Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Consensus molecular subtyping in colon cancer: from the design of in-house assay to prognostic nomogram

Mariapia Caputo
IRCCS Istituto Tumori "Giovanni Paolo II"
Leonarda Maurmo
LHA BAT
Debora Traversa
IRCCS Istituto Tumori "Giovanni Paolo II"
Tommaso Maria Marvulli
IRCCS Istituto Tumori "Giovanni Paolo II"
Eliseo Mattioli
IRCCS Istituto Tumori "Giovanni Paolo II"
Francesco Alfredo Zito
IRCCS Istituto Tumori "Giovanni Paolo II"
Raffaele De Luca
IRCCS Istituto Tumori "Giovanni Paolo II"
Graziana Barile
S.S.D. Chirurgia Generale ad Indirizzo Senologico, IRCCS Istituto Tumori “Giovanni Paolo II”
Angela Ricco
IRCCS Istituto Tumori "Giovanni Paolo II"
Fulvia Colonna
IRCCS Istituto Tumori "Giovanni Paolo II"
Brunella Pilato
IRCCS Istituto Tumori "Giovanni Paolo II"
Giuseppina Matera
IRCCS Istituto Tumori "Giovanni Paolo II"
Stefania Tommasi
IRCCS Istituto Tumori "Giovanni Paolo II"
Simona De Summa
s.desumma@oncologico.bari.it

IRCCS Istituto Tumori "Giovanni Paolo II"

Research Article

Keywords: CMS, colon cancer, in house panel, transcriptomics, nomogram

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9178611/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background.

Consensus Molecular Subtypes (CMS) provide a biologically robust classification of colorectal cancer (CRC) with prognostic relevance; however, their clinical implementation remains limited by reliance on genome-wide transcriptomics and platform-dependent classifiers. We aimed to develop and validate a compact, FFPE-compatible in-house CMS assay and to integrate CMS information into a clinically applicable prognostic nomogram.

Methods.

An in silico training cohort was generated from TCGA-COAD and TCGA-READ datasets, including transcriptomic profiles, somatic mutations, and microsatellite instability (MSI) status. Feature selection was performed using a minimum Redundancy Maximum Relevance (mRMR) wrapper approach combined with Random Forest and Support Vector Machine classifiers to identify a minimal CMS-informative feature set. Based on selected features, a custom amplicon-based DNA/RNA NGS assay was designed. The assay was validated in an independent cohort of 100 CRC patients with bulk RNAseq–based CMS classification. Classification performance was assessed against CMSclassifier and CMScaller. A multivariable Cox proportional hazards model incorporating CMS, clinicopathological variables, MSI, and driver mutations was used to develop a prognostic nomogram for 3- and 5-year overall survival.

Results.

The optimal model achieved a global accuracy of 0.76 using 70 features, comprising 7 coding genes, 32 lncRNAs, 30 genomic alterations, and MSI status. The in-house CMS assay demonstrated high balanced accuracy for CMS1 and CMS2, with lower performance for CMS3 and CMS4. Integration of CMSassay classification into the prognostic model showed that CMS classification and tumor stage contributed the largest effects on survival risk. The resulting nomogram provided individualized 3- and 5-year overall survival estimates with clear risk stratification.

Conclusions.

We developed a clinically feasible, FFPE-compatible CMS assay that preserves the prognostic relevance of transcriptome-based CMS classification while reducing technical and computational complexity. When integrated into a multivariable nomogram, CMS provides substantial prognostic value beyond standard molecular and clinicopathological factors. This approach facilitates the translation of CMS into routine clinical workflows and supports its use as a quantitative prognostic tool in colon cancer.

Background

Colorectal cancer (CRC) represents the third most commonly diagnosed cancer and the second leading cause of cancer-related deaths [1, 2], according to the global statistics.

CRC is a highly heterogeneous disease, characterized by complex interactions between genetic, epigenetic and environmental factors. This heterogeneity highlights significant challenges for diagnosis, prognosis and treatment [3].

One of the innovative solutions to address these unmet clinical needs, consists in the development of the molecular classification systems to stratify CRC into biologically different subtypes with clinical relevance. Among these, the Consensus Molecular Subtypes (CMS) classification has emerged as the most robust and widely accepted framework.

The CMS classification, developed by Guinney et al. [4] in 2015, categorizes CRC into four distinct subtypes based on transcriptomic data: CMS1 (Microsatellite Instability Immune, MSI), CMS2 (Canonical), CMS3 (Metabolic), and CMS4 (Mesenchymal) [4].

Each CMS subtype is associated with unique molecular, histological, and clinical characteristics. CMS1 is characterized by MSI [5], strong immune infiltration, and a favorable prognosis but poor response to conventional chemotherapy. CMS2, the most prevalent subtype, exhibits chromosomal instability, epithelial differentiation, and activation of the WNT and MYC signaling pathways [6], displaying a better response to chemotherapy and improved survival. CMS3 shows dysregulated metabolic pathways, particularly in glucose metabolism, and tends to occur in tumors with KRAS mutations [7]. CMS4, associated with mesenchymal activation, stromal invasion, and angiogenesis, carries the worst prognosis due to its aggressive nature and resistance to standard treatments [6, 8, 9].

The CMS framework has provided valuable insights into the biological bases of CRC and has facilitated the development of precision medicine strategies. By identifying molecular subtypes, clinicians and researchers can better understand tumor behavior, predict treatment responses, and design targeted therapies.

Several challenges still remain in the clinical implementation of CMS classification, including the integration of transcriptomic profiling into routine practice and the need for subtype-specific therapeutic options.

In these years, various algorithms have been developed to improve the CMS classification in clinical and research settings, including the *CMScaller* [10], *CMSclassifier* [11], frameworks. Each of these algorithms offers unique strengths and limitations.

The *CMScaller* is a computational tool that uses gene expression profiles to assign CRC tumors to one of the four CMS categories [10]. It relies on a predefined set of gene expression signatures and applies statistical methods to ensure accurate subtype assignment. The *CMScaller* is known for its high

reproducibility and robustness, making it a preferred choice in transcriptomic studies. However, its reliance on gene expression data necessitates high-quality RNA samples, which can be a limitation in certain clinical scenarios.

The *CMSclassifier*, similar to the *CMScaller*, employs a supervised machine learning approach to classify CRC tumors [11]. It incorporates additional biological parameters, such as mutation status and pathway activation scores, to enhance the accuracy of subtype assignments. This algorithm is particularly advantageous in settings where multi-omic data are available, as it provides a more integrative understanding of tumor biology. However, the *CMSclassifier* requires more computational resources and data integration expertise, potentially limiting its accessibility for routine clinical use.

While the *CMScaller* and *CMSclassifier* share a common goal of implementing the CMS framework, their methodological differences highlight the importance of algorithm selection based on the research or clinical context. Together, these tools could provide a multifaceted approach to understanding CRC heterogeneity and advancing precision oncology.

