GENETICS AND GENOMICS

original reports abstract

# Deviating Alternative Splicing as a Molecular Subtype of Microsatellite Stable Colorectal Cancer

Jonas Meier Strømme, MSc, PhD[1,2]; Bjarne Johannessen, MSc, PhD[1]; and Rolf I. Skotheim, MSc, PhD[1,2]

**PURPOSE** Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide. Biomarkers to aid in prognostication and treatment decisions are in high demand, and to facilitate their development, a better understanding of the underlying biology of the highly heterogeneous disease is needed.

**METHODS** A genome-scale alternative splicing (AS) analysis using RNA-sequencing data from primary microsatellite stable (MSS) CRCs from 127 patients was performed. Splice variant–specific expression levels of individual cancer samples were compared with the total set of samples, and a metric for a tumor sample's global amount of deviating AS was developed. This metric varied considerably across the cohort and ranged from 6 to 282 deviating AS events per tumor sample. A threshold of 45 or more deviating events was set to distinguish cancers with high (n = 44) and low (n = 83) levels of deviating AS.

**RESULTS** Patients with high amounts of AS deviations had significantly shorter time to relapse compared with patients with fewer deviations ($P = .04$). Furthermore, differential gene expression analysis revealed nine known cancer-critical genes that are significantly upregulated in samples with high amounts of deviating AS. Validation of the results in an independent cohort of MSS CRCs showed the same tendency toward shorter progression-free survival among the high-deviation group. In both cohorts, enrichment for growth factors was identified among upregulated genes associated with this phenotype.

**CONCLUSION** There is a large variation in the amount of deviating AS among MSS CRCs, and we provide evidence that those with high amounts of deviations represent different cancer biology.

**JCO Clin Cancer Inform 7:e2200159. © 2023 by American Society of Clinical Oncology**



## INTRODUCTION

Colorectal cancer (CRC) remains a global health challenge because of exceptionally high incidence and mortality rates worldwide.[1-3] Primary CRCs are commonly divided into two major phenotypes: the microsatellite instability type, caused by deficiencies in the DNA mismatch repair system, and the microsatellite stable (MSS) type, which are characterized by larger chromosomal rearrangements. Despite large variability in the underlying mechanisms that cause development of colorectal tumors, several molecular biomarkers have been discovered that can aid in early detection and estimation of progression of this disease.[4] Particularly well studied is the expression and mutational landscape of oncogenes such as *KRAS*[5] and *BRAF*,[6] and tumor suppressor genes such as *TP53*.[7] Less elucidated is the role of alternative splicing (AS) in the development of CRC. We have previously studied how AS in *KRAS* can affect the prognosis of patients with CRC,[8] and the recent adaptation of high-throughput sequencing technologies has enabled investigations into the landscape of AS on a scale that spans the entire genome. It has previously been found that cancers originating from similar tissue, such as colon and rectum adenocarcinomas, form clusters on the basis of how the AS patterns differ between corresponding normal and tumor tissue, and also that AS affects cancer development to varying extents depending on the cancer type.[9] Moreover, a study presenting the genome-wide AS landscape in CRC found differences in splicing patterns between normal and tumor tissues, attesting that AS is a key characteristic behind tumor progression.[10]

We have previously published a study comparing the splicing patterns within a cohort of MSS primary CRCs.[11] Using exon-level microarrays, deviating exon usage was found to be correlated with the expression of splicing factors and associated with poor patient prognosis. The aim of this study was to evaluate genome-wide AS patterns in CRC by using RNA-sequencing technology on a similar set of MSS CRCs.

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on January 19, 2023 and published at ascopubs.org/journal/cci on February 23, 2023: DOI https://doi.org/10.1200/CCI.22.00159

## CONTEXT

**Key Objective**

Can identification and quantification of deviating alternative splicing in microsatellite stable colorectal cancer reflect meaningful tumor biology and provide relevant information on patient prognosis?

**Knowledge Generated**

The amount of deviating splicing varies between patients. Patients with high amounts of deviating splicing have a particularly poor prognosis, and their tumors express genes that are associated with more aggressive cancer phenotypes.

**Relevance**

Patients with microsatellite stable colorectal cancer can be stratified on the basis of their cancers' RNA transcript variation. This will inform on their prognosis, and therefore also influence on treatment decisions.

## METHODS

### Material

This study included RNA-sequencing data from 127 patients with CRC, 77 of whom were also included in the previous exon microarray study.[11] The RNA-sequencing data were prepared as previously described.[12] All tissue samples were of the MSS subtype. From all patients, we used RNA-sequencing data from primary cancerous tissue (Data Supplement). Average age at diagnosis was 70.5 years, and the majority had stage II or III CRC. Sixty-six patients (52%) were male and 61 (48%) were female. The protocol was approved by the Regional Committee for Medical and Health Research Ethics (REC numbers 1.2005.1629 and 2010/1805). All patients provided written informed consent, and the study was conducted in accordance with the Declaration of Helsinki. The research biobanks were registered according to national legislation.

### Analysis of RNA-Sequencing Data

***Quality control.*** All samples, with an average of 88.95 million reads, were assessed for adapter contamination using the FastQC software program[13] and aggregated by the MultiQC software program[14] (Data Supplement). All samples had $< 10\%$ of reads containing adapter sequences. Two CRC samples with outlier quality control values were excluded from the study.

***Read alignment and splice event identification.*** Reads were processed with the SpliceSeq software program[15] (version 2.1), which infers AS events on the basis of read alignment to a reference splice graph. The reference splice graph database was constructed using a proprietary software tool (acquired by correspondence with the author of SpliceSeq) and was based on the hg38 reference genome downloaded from the National Center for Biotechnology Information File Transfer Protocol server[16] and feature annotations from Ensembl (Ensembl Genes 85 database, data set GRCh38.p7[17]). Only features from protein coding genes were included in the analysis. SpliceSeq was configured to use Bowtie[18] version 1.0.0 with default parameters for read alignment.

