nature medicine

The consensus molecular subtypes of colorectal cancer

Justin Guinney[1,21], Rodrigo Dienstmann[1,2,21], Xin Wang[3,4,21], Aurélien de Reyniès[5,21], Andreas Schlicker[6,21], Charlotte Soneson[7,21], Laetitia Marisa[5,21], Paul Roepman[8,21], Gift Nyamundanda[9,21], 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[16,17,22], Jan Paul Medema[3,22], Anguraj Sadanandam[9,22], Lodewyk Wessels[6,22], Mauro Delorenzi[7,18,19,22], Scott Kopetz[10,22], Louis Vermeulen[3,22] & Sabine Tejpar[20,22]

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 (CMSs) with distinguishing features: CMS1 (microsatellite instability immune, 14%), hypermutated, microsatellite unstable and strong immune activation; CMS2 (canonical, 37%), epithelial, marked WNT and MYC signaling activation; CMS3 (metabolic, 13%), epithelial and 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 intratumoral 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.

Gene expression–based subtyping is widely accepted as a relevant source of disease stratification[1]. Despite the technique's widespread use, its translational and clinical utility is hampered by discrepant results, which are probably 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 CRC.

Inspection of the published gene expression–based CRC classifications[2–9] revealed only superficial similarities. For example, all of the groups identified one tumor subtype enriched for microsatellite instability (MSI) and one subtype characterized by high expression of mesenchymal genes, but they 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 the 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 or 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 and 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 for other malignancies as well.

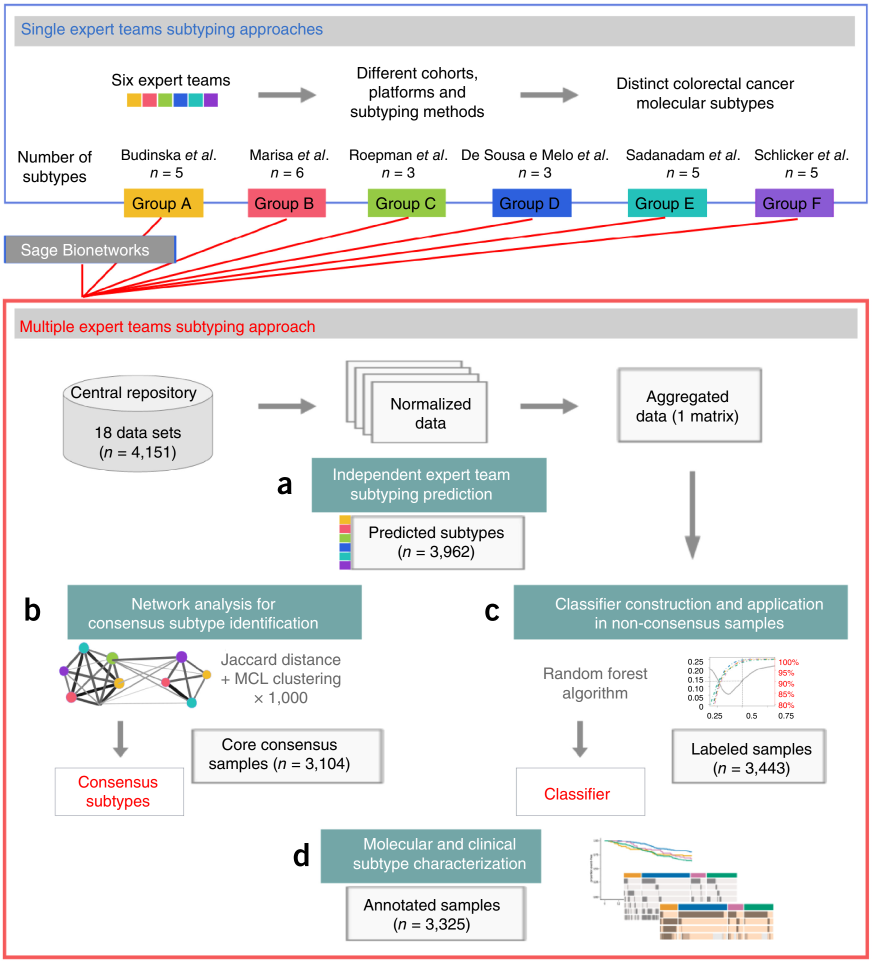
RESULTS

Comparison of published molecular subtyping platforms

We evaluated the results of six CRC subtyping algorithms[3–8], each developed independently using different gene expression data sets and analytical approaches (**Supplementary Tables 1** and **2**). **Figure 1** summarizes the workflow of our analysis. A total of 18 CRC data sets (n = 4,151 patients) from both public (GSE42284, GSE33113, GSE39582, GSE35896, GSE13067, GSE13294, GSE14333, GSE17536, GSE20916, GSE2109 and The Cancer Genome Atlas (TCGA)) and proprietary[3,10] sources (**Supplementary Table 3**)—which consisted of multiple gene expression platforms (Affymetrix, Agilent and RNA-sequencing), sample types (fresh-frozen samples and formalin-fixed paraffin-embedded (FFPE) samples) and study designs (retrospective and prospective series and one clinical trial[10])—were uniformly preprocessed and normalized from the raw formats to reduce technical variation. The six expert groups applied their

A full list of affiliations appears at the end of the paper.

Received 6 March; accepted 6 September; published online 12 October 2015; doi:10.1038/nm.3967



Single expert teams subtyping approaches
Six expert teams
Different cohorts, platforms and subtyping methods
Distinct colorectal cancer molecular subtypes
Number of subtypes
Budinska *et al.* n = 5
Marisa *et al.* n = 6
Roepman *et al.* n = 3
De Sousa e Melo *et al.* n = 3
Sadanadam *et al.* n = 5
Schlicker *et al.* n = 5
Group A
Group B
Group C
Group D
Group E
Group F
Sage Bionetworks
Multiple expert teams subtyping approach
Central repository
18 data sets (n = 4,151)
Normalized data
Aggregated data (1 matrix)
a
Independent expert team subtyping prediction
Predicted subtypes (n = 3,962)
b
Network analysis for consensus subtype identification
Jaccard distance + MCL clustering × 1,000
Core consensus samples (n = 3,104)
Consensus subtypes
c
Classifier construction and application in non-consensus samples
Random forest algorithm
0.25
0.20
0.15
0.10
0.05
0
0.25
0.5
0.75
100%
95%
90%
85%
80%
Labeled samples (n = 3,443)
Classifier
d
Molecular and clinical subtype characterization
Annotated samples (n = 3,325)

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 clusters. (**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.