Materials and Methods

TCGA-COAD and TCGA-READ Data retrieval

RNA expression data and corresponding clinical information from patients with colorectal cancer were downloaded from the Genomic Data Commons (GDC) Data Portal [12]. The datasets were obtained from The Cancer Genome Atlas (TCGA) project, specifically from the TCGA-COAD (colon adenocarcinoma) and TCGA-READ (rectum adenocarcinoma) cohorts. *TCGAbiolinks* R package [13] has been used to retrieve clinical data, HT-Seq count transcriptome data and miRNA Expression Quantification data. Ensemble ID has been converted into HUGO gene symbols by *BiomaRt* R package [14]. LncRNA expression data has been extrapolated by LNCpedia v5.2 conversion table [15]. *TCGAbiolinks* R package has been also used to retrieve alterations data, thus obtaining MAF-files. MSI status has been obtained by *GDCquery* (data.category='Other' and clinical.info='msi') function. CMSclassifier was used for the classification of samples in CMSs. It was implemented by Guinney et al.[4],and it is based on the 273 genes included in the original random forest (RF) CMS classifier.

Pre-processing: sample filtering and data adjusting

For the analysis of count data from RNA-seq for the detection of differentially expressed genes, has been used *DESeq2* R package [16, 17]. The count data, previously downloaded, are presented as a table (the rows correspond to the genes and the columns to the patients) which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Firstly, a Summarized Experiment (SE) has been obtained by *GDCprepare* (summarizedExperiment=TRUE) function. This SE was used to create a *DESeqDataSet* class. While it is not necessary to pre-filter low count genes before running the *DESeq2* functions, there are two reasons which make pre-filtering useful: by removing rows in which

there are very few reads, the memory size of *DESeqDataSet* object is reduced, and the speed of the transformation and testing functions within *DESeq2* is increased. In our study, has been performed a minimal pre-filtering to keep only rows that have at least 10 reads total. After that, it has been used the *varianceStabilizingTranformation* function which removes the dependence of the variance on the mean and then transforms the count data, obtaining a matrix of values which have constant variance along the range of mean values[16].

To reduce the number of miRNAs and lncRNAs, before data normalization, further filtering was applied: all rows with reads count < 10 in more than 10% of patients were eliminated.

For data mutations we selected the clinically relevant mutations by *OncoKB*, a precision oncology knowledge base [18]. After that, have been selected only *'Nonsense'* and *'Missense'* mutations grouped by genes and has been created a matrix which reports, for each sample, the presence or absence of the mutation for each gene.

Feature Selection and classification

Wrapper features selection method

To reduce the computational burden of classification and obtain a compact feature set that could be more easily applied in clinical practice, a feature selection step was performed. We adopted a wrapper-based feature selection approach, which evaluates subsets of features using a machine learning algorithm combined with a search strategy that explores the space of possible feature combinations. Each subset is assessed according to the predictive performance achieved by the selected algorithm. In general, wrapper methods follow these steps: (1) a subset of features is selected from the full set using a search strategy; (2) a machine learning model is trained on the selected subset; (3) model performance is evaluated using a predefined metric; and (4) the procedure is repeated iteratively across different feature subsets and corresponding trained models. In our study, we used minimum Redundancy Maximum Relevance (mRMR) [19] for feature selection together with two classifiers, Support-Vector-Machine (SVM) [20] and Random Forest (RF) [21], both of which are well suited for multi-label classification. Our specific aim was to assign patients to the four CMS classes.

mRMR Feature Selection Method

Feature selection was performed using the *mRMR.classic* fuction, implemented in the *mRMRe* R package [19], which enables efficient identification of relevant and non-redundant features. Let *y* denote the output variable and *X*= {x1,.., xn} the set of *n* input features. The method ranks features in *X* by maximizing their mutual information (MI) with *y* (maximum relevance) while minimizing the average MI with features already selected (minimum redundancy). Mutual information between two datasets is estimated on the basis of the correlation between variables Pearson's or Spearman's correlation

coefficients are used for continuous variables, Cramer's V for discrete variables, and Somers' Dxy index for the association between continuous variables and survival data.

Support Vector Machine and Random Forest

SVMs are supervised learning methods commonly used for classification and regression tasks [20]. By means of kernel functions, they allow efficient classification of non-linear data. In this framework, input samples are represented as points in an n-dimensional space, where each dimension corresponds to one feature. The algorithm then iteratively identifies a function defining a hyperplane that best separates the regions occupied by the different output classes. In our study, SVM classification was performed using the *svm()* function from the *e1071* R package. The RF algorithm, instead, combines multiple randomized decision trees and aggregates their predictions by majority voting [21]. This approach has shown strong performance, particularly in settings where the number of variables greatly exceeds the number of observations. Its general workflow includes drawing random samples from the original dataset, building a decision tree for each sample, generating predictions from each tree, and selecting the class receiving the highest number of votes as the final output. In our study, RF classification was carried out using the *randomForest* function from the *randomForest* R package, based on Breiman's original implementation [21].

In-house CMSassay: DNA/RNA custom panel design and bioinformatic analyses

The in-house CMSassay included two panels, both of them amplicon-based and designed through Ion Ampliseq Designer web tool (ThermoFisher). RNA panel required a great effort because it was built through multiple rounds of tiling/pooling in order to increase coverage rate. No SNPs were allowed under primers for the bulk of the design. Quality Control (QC) to avoid cross-binding of primers was performed for each panel. Two lncRNAs were excluded from the final design due technical issues in the identification of primers able to correctly address them. Alignment and variant calling was performed through Torrent Server plugins, Torrent Variant Caller in somatic configuration. From Ion Torrent Server we retrieved VCF files for each sample, which were annotated by OncoKB™ annotator [22].

RNA panel was analyzed with Ampliseq RNA plugin to obtain read count matrices, that were exported and normalized through *DESeq2* R package.

DNA/RNA extraction and sequencing

Nucleic acid extraction was performed using the MagMAX™ FFPE DNA/RNA Ultra Kit (ThermoFisher Scientific) on the KingFisher™ Duo Prime System (ThermoFisher Scientific) following the manufacturer's instructions. Briefly, 6 µm-thick formalin-fixed paraffin-embedded (FFPE) tissue sections were deparaffinized and subjected to proteinase K digestion at 56°C overnight to ensure effective nucleic acid

release. Nucleic acids were captured using paramagnetic beads and washed with buffers provided in the kit to remove contaminants. The extracted DNA and RNA were eluted in 50 µL of the elution buffer provided and quantified using Qubit™ Fluorometric Quantification (ThermoFisher Scientific) to assess concentration. Extracted nucleic acids were stored at -20°C and – 80°C, respectively, until further processing.

