

# Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer

Eva Budinska,<sup>1,2\*</sup> Vlad Popovici,<sup>1,2</sup> Sabine Tejpar,<sup>3</sup> Giovanni D'Ario,<sup>1</sup> Nicolas Lapique,<sup>1</sup> Katarzyna Otylia Sikora,<sup>1</sup> Antonio Fabio Di Narzo,<sup>1</sup> Pu Yan,<sup>4</sup> John Graeme Hodgson,<sup>5</sup> Scott Weinrich,<sup>5</sup> Fred Bosman,<sup>5</sup> Arnaud Roth<sup>6,7</sup> and Mauro Delorenzi<sup>1,8</sup>

<sup>1</sup> Bioinformatics Core Facility, Swiss Institute of Bioinformatics (SIB), Lausanne, 1015, Switzerland

<sup>2</sup> Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

<sup>3</sup> Department of Oncology, University Hospital Gasthuisberg, Katholiek Universiteit Leuven, Belgium

<sup>4</sup> University Institute of Pathology, Lausanne University Medical Centre, Switzerland

<sup>5</sup> Pfizer Inc., Worldwide Research and Development, Oncology Research Unit, La Jolla, CA, USA

<sup>6</sup> Oncosurgery, Geneva University Hospital, Switzerland

<sup>7</sup> Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland

<sup>8</sup> Département de Formation et Recherche, Lausanne University Medical Centre, Switzerland

\*Correspondence to: Eva Budinska, Institute of Biostatistics and Analyses, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic  
e-mail: budinska@iba.muni.cz

## Abstract

The recognition that colorectal cancer (CRC) is a heterogeneous disease in terms of clinical behaviour and response to therapy translates into an urgent need for robust molecular disease subclassifiers that can explain this heterogeneity beyond current parameters (MSI, *KRAS*, *BRAF*). Attempts to fill this gap are emerging. The Cancer Genome Atlas (TCGA) reported two main CRC groups, based on the incidence and spectrum of mutated genes, and another paper reported an EMT expression signature defined subgroup. We performed a prior free analysis of CRC heterogeneity on 1113 CRC gene expression profiles and confronted our findings to established molecular determinants and clinical, histopathological and survival data. Unsupervised clustering based on gene modules allowed us to distinguish at least five different gene expression CRC subtypes, which we call surface crypt-like, lower crypt-like, CIMP-H-like, mesenchymal and mixed. A gene set enrichment analysis combined with literature search of gene module members identified distinct biological motifs in different subtypes. The subtypes, which were not derived based on outcome, nonetheless showed differences in prognosis. Known gene copy number variations and mutations in key cancer-associated genes differed between subtypes, but the subtypes provided molecular information beyond that contained in these variables. Morphological features significantly differed between subtypes. The objective existence of the subtypes and their clinical and molecular characteristics were validated in an independent set of 720 CRC expression profiles. Our subtypes provide a novel perspective on the heterogeneity of CRC. The proposed subtypes should be further explored retrospectively on existing clinical trial datasets and, when sufficiently robust, be prospectively assessed for clinical relevance in terms of prognosis and treatment response predictive capacity. Original microarray data were uploaded to the ArrayExpress database (<http://www.ebi.ac.uk/arrayexpress/>) under Accession Nos E-MTAB-990 and E-MTAB-1026.

© 2013 Swiss Institute of Bioinformatics. *Journal of Pathology* published by John Wiley & Sons Ltd on behalf of Pathological Society of Great Britain and Ireland.

**Keywords:** colorectal cancer; histopathology; gene expression; molecular heterogeneity

Received 3 February 2013; Revised 10 May 2013; Accepted 14 May 2013

Conflict of interest: ST and FTB have received research funds from Pfizer; and AR is a compensated consultant for Pfizer.

## Introduction

Current classifications of sporadic colorectal cancer take into consideration stage, histological type and grade [1]. Colorectal cancer (CRC) is a highly heterogeneous disease, with clinicopathologically similar tumours differing strikingly in treatment response and patient survival. These differences are only partly explained by current concepts regarding the molecular events leading to CRC. In recent years, microsatellite

instability (MSI) emerged as an important classifier with significant prognostic impact and potential for patient stratification for therapy [2,3]. Some molecular markers, as well as the mutation status of *BRAF* or *KRAS* genes (predictive for anti-EGFR [4]), are in use for treatment decisions and patient stratification. However, patient groups defined by these molecular markers still differ remarkably in behaviour and therapy response [5,6]. Several approaches to further subtype CRC have been proposed, based on combinations

