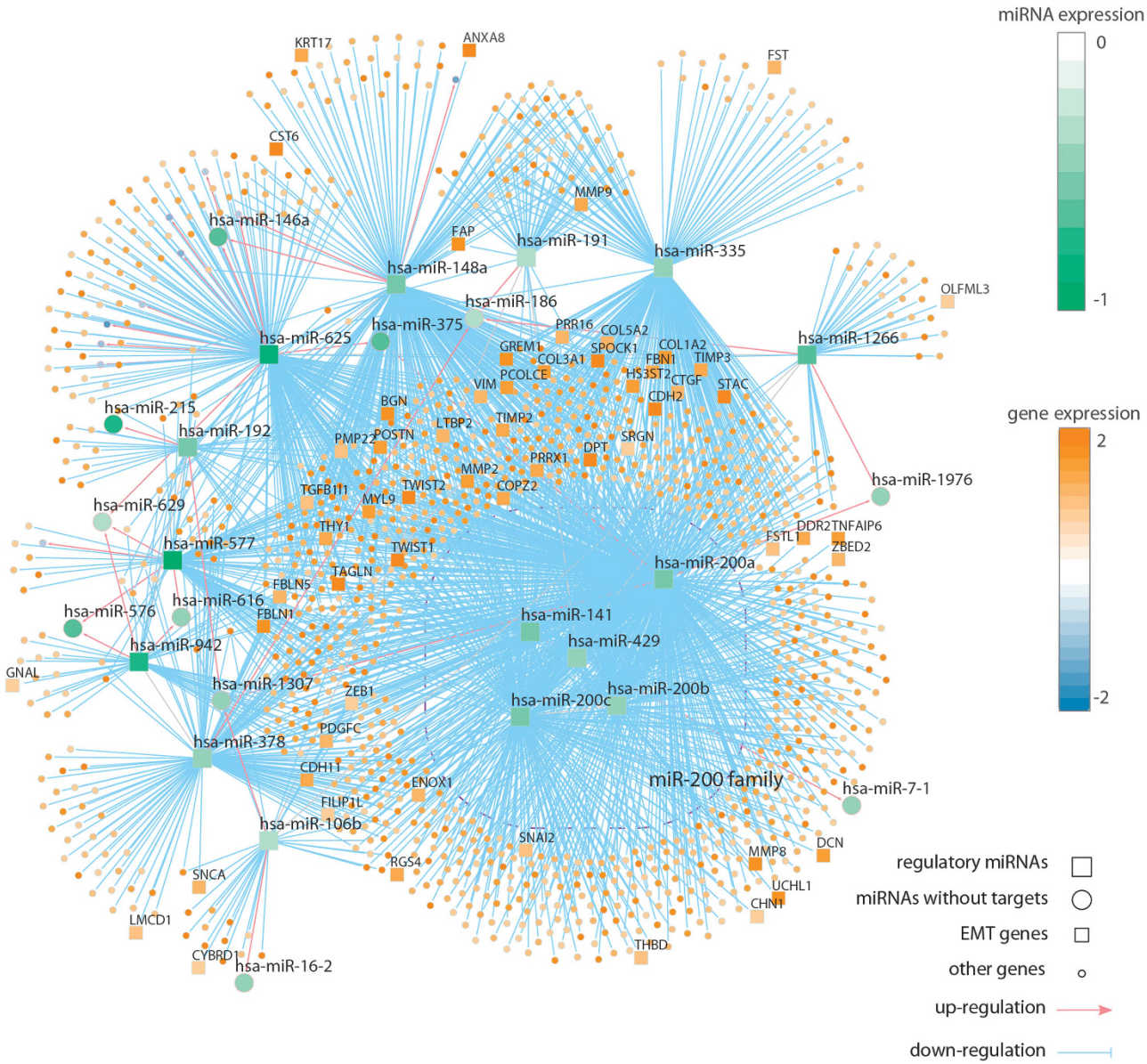


Figure 2. Network analysis identifies the miR-200 family as the main regulatory network of CMS4- and EMT-related genes. Displayed is the differential miR expression between CMS1–3 and CMS4 tumors of the TCGA data set (green: lowly expressed in CMS4). Genes predicted to be regulated by the miRs are shown based on their differential expression between CMS1–3 and CMS4 tumors (blue: lowly expressed in CMS4, orange: highly expressed in CMS4). The connection between miRs and genes is depicted in red (induction) or blue (repression) based on the influence of the miR on gene expression. EMT-associated genes from the Taube EMT signature are highlighted as rectangles.

display any miR-200 promoter methylation, this signal probably reflects the stromal content of these cancers. Yet, if miR-200 promoter methylation displays high levels in the primary tumors, it can also be detected in matching cell lines and xenografts, therefore methylation signals derived from the epithelial tumor cell fraction importantly contribute to this signal (Supplementary Figure S8). To study miR-200 promoter methylation in the epithelial compartment more closely we made use of CRC cell lines stratified into the CMSs using gene expression data. Also in this *in vitro* culture system, we could detect methylation of the miR-200 promoter regions. Seventy-one percent of cell lines classified as CMS4 (5/7) presented with high miR-200 promoter methylation levels, whereas only one CMS1 cell line (14%) and none of the cell lines classified as CMS2 and CMS3 displayed these methylation marks (Figure 4a and Supplementary Figure S7). The highest levels of miR-200 promoter methylation were detected in cell lines that also displayed low levels of miR-200 expression (Figure 4b). Using decitabine the miR-200 loci were demethylated in two cell lines belonging to the mesenchymal subtype, which resulted in the upregulation of miR-200 family member expression (Figure 4c). Interfering with another component of epigenetic gene regulation, namely acetylation, by histone deacetylase inhibition using Panobinostat, did not modulate miR-200 family expression in a meaningful and consistent way (Supplementary Figure S9). These results indicate that repression of the miR-200 regulatory network by methylation is instrumental in controlling miR-200 expression and is indeed active in the epithelial compartment of CMS4 tumors.

miR-200 family members are instrumental for subtype-associated gene expression

To confirm the role of miR-200 family members in regulating subtype-specific gene expression, miR-200 family members were introduced in three CMS4 cell lines resulting in high levels of

© 2016 Macmillan Publishers Limited, part of Springer Nature.

