Published in final edited form as:

Nat Med. 2015 November ; 21(11): 1350–1356. doi:10.1038/nm.3967.

The Consensus Molecular Subtypes of Colorectal Cancer

Justin Guinney[1,*], **Rodrigo Dienstmann**[1,2,*], **Xin Wang**[3,4,*], **Aurélien de Reyniès**[5,*], **Andreas Schlicker**[6,*], **Charlotte Soneson**[7,*], **Laetitia Marisa**[5,*], **Paul Roepman**[8,*], **Gift Nyamundanda**[9,*], **Paolo Angelino**[7], **Brian M. Bot**[1], **Jeffrey S. Morris**[10], **Iris M. Simon**[8], **Sarah Gerster**[7], **Evelyn Fessler**[3], **Felipe de Sousa e Melo**[3], **Edoardo Missiaglia**[7], **Hena Ramay**[7], **David Barras**[7], **Krisztian Homicsko**[11], **Dipen Maru**[10], **Ganiraju C. Manyam**[10], **Bradley Broom**[10], **Valerie Boige**[12], **Beatriz Perez-Villamil**[13], **Ted Laderas**[1], **Ramon Salazar**[14], **Joe W. Gray**[15], **Douglas Hanahan**[11], **Josep Tabernero**[2], **Rene Bernards**[6], **Stephen H. Friend**[1], **Pierre Laurent-Puig**[16,§], **Jan Paul Medema**[3,§], **Anguraj Sadanandam**[9,§], **Lodewyk Wessels**[6,§], **Mauro Delorenzi**[7,17,§], **Scott Kopetz**[10,§], **Louis Vermeulen**[3,§], and **Sabine Tejpar**[18,§]

[1]Sage Bionetworks, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA [2]Vall d'Hebron Institute of Oncology (VHIO), Universitat Autònoma de Barcelona, Barcelona, Spain [3]LEXOR, Center for Experimental Molecular Medicine (CEMM), Academic Medical Center (AMC), University of Amsterdam, Amsterdam, Netherlands [4]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong [5]Ligue Nationale Contre le Cancer, Paris, France [6]The Netherlands Cancer Institute (NKI), Amsterdam, Netherlands [7]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland [8]Agendia NV, Amsterdam, Netherlands [9]The Institute of Cancer Research, London, UK [10]The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA [11]Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland [12]Gustave Roussy, Villejuif, France [13]Laboratorio de Genomica y Microarrays, Instituto de Investigación Sanitaria San Carlos, Hospital Clinico San Carlos, Madrid, Spain [14]Institut Catala d'Oncologia, L'Institut d'Investigació Biomèdica de Bellvitge, Barcelona, Spain [15]Oregon Health

Corresponding authors: Justin Guinney, justin.guinney@sagebase.org, Louis Vermeulen, l.vermeulen@amc.uva.nl, Sabine Tejpar, sabine.tejpar@uzleuven.be.

*co-first authors

§co-senior authors

Accession codes

- doi:10.7303/syn2623706: normalized gene expression data, CMS subtyping calls, and sample annotation from public data sets used in the consortium
- https://github.com/Sage-Bionetworks/crcsc: cripts and code for the Random Forest CMS classifier
- *CMSclassifier* R package

Author contributions

Conception and design: J.G., R.D., J.P.M., A.S., L.W., M.D., S.K., L.M., L.V., S.T., S.H.F.
Provision of study materials: A.d.R., P.R., P.L.P. I.S., E.F., F.S.M., E.M., D.B., K.H.,J.W.G., D.H., J.T., R.B., J.P.M.,A.S., L.W., M.D., S.K., L.V., S.T.
Collection and assembly of data: J.G., R.D., P.A., B.B., S.G., E.F., D.B., K.H., D.M., G.M., B.B.
Data analysis and interpretation: J.G., R.D., X.W., A.d.R., A.S., C.S., L.M., G.N., P.A., B.B., J.M., T.L., L.V., A.S., M.D.
Manuscript writing: J.G. R.D., X.W., A.d.R, A.S., C.S., L.M.. J.T., J.P.M., A.S., M.D., S.K., L.V., S.T.
All authors contributed to final approval of the manuscript

Competing Financial Interest: Paul Roepman, Iris Simon, and Rene Bernards are employees of Agendia NV.

Sciences University, Portland, OR, USA [16]Universite Paris Descartes, Paris, France [17]University of Lausanne, Lausanne, Switzerland [18]Universitair ziekenhuis Leuven, Leuven, Belgium

Abstract

Colorectal cancer (CRC) is a frequently lethal disease with heterogeneous outcomes and drug responses. To resolve inconsistencies among the reported gene expression–based CRC classifications and facilitate clinical translation, we formed an international consortium dedicated to large-scale data sharing and analytics across expert groups. We show marked interconnectivity between six independent classification systems coalescing into four consensus molecular subtypes (CMS) with distinguishing features: CMS1 (MSI Immune, 14%), hypermutated, microsatellite unstable, strong immune activation; CMS2 (Canonical, 37%), epithelial, chromosomally unstable, marked WNT and MYC signaling activation; CMS3 (Metabolic, 13%), epithelial, evident metabolic dysregulation; and CMS4 (Mesenchymal, 23%), prominent transforming growth factor β activation, stromal invasion, and angiogenesis. Samples with mixed features (13%) possibly represent a transition phenotype or intra-tumoral heterogeneity. We consider the CMS groups the most robust classification system currently available for CRC – with clear biological interpretability – and the basis for future clinical stratification and subtype–based targeted interventions.

Introduction

Gene expression-based subtyping is widely accepted as a relevant source of disease stratification[1]. Despite the widespread use, its translational and clinical utility is hampered by discrepant results, likely related to differences in data processing and algorithms applied to diverse patient cohorts, sample preparation methods, and gene expression platforms. In the absence of a clear methodological gold standard to perform such analyses, a more general framework that integrates and compares multiple strategies is needed to define common disease patterns in a principled, unbiased manner. Here, we describe such a framework and its application to elucidate the intrinsic subtypes of colorectal cancer (CRC).

Inspection of the published gene expression-based CRC classifications[2–9] revealed only superficial similarities. For example, all groups identified one tumor subtype enriched for microsatellite instability (MSI) and one subtype characterized by high expression of mesenchymal genes, but failed to achieve full consistency among the other subtypes. We envisioned that a comprehensive cross-comparison of subtype assignments obtained by the various approaches on a common set of samples could resolve inconsistencies in both the number and interpretation of CRC subtypes. The CRC Subtyping Consortium (CRCSC) was formed to assess the presence or absence of core subtype patterns among existing gene expression-based CRC subtyping algorithms. Recognizing that transcriptomics represents the level of high-throughput molecular data that is most intimately linked to cellular/tumor phenotype and clinical behavior, we also wanted to characterize the key biological features of the core subtypes, integrate and confront all other available data sources (mutation, copy number, methylation, microRNA, proteomics), and assess whether the subtype assignment

correlated with patient outcome. Furthermore, our aim was to establish an important paradigm for collaborative, community-based cancer subtyping that will facilitate the translation of molecular subtypes into the clinic, not only for CRC but other malignancies as well.

Results

Comparison of published molecular subtyping platforms

We evaluated the results of six CRC subtyping algorithms[3–8], each developed independently utilizing different gene expression data sets and analytical approaches (Supplementary Tables 1 and 2). Figure 1 summarizes the workflow of our analysis. Eighteen CRC data sets (n = 4,151 patients), both public (GSE42284, GSE33113, GSE39582, GSE35896, GSE13067, GSE13294, GSE14333, GSE17536, GSE20916, GSE2109, GSE2109, TCGA) and proprietary[3,10] (Supplementary Table 3), consisting of multiple gene expression platforms (Affymetrix, Agilent, and RNA-sequencing), sample types (fresh-frozen and formalin-fixed paraffin-embedded [FFPE]), and study designs (retrospective and prospective series, and one clinical trial[10]) were uniformly pre-processed and normalized from raw formats to reduce technical variation. The six expert groups applied their subtyping classification algorithm to each of the data sets separately to ensure correct method utilization and interpretation of results. The output of this workflow was six different subtype labels per sample.

We developed a network-based approach to study the association among the six CRC classification systems, each consisting of three to six subtypes and collectively numbering 27 unique subtype labels. In this association network, nodes corresponded to the union of all group subtypes (n = 27), and weighted edges encoded the Jaccard similarity coefficients between nodes. We then applied a Markov Cluster (MCL) algorithm[11,12] to this network to detect the presence of robust network substructures that would indicate recurring subtype patterns. During network clustering using MCL, network granularity is controlled by inflation factor f, which is associated with the number of clusters[11,12]. For varying inflation factors, we compared the corresponding clustering performances using *weighted silhouette width* (**Online Methods**). Using the optimal inflation factor (Supplementary Fig. 1), we identified four robust consensus molecular subtypes (CMS) with significant interconnectivity (P <0.001, hypergeometric test) among the six independent classification systems (Fig. 2a,b). The network-based approach revealed a set of core consensus samples, i.e., tumors representative of each CMS (3,104 of 3,962 samples [78%]) with a high concordance in subtype labels among the groups (P <0.05, hypergeometric test). The remaining unlabeled (non-consensus) samples, which did not have a consistent pattern of subtype label association, represented a substantial proportion of primary tumors (n = 858 [22%]) (Fig. 2b). Importantly, these samples were distributed across all data sets (Supplementary Fig. 2). In addition, visualization of the global patient network revealed that non-consensus samples remained scattered between the four large primary modules (Fig. 2c).

Consensus molecular subtype classification

Using the CMS labels of the core consensus samples as a 'gold standard', we developed a novel classification framework for predicting CMS subtypes using aggregated gene expression data from all cohorts (**Online Methods**). CMS labeled samples were split into two equal partitions for training and validation, and a Random Forest classifier was generated from 500 balanced bootstraps of the training data. When applied to the validation data, the classifier demonstrated robust performance across gene expression platforms (Affymetrix, Agilent, and RNA-sequencing) and sample collections (FFPE, fresh-frozen) with a >90% balanced accuracy across all subtypes (Supplementary Table 4, Supplementary Fig. 3). This corroborates both the portability of the classifier as well as the evident subtype-specific signals across datasets.

The CMS classifier allowed characterization of the originally unlabeled samples from network analysis (n = 858). Using a conservative posterior probability threshold with high specificity (**Online Methods**), we were able to assign 40% of these samples (n = 339) to a single subtype (Supplementary Fig. 4) and the remaining unclassified samples (n = 519 [13% of the overall population]) had heterogeneous patterns of CMS mixtures (Supplementary Fig. 5). We confirmed that 'mixed' samples were not outliers and did not represent a fifth independent subtype (Supplementary Fig. 5), although the quality of gene expression data could have affected a small subset of samples (**Online Methods**). The final distribution of the CMS groups is shown in Fig. 2d, including 'mixed' or indeterminate samples.

Biological characterization of the consensus molecular subtypes

We studied additional molecular data that was available for a subset of the samples in our cohort (Supplementary Table 3) to delineate the biological characteristics of each CMS group. With respect to genomic aberrations, CMS1 samples were hypermutated and had low prevalence of somatic copy number alterations (SCNAs) (Fig. 3a–c and e, Supplementary Tables 5 and 6). CMS1 encompassed the majority of MSI tumors and had overexpression of proteins involved in DNA damage repair in reverse phase protein array (RPPA) analysis, consistent with defective DNA mismatch repair (Supplementary Table 7). As expected, analysis of methylation profiles in TCGA showed that CMS1 tumors display a widespread hypermethylation status (Fig. 3f, Supplementary Fig. 6). Conversely, CMS2-4 displayed higher chromosomal instability (CIN) as measured by SCNA counts (Fig. 3b, Supplementary Table 5). We detected more frequent copy number gains in oncogenes and losses in tumor suppressor genes in CMS2 than in the other subtypes (Supplementary Table 6). Notably, CMS3 samples had a distinctive global genomic and epigenomic profile as compared with other CIN tumors: (i) consistently fewer SCNAs (Fig. 3b,c and e, Supplementary Table 5), an association not explained by differences in tumor purity (Supplementary Fig. 7, Supplementary Table 5); (ii) nearly 30% were hypermutated (Fig. 3c, Supplementary Table 5), which overlapped with MSI status (Supplementary Fig. 7); and (iii) higher prevalence of CpG Island Methylator Phenotype (CIMP) low cluster in TCGA samples (Fig. 3c, Supplementary Table 5), with intermediate levels of gene hypermethylation (Fig. 3f).