Library preparation was performed using the Ion AmpliSeq™ Library Kit Plus (ThermoFisher Scientific) with a custom gene panel targeting regions of interest. Briefly, the extracted material was used for multiplex PCR amplification of target regions, followed by partial digestion of primer sequences using FuPa reagent. The amplicons were then ligated with Ion-compatible barcoded adapters using DNA Ligase and purified with magnetic bead-based selection. The final libraries were quantified using the Ion Library TaqMan® Quantitation Kit and diluted to a final concentration of 100 pM for template preparation.

The prepared library pool was loaded onto the Ion Chef™ System (ThermoFisher Scientific) for automated loading onto the Ion 530™ Chip. Sequencing was performed on the Ion S5™ System (ThermoFisher Scientific) following standard protocols for sequencing run setup and data acquisition.

Microsatellite Instability (MSI) evaluation

MSI status was assessed using the Idylla™ MSI Test (Biocartis), an automated, real-time PCR–based assay designed for the qualitative detection of MSI in solid tumors. FFPE tumor tissue sections were used as input material according to the manufacturer's instructions.

Bulk RNAseq of the validation cohort

A local cohort including 100 patients was enrolled and bulk RNAseq was performed to perform CMS subtyping.

The study was approved by the Institutional Ethics Committee "Gabriella Serio" of the IRCCS Istituto Tumori "Giovanni Paolo II", and all patients provided written informed consent. In light of the retrospective design, a Data Protection Impact Assessment was prepared by the Institutional Data Protection Officer in accordance with Article 36 of the EU Genaral Data Protection Regulation (Regulation (EU) 2016/679). This document was subsequently reviewed and approved by the Institutional Ethics Committee "Gabriella Serio" (Prot. n. 780/CE). The study was conducted in accordance with the Declaration of Helsinki.

In detail, Next Generation Sequencing (NGS) experiments were performed by Genomix4life S.R.L. (Baronissi, Salerno, Italy) as described in [23].

Bioinformatic analysis and CMS classification

Once the quality of all RNAseq data was confirmed to be adequate, the reference genome (hg38, downloaded from UCSC (https://genome.ucsc.edu/), was indexed using STAR (v.2.7.11b), and alignment

was performed. Following alignment, BAM files were quantified using RSEM[24], generating two types of quantification for each sample: one at the gene level and one at the isoform level.

Raw read count gene level matrix was used to run *CMSclassifier* R package [25] *CMSclassifier* accepted as input log2_scaled Gene Expression Profiles data values. Thus, we prepared two matrices, one log2 transformed and the other normalized using the Variance Stabilizing Transformation (VST) method, both obtained through *DESeq2* R package. We use *RandomForest* methods in *CMSclassify* function. In addition, *CMScaller* classification was also performed using the *CMScaller* package[26]. CMScaller accepts as input only RNA-seq data and it can be run with counts or RSEM values by setting the parameter *RNAseq=TRUE*; for VST normalized counts, the parameter was set to *RNAseq=FALSE*.

Statistical analyses and Development of the prognostic nomogram

To calculate metrics to evaluate validation performance of the in-house CMSassay, sensitivity, specificity, accuracy was calculated in R. Heatmaps with discordant cases were generated through *pheatmap* R package.

Survival model construction

Overall survival (OS) was defined as the time from diagnosis to death from any cause or last follow-up, expressed in months. Survival status was coded as an event for deceased patients and as censored for patients alive at last follow-up. A multivariable Cox proportional hazards regression model [27] was fitted using the *cph* function from the *rms* R package. The model included the following covariates selected a priori based on clinical relevance and biological plausibility: in-house CMSassay classification, sex, tumor sidedness, pathological stage, MSI status, and BRAF, KRAS, and PIK3CA mutational status. All categorical variables were modeled as factors. Model fitting retained the design matrix and response (*x = TRUE, y = TRUE*) and enabled survival probability estimation (*surv = TRUE*). The proportional hazards assumption was assessed graphically and found to be acceptable for all included covariates.

Nomogram development

Based on the fitted Cox model, a prognostic nomogram was constructed using the *nomogram* function [28] of the *rms* R package to provide individualized predictions of 3-year and 5-year overall survival. For each predictor, the nomogram assigns a number of points proportional to its regression coefficient. The sum of points across all variables yields a total prognostic score, which is mapped to the model's linear predictor (LP) and corresponding survival probabilities at predefined time horizons (36 and 60 months). Survival probabilities at 3 and 5 years were derived from the Cox model using the baseline survival function estimated by the *Survival()* function in *rms* R package. To visualize the relationship between prognostic risk and outcome, survival probabilities were plotted as continuous functions of the LP. Additionally, a transformation was applied to convert the LP into total nomogram points, allowing a dual-axis graphical representation linking LP values, total point scores, and absolute survival probabilities.

This visualization enables direct translation of the cumulative nomogram score into individualized 3- and 5-year OS estimates.

All analyses and graphical outputs were performed with *ggplot2* R package [29].

Results

Best model identification

To be able to design a custom NGS panel usable in clinical routine for CMS identification, the bioinformatic workflow was depicted in Fig. 1a. At the very first instance, an *in silico* cohort was built up starting from TCGA-COAD and TCGA-READ cohorts.

Patients to be included needed to have transcriptomic and mutational data available and MSI status information. Thus, the final matrix included 232 samples and 3804 features which correspond to 273 genes, 3491 lncRNAs, 39 alterations and MSI status (Table 1).

mRMR technique was used to perform feature selection through two ML models, SVM and RF.

Table 1
Number of original features and final selected features used for the design of the CMS assay

Data Category	Number features original dataset	Final features number
Gene Expression	273	7
Lnc RNAs	3491	32
MSI status	1	1
Alterations	39	30

As mentioned above, we used a wrapper feature selection method: we selected features from the overall dataset to form a small feature set. The variables were sorted by their scores based on mRMR algorithm; the first *N* variables were selected as our features. The best *N* was searched by step 10, from 10 to 200. For each subset of *N* features the classification result was observed. The best global accuracy achieved was 0.76 obtained when *N* = 70 with RF classifier (Fig. 1b). The final model included 39 expression-related features (7 protein-coding genes and 32 lncRNAs), 30 DNA-related features, together with MSI status (Table 2). In Fig. 1c, precision, recall, F1 and accuracy per CMS class are shown, highlighting a poorer performance for CMS2/3 than for CMS1/4.

Table 2
List of final lncRNA, gene expression and mutation features selected for the in-house CMSassay panel.