of clinical, histopathological, gene expression, CNV, epigenetic and single gene parameters [7–13]. Each of these different modalities provides its own perspective on the same underlying biological reality. The CpG island methylator phenotype (CIMP) status is emerging as important molecular determinant of CRC heterogeneity [11]. The cancer genome atlas (TCGA) analysis identified a hypermutant group not entirely captured by MSI status [13]. Several studies have addressed CRC subtyping using genome-wide gene expression profiling of relatively large patient cohorts [12,14]. One study used unsupervised clustering of stage II and III CRCs to identify three stage-independent subtypes, with *BRAF* mutation and MSI status dominating one of the subtypes [14]. A study of stage I–IV CRC samples segregated CRC into two prognostic subtypes with epithelial–mesenchymal transition (EMT) as a main determinant [12]. Another study on 88 stage I–IV samples identified four subtypes, one correlated with MSI, *BRAF* mutation and mucinous histology, two with stromal component and one with high nuclear  $\beta$ -catenin expression [15].

We recently reported CRC expressing a *BRAF*-mutated signature [6], which strongly overlaps with the methylation-based group of Hinoue [11], and a MSI-like gene expression group that captures the hypermutant tumours of TCGA [13], indicating the potential for identification of robust biological subgroups. We now describe CRC subtypes based upon unsupervised clustering of genome-wide expression patterns. We characterized these subtypes in terms of biological motifs, common clinical variables, association with known CRC molecular markers and morphological patterns. A key element in our approach was the use of a system of unsupervised gene modules—groups of genes with correlated expression. They are more resistant to noise and have a higher chance of having at least a few members represented on various platforms. In addition, as each gene module is represented by its median expression, the modules with fewer genes contribute equally to the subtype definition. We and others have successfully used similar strategies previously [16–18]. We validated the existence of the subtypes and their respective clinical and molecular marker characteristics in an independent dataset. Ultimately, it will be mandatory to integrate the various sources of information on CRC heterogeneity into an integrative, robust and reproducible subclassifier that can become a tool for clinical use.

## Materials and methods

A detailed description of all the datasets and analysis procedures is given in Supplementary methods and results (see Supplementary material).

### Data acquisition and processing

We have built two non-overlapping data collections: a discovery collection, comprising four publicly

available (425 samples) and two previously unpublished datasets (688 samples with 10 year follow-up in a clinical trial setting and 64 normal samples) with known stage status, and a validation collection of eight publicly available datasets (720 CRC samples) (see Supplementary material, Supplementary methods and results). Observations derived from the analysis of 64 normal samples were further validated on five publicly available datasets, with both carcinoma and normal samples available in one batch (totalling 205 normal/adenoma/carcinoma samples). Copy number data was available for 154 of the PETACC3, as in [19]. Our analysis included a total of 2102 samples.

The discovery collection contained the previously unpublished 688 CRC formalin-fixed, paraffin-embedded (FFPE) samples of PETACC3 [6] and 64 FFPE normal colon tissue samples from Centre Hospitalier Universitaire Vaudois's Biobank, which were uploaded to ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>), under Accession Nos E-MTAB-990 and E-MTAB-1026, respectively. Gene expression data were processed by standard tools to obtain normalized, probeset-level expression data. For each EntrezID in the datasets, the probeset with the highest variability was selected as representative and the number of EntrezIDs entering the analysis was reduced to 3025 by applying non-specific filtering. For PETACC3 and normal colon samples, patients signed an informed consent form in which the use of tissue specimens was included, and all marker study proposals were subjected to the approval of the trial steering committee.

### Subtype definition and validation

For model development (gene modules and subtype definition, classifier training, identification of subtype-specific genes) only the 1113 CRC samples of the discovery set were used, no sample in the validation collection being used for any model tuning. Hierarchical clustering (complete linkage, Pearson correlation similarity measure) and dynamic cut tree [20] were used to produce *gene modules* (groups of genes with correlated expression), from which non-robust modules (see Supplementary material, Supplementary methods and results) and a gender-related module were discarded. Each expression profile was then reduced to a vector of *meta-genes* by taking the median of the values of genes in each gene module. The meta-genes were then further grouped into clusters using hierarchical clustering.

The subtypes were defined in terms of *core samples*—those samples from the discovery collection that were assigned to clusters by hierarchical clustering, using a consensus distance [21] followed by pruning of the dendrogram (see Supplementary material, Supplementary methods and results). The clusters to which the core samples were assigned were called

*subtypes*. The rest of the samples from the discovery collection, not assigned to subtypes by this procedure, were called *non-core samples*. This approach allowed the reduction of noise in subtype-defining samples, and thus a higher consistency of the resulting subtypes defining the ground truth for downstream analyses. The stability of the obtained clusters was assessed under different perturbations of the processing pipeline (different parameters and clustering methods) to ensure that the results were not simple artefacts (see Supplementary material, Supplementary methods and results). A multiclass linear discriminant (LDA) [22] was trained on core samples with meta-genes as variables to assign new samples to one of the subtypes. Minimal gene sets characteristic to each subtype were identified using ElasticNet [23] on gene-level data.

