Supporting Information: Long non-coding RNA *MIR31HG* is a bona fide prognostic marker with colorectal cancer cell-intrinsic properties

Peter W. Eide[1,2,3], Ina A. Eilertsen[1,2,3], Anita Sveen[1,2,3] and Ragnhild A. Lothe[1,2,3,*]

[1] Department of Molecular Oncology, Institute for Cancer Research and [2] K.G.Jebsen Colorectal Cancer Research Centre, Oslo University Hospital, Oslo, NO-0424, Norway. [3] Institute for Clinical Medicine, University of Oslo, Oslo, NO-0318, Norway.

[*] corresponding author. Ragnhild A. Lothe, Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital, PO Box 4953 Nydalen, Oslo NO-0424, Norway. Phone: 47-2278-1728; Fax: 47-2278-1745; E-mail: rlothe@rr-research.no

1 Tables

Table S 1: Baseline characteristics of CIT [Marisa 2013], LICR [Jorissen 2009] and Oslo [Sveen 2018] cohorts included in the survival analyses. The numbers in parentheses indicate percentages. adj: adjuvant; CMS: consensus molecular subtypes; MSI/MSS; micro-satellite instable/stable; OS: overall survival; RFS: relapse-free survival

	total (N = 1097)	CIT (N = 562)	LICR (N = 126)	OSLO (N = 409)
age–years				
median	69	68	68	72
range	22-97	22-97	30-92	27-97
gender–no.				
male	581 (53)	309 (55)	70 (56)	202 (49)
female	516 (47)	253 (45)	56 (44)	207 (51)
site–no.				
left	510 (47)	342 (61)	46 (37)	122 (30)
right	444 (41)	220 (39)	51 (41)	173 (43)
rectum	138 (13)	0 (0)	27 (22)	111 (27)
stage–no.				
1	148 (13)	40 (7)	24 (19)	84 (21)
2	467 (43)	257 (46)	57 (45)	153 (37)
3	364 (33)	204 (36)	45 (36)	115 (28)
4	118 (11)	61 (11)	0 (0)	57 (14)
adjuvant chemo–no.				
yes	334 (34)	233 (41)	NA	101 (25)
no	637 (66)	329 (59)	NA	308 (75)
OS–months				
median	55	51	NA	60
range	0-200	0-200	NA–NA	0.66-120
RFS–months				
median	46	43	38	55
range	0-200	0-200	1.6-110	0.66-120
mutated–no.				
KRAS	349 (37)	216 (40)	NA	133 (33)
BRAF	118 (13)	49 (10)	NA	69 (17)
TP53	434 (57)	190 (54)	NA	244 (60)
MSS–no.				
MSI	144 (16)	72 (14)	NA	72 (18)
CMS–no.				
1:immune	173 (18)	89 (17)	20 (19)	64 (20)
2:canonical	413 (44)	232 (45)	42 (40)	139 (43)
3:metabolic	144 (15)	69 (13)	19 (18)	56 (17)
4:mesenchymal	215 (23)	126 (24)	25 (24)	64 (20)
not assigned	152 (14)	46 (8)	20 (16)	86 (21)

Table S 2: Relapse-free survival modeled using Cox proportional univariate hazard function with dichotomized $MIR31HG$ expression. Crude and adjusted values represent univariate and multivariate estimates respectively. MIR31 normal-like, BRAF wt, CMS1 and stage I were used as reference groups. R2=0.19; Wald test p-value<0.001.

	crude (univariate)		adjusted (multivariate)		
	HR	CI95%	HR	CI95%	p-value
MIR31 outlier	2.3	1.7 to 2.9	2.2	1.6 to 3.0	4.4e-06
BRAF V600	1.1	0.8 to 1.5	1.4	0.9 to 2.2	9.0e-02
CMS (CMS1)					
CMS2	1.1	0.8 to 1.6	1.7	1.1 to 2.6	2.5e-02
CMS3	1.0	0.7 to 1.5	1.4	0.8 to 2.3	2.3e-01
CMS4	1.9	1.4 to 2.7	2.2	1.4 to 3.3	4.5e-04
stage (I)					
II	1.9	1.2 to 3.1	2.7	1.5 to 5.1	1.8e-03
III	3.5	2.2 to 5.5	3.8	2.0 to 7.1	3.6e-05
IV	11.4	7.1 to 18.5	13.1	6.9 to 24.9	3.8e-15

Table S 3: Relapse-free survival modeled using Cox proportional univariate hazard function with continious $MIR31HG$ expression (m31). Crude and adjusted values represent univariate and multivariate estimates respectively. BRAF wt, CMS1 and stage I were used as reference groups. R2=0.18; Wald test p-value<0.001.

	crude (univariate)		adjusted (multivariate)		
	HR	CI95%	HR	CI95%	p-value
m31	1.3	1.2 to 1.4	1.2	1.1 to 1.3	4.3e-05
BRAF V600	1.1	0.8 to 1.5	1.4	0.9 to 2.1	9.4e-02
CMS (CMS1)					
CMS2	1.1	0.8 to 1.6	1.7	1.1 to 2.7	2.2e-02
CMS3	1.0	0.7 to 1.5	1.4	0.9 to 2.4	1.7e-01
CMS4	1.9	1.4 to 2.7	2.2	1.4 to 3.3	3.8e-04
stage (I)					
II	1.9	1.2 to 3.1	2.7	1.4 to 5.0	2.1e-03
III	3.5	2.2 to 5.5	3.8	2.0 to 7.2	3.1e-05
IV	11.4	7.1 to 18.5	12.6	6.6 to 24.0	1.0e-14

Table S 4: Recurrence type according to $MIR31HG$ dichotomized expression for stage II+III colorectal cancers. Numbers refer to cases and column percentages are indicated in parentheses. Data are for Oslo cohort only [Sveen 2018].

	MIR31 normal-like	MIR31 outlier
distant	32 (14%)	10 (28%)
localized	2 (0.86%)	0 (0%)
localized&distant	7 (3%)	3 (8.3%)
none	191 (82%)	23 (64%)

2 Figures and text

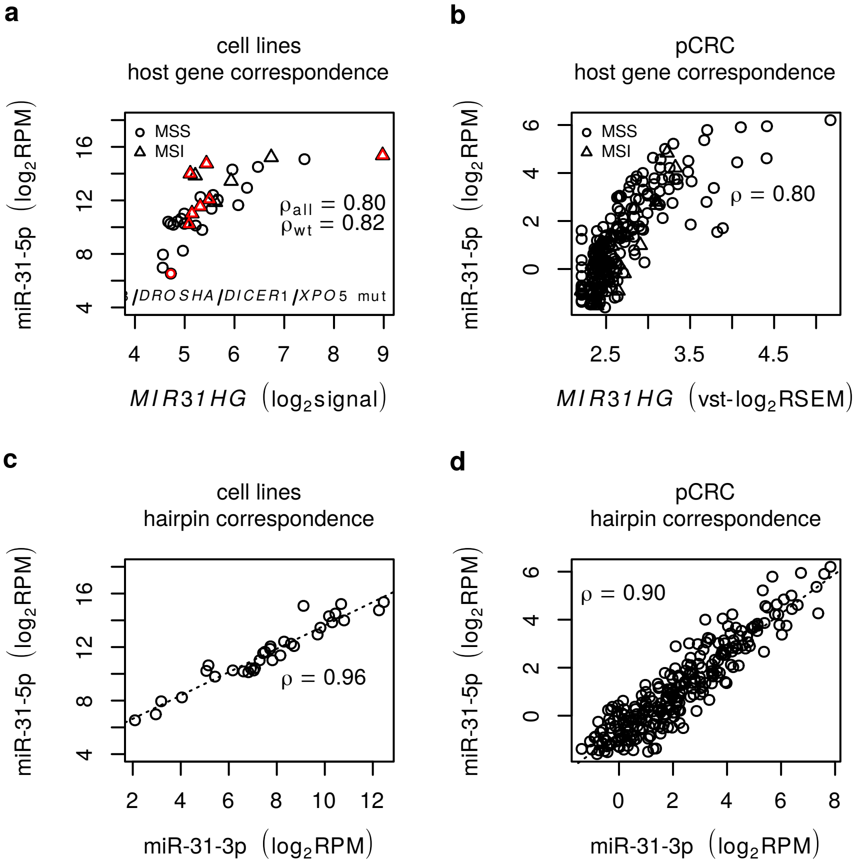

Figure S 1: ***MIR31HG*****, miR-31-5p and miR-31-3p are highly correlated in CRC cell lines and primary tumors** (**a**) Scatter plot shows *MIR31HG*/mir-31-5p expression in CRC cell lines[1]. MSI/MSS samples are indicated with triangles/circles. Cell lines carrying mutations in indicated miRNA-processing genes are highlighted in red. ρ represents the corresponding Spearman's correlations. (**b**) Scatter plot shows *MIR31HG*/mir-31-5p expression in TCGA primaries[2]. MSI/MSS samples are indicated with triangles/circles and ρ is the Spearman's correlation. (**c**) Scatter plot show miR-31-3p/mir-31-5p expression in CRC cell lines[1]. (**d**) Scatter plot shows miR-31-3p/mir-31-5p expression in TCGA primaries[2]. ρ indicates the Spearman's rank correlation. mut; mutated; RPM: reads per millions miRNA mapped; RSEM: RNA-Seq by expectation maximization[3]; vst: variance stabilizing transformed[4]