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 the 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 (CMSs) 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**). Notably, 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 of the 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 and fresh-frozen), with a >90% balanced accuracy across all subtypes (**Supplementary Table 4** and **Supplementary Fig. 3**). This corroborates both the portability of the classifier and the evident subtype-specific signals across data sets.

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 **Figure 2d**, including the 'mixed or indeterminate' samples.

Biological characterization of the consensus molecular subtypes

We studied additional molecular data that were 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 (**Fig. 3**), CMS1 samples were hypermutated and had low prevalence of somatic copy number alterations (SCNAs) (**Fig. 3a–c,e** and **Supplementary Tables 5** and **6**). CMS1 encompassed the majority of MSI tumors and had overexpression of proteins involved in DNA damage repair, as determined by reverse-phase

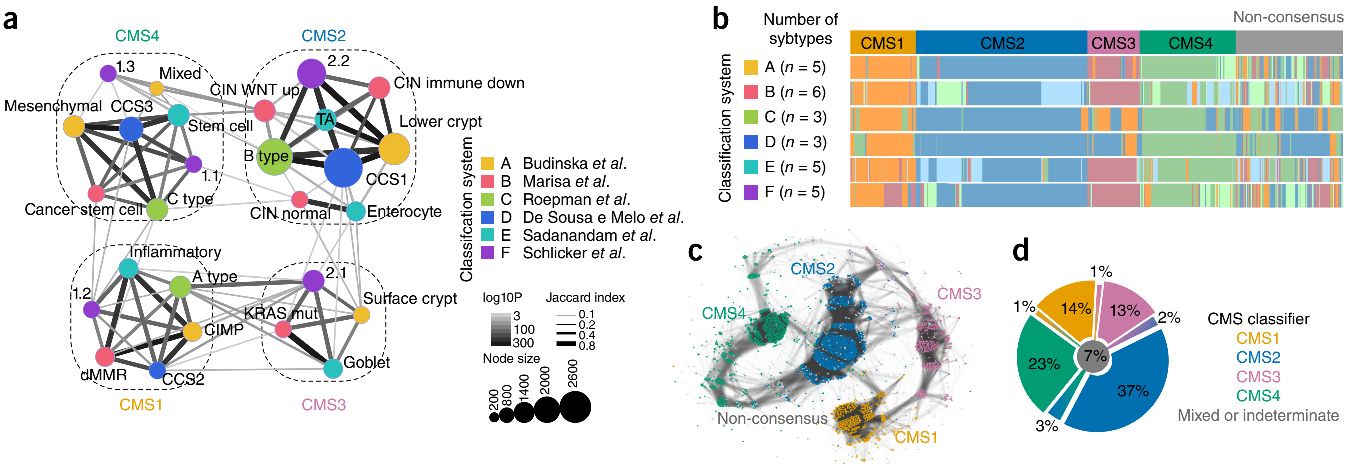


Figure 2 Identification of the consensus subtypes of colorectal cancer and application of classification framework in non-consensus samples. (**a**) Network of CRC subtypes across six classification systems: each node corresponds to a single subtype (colored according to group) and edge width corresponds to the 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. The *n* values shown in **b** correspond to the number of subtypes in the original independent classification published by each group. (**c**) Patient network: each node represents a single patient sample (*n* = 3,962). Network edges correspond to highly concordant (5 of 6) subtyping calls between samples. Nodes are colored according to their CMSs, with non-consensus samples in gray. (**d**) Final distribution of the CMS1–4 groups (solid colors), 'mixed' samples (gradient colors) and indeterminate samples (gray color) as per the classification framework.

protein array (RPPA) analysis, consistent with defective DNA mismatch repair (**Supplementary Table 7**). As expected, the analysis of methylation profiles in TCGA showed that CMS1 tumors display a widespread hypermethylation status (**Fig. 3f** and **Supplementary Fig. 6**). Conversely, CMS2–CMS4 displayed higher chromosomal instability (CIN), as measured by SCNA counts (**Fig. 3b** and **Supplementary Table 5**). We detected more frequent copy number gains in oncogenes and copy number 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 to the other CIN tumors: (i) there were consistently fewer SCNAs (**Fig. 3b**,**c**,**e** and **Supplementary Table 5**), an association not explained by differences in tumor purity (**Supplementary Fig. 7** and **Supplementary Table 5**); (ii) nearly 30% of the samples were hypermutated (**Fig. 3c** and **Supplementary Table 5**), which overlapped with MSI status (**Supplementary Fig. 7**); and (iii) there was a higher prevalence of CpG island methylator phenotype (CIMP)-low clusters in TCGA samples (**Fig. 3c** and **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** and **Supplementary Tables 5** and **8**), such as the frequent occurrence of *BRAF* mutations in CMS1 (in line with the known association of this event with MSI tumors[2]) and the overrepresentation of *KRAS* mutations in CMS3, none of the subtypes was defined by an individual event, and no genetic aberration was limited to a subtype. Similarly, we detected no unique and recurrent SCNA that strongly associated with a subtype, although amplifications of the gene encoding the transcription factor HNF4A were enriched in CMS2 (**Supplementary Tables 5** and **6**). Because 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** and **Supplementary Table 5**). This supports the notion that tumors harboring commonly assumed driver events in CRC still vary markedly 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** and **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** and **Supplementary Table 9**). CMS2 tumors displayed epithelial differentiation and strong upregulation of WNT and MYC downstream targets, both of which have classically been implicated in CRC carcinogenesis (**Fig. 3i** and **Supplementary Table 9**). In contrast, enrichment for multiple metabolism signatures was found in CMS3 epithelial CRCs, in line with the occurrence of *KRAS*-activating mutations that have been described as inducing prominent metabolic adaptation[14–17] (**Fig. 3i** and **Supplementary Table 9**). Of note, CMS3 tumors displayed similarities with a 'metabolic', genomically stable subtype that was recently described in gastric cancer[18,19]. Finally, CMS4 tumors showed clear upregulation of genes implicated in epithelial-to-mesenchymal transition (EMT) and of signatures associated with the activation of transforming growth factor (TGF)-β signaling, angiogenesis, matrix remodeling pathways and the complement-mediated inflammatory system (**Fig. 3i** and **Supplementary Table 9**). In addition, CMS4 samples exhibited a gene expression profile compatible with stromal infiltration (**Fig. 3i** and **Supplementary Table 9**), overexpression of extracellular matrix proteins using RPPA analysis (**Supplementary Table 7**) and higher admixture with non-cancer cells, as measured by the ABSOLUTE algorithm[20] (**Supplementary Fig. 7** and **Supplementary Table 5**).