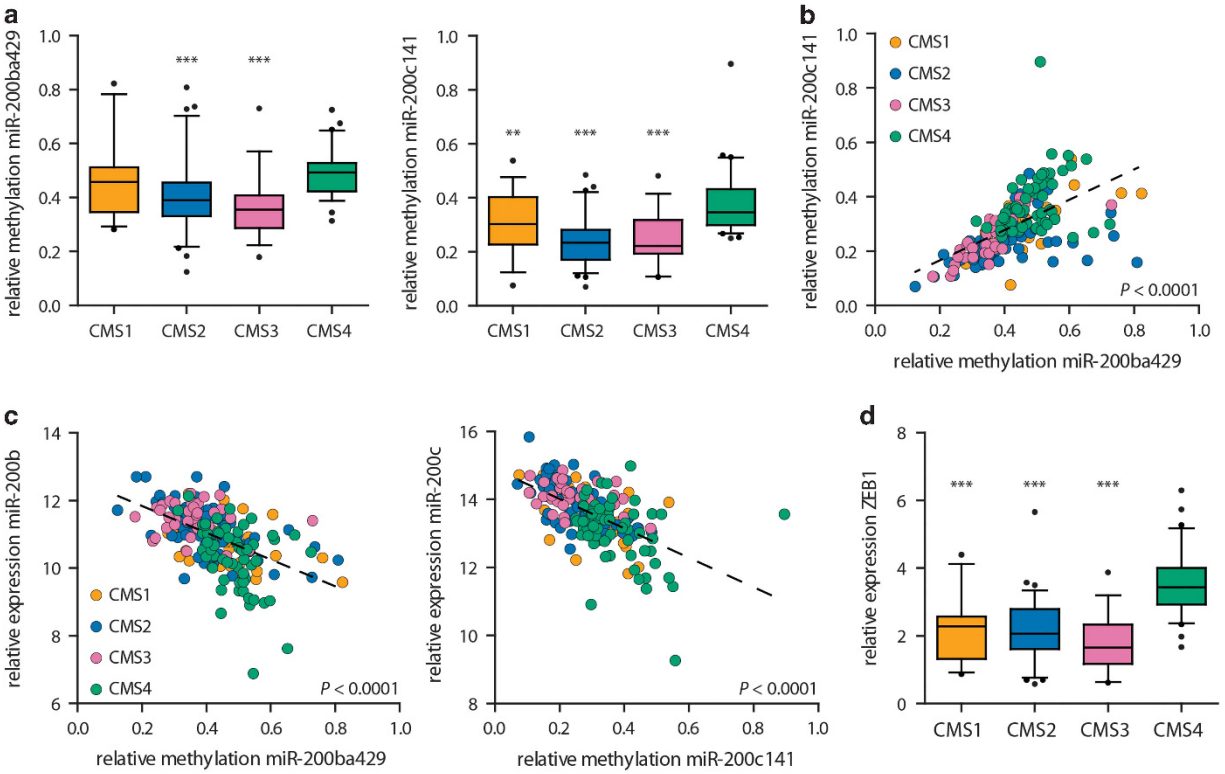


Figure 3. miR-200 loci are highly methylated in tumors of the mesenchymal CMS4. (**a**) Relative methylation of the miR-200ba429 and miR-200c141 loci in CMS1–4 tumor samples of the TCGA data set. For each sample the methylation of the two or three miRs belonging to one locus is averaged. Boxes extend from the 25th to the 75th percentile; the lines indicate the median. The whiskers are drawn to the 5th and the 95th percentile, outliers are depicted as dots. ($n = 32$ for CMS1, $n = 66$ for CMS2, $n = 35$ for CMS3, and $n = 61$ for CMS4. $^{**}P < 0.01$, $^{***}P < 0.001$, asterisks indicate the significance of the respective subtype compared with CMS4; *P*-values are based on two-tailed Student's *t*-tests). (**b**) The methylation of the miR-200ba429 and miR-200c141 loci is highly correlated. (**c**)The methylation of the miR-200ba429 locus is highly correlated with the expression of miR-200b and the methylation of the miR-200c141 locus with the expression of miR-200c. (**d**) CMS4 tumors display significantly higher expression of ZEB1 compared with CMS1–3 tumor samples ($^{***}P < 0.001$, asterisks indicate the significance of the respective subtype compared with CMS4; *P*-values are based on two-tailed Student's *t*-tests. Box plots represent the same parameters as described in (**a**)).

miR-200 family member expression, downregulation of ZEB1 and upregulation of E-Cadherin (Figure 5a and Supplementary Figure S10a). Gene expression profiles were derived from these lines and gene set enrichment analysis[35,36] revealed that genes that were putatively regulated by miR-200 family members, that is, the miR-200 family regulon in the network analysis, were significantly enriched in the gene set that is regulated by miR-200 expression (Supplementary Figure S10b). Moreover, miR-200 overexpression in cell lines suppressed EMT-associated genes (Figure 5b) and genes involved in matrix remodeling, migration and transforming growth factor-β signaling (Supplementary Figures S10c–e). Importantly, genes upregulated in CMS4 compared with CMS1–3 tumors were significantly inhibited in the miR-200-overexpressing cell lines (Figure 5c). To assess whether overexpression of miR-200 family members also has a functional impact, we introduced the two clusters into the primary colon cancer cell line RC511 (Supplementary Figure S11a). RC511 cells show high methylation of miR-200 promoter regions (Supplementary Figure S8) and classify as CMS4 based on gene expression (data not shown). Overexpression of miR-200 family members resulted in similar gene expression changes as observed in the above-mentioned cell lines: the EMT-inducing TF ZEB1 was downregulated and E-Cadherin expression was induced (Supplementary Figure S11b). Subcutaneous injection of RC511 cells into NSG mice revealed that overexpression of miR-200 family members renders the cells less aggressive, evidenced by slower tumor growth (Supplementary Figure S11c) and longer survival time (Supplementary Figure S11d). In combination, these findings validate the notion that the miR-200 family represents an important determinant of subtype-specific gene expression and functional properties.

Methylation of the miR-200 loci is a determinant of CMS4 affiliation and is predictive of DFS

As miR-200 family members orchestrate subtype-specific gene expression, we speculated that the methylation of the miR-200 loci is a predictor of disease subtype and in consequence identifies patients with poor prognosis. We computed a binary logistic regression model using both miR-200 loci methylation levels as covariates. The resulting predicted probabilities were used to calculate the receiver operator characteristics curve, revealing that miR-200 loci methylation was indeed able to identify CMS4 affiliation in the TCGA data set (Figure 6a). Next, we sought to validate the utility of this test in an independent patient series. We therefore determined the methylation of the miR-200 promoter regions in the AMC-AJCCII-90 data set comprising only stage II CRC patients (Supplementary Figures S1a and b). Tumors of the CMS4 group showed a significantly higher methylation of the miR-200 loci compared with tumors belonging to the other subtypes (Figure 6b). The methylation of these two loci was highly correlated (Figure 6c); however, the methylation levels were not influenced by the CIMP status of the tumor (Figure 6d), indicating that it is not a feature of tumors with overall higher DNA

© 2016 Macmillan Publishers Limited, part of Springer Nature.