Next, we sought to identify mutations that specifically associate with the CMS groups. Although we found clear enrichment of certain mutations within subtypes (Fig. 3d, Supplementary Tables 5 and 8), e.g. *BRAF* mutations frequently occurring in CMS1 – in line with the known association of this event with MSI tumors[2] – and *KRAS* mutations overrepresented in CMS3, none of the subtypes is defined by an individual event, nor is any genetic aberration limited to a subtype. Similarly, we detected no unique and recurrent SCNA that strongly associated with a subtype, albeit amplifications of the transcription factor *HNF4A* were enriched in CMS2 (Supplementary Tables 5 and 6). Since single genomic aberrations do not clearly delineate the CMS groups, we performed an integrative analysis of mutations and copy number events using TCGA data to find signal transduction cascades that might underlie the biology of the various subtypes. Apart from the nearly universal genetic activation of the receptor tyrosine kinase (RTK) and mitogen activated protein kinase (MAPK) pathways in CMS1 and CMS3, no specific associations were identified (Supplementary Fig. 7, Supplementary Table 5). This supports the notion that tumors harboring commonly assumed driver events in CRC still vary significantly in their biology and highlights the very poor genotype-phenotype correlations in this disease.

We then focused on the gene expression data and performed gene set enrichment analysis using previously described signatures of pathway activity and well-characterized cellular processes. These analyses provided substantial insight into the biological understanding of the CMS groups (Fig. 3i, Supplementary Table 9). CMS1 is characterized by increased expression of genes associated with a diffuse immune infiltrate, mainly composed of T_H1 and cytotoxic T cells, along with strong activation of immune evasion pathways, an emerging feature of MSI CRC[13] (Fig. 3i, Supplementary Table 9). CMS2 tumors displayed epithelial differentiation and strong upregulation of WNT and MYC downstream targets, classically implicated in CRC carcinogenesis (Fig. 2i, Supplementary Table 9). In contrast, enrichment for multiple metabolism signatures was pronounced in CMS3 epithelial CRCs, in line with the occurrence of *KRAS* activating mutations described as inducing prominent metabolic adaptation[14–17] (Fig. 3i, Supplementary Table 9). Interestingly, CMS3 tumors displayed similarities with a ‘metabolic’, genomically stable subtype recently described in gastric cancer[18,19]. Finally, CMS4 tumors showed clear upregulation of genes implicated in epithelial mesenchymal transition (EMT) and signatures associated with the activation of transforming growth factor β (TGF β) signaling, angiogenesis, matrix remodeling pathways and complement inflammatory system (Fig. 3i, Supplementary Table 9). In addition, CMS4 samples exhibited a gene expression profile compatible with stromal infiltration (Fig. 3i, Supplementary Table 9), overexpression of extracellular matrix proteins on RPPA analysis (Supplementary Table 7), and higher admixture with non–cancer cells, as measured by the ABSOLUTE algorithm[20] (Supplementary Fig. 7, Supplementary Table 5).

To assess whether gene expression–based subtypes are recapitulated at the protein level, we compared our CMS groups with the recently characterized proteomic clusters in TCGA samples ($n = 81$)[21]. We observed a partial concordance between the two classification systems and could describe an approximate mapping between the subtype groups (Supplementary Table 10). In a supervised analysis (Fig. 3g), CMS1 tumors showed upregulation of proteins involved in immune response pathways, while CMS4 samples had

significant overexpression of proteins implicated in stromal invasion, mesenchymal activation, and complement pathways (Fig. 3j, Supplementary Table 11).

In addition, to interrogate posttranscriptional regulation of gene expression across CMS groups, we performed supervised microRNA (miR) analysis, identifying significant subtype specific miR regulation changes (Fig. 3h, Supplementary Fig. 8, Supplementary Table 12). Of particular note, CMS2 tumors showed upregulation of the miR-17-92 cluster, a direct transcriptional target of MYC[22]; CMS3 samples had low expression of the let-7 miR family, which is accompanied by high *KRAS* expression levels, whereas the miR-200 family, previously implicated in regulation of EMT, showed clear downregulation in CMS4[23,24].

Finally, we also compared gene expression patterns of CRC tumors with: (a) adjacent normal colon tissue from patients with colon cancer, (n = 19); and (b) left colon tissue from non-cancer individuals (n = 64) (**Online Methods**). Global PCA analysis revealed that normal samples were clearly differentiated from tumor samples in both cohorts (Supplementary Fig. 9). Although CMS3 tumors appeared more ‘normal’-like at the gene expression level (Supplementary Fig. 9), we did not find greater normal tissue contamination in CMS3 group after pathological review of a subset of samples from PETACC-3 clinical trial[10], as well as ABSOLUTE tumor purity scores in TCGA data (Supplementary Fig. 7, Supplementary Table 5).

Clinical and prognostic associations of the consensus molecular subtypes

We also found important associations between CMS groups and clinical variables (Fig. 4, Supplementary Table 5). CMS1 tumors were frequently diagnosed in females with right-sided lesions (Fig. 4a,b, Supplementary Fig. 10 and Supplementary Table 5) and presented with higher histopathological grade (Fig. 4d, Supplementary Table 5). Conversely, CMS2 tumors were mainly left-sided (Fig. 4b, Supplementary Fig. 10 and Supplementary Table 5). CMS4 tumors tended to be diagnosed at more advanced stages (III and IV) (Fig. 4c, Supplementary Table 5). To determine whether the CMS groups differed in outcome, we performed a Cox Proportional Hazards analysis on the combined data sets and separately in the subset of patients enrolled in a clinical trial with uniform follow-up (PETACC-3 clinical trial[10]) (Supplementary Table 13, Supplementary Fig. 11). Irrespective of patient cohort, CMS4 tumors displayed worse overall survival (Fig. 4e) and relapse-free survival (Fig. 4f) in both univariate and multivariate analyses, after adjustment for clinicopathological features, MSI status, and *BRAF* and *KRAS* mutations (Supplementary Table 13). We also found superior survival after relapse in CMS2 patients (Fig. 4g), with a larger proportion of long-term survivors in this subset. Interestingly, the CMS1 population had very poor survival after relapse (Fig. 4g), in agreement with recent studies showing worse prognosis of patients with MSI and *BRAF*-mutated CRC that recur[25–27]. These differences in prognosis with unsupervised gene expression signatures confirm the clinical relevance of the intrinsic biological processes implicated in each CMS.

Discussion

This report is a unique example of a discovery effort performed by a community of experts to identify a consensus gene expression–based subtyping of CRC. Thanks to collaborative

bioinformatics work on the largest collection of CRC cohorts with molecular annotation to date, and building upon previous efforts by the independent researchers, the consortium resulted in a consensus molecular classification system that allows the categorization of most tumors into one of four robust subtypes. Marked differences in the intrinsic biological underpinnings of each subtype support the new taxonomy of this disease (Fig. 5) that will facilitate future research in this field and should be adopted by the community for CRC stratification: CMS1 (MSI Immune), CMS2 (Canonical), CMS3 (Metabolic), and CMS4 (Mesenchymal). From a biological perspective, we were able to refine the number and interpretation of the 'non-MSI' subtypes, which represent nearly 85% of the primary CRC samples. We also describe strong molecular associations where previous work was fragmented and inconsistent, particularly in samples lacking a mesenchymal phenotype. From a clinical perspective, in CRC as for many cancer types, it remains unclear what features will provide the most relevant subclassification tool. Gene expression subtypes have been extensively investigated in breast cancer, gene mutations and fusions in lung cancer, chromosomal alterations in hematological malignancies and histological features in sarcomas, but whether combinations of these features is needed for accurate prediction of prognosis or drug responses is still unknown. In CRC, few biomarkers have been translated to patient care, including *RAS*, *BRAF* mutations, MSI and CIMP status. It is important to emphasize that even though the CMS groups are enriched for some genomic and epigenomic markers, their associations described here are weak and do not allow categorization of gene expression subtypes, reinforcing the notion that transcriptional signatures allow refinement of disease subclassification beyond what can be achieved by currently validated biomarkers[28]. For example, *RAS* wild type tumors are considered a homogenous entity for therapeutic decisions in the advanced setting, despite being found across distinct CMS groups with profound biological differences, which are expected to translate into heterogeneous drug responses.

Future steps and resources

Qualitative and clinically relevant disease subtyping takes time and multiple resources. Our effort of CRC subclassification is a stepwise process aiming to involve a large number of relevant researchers from the CRC research community at first, and then subsequently cooperative groups, pharmaceutical companies and regulatory agencies. We postulate that the identification of molecularly homogeneous subsets of CRC tumors – and the characterization of potential driver events in these samples – will advance effective drug development strategies. Recently, MSI status was found to predict benefit of immune checkpoint blockade in advanced CRC, corroborating the value of integrating knowledge on the underlying biology with drug development strategies[29]. While admittedly speculative at this point, oncogene amplifications found in CMS2 samples and the prominent metabolic activation of CMS3 tumors have strong potential for novel targeted therapy development in CRC, and yield a well-defined and reasonably sized group in which to test these hypotheses.

Subclassification per se, even when built on what are believed to be relevant features of cancer cells such as expression of cancer pathway components or driver gene mutations, may still not be predictive of differential drug responses. This can be due to the drugs themselves, with promiscuous mechanisms of action that may not track well with single

pathway descriptors, or our inability to properly define pathway engagement or cross-talk using static omics data. Reanalysis of relevant clinical trials using semi-supervised approaches based on pre-defined patient subgroups and allowing for further discovery based on observed outcomes may be the best alternative for the research community. Our current work, providing the consensual best description of CRC heterogeneity available today, aims at providing exactly that tool for systematic interrogation in different clinical settings. It will also accelerate the application of gene classifications to cell lines, organoids and patient-derived xenograft models with drug sensitivity data.

To enable retrospective and prospective stratified explorations, we are releasing a set of CMS classifiers to be used by the community as research tools (R package available for download, see **Online Methods**), either in the context of population studies (original Random Forest classifier described above, which requires data normalization) or for use in a single sample setting (alternative Pearson-based predictor, optimized to be less dependent on pre-processing of gene expression data). Importantly, samples that do not fall within the four CMS groups should be considered separately as indeterminate subtypes, yet of unknown biological and clinical behavior.

To conclude, we believe that the framework presented here provides a common foundation of CRC subtyping, to be further refined in the future as other sources of ‘omics’ data are integrated and clinical outcomes under specific drug interventions become available. We hope that this model of expert collaboration and data sharing among independent groups with strong clinical and preclinical expertise will be emulated by other disease areas to accelerate our understanding of tumor biology.

Online Methods

1. Overall design

The design and workflow of this project is described in Figure 1. There were six participating groups, each who had previously developed and published a methodology for classifying CRC samples using gene expression data (described below). An additional group was designated as an “evaluation group” (Sage Bionetworks) to run an unbiased comparative analysis. All public and proprietary data sets (Supplementary Table 1) were uploaded into a common data repository (www.synapse.org)[31]. This project focused on the secondary analysis of existing de-identified genomic and clinical data. No readily identifiable information was included in these data sets and all patients had previously given informed consent for use of the data in future CRC research at time of specimen collection. Gene expression data was accessible to all groups, and non-expression data (i.e. clinical, molecular annotations) were accessible only to the evaluation group. Each data set was processed and normalized once, using a single protocol per platform (see section on “Gene expression processing and normalization”). While this decision precluded an analysis of the impact of gene expression normalization on subtyping, it significantly reduced the number of cross-group comparisons, and allowed this study to focus on biological interpretations of the different subtypes rather than on bioinformatic procedures. Each group then applied their subtyping classifier to the data sets in the common repository. Of note, the distribution of subtypes labels from each group as reported in corresponding subtyping publications was

maintained in the collection of data sets form the consortium (Supplementary Fig. 12). All results were deposited in Synapse, allowing for an automated evaluation of all results.