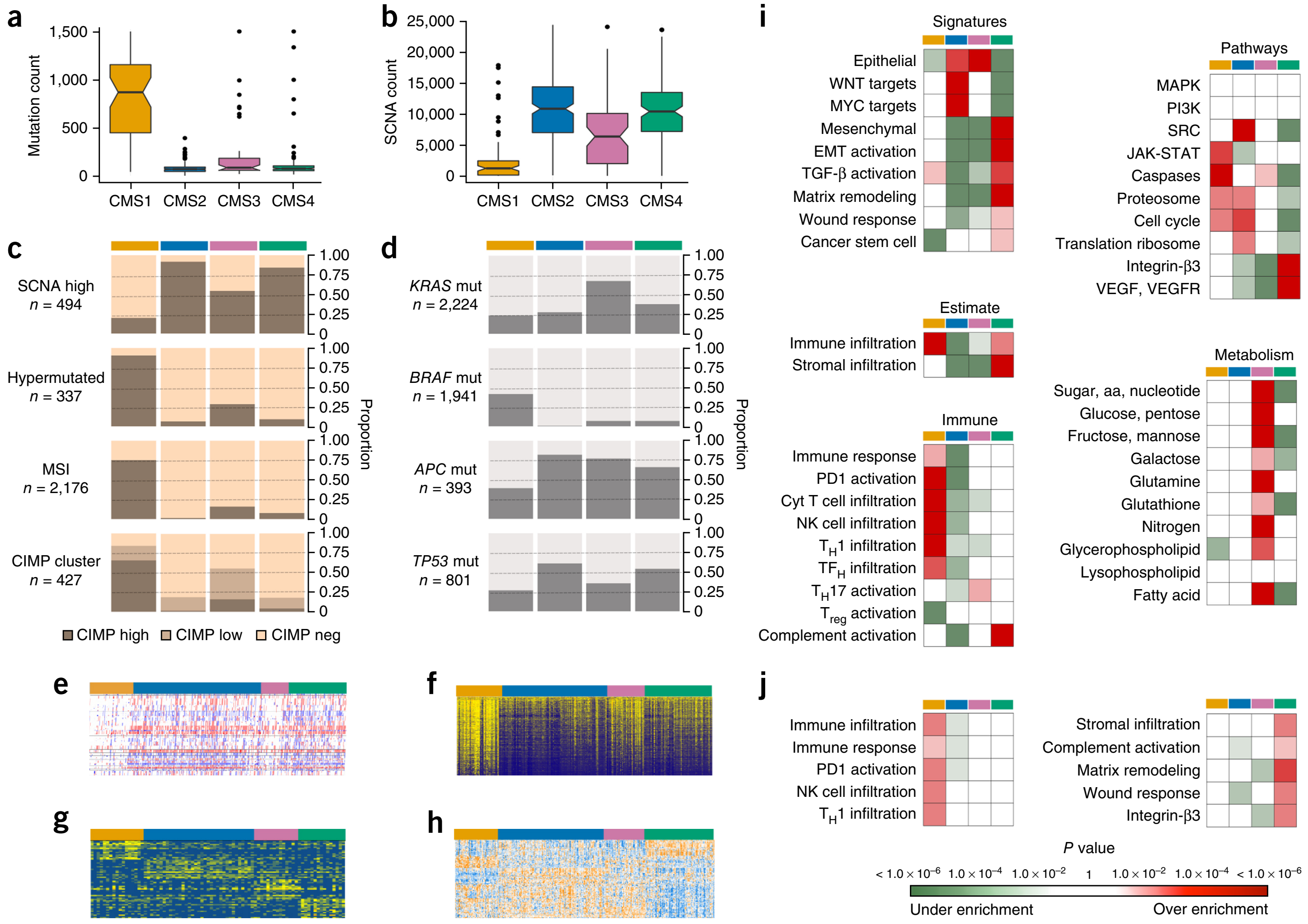


Figure 3 Molecular associations of consensus molecular subtype groups. (**a**,**b**) Distribution of nonsynonymous somatic mutation events ($n = 337$) (**a**) and SCNAs ($n = 494$; defined as nonzero GISTIC (genomic identification of significant targets in cancer) scores in the TCGA data set) (**b**), across consensus subtype samples (median, lower and upper 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; $n = 494$), hypermutation (≥180 events in exome sequencing; $n = 337$), MSI high ($n = 2,176$) or CIMP-cluster high ($n = 427$). (**d**) Mutation profile, with darker gray representing positivity for *KRAS* ($n = 2,224$), *BRAF* ($n = 1,941$), *APC* ($n = 393$) and *TP53* ($n = 801$) mutations. (**e**) Heat map of copy number changes of the 22 autosomes, with shades of red for gains and blue for losses ($n = 494$). CMS1 samples have fewer SCNAs, and CMS3 samples show an intermediate pattern. (**f**) Heat map representation of DNA methylation β-values (estimates of methylation levels using the ratio of intensities between methylated and unmethylated alleles) of most variable probes. Yellow denotes high DNA methylation; blue denotes low methylation ($n = 376$). CMS1 samples show a distinguished hypermethylation profile, and CMS3 samples show an intermediate pattern. (**g**) Heat map of top differentially expressed proteins in TCGA, colored with a gradient from blue (low expression) to yellow (high expression) ($n = 81$). (**h**) Heat map of top differentially expressed microRNAs in TCGA, with shades of blue for downregulation and orange for upregulation ($n = 397$). (**i**) Gene set mRNA enrichment analysis showing signatures of special interest in CRC, including canonical pathways, immune signatures, immune and stromal cell admixture in tumor samples (inferred by the ESTIMATE algorithm[30]) and metabolic pathways ($n = 3,301$). (**j**) Gene set enrichment analysis of proteomic TCGA data ($n = 81$). Detailed statistics are in **Supplementary Tables 5**,**8**,**9** and **11**.