Additional data from the SpliceSeq database that were not included in the standard data export, such as information about exons upstream and downstream of the splice sites, and expression estimates for the exons affected by splicing, were extracted from the database using MySQL queries. In 90.5% of the 51,750 genomic loci that were revealed with AS, percent spliced in (PSI) values, representing the exon inclusion ratios, were calculated for all samples. Splice sites were included in the AS analysis if it was possible to obtain a PSI value in at least 80 of the 127 samples.

Exon expression values, measured in reads per kilobase of transcript per million reads mapped, were averaged for multiexon AS events. For every AS event, median expression values from every tumor sample were calculated for the exon or exons affected. Median expression values were also calculated for the parent gene. Median PSI values were calculated for every AS event, and across all samples, a delta-PSI was calculated as the difference to the median value. The 25th and 75th percentiles of PSI values were calculated for every AS event, along with the interquartile range (IQR). Finally, exon and gene expression in individual samples were normalized for the median expression in all samples, and a ratio between normalized exon expression and normalized gene expression was computed for every splicing event in each sample.

***Identification of samples with high levels of deviating AS.*** Data filters were tailored to target two main classes of AS deviations. Deviating inclusion events are AS events where a subset of samples exhibits an inclination toward the inclusion of an exon that is excluded in the majority of samples. Deviating exclusion events are AS events where the exon is excluded in a minority of samples, while included in the majority. Outlier detection was performed on PSI values to indicate relative deviating inclusion or exclusion of a given event. A PSI value outlier in the higher or lower end of the spectrum implies deviating inclusion or exclusion, respectively. The following rules were applied, inspired by a previous study[9]:

$$\text{inclusion} = \text{PSI} > Q3 + 1.5\,IQR$$

$$\text{exclusion} = \text{PSI} < Q1 - 1.5\,IQR$$

 

where *Q*1 is the 25th percentile of PSI values, *Q*3 is the 75th percentile of PSI values, and *IQR* is the IQR for PSI values, derived from *Q*3-*Q*1. Individual deviating exon inclusion and exon exclusion events were identified using the criteria outlined in Table 1. For inclusion events, minimum requirements for reads per kilobase of transcript per million reads mapped and included counts ensure adequate evidence that the spliced exon and its gene of origin are expressed. Furthermore, PSI outliers with a minimum absolute difference from the median in the cohort are indicative of relative exon usage where outlier detection is not compromised by lack of variation. Finally, a fold change greater than two for the median-normalized exon/gene expression ratio ensures that the splicing pattern of an event in a sample is considerably deviant from the background level. Similar filters were used for the identification of exon-exclusion events, but the parameters were altered to target events with expression and PSI values below the baseline, rather than above.

Samples were categorized into two groups on the basis of the total number of AS deviations using the expectation-maximization algorithm from the R package mixtools[19] (version 1.2.0). The intersection of the two distributions was used as a numerical threshold for classifying samples with high levels of deviating AS.

***Additional statistical tests.*** Five-year relapse-free survival times for low- and high-deviation sample groups were estimated using the Kaplan-Meier method,[20] implemented in the survival[21] R package (version 3.1-8). Events were considered as relapse or death from any cause and censored in the case of no event within five years or when the patient was lost to follow-up. Twelve patients with TNM stage IV were excluded from the survival analysis, as metastatic CRCs can be interpreted as already having relapsed at time point zero. Univariable and multivariable Cox proportional hazards regression was performed using the survival[21] R package (version 3.1-8). The survminer[22] R package (version 0.4.6) was used to draw Kaplan-Meier curves. Fisher's exact test was run in R (version 3.6.0) with two-sided significance values on a two-by-two contingency table. Additional procedures are described in the Data Supplement.

## RESULTS

From the RNA-sequencing data of 127 CRCs, a total of approximately 6.5 million AS data points were quantified. Among these, 6,326 deviating AS events, that is, individual cancer samples with values for a particular splicing site that deviate from the value distribution in the other cancer samples, were identified. The division of these into particular types of AS showed that cassette exon and alternate promoter usage were the most common events (2,081 and 2,065, respectively; Fig 1A). Among the cassette exons, deviating inclusion events were more common than deviating exclusion events (1,763 inclusions and 318 exclusions). An example of a deviating inclusion cassette exon is

**TABLE 1.** Filtering Criteria Used to Identify Deviating Exon Inclusion and Exclusion Events

| Parameter | Inclusions Threshold | Exclusions Threshold | Description |
|---|---|---|---|
| Gene RPKM | > 5 | > 5 | Minimum expression of the exon's gene of origin in the particular sample |
| Median gene RPKM | > 5 | > 5 | Minimum median expression of the exon's gene of origin in all samples |
| Exon RPKM | > 1<br>> Median exon RPKM | | Minimum expression of the exon that is included in the particular sample (averaged for all exons in multiexon splicing events) |
| Median exon RPKM | | > 5 | Minimum median expression of the exon across all samples |
| PSI | > Q3 + (1.5 IQR) | < Q1 – (1.5 IQR) | PSI for the event must be > 75% for inclusion events and lower than 25% for exclusion events |
| Included counts | > 5 | | The minimum number of reads supporting the inclusion of the exon in the particular sample |
| Excluded counts | | > 5 | The minimum number of reads supporting the exclusion of the spliced exon in the particular sample |
| Delta-PSI (absolute) | > 0.2 | > 0.2 | The PSI of the exon in the particular sample must be more than 0.2 greater than the median PSI in all samples |
| Normalized exon/gene ratio ($log_2$) | > 1 | < –1 | The $log_2$-transformed median-normalized exon/gene expression ratio must be > 1 for inclusion events and lower than –1 for exclusion events |

NOTE. IQR is derived from quartile 3 to quartile 1 (Q3-Q1).