2. Colorectal cancer subtyping platforms

A summary of the six subtyping platforms is provided in Supplementary Tables 1 and 2. This includes enumeration of the methodologies and data used to define CRC subtypes, and molecular characterization of each of these subtypes.

2.1. Group A—Budinska, et al, 2013[4]: Based on a discovery data set consisting of 1,113 CRC samples and 3,025 genes with variance exceeding a given threshold, we applied hierarchical clustering to the genes, followed by dynamic tree cut to produce 54 gene modules containing in total 658 genes, as described by Budinska et al. For each sample we then computed a vector of meta-gene scores by taking the median of the expression values for the genes in each module. On the resulting meta-gene expression matrix we applied hierarchical clustering using a consensus distance, followed by dendrogram pruning, which identified five distinct subtypes. A subset of the samples, which were reliably assigned to a subtype (so called core samples), was used to define a classifier. To build the classifier, we first converted the expression values for each gene to z-scores by subtracting the mean and dividing by the standard deviation across the core samples. Then, we computed meta-gene scores by taking the median of the expression values for each sample across the genes in each of the previously defined modules. The resulting meta-gene expression matrix was used as the input to train a linear discriminant analysis (LDA) classifier for the five subtypes. To subtype the samples of an independent data set, we first computed z-scores for each gene across all samples, followed by meta-gene score computation as described above. After this preparation, the independent data set was submitted to the pre-trained LDA. For each sample, this returns the probability of belonging to each of the subtypes. In cases where a non-probabilistic partition of the samples into groups is sought, each sample is assigned to the subtype with the highest probability.

2.2. Group B—Marisa, et al, 2013[8]: A multicenter series of 556 fresh frozen tumor samples of patients with stage I to IV colon cancer, mainly retrospectively collected, was used (GSE39582, Affymetrix U133plus2 platform). All the expression profiles were normalized together using the RMA method. The ComBat method[32] was then used to correct technical batch effects. The resulting matrix was row-mean-centered. Our series was then split into a training set (n = 433) and a validation set (n = 123). CRC subtypes were derived from the training set, by applying consensus hierarchical clustering (consensus cluster plus procedure) to the expression profiles reduced to the most variant probe sets (n = 1,459). The consensus was calculated across 1,000 resampling iterations of the hierarchical clustering (linkage: Ward; inter-individual distance: 1 - Pearson correlation coefficient), each iteration being based on a random selection of 90% of the samples and 90% of the probe sets. To predict subtypes in independent data sets, we developed a centroid-based predictor using the most discriminative genes (57 genes). A tumor was assigned to the subtype of the closest centroid using diagonal LDA distance for Affymetrix data set and (1-Pearson correlation) for non-Affymetrix data sets. The confidence call of the prediction

(posterior probability approach) was determined using the distribution of the difference between the two nearest centroids on the training set.

2.3. Group C—Roepman, et al, 2014[3]: Using Agilent microarray based full genome expression data of 188 stage I–IV CRC patients, an unsupervised clustering revealed three major subtypes (A-, B-, C-type). A single sample molecular subtype classifier (Pearson correlation based nearest centroid model) was developed and validated in 543 stage II and III patients. In this consensus effort, additional CRC sample that were hybridized onto the same Agilent platform were analyzed using the exact same method as described in detail in Roepman et al, 2014. CRC samples analyzed on the Agilent platform were preprocessed by median centering within each of the Agilent data set. Following median centering, subtype similarity scores for A-type, B-type and C-type were processed similarly as the Agilent derived data.

2.4. Group D—De Sousa E Melo, et al, 2013[7]: A colon cancer subtype (CCS) classifier was derived from unsupervised classification of the core data set AMC-AJCCII-90, consisting of 90 stage II colon cancer patients (GSE33113). The microarray data were first normalized using the frozen robust multiarray analysis (fRMA)[33], with gene expression presence and absence called using the barcode algorithm[34]. After filtering out genes not present in at least one sample, 7,846 probe sets of top variability (median absolute deviation >0.5) were kept and median centered. Based on consensus clustering (1000 iterations, 0.98 subsampling ratio) and GAP statistics, we identified three robust clusters. Eighty-five samples with positive silhouette width were considered as the most representative samples and retained for following analysis. In order to allow cross-platform classification, we mapped probe sets to unique genes: for each gene we kept its corresponding probe set with highest overall expression. Significance analysis of microarrays (SAM)[35] and AUC (area under ROC curve) scores were employed to identify the most discriminative genes. Prediction analysis for microarrays (PAM)[36] was subsequently performed with tenfold cross-validation over a range of centroid shrinkage thresholds for 1000 iterations. Finally, a PAM classifier of 146 unique genes was built with the optimal threshold for centroid shrinkage selected based on a trade-off between classification performance (error rate <2%) and the size of gene signature. To use the CCS classifier, expression profiles obtained after normalization were median centered across cancer samples. For microarray data generated based on platforms other than Affymetrix Human Genome U133 Plus 2.0, probe sets were mapped to gene symbols. Signature genes without annotation were substituted by genes with highest correlation as calculated from our core data set. Median centered expression profiles of signature genes were subjected to CCS classifier for subtype prediction, which returns the posterior probability that a cancer sample belongs to each subtype. Each cancer sample is subsequently classified to the subtype with the highest probability.

2.5. Group E—Sadanandam, et al, 2013[6]: The five CRCassigner subtypes were defined using non-negative matrix factorization (NMF)-based consensus[37] clustering of two publicly available gene expression profile data sets (GSE13294 and GSE14333) merged using the distance weighted discrimination method[38]. Statistical analysis of microarrays (SAM)[35] was then used to identify the most significant differentially expressed genes between subtypes.

The prediction analysis for microarrays (PAM)-based shrunken centroid method[36] (with ten-fold cross validation) was used to define a 786-gene classifier (CRCassigner-786; PAM centroids) to assign individual CRC samples to one of five CRCassigner subtypes[6]. Here we classified the samples into five subtypes using the PAM centroids for CRCassigner-786 genes and Pearson correlations, which is different from our original publication. This method was chosen to unambiguously assign each sample to one of the five subtype labels based on correlations and to ensure consistency between the methodologies used by the groups in this consortium. Before subtyping, probe sets were mapped to their corresponding HUGO gene nomenclature committee (HGNC)-based official gene symbols. We also: (i) removed probes that did not map to any known gene symbol; (ii) removed duplicate genes by selecting probes with highest variability; and (iii) performed row (across samples) median centering for each data set. Finally, the CRCassigner-786 genes were selected from the data sets. Pearson correlations between median-centered CRCassigner-786 gene expression data for each sample and the PAM centroids were estimated for a given data set.

2.6. Group F—Schlicker, et al, 2012[5]: We derived CRC subtypes by applying iterative non-negative matrix factorization (iNMF) to data set GSE35896. Raw gene expression data were first normalized using the RMA procedure and subsequently mean-centered. Probes that were not expressed in any tumor sample were removed from the data set. Briefly, iNMF proceeds in the following steps. First, we applied non-negative matrix factorization (NMF) to 100 randomly selected groups of probe sets. Second, we hierarchically clustered samples based on how often they co-clustered in the 100 NMF runs and selected core clusters consisting of frequently co-clustering samples. Third, probe sets that were differentially expressed between the core clusters were selected as subtype signatures and all samples were assigned to subtypes by hierarchical clustering. Iteratively applying this procedure resulted in identification of five CRC subtypes. Independent data sets are subtyped by hierarchically clustering the samples using the expression signatures. In order to derive a probability value for a subtype assignment, we performed the hierarchical clustering on 10000 randomly selected bootstraps. The subtype probability is then defined as relative frequency with which a sample has been assigned to each subtype.

3. Gene expression data processing and normalization

The publicly available data sets with CRC tumor samples from the Gene Expression Omnibus (Supplementary Table 3) were normalized using the robust multi-array average (RMA) method as implemented in the *affy* package[39]. Overlapping samples in GSE14333 and GSE17536 were excluded from GSE14333. For consensus network analysis and training a consensus subtype classifier, all private and public Affymetrix data sets were renormalized using the single-sample frozen RMA method[33] as implemented in the *frma* package for R/Bioconductor.

Several of the CRC tumor sets were analysed on full genome Agilent microarrays (Agilent, Santa Clara). Samples were hybridized against a common CRC reference pool, and full genome data was normalized using loess and local background subtraction (*limma* package). Details about sample processing and microarray analysis can be found in Roepman et al.[3].

Level 3 TCGA RNA-seq data for colon and rectal was downloaded from the TCGA data portal (January 2014). RSEM normalized data[40] was further log transformed, and non-tumor samples were removed. Principal component analysis (PCA) revealed no clear differences between rectal and colon samples (data not shown) and samples were combined without adjustment. PCA showed a strong separation between GA and HiSeq samples, and corrected using the ComBat method[32].

We additionally performed outlier sample detection within each data set using 2 methods: a method based on PCA, and the *arrayQualityMetrics* R package[41]. For the PCA approach, we took into account the first two principal components and marked all samples with a distance greater than 2.5 as potential outliers. We next employed *arrayQualityMetrics* to flag outliers based on pairwise sample distances, gene expression value distributions and MA plots (MA plots were not investigated for Agilent-based expression data sets). Overall, a sample was classified as outlier if it was flagged based on the distribution of gene expression values and either pairwise distances to other samples or the PCA criterion. Outliers were removed from further analysis.

4. Network analysis of subtype association

To study the association between the six CRC classification systems (A to F, each consisting of 3, 5 or 6 subtypes and totaling 27) we employed a network-based approach. The network encodes on nodes the information of subtype prevalence and on edges their association calculated based on Jaccard similarity coefficient, which is defined by the size of the intersection between two sample sets over the size of their union. To quantify the statistical significance of subtype associations, we performed hypergeometric tests for overrepresentation of samples classified to one subtype in another. The resulting *P* values were adjusted for multiple hypotheses testing using the Benjamini–Hochberg (BH) method. Using this approach, we built a network consisting of the total 27 subtypes defined in the six different subtyping systems, interconnected by 96 significant (BH-corrected *P* value <0.001) edges.

4.1. Identification of consensus subtypes—To identify consensus groups from the network of subtype association, we used a consensus clustering approach involving the following steps:

- **a.** Network construction. Using the approach described above, 80% patient samples are randomly selected to generate a network of subtype association.
- **b.** Network clustering. The network generated is partitioned into clusters using MCL (Markov cluster algorithm)[11,12], which is a scalable and efficient unsupervised cluster algorithm for networks.
- **c.** Cluster evaluation. Steps (1) and (2) are repeated for $n = 1000$ times. From all clustering results, we calculated a 27×27 consensus matrix, defined by the frequency that each pair of subtypes is partitioned into the same cluster. Based on the consensus matrix, we assessed the robustness of each subtype with a stability score, which is the average frequency that its within-cluster association with other subtypes is the same as predicted by MCL on the network generated with all

samples. For evaluation of clustering performance, we employed weighted Silhouette width (R package *WeightedCluster*), which extends Silhouette width by giving more weights to subtypes that are more representative of their assigned clusters. Here, we used stability scores as weights to calculate weighted Silhouette width and took the median over all subtypes as a measure of clustering performance, which was used to evaluate the optimal number of clusters.