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 ($n = 81$) (ref. 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, whereas CMS4 samples had significant overexpression of proteins implicated in stromal invasion, mesenchymal activation and complement pathways (**Fig. 3j** and **Supplementary Table 11**).

To interrogate post-transcriptional regulation of gene expression across CMS groups, we performed supervised microRNA (miR) analysis and identified significant subtype-specific miR regulation changes (**Fig. 3h**, **Supplementary Fig. 8** and **Supplementary Table 12**). Of particular note, CMS2 tumors showed upregulation of the miR-17–92 cluster (a direct transcriptional target of MYC[22]) and 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)[23,24] showed clear downregulation in CMS4.

Finally, we also compared gene expression patterns of CRC tumors with (i) adjacent normal colon tissue from patients with colon cancer ($n = 19$) and (ii) left colon (splenic flexure, descending and sigmoid colon) tissue from cancer-free individuals ($n = 64$) (Online Methods). Global principal component analysis (PCA) revealed that normal samples were clearly differentiated from tumor samples in both cohorts

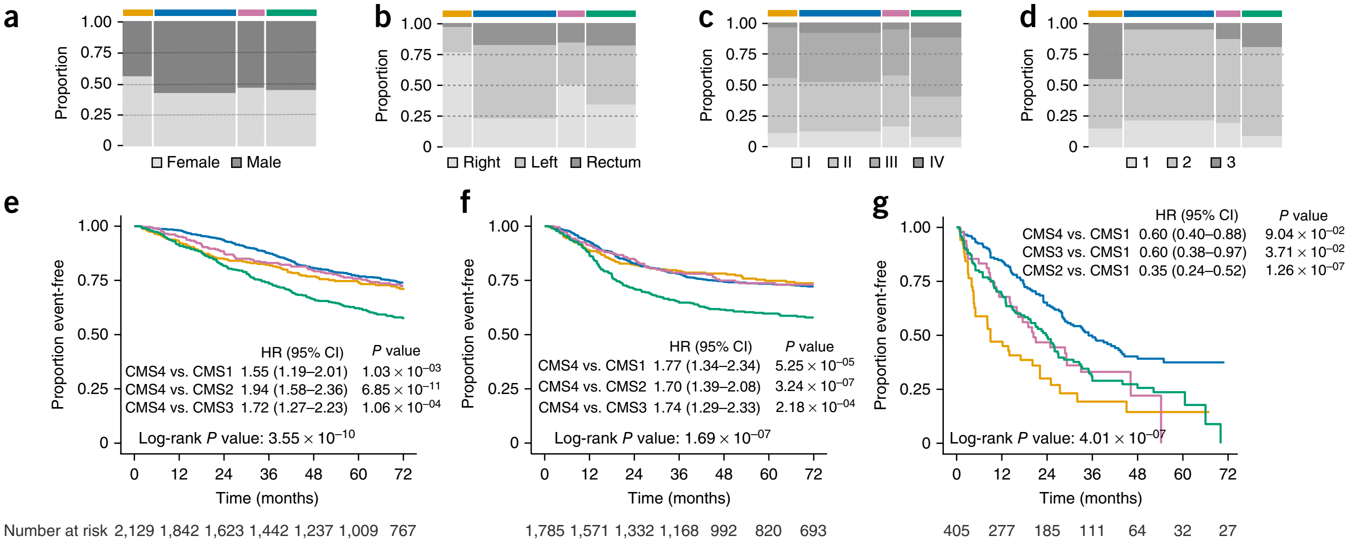


Figure 4 Clinicopathological and prognostic associations of consensus molecular subtype groups. (**a**–**d**) Distribution of gender (n = 2,844) (**a**), tumor site location (n = 2,641) (**b**), stage at diagnosis (n = 2,952) (**c**) and histopathological grade (n = 747) (**d**) across consensus subtype samples, represented by the colored bars CMS1, yellow; CMS2, blue; CMS3, pink; CMS4, green. (**e**–**g**) Prognostic value of CMS1 (yellow), CMS2 (blue), CMS3 (pink) and CMS4 (green) with Kaplan-Meier survival analysis in the aggregated cohort for overall survival (n = 2,129) (**e**), relapse-free survival (n = 1,785) (**f**) and survival after relapse (n = 405) (**g**). The hazard ratios (HR) and 95% confidence intervals (CI) for significant pairwise comparisons in univariate analyses (log-rank test) are displayed in each Kaplan-Meier plot. Numbers below the x axes represent the number of patients at risk at the selected time points. Detailed statistics are in **Supplementary Tables 5** and **13**.

(**Supplementary Fig. 9**). Although CMS3 tumors appeared more 'normal'-like at the gene expression level (**Supplementary Fig. 9**), we did not find greater contamination with non-cancer tissue in tumors of the CMS3 group as compared to tumors from the other consensus subtypes after pathological review of a subset of samples from the PETACC-3 clinical trial[10] as well as an assessment of ABSOLUTE tumor purity scores in TCGA data (**Supplementary Fig. 7** and **Supplementary Table 5**).

Clinical and prognostic associations of the consensus molecular subtypes

We also found important associations between the CMS groups and clinical variables (**Fig. 4** and **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** and **Supplementary Table 5**). Conversely, CMS2 tumors were mainly left-sided (**Fig. 4b**, **Supplementary Fig. 10** and **Supplementary Table 5**), whereas CMS4 tumors tended to be diagnosed at more advanced stages (III and IV) (**Fig. 4c** and **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 Fig. 11** and **Supplementary Table 13**). Irrespective of patient cohort, CMS4 tumors resulted in worse overall survival (**Fig. 4e**) and worse relapse-free survival (**Fig. 4f**) in both univariate and multivariate analyses, after adjustment for clinico-pathological features, MSI status and presence of *BRAF* or *KRAS* mutations (**Supplementary Table 13**). We also found superior survival rates after relapse in CMS2 patients (**Fig. 4g**), with a larger proportion of long-term survivors in this subset. Notably, the CMS1 population had a very poor survival rate after relapse (**Fig. 4g**), in agreement with recent studies showing worse prognosis of patients with MSI and *BRAF*-mutated CRCs 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 classification system for CRC. Thanks to the 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 analyses by members of 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**). We believe that this new taxonomy (CMS1 (MSI immune), CMS2 (canonical), CMS3 (metabolic) and CMS4 (mesenchymal)) will facilitate future research in the CRC field and should be adopted by the community for CRC stratification. From a biological perspective, we

CMS1 MSI immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high, hypermutation	SCNA high	Mixed MSI status, SCNA low, CIMP low	SCNA high
BRAF mutations		*KRAS* mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration, TGF-β activation, angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

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.