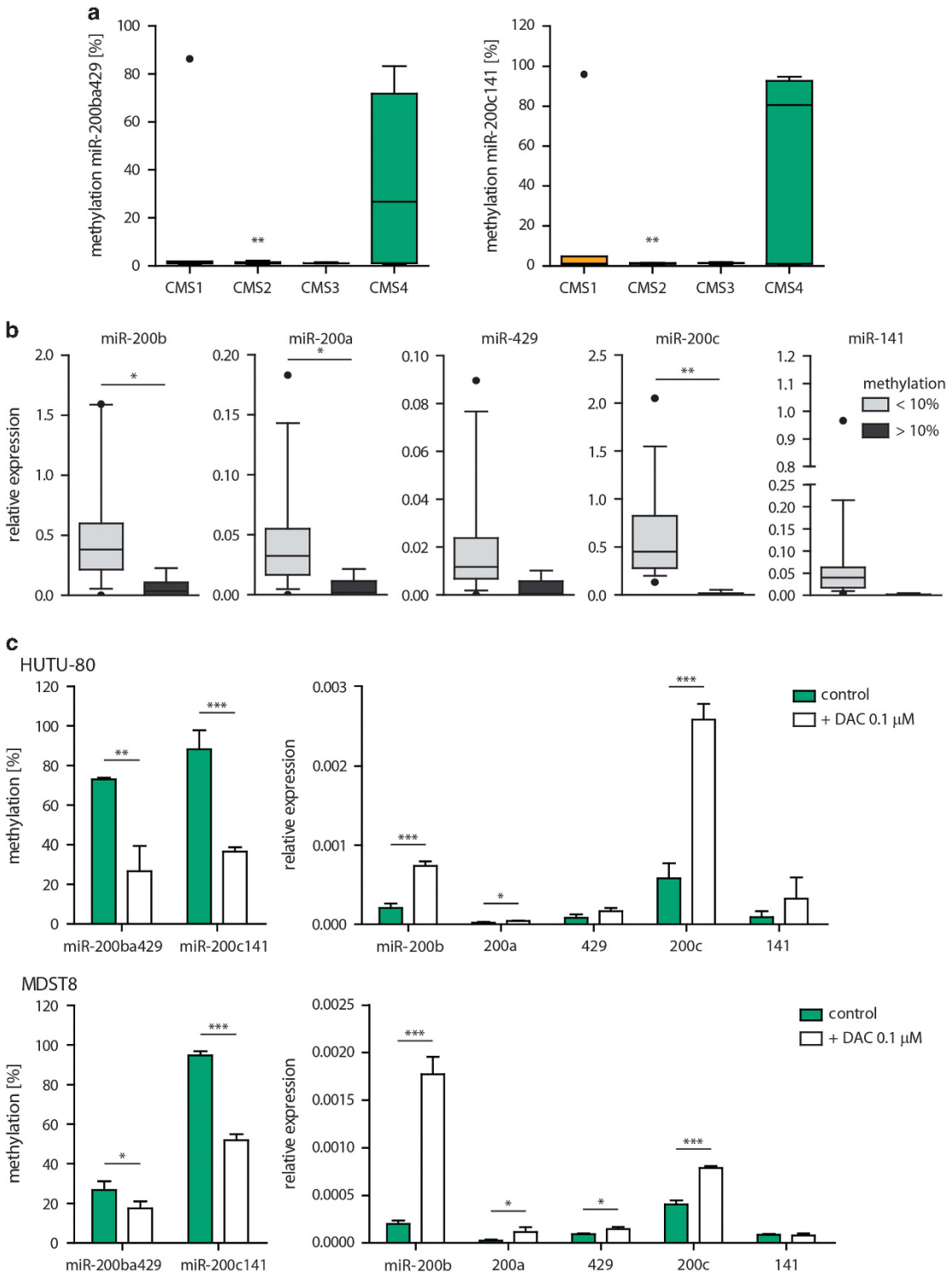


Figure 4. CMS4 cell lines often display high miR-200 promoter methylation levels, which are instrumental for miR-200 repression. (**a**) Promoter methylation levels of miR-200 family members in CMS1–4 CRC cell lines (**$P < 0.01$, asterisks indicate the significance of the respective subtype compared with CMS4; *P*-values are based on two-tailed Student's *t*-tests; $n = 7$ for CMS1, $n = 12$ for CMS2, $n = 3$ for CMS3, $n = 7$ for CMS4. Boxes extend from the 25th to the 75th percentile and the lines indicate the median. The whiskers and outliers are plotted using the Tukey method). (**b**) Cell lines that display high miR-200 promoter methylation (>10%) express low levels of the miR-200 family members ($n = 38$ for methylation <10% and $n = 6$ for methylation >10%; *$P < 0.05$, **$P < 0.01$; *P*-values are based on two-tailed Student's *t*-tests. Boxes extend from the 25th to the 75th percentile and the lines indicate the median. The whiskers are drawn to the 5th and the 95th percentile, outliers are depicted as dots. (**c**) Decitabine (DAC) treatment of the HUTU-80 and MDST8 cell lines reveals that a reduction in methylation results in re-expression of the miR-200 family members ($n = 3$, *P*-values are based on two-tailed Student's *t*-tests, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

© 2016 Macmillan Publishers Limited, part of Springer Nature.

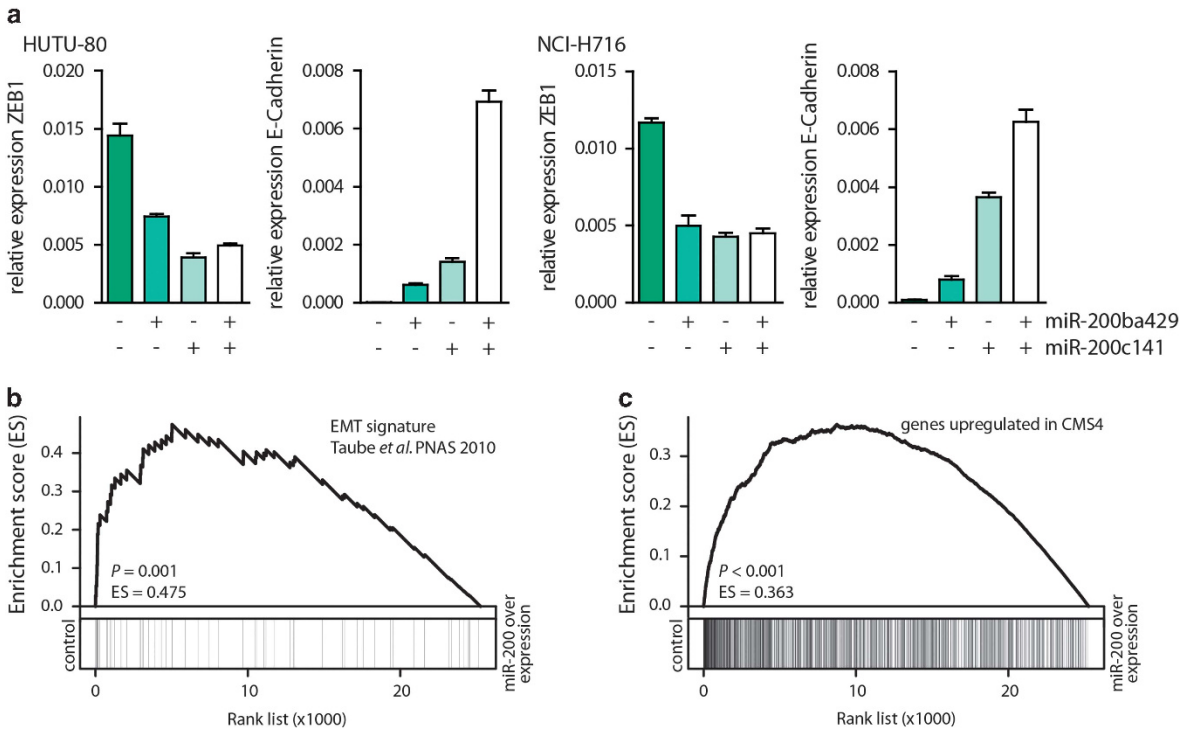


Figure 5. miR-200 family members regulate EMT- and CMS4-associated genes. (**a**) ZEB1 expression is reduced and E-Cadherin expression induced in two CMS4 cell lines following overexpression of miR-200ba429, miR-200c141 or the combination of both clusters, indicating an overlap in target genes. (**b**, **c**) Gene set enrichment analysis (GSEA) of miR-200-overexpressing cell lines confirms that these miRs significantly regulate (**b**) EMT-associated genes[31] and (**c**) the CMS4-specific gene expression program (log2FC (CMS4 vs CMS1–3) > 0.85, FDR < 0.001). ES, enrichment score.