It should be noted that during network clustering, network granularity is controlled by inflation factor *f* in MCL, which is associated with the number of clusters k. No network substructure is recognized by MCL with $f < 1.6$, while $f > 10$ MCL does not provide any conceivable clustering. Therefore, we enumerated f from 1.6 to 10 and performed the three steps described above to compare their clustering performances. We selected as the optimal $f = 3.8$, which gives the highest median weighted Silhouette width (Supplementary Fig. 1), and generated four consensus molecular subtypes (CMS) using MCL. Representative consensus matrices illustrating robustness of clustering based on $f = 1.6$, 3.8 and 10, resulted in 3, 4 and 5 clusters, respectively, are shown in heatmaps ordered by identified CMS groups (Supplementary Fig. 1).

4.2. Identification of core consensus samples—For each CRC sample ($n = 3,962$), we performed a hypergeometric test for overrepresentation of assigned subtypes in the set of subtypes associated with each CMS. The CRC sample is assigned to a CMS if corresponding overrepresentation test is significant (P value <0.05). Using this strategy, 78% of the samples are identified to be highly representative of that particular consensus subtype and are considered core *consensus* samples. These samples have been taken to train a classifier using a Random Forest algorithm to apply the consensus classification to the non-consensus samples (details in “6. Classification” section). The distribution of unlabeled samples per dataset is shown in Supplementary Figure 2.

5. Data aggregation

In order to construct the classifier described in the main article, the private (shared amongst the consortium members) and public gene expression data sets had to be aggregated into a single matrix. These data sets were generated on different platforms, in different labs and at different time points, and thus we expect strong batch effects that, if not addressed, prevent efficient merging. Moreover, not all genes are measured on all platforms, and those that are may be represented by different probes, which can give rise to inconsistent or even contradictory measurements and thus further highlights the need for careful data preprocessing before the merge. We devised an algorithm suited for this aggregation, which is explained step by step below. The complete workflow is illustrated in Supplementary Figure 13.

Detailed strategy:

- **a.** Remove outlier samples from each data set separately (see section “3. Gene expression data processing and normalization” for details)
- **b.** Create a collection of reference genes (G_{REF}): 5,000 genes with largest median absolute deviation (MAD) were selected among those that were measured by at

least one probeset in all data sets. Each of these genes was represented by the corresponding probeset with the largest MAD.

c. Select a reference data set (referred onwards as D_{REF}). In our case we chose the largest Affymetrix data set[8]. In this data set, each gene in G_{REF} was represented by the probeset with the largest MAD.

d. For each of the other data sets, we used a consistency criterion to select the probeset to represent each gene. First, for each probeset, we calculated the correlation between the expression of the probeset and the reference genes in the same data set. This gives, for each probeset, a correlation vector C of length $|G_{REF}|$.

e. To select the probeset that was used to represent a gene *g* in data set D, we computed the correlation (c) between the correlation vector C for each of the corresponding probesets and the correlation vector for gene *g* in D_{REF}.

f. To select the probeset that represented gene *g* in the reference data set D_{REF}, we chose the probeset with the highest correlation with most of the other data sets. Therefore, for each data set D, we selected the probeset in D_{REF}, which has the largest "correlation of correlations" value from "V" with the probesets in D. The probeset selected is the one chosen to represent *g* in D_{REF}.

g. For each other data set D, the probeset with the highest value of the "correlation of correlations" with the chosen probeset in D_{REF} was selected to represent *g*.

h. At this stage we had, for each data set an expression matrix with a number of rows equal to the number of genes that are measured in all data sets. We then merged all these matrices to form a new expression matrix containing all the samples.

i. We used ComBat[32] to remove the per data set "(batch) effect", adding MSI status as a covariate. For data that did not have MSI status, we imputed MSI status using the MSI signature score[42].

j. We filtered the aggregated data set further based on the quantile range and the correlations calculated in step "e". We kept genes for which the difference between the 0.95 and 0.05 expression quantiles exceeded 0.75 in all data sets, and when the correlation c exceeded 0.5 in all data sets.

6. Consensus Molecular Subtype Classifier (Random Forest)

Using the aggregated gene expression data set, we developed a multi-class classifier to predict CMS subtypes in new samples. To train and validate our classifier, we used the core *consensus samples* (n = 3,104), i.e. those samples that are strongly representative of each of the CMS subtypes. We trained and validated our models using the aggregated data set (see "5. Data aggregation" section), which includes 5,972 genes that were observed to have gene level consistency as measured by correlation and variance across the multiple data sets in this study.

To train the classifier(s), we used the Random Forest (RF) algorithm[43], a widely used machine learning method that operates by generating multiple bootstrapped versions of the training data, and fitting a decision tree to each of these bootstraps. The final classifier is

then an ensemble of each of these decision trees. The RF algorithm has been well studied in the context of gene expression classifiers as it performs well with highly correlated, high-dimensional data, and is less prone to overfitting due to the averaging effect across many models[44]. Although the CMS subtypes do not occur with equal proportions, we trained our classifier using a *balanced* model approach, i.e. our model does not make *a priori* based assumptions about the frequency of each subtype. Therefore, for each iteration of the RF bootstrap, we randomly sample from each subtype in equal proportions. We parameterized the *forest* to have 500 trees with an average of 70 nodes per tree.

6.1. Global classifier—To assess feasibility of developing a CMS classifier, we randomly split our aggregated gene expression data matrix into ⅔ training and ⅓ validation using the core *consensus* samples from all data sets. After model training, we applied the classifier to the validation samples and computed performance metrics (sensitivity, specificity, and balanced accuracy) for each CMS (Supplementary Table 4) and per data set. While overall performance was robust (Supplementary Fig. 3a), we observed that the 4 data sets utilizing the Agilent platform had significantly lower performance metrics (Supplementary Fig. 3b).

6.2. Affymetrix (and RNAseq) classifier—We repeated the above procedure using only the core consensus samples profiled on the Affymetrix and RNAseq platforms (n = 2,688). Overall performance metrics improved compared to the global classifier (Supplementary Fig. 3c,d).

6.3. Agilent classifier—We repeated the above procedure using core consensus samples profiled on the Agilent platform (n = 416). Performance metrics were improved relative to the Agilent metrics from the global classifier. However, overall performance was below the Affymetrix/RNAseq classifier (Supplementary Fig. 3e,f). Given the smaller number of samples available to train this model, the lower performance is not unexpected.

6.4. Data set splits—The previous classifiers were developed by randomly sampling from all data sets, and partitioning into training and validation sets. To evaluate classifier performance across data sets (i.e. training in one set of data sets, and validating in an independent set of data sets), we performed two independent experiments. The first experiment utilized the GSE39582 (Affymetrix, fresh-frozen, n = 466), the TCGA (RNAseq, n = 459), and GSE17536 (Affymetrix, n = 147) data sets for model validation. Results are shown in Supplementary Figure 3g. In this experiment, no RNAseq data was used in training of the classifier and yet we observed that balanced accuracy in all CMS groups was >0.9 and comparable to the Affymetrix data sets. Overall, we observed robust performance metrics in these validation data sets.

Our second data split experiment was to separate the PETACC3 (n = 526) data set for validation, composed of Formalin Fixed Paraffin Embedded (FFPE) samples. This experiment allowed performance assessment of a fresh-frozen model applied to FFPE samples. Results are shown in Supplementary Figure 3h. In general, performance metrics were robust with the exception of CMS3. Notably, sensitivity/specificity for CMS3 was 0.70/.98. The high type II error rate in CMS3 suggests some biological differences between

FFPE and fresh-frozen samples, and underscores the importance of utilizing FFPE samples for training a classifier in this context.

6.5. Classification of non-consensus samples—We developed final classifiers separately for the Agilent and the Affymetrix/RNAseq data sets using all core *consensus samples* for model training. We then applied the classifiers on the unlabeled (*non-consensus*) samples. Recognizing that the samples may not be robustly classifiable, we set a minimum threshold of a 0.5 posterior probability (output from the Random Forest model) to assign a sample to a CMS group (specificity analysis revealed this threshold choice to be conservative with few false positives, as seen in Supplementary Figure 4). Using this criterion, we were able to assign 279 samples (39% of the unlabeled Affymetrix/RNAseq samples) and 60 samples (40% of the unlabeled Agilent samples) to a single subtype.

A comparison of the major clinicopathological and molecular traits between the classified samples (combination of core consensus samples plus non-consensus samples with CMS label after Random Forest classifier) versus unclassified samples revealed no significant differences between these two groups (Supplementary Table 14). In addition, an intra-subtype comparison confirmed that the clinicopathological and molecular associations of the core consensus samples are recapitulated in the newly classified samples (Supplementary Table 15).

For the remaining unclassified samples (n = 519), we examined the presence of any pattern in the subtype probability scoring that would indicate which subtype pairs present a challenge for disambiguating. We observed a strong negative correlation between CMS1 and CMS2 (R=−0.60, P <1e–16) and CMS3 and CMS4 (R=−0.76, P < 1e–16) indicating that these pairs are more easily separable. Conversely, the near-zero correlation between CMS2 and CMS3 (R=−0.06) suggests that this pair may be the most challenging to disambiguate.

Using the aggregated gene expression data, we further examined the unclassified samples with PCA and sparse Bayesian factor analysis (sBFA). A plot of the first four PCs confirms that unclassified samples are not outliers, but are instead heavily concentrated in the regions between the CMS distributed samples (Supplementary Fig. 5a), corroborating the distribution of the *non-consensus* samples in Figure 2c. Next, we selected the most variable genes across samples using a standard deviation cut-off of one and fitted the factor analysis model to this dataset using Bayesian framework. By introducing sparsity in the feature space through priors, the sBFA improves clustering of samples and allows identification of a latent or “hidden” variable that may discriminate unclassified samples from the CMS samples[45,46]. The projected data in the three-dimensional latent space shows that the unclassified samples in red are not separate from the CMS classified samples in black (Supplementary Fig. 5b). These analyses suggest that many of these unclassified or mixed samples are not necessarily technical outliers or new (and yet undetected) subtypes but instead potential mixtures or indeterminate CMS subtypes.

We next clustered the posterior probabilities of these unclassified samples to examine any potential pattern of subtype mixtures. We observed distinctive patterns including two or

more subtypes (Supplementary Fig. 5c), with CMS2-CMS4 comprising over 23% of the unclassified samples, followed by CMS2–CMS3 mixed with 17% (Supplementary Fig. 5d).

7. Clinical and molecular correlative analyses

Samples and data sets with clinical and molecular annotation are described in Supplementary Table 3. The distribution of clinical and molecular data by the four consensus subtypes is shown in Supplementary Table 5. Data was generated by each independent group or TCGA and aggregated with standardization as described below. We performed non-parametric tests for comparisons of continuous values (Kruskal-Wallis) and discrete counts (Fisher's exact test). Samples from each CMS were compared with the remaining samples, after confirming similar variance of the groups been compared. *P* values were adjusted for multiple comparisons as detailed in each section. All correlative analyses were carried out using R statistical software version 3.1.1.

7.1. Mutation profile

- *KRAS, BRAF, PIK3CA, PTEN, APC, TP53* mutation detection: for sequencing platform details refer to individual groups publications. In summary, in data sets other than TCGA, targeted sequencing was performed (codons or specific variants in oncogenes - *KRAS, BRAF, PIK3CA* - and most frequently mutated exons in tumor suppressors - *PTEN, APC, TP53*). For TCGA samples, somatic mutations and indels called from exome sequencing of matched tumor and normal genome pairs were aggregated using mutation annotation format (MAF) files from Synapse TCGA Live data portal (doi:10.7303/syn300013; September 2014). Silent mutations were excluded.
- Other genes (exome level): available in TCGA data set, as described above.
- Hypermutation class: available in TCGA data set, defined based on whole exome mutation count distribution using the same threshold as in the original publication (>180 events per exome as hypermutated sample)[2].
- Mutation in cancer driver genes analysis: In TCGA samples, we identified non-silent somatic mutations and indels in a selected list of significantly mutated cancer drivers[47]. We performed a supervised analysis of mutations in these genes and consensus subtypes. A Fisher's exact test comparing prevalence of mutation events in all samples from each CMS and the remaining samples was conducted and the resulting *P* values were adjusted for multiple comparisons using Benjamini-Hochberg method. Results can be found in Supplementary Table 8. A clear pattern of over-enrichment of mutations in cancer drivers is seen in CMS1, with the exception of *APC* and *TP53*. *APC* mutations are significantly enriched in CMS2, as are *KRAS* mutations in CMS3.