LncRNA	Gene expression	Alterations
lnc-COX7C-7	DPYSL3	MTOR
lnc-MBL2-2	SLIT2	NRAS
LINC01980	RECK	FANCL
LINC00397	MRVI1	BARD1
LINC00355	IGF1	IDH1
lnc-SHISA9-3	SYNM	RAF1
lnc-EVX1-14	UCHL1	PIK3CA
lnc-KCNC2-2		TSC1
LINC02178		PTCH1
POU6F2-AS2		FGFR2
lnc-SNCA-5		PTEN
LINC00393		HRAS
lnc-CDH8-10		CHEK1
C21orf62-AS1		ATM
lnc-TSPY2-6		KRAS
LINC01036		FLT3
SERTAD4-AS1		BRCA2
lnc-TACR3-1		RAD51B
lnc-PWWP2B-4		TSC2
LINC02405		CDK12
lnc-FAT1-1		RAD51C
RPARP-AS1		NF1
lnc-HOXC8-1		BRIP1
lnc-CIT-1		BRCA1
lnc-TMEM132B-2		ERCC2
LINC00639		SMARCB1

LncRNA	Gene expression	Alterations
LINC01967		CHEK2
SREBF2-AS1		KDM6A
FAM66C		BRAF
SNHG1		MAP2K1
lnc-ALX1		
lnc-ALDH1A3		

Design of the CMS in-house assay

The selected features by the approach mRMR + RF were used to design a custom amplicon-based NGS assay. In detail, the assay is a combination of two custom panels, one for the evaluation of genomic alterations and the other for expression, both coding genes and lncRNAs. As described in Methods section, MSI status was evaluated through an IVD assay.

The DNA panel included the full coding sequence of 30 genes (Table 2), with a final coverage of 99.21% with 110690 bases covered over 111562 bases.

The design included 1508 amplicons with an insert size ranging 125-175bp. In Fig. 2, it is possible to observe the distribution of amplicon insert size which is tight and unimodal and with no long tails beyond ~ 175 bp. This design was considered optimal for FFPE samples to obtain a uniform amplification and a balanced coverage across targets. Indeed, we observed almost $3800*10^3$ mapped reads, 98% of reads were on target and almost a mean depth of 2300.

Regarding RNA-based panel, one/amplicon per target was designed taking into account the use of FFPE samples. We were able to obtain 100% of targets detected with almost 70% of valid reads and $2500*10^3$ of mapped ones.

Validation in a local independent cohort

To validate the in-house CMSassay, an internal retrospective cohort including 100 patients has been enrolled. For the internal cohort, bulk RNA-seq was performed to be able to run *CMSclassifier*.

As displayed in Fig. 3a-b, the CMS distribution is very similar in our cohort and in TCGA cohort, which was used to identify the model and the design the in-house CMSassay. The general features of the local cohort are listed in Table 3.

The performances of in-house CMSassay were CMS-related. Indeed, we observed very good performances for CMS1 and CMS2 reaching a balanced accuracy of 0.87 and 0.82 and a specificity of 0.88 and 0.64, respectively (Table 4).

Table 3
Features of internal validation cohort.

Characteristic	N = 100[1]
Sex	
F	48 (48%)
M	52 (52%)
Age	74 (65, 79)
Sideness	
Left	47 (47%)
Right	53 (53%)
Pathological Stage	
Stage I	8 (8.0%)
Stage II	23 (23%)
Stage III	38 (38%)
Stage IV	31 (31%)
CMS	
CMS1	17 (17%)
CMS2	37 (37%)
CMS3	14 (14%)
CMS4	32 (32%)

[1]n (%); Median (Q1, Q3)

To deeply dive into classification performances in terms of discordant cases per CMS, the raw read count martix of the internal cohort was normalized using *VST* function in *DESeq2* R package and log2 transformation. These two matrices were used to run *CMScaller* algorithm and *CMSclassifier*, the latter being run on the VST-normalized matrix.

We observed that, across all classification methods, the largest number of discordant cases involved misclassification from CMS2 to CMS3 and from CMS2 to CMS4. Notably, *CMScaller*, in both configurations, showed poor performance in distinguishing CMS3 from CMS1 and CMS2. Moreover, CMSclassifier applied to VST-normalized data misclassified four CMS4 cases as CMS3 (Fig. 3b). Overall,

these findings suggest the presence of method- and normalization-dependent internal biases in CMS classification.

Table 4
Sensitivity, specificity and balanced accuracy for each CMS subtypes obtained with the in-house CMSassay

sensitivity	**CMS1**	**CMS2**	**CMS3**	**CMS4**
	0,857143	**1**	**0,145833**	**0,340909**
specificity	0,88172	0,636364	0,865385	0,696429
balanced_accuracy	0,869432	0,818182	0,505609	0,518669

Biological insights of in-house CMS assay and prognostic nomogram

Biologically speaking CMS classification has intrinsically both a predictive and a prognostic function. Indeed, CMS1 immune subtype includes patients which could benefit from immunecheckpoint inhibitors and CMS4 patients have a very poor prognosis. Thus, we explore the distribution of in-house CMSassay classification with respect to routinely evaluated diagnostic (MS status, BRAF, KRAS and PIK3CA mutational status) and known prognostic features (sideness and stage). CMS1 is enriched of patients with II stage, right-sided tumors, BRAF and PIK3CA alterations, besides the predictable high number of MSI. CMS3 and CMS4 patients were found to be enriched of KRAS and PIK3CA alterations (Fig. 4a).

The chance to include in-house CMSassay into the diagnostic routine led us to the built-up of a nomogram. The nomogram (Fig. 4b) incorporates in-house CMSassay classification, sex, tumor sidedness, pathological stage, MSI status, and BRAF, KRAS, and PIK3CA mutational status. Each covariate contributes a variable-specific number of points proportional to its effect size in the Cox proportional hazards model. The cumulative score, obtained by summing individual points across all predictors, corresponds to a linear predictor value, which is subsequently translated into predicted 3- and 5-year overall survival probabilities.

Among all variables, tumor stage and CMS classification contributed the largest range of points, indicating a dominant impact on survival risk. Molecular alterations (MSI and oncogenic mutations) and clinicopathological features provided additional, incremental prognostic information.

Figure 4c illustrates the continuous relationship between the linear predictor derived from the nomogram and the corresponding predicted survival probabilities at 3 years (solid line) and 5 years (dashed line).

Increasing LP values—reflecting higher total point scores—were associated with a monotonic decrease in survival probability, with a steeper decline observed for 5-year survival compared to 3-year survival. The upper x-axis translates LP values into total nomogram points, enabling direct visual conversion from the cumulative nomogram score to absolute survival probabilities. Patients with low total point scores

exhibited high predicted survival at both time horizons, whereas high-risk patients showed markedly reduced long-term survival, approaching zero at extreme LP values.

Discussion

The CMS taxonomy has consolidated transcriptomic heterogeneity of colorectal cancer into four biologically and clinically meaningful groups. Indeed, CMS framework has profoundly shaped the biological interpretation of colorectal cancer heterogeneity, providing a reproducible transcriptomic taxonomy with well-defined associations to prognosis, tumor microenvironment, and key molecular alterations [4, 30, 31].