Abbreviations: IQR, interquartile range; PSI, percent spliced in; Q1, 25th percentile; Q3, 75th percentile; RPKM, reads per kilobase of transcript per million reads mapped.

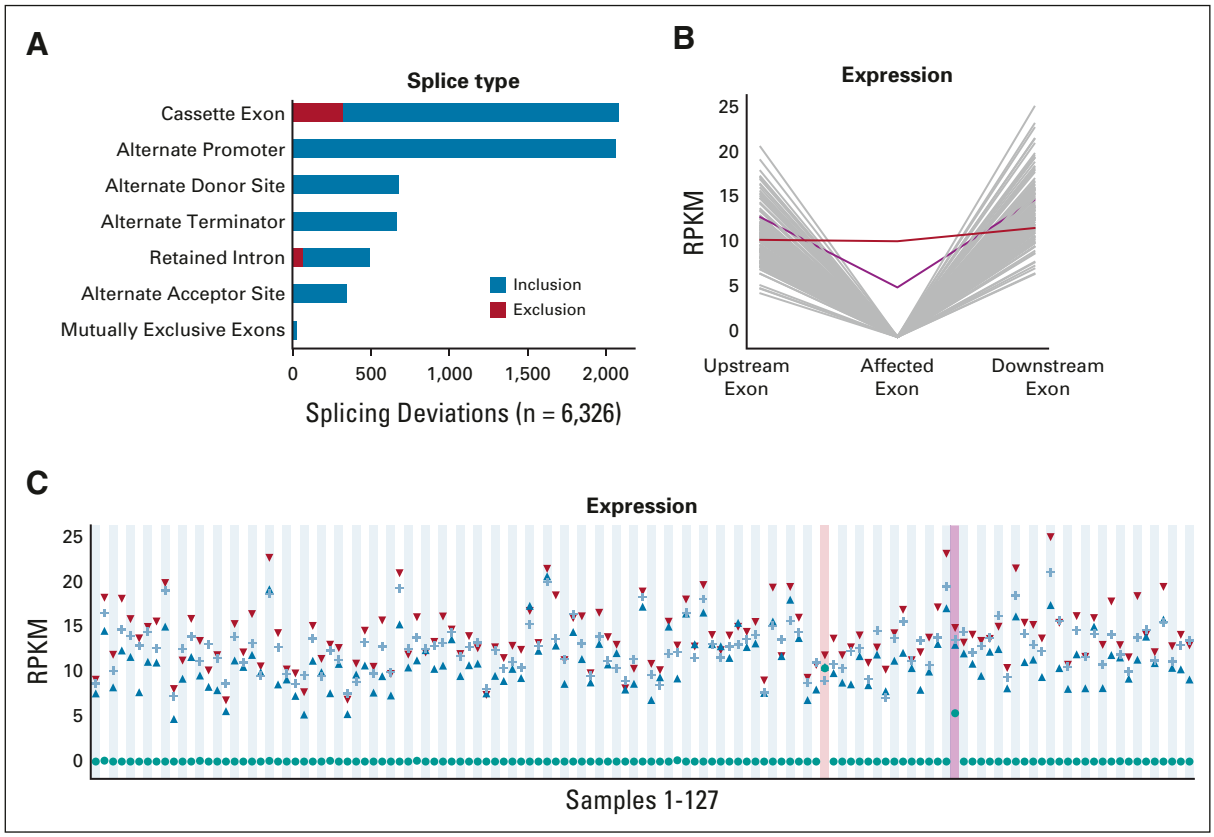


**FIG 1.** Identification of deviating AS events. (A) Distribution of splice types among 6,326 identified deviating AS events. Deviating inclusion of exon 2 in *PAPD4*: (B) Line plot showing expression in RPKM on the *y*-axis and exon 2 and its immediate upstream and downstream exons on the *x*-axis. The red line indicates sample 98, the purple line sample 113, and gray lines are the remaining samples in the cohort. (C) Scatter plot of the AS event, with the *y*-axis displaying expression in RPKM, and the *x*-axis representing one sample per tick, with sample 98 highlighted with a red background and sample 113 highlighted with a purple background. Green dots represent the spliced exon, red down-oriented triangles represent the downstream exon, blue up-oriented triangles represent the upstream exon, and black plus signs represent the gene-level expression, here *PAPD4*. AS, alternative splicing; RPKM, reads per kilobase of transcript per million reads mapped.

shown in Figures 1B and 1C, where exon 2 in the *PAPD4* gene is included in sample 98 (red) and sample 113 (purple), while the same exon is excluded in the majority of samples. Alternative promoter deviations were identified in 826 different genes. In 637 (77%) of those genes, differential usage was identified between two alternative promoters, but we observed deviating usage of up to four different promoters (for the gene *OXR1*). The majority (571; 27%) of alternative promoter deviations occurred uniquely in one individual sample, whereas the total number of alternative promoter deviations per sample ranged from 1 to 81.

### Identification of Samples With High Levels of Deviating Alternative Splicing

In total, 6,326 events were detected from the 127 samples, ranging from 6 to 282 deviations per sample. No correlation was found between the number of AS deviations identified and the sequencing depth (Pearson's r = –0.03; $P$ = .78; Data Supplement). The expectation-maximization algorithm was applied to identify a threshold of 44.5 AS deviations per sample, which separated 44 samples with high levels of deviation from 83 samples with low levels of deviation (Fig 2A). No particular skewedness was found for sex, age, or sidedness when comparing the high versus low AS deviation groups (Fig 2B). However, nine of 12 TNM stage IV cancers, versus 30% of the TNM stage I-III cancers belonged to the high-level group (Fisher's exact; $P$ = .004). Distributions of additional clinical parameters are shown in the Data Supplement.