In order to validate the existence of subtypes (and their independence on data selection) and the modelling choices in subtype discovery, we applied the same subtyping procedure (including parameters) to the validation collection. The clusters identified in the validation collection were put in correspondence with the subtypes in the training set by LDA predictions and correlations of subtype-specific moderated *t* statistic [24] values, corresponding to the gene-wise comparison of the respective subtype with the other subtypes (one-versus-all comparison). A simple classifier application would have led the validation samples to be classified as one of the subtypes, but it would have not informed us of possible over-fitting of the data in the discovery procedure.

### Subtype characterization

If not specified differently, all the reported *p* values were adjusted for multiple hypothesis testing, using the Benjamini–Hochberg procedure. Significance level was set at 0.1. Pathway analysis for each set of gene modules was carried out using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [25]. Gene set enrichment analysis of gene signatures was performed using the mygsea2 tool, in each subtype and normal samples, on average expression-ordered median-centred lists of genes. Differential expression analysis was performed using limma [24] and sign test using BSDA [26]. The Cox proportional hazards model was used to analyse the prognostic value of interquartile range (IQR)-standardized values of meta-genes, for overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR), stratified by dataset. The Wald test was used to assess the global significance of the models. Pairwise differences in survival were assessed using the log-rank test. For subtype comparison, the survival was truncated at 7 years. Subtype enrichment for clinical or molecular markers was assessed by the Fisher test to the baseline, defined as the proportion of the marker in the whole dataset. Morphological pattern differences were assessed pairwise by Fisher test.

### Histology

The identified subtypes were characterized histologically in terms of six different architectural patterns: complex tubular; solid/trabecular; mucinous; papillary; desmoplastic; and serrated (Figure 4A), which were called dominant or secondary depending on their presence in the histology slides (for details on immunohistochemistry, see Supplementary material, Supplementary methods and results).

## Results

### Gene modules and subtype definition

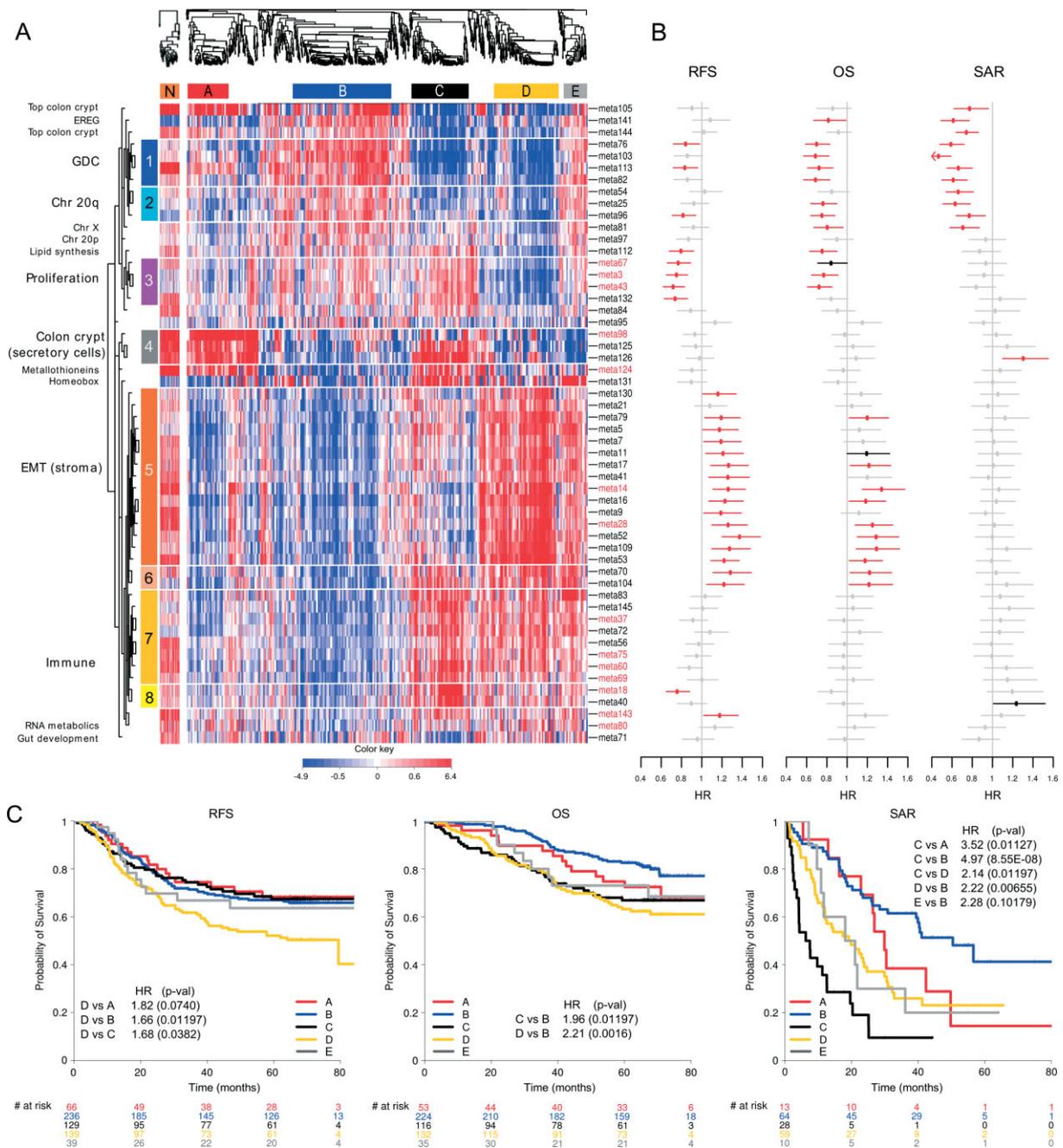
We identified 54 gene modules, reproducible across all datasets in the discovery collection, comprising 658 genes from an initial list of 3025 identified as the most variable. The assignment of genes to gene modules and gene module clusters is listed in Table S1 (see Supplementary material); meta-gene expression profiles for the discovery set are shown in Figure 1A; and between meta-gene correlations in Figure S1C (see Supplementary material). Based on gene modules, we identified five major subtypes: surface crypt-like (A), lower crypt-like (B), CIMP-H-like (C), mesenchymal (D) and mixed (E), totalling 765 samples (69% of discovery data; see Supplementary material, Supplementary methods and results).