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, particularly for samples lacking a mesenchymal phenotype. From a clinical perspective, in CRC, as for many other 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; however, it is still unknown whether combinations of these features are needed for accurate prediction of prognosis or drug responses. In CRC, few biomarkers (including *RAS* and *BRAF* mutations and MSI and CIMP status) have been translated to patient care. It is important to emphasize that although the CMS groups are enriched for some genomic and epigenomic markers, their associations described here do not allow categorization of gene expression subtypes, thus reinforcing the notion that transcriptional signatures allow refinement of disease subclassification beyond what can be achieved by currently validated biomarkers[28]. For example, although tumors with wild-type *RAS* are considered to be a homogeneous entity for the purpose of making therapeutic decisions in the setting of advanced cancer, they were found across distinct CMS groups with profound biological differences that are expected to translate into heterogeneous drug responses.

Future steps and resources

Qualitative and clinically relevant disease subtyping takes time and multiple resources. Our CRC subclassification effort is a stepwise process that aims to involve a large number of relevant researchers from the CRC research community at first and subsequently involve 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 be predictive of the benefit of immune checkpoint blockade in advanced CRC, corroborating the value of integrating knowledge of the underlying biology with drug development strategies[29]. Although this is admittedly speculative at this point, the oncogene amplifications that were found in CMS2 samples, the prominent metabolic activation of CMS3 tumors and the TGF-β signaling dependence of CMS4 malignancies have strong potential for the development of novel targeted therapies in CRC, yielding well-defined and reasonably sized groups 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 to our inability to properly define pathway engagement or cross-talk using static 'omics' data. Reanalysis of relevant clinical trials using semisupervised approaches that are based on predefined patient subgroups and allowing for further discovery on the basis of observed outcomes may be the best alternative for the research community. Our current work, which provides the consensual best description of CRC heterogeneity available today, aims at delivering 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 that can be used by the community as research tools (R package available for download; see Online Methods), either in the context of population studies (random forest classifier, as described above, which requires data normalization) or for use in a single-sample setting (alternative Pearson-based predictor, which has been optimized to be less dependent on preprocessing of gene expression data). Of note, samples that do not fall within the four CMS groups should be considered separately as indeterminate subtypes, of yet unknown biological and clinical behavior.

To conclude, we believe that the framework presented here provides a common foundation for CRC subtyping and is 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 expertises will be emulated by other disease areas to accelerate our understanding of tumor biology.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Normalized gene expression data, CMS subtyping calls and sample annotation from public data sets used in the consortium are available at Synapse (Synapse ID syn2623706).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to acknowledge 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 (grant 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 and UVA2014-7245), Worldwide Cancer Research (14-1164), the Maag Lever Darm Stichting (MLDS) (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) and the MLDS (FP012). S.K. is supported by the US National Institutes of Health (grants R01CA172670, R01CA184843, R01 CA187238 and P30CA016672 (Biostatistic and Bioinformatic Core)). A. Sadanandam and G.N. acknowledge support from the National Health Service. S.T. is supported by the Katholieke Universiteit Leuven GOA/12/2106 grant, the EU FP7 Coltheres grant, the Research Foundation Flanders and the Belgian National Cancer Plan.

AUTHOR CONTRIBUTIONS

J.G., R.D., J.P.M., A. Sadanandam, L.W., M.D., S.K., L.M., L.V., S.T. and S.H.F. conceived and designed the study. A.d.R., P.R., P.L.-P., I.M.S., E.F., F.D.S.E.M., E.M., D.B., K.H., J.W.G., B.B., D.H., J.T., R.B., J.P.M., A. Sadanandam, L.W., M.D., S.K., L.V., V.B. and S.T. provided study materials. J.G., R.D., P.A., B.B., S.G., E.F., D.B., K.H., D.M., G.C.M. and B.M.B. collected and assembled the data. J.G., R.D., X.W., A.d.R., A. Schlicker, C.S., L.M., G.N., P.A., B.M.B., J.M., T.L., L.V., A. Schlicker, J.S.M., B.P.-V., R.S. and M.D. analysed and interpreted the data. J.G., R.D., X.W., A.d.R., A. Sadanandam, C.S., L.M., J.T., R.S., J.P.M., A. Schlicker, M.D., S.K., L.V. and S.T. wrote the manuscript. All authors contributed to the final approval of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Hoadley, K.A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

2. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
3. Roepman, P. *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int. J. Cancer* **134**, 552–562 (2014).
4. Budinska, E. *et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **231**, 63–76 (2013).
5. Schlicker, A. *et al.* Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med. Genomics* **5**, 66 (2012).
6. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**, 619–625 (2013).
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.* **19**, 614–618 (2013).
8. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation and prognostic value. *PLoS Med.* **10**, e1001453 (2013).
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* **12**, 260 (2012).
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.* **27**, 3117–3125 (2009).
11. Van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).
12. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
13. Llosa, N.J. *et al.* The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov.* **5**, 43–51 (2015).
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* **5**, 4722–4731 (2014).
15. Son, J. *et al.* Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. *Nature* **496**, 101–105 (2013).
16. Kamphorst, J.J. *et al.* Hypoxic and Ras-transformed cells support growth by scavenging unsaturated fatty acids from lysophospholipids. *Proc. Natl. Acad. Sci. USA* **110**, 8882–8887 (2013).
17. Ying, H. *et al.* Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell* **149**, 656–670 (2012).
18. Lei, Z. *et al.* Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* **145**, 554–565 (2013).
19. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
20. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
21. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
22. Li, Y., Choi, P.S., Casey, S.C., Dill, D.L. & Felsher, D.W. MYC through miR-17–92 suppresses specific target genes to maintain survival, autonomous proliferation and a neoplastic state. *Cancer Cell* **26**, 262–272 (2014).
23. Park, S.-M., Gaur, A.B., Lengyel, E. & Peter, M.E. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev.* **22**, 894–907 (2008).
24. Carmona, F.J. *et al.* A comprehensive DNA methylation profile of epithelial-to-mesenchymal transition. *Cancer Res.* **74**, 5608–5619 (2014).
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* **117**, 4623–4632 (2011).
26. Gavin, P.G. *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.* **18**, 6531–6541 (2012).
27. Popovici, V. *et al.* Context-dependent interpretation of the prognostic value of BRAF and KRAS mutations in colorectal cancer. *BMC Cancer* **13**, 439 (2013).
28. Sinicrope, F.A. *et al.* Molecular markers identify subtypes of stage III colon cancer associated with patient outcomes. *Gastroenterology* **148**, 88–99 (2015).
29. Le, D.T. *et al.* PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
30. Yoshihara, K. *et al.* Inferring tumor purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