### Gene Expression Associated With Levels of Deviating Alternative Splicing

Gene set enrichment analysis on gene expression ratios between the low- and high-deviation groups showed differences in gene sets related to metabolic processes and *MYC* targets, among several other biological processes (Fig 3). Analysis of individual genes revealed a total of 755 differentially expressed genes (Data Supplement). Of these, 749 genes had higher expression, and six genes had lower expression in the high-deviation group compared with the low-deviation group. Nine of the differentially expressed

 

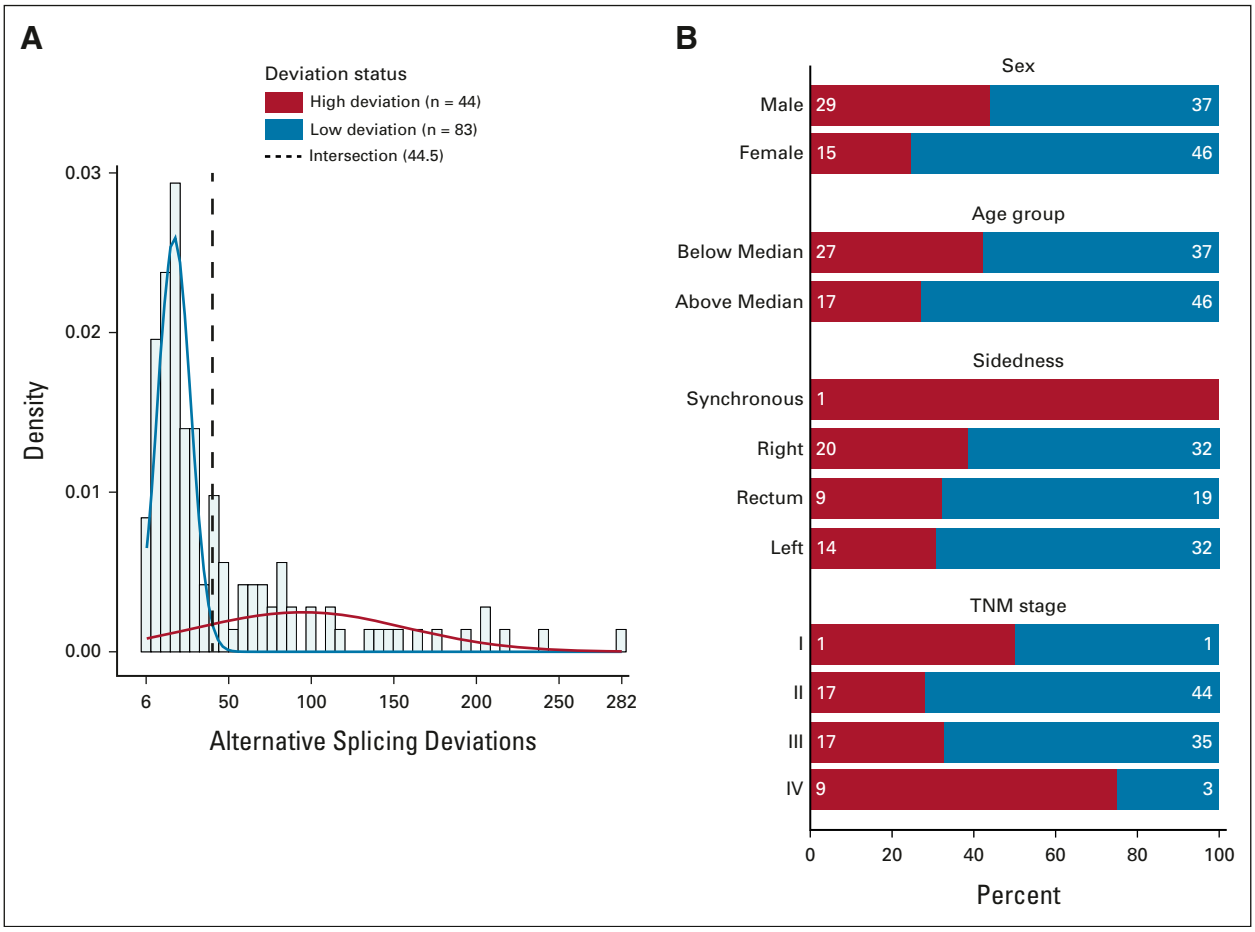


**FIG 2.** Patient subtyping on the basis of AS deviation levels. (A) Subtyping of 127 patients on the basis of the number of deviating AS events. The *x*-axis represents numbers of AS deviations, and the *y*-axis represents the density of patients. The expectation-maximization algorithm was used to identify two subdistributions of patients: those with high levels of deviations (red line) and those with low levels of deviations (blue line). The intersection of the two distributions (44.5) was used as a numerical threshold to separate the patients. (B) Distribution of clinical variables in the identified subgroups. The *x*-axis represents the total percentage distribution of each parameter, and absolute values for the number of patients in each subgroup are written in white. AS, alternative splicing.

protein-coding genes are part of the Cancer Gene Census,[23] and all of them had higher expression in the high-deviation group compared with the low-deviation group (Fig 4A).

We also compared differential AS between the high- and low-deviation groups and identified a total of 414 significantly differentially alternatively spliced sites. The distribution of splice types among the significant AS events is displayed in Figure 4B. Eighteen of the events were detected in known cancer genes (Data Supplement), including in *ATM* and *CHEK2*.

### Association With Relapse-Free Survival

There was a significant difference in time to relapse between the low-deviation and high-deviation groups ($P$ = .037). Univariable survival analyses found that patients with high levels of AS deviations (n = 35) had poorer prognosis (Cox proportional hazard ratio [HR], 0.55; 95% CI, 0.31 to 0.97; $P$ = .039) than patients with low levels (n = 80). Multivariable Cox regression analysis supported the effect of AS deviation levels on relapse or death when more covariates were considered (Table 2). The Kaplan-Meier plot for five-year relapse-free survival analysis is shown in Figure 4C.