methylation, but that these regions are specifically methylated. DNA methylation in the patients displayed a gradient for both loci, in which CMS4 samples were enriched in the quartile displaying the highest DNA methylation (Supplementary Figures S12a and b). Also in this data set, the combined methylation of the miR-200 loci was able to predict CMS4 affiliation with high confidence (Figure 6e).

Because detection of stage II CRC patients that harbor a poor prognosis is of utmost importance as this provides the opportunity to treat these patients with adjuvant therapy, we investigated if miR-200 methylation directly predicts patient prognosis.

DFS data are available for all patients of the AMC-AJCCII-90 set, in which disease recurrence largely represents detection of metastases in the liver, the lung or the peritoneal cavity. miR-200 promoter methylation was able to identify patients developing recurrences using a Cox proportional hazards regression model (Figure 7a). A Kaplan–Meier (KM) curve was generated making use of a cutoff determined using this receiver operator characteristics curve. Indeed patients containing highly methylated tumors displayed a significantly poorer DFS compared with patients whose tumors showed lower methylation of the miR-200 loci (Figure 7b). To exclude that the miR-200-based stratification leads to a separation of MSI samples that are known to have a better prognosis, the above-described analyses were performed using only microsatellite stable (MSS) patients of the TCGA and AMC-AJCCII-90 data sets, resulting in similar outcomes (Supplementary Figure S13).

Multivariate analyses confirmed that miR-200 promoter methylation is an independent prognostic factor ($P = 0.002$) in the AMC-AJCCII-90 data set also when taking T-stage, differentiation grade, MSS/MSI and *BRAF* mutation status into account (Figure 7c). Hence, we identified the miR-200 promoter methylation as a critical molecular determinant of the mesenchymal CMS, therefore strongly correlating with the patient group that is at high risk for developing recurrences.

DISCUSSION

Gene expression profiling is a powerful tool to detect different patient subgroups and four distinct CRC CMSs have been identified in a large international multicenter study.[7] Tumors belonging to the different CMSs present distinct entities characterized by unifying biological programs and specific survival rates. Yet, it is not clear whether there is a common underlying mechanism for differences in gene expression or distinct clinical outcome. Herein, a network-based approach was applied to investigate the drivers of specific CRC subgroups. Making use of the TCGA data set, we studied gene expression regulation on several dimensions, such as methylation and miRs. As CMS4 tumors display worse prognosis and are thus in need of early identification, we chose to validate our approach using the CMS1–3 vs CMS4 distinction. This way, the miR-200 family was identified as subtype determinant.

The idea of an aggressive mesenchymal CMS in which epithelial cells gain malignant features by undergoing an EMT program has recently been challenged. Two studies report that not the epithelium, but the abundance or activation of the stroma causes the mesenchymal appearance of CMS4 tumors.[33,34] Indeed, stromal cells display methylation of the miR-200 loci and thus likely contribute to the observed miR-200 promoter methylation levels in primary tumors.[25] However, we demonstrate that this epigenetic regulatory loop is also active in epithelial cells, as the miR-200 promoter regions display high methylation levels in cell lines mainly belonging to the CMS4 subtype. We speculate that a shared, yet unidentified upstream mechanism is responsible for

© 2016 Macmillan Publishers Limited, part of Springer Nature.