In the present study, we demonstrate that CMS classification retains robust prognostic relevance when implemented through a compact, FFPE-compatible in-house assay and integrated into a multivariable survival model alongside established clinicopathological and molecular factors.

In line with previous reports, the in-house CMSassay recapitulated these patterns, with CMS1 enriched for right-sided, MSI-high, BRAF/PIK3CA-mutant, advanced-stage tumors and CMS4 associated with adverse clinicopathological features and poorer survival, supporting the biological plausibility of the panel and its potential integration into routine FFPE-based workflows [31].

Systematic reviews and meta-analyses confirm CMS as a prognostic biomarker across stages, with CMS4 linked to worst outcomes and CMS2 to best, independent of standard clinicopathological factors [32].

Emerging evidence also suggests predictive utility, such as CMS1 enrichment in immunotherapy responders and CMS2 benefit from anti-EGFR agents, though prospective trials validating CMS-guided therapy allocation remain scarce. By enabling CMS assessment from targeted FFPE panels, the assay bridges this gap toward precision oncology implementation [33].

Current implementations of CMS in clinical research mainly rely on *CMSclassifier* [4] and *CMScaller* [10], both of which were derived from, or benchmarked on, large meta-cohorts profiled predominantly with microarray or high-quality bulk RNA-seq and assume well-normalized, genome-wide gene expression matrices.

In our study, direct comparison of *CMSclassifier* and *CMScaller* on the same validation cohort revealed substantial discordance, particularly at the CMS2–CMS3 and CMS2–CMS4 boundaries. Moreover, CMS assignments proved sensitive to data preprocessing choices, with normalization strategy (VST versus log2 transformation) influencing subtype calls. These findings are consistent with previous reports showing reduced classification stability for CMS3 and CMS4 and increased uncertainty near subtype borders[34].

Multiple independent cohorts and meta-analyses have established CMS as an independent prognostic factor in both early-stage and metastatic colorectal cancer, with CMS1 and CMS4 generally associated

with worse long-term outcomes compared with CMS2 and, variably, CMS3 [32]. The present study confirms this prognostic impact in a real-world FFPE cohort, where CMS classification derived from the in-house assay contributed one of the largest point ranges in the multivariable Cox-based nomogram, second only to pathological stage, and provided incremental risk stratification beyond MSI, BRAF, KRAS and PIK3CA status.

By integrating molecular (CMS, MSI, driver mutations) and clinicopathological variables (sex, sidedness, stage), the nomogram translates CMS information into individualized 3- and 5-year overall survival estimates, which can be readily interpreted at the bedside. Thus, this approach moves from subtype labels to quantitative survival probabilities that may support patient counseling, follow-up intensity, and trial stratification.

Despite its robust prognostic value, the direct impact of CMS on systemic therapy selection remains modest compared with single-marker biomarkers such as RAS, BRAF, MSI and HER2 [35]. Retrospective analyses suggest differential chemosensitivity and variable benefit from targeted agents across CMS groups, but the benefit of intensified regimens such as FOLFOXIRI plus bevacizumab appears largely independent of CMS, and prospective trials specifically powered to allocate therapy based on CMS are still lacking [33]. In this context, the in-house CMSassay should currently be viewed primarily as a high-resolution prognostic and biological stratification tool, rather than a stand-alone predictive test guiding drug choice.

This study has several limitations. First, discordance observed between *CMSclassifier* and *CMScaller* highlights the inherent instability of current CMS tools when applied across platforms and normalization strategies. This limitation reflects broader challenges in CMS standardization rather than assay-specific shortcomings, but it underscores the need for assay-calibrated classifiers.

Finally, the prognostic nomogram was internally derived and requires external validation in independent, multi-center cohorts before clinical adoption. Prospective studies will be necessary to determine whether CMS-integrated prognostic tools can meaningfully influence clinical decision-making and patient outcomes.

Conclusion

By enabling CMS classification from routine FFPE material using a compact DNA/RNA panel combined with MSI testing, the present assay addresses a key translational barrier to CMS implementation. While this approach does not overcome the current limitations of CMS-guided therapy selection, it provides a clinically feasible platform to incorporate CMS into prognostic algorithms and real-world datasets.

Future advances are likely to emerge from the integration of CMS with additional layers of tumor biology, including immune composition, stromal architecture, and functional drug sensitivity profiling. Such integrative strategies may ultimately convert the currently limited therapeutic output of CMS into actionable subtype-specific treatment algorithms.

Declarations

Ethics approval and consent to participate

The Institutional Ethics Committee “Gabriella Serio” of the IRCCS Istituto Tumori “Giovanni Paolo II” approved the study Given the retrospective nature of the study, a Data Protection Impact Assessment document was redacted by the institutional Data Protection Officer in compliance with Article 36 of the EU General Data Protection Regulation (Regulation (EU) 2016/679). The document was evaluated and approved by the Institutional Ethics Committee “Gabriella Serio” (Prot n. 780/CE). The study is compliant with the Declaration of Helsinki.

The authors affiliated to the IRCCS Istituto Tumori “Giovanni Paolo II”, Bari, are responsible for the views expressed in this article, which do not necessarily represent the Institute.

Clinical trial number: not applicable.

Consent for publication

Not applicable

Availability of data and materials

Bulk RNAseq raw data were deposited in SRA repository under the accession number PRJNA1304982 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1304982).

Competing interests

The authors report no disclosures relating to this work.

Funding

“Tecnopolo per la Medicina di Precisione”, CUP: B84I18000540002.

Authors' contributions

MC performed bioinformatic analyses and feature selection; LM: data management; DT, BP, GM performed all sequencing experiments; TMM validation analyses; EM, FAZ, AR, FC histological revision and FFPE sample retrieval; RDL, GB enrollment; ST, SDS, conceived the study and supervised the project.

Acknowledgements

All the patients.