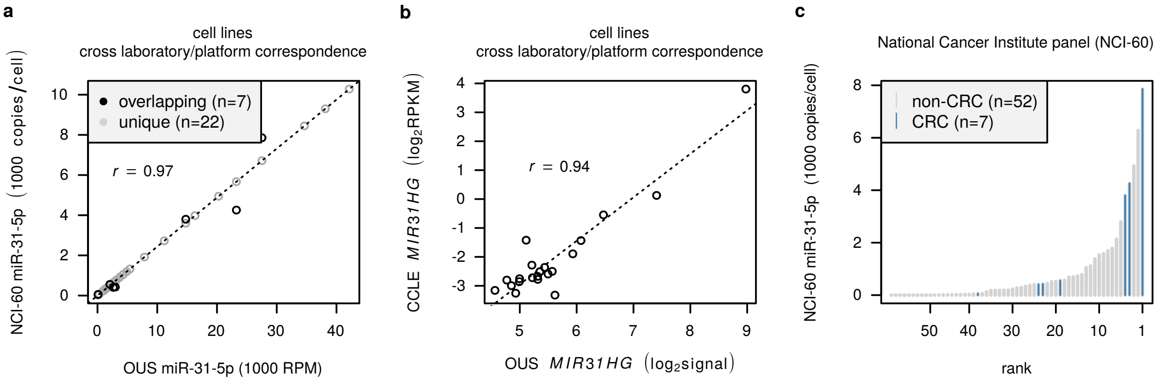

Figure S 2: **Colorectal cancer cell line miR-31-5p and *MIR31HG* expression estimates are consistent across research laboratories and technological platforms and present large ranges in expression values.** (**a**) Scatter plot shows correspondence between in-house smRNA-seq reads per million (RPM) and Gaur *et al.*[5] qRT-PCR based cellular abundance estimates for 7 overlapping cell lines (black circles). Gray circles represent non-overlapping cell lines with in-house RPM values and copies/cell estimates from the linear least squares regression model (dotted line). The Pearson's correlation coefficient (*r*) for overlapping samples are indicated. (**b**) Scatter plot depicts in-house Affymetrix microarray and CCLE[6] RNA-seq derived *MIR31HG* expression values for 21 overlapping CRC cell lines with the corresponding Pearson's correlation coefficient (*r*). A pseudo count of 1/10 was added to RNA-seq values to avoid *log* of zero. (**c**) Barplot illustrates miR-31-5p expression across the NCI-60 panel with CRC samples highlighted in blue. qRT-PCR based cellular abundance estimates are from Gaur *et al.*[5].

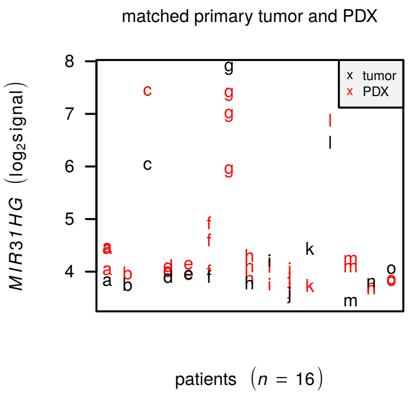

Figure S 3: **MIR31 expression levels are maintained in xenografts indicating cancer cell-intrinsic expression**. Letters indicate separate patients with black and red indicating primary tumor and PDX samples respectively. Data are from Linnekamp *et al.* and were retrieved from GEO with accession identifier GSE100480[7]. PDX: patient-derived xenograft

2.1 *MIR31HG* and miR-31-5p are analytically robust

To assess analytical robustness, we tested whether MIR31HG/miR-31-5p expression in CRC cell lines is either overtly responsive to environmental changes (*e.g.* differences in culturing between laboratories) or maintained across time and conditions (*i.e.* whether MIR31 activity represents a cell state marker). To answer this, we compared in-house miR-31-5p and *MIR31HG* expression levels against public datasets with overlapping cell lines. For miR-31-5p the Pearson's correlation among seven cell lines quantified by qRT-PCR as part of the NCI60 panel[5] and our smRNA-seq-derived estimates was 0.97 (Supplementary Fig.S2a)[5]. Similarly, the correlation in *MIR31HG* expression between in-house Affymetrix microarrays and the total RNA-seq data from Cancer Cell Line Encyclopedia (CCLE)[6] was 0.94 (*n*=21, Supplementary Fig.S2b)[6]. Thus, differences in *MIR31HG*/miR-31-5p expression between CRC cell lines are robustly recapitulated across molecular levels, research groups and technological platforms.

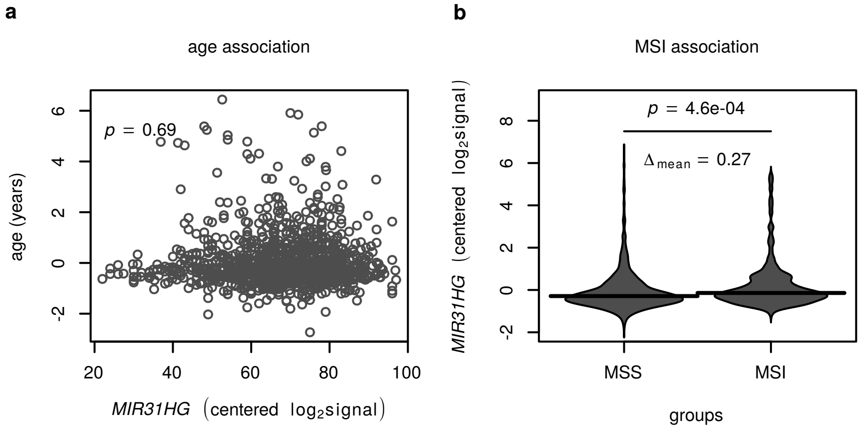

Figure S 4: ***MIR31HG* expression is not associated with age, but MSI-status**. (**a**) Scatterplot for age and *MIR31HG* expression. The *p*-value is from Pearson's correlation test. (**b**) Beanplot depicts distribution in *MIR31HG* expression for MSS and MSI samples. The horizontal bars represent the group-wise medians and the *p*-value is from Wilcoxon rank sum test. Δ_{mean} states the mean log_2 fold-change. Data are from CIT[8], LICR[9] and Oslo[10] cohorts. MSI/MSS: micro-satellite instable/stable

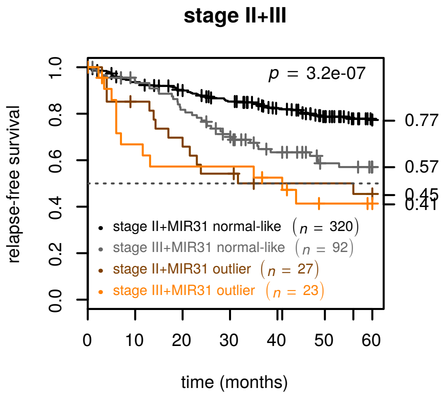

Figure S 5: **MIR31 outlier status stratifies non-adjuvant treated stage II and III primary colorectal cancers**. Kaplan-Meier plot shows relapse-free survival for non-adjuvant chemotherapy treated stage II+III pCRCs stratified according to stage and MIR31 status. The *p*-value is from Wald tests for Cox model. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

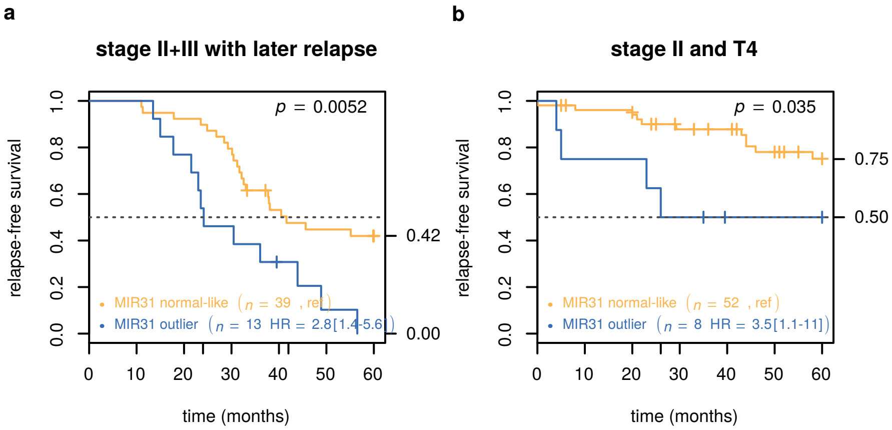

Figure S 6: **MIR31 outlier status stratifies subgroups of patients with primary colorectal cancers.** (**a**) Kaplan-Meier plot shows relapse-free survival (RFS) for stage II+III pCRCs experiencing distant relapse stratified according to stage and MIR31 status. (**b**) Kaplan-Meier plot shows RFS for stage II/T4 pCRCs stratified according to MIR31 status. The *p*-value is from Wald tests for Cox model. Data are from Oslo[10] and CIT[8] cohorts. Recurrence type was available only for the Oslo cohort. LICR cohort lacked information on T-stage.