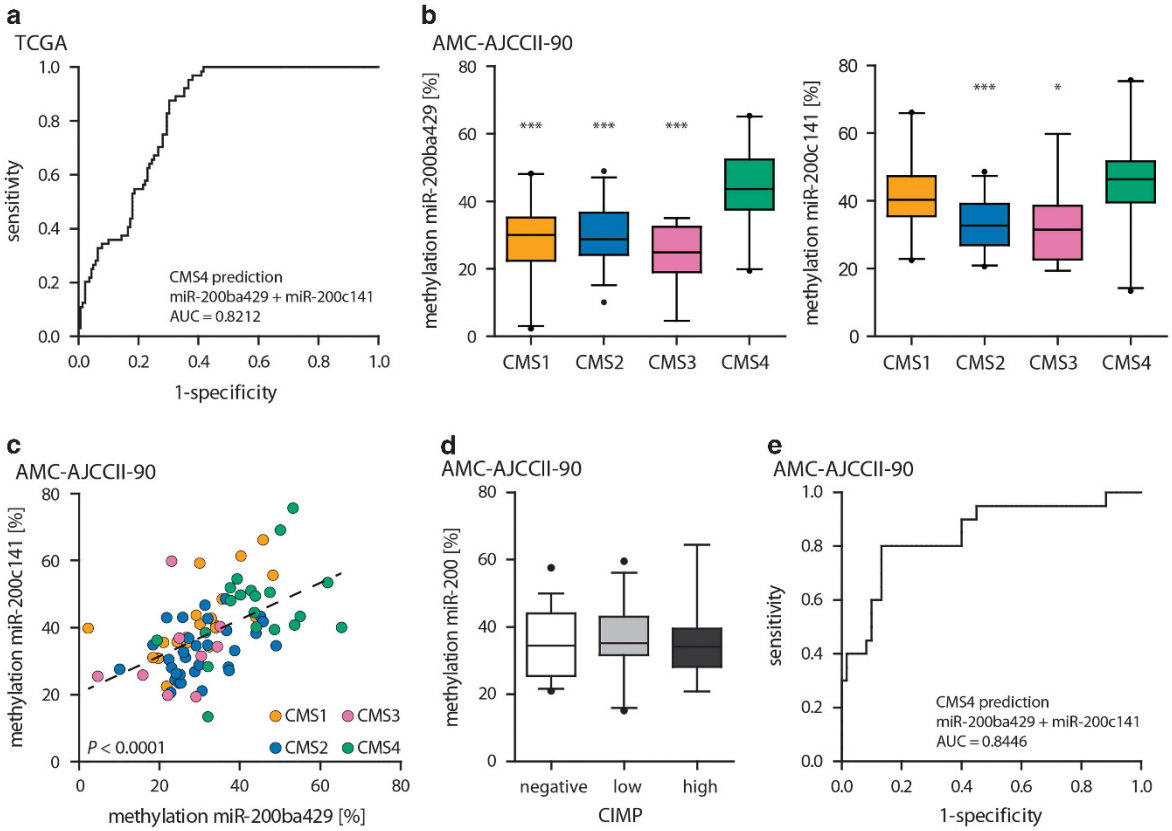


Figure 6. miR-200 loci methylation is predictive of CMS4 affiliation in the TCGA and AMC-AJCCII-90 data sets. (**a**) The receiver operator characteristics (ROC) curve showing the prediction of CMS4 affiliation in the TCGA data set using miR-200 loci methylation (AUC = area under the curve). (**b**) The promoter methylation of the miR-200ba429 and the miR-200c141 loci of 80 patients of the AMC-AJCCII-90 set shows significantly higher levels in tumors of the CMS4 class (*$P < 0.05$, ***$P < 0.001$, asterisks indicate the significance of the respective subtype compared with CMS4; *P*-values are based on two-tailed Student's *t*-tests. Boxes extend from the 25th to the 75th percentile, the lines indicate the median. The whiskers are drawn to the 5th and the 95th percentile, outliers are depicted as dots; $n = 20$ for CMS1, $n = 31$ for CMS2, $n = 9$ for CMS3 and $n = 20$ for CMS4). (**c**) The methylation of the miR-200c141 and miR-200ba429 loci is highly correlated, (**d**) but is not influenced by the CIMP status of the tumor. The average methylation of both miR-200 loci taken together is shown (box plots represent the same parameters as described in (**b**)). (**e**) The ROC curve showing the prediction of CMS4 affiliation in the AMC-AJCCII-90 data set using miR-200 loci methylation.

EMT induction in tumor epithelial cells, as well as for attracting abundant and activated mesenchymal cells to the cancer site.

The EMT is a process during which epithelial cells lose their cell–cell junctions, and gain a mesenchymal, migratory and invasive phenotype, allowing them to leave the primary site and spread to distant organs.[37] Indeed, EMT has been linked to dismal outcome before.[38,39] Two recent reports using mouse models enforce this notion, yet link EMT to therapy resistance rather than enhanced metastatic spread.[40,41] Notwithstanding the exact mechanism, the mesenchymal phenotype has been linked to poor prognosis in patients for many cancer types, including CRC.[7]

As we describe here, the methylation of the miR-200 promoter regions—an epigenetic modification allowing for the activation of the EMT program—could serve as a tool to identify poor prognosis CRC lesions early on. This epigenetic mark is prognostic in stage II CRCs and could be used to select individuals that are likely to benefit from more aggressive therapy. It remains to be examined whether high methylation levels could also be used as a marker to identify patients within this group that benefit from adjuvant therapy. In this regard, the simplicity and highly robust nature of the miR-200 promoter methylation assay might be of great benefit to the field as this will allow for improved retrospective analysis of clinical studies and design of novel trials aiming to improve therapeutic outcome in these patients.

Taken together, our data demonstrate that employing networks to study subtype-specific gene expression is a powerful approach to identify drivers of distinct cancer subgroups. The regulators determined in this way grant insight into the biological behavior of the subtype. Herein, we show that the miR-200 network is a determinant for mesenchymal CRCs and that its activity is regulated by promoter methylation. This methylation mark is predictive of disease outcome, can be detected already early in the development of CRC and thus presents a tool to identify poor prognosis patients at early stages.

MATERIALS AND METHODS

Sample collection

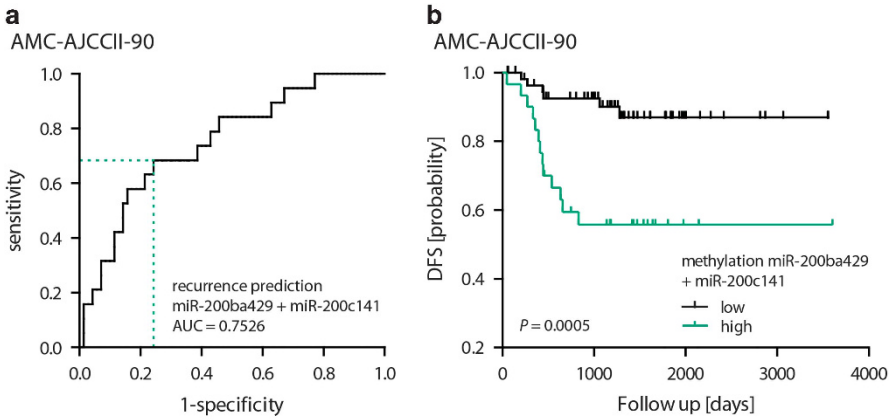
Frozen tissues were collected from pathology (AMC), following the institute's guidelines (Medical Ethical Committee, AMC, Amsterdam, the Netherlands). No approval was needed for samples of the AMC-AJCCII-90 data set; the material was collected as rest material in accordance with the Dutch legislation.

Clinical characteristics of the AMC-AJCCII-90 data set, processing of samples for microarray (GSE33113), gDNA extraction, bisulfite conversion and CIMP analysis were described before.[2]

Consensus molecular subtype (CMS) classification

CIT data set: microarrays (GSE39582) were first normalized and summarized using the robust multi-array average method. Non-biological effects across batches were detected using principal component analysis and

© 2016 Macmillan Publishers Limited, part of Springer Nature.



c

	P-value	Hazard Ratio	95% Confidence Interval
Univariate analyses			
Gender	0.855	1.088	0.438-2.707
Age at operation	0.571	0.771	0.313-1.897
Location	0.847	0.914	0.368-2.274
Differentiation grade	0.297	1.643	0.646-4.179
T-stage	0.355	1.792	0.520-6.176
MSI	0.988	1.008	0.363-2.799
***BRAF*^{V600E} MSS**	0.012	4.997	1.430-17.464
miR-200ba429 + miR-200c141 methylation	0.001	4.902	1.860-12.921
Multivariate analysis			
Gender	0.834	0.894	0.314-2.545
Age at operation	0.436	0.653	0.224-1.905
Location	0.852	0.898	0.289-2.788
Differentiation grade	0.532	1.397	0.489-3.987
T-stage	0.146	3.004	0.681-13.249
MSI	0.982	0.985	0.259-3.745
***BRAF*^{V600E} MSS**	0.027	4.728	1.191-18.772
miR-200ba429 + miR-200c141 methylation	0.002	4.786	1.780-12.869

Figure 7. miR-200 loci methylation is predictive of prognosis in the AMC-AJCCII-90 data set. (**a**) The receiver operator characteristics (ROC) curve displaying the survival prediction in the AMC-AJCCII-90 data set. Green lines indicate cutoff used for the KM curve shown in (**b**). (**b**) The KM curve depicting DFS survival of the AMC-AJCCII-90 data set separated based on the cutoff chosen using the ROC curve shown in (**a**) (*P*-value is based on log-rank test). (**c**) Univariate (top) and multivariate (bottom) analyses of prognostic features in the AMC-AJCCII-90 set. Also in a multivariate analysis miR-200 promoter methylation remains an independent prognostic factor ($P = 0.002$).

were corrected using ComBat.[42] The expression profiles were standardized and then subjected to classification using the CMS classifier as described.[7]

Using the same strategy, we classified the Sanger CRC cell line panel ($n = 43$). This was performed using multiple publicly available gene expression data sets ArrayExpress database (E-MTAB-783), the Cancer Cell Line Encyclopedia[43] ($n = 55$, GSE36133), Wagner *et al.*[44] ($n = 38$, GSE8332) and Medico *et al.*[45] ($n = 155$, GSE59857). Additionally, we derived a classifier that was optimized by first identifying epithelial-expressed genes and subsequent training on the CIT data set. This epithelial-specific classifier was applied to the same data sets as described above. Using all data sets and the different classification methods revealed that most of the lines were consistently classified. However, for some cell lines, this resulted in different CMS affiliations. Cell lines that could not be classified with a confidence of $>66.6\%$ were excluded. Using this approach 29 cell lines were faithfully classified.