[1]Sage Bionetworks, Seattle, Washington, USA. [2]Vall d'Hebron Institute of Oncology (VHIO), Universitat Autònoma de Barcelona, Barcelona, Spain. [3]Laboratory for Experimental Oncology and Radiobiology (LEXOR), Center for Experimental Molecular Medicine (CEMM), Academic Medical Center (AMC), University of Amsterdam, Amsterdam, the Netherlands. [4]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong. [5]Ligue Nationale Contre le Cancer, Paris, France. [6]Netherlands Cancer Institute (NKI), Amsterdam, the Netherlands. [7]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland. [8]Agendia NV, Amsterdam, the Netherlands. [9]Institute of Cancer Research, London, UK. [10]The University of Texas, M.D. Anderson Cancer Center, Houston, Texas, USA. [11]École Polytechnique Fédérale 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]Biomedical Engineering, Oregon Health Sciences University, Portland, Oregon, USA. [16]Université Paris Descartes, Paris, France. [17]Department of Biology, Hôpital Européen Georges-Pompidou, Assistance Publique - Hôpitaux de Paris, Paris, France. [18]Ludwig Center for Cancer Research, University of Lausanne, Lausanne, Switzerland. [19]Department of Oncology, University of Lausanne, Lausanne, Switzerland. [20]Universitair ziekenhuis Leuven, Leuven, Belgium. [21]These authors contributed equally to this work. [22]These authors jointly directed this work. Correspondence should be addressed to J.G. (justin.guinney@sagebase.org), L.V. (l.vermeulen@amc.uva.nl) or S.T. (sabine.tejpar@uzleuven.be).

ONLINE METHODS

Overall design. The design and workflow of this project are described in **Figure 1**. There were six participating groups, each of which 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 (http://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 the time of specimen collection. Gene expression data was accessible to all groups, and non-expression data (i.e., clinical and 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'). Although 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 subtype labels from each group as reported in corresponding subtyping publications was maintained in the collection of data sets from the consortium (**Supplementary Fig. 12**). All results were deposited in Synapse, allowing for an automated evaluation of all results.

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.

Group A. Budinska *et al.*[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 in ref. 4. 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 s.d. 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.

Group B. Marisa *et al.*[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 of the expression profiles were normalized together using the robust multi-array average (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; interindividual 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.

Group C. Roepman *et al.*[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- and 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 ref. 3. CRC samples analyzed on the Agilent platform were preprocessed by median centering within each of the Agilent data sets. Following median centering, subtype similarity scores for A-type, B-type and C-type were processed similarly as in the Agilent-derived data.

Group D. De Sousa E Melo *et al.*[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. On the basis of consensus clustering (1,000 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 analysis. 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 receiver operating characteristic (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 1,000 iterations. Finally, a PAM classifier of 146 unique genes was built with the optimal threshold for centroid shrinkage selected on the basis of a trade-off between classification performance (error rate < 2%) and the size of the gene signature. To use the CCS classifier, expression profiles obtained after normalization were median-centered across cancer samples. For microarray data generated from 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.

Group E. Sadanandam *et al.*[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 tenfold 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 on the basis of 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.

Group F. Schlicker *et al.*[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 on the basis of 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. To derive a probability value for a subtype assignment, we performed the hierarchical clustering on 10,000 randomly selected bootstraps. The subtype probability is then defined as relative frequency with which a sample has been assigned to each subtype.

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 analyzed 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 ref. 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 genome analyzer (GA) and HiSeq samples and was batch-corrected using the ComBat method[32].

We additionally performed outlier sample detection within each data set using two methods: a method based on PCA and one using 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 on the basis of 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 on the basis of the distribution of gene expression values and either pairwise distances to other samples or to the PCA criterion. Outliers were removed from further analysis.

Network analysis of subtype association. To study the association between the six CRC classification systems (A to F, each consisting of three, five or six 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 on the basis of 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.

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% of 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 (a) and (b) are repeated for n = 1,000 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. On the basis of 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 $f < 1.6$, whereas $f > 10$ 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 three, four and five clusters, respectively, are shown in heat maps ordered by identified CMS groups (**Supplementary Fig. 1**).

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 the 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 the 'Classification' section). The distribution of unlabeled samples per data set is shown in **Supplementary Figure 2**.

Data aggregation. 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 of the 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 was as follows. (i) Remove outlier samples from each data set separately (see section 'Gene expression data processing and normalization' for details). (ii) 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. (iii) Select a reference data set (referred onwards as D_{REF}). In our case we chose the largest Affymetrix data set[8], and in this data set, each gene in G_{REF} was represented by the probeset with the largest MAD. (iv) For each of the other data sets, we used a consistency criterion to select the probe set to represent each gene. First, for each probe set, we calculated the correlation between the expression of the probeset and the reference genes in the same data set. This gives, for each probe set, a correlation vector C of length $|G_{REF}|$. (v) To select the probe set 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 probe sets and the correlation vector for gene g in D_{REF}. (vi) To select the probe set that represented gene g in the reference data set D_{REF}, we chose the probe set with the highest correlation with most of the other data sets. Therefore, for each data set D, we selected the probe set in D_{REF}, which has the largest 'correlation of correlations value from 'V' with the probe sets in D. The probe set selected is the one chosen to represent g in D_{REF}. (vii) For each other

doi:10.1038/nm.3967

data set D, the probe set with the highest value of the 'correlation of correlations' with the chosen probe set in D_{REF} was selected to represent *g*. (viii) 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. (ix) 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]. (x) We filtered the aggregated data set further on the basis of the quantile range and the correlations calculated in step (v). 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.

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 '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 (scripts and code for the fandom forest CMS classifier are available at Github, https://github.com/Sage-Bionetworks/crcsc). 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.