7.2. Copy number events profile

- Arm level copy number changes were visualized by using the GISTIC scores and CMS labels with the UCSC cancer genome browser. Focal (gene-level) copy differences were compared between subtypes by first mapping the genomic

coordinates of the segmented means to single genes using the *GenomicRanges* Bioconductor package. For a selected list of significantly altered oncogenes or tumor suppressors according to TCGA, we performed a supervised analysis of copy number counts and consensus subtypes (n = 485). A Student's t-test between the copy mean of all samples within a CMS and the copy mean of the remaining samples was conducted and the resulting P values were adjusted for multiple comparisons using Benjamini-Hochberg method. Results can be found in Supplementary Table 6. In CMS2 samples, copy number counts were consistently higher in oncogenes and lower in tumor suppressors. The opposite trend is seen in CMS1 samples while CMS4 tumors displayed no significant enrichments for copy number events in candidate driver genes.

- Somatic copy number alterations (SCNA) count and class: available in TCGA data set. Whole genome copy number GISTIC scores were downloaded from the Firehose Broad website (http://gdac.broadinstitute.org/; Sept 2014). We counted GISTIC scores −2/−1/+1/+2 as events for SCNA estimation (<Q1 was considered low and ≥Q1 was considered high).
- High-level amplifications and homozygous deletions: for a targeted list of significantly altered oncogenes or tumor suppressors according to TCGA[2] (*MYC, HNF4A, CDK8, FGFR1, ERBB2, IGF2, PTEN, SMAD4, APC, TCF7L2*), high level amplification was defined as GISTIC scores +2 and homozygous deletion as GISTIC scores −2.

7.3. Microsatellite status—Microsatellite status was determined using either using a panel of five microsatellite loci from the Bethesda reference panel[48] or immunohistochemistry markers[49]. For consistency, only samples with high-level microsatellite instability were considered instable (MSI).

7.4. Methylation data analysis—For characterization of the four CMS groups with DNA methylation data, we used TCGA defined four DNA-methylation subgroups (CIMP-H, CIMP-L, cluster3 and cluster4) in their 27K subseries by unsupervised analysis (see Supplementary Table 1 in TCGA CRC[2]) and extended this analysis with an additional 450K data set as detailed below.

We downloaded Level3 β-values based on Illumina Infinium HumanMethylation450 Array platform. The data set consists of in total 301 tumors and 38 normal samples. We employed hierarchical clustering and PCA to assess if there is any potential non-biological batch effect with respect to tissue source site (TSS) and batch variables. The hierarchical clustering was performed based on the Ward's linkage algorithm, with dissimilarity scores calculated from 1-Pearson correlation coefficients. As shown in Supplementary Figure 6a, samples are well mixed among various tissue source sites and batches.

To determine CpG Island Methylator Phenotype (CIMP) status, we first reduced data to the probes present in the 27K version beadchip (n = 25,978 probes). We then applied the same filters (removing probes with any NA values and probes designed on X and Y chromosomes) and performed recursively partitioned mixture model (RPMM) clustering

approach on the 10% most variant probes across tumors based on standard deviations (n = 1,486; SD >0.18) using *RPMM* R/Bioconductor package (http://CRAN.R-project.org/package=RPMM) with default parameters. RPMM returned, as for the 27K subseries, four clusters. We then drew the heatmap of β-values as in the original article (using R packages *heatmap.plus* and *seriation*, Supplementary Fig. 6b). Considering the methylome patterns of the four subgroups from the 27K subseries, we could assign the cluster 1 to CIMP-H, the cluster 2 to CIMP-L and the other 2 clusters to cluster3/cluster4.

For differential methylation analysis, we used 187 tumor samples that have classification labels based on classification of the TCGA gene expression data. We first calculated the methylation level for each gene by taking the median β-value over all corresponding annotated probes. Next, we performed differential methylation analyses based on two sample t-tests, comparing each CMS with the other CMS groups. Out of the total 21,231 genes, we identified 1664 genes differentially methylated (Benjamini–Hochberg-corrected P value < .05 and |log2 fold change| >0.5) between at least one CMS and the others (heatmap shown in Supplementary Fig. 6c). As expected, most of the differentially methylated genes (n = 1,262) have significant higher methylation in CMS1 tumors, which is consistent with their CIMP-H status. Nonetheless, we also observed genes specifically hyper- or hypo-methylated in the other three CMS groups, suggesting subtype-specific epigenetic regulation of the identified four CMS groups (data not shown).

We also performed a combined CIMP status analysis with TCGA results added to the panel of five markers as previously described[50] available in other data sets (Supplementary Table 3). For consistency, in the combined analysis only samples with high level methylation were considered CIMP-high and the remaining were classified as CIMP negative. Results are described in Supplementary Table 5, with enrichment for CIMP-high in CMS1.

7.5. Integrative analysis—We performed integrative analysis in TCGA data set only, using the same strategy as described in the original TCGA publication[2] with regards to mutation, copy number and gene expression changes in targeted genes and pathways (Supplementary Fig. 7c). To summarize, for mutations only non-silent events were considered activating/inactivating alterations. For copy number events, only high-level amplifications or homozygous deletions were defined as alterations. In some cases, up- or down-regulation of gene expression was also considered a pathway alteration: *IGF2*, *FZD10*, *SMAD4* genes.

7.6. Pathway analysis—Genesets of interest were identified by the consortium and separated in five main groups, as detailed in Supplementary Table 9 and below:

- **a.** ESTIMATE algorithm: method that uses gene expression signatures to infer the fraction of stromal and immune cells in tumor samples[30];
- **b.** Curated signatures: upper and lower normal colon crypt compartments[51], epithelial and mesenchymal markers[7], WNT[52] and MYC downstream target[53], epithelial-mesenchymal transition core genes and TGFβ pathway[54], intestinal stem cells[55], matrix remodeling (REACTOME) and wound-response (GO BP);

c. Canonical genesets: MAPK and PI3K (GO BP), SRC, JAK-STAT, caspases (BIOCARTA), proteosome (KEGG), Notch, cell cycle, translation and ribosome, integrin beta3, VEGF/VEGFR interactions (REACTOME);

d. Immune activation: immune response (GO BP), PD1 activation (REACTOME), infiltration with T cytotoxic cells (CD8)[56] and T helper cells (T_H1) in cancer samples[57,58], infiltration with Natural Killer (NK) cells[59] and follicular helper T (TF_H) cells[60] in cancer samples, activation of T helper 17 (T_H17) cells[61], regulatory T cells (Treg)[62] or myeloid-derived suppressor cells (MDSC)[63];

e. Metabolic activation: sugar, amino acid, nucleotide, glucose, pentose, fructose, mannose, starch, sucrose, galactose, glutathione, nitrogen, tyrosine, glycerophospholipid, fatty acid, arachnoid acid, linoleic acid (KEGG), glutamine (GO BP), lysophospholipid (PID).

Gene symbols were mapped to Entrez IDs to determine overlap in each individual data set that was evaluated for geneset enrichment. Geneset enrichment was tested for each subtype as compared to all other subtypes using the GSA[64] method and was performed for each geneset by data set combination using two-class unpaired tests with 10,000 permutations. A single *P* value per geneset was computed - consolidated across data sets - using Fisher's combined probability test.

7.7. Proteomic analysis—For reverse-phase protein array (RPPA), normalized measurements of 187 proteins were downloaded from TCPA website (http://app1.bioinformatics.mdanderson.org/tcpa/, Sept 2014). We performed a supervised analysis of RPPA levels and consensus subtypes ($n = 382$). A Kruskal-Wallis test comparing median protein expression values in all samples from each CMS and the remaining samples was conducted and the resulting *P* values were adjusted for multiple comparisons using Benjamini-Hochberg method. Results can be found in Supplementary Table 7. We identified 145 protein features that were significantly associated (*P* value <0.05) with consensus subtypes. Of note, CMS1 samples had elevated expression of proteins involved in apoptosis (Caspase 7, Rad51), cell cycle (Cyclins D1, E1, E2) and DNA damage repair (Chk1), while CMS4 samples were mainly enriched for microenvironment proteins (Collagen, Fibronectin).

We also obtained liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic quantile-normalized and log-transformed data for 95 TCGA tumor samples[21]. Heatmap of top differentially expressed proteins in TCGA colored with a gradient from blue (low expression) to yellow (high expression) is shown in Figure 3g. Overall, 81 samples were assigned to one of the four CMS identified here. Geneset enrichment was tested for each subtype as compared to all other subtypes using the GSA[64] method, as described above. Results are summarized in Supplementary Table 10.

7.8. MicroRNA data analysis—For miRNA characterization of the four CMS groups, we used two independent data sets obtained from TCGA. Data set 1 includes Illumina GA sequencing data for 255 primary colorectal tumors, whereas data set 2 consists of Illumina HiSeq sequencing data for 241 primary colorectal tumors. For both data sets, we obtained

Level 3 RPM (reads per million miRNA mapped) data from the TCGA data portal. The RPM data were log2-transformed after adding 1 pseudocount for the following analyses.

It has been confirmed previously that data set 1 has no serious batch effect[2]. For data set 2, we examined potential non-biological batch effects with respect to tissue source site (TSS) and batch variables. For hierarchical clustering, the Ward's linkage algorithm was performed with dissimilarity scores calculated from 1-Pearson correlation coefficients. Overall, the hierarchical clustering results show that samples are well mixed among various tissue source sites and batches (Supplementary Fig. 8a).

For differential expression analysis, we first filtered out samples that do not have a CMS assigned due to lack of mRNA expression data availability. The filtering step resulted in 197 samples for data set 1 and 200 samples for data set 2. For each data set, we performed differential expression analyses based on two sample *t*-tests, comparing each CMS with the other CMS groups. A high Pearson correlation coefficient was observed in the log2 fold change between data set 1 and 2 for each CMS (Supplementary Fig. 8b), suggesting a high concordance between the two independent data sets. In both data sets 110 miRNAs are differentially expressed (Benjamini–Hochberg-corrected P value < 0.05 and |log2 fold change| > 0.5) between at least one CMS and the others.

Differentially expressed miRNAs between CMSs were illustrated in a heatmap (Supplementary Fig. 8c). CMS2 can be characterized by the up-regulated mir-17-92 cluster, which is known to be bound and regulated by MYC[22]. The upregulation of the mir-17-92 cluster is consistent with the fact that MYC signaling is promoted in CMS2. Out of the total six miRNAs down-regulated in CMS3, hsa-mir-143 and four miRNAs belonging to the let-7 family are known to bind and regulate the expression of RAS[65,66]. The five miRNAs can be used for characterizing CMS3, which is featured with more activated RAS and MAPK signaling. CMS4 is enriched for downregulated miRNAs (e.g., hsa-mir-148a, the miR-192 and miR-200 families) that are known for tumor suppression. The miR-200 and miR-192 families regulate epithelial mesencgymal transition (EMT) pathway by targeting ZEB1 and/or ZEB2[23,67], whereas hsa-mir-148a is predicted by TargetScan[68] to regulate MMP13 and TGFB2, which are important for matrix remodeling and TGFβ pathways. Taken together, the downregulation of miRNAs associated with suppression of the EMT, MR and TGFβ associated signatures could explain why CMS4 is more aggressive and metastatic than the other CMSs.

7.9. Clinical and pathological variables—Data from different data sets was standardized as described below:

- Site: right colon (cecum, ascending, hepatic flexure and transverse colon); left (splenic flexure, descending and sigmoid colon); and rectum (Supplementary Fig. 10)
- Stage: assignments were defined using the latest edition of AJCC Cancer Staging Manual available at the time of diagnosis (3rd – 6th). For consistency, we only investigated the major stage (I, II, III or IV), whose definition does not change in these different staging systems.