Two independent TCGA RNASeq data sets, based on Illumina GA and HiSeq platforms, were used for constructing a regulatory network and for correlation analysis, respectively. CMS classification for the two TCGA data sets and the AMC-AJCCII-90 data set were obtained from the CRCSC.[7]

Regulatory network inference

We employed a network-based approach to study the regulatory relationships between miRs and potential target genes. Gene expression data are log2-transformed RPKM profiles ($n = 270$) in the TCGA Illumina GA data set, while the miR data are log-transformed RPM (reads per million miR mapped, $n = 255$) obtained from The Cancer Genome Atlas Network.[28] Together, 200 patient samples have both miR and gene expression data. We focused on 30 miRs downregulated (fold change <0.71, false discovery rate (FDR) <0.001) in CMS4 vs CMS1–3, and 1437 genes differentially expressed between CMS4 and CMS1–3 ($|\log2FC| > 0.85$, FDR < 0.001). The miR expression data and gene expression data were standardized independently and merged for network inference. The bioconductor *RTN* package was employed to infer the regulatory network.

We tested the statistical significance of overrepresentation of EMT signature genes[31] in each miR's regulon. Twelve miRs of top significance (Benjamini–Hochberg-adjusted *P*-value < 0.05) were selected as master regulators.

Inference of transcription factor regulatory network

From the total number of 1544 transcription factors (TFs) obtained from AnimalTFDB,[46] we first selected 90 TFs significantly differentially expressed between CMS4 and CMS1–3 ($|\log2FC| > 0.85$, FDR < 0.001). The same approach used for inference of the miR regulatory network was employed here to infer a network encoding regulatory relationships between TFs and target genes.

© 2016 Macmillan Publishers Limited, part of Springer Nature.

Inference of DNA methylation regulatory network

Forty-five genes and nine miRs were selected as regulators based on two criteria: (1) significant inverse correlation between their expression profiles and DNA methylation profiles (Pearson correlation coefficient < -0.5, *P*-value < 0.001); (2) significant differential DNA methylation between CMS4 and CMS1–3 samples ($|\log2FC| > \log2(1.1)$, FDR < 0.001). A total of 2273 genes significantly differentially expressed between CMS4 and CMS1–3 ($|\log2FC| > 0.5$, FDR < 0.001) were selected as target genes of the total 54 regulators. Using the same approach for miR network inference, we predicted a network representing how DNA methylation regulates gene expression.

TCGA data

Multiple TCGA data sets were used for analyzing the correlation between expression and promoter methylation of the miR-200 family members. From the TCGA data portal, we obtained mRNA expression, miRNA expression as well as DNA methylation profiles (all level-3 data, from 194 patients). For each miR-200 family member, we took the median β-value over all annotated CpG sites as its promoter methylation level. The mean methylation levels over corresponding individual miR-200 family members were used for the miR-200ba429 and miR-200c141 clusters, respectively.

EMT heatmap

Expression of EMT-associated genes of the CIT data set was imported into MultiExperiment Viewer (http://www.tm4.org/mev.html).[6,31] Genes were normalized by rows, median centered, and the gene tree was hierarchically clustered using Pearson correlation.

Cells

Forty-four CRC cell lines were a kind gift from the Sanger Institute (Cambridge, UK; authenticated by STR Genotyping). Cell lines were mycoplasma negative, cultured in Dulbecco's modified Eagle's medium/F-12 medium with L-glutamine, 15 mM HEPES (Thermo-Fisher Scientific, Bleiswijk, The Netherlands), 8% fetal calf serum (Lonza, Breda, The Netherlands), or in RPMI1640 medium with L-glutamine, 25 mM HEPES (Thermo-Fisher Scientific), 8% fetal calf serum, 1% D-glucose solution plus (Sigma-Aldrich, Zwijndrecht, The Netherlands) and 100 μM sodium pyruvate (Life Technologies, Bleiswijk, The Netherlands).

Cells were plated in medium containing 0.1μM decitabine (5-aza-2′-deoxycytidine; Sigma-Aldrich) for 96 h, 10 nM Panobinostat (LBH589; Selleck Chemicals, Munich, Germany) for 36 h or dimethylsulfoxide. Decitabine and dimethylsulfoxide were refreshed daily.

The primary cell lines RC511 and Co147 were derived from a colon cancer patient as described before[47] and cultured in advanced Dulbecco's modified Eagle's medium/F-12 (Thermo-Fisher Scientific), supplemented with N2 supplement (Invitrogen, Waltham, MA, USA), 2 mM L-glutamine, 0.15% D-glucose (Sigma-Aldrich), 100 μM β-mercaptoethanol (Sigma-Aldrich), trace elements B and C (Thermo-Fisher Scientific), 5 mM HEPES (Life Technologies), 2 μg/ml heparin (Sigma-Aldrich), 10 μg/ml insulin (Sigma-Aldrich), 10 ng/ml human bFGF and 20 ng/ml human EGF (Peprotech, London, UK) in ultra-low attachment flasks (Corning, Amsterdam, The Netherlands). Spheroids were dissociated manually and replated in fresh medium twice weekly.

miR-200 overexpression

HUTU-80, MDST8 and NCI-H716 cells were transduced with lentiviral particles containing lenti-miR-200b+200a+429 or lenti-miR-200c+141 (System Biosciences, Huissen, The Netherlands) constructs and sorted on GFP after 48 h. Twenty-four hours after sorting, a fraction of the miR-200ba429-transduced cells was transduced with miR-200c141 lentivirus. RNA was extracted 7 days after transduction. RC511 primary cells were transduced with lentiviral particles containing lenti-miR-200b+200a+429 and sorted on GFP. Two weeks after sorting, a fraction of the miR-200ba429-transduced cells was transduced with miR-200c141 lentivirus.

For microarrays the GeneTitanTM MC system from Affymetrix (Santa Clara, CA, USA) was used according to the standard protocols of the Cologne Center for Genomics, University of Cologne, Germany (GEO accession GSE65551).