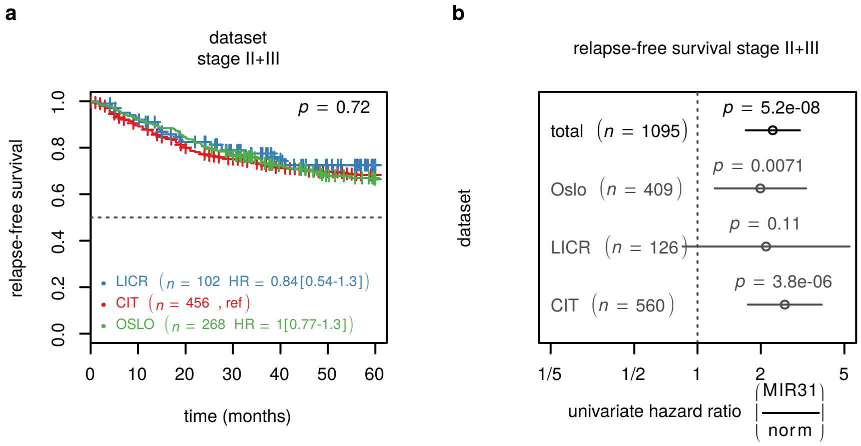

Figure S 7: **Relapse-free survival characteristics are largely consistent across the Oslo, CIT and LICR cohorts**. (**a**) Kaplan-Meier plot shows relapse-free survival for stage II+III samples stratified by cohort. (**b**) Plot visualizes pooled and per cohort hazard ratios with 95% confidence interval associated with MIR31 outlier status. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

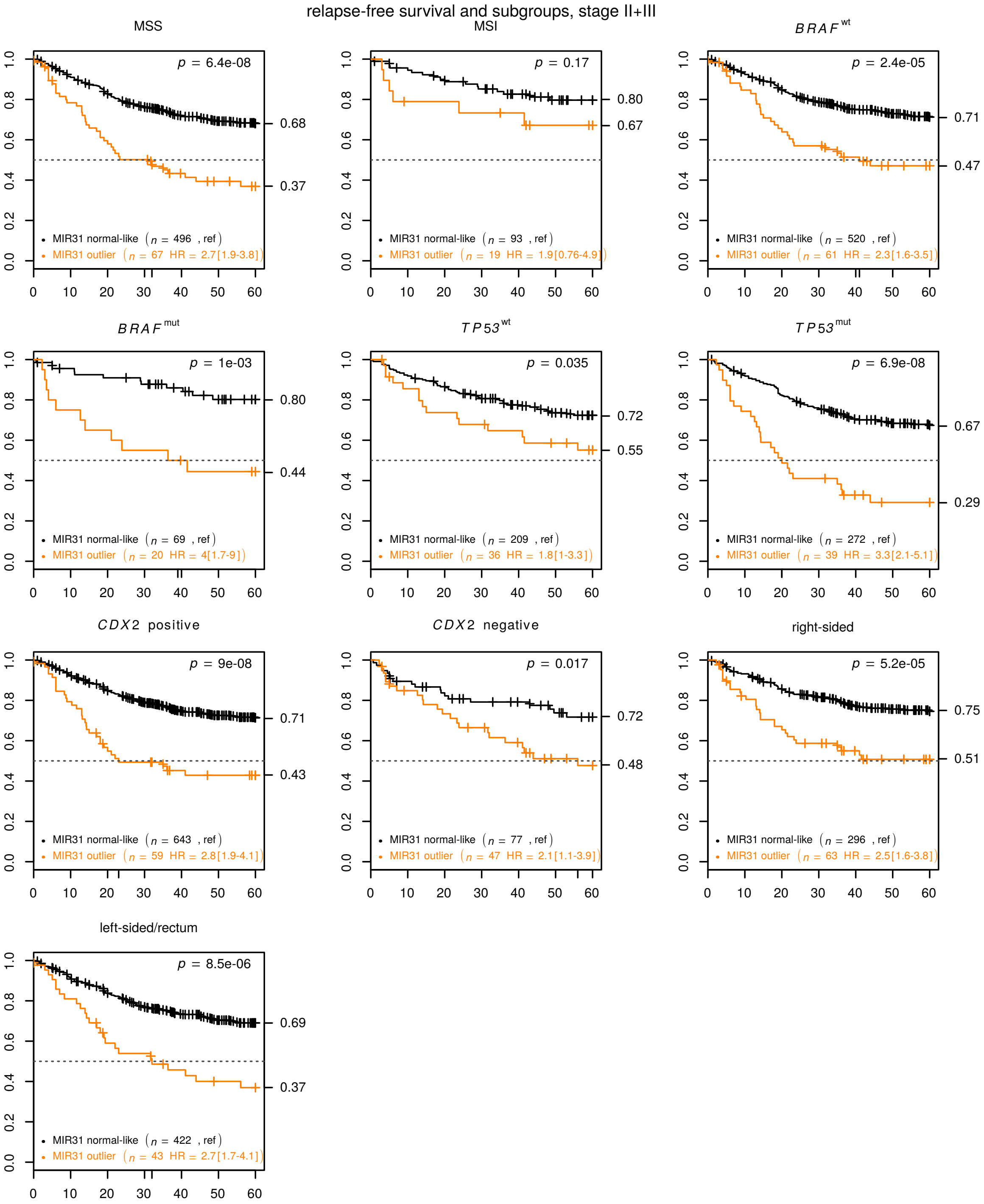

Figure S 8: **MIR31 outliers present inferior relapse-free survival within important subgroups**. Kaplan-Meier plots for different patient stage II+III subsets. The Kaplan-Meier titles indicate sample subset included. For instance, in the upper left plot, samples were stratified by MSS/MIR31. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

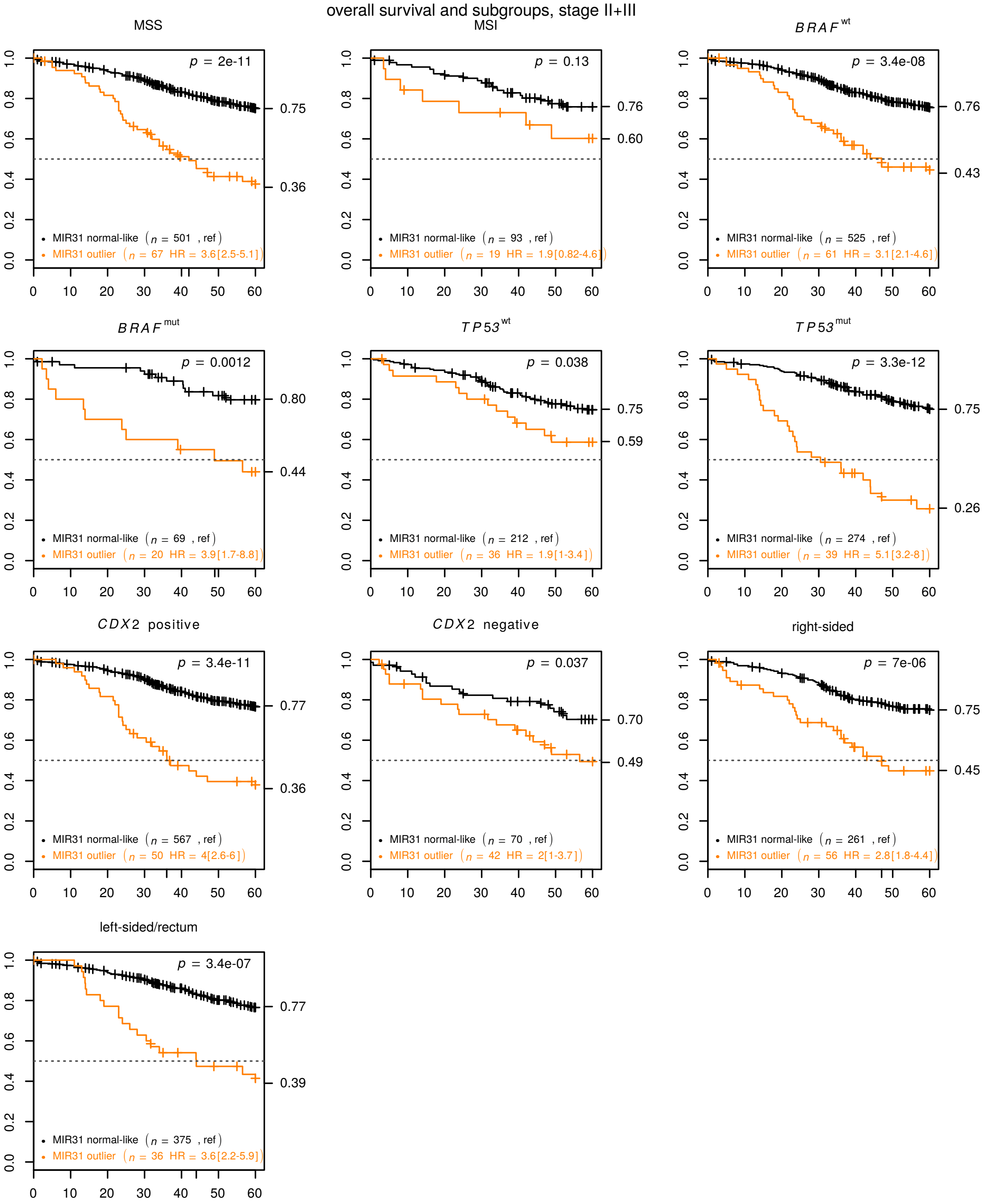

Figure S 9: **MIR31 outliers present inferior overall survival within important subgroups**. Kaplan-Meier plots for different patient stage II+III subsets. The Kaplan-Meier titles indicate sample subset included. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