- Grade: 1 (well differentiated), 2 (moderately differentiated), 3 (poorly differentiated) carcinomas, according to pathology review performed by each independent institution.

7.10. Tumor purity analysis—We obtained the tumor purity estimation of CRC samples in TCGA data set as defined by the ABSOLUTE algorithm[20] (doi:10.7303/syn1710466.2). As seen in Supplementary Figure 7d and Supplementary Table 14, classified and unclassified samples did not have significant differences in tumor purity. We did observe reduced proportion of cancer cells (less tumor purity) in CMS4 samples, as shown in Supplementary Figure 7e and Supplementary Table 5. This finding is in line with the higher stromal and immune infiltration scores in CMS4 samples as per ESTIMATE algorithm[27] (Fig. 3i).

7.11 Tumor vs normal analysis—We assessed the distribution of normal samples obtained from the GSE39582 (n = 19 normal) and PETACC-3 (n = 64 normal) data sets. The gene expression data from each cohort was re-normalized (see previous description of data normalization) including normal samples. PCA was then applied to each data set, and expectedly, tumor samples were clearly differentiable from normal samples using the top two PCs (Supplementary Fig. 9a,c). We next interrogated which of the CMS groups were more ‘normal’-like. We trained a Support Vector Machine to find the optimal hyperplane separating tumor vs normal, and then computed the distance from all tumor samples to the hyperplane. Overall distance distributions by CMS groups are depicted in Supplementary Figure 9b,c.

7.12. Survival analyses—Overall survival (OS) and relapse-free survival (RFS) times were calculated based on dates of cancer diagnosis or time of surgery, death due to any cause and disease relapse. For RFS analysis, patients that died without a relapse event were censored at the time of death. Relapse event was defined as clinical or radiological evidence of disease recurrence. Survival after relapse (SAR) was defined as time from relapse until death due to any cause. Data were censored based upon last known clinical follow-up and patients with less than 1 month of follow-up were excluded from all survival analyses. Supplementary Table 13 summarizes follow-up time, number of events, number of patients at risk and survival estimates for the entire population and patients assigned each CMS.

We performed Cox Proportional Hazards modeling in the aggregated data sets after confirming proportionality of hazards across patient cohorts. OS models included all stage I–IV patients while both RFS and SAR analyses were limited to patients with stage I, II or III at diagnosis. Both univariate and multivariate models were stratified by data set. We also performed univariate survival modeling separately in the subset of patients enrolled in the PETACC-3 study[10], as one can expect closer follow-up for relapse and death events in a clinical trial (Supplementary Fig. 11a). Detailed description of survival models can be found in Supplementary Table 13.

In order to evaluate the performance of survival models, we split the data sets into ⅔ and ⅓ for training and validation and computed the time-dependent area under the curve (tAUC), which measures the ability to distinguish the individuals who will experience a relapse or

death event. Results are summarized in Supplementary Table 13 and Supplementary Figure 11b. Indeed, when the CMS classification was added to multivariate clinico-molecular survival models, we still observe a significant discriminative contribution by the CMS subtypes in predicting outcome.

All survival analyses were carried out using *survival* and *survAUC* packages for R statistical software version 3.1.1[69]. We calculated log-rank *P* values in survival models and compared multivariate models with and without CMS classification by performing ANOVA. Paired Student's t-test was used to compare tAUCs estimates.

8. Data, code sharing, and *CMSclassifier* R package (Random Forest and Single Sample Predictor)

As a resource for the community, for all public data sets used in the consortium we have provided normalized gene expression data, CMS subtyping calls, and sample annotation for download through the Synapse platform (doi:10.7303/syn2623706). Additionally, scripts and code for the Random Forest CMS classifier are available for download: (https://github.com/Sage-Bionetworks/crcsc).

We also provide a downloadable R package (*CMSclassifier*) which includes the Random Forest classifier described previously, as well as a "Single Sample Predictor" (SSP) classifier. By definition a SSP makes possible to predict a unique sample, and its output considering any given sample has to remain constant whether it is predicted alone or within a series of samples. A typical requirement of SSP is that they cannot be based on (intra-series) row-centered data, because row-centering is impacted by the composition of the series. Here the proposed SSP is multi-platform (RNA seq / single color microarray / two colors microarray) and as such doesn't include any normalization procedure (such procedures are platform dependent), meaning that the user has to provide normalized data, with a normalization procedure respecting the Single Sample 'spirit' (such as single sample frozen RMA for Affymetrix microarrays, for example). Of note, the SSP reported here can be used on row-centered data with satisfactory results in most situations, however in such a case it cannot be any more seen as a Single Sample Predictor.

The SSP is implemented in the *CMSclassifier* R package. It is based on a similarity-to-centroid approach, with Pearson coefficient as similarity measure. It uses centroids of the CMS calculated for 693 discriminant genes (Entrez Ids), which were selected using the GSE39582 series based on AUC and fold change criterion. The CMS centroids were obtained for five series (TCGA COAD 'RNASeq V2 GA', TCGA COAD 'RNASeq V2 HiSeq', TCGA COAD 'Agilent', GSE39582, EMTAB990), yielding 20 centroids $C_{i,j}$ (i: CMS 1..4; j: series 1..5). To classify a given CRC sample, the SSP first calculates the similarity $S_{i,j}$ of the CRC sample expression profile (for the 693 discriminant genes) to the 20 centroids. The minimal similarity S_i to each CMS in the 5 series is then reported ($S_i = \mathrm{Min}_{j=1..5}\, S_{i,j}$). Then the nearest CMS i* is reported ($S_{i*} = \mathrm{Max}_{i=1..4}\, S_i$). The similarity difference D between the two nearest CMS is also reported (D = Si* − Si', with i' being the second nearest CMS). Then if both S_{i*} is above 0.15 and D is above 0.06 the sample is classified in CMS i*, otherwise its label is said undetermined.

The performance metrics of Random Forest and SSP classifiers using the consensus network class as “gold-standard” (*n* = 3,104 samples) is summarized in Supplementary Table 16.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the goodwill and generosity of the colorectal research community who made this study possible. J.G. and S.H.F. are supported by the Integrative Cancer Biology Program of the National Cancer Institute (U54CA149237). R.D. is supported by “La Caixa International Program for Cancer Research & Education”. L.V. is supported by grants from the Dutch Cancer Society (UVA2011-4969, UVA2014-7245), Worldwide Cancer Research (14-1164), the Maag Lever Darm Stichting (MLDS-CDG 14-03) and the European Research Council (ERG-StG 638193). J.P.M. is supported by grants from the Dutch Cancer Society UVA2012-573, UVA2013-6331 and UVA2015-7587, MLDS grant FP012 and NWO gravitation. S.K. is supported by NIH (R01CA172670, R01CA184843, R01 CA187238) and P30CA016672 (Biostatistic and Bioinformatic Core). A.S. and G.N. acknowledge support from the NHS. ST is supported by the KULeuven GOA/12/2106 grant, the EU FP7 Coltheres grant, the Research Foundation Flanders-FWO and the Belgian National Cancer Plan.

References

1. Hoadley KA, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell. 2014; 158:929–944. [PubMed: 25109877]
2. Muzny DM, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. [PubMed: 22810696]
3. Roepman P, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int. J. Cancer. 2014; 134:552–562. [PubMed: 23852808]
4. Budinska E, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J. Pathol. 2013; 231:63–76. [PubMed: 23836465]
5. Schlicker A, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med. Genomics. 2012; 5:66. [PubMed: 23272949]
6. Sadanandam A, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat. Med. 2013; 19:619–625. [PubMed: 23584089]
7. De Sousa E Melo F, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat. Med. 2013; 19:614–618. [PubMed: 23584090]
8. Marisa L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013; 10:e1001453. [PubMed: 23700391]
9. Perez-Villamil B, et al. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. BMC Cancer. 2012; 12:260. [PubMed: 22712570]
10. Van Cutsem E, et al. Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. J. Clin. Oncol. 2009; 27:3117–3125. [PubMed: 19451425]
11. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. SIAM J. Matrix Anal. Appl. 2008; 30:121–141.
12. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30:1575–1584. [PubMed: 11917018]
13. Llosa NJ, et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. Cancer Discov. 2015; 5:43–51. [PubMed: 25358689]

14. Brunelli L, Caiola E, Marabese M, Broggini M, Pastorelli R. Capturing the metabolomic diversity of KRAS mutants in non-small-cell lung cancer cells. Oncotarget. 2014; 5:4722–4731. [PubMed: 24952473]
15. Son J, et al. Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. Nature. 2013; 496:101–105. [PubMed: 23535601]
16. Kamphorst JJ, et al. Hypoxic and Ras-transformed cells support growth by scavenging unsaturated fatty acids from lysophospholipids. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:8882–8887. [PubMed: 23671091]
17. Ying H, et al. Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. Cell. 2012; 149:656–670. [PubMed: 22541435]
18. Lei Z, et al. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology. 2013; 145:554–565. [PubMed: 23684942]
19. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014; 513:202–209. [PubMed: 25079317]
20. Carter SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 2012; 30:413–421. [PubMed: 22544022]
21. Zhang, B., et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014. at <http://www.ncbi.nlm.nih.gov/pubmed/25043054>
22. Li Y, Choi PS, Casey SC, Dill DL, Felsher DW. MYC through miR-17-92 suppresses specific target genes to maintain survival, autonomous proliferation, and a neoplastic state. Cancer Cell. 2014; 26:262–272. [PubMed: 25117713]
23. Park S-M, Gaur AB, Lengyel E, Peter ME. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev. 2008; 22:894–907. [PubMed: 18381893]
24. Carmona FJ, et al. A Comprehensive DNA Methylation Profile of Epithelial-to-Mesenchymal Transition. Cancer Res. 2014; 74:5608–5619. [PubMed: 25106427]
25. Tran B, et al. Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. Cancer. 2011; 117:4623–4632. [PubMed: 21456008]
26. Gavin PG, et al. Mutation profiling and microsatellite instability in stage II and III colon cancer: an assessment of their prognostic and oxaliplatin predictive value. Clin. Cancer Res. 2012; 18:6531–6541. [PubMed: 23045248]
27. Popovici V, et al. Context-dependent interpretation of the prognostic value of BRAF and KRAS mutations in colorectal cancer. BMC Cancer. 2013; 13:439. [PubMed: 24073892]
28. Sinicrope FA, et al. Molecular markers identify subtypes of stage III colon cancer associated with patient outcomes. Gastroenterology. 2015; 148:88–99. [PubMed: 25305506]
29. Le DT, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N. Engl. J. Med. 2015; 372:2509–2520. [PubMed: 26028255]
30. Yoshihara K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. 2013; 4:2612. [PubMed: 24113773]

References

31. Derry JMJ, et al. Developing predictive molecular maps of human disease through community-based modeling. Nat. Genet. 2012; 44:127–130. [PubMed: 22281773]
32. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–127. [PubMed: 16632515]
33. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010; 11:242–253. [PubMed: 20097884]
34. Zilliox MJ, Irizarry RA. A gene expression bar code for microarray data. Nat. Methods. 2007; 4:911–913. [PubMed: 17906632]
35. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. 2001; 98:5116–5121. [PubMed: 11309499]

36. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U. S. A. 2002; 99:6567–6572. [PubMed: 12011421]
37. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A. 2004; 101:4164–4169. [PubMed: 15016911]
38. Marron JS, Todd MJ, Ahn J. Distance-Weighted Discrimination. J. Am. Stat. Assoc. 2007; 102:1267–1271.
39. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004; 20:307–315. [PubMed: 14960456]
40. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]
41. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics. 2009; 25:415–416. [PubMed: 19106121]
42. Tian S, et al. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. J. Pathol. 2012
43. Breiman L. Random forest. Mach. Learn. 2001; 45:5–32.
44. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012; 99:323–329. [PubMed: 22546560]
45. Murray JS, Dunson DB, Carin L, Lucas JE. Bayesian Gaussian Copula Factor Models for Mixed Data. J. Am. Stat. Assoc. 2013; 108:656–665. [PubMed: 23990691]
46. Ghosh J, Dunson DB. Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis. J. Comput. Graph. Stat. 2009; 18:306–320. [PubMed: 23997568]
47. Tamborero D, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci. Rep. 2013; 3:2650. [PubMed: 24084849]
48. Umar A, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J. Natl. Cancer Inst. 2004; 96:261–268. [PubMed: 14970275]
49. Lindor NM, et al. Immunohistochemistry versus microsatellite instability testing in phenotyping colorectal tumors. J. Clin. Oncol. 2002; 20:1043–1048. [PubMed: 11844828]
50. Weisenberger DJ, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. 2006; 38:787–793. [PubMed: 16804544]
51. Kosinski C, et al. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proc. Natl. Acad. Sci. U. S. A. 2007; 104:15418–15423. [PubMed: 17881565]
52. Van der Flier LG, et al. The Intestinal Wnt/TCF Signature. Gastroenterology. 2007; 132:628–632. [PubMed: 17320548]
53. Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. Genome Biol. 2003; 4:R69. [PubMed: 14519204]
54. Loboda A, et al. EMT is the dominant program in human colon cancer. BMC Med. Genomics. 2011; 4:9. [PubMed: 21251323]
55. Merlos-Suárez A, et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. Cell Stem Cell. 2011; 8:511–524. [PubMed: 21419747]
56. Mlecnik B, et al. Biomolecular network reconstruction identifies T-cell homing factors associated with survival in colorectal cancer. Gastroenterology. 2010; 138:1429–1440. [PubMed: 19909745]
57. Tosolini M, et al. Clinical impact of different classes of infiltrating T cytotoxic and helper cells (Th1, th2, treg, th17) in patients with colorectal cancer. Cancer Res. 2011; 71:1263–1271. [PubMed: 21303976]
58. Galon J, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science. 2006; 313:1960–1964. [PubMed: 17008531]
59. Ascierto ML, et al. Molecular signatures mostly associated with NK cells are predictive of relapse free survival in breast cancer patients. J. Transl. Med. 2013; 11:145. [PubMed: 23758773]

60. Gu-Trantien C, et al. CD4+ follicular helper T cell infiltration predicts breast cancer survival. J. Clin. Invest. 2013; 123:2873–2892. [PubMed: 23778140]
61. Keerthivasan S, et al. β-Catenin promotes colitis and colon cancer through imprinting of proinflammatory properties in T cells. Sci. Transl. Med. 2014; 6:225ra28.
62. Stockis J, et al. Comparison of stable human Treg and Th clones by transcriptional profiling. Eur. J. Immunol. 2009; 39:869–882. [PubMed: 19224638]
63. Fridlender ZG, et al. Transcriptomic analysis comparing tumor-associated neutrophils with granulocytic myeloid-derived suppressor cells and normal neutrophils. PLoS One. 2012; 7:e31524. [PubMed: 22348096]
64. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann. Appl. Stat. 2007; 1:107–129.
65. Wang L, et al. MiR-143 acts as a tumor suppressor by targeting N-RAS and enhances temozolomide-induced apoptosis in glioma. Oncotarget. 2014; 5:5416–5427. [PubMed: 24980823]
66. Johnson SM, et al. RAS is regulated by the let-7 microRNA family. Cell. 2005; 120:635–647. [PubMed: 15766527]
67. Kim T, et al. p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. J. Exp. Med. 2011; 208:875–883. [PubMed: 21518799]
68. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120:15–20. [PubMed: 15652477]
69. Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. Bioconductor. 2004; 5:R80.

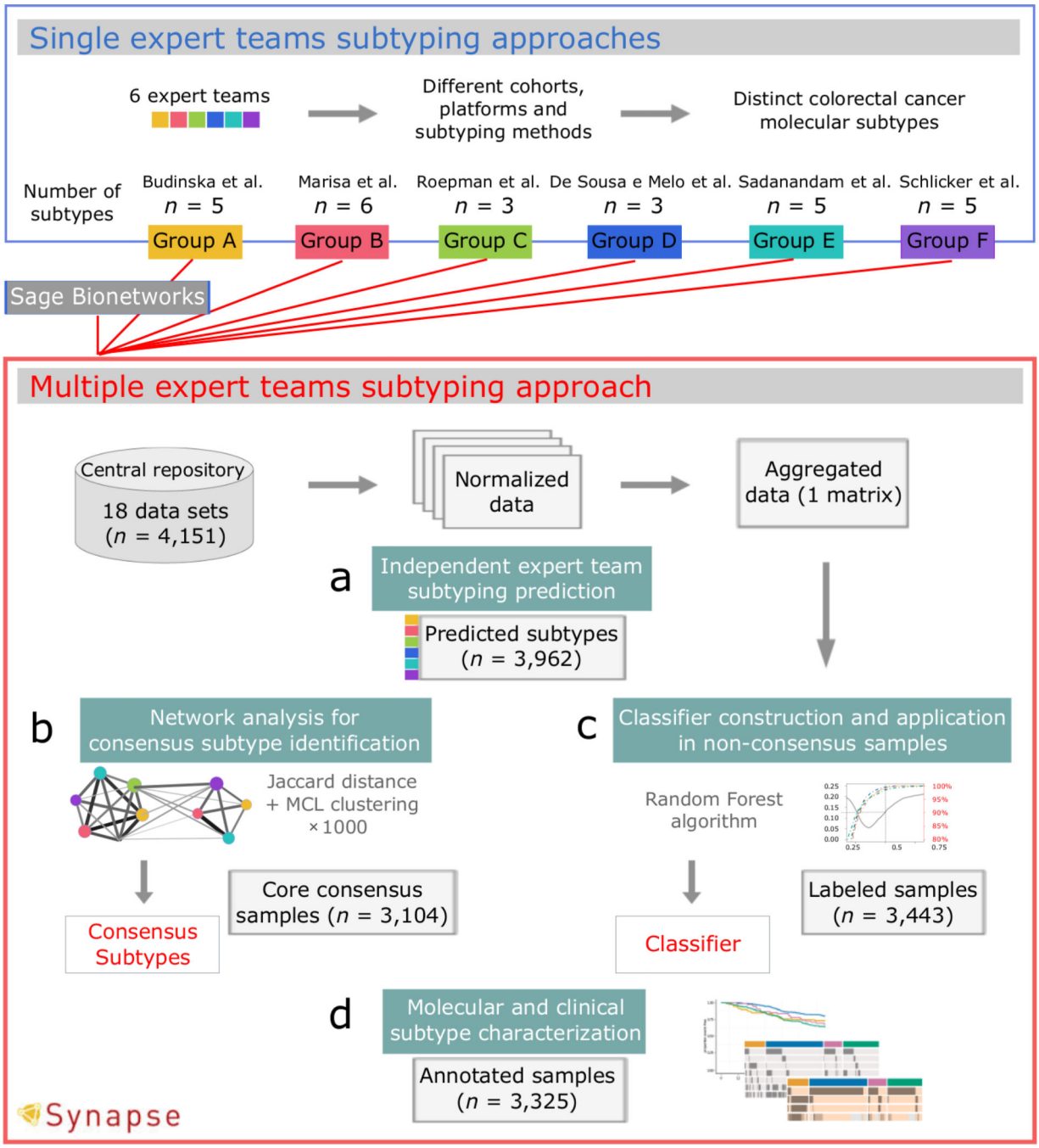

Figure 1. Analytical workflow of the Colorectal Cancer Subtyping Consortium

(**a**) Subtype classification on 18 shared data sets across six groups. (**b**) Concordance analysis of the six subtyping platforms, and application of a network analytical method to identify consensus subtype cluster. (**c**) Development of a consensus subtype classifier from an aggregated gene expression data set and the consensus subtype labels. (**d**) Biological and clinical characterization of the consensus subtypes.

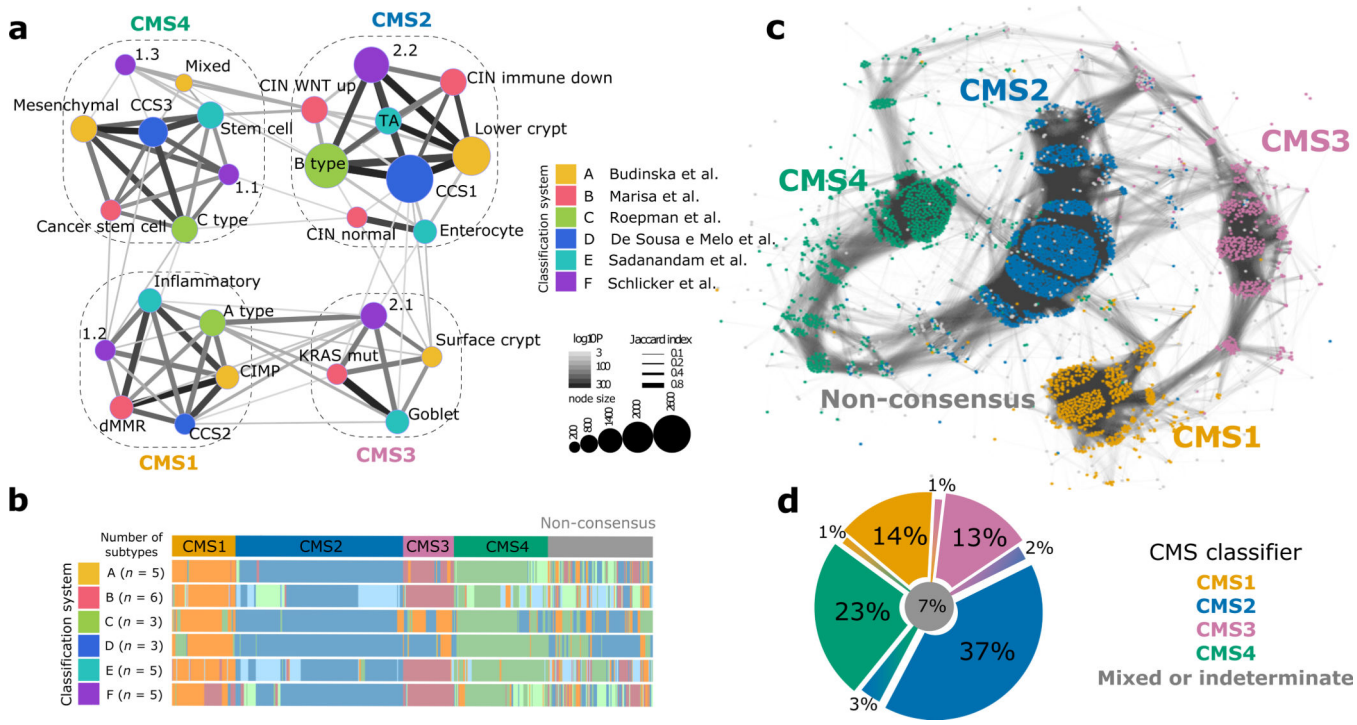

Figure 2. Identification of the consensus subtypes of colorectal cancer and application of classification framework in non-consensus samples

(**a**) Network of CRC subtypes across six classification systems: each node corresponds to a single subtype (colored according to group) and edge width corresponds to Jaccard similarity coefficient. The four primary clusters – identified from the Markov cluster algorithm – are highlighted and correspond to the four CMS groups. (**b**) Per sample distribution of each of the six CRC subtyping systems (A–F), grouped by the four consensus subtyping clusters (n = 3,104), and the unlabeled non-consensus set of samples (n = 858). Colors within each row represent a different subtype. (**c**) Patient network: each node represents a single patient sample (n = 3,962). Network edges correspond to highly concordant (5/6 of 6) subtyping calls between samples. Nodes are colored according to their CMS, with non-consensus samples gray. (**d**) Final distribution of the CMS1–4 groups (solid colors), ‘mixed’ samples (gradient colors) or indeterminate samples (gray color) as per classification framework.