Microarrays ($n = 12$) were first normalized and summarized using the robust multi-array average method.[48] To identify differentially expressed genes, one-class Rank Product analysis was employed,[49] using log2-fold changes of gene expression between each cell line overexpressing miR-200 members and corresponding control. In the gene set enrichment analysis,[35,36] we used as phenotype the ranks of the rank product of genes (*P*-values were adjusted using the Benjamini–Hochberg method). Of the EMT signature[31] only 59 genes differentially expressed between CMS4 and CMS1–3 ($|\log2FC| > 0.85$, FDR < 0.001) were included.

Quantitative real-time PCR

Total RNA was extracted using the NucleoSpin miRNA kit (Macherey-Nagel, Düren, Germany).

miRNA. One microgram of total RNA was reverse transcribed using the HiFlex buffer of the miScript II RT Kit (Qiagen, Venlo, The Netherlands). Quantitative real-time PCR (qRT–PCR) was performed using the miScript SYBR Green PCR kit (Qiagen). miScript Primer assays: Hs_RNU6-2_11, Hs_SNORA74A_11, Hs_miR-200b_3, Hs_miR-200a_1, Hs_miR-429_1, Hs_miR-200c_1, and Hs_miR-141_1. Data shown were normalized to the expression of SNORA74A, and normalization to RNU6-2 yielded comparable results.

mRNA. One microgram of total RNA was reverse transcribed using Superscript III (Invitrogen). qRT–PCR was performed using SYBR Green and a Roche Light Cycler 480 II (Roche, Almere, The Netherlands). Values for cell lines were normalized to *GAPDH* expression; for the primary cell line RC511 values were normalized to *B2M* expression. Primer sequences: *GAPDH*-forward: 5′-AATCCCATCACCATCTTCCA, *GAPDH*-reverse: 5′-TGGACTCCAC GACGTACTCA; *B2M*-forward: 5′-GTCTTTCAGCAAGGACTGGTC, *B2M*-reverse: 5′-CTTCAAACCTCCATGATGC; *ZEB1*-forward: 5′-GCACAAGAAGAGCCACA AGTA, *ZEB1*-reverse: 5′-GCAAGACAAGTTCAAGGGTTC; *E-Cadherin*-forward: 5′-TGGAGGAATTCTTGCTTTGC, *E-Cadherin*-reverse: 5′-CGCTCTCCTCCGAAG AAAC; *p21*-forward: 5′-CAGGCTGAAGGGTCCCCA, *p21*-reverse: 5′-TCAGCCG GCGTTTGGAGTGG.

miR-200 promoter methylation analysis

gDNA was extracted using the High Pure PCR Template Preparation Kit (Roche). Two micrograms of gDNA were bisulfite converted using the EpiTect Bisulfite Kit (Qiagen). Methylation was detected using the PyroMark PCR system (Qiagen). PyroMark Assay Design Software 2.0 (Qiagen) was used for primer design (miR-200ba429 promoter region: 110 basepairs (6 CpGs), miR-200c141 promoter region: 140 basepairs (11 CpGs)). Depicted is the average methylation of all 6 or 11 CpGs combined. Primer sequences: *miR-200ba429*-forward: 5′-GGTTTGAATTGATTTTTTGTGTTAGG; *miR-200ba429*-reverse: 5′-CCTCAACCAAAATCAAACCTCA. *miR-200ba429*-sequencing primer: 5′-ATTTTTTGTGTTAGGGTTT. *miR-200c141*-forward: 5′-ATTGTAGAGGGGGGA TGAG; *miR-200c141*-reverse: 5′-CCAAATTACAATCCAAACAAACC. *miR-200c141*-sequencing primer: 5′-GATGAGGGTGGGTAA. All analyses were performed blinded with respect to subtype and DFS data. Prior to the study, the reproducibility of the miR-200 promoter methylation assay was tested using control cell lines.

In vivo tumor growth

All animal experimentation was approved by the animal ethics committee of the AMC, University of Amsterdam and the University of Palermo. In all, 100 000 primary colon cancer cells (RC511, Co149, DA13) were, after extensive washing with phoshate-buffered saline, injected subcutaneously into the right flank of 13- to 14-week-old male NSG mice (NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ, in house breeding) in 50% growth factor reduced BD matrigel matrix (BD Biosciences, Breda, The Netherlands). Tumors were measured twice weekly in a non-blinded way using a caliper and the formula 'tumor volume = (length × width^2)/2' was used to calculate the volumes. Animals were killed when the tumor volume exceeded 1000mm^3 and tumors were excised for DNA and RNA isolation. To determine the difference in growth speed between RC511 control and RC511+miR-200ba429+miR-200c141, group size was first determined using a power calculation analysis (NQuery, using a power of 80%, a confidence of 0.05 and a variation estimation within the group of 20%). Based on this calculation, six mice per group were subcutaneously injected non-randomized, and subsequent growth and survival curves were generated for mice over a period of 50 days (in GraphPad Prism; significance for survival was calculated using the log-rank test).

© 2016 Macmillan Publishers Limited, part of Springer Nature.

Statistics

Predicted probabilities for CMS4 affiliation were calculated using a binary logistic regression with miR-200ba429 and miR-200c141 methylation as covariates for 80 patients of the AMC-AJCCII-90 data set (nine patients could not be classified into one of the CMSs, one patient had poor gDNA quality). The $x-\beta$ values for predicting recurrence were calculated using a Cox proportional hazards regression model with miR-200ba429 and miR-200c141 methylation as covariates for 89 patients. The sensitivity and specificity of CMS4 and recurrence prediction were calculated by plotting the receiver operator characteristics curves of the predicted probabilities and $x-\beta$ values, respectively, and calculating the AUC using IBM SPSS Statistics version 22 software (IBM, Amsterdam, The Netherlands).

Kaplan–Meier curves display survival; significance was evaluated by log-rank tests. DFS was measured from the day of surgery until recurrence detection. The $x-\beta$ value 0.293 was chosen to divide the patients into methylation low (<0.293) and methylation high (>0.293) (sensitivity: 68.4%, specificity: 75.7%).

Univariate and multivariate Cox proportional hazards regression models were used to test the prognostic relevance of clinical factors and miR-200 methylation. Variables were dichotomized as follows: gender (male vs female), age at operation (mean-dichotomized, <70.24 vs >70.24), location (right- vs left-sided), differentiation grade (well and moderate vs poor), T-stage (T3 vs T4), MSI vs MSS, $BRAF^{V600E}$ and MSS vs the rest, miR-200ba429+miR-200c141 methylation (predicted probability of <0.293 vs >0.293). Three patients (including the patient with insufficient gDNA quality) of the AMC-AJCCII-90 data set were excluded from the uni- and multivariate analyses for lack of differentiation grade information.

Statistics were used as indicated in the figure legends. *P*-values <0.05 were considered significant (Benjamini–Hochberg-corrected, where applicable).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by an AMC Graduate School PhD Scholarship (EF), KWF grants (UvA2009-4416 and UvA2012-5735) KWF/Alpe d'Huzes program CONNECTION, MLDS focus project FP13-07 and NWO gravitation (JPM), KWF grants (UVA2011-4969 and UVA2014-7245), Worldwide Cancer Research (14-1164), the Maag Lever Darm Stichting (MLDS-CDG 14-03) and the European Research Council (ERG-StG 638193) (LV), CityU Start-up Grant for New Faculty (7200455) and an early career scheme grant from the Research Grants Council of Hong Kong SAR (9048033) (XW).