Global classifier. To assess feasibility of developing a CMS classifier, we randomly split our aggregated gene expression data matrix into two-thirds training and one-third 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. Although overall performance was robust (**Supplementary Fig. 3a**), we observed that the four data sets generated using the Agilent platform had significantly lower performance metrics (**Supplementary Fig. 3b**).

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**).

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.

Data set splits. The previous classifiers were developed by randomly sampling from all data sets and partitioning them 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 used the GSE39582 (Affymetrix, fresh-frozen, n = 466), the TCGA (RNAseq, n = 459) and the 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 that in the Affymetrix data sets. Overall, we observed robust performance metrics in these validation data sets.

Our second data split experiment was to separate the PETACC-3 (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 0.70/0.98. The high type II error rate in CMS3 suggests some biological differences between FFPE and fresh-frozen samples and underscores the importance of using FFPE samples for training a classifier in this context.

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 Fig. 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 intrasubtype 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 ($R = -0.60$, $P < 1 \times 10^{-16}$) and CMS3 and CMS4 ($R = -0.76$, $P < 1 \times 10^{-16}$) 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 an s.d. cut-off of 1 and fitted the factor analysis model to this data set 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 are not separate from the CMS-classified samples (**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 are 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**).

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 nonparametric 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 being 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.

Mutation profile. KRAS, *BRAF*, *PIK3CA*, *PTEN*, *APC* and *TP53* mutation detection: for sequencing platform details refer to publications from the

doi:10.1038/nm.3967

individual groups. In summary, in data sets other than TCGA, targeted sequencing was performed (codons or specific variants in oncogenes—*KRAS*, *BRAF* and *PIK3CA*—and most frequently mutated exons in tumor suppressors—*PTEN*, *APC* and *TP53*). For TCGA samples, somatic mutations and indels (insertions and deletions) 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 (https://www.synapse.org/#!Synapse:syn300013/files/; 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 on the basis of 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 nonsilent 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.

Copy number events profile. Arm level copy number changes were visualized by using the GISTIC scores and CMS labels with the University of California Santa Cruz 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, whereas 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 (ref. 2) (*MYC*, *HNF4A*, *CDK8*, *FGFR1*, *ERBB2*, *IGF2*, *PTEN*, *SMAD4*, *APC* and *TCF7L2*), high level amplification was defined as GISTIC scores +2 and homozygous deletion as GISTIC scores −2.

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).

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 (ref. 2)) and extended this analysis with an additional 450K data set as detailed below.

We downloaded Level3 β-values from the 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 nonbiological batch effect with respect to tissue source site (TSS) and batch variables. The hierarchical clustering was performed on the basis of 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 on the basis of s.d. (n = 1,486; s.d. > 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 heat map 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 two clusters to cluster3-cluster4.

For differential methylation analysis, we used 187 tumor samples that have classification labels on the basis of TCGA gene expression CMS 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 on the basis of two sample t-tests, comparing each CMS with the other CMS groups. Out of the total 21,231 genes, we identified 1,664 genes that were differentially methylated (Benjamini–Hochberg-corrected *P* value < 0.05 and $|\log_2$ fold change$| > 0.5$) between at least one CMS and the others (heat map shown in **Supplementary Fig. 6c**). As expected, most of the differentially methylated genes (n = 1,262) have significantly higher methylation in CMS1 tumors, which is consistent with their CIMP-H status. Nonetheless, we also observed genes that were specifically hyper- or hypomethylated 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.

Integrative analysis. We performed integrative analysis in the 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 nonsilent events were considered activating or inactivating alterations. For copy number events, only high-level amplifications or homozygous deletions were defined as alterations. In some cases, up- or downregulation of gene expression was also considered a pathway alteration (*IGF2*, *FZD10* and *SMAD4* genes).

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

(i) ESTIMATE algorithm: method that uses gene expression signatures to infer the fraction of stromal and immune cells in tumor samples[30].

(ii) 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 TGF-β pathway[54], intestinal stem cells[55], matrix remodeling (REACTOME) and wound-response (GO BP).

(iii) Canonical genesets: MAPK and PI3K (GO BP), SRC, JAK-STAT, caspases (BIOCARTA), proteosome (KEGG), Notch, cell cycle, translation and ribosome, integrin-β3 and vascular endothelial growth factor (VEGF) and VEGF receptor (VEGFR) interactions (REACTOME).

(iv) 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 (T_{FH}) cells[60] in cancer samples, activation of T helper 17 (T_H17) cells[61], regulatory T cells (T_{reg})[62] or myeloid-derived suppressor cells (MDSC)[63].

(v) 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) and 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.

Proteomic analysis. For reverse-phase protein array (RPPA), normalized measurements of 187 proteins were downloaded from the 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 and Rad51), cell cycle (cyclins D1, E1 and E2) and DNA damage repair (Chk1), whereas CMS4 samples were mainly enriched for microenvironment proteins (collagen and fibronectin).

We also obtained liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic quantile-normalized and log-transformed data for 95 TCGA tumor samples[21]. Heat map 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 CMSs 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**.

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 $\log_2$-transformed after adding one 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 nonbiological 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 on the basis of two sample t-tests, comparing each CMS with the other CMS groups. A high Pearson correlation coefficient was observed in the $\log_2$ fold change between data sets 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 $|\log_2$ fold change$| > 0.5$) between at least one CMS and the others.

Differentially expressed miRNAs between CMSs were illustrated in a heat map (**Supplementary Fig. 8c**). CMS2 can be characterized by the upregulated 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. Of the total of six miRNAs downregulated 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 (for example, hsa-mir-148a and the miR-192 and miR-200 families) that are known for tumor suppression. The miR-200 and miR-192 families regulate epithelial-to-mesenchymal transition (EMT) pathway by targeting ZEB1 and/or ZEB2 (refs. 23,67), whereas hsa-mir-148a is predicted by TargetScan[68] to regulate MMP13 and TGFB2, which are important for the matrix remodeling (MR) and TGF-β pathways. Taken together, the downregulation of miRNAs associated with suppression of the EMT-, MR- and TGF-β–associated signatures could explain why CMS4 is more aggressive and metastatic than the other CMSs.