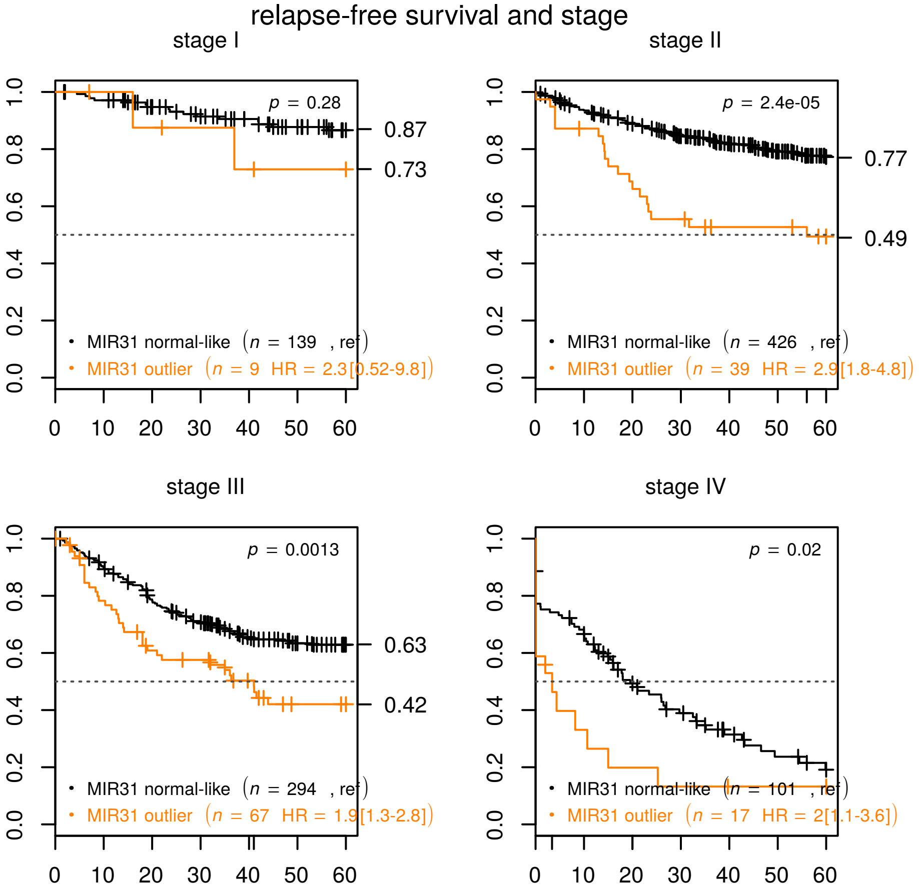

Figure S 10: **MIR31 status is associated with significantly shorter relapse-free survival for stage II–IV colorectal cancers**. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

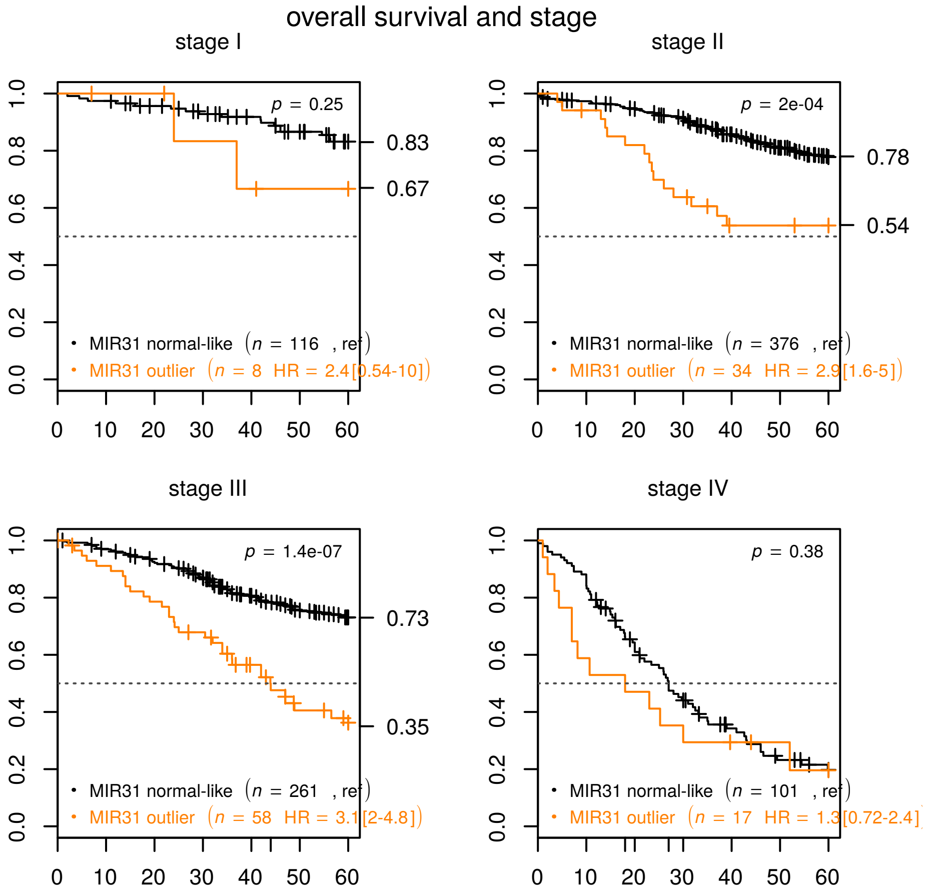

Figure S 11: **MIR31 status is associated with significantly shorter overall survival for stage II–III colorectal cancers**. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

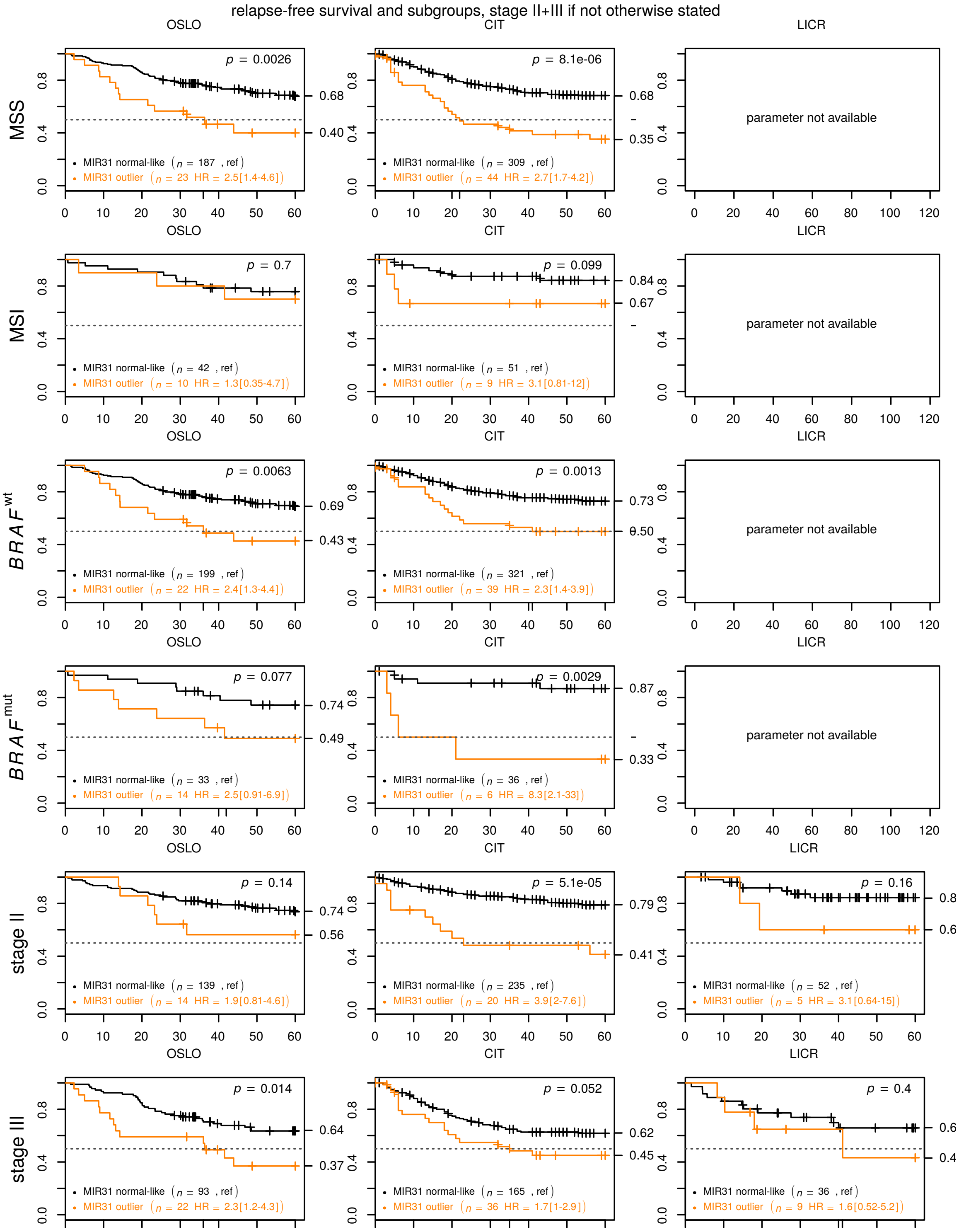

Figure S 12: **Kaplan-Meier plots for subgroups per dataset.** The Kaplan-Meier titles indicate dataset while subset are indicated to the left of each row. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