REFERENCES

1 Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S *et al.* Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics* 2012; **5**.
2 De Sousa E, Melo F, Wang X, Jansen M, Fessler E, Trinh A *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med* 2013; **19**: 614–618.
3 Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* 2013; **19**: 619–625.
4 Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO *et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol* 2013; **231**: 63–76.
5 Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer* 2014; **134**: 552–562.
6 Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* 2013; **10**.
7 Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; **21**: 1350–1356.
8 Wang X, Markowetz F, De Sousa E, Melo F, Medema JP, Vermeulen L. Dissecting cancer heterogeneity—an unsupervised classification approach. *Int J Biochem Cell Biol* 2013; **45**: 2574–2579.
9 Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 2010; **463**: 318–325.
10 Yang D, Sun Y, Hu L, Zheng H, Ji P, Pecot CV *et al.* Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell* 2013; **23**: 186–199.
11 Song F, Yang D, Liu B, Guo Y, Zheng H, Li L *et al.* Integrated microRNA network analyses identify a poor-prognosis subtype of gastric cancer characterized by the miR-200 family. *Clin Cancer Res* 2014; **20**: 878–889.
12 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004; **116**: 281–297.
13 Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature* 2008; **455**: 58–63.
14 Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature* 2008; **455**: 64–71.
15 Nicoloso MS, Spizzo R, Shimizu M, Rossi S, Calin GA. MicroRNAs—the micro steering wheel of tumour metastases. *Nat Rev Cancer* 2009; **9**: 293–302.
16 Thiery JP, Acloque H, Huang RYJ, Nieto MA. Epithelial-mesenchymal transitions in development and disease. *Cell* 2009; **139**: 871–890.
17 Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G *et al.* The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* 2008; **10**: 593–601.
18 Park S-M, Gaur AB, Lengyel E, Peter ME. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev* 2008; **22**: 894–907.
19 Tian Y, Pan Q, Shang Y, Zhu R, Ye J, Liu Y *et al.* MicroRNA-200 (miR-200) cluster regulation by achaete scute-like 2 (Ascl2): impact on the epithelial-mesenchymal transition in colon cancer cells. *J Biol Chem* 2014; **289**: 36101–36115.
20 Paterson EL, Kazenwadel J, Bert AG, Khew-Goodall Y, Ruszkiewicz A, Goodall GJ. Down-regulation of the miRNA-200 family at the invasive front of colorectal cancers with degraded basement membrane indicates EMT is involved in cancer progression. *Neoplasia* 2013; **15**: 180–191.
21 Hur K, Toiyama Y, Takahashi M, Balaguer F, Nagasaka T, Koike J *et al.* MicroRNA-200c modulates epithelial-to-mesenchymal transition (EMT) in human colorectal cancer metastasis. *Gut* 2013; **62**: 1315–1326.
22 Neves R, Scheel C, Weinhold S, Honisch E, Iwaniuk KM, Trompeter H-I *et al.* Role of DNA methylation in miR-200c/141 cluster silencing in invasive breast cancer cells. *BMC Res notes* 2010; **3**.
23 Vrba L, Jensen TJ, Garbe JC, Heimark RL, Cress AE, Dickinson S *et al.* Role for DNA methylation in the regulation of miR-200c and miR-141 expression in normal and cancer cells. *PloS One* 2010; **5**.
24 Wiklund ED, Bramsen JB, Hulf T, Dyrskjøt L, Ramanathan R, Hansen TB *et al.* Coordinated epigenetic repression of the miR-200 family and miR-205 in invasive bladder cancer. *Int J Cancer* 2011; **128**: 1327–1334.
25 Davalos V, Moutinho C, Villanueva A, Boque R, Silva P, Carneiro F *et al.* Dynamic epigenetic regulation of the microRNA-200 family mediates epithelial and mesenchymal transitions in human tumorigenesis. *Oncogene* 2012; **31**: 2062–2074.
26 Sugai T, Habano W, Jiao Y-F, Tsukahara M, Takeda Y, Otsuka K *et al.* Analysis of molecular alterations in left- and right-sided colorectal carcinomas reveals distinct pathways of carcinogenesis. *J Mol Diagn* 2006; **8**: 193–201.
27 Markowitz SD, Bertagnolli MM. Molecular basis of colorectal cancer. *N Engl J Med* 2009; **361**: 2449–2460.
28 The Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330–337.
29 Howe EN, Cochrane DR, Richer JK. Targets of miR-200c mediate suppression of cell motility and anoikis resistance. *Breast Cancer Res* 2011; **13**.
30 Burk U, Schubert J, Wellner U, Schmalhofer O, Vincan E, Spaderna S *et al.* A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *EMBO Rep* 2008; **9**: 582–589.
31 Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J *et al.* Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc Natl Acad Sci* 2010; **107**: 15449–15454.
32 Brabletz S, Brabletz T. The ZEB/miR-200 feedback loop--a motor of cellular plasticity in development and cancer? *EMBO Rep* 2010; **11**: 670–677.
33 Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* 2015; **47**: 312–319.
34 Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* 2015; **47**: 320–329.
35 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**: 15545–15550.

© 2016 Macmillan Publishers Limited, part of Springer Nature.

36 Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003; **34**: 267–273.
37 Thiery JP. Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer* 2002; **2**: 442–454.
38 Shioiri M, Shida T, Koda K, Oda K, Seike K, Nishimura M *et al.* Slug expression is an independent prognostic parameter for poor survival in colorectal carcinoma patients. *Br J Cancer* 2006; **94**: 1816–1822.
39 Spaderna S, Schmalhofer O, Hlubek F, Berx G, Eger A, Merkel S *et al.* A transient, EMT-linked loss of basement membranes indicates metastasis and poor survival in colorectal cancer. *Gastroenterology* 2006; **131**: 830–840.
40 Fischer KR, Durrans A, Lee S, Sheng J, Li F, Wong ST *et al.* Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature* 2015; **527**: 472–476.
41 Zheng X, Carstens JL, Kim J, Scheible M, Kaye J, Sugimoto H *et al.* Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature* 2015; **527**: 525–530.
42 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**: 118–127.
43 Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012; **483**: 603–607.
44 Wagner KW, Punnoose EA, Januario T, Lawrence DA, Pitti RM, Lancaster K *et al.* Death-receptor O-glycosylation controls tumor-cell sensitivity to the proapoptotic ligand Apo2L/TRAIL. *Nat Med* 2007; **13**: 1070–1077.
45 Medico E, Russo M, Picco G, Cancelliere C, Valtorta E, Corti G *et al.* The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat Commun* 2015; **6**.
46 Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* 2012; **40**: D144–D149.
47 Prasetyanti PR, Zimberlin C, De Sousa EMF, Medema JP. Isolation and propagation of colon cancer stem cells. *Methods Mol Biol* 2013; **1035**: 247–259.
48 Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003; **31**.
49 Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 2006; **22**: 2825–2827.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

Supplementary Information accompanies this paper on the Oncogene website (http://www.nature.com/onc)