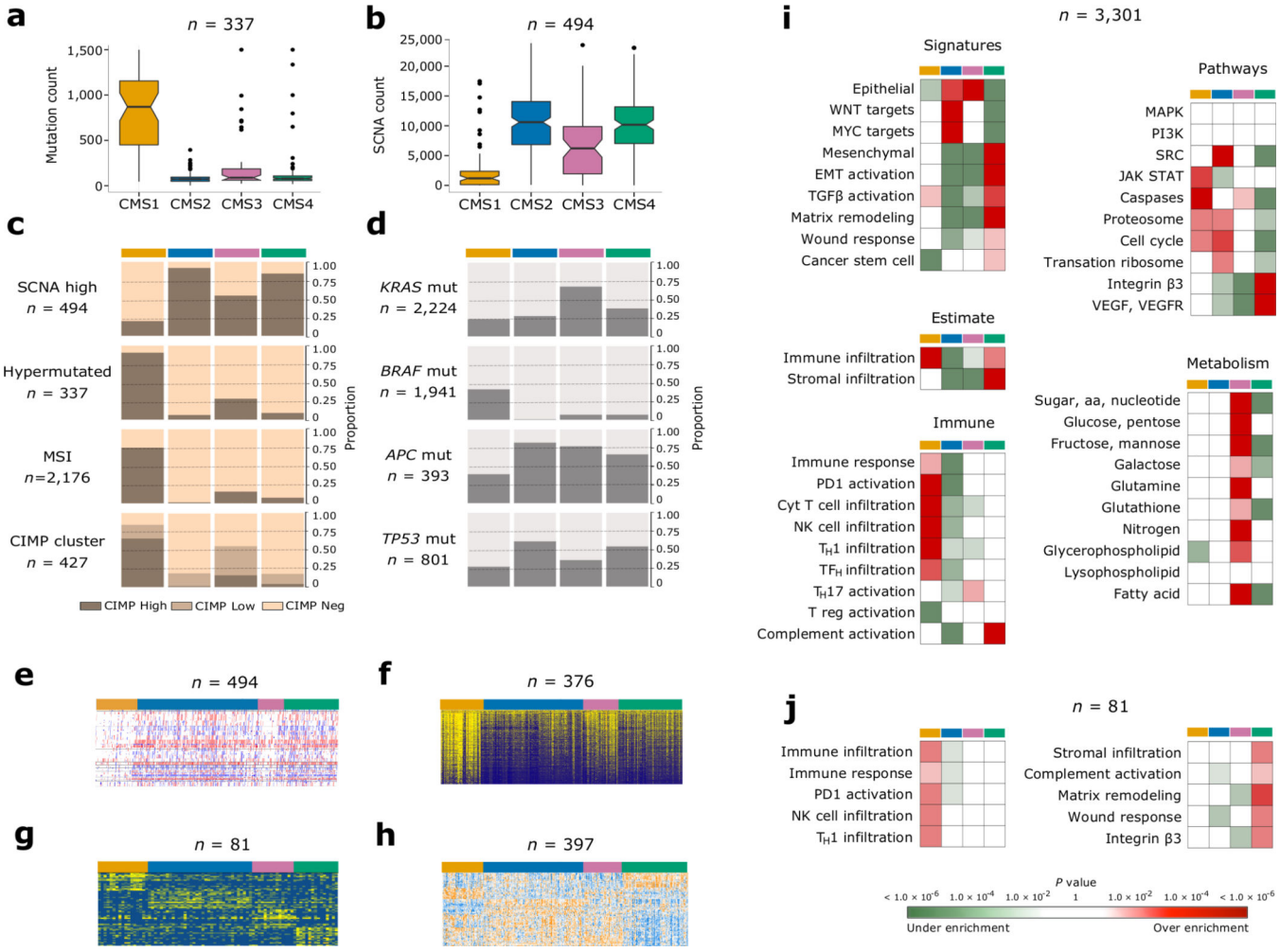

Figure 3. Molecular associations of consensus molecular subtype groups

(**a**) Distribution of non–synonymous somatic mutation events; and (**b**) somatic copy-number alterations (SCNAs), defined as non-zero GISTIC scores in TCGA data set, across consensus subtype samples (median, lower [Q1] and upper [Q3] quartiles, horizontal lines define minimum and maximum, dots define outliers). (**c**) Key genomic and epigenomic markers, with darker brown representing positivity for SCNA high (≥Q1 for non–zero GISTIC score events), hypermutation (≥180 events in exome sequencing), microsatellite instability (MSI) high or CpG Island Methylator Phenotype (CIMP) cluster high. (**d**) Mutation profile, with darker gray representing positivity for *KRAS*, *BRAF*, *APC* and *TP53* mutations. (**e**) Heatmap of copy number changes of the 22 autosomes, with shades of red for gains and blue for losses. CMS1 samples have fewer SCNAs and an intermediate pattern is seen in CMS3. (**f**) Heatmap representation of DNA methylation beta-values of most variable probes with yellow denoting high DNA methylation and blue low methylation. CMS1 samples show a distinguished hypermethylation profile and an intermediate pattern is seen in CMS3. (**g**) Heatmap of top differentially expressed proteins in TCGA colored with a gradient from blue (low expression) to yellow (high expression). (**h**) Heatmap of top differentially expressed microRNAs in TCGA with shades of blue for downregulation and

orange for upregulation. (**i**) Gene set mRNA enrichment analysis: signatures of special interest in CRC, ESTIMATE algorithm[30] to infer immune and stromal cell admixture in tumor samples, canonical pathways, immune signatures and metabolic pathways. (**j**) Gene set enrichment analysis of proteomic TCGA data. Detailed statistics in Supplementary Tables 5, 8, 9 and 11.

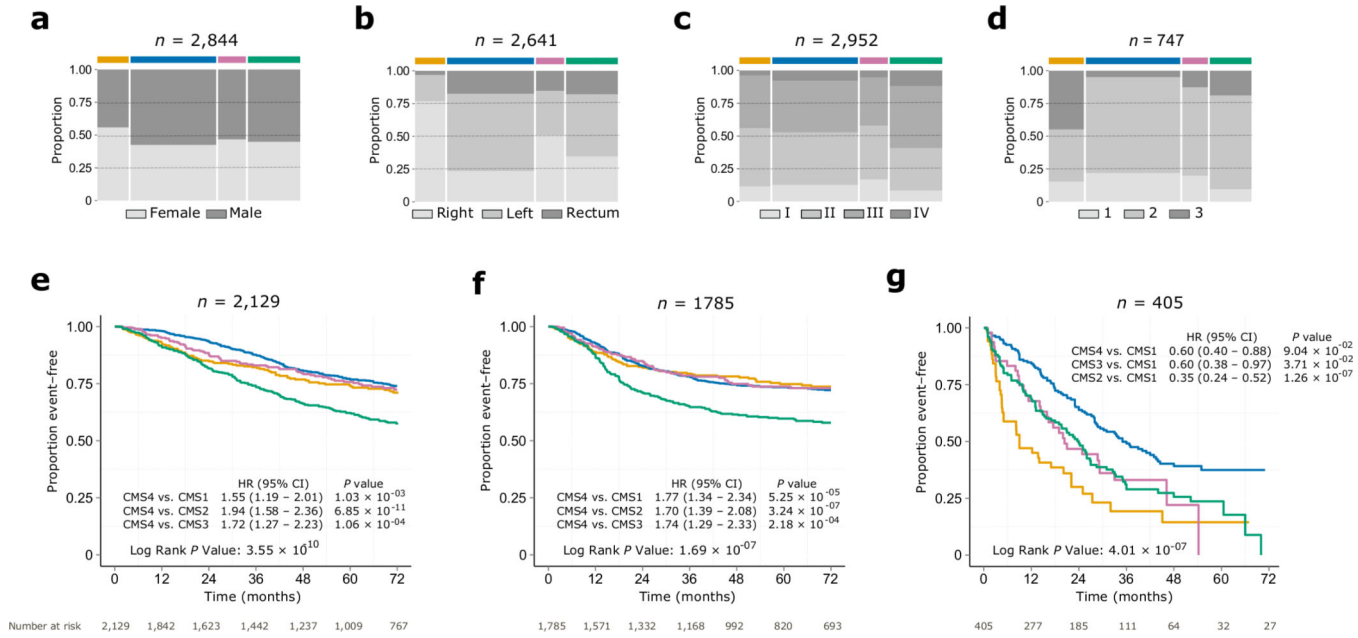

Figure 4. Clinicopathological and prognostic associations of consensus molecular subtype groups

(**a**) Distribution of gender; (**b**) Tumor site location; (**c**) Stage at diagnosis; and (**d**) Histopathological grade across consensus subtype samples. Prognostic value of CMS groups with Kaplan-Meier survival analysis in the aggregated cohort for (**e**) overall survival, (**f**) relapse-free survival and (**g**) survival after relapse with hazard ratios (HR) and 95% Confidence Interval (CI) for significant pairwise comparisons in univariate analyses (log-rank test). Numbers below the *x axes* represent patients at risk at selected time points. Detailed statistics in Supplementary Tables 5 and 13.

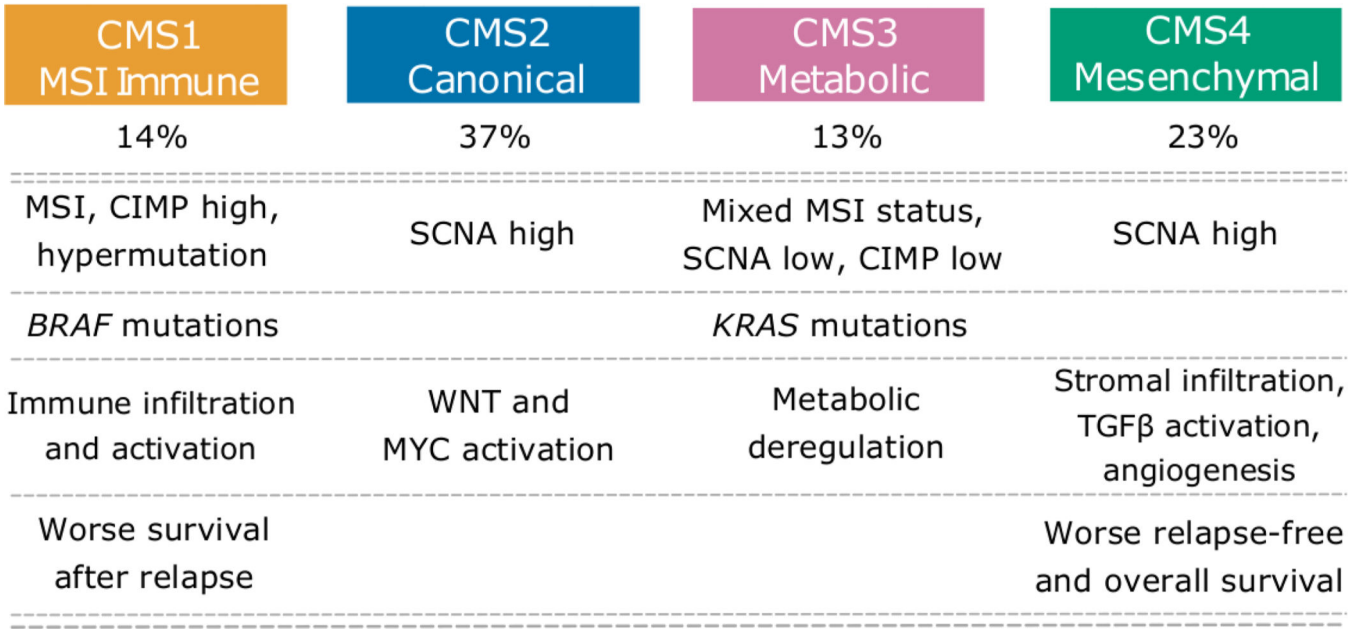

Figure 5. Proposed taxonomy of colorectal cancer reflecting significant biological differences in the gene expression-based molecular subtypes

CIMP, CpG Island Methylator Phenotype; MSI, microsatellite instability; SCNA, somatic copy number alterations; TGF, transforming growth factor.