### Validation in an Independent Data Set

RNA-sequencing data from a total of 426 MSS CRC tissue samples from The Cancer Genome Atlas (TCGA) were included in an external validation. Two hundred eighty-two samples were classified as having low levels of AS deviations and 144 samples as having high levels of AS deviations. Percentagewise, this is the closest possible distribution to that of the in-house cohort (66% *v* 34% low and high levels of deviation, respectively, in the TCGA cohort, compared with 65% *v* 35%, respectively, in the in-house cohort). Both cohorts produced similar distributions, with a substantial body of low-level samples and a long tail of high-deviation samples (Data Supplement).

Differential gene expression analysis (DGEA) revealed 538 significantly differentially expressed genes (510 upregulated and 28 downregulated in high-deviation samples compared

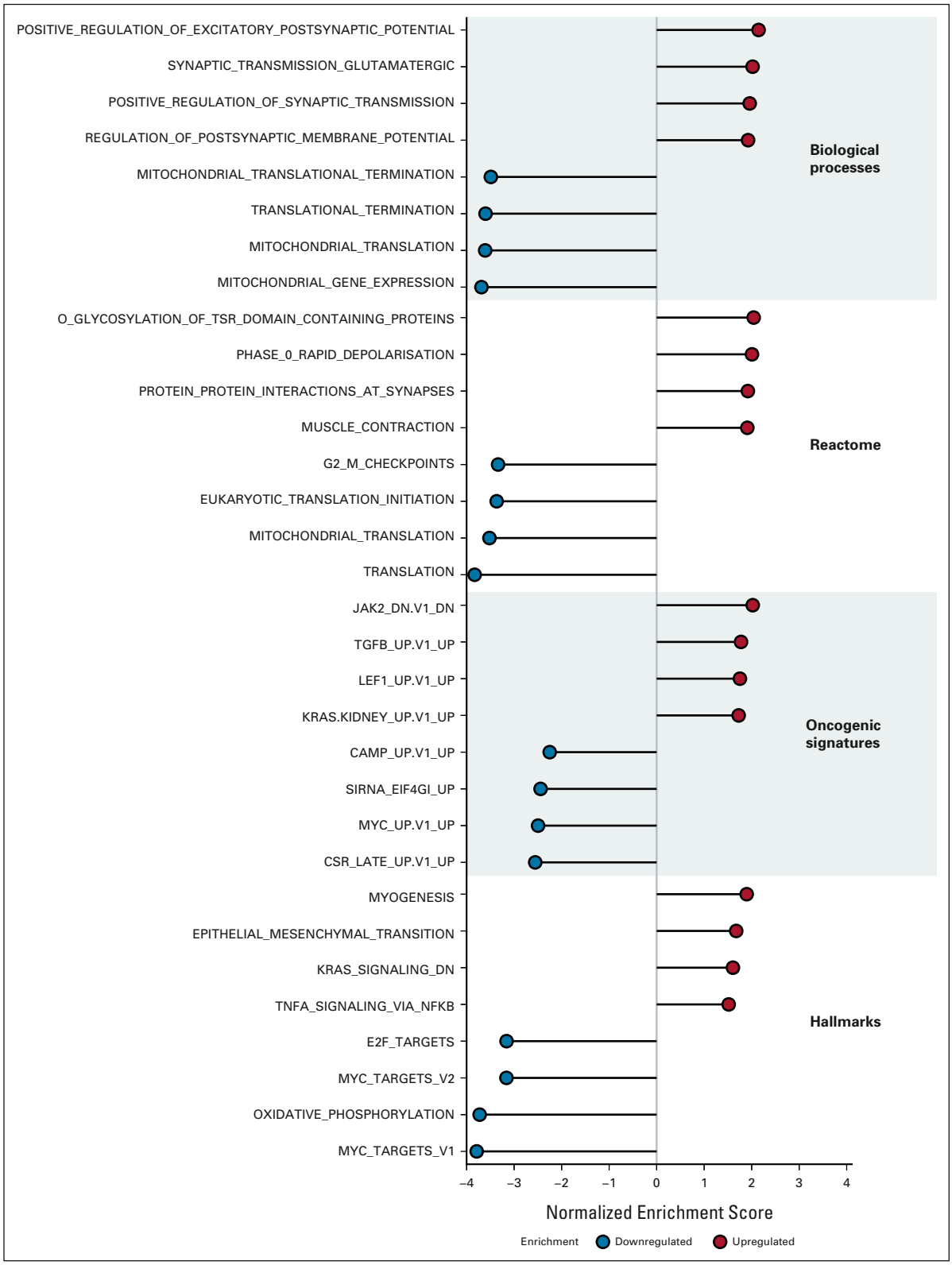


**FIG 3.** Gene set enrichment analysis of differentially expressed genes between the low- and high-deviation groups. Included here are the top four gene sets with higher expression in the high-deviation group and the top four gene sets with lower expression in the high-deviation group from each of the following gene set collections from MSigDB release 7.0: C5 biological processes; C2 reactome; C6 oncogenic processes; and hallmarks. The low-deviation group serves as reference for the direction.