References

1. Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. CA Cancer J Clin. 2014;64:104–17.
2. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. Cancer J Clin. 2021;71:209–49.
3. Dang Q, Zuo L, Hu X, et al. Molecular subtypes of colorectal cancer in the era of precision oncotherapy: Current inspirations and future challenges. Cancer Med. 2024;13:e70041.
4. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21:1350–6.
5. Eso Y, Shimizu T, Takeda H, et al. Microsatellite instability and immune checkpoint inhibitors: toward precision medicine against gastrointestinal and hepatobiliary cancers. J Gastroenterol. 2020;55:15–26.
6. Ding X, Huang H, Fang Z, et al. From Subtypes to Solutions: Integrating CMS Classification with Precision Therapeutics in Colorectal Cancer. Curr Treat Options Oncol. 2024;25:1580–93.
7. Valenzuela G, Canepa J, Simonetti C, et al. Consensus molecular subtypes of colorectal cancer in clinical practice: A translational approach. World J Clin Oncol. 2021;12:1000–8.
8. Feliu J, Gámez-Pozo A, Martínez-Pérez D, et al. Functional proteomics of colon cancer Consensus Molecular Subtypes. Br J Cancer. 2024;130:1670–8.
9. Chong W, Zhu X, Ren H, et al. Integrated multi-omics characterization of KRAS mutant colorectal cancer. Theranostics. 2022;12:5138–54.
10. Eide PW, Bruun J, Lothe RA, et al. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. 2017;7:16618.
11. Quinn GP, Sessler T, Ahmaderaghi B, et al. classifieR a flexible interactive cloud-application for functional annotation of cancer transcriptomes. BMC Bioinformatics. 2022;23:114.
12. GDC Data Portal Homepage. https://portal.gdc.cancer.gov/ (accessed 21 January 2026).
13. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44:e71.
14. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40.
15. Volders P-J, Anckaert J, Verheggen K, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. 2019;47:D135–9.
16. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
18. Chakravarty D, Gao J, Phillips S et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol 2017; 1: PO.17.00011.
19. De Jay N, Papillon-Cavanagh S, Olsen C, et al. mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics. 2013;29:2365–8.

20. Mayoraz E, Alpaydin E. Support vector machines for multi-class classification. In: Mira J, Sánchez-Andrés JV, editors. Engineering Applications of Bio-Inspired Artificial Neural Networks. Berlin, Heidelberg: Springer; 1999. pp. 833–42.
21. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
22. OncoKB. ™ · GitHub, https://github.com/oncokb/oncokb-annotator (accessed 19 January 2026).
23. Monaco D, Traversa D, Mattioli E, et al. Biological and prognostic relevance of A-to-I RNA editing across consensus molecular subtypes of colon cancer. Sci Rep. 2026;16:4018.
24. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
25. Sage-Bionetworks/CMSclassifier. https://github.com/Sage-Bionetworks/CMSclassifier (2025, accessed 20 January 2026).
26. peterawe. peterawe/CMScaller. https://github.com/peterawe/CMScaller (2025, accessed 20 January 2026).
27. Cox DR. Regression Models and Life-Tables. J Roy Stat Soc: Ser B (Methodol). 1972;34:187–202.
28. Zhang Z, Kattan MW. Drawing Nomograms with R: applications to categorical outcome and survival data. Ann Transl Med. 2017;5:211.
29. Wilkinson L. WICKHAM Biometrics. 2011;67:678–9. ggplot2: Elegant Graphics for Data Analysis by H.
30. Dienstmann R, Vermeulen L, Guinney J, et al. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer. 2017;17:79–92.
31. Sawayama H, Miyamoto Y, Ogawa K, et al. Investigation of colorectal cancer in accordance with consensus molecular subtype classification. Ann Gastroenterol Surg. 2020;4:528–39.
32. Ten Hoorn S, de Back TR, Sommeijer DW, et al. Clinical Value of Consensus Molecular Subtypes in Colorectal Cancer: A Systematic Review and Meta-Analysis. J Natl Cancer Inst. 2022;114:503–16.
33. Borelli B, Fontana E, Giordano M, et al. Prognostic and predictive impact of consensus molecular subtypes and CRCAssigner classifications in metastatic colorectal cancer: a translational analysis of the TRIBE2 study. ESMO Open. 2021;6:100073.
34. Trinh A, Trumpi K, De Sousa E, Melo F, et al. Practical and Robust Identification of Molecular Subtypes in Colorectal Cancer by Immunohistochemistry. Clin Cancer Res. 2017;23:387–98.
35. Martelli V, Pastorino A, Sobrero AF. Prognostic and predictive molecular biomarkers in advanced colorectal cancer. Pharmacol Ther. 2022;236:108239.