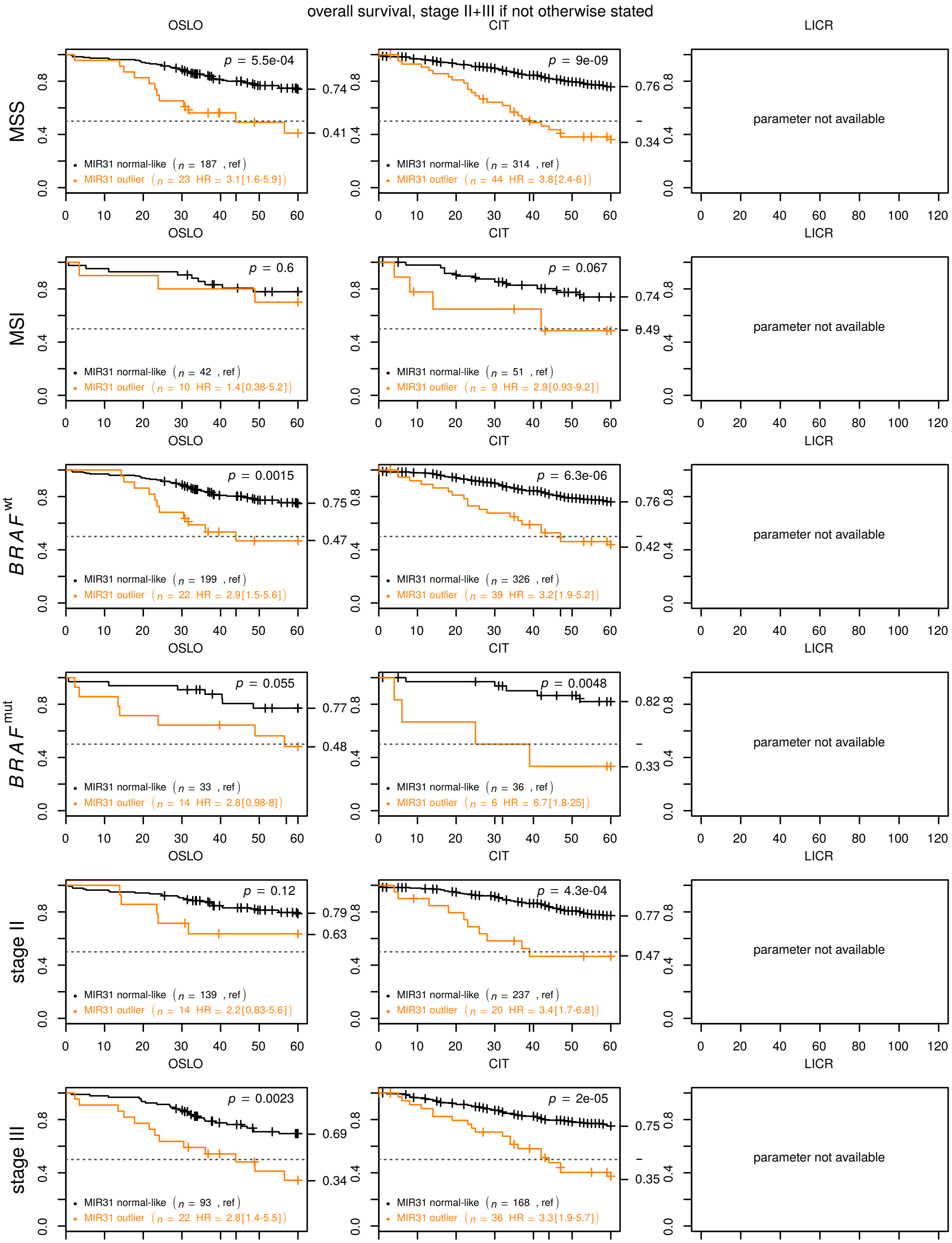

Figure S 13: **Kaplan-Meier plots for subgroups per dataset.** The Kaplan-Meier titles indicate dataset while subset are indicated to the left of each row. Data are from CIT[8], LICR[9] and Oslo[10] cohorts.

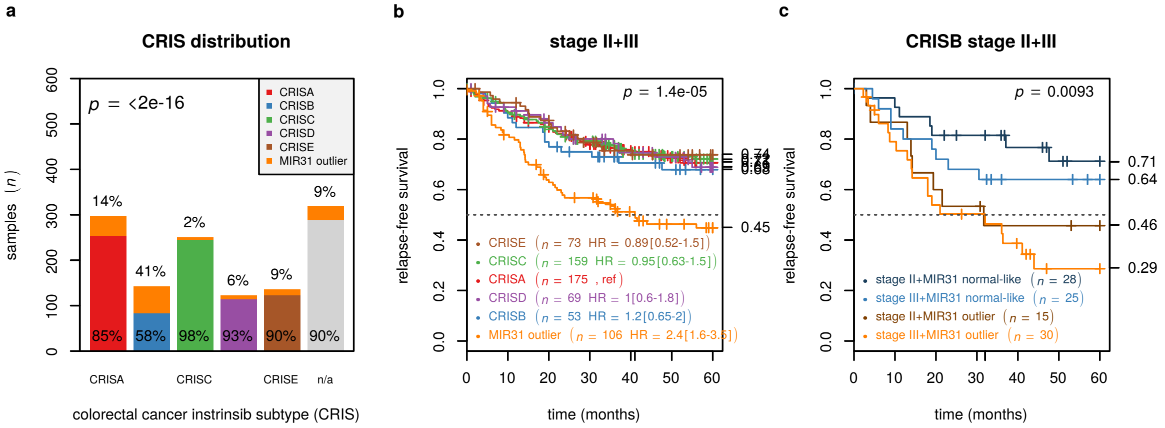

Figure S 14: **MIR31 outlier samples are predictive within the poor prognostic Isella *et al*. CRC intrinsic subtype B.** (**a**) Barplot illustrates distribution of CRIS[11] and MIR31 outliers. The *p*-value is from χ^2test. (**b**) Kaplan-Meier plot shows relapse-free survival for stage II+III pCRC patients stratified by CRIS and MIR31 outlier status. (**c**) Kaplan-Meier plot shows relapse-free survival for poor prognostic CRISB stage II+III pCRC patients stratified by MIR31 and stage. Data are from CIT[8], LICR[9] and Oslo[10] cohorts. CRIS: CRC intrinsic subtype; HR: hazard ratio; MAD: median absolute deviation; n/a: not assigned

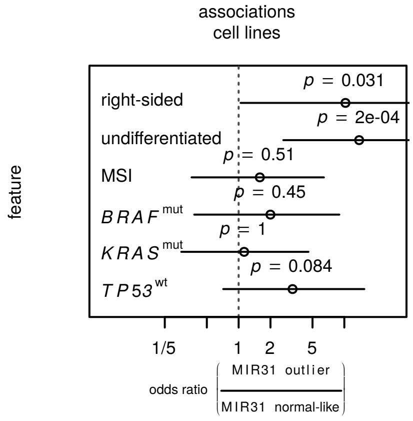

Figure S 15: **Plot visualizes odds ratios with 95% confidence intervals for cell line panel (*n*=78) and selected clinicopathological and molecular variables**. *p*-values are from Fisher's exact tests. Panel is a merged set of Oslo[1], CCLE[6] and Astra-Zeneca [no reference] cell lines. MSI: micro-satellite instable; mut: mutated; wt: wild type

2.2 Cell lines are representative models

As for pCRC, cell line *MIR31HG* expression distribution was heavily right-skewed with mode near microarray background. Using the same thresholding as applied to the pCRCs, 13/78 (17%) cell lines were classified as MIR31 outliers (Manuscript Fig.4a). For gene expression-based undifferentiated state, a proxy for low differentiation grade, 2/49 (4%) colon-like and 11/29 (39%) undifferentiated cell lines displayed MIR31 outlier expression (OR=11 [2.6–141], p=2×10^{-4}, Fisher's exact test, Supplementary Fig.S15). Considering MSS only, 2/38 (5%) colon-like and 6/16 (38%) undifferentiated cell lines were MIR31 outliers (OR=10 [1.5–120], p=6×10^{-3}). Likewise, for 38 samples where details on anatomical sub-site were available, 6/7 (86%) MIR31 outliers were cell lines derived from right-sided tumors compared to 1/19 (5%) for non-outliers (OR=10 [1–525], *p*=0.03). With regards to MSI and *BRAF*-status, the highest median *MIR31HG* expression was observed in MSI/*BRAF*$^{\mathrm{V600}}$ samples (Supplementary Fig.S16. However, many distinct *MIR31HG* outliers were both MSS and *BRAF*$^{\mathrm{wt}}$. Thus, *BRAF* mutation is unlikely to be the (sole) causal driver of *MIR31HG* activation. We have previously reported CRC cell line consensus molecular subtype (CMS) classification[10]. As expected based on the pCRCs and differences in differentiation-state, MIR31 status was also significantly associated with CMS (*p*=<2e-16, Fisher's exact test, Manuscript Fig.2b). Specifically, 0/23 CMS2-like and 20-29% of non-CMS2 subtypes were classified as MIR31 outliers. Summarized, the CRC cell line panel recapitulates MIR31 associations described for pCRCs.