 

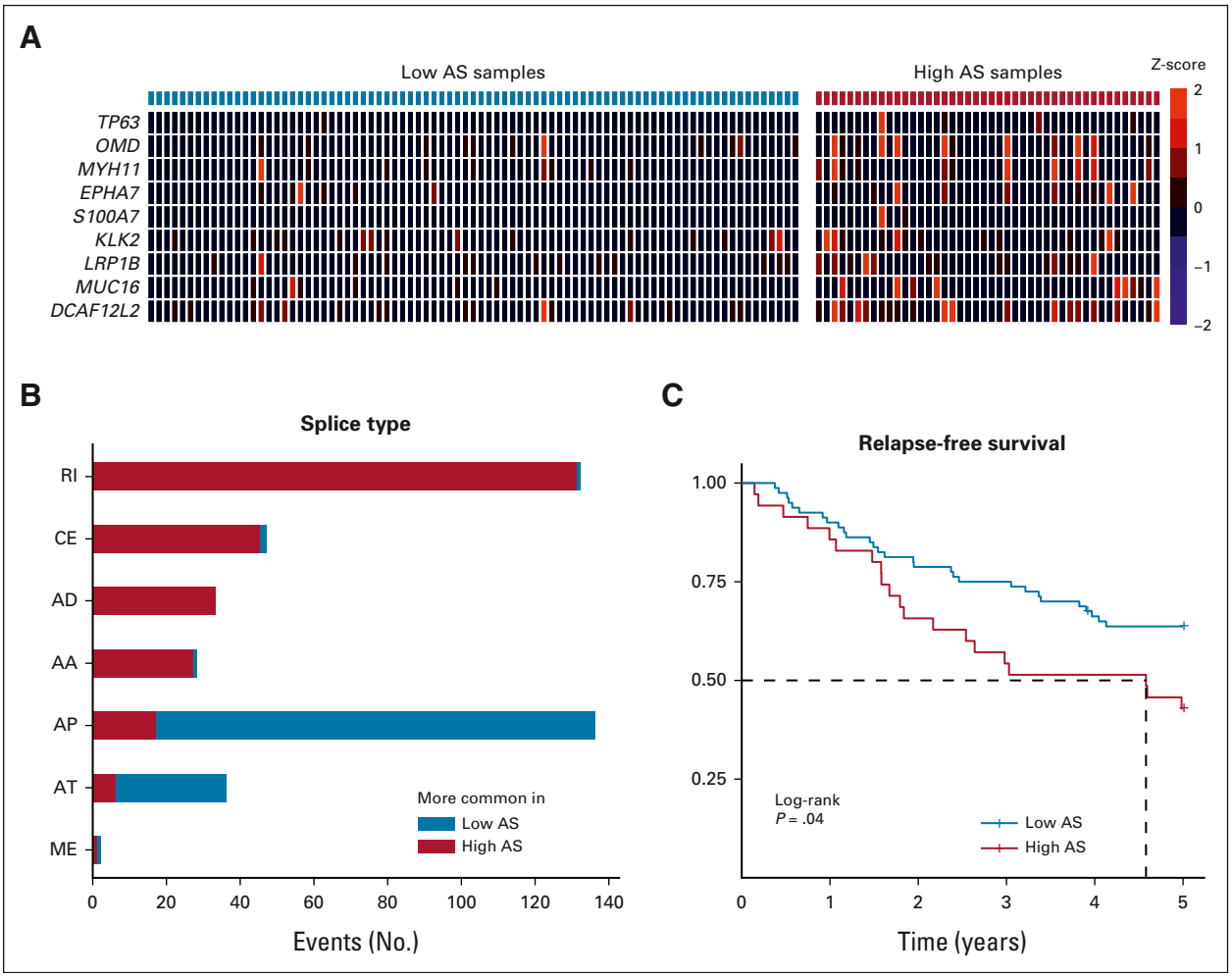


**FIG 4.** Differences between samples with high and low levels of AS deviations. (A) Heatmap showing expression (as Z-scores) of nine genes from the cancer gene census with higher expression in samples with high levels of deviation. Samples on the *x*-axis are colored by deviation level status (low deviation in blue, high deviation in red). Z-scores range from –0.41 to 11.18, and extremes have been aggregated for visualization purposes. (B) Distribution of splice types among 414 differentially alternatively spliced sites between samples with high and low levels of deviating AS. (C) Relapse-free survival analysis of 115 patients (12 stage IV patients were not included) with high and low levels of deviating AS. Patients with high levels (n = 35) have significantly poorer survival than patients with low levels (n = 80). AA, alternate acceptor site; AD, alternate donor site; AP, alternate promoter; AS, alternative splicing; AT, alternate terminator; CE, cassette exon; ME, mutually exclusive exon; RI, retained intron.

with low-deviation samples; Data Supplement). Sixty (46 protein coding and 14 lncRNAs) of the genes were also significantly differentially expressed in the in-house cohort (Data Supplement).

By applying functional annotations tools from the Database for Annotation, Visualization, and Integrated Discovery to the 59 (one gene was missing in Database for Annotation, Visualization, and Integrated Discovery), six annotation clusters were defined. Notably, the combined set of genes was significantly enriched for growth factor genes (Data Supplement), including *VGF*, *MIA*, and *INHA*.

Analyses including clinical data from the TCGA cohort did not yield significant association between AS deviation and progression-free survival (Cox proportional HR, univariable HR, 0.75; 95% CI, 0.45 to 1.26; $P = .28$; multivariable HR, 0.80; 95% CI, 0.48 to 1.33; $P = .384$).