### Subtype reproducibility in an independent validation set

In the validation set of 720 CRC samples we identified a set of subtypes comprising 602 samples (83.6% of the validation set) and associated them with our discovery subtypes using the subtype classifier (see Supplementary material, Table S2) and correlations of subtype-specific patterns based on moderated *t* statistic (see Supplementary material, Table S3). All five major subtypes reappeared in the validation set, confirming the robustness of our approach. Figure S2 (see Supplementary material) presents gene expression profiles of both discovery and validation sets. Two notable differences were observed: (i) subtype B in the validation set was split into two subgroups (B1, B2), as observed in the discovery set too, but only at lower pruning height; (ii) another cluster passed the minimal size criteria, corresponding to the small subtype (F) which, in the discovery set, was not considered for further characterization because of small sample size. Validation of other subtype characteristics (to the extent of available information) is described in each of the respective sections.

### Subtypes are characterized by distinct biological components

We set out to assign biological labels to gene modules that define the subtypes (Table 1; see also Supplementary material, Table S1). Of the 54 meta-genes,



**Figure 1.** Meta-gene expression pattern in subtypes, connected with prognostic effect of subtypes and meta-genes, in the discovery set. (A) Two heat maps clustering normal (left) and CRC (right) samples (columns) and meta-genes (rows). Colours represent decreased (blue) or increased (red) meta-gene expression relative to their medians. Normal samples were clustered independently on meta-genes centred to CRC meta-gene medians. For comparative purposes, ordering of meta-genes in normal samples is imposed to correspond to that of CRC samples. White horizontal lines denote eight unsupervised clusters of meta-genes, each assigned a colour bar on the left; meta-genes not belonging to a cluster have no colour bar. Names of the meta-genes corresponding to gene modules with gene–gene correlations in normal samples comparable to those in cancer samples are marked red (see Supplementary material, Figure S1D). (B) Effect of inter-quartile range (IQR) standardized expression of meta-genes on RFS, OS and SAR. Points represent estimated hazard ratio (HR), bars represent 95% CI. Bold lines represent effects significant at 5% without adjustment for multiple hypothesis testing; red lines represent effects significant at FDR < 10%; details are provided in Table S6 (see Supplementary material). (C) Kaplan–Meier plots for RFS, OS and SAR, with HR for significant pairwise comparisons ( $p$  values adjusted for FDR). Numbers below  $x$  axes represent number of patients at risk at selected time points.