TargetScan records 867 human transcripts with miR-31-5p seed matches whereof 464 show evidence of evolutionary conservation[12]. Despite the large difference in *MIR31HG*/miR-31-5p abundance between the two groups, there was no significant overall depletion of target transcripts in MIR31 outlier samples (Supplementary Fig.S17a). We further tested whether repression could be observed at the protein-level by taking advantage of mass spectroscopy derived protein abundances for 41 overlapping CRC cell lines[13]. As for the RNA-level, miR-31-5p targets showed no trend towards having lower relative protein expression in MIR31 outlier cell lines (Supplementary Fig.S17b).

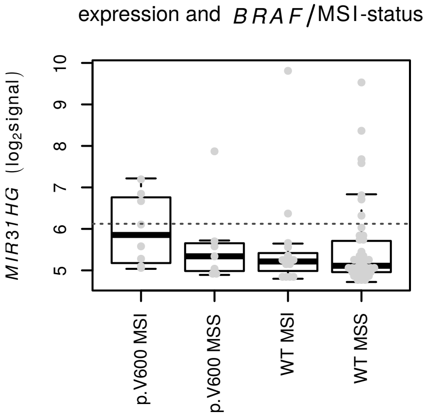

Figure S 16: **Boxplot shows *MIR31HG* expression stratified by *BRAF*-mutation and MSI-status**. The dotted line represents the dichotomization threshold. Panel is a merged set of Oslo[1], CCLE[6] and Astra-Zeneca [no reference] cell lines.

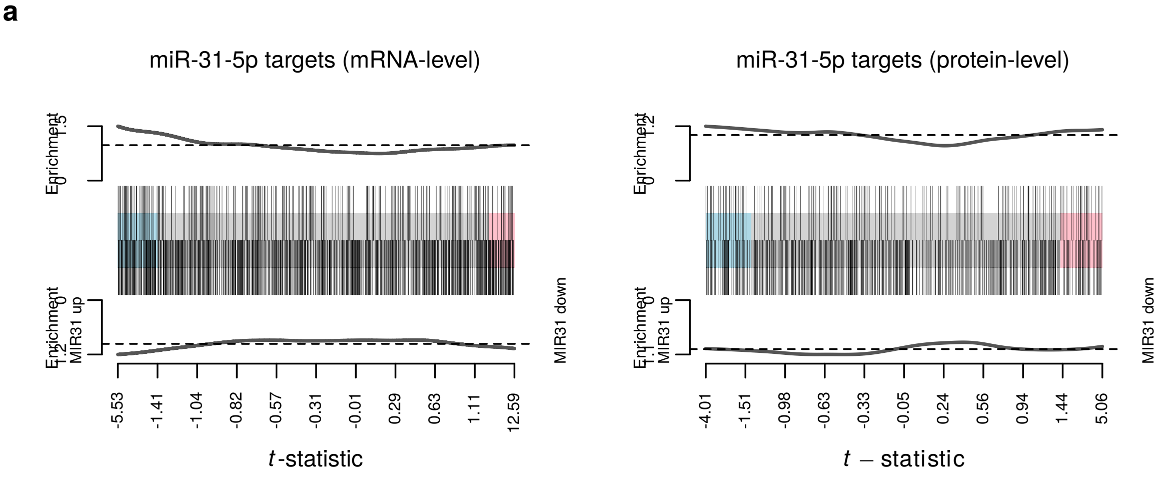

Figure S 17: **No overall difference in expression of miR-31-5p targets between MIR31 outlier subgroups at either the RNA or protein levels were observed for cell lines.** Barcodeplots show enrichment scores for ranked *t*-statistic from differential expression analysis comparing MIR31 outliers against remaining samples. Top panels include only genes with evolutionary conserved miR-31-5p binding sites while lower panels include all predicted targets. Gene lists are from TargetScan[12]. 78 and 41 cell lines were included in the mRNA and protein level analysis, respectively. RNA expression data are a merged set of Oslo[1], CCLE[6] and Astra-Zeneca [no reference] cell lines. Protein expression data are from Supplementary Information in Roumeliotis *et al.*[13].

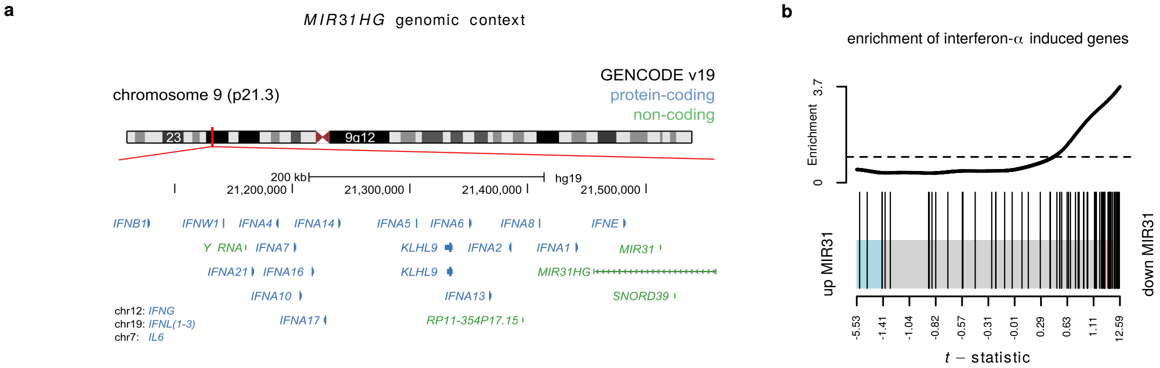

Figure S 18: ***MIR31HG* is located within an interferon gene cluster and cell line *MIR31HG* expression is associated with IFNA responsive genes.** (**a**) Figure shows interferon gene cluster with *MIR31HG* in immediate proximity. (**b**) Barcodeplot visualizes the differential expression enrichment of genes induced by IFN-α. Gene expression data are for Oslo[1], CCLE[6] and Astra-Zeneca [no reference] cell lines. Moserle *et al.* HGU133p2 gene expression data of ovarian cancer cell lines exposed to IFN-α were used to identify genes upregulated upon exposure (log_2fold-change>1 and FDR adjusted-p<0.05, GEO identifier GSE10943)[14]. chr: chromosome; IFNA: interferon-α; kb: kilobases

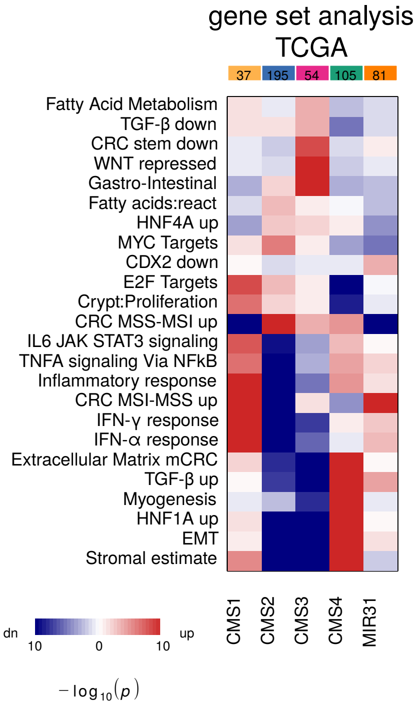

Figure S 19: **MIR31 outlier samples present upregulation of interferon-α/γ signatures compared to non-CMS1 subgroups**. Heatmap visualizes results from Camera[15] gene set analysis comparing CMS and MIR31 outliers. Color saturation indicates increasing significance and red and blue relative up- and downregulation, respectively. Gene signatures are from MSigDB Hallmarks[16,17] (v6.2). Gene expression data are from TCGA[2]. CMS: consensus molecular subtypes; EMT: epithelial mesenchymal transition; FDR: false discovery rate[18]; IFNA/G: interferon-α/γ; MAD: median absolute deviation; MSS/MSI; micro-satellite stable/instable; pCRC; primary colorectal cancer

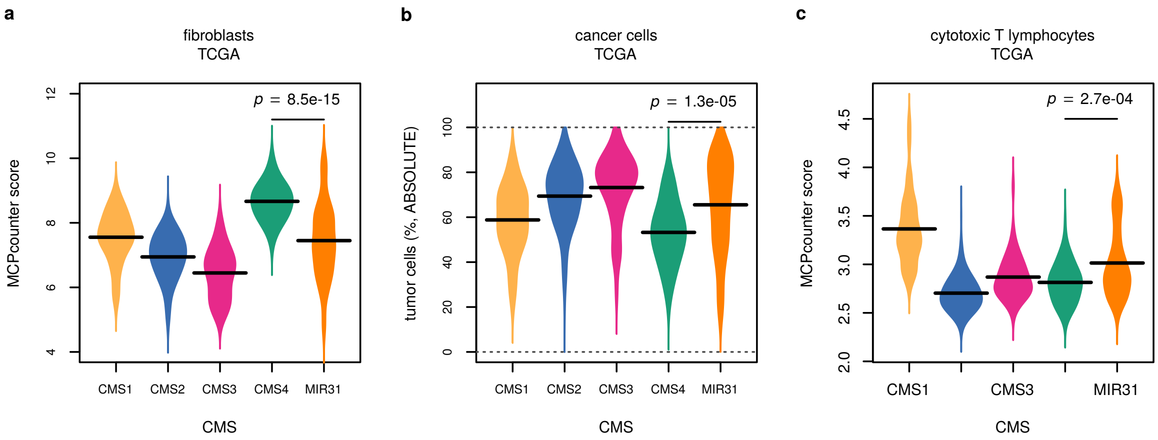

Figure S 20: **MIR31 outlier tumors are less stromal than CMS4.** (**a**) Beanplot shows MCPcounter[19] fibroblast infiltration score stratified by CMS and MIR31 outlier expression for TCGA cohort[2]. (**b**) Plot depicts fibroblast infiltration score for CIT cohort[8]. (**c**) Beanplot shows MCPcounter[19] cytotoxic T lymphocyte infiltration score stratified by CMS and MIR31 outlier expression. The *p*-values are from Wilcoxon rank sum tests. Gene expression data are from TCGA[2].