## DISCUSSION

We have provided evidence for a relevant subgrouping of MSS CRCs with high amounts of AS, which deviates from the majority of CRC samples. From RNA-sequencing data, we have established a method for quantifying deviating AS in a tumor setting on a per-sample basis. We have previously analyzed a part of this series of CRC[11] and other types of cancer[24] by a similar approach using exon microarray data. These analyses were performed with the Affymetrix GeneChip Human Exon 1.0 ST Array, which provides exon-level expression data, but with no measurements of exon-exon junctions. The present use of RNA-sequencing data

**TABLE 2.** Multivariable Analysis of HRs for 5-Year Relapse or Death

| Variable | HR | 95% CI | *P* |
|---|---|---|---|
| Deviation status | | | |
| High (n = 35) | 1.00 | | |
| Low (n = 80) | 0.54 | 0.30 to 0.99 | .048 |
| Age group | | | |
| Above median (n = 62) | 1.00 | | |
| Below median (n = 53) | 0.82 | 0.46 to 1.49 | .522 |
| Tumor location | | | |
| Left + rectum (n = 65) | 1.00 | | |
| Right (n = 50) | 0.73 | 0.40 to 1.32 | .297 |
| TNM stage | | | |
| I (n = 2) | 1.00 | | |
| II (n = 61) | 0.63 | 0.08 to 4.90 | .656 |
| III (n = 52) | 0.95 | 0.12 to 7.23 | .957 |
| Sex | | | |
| Female (n = 59) | 1.00 | | |
| Male (n = 56) | 1.38 | 0.76 to 2.51 | .285 |

NOTE. Deviation status is significantly predictive of patient prognosis, and patients with low levels of AS deviations have approximately half the risk of relapse or death, compared with those with high levels of AS deviation.

Abbreviations: AS, alternative splicing; HR, hazard ratio.

enabled a genome-scale study of the particular splice junctions and the ability to distinguish between different types of AS. We found exon inclusion and alternate promoter usage to be the most common deviations.

Interestingly, we found that the number of AS deviations varies greatly within the cohort, and patients with TNM stage IV were found to be associated with a higher number of deviating AS than patients with stage I-III. However, even within the TNM stage I-III group, survival analysis between samples with high and low numbers of deviations showed that it had predictive value for patient prognosis.

DGEA between the groups showed significant difference in 755 genes. Among them were known cancer genes (*DCAF12L2*, *EPHA7*, *KLK2*, *LRP1B*, *MUC16*, *OMD*, *S100A7*, and *TP63*), all of which had stronger expression in samples with high number of deviating AS, compared with the low-deviation subgroup.

Differential AS analysis showed significant differences in splicing of 414 splice sites, 18 of which correspond to known cancer driver genes.[23] High-deviation samples, more often than low-deviation samples, had a retained intron between exons 1 and 2 of *ATM*, and skipped exon 2 in *CHEK2* (Data Supplement). *ATM* and *CHEK2* both encode cell-cycle checkpoint kinases involved in DNA damage response. The proteins encoded by both genes interact with the proteins encoded by the tumor suppressor genes *TP53* and *BRCA1*, as well as one another. Mutations in both genes are linked to cancer progression in various tissue types, such as colon, rectum, breast, thyroid, lymphoid, and skin.[25-27]

Gene set enrichment analysis identified systematic changes on larger networks of genes. High-deviation samples showed low expression of genes associated with mitochondrial gene expression and translation, and oxidative phosphorylation, suggesting that there are differences in energy metabolism between the groups, and possibly that high-deviation cancers increasingly exhibit the Warburg effect[28] known to facilitate oncogenesis, tumor progression, and motility.[29] *MYC* targets were also less expressed in high-deviation samples. Deregulation of *MYC* is known to affect energy metabolism in cancer.[30] Furthermore, samples with high levels of AS deviation also showed higher expression of genes associated with epithelial-mesenchymal transition, which increases cell motility and is associated with the initiation of metastasis.[31] Combined, these characteristics suggest that high levels of AS deviation are linked to tumor aggressiveness, which could explain the increased prevalence of AS deviation in patients with stage IV CRC and the observed difference in relapse-free survival between the subgroups.

By comparing our results with an external cohort of MSS CRC tissue samples, we found the same tendency of association between high amounts of deviating AS and disease progression, but the results did not reach statistical significance. One possible explanation can be that the external cohort had only a median follow-up time of 1.7 years, compared with 5 years in the in-house cohort. We also used the external data in DGEA and identified genes that were differentially regulated in association with high amount of deviating AS in both cohorts. Functional annotation of these genes showed upregulation of genes associated with growth factors, notably the genes *VGF*, *MIA*, and *INHA*. Associations have been found between upregulation of *VGF* and disease aggressiveness in lung adenocarcinomas,[32] *MIA* is associated with aggressive malignant melanoma,[33] and *INHA* is known to be upregulated in some ovarian cancers.[34]

In conclusion, we have demonstrated that the transcriptome-wide amount of deviating AS in MSS CRC is highly variable, and that the cancers with the highest number of deviations are associated with shorter time to disease relapse. Furthermore, these cancers show lower expression of genes related to energy metabolism and higher expression of growth factor genes, but further studies are needed to assess the association between levels of deviating AS and aggressive cancer phenotypes.

 

## AFFILIATIONS

[1]Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital, Radiumhospitalet, Oslo, Norway
[2]Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway

## CORRESPONDING AUTHOR

Rolf I. Skotheim, MSc, PhD, P.O. Box 4953 Nydalen, NO-0424 Oslo, Norway; Twitter: @rolfsk; e-mail: rolf.i.skotheim@rr-research.no.



## SUPPORT

Supported by the Southern and Eastern Norway Regional Health Authority (project number 2016003), the Norwegian Cancer Society (208197), and the Research Council of Norway (262529), all with R.I.S. as the grant recipient. The study was granted secure storage of computer files and high-performance computation resources from Sigma2 and University of Oslo's Services for Sensitive Data (NS9013S).


## DATA SHARING STATEMENT

Two independent genomics data sets were included in the analyses, both of which are referred to in the original publications.[12,35] No new molecular data were generated for the current manuscript. Any other data and results that support the results of this study are available from the corresponding author, R.I.S., upon reasonable request.

## AUTHOR CONTRIBUTIONS

**Conception and design:** All authors
**Financial support:** Rolf I. Skotheim
**Administrative support:** Rolf I. Skotheim
**Collection and assembly of data:** Jonas Meier Strømme, Bjarne Johannessen
**Data analysis and interpretation:** All authors
**Manuscript writing:** All authors
**Final approval of manuscript:** All authors
**Accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

No potential conflicts of interest were reported.