Figures

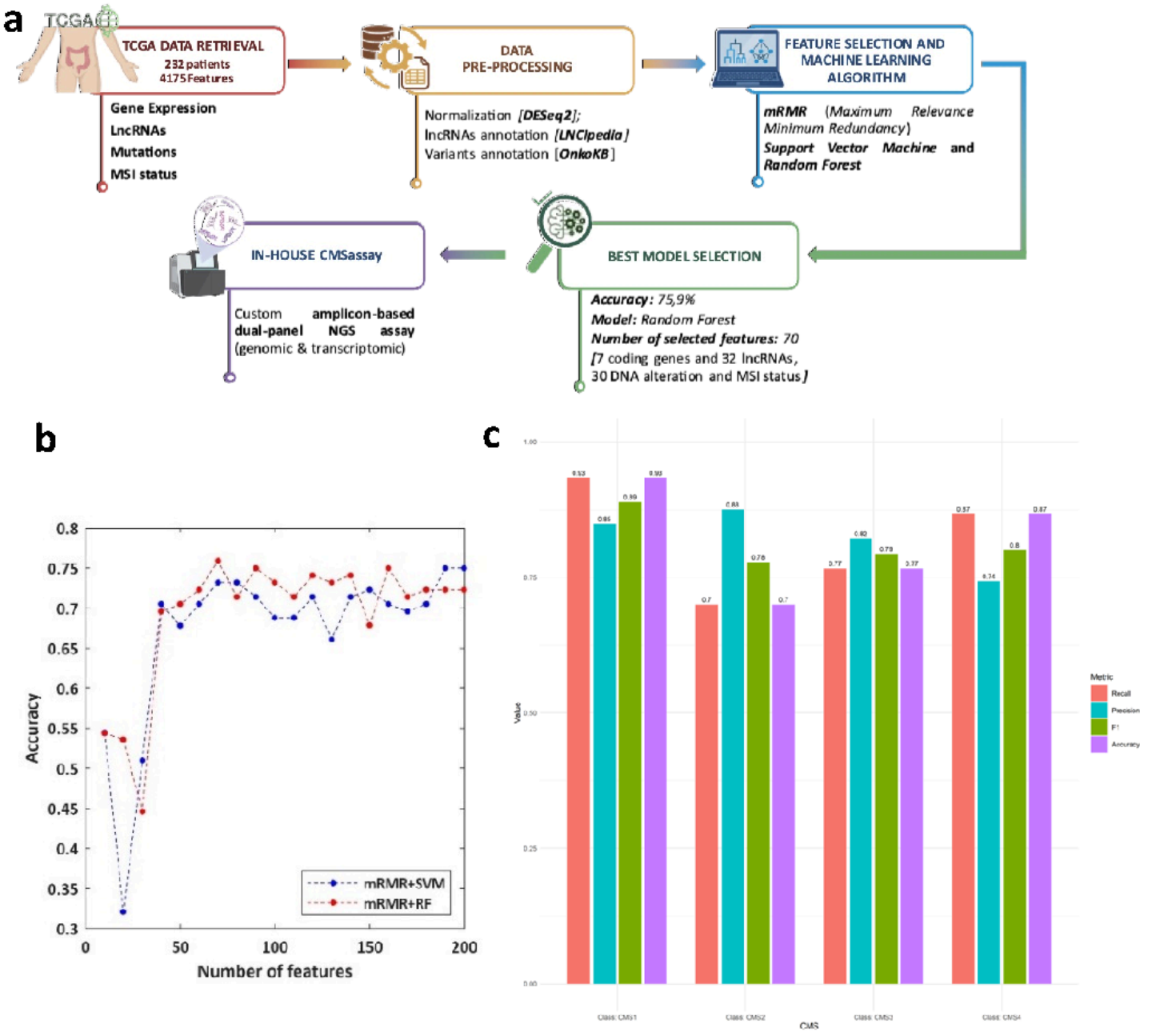


Figure 1

(a) Bioinformatic workflow for the development of the in-house CMSassay. (b) Classification accuracy as function of the number of selected features using the mRMR features selection method for SVM classifier (blue dots) and RF classifier (red dots). (c) Precision, recall, F1 score and accuracy obtained by selecting 70 features using the RF classifier for each CMS subtypes.

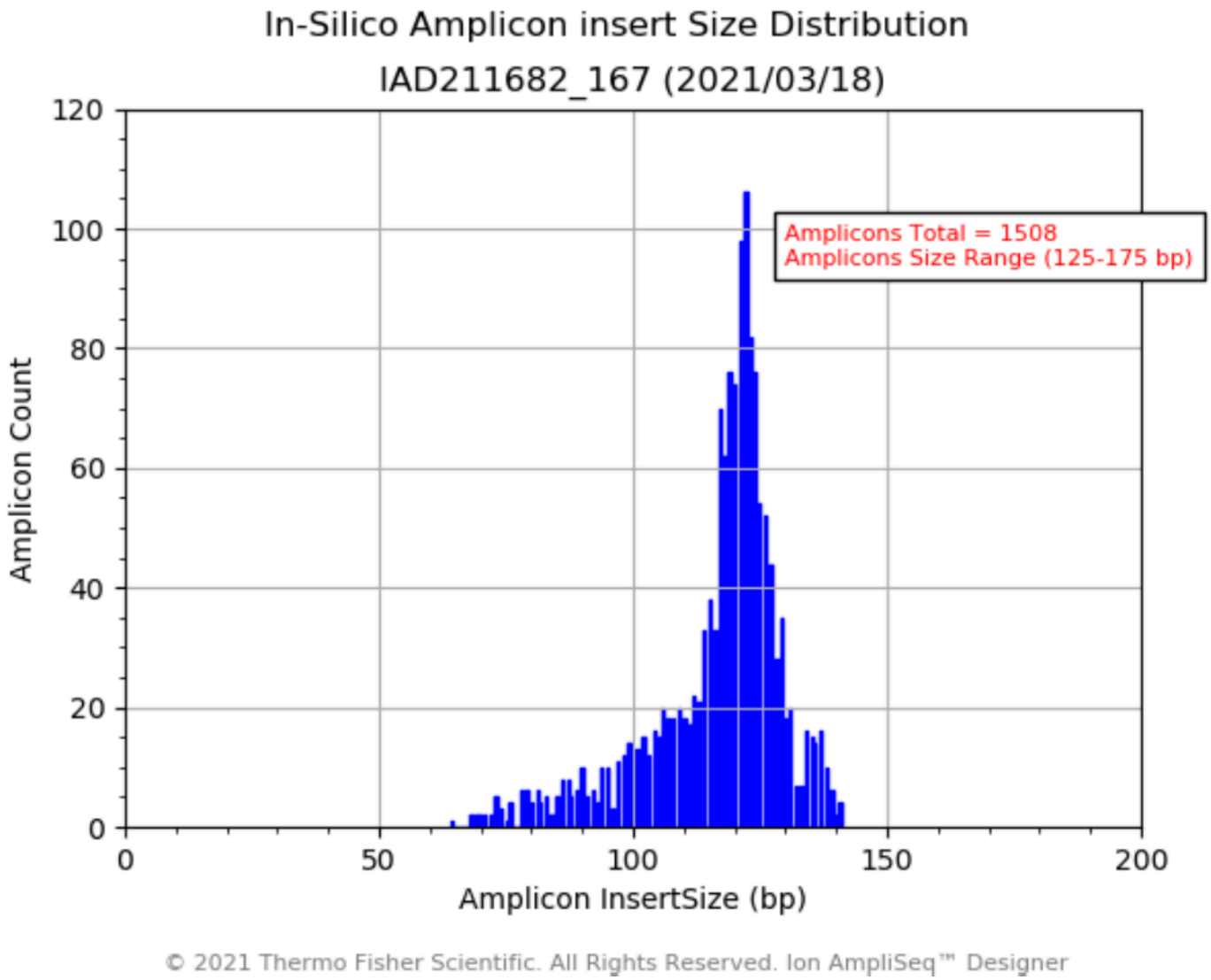


Figure 2

Distribution of amplicon insert sizes for the DNA panel.