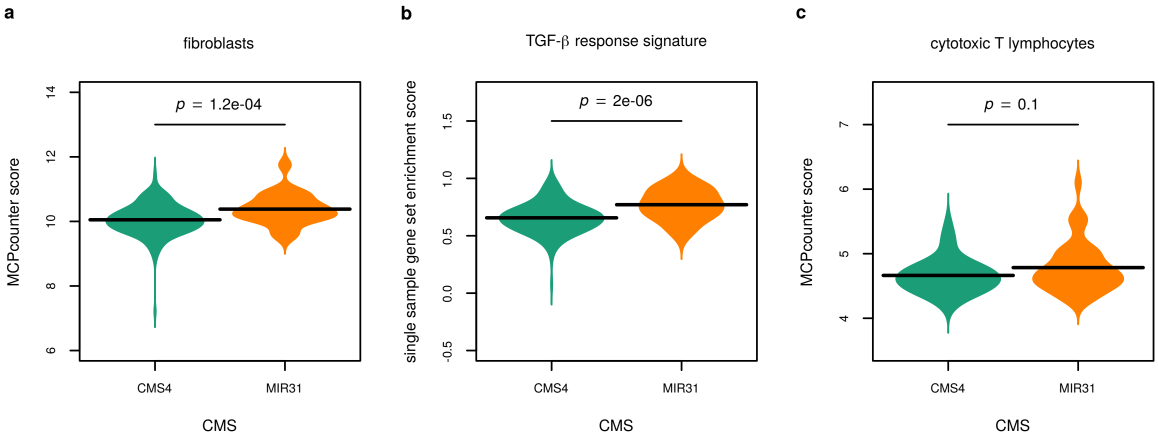

Figure S 21: **Among CMS4 tumors, MIR31 outliers exhibhit more stromal infiltration than samples with normal-like *MIR31HG* expression**. (**a**) Beanplot shows MCPcounter[19] fibroblast infiltration score stratified by CMS and MIR31 outlier expression. (**b**). Beanplot depicts distributions in single sample gene set enrichment scores for normal colonic fibroblast TGF-β response signature stratified by MIR31 outlier expression.[20]. (**c**) Beanplot shows MCPcounter[19] cytotoxic T lymphocyte infiltration score stratified by CMS and MIR31 outlier expression. The *p*-values are from Wilcoxon rank sum tests. Data include 226 CMS4 samples from CIT[8], LICR[9] and Oslo[10] cohorts.

2.3 Survival models

Adding patient gender, patient age (<=70 vs >70), tumor localization (right vs left/rectum), adjuvant chemotherapy (yes/no), MSI, *CDX2* expression (dichotomized at 15.6 percentile[pilati_cdx2_2017]) and *KRAS*-status did not substantially improve the presented multivariable model. Replacing with CRC intrinsic subtypes (CRIS)[11] yielded comparable results (see output below).

```
Call:
survival::coxph(formula = survFit ~ mir + braf_mut + cms + survival::strata(stage) +
    gender + agegrp + side + chemo + msi + cdxneg + kras_mut,
    data = pool)

  n= 735, number of events= 255
   (530 observations deleted due to missingness)

                            coef exp(coef) se(coef)      z Pr(>|z|)
mirMIR31 outlier          0.6549    1.9250   0.1867  3.51  4.5e-04 ***
braf_mut1                 0.4865    1.6267   0.2587  1.88  6.0e-02 .
cmsCMS2                   0.5259    1.6919   0.3036  1.73  8.3e-02 .
cmsCMS3                   0.2810    1.3244   0.3169  0.89  3.8e-01
cmsCMS4                   0.8372    2.3098   0.2773  3.02  2.5e-03 **
survival::strata(stage)2  1.0864    2.9635   0.3224  3.37  7.5e-04 ***
survival::strata(stage)3  1.4975    4.4706   0.3346  4.48  7.6e-06 ***
survival::strata(stage)4  2.7084   15.0056   0.3411  7.94  2.0e-15 ***
gendermale                0.1815    1.1990   0.1329  1.37  1.7e-01
agegrp(70,100]            0.0818    1.0853   0.1385  0.59  5.5e-01
sideright                -0.0840    0.9194   0.1532 -0.55  5.8e-01
chemoyes                 -0.3597    0.6979   0.1581 -2.28  2.3e-02 *
msi                      -0.1938    0.8239   0.2769 -0.70  4.8e-01
cdxnegTRUE                0.1426    1.1533   0.2086  0.68  4.9e-01
kras_mut1                 0.0862    1.0900   0.1508  0.57  5.7e-01
---
Signif. codes:  0 '***' 1e-03 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
mirMIR31 outlier             1.925     0.5195     1.335     2.776
braf_mut1                    1.627     0.6148     0.980     2.701
cmsCMS2                      1.692     0.5910     0.933     3.068
cmsCMS3                      1.324     0.7551     0.712     2.465
cmsCMS4                      2.310     0.4329     1.341     3.977
survival::strata(stage)2     2.963     0.3374     1.575     5.575
survival::strata(stage)3     4.471     0.2237     2.320     8.614
survival::strata(stage)4    15.006     0.0666     7.689    29.284
gendermale                   1.199     0.8340     0.924     1.556
agegrp(70,100]               1.085     0.9214     0.827     1.424
sideright                    0.919     1.0876     0.681     1.242
chemoyes                     0.698     1.4330     0.512     0.951
msi                          0.824     1.2138     0.479     1.418
cdxnegTRUE                   1.153     0.8671     0.766     1.736
kras_mut1                    1.090     0.9174     0.811     1.465

Concordance= 0.709  (se = 0.017 )
Rsquare= 0.195   (max possible= 0.987 )
Likelihood ratio test= 159  on 15 df,   p=<2e-16
Wald test            = 173  on 15 df,   p=<2e-16
Score (logrank) test = 207  on 15 df,   p=<2e-16
```

```
Call:
survival::coxph(formula = survFit ~ mir + braf_mut + cris + survival::strata(stage) +
    gender + agegrp + side + chemo + msi + cdxneg + kras_mut,
    data = pool)

  n= 638, number of events= 236
   (627 observations deleted due to missingness)

                             coef exp(coef) se(coef)     z Pr(>|z|)
mirMIR31 outlier          0.57286   1.77333  0.20474  2.80  5.1e-03 **
braf_mut1                 0.21801   1.24360  0.25641  0.85  4.0e-01
crisCRISB                 0.09177   1.09611  0.21927  0.42  6.8e-01
crisCRISC                -0.03360   0.96696  0.21747 -0.15  8.8e-01
crisCRISD                 0.05960   1.06141  0.22278  0.27  7.9e-01
crisCRISE                -0.17517   0.83932  0.21979 -0.80  4.3e-01
survival::strata(stage)2  0.65210   1.91956  0.28543  2.28  2.2e-02 *
survival::strata(stage)3  1.02528   2.78788  0.30192  3.40  6.8e-04 ***
survival::strata(stage)4  2.38076  10.81308  0.30899  7.71  1.3e-14 ***
gendermale                0.17557   1.19193  0.13607  1.29  2.0e-01
agegrp(70,100]            0.07321   1.07595  0.14711  0.50  6.2e-01
sideright                 0.04576   1.04682  0.15614  0.29  7.7e-01
chemoyes                 -0.14765   0.86273  0.16885 -0.87  3.8e-01
msi                      -0.63169   0.53169  0.27041 -2.34  1.9e-02 *
cdxnegTRUE                0.21914   1.24501  0.22236  0.99  3.2e-01
kras_mut1                 0.00858   1.00861  0.16156  0.05  9.6e-01
---
Signif. codes:  0 '***' 1e-03 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
mirMIR31 outlier             1.773     0.5639     1.187     2.649
braf_mut1                    1.244     0.8041     0.752     2.056
crisCRISB                    1.096     0.9123     0.713     1.685
crisCRISC                    0.967     1.0342     0.631     1.481
crisCRISD                    1.061     0.9421     0.686     1.643
crisCRISE                    0.839     1.1914     0.546     1.291
survival::strata(stage)2     1.920     0.5210     1.097     3.359
survival::strata(stage)3     2.788     0.3587     1.543     5.038
survival::strata(stage)4    10.813     0.0925     5.901    19.813
gendermale                   1.192     0.8390     0.913     1.556
agegrp(70,100]               1.076     0.9294     0.806     1.436
sideright                    1.047     0.9553     0.771     1.422
chemoyes                     0.863     1.1591     0.620     1.201
msi                          0.532     1.8808     0.313     0.903
cdxnegTRUE                   1.245     0.8032     0.805     1.925
kras_mut1                    1.009     0.9915     0.735     1.384

Concordance= 0.705  (se = 0.018 )
Rsquare= 0.195   (max possible= 0.989 )
Likelihood ratio test= 139  on 16 df,   p=<2e-16
Wald test            = 161  on 16 df,   p=<2e-16
Score (logrank) test = 195  on 16 df,   p=<2e-16
```

3 R packages

Packages explicitly loaded include beanplot[21], Biobase[22], CMSclassifier[23], genefilter[24], Greg[25], kableExtra[26], knitr[27], limma[28], MCPcounter[19], qwraps2[29], RColorBrewer[30]; survival[31], and sva[32].