**Rolf I. Skotheim**

This author is a member of the *JCO Clinical Cancer Informatics* Editorial Board. Journal policy recused the author from having any role in the peer review of this manuscript


## ACKNOWLEDGMENT

The authors are grateful to Ragnhild A. Lothe and Anita Sveen for contributions of data and critical discussions of this work. The authors are grateful to the patients with colorectal cancer for contributing to this study.


## REFERENCES


1. Global Burden of Disease Cancer Collaboration: Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: A systematic analysis for the global burden of disease study. JAMA Oncol 5:1749-1768, 2019
2. Vega P, Valentín F, Cubiella J: Colorectal cancer diagnosis: Pitfalls and opportunities. World J Gastrointest Oncol 7:422-433, 2015
3. Levin B, Lieberman DA, McFarland B, et al: Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. CA Cancer J Clin 58:130-160, 2008
4. Tanaka T, Tanaka M, Tanaka T, et al: Biomarkers for colorectal cancer. Int J Mol Sci 11:3209-3225, 2010
5. Fearon ER, Vogelstein B: A genetic model for colorectal tumorigenesis. Cell 61:759-767, 1990
6. Taieb J, Le Malicot K, Shi Q, et al: Prognostic value of BRAF and KRAS mutations in MSI and MSS stage III colon cancer. J Natl Cancer Inst 109:djw272, 2017
7. Iacopetta B: TP53 mutation in colorectal cancer. Hum Mutat 21:271-276, 2003
8. Eilertsen IA, Sveen A, Strømme JM, et al: Alternative splicing expands the prognostic impact of KRAS in microsatellite stable primary colorectal cancer. Int J Cancer 144:841-847, 2018
9. Li Y, Sahni N, Pancsa R, et al: Revealing the determinants of widespread alternative splicing perturbation in cancer. Cell Rep 21:798-812, 2017
10. Xiong Y, Deng Y, Wang K, et al: Profiles of alternative splicing in colorectal cancer and their clinical significance: A study based on large-scale sequencing data. EBioMedicine 36:183-195, 2018
11. Sveen A, Agesen TH, Nesbakken A, et al: Transcriptome instability in colorectal cancer identified by exon microarray analyses: Associations with splicing factor expression levels and patient survival. Genome Med 3:32, 2011
12. Sveen A, Johannessen B, Eilertsen IA, et al: The expressed mutational landscape of microsatellite stable colorectal cancers. Genome Med 13:1-18, 2021
13. Andrews S: FastQC: A Quality Control Tool for High Throughput Sequence Data 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
14. Ewels P, Magnusson M, Lundin S, et al: MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047-3048, 2016
15. Ryan MC, Cleland J, Kim R, et al: SpliceSeq: A resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics 28:2385-2387, 2012
16. National Center for Biotechnology Information: NCBI FTP Site. Viewed 2017-01-04. n.d. ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/
17. Ensembl: Ensembl FTP Site. Viewed 2017-01-04. 2016. ftp://ftp.ensembl.org/pub/release-85/gtf/homo_sapiens
18. Langmead B, Trapnell C, Pop M, et al: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25, 2009
19. Benaglia T, Chauveau D, Hunter D, et al: mixtools: An R package for analyzing finite mixture models. J Stat Softw 32:1-29, 2009
20. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457-481, 1958
21. Therneau TM: A Package for Survival Analysis in R, 2019. https://CRAN.R-project.org/package=survival

22. Kassambara A, Kosinski M, Biecek P: Survminer 2019. https://cran.r-project.org/package=survminer
23. Futreal PA, Coin L, Marshall M, et al: A census of human cancer genes. Nat Rev Cancer 4:177-183, 2004
24. Sveen A, Johannessen B, Teixeira MR, et al: Transcriptome instability as a molecular pan-cancer characteristic of carcinomas. BMC Genomics 15:672, 2014
25. Choi M, Kipps T, Kurzrock R: ATM mutations in cancer: Therapeutic implications. Mol Cancer Ther 15:1781-1791, 2016
26. Pearlman R, Frankel WL, Swanson B, et al: Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. JAMA Oncol 3:464-471, 2017
27. Sabapathy K, Lane DP: Therapeutic targeting of p53: All mutants are equal, but some mutants are more equal than others. Nat Rev Clin Oncol 15:13-30, 2018
28. House SW, Warburg O, Burk D, et al: On respiratory impairment in cancer cells. Science 124:267-272, 1956
29. Kim JW, Dang CV: Cancer's molecular sweet tooth and the Warburg effect. Cancer Res 66:8927-8930, 2006
30. Stine ZE, Walton ZE, Altman BJ, et al: MYC, metabolism, and cancer. Cancer Discov 5:1024-1039, 2015
31. Thiery JP: Epithelial–mesenchymal transitions in tumour progression. Nat Rev Cancer 2:442-454, 2002
32. Hwang W, Chiu Y-F, Kuo M-H, et al: Expression of neuroendocrine factor VGF in lung cancer cells confers resistance to EGFR kinase inhibitors and triggers epithelial-to-mesenchymal transition. Cancer Res 77:3013-3026, 2017
33. Riechers A, Bosserhoff AK: Melanoma inhibitory activity in melanoma diagnostics and therapy—A small protein is looming large. Exp Dermatol 23:12-14, 2014
34. Cooke I, O'Brien M, Charnock F, et al: Inhibin as a marker for ovarian cancer. Br J Cancer 71:1046-1050, 1995
35. Weinstein JN, Collisson EA, Mills GB, et al: The cancer genome atlas pan-cancer analysis project. Nat Genet 45:1113-1120, 2013