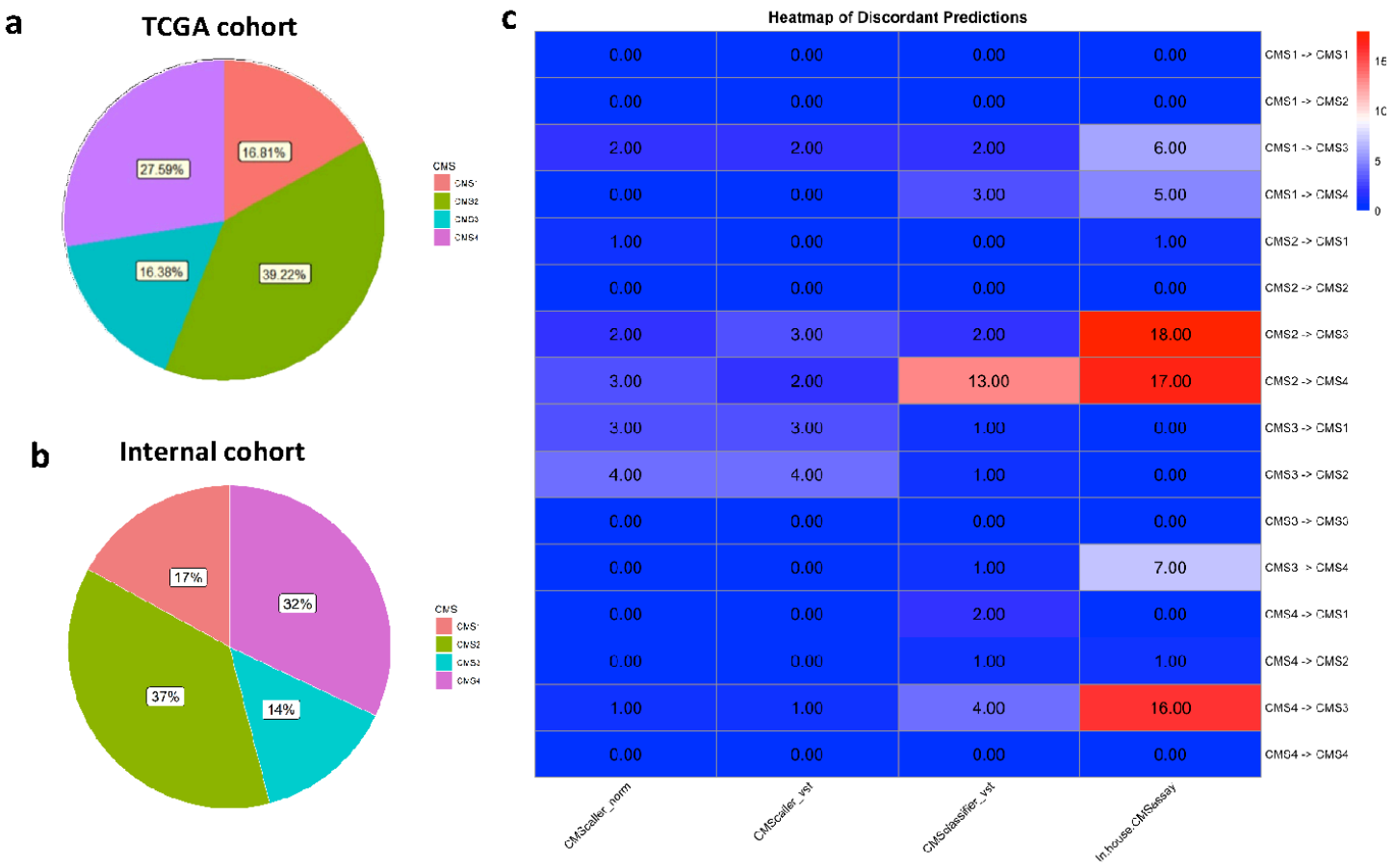


Figure 3

Distribution of CMS subtypes in (a) the TCGA cohort and in (b) the internal validation cohort. (c) Heatmap of discordant predictions obtained using log2-normalized and VST-normalized data classified with CMScaller, VST-normalized data classified with CMSclassifier, and with the in-house CMSassay.

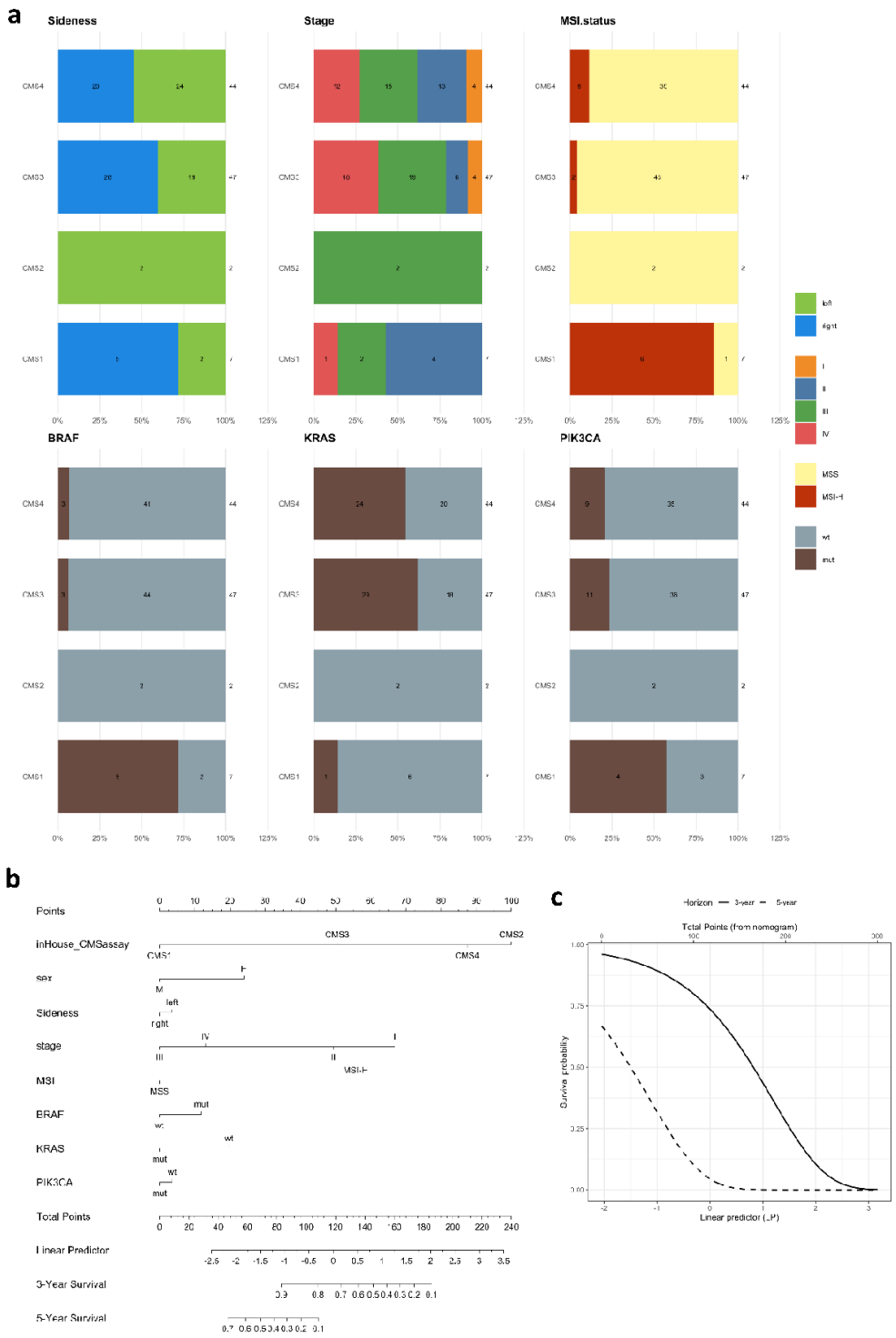


Figure 4

(a) Biological insights from the in-house CMSassay, showing the distribution of tumor sidedness, pathological stage, MSI status, and BRAF, KRAS and PIK3CA mutations across CMS subtypes. (b) Nomogram incorporating in-house CMSassay classification, sex, tumor sidedness, pathological stage, MSI status, and BRAF, KRAS, and PIK3CA mutational status. (c) Continuous relationship between the

linear predictor derived from the nomogram and the corresponding predicted 3-year (solid line) and 5-year (dashed line) survival probabilities.