4 References

1. Berg KCG, Eide PW, Eilertsen IA, Johannessen B, Bruun J, Danielsen SA, Bjørnslett M, Meza-Zepeda LA, Eknæs M, Lind GE, Myklebost O, Skotheim RI, Sveen A, Lothe RA. Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies. *Mol Cancer [Internet]* 2017 [cited 2017 Sep 15];16:116. Available from: https://link.springer.com/article/10.1186/s12943-017-0691-y

2. TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature [Internet]* 2012;487:330–7. Available from: http://dx.doi.org/10.1038/nature11252

3. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform [Internet]* 2011 [cited 2018 Apr 9];12:323. Available from: https://doi.org/10.1186/1471-2105-12-323

4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol [Internet]* 2014 [cited 2015 Oct 17];15:550. Available from: http://genomebiology.com/2014/15/12/550/abstract

5. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, Ambros VR, Israel MA. Characterization of MicroRNA expression levels and their biological correlates in human cancer cell lines. *Cancer Res [Internet]* 2007;67:2456–68. Available from: http://cancerres.aacrjournals.org/content/67/6/2456.abstract

6. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Silva M de, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature [Internet]* 2012;483:603–307. Available from: http://dx.doi.org/10.1038/nature11003

7. Linnekamp JF, Hooff SR van, Prasetyanti PR, Kandimalla R, Buikhuisen JY, Fessler E, Ramesh P, Lee KAST, Bochove GGW, Jong JH de, Cameron K, Leersum R van, Rodermond HM, Franitza M, Nürnberg P, Mangiapane LR, Wang X, Clevers H, Vermeulen L, Stassi G, Medema JP. Consensus molecular subtypes of colorectal cancer are recapitulated in in vitro and in vivo models. *Cell Death Differ [Internet]* 2018 [cited 2018 Mar 6];25:616–33. Available from: https://www.nature.com/articles/s41418-017-0011-5

8. Marisa L, Reyniès A de, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi M-C, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou J-F, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V. Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PLoS Med [Internet]* 2013 [cited 2015 Mar 2];10:e1001453. Available from: http://dx.doi.org/10.1371/journal.pmed.1001453

9. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, Kerr D, Aaltonen LA, Arango D, Kruhøffer M, Orntoft TF, Andersen CL, Gruidl M, Kamath VP, Eschrich S, Yeatman TJ, Sieber OM. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage b and c colorectal cancer. *Clin Cancer Res [Internet]* 2009;15:7642–51. Available from: http://clincancerres.aacrjournals.org/content/15/24/7642

10. Sveen A, Bruun J, Eide PW, Eilertsen IA, Ramirez L, Murumägi A, Arjama M, Danielsen SA, Kryeziu K, Elez E, Tabernero J, Guinney J, Palmer HG, Nesbakken A, Kallioniemi O, Dienstmann R, Lothe RA. Colorectal cancer consensus molecular subtypes translated to preclinical models uncover potentially targetable cancer cell dependencies. *Clin Cancer Res [Internet]* 2018 [cited 2018 Feb 16];24:794–806. Available from: http://clincancerres.aacrjournals.org/content/24/4/794

11. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, Petti C, Fiori A, Orzan F, Senetta R, Boccaccio C, Ficarra E, Marchionni L, Trusolino L, Medico E, Bertotti A. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat Commun [Internet]* 2017 [cited 2017 Sep 11];8:ncomms15107. Available from: https://www.nature.com/articles/ncomms15107

12. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife [Internet]* 2015 [cited 2016 Mar 3];4:e05005. Available from: http://elifesciences.org/content/4/e05005v1

13. Roumeliotis TI, Williams SP, Gonçalves E, Alsinet C, Del Castillo Velasco-Herrera M, Aben N, Ghavidel FZ, Michaut M, Schubert M, Price S, Wright JC, Yu L, Yang M, Dienstmann R, Guinney J, Beltrao P, Brazma A, Pardo M, Stegle O, Adams DJ, Wessels L, Saez-Rodriguez J, McDermott U, Choudhary JS. Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell Reports [Internet]* 2017 [cited 2017 Nov 9];20:2201–14. Available from: http://www.sciencedirect.com/science/article/pii/S2211124717311002

14. Moserle L, Indraccolo S, Ghisi M, Frasson C, Fortunato E, Canevari S, Miotti S, Tosello V, Zamarchi R, Corradin A, Minuzzo S, Rossi E, Basso G, Amadori A. The side population of ovarian cancer cells is a primary target of IFN-alpha antitumor effects. *Cancer Res [Internet]* 2008 [cited 2017 Nov 25];68:5658–68. Available from: http://cancerres.aacrjournals.org/content/68/14/5658

15. Wu D, Smyth GK. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucl Acids Res [Internet]* 2012;40:e133. Available from: https://academic.oup.com/nar/article/40/17/e133/2411151

16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS [Internet]* 2005 [cited 2017 Oct 22];102:15545–50. Available from: http://www.pnas.org/content/102/43/15545

17. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Systems [Internet]* 2015 [cited 2017 Oct 22];1:417–25. Available from: http://www.sciencedirect.com/science/article/pii/S2405471215002185

18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc [Internet]* 1995 [cited 2017 Sep 13];57:289–300. Available from: http://www.jstor.org/stable/2346101

19. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, Reyniès A de. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol [Internet]* 2016 [cited 2017 Nov 25];17:218. Available from: https://doi.org/10.1186/s13059-016-1070-5

20. Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M, Sevillano M, Palomo-Ponce S, Tauriello DVF, Byrom D, Cortina C, Morral C, Barceló C, Tosi S, Riera A, Attolini CS-O, Rossell D, Sancho E, Batlle E. Stromal gene expression defines poor-prognosis subtypes in

colorectal cancer. *Nat Genet [Internet]* 2015 [cited 2017 May 8];47:320–9. Available from: https://www.nature.com/ng/journal/v47/n4/full/ng.3225.html

21. Kampstra P. Beanplot: A boxplot alternative for visual comparison of distributions. *J Stat Softw [Internet]* 2008;28:1–9. Available from: http://www.jstatsoft.org/v28/c01/

22. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Meth [Internet]* 2015 [cited 2017 Feb 22];12:115–21. Available from: http://www.nature.com/nmeth/journal/v12/n2/abs/nmeth.3252.html

23. Guinney J, Dienstmann R, Wang X, Reyniès A de, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, Melo FDSE, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J, Bernards R, Friend SH, Laurent-Puig P, Medema JP, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S. The consensus molecular subtypes of colorectal cancer. *Nat Med [Internet]* 2015 [cited 2017 May 8];21:1350–6. Available from: https://www.nature.com/nm/journal/v21/n11/full/nm.3967.html

24. Gentleman R, Carey V, Huber W, Hahne F. Genefilter: Genefilter: Methods for filtering genes from high-throughput experiments [Internet]. 2017. Available from: https://bioconductor.org/packages/release/bioc/html/genefilter.html

25. Gordon M, Seifert R. Greg: Regression helper functions [Internet]. 2016. Available from: https://CRAN.R-project.org/package=Greg

26. Zhu H. kableExtra: Construct complex table with 'kable' and pipe syntax [Internet]. 2018. Available from: https://CRAN.R-project.org/package=kableExtra

27. Xie Y. Dynamic documents with r and knitr, second edition. 2 edition. Boca Raton: Chapman; Hall/CRC, 2015. 294p

28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl Acids Res [Internet]* 2015;43:e47. Available from: https://academic.oup.com/nar/article/43/7/e47/2414268

29. DeWitt P. Qwraps2: Quick wraps 2 [Internet]. 2018. Available from: https://CRAN.R-project.org/package=qwraps2

30. Neuwirth E. RColorBrewer: ColorBrewer palettes [Internet]. 2014. Available from: http://CRAN.R-project.org/package=RColorBrewer

31. Therneau TM. A package for survival analysis in s [Internet]. 2015. Available from: https://CRAN.R-project.org/package=survival

32. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet [Internet]* 2010 [cited 2014 Nov 21];11:733–9. Available from: http://www.nature.com/nrg/journal/v11/n10/abs/nrg2825.html