Clinical and pathological variables. Data from different data sets were standardized as described as: (i) site: right colon (cecum, ascending, hepatic flexure and transverse colon), left (splenic flexure, descending and sigmoid colon) and rectum (**Supplementary Fig. 10**). (ii) Stage: assignments were defined using the latest edition of the American Joint Committee on Cancer's Cancer Staging Manual available at the time of diagnosis (third to sixth). For consistency, we only investigated the major stage (I, II, III or IV), whose definition does not change in these different staging systems. (iii) Grade: 1 (well-differentiated), 2 (moderately differentiated) and 3 (poorly differentiated) carcinomas, according to pathology review performed by each independent institution.

Tumor purity analysis. We obtained the tumor purity estimation of CRC samples in the TCGA data set as defined by the ABSOLUTE algorithm[20] (https://www.synapse.org/#!Synapse:syn1710466). 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 a reduced proportion of cancer cells (i.e., 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 the ESTIMATE algorithm[27] (**Fig. 3i**).

Analysis of tumor versus normal non-cancer tissue. 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 renormalized (see previous description of data normalization) including those from 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 versus normal samples, 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**.

Survival analyses. Overall survival (OS) and relapse-free survival (RFS) times were calculated on the basis of 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, whereas both RFS and SAR analyses were limited to patients with stage I, II or III tumors 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**.

To evaluate the performance of survival models, we split the data sets into two-thirds and one-third for training and validation, respectively, 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 (ref. 69). We calculated log-rank P values in survival models and compared multivariate models with and without CMS classification by performing analysis of variance (ANOVA). Paired Student's t-test was used to compare tAUC estimates.

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 (https://www.synapse.org/#!Synapse:syn2623706/wiki/). 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') that includes the random forest classifier described previously, as well as a 'single-sample predictor' (SSP) classifier. By definition, an SSP makes it 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 (intraseries) 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 or two-color 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). 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 seen any more as a single-sample predictor.

The SSP is implemented in the 'CMSclassifier' R package. It is based on a similarity-to-centroid approach, with the Pearson coefficient as a similarity measure. It uses centroids of the CMS calculated for 693 discriminant genes (Entrez Ids), which were selected using the GSE39582 series on the basis of 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 and E-MTAB-990), 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 five 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 CMSs is also reported (D = $Si^* – Si'$, with i' being the second-nearest CMS). Then if S_{i^*} is above 0.15 and D is above 0.06, the sample is classified in CMS i^*, otherwise its label is '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**.

31. Derry, J.M.J. *et al.* Developing predictive molecular maps of human disease through community-based modeling. *Nat. Genet.* **44**, 127–130 (2012).
32. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
33. McCall, M.N., Bolstad, B.M. & Irizarry, R.A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242–253 (2010).
34. Zilliox, M.J. & Irizarry, R.A. A gene expression bar code for microarray data. *Nat. Methods* **4**, 911–913 (2007).
35. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
36. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002).
37. Brunet, J.-P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
38. Marron, J.S., Todd, M.J. & Ahn, J. Distance-weighted discrimination. *J. Am. Stat. Assoc.* **102**, 1267–1271 (2007).
39. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
40. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
41. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2009).
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.* **228**, 586–595 (2012).
43. Breiman, L. Random forest. *Mach. Learn.* **45**, 5–32 (2001).
44. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323–329 (2012).
45. Murray, J.S., Dunson, D.B., Carin, L. & Lucas, J.E. Bayesian Gaussian copula factor models for mixed data. *J. Am. Stat. Assoc.* **108**, 656–665 (2013).
46. Ghosh, J. & Dunson, D.B. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Stat.* **18**, 306–320 (2009).
47. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
48. Umar, A. *et al.* Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J. Natl. Cancer Inst.* **96**, 261–268 (2004).
49. Lindor, N.M. *et al.* Immunohistochemistry versus microsatellite instability testing in phenotyping colorectal tumors. *J. Clin. Oncol.* **20**, 1043–1048 (2002).
50. Weisenberger, D.J. *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with *BRAF* mutation in colorectal cancer. *Nat. Genet.* **38**, 787–793 (2006).
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. USA* **104**, 15418–15423 (2007).
52. Van der Flier, L.G. *et al.* The intestinal Wnt/TCF signature. *Gastroenterology* **132**, 628–632 (2007).
53. Zeller, K.I., Jegga, A.G., Aronow, B.J., O'Donnell, K.A. & Dang, C.V. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol.* **4**, R69 (2003).
54. Loboda, A. *et al.* EMT is the dominant program in human colon cancer. *BMC Med. Genomics* **4**, 9 (2011).
55. Merlos-Suárez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).
56. Mlecnik, B. *et al.* Biomolecular network reconstruction identifies T cell homing factors associated with survival in colorectal cancer. *Gastroenterology* **138**, 1429–1440 (2010).
57. Tosolini, M. *et al.* Clinical impact of different classes of infiltrating T cytotoxic and helper cells (T_H1, T_H2, T_{reg}, T_H17) in patients with colorectal cancer. *Cancer Res.* **71**, 1263–1271 (2011).
58. Galon, J. *et al.* Type, density and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964 (2006).
59. Ascierto, M.L. *et al.* Molecular signatures mostly associated with NK cells are predictive of relapse-free survival in breast cancer patients. *J. Transl. Med.* **11**, 145 (2013).
60. Gu-Trantien, C. *et al.* $CD4^+$ follicular helper T cell infiltration predicts breast cancer survival. *J. Clin. Invest.* **123**, 2873–2892 (2013).
61. Keerthivasan, S. *et al.* β-Catenin promotes colitis and colon cancer through imprinting of proinflammatory properties in T cells. *Sci. Transl. Med.* **6**, 225ra28 (2014).
62. Stockis, J. *et al.* Comparison of stable human T_{reg} and T_H clones by transcriptional profiling. *Eur. J. Immunol.* **39**, 869–882 (2009).
63. Fridlender, Z.G. *et al.* Transcriptomic analysis comparing tumor-associated neutrophils with granulocytic myeloid-derived suppressor cells and normal neutrophils. *PLoS ONE* **7**, e31524 (2012).
64. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129 (2007).
65. Wang, L. *et al.* miR-143 acts as a tumor suppressor by targeting N-RAS and enhances temozolomide-induced apoptosis in glioma. *Oncotarget* **5**, 5416–5427 (2014).
66. Johnson, S.M. *et al.* RAS is regulated by the let-7 microRNA family. *Cell* **120**, 635–647 (2005).
67. Kim, T. *et al.* p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. *J. Exp. Med.* **208**, 875–883 (2011).
68. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
69. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).