41 could be further grouped into eight gene module clusters; 13 meta-genes remained ungrouped, each possibly representing a distinct biological motif. Pathway analysis characterized five of eight gene module clusters by the following biological motifs: chromosome

20q (cluster 2), proliferation (cluster 3), EMT/stroma (cluster 5) and immune response (clusters 7 and 8). Literature searching identified biological motifs associated with other gene modules. We labelled cluster 1 as GDC (genes differentially expressed in CRC), as

Table 1. Biological identification of gene modules

| Cluster name  | Number of genes | Pathway analysis result (number of overlapping genes, <i>p</i> value) OR description based on literature search   | Selected genes   |
|---|-----------------|---|--|
| 1. GDC  | 27              | Genes involved in differentiation of colon crypt and/or whose expression was reported to be affected in colorectal cancer and/or with prognostic effect in CRC  | Intestinal differentiation genes: <i>CDX2</i> [45], <i>IHH</i> [46], <i>VAV3</i> [47], <i>ASCL2</i> [35], <i>PLAGL2</i> [48]<br>Genes reported altered in colorectal cancer with prognostic effect: <i>PITX2</i> [49], <i>DDC</i> [50], <i>PRLR</i> [51], <i>SPINK1</i> [52]<br>Other genes connected to CRC:<br><i>GGH</i> –connected to CIMP <sup>+</sup> phenotype [53]<br><i>NR1I2</i> –connected to chemoresistance [54]  |
| 2. Chromosome 20q genes   | 33              | Chromosome 20 (26 genes, 9.2E-34)   | Other, non-20q genes: <i>TP53RK</i> , <i>AN09</i> , <i>NEU1</i> , <i>CLDN3</i> , <i>PRSS8</i>  |
| 3. Proliferation  | 83              | Cell cycle (36 genes, 3.0E-33)<br>Mitosis (26 genes, 1.4E-29)<br>Chromosome (26 genes, 2.5E-17)<br>DNA metabolic process (20 genes, 4.9E-10)<br>Lipid synthesis (4 genes, 5.0E-2)   | Mitotic checkpoint kinases: <i>BUB1</i> , <i>BUB1B</i><br>Cyclins: <i>CCNA2</i> , <i>CCNB2</i> Centromere proteins: <i>CENPA</i> , <i>CENPE</i> , <i>CENPN</i><br>Kinesins: <i>KIF11</i> , <i>KIF23</i> , <i>KIF4A</i><br>Topoisomerase II ( <i>TOP2A</i> )<br>Cell division cycle 2 <i>CDC2</i>   |
| 4. Colon crypt markers (secretory cells)  | 16              |   | <i>AGR2</i> [55], <i>AGR3</i> , <i>MUC2</i> , <i>SPINK4</i> [56], <i>RETNLB</i> [57], <i>REG4</i> [58]   |
| 5. EMT/stroma   | 310             | Extracellular region part (90 genes) 2.7E-36<br>Cell adhesion (57 genes) 1.2E-17<br>Extracellular matrix (44 genes) 5.3E-30<br>Collagen (16 genes) 1.2E-15<br>EGF-like domain (26 genes) 1.6E-12<br>Cell motion (33 genes) 7.2E-8<br>Blood vessel development (25 genes) 1.1E-8<br>Growth factor binding (6 genes) 6.0E-5<br>Frizzled related (5 genes) 6.7E-3<br>Cell junction organization (7 genes) 1.8E-2<br>WNT receptor signalling pathway (8 genes) 1.4E-1 | Inhibitors of $\beta$ -catenin-dependent canonical WNT: <i>SFRP1</i> , <i>SFRP2</i> , <i>SFRP4</i> , <i>DKK3</i> , <i>FZD1,7</i> , <i>PRICKLE1</i> , <i>NXN</i><br>Mesenchymal markers: N-cadherin, OB cadherin, <i>SPARC</i> , <i>DDR2</i><br>EMT inducers(TFs): <i>SNAI2</i> , <i>ZEB1</i> , <i>ZEB2</i> , <i>TWIST1</i> , <i>CDH11</i><br>ECM remodelling and invasion: <i>MMP14</i> , <i>VIM</i> ECM proteins: fibronectin 1, collagens<br>Angiogenesis: <i>PLAT</i> , <i>PLAU</i> , <i>NRP1</i> , <i>NRP2</i> , <i>THBS1</i> , <i>THBS2</i> , <i>THBS4</i><br>TGFs, their receptors and binding proteins: <i>IGF1</i> , <i>IGFBP5</i> , <i>IGFBP7</i> , <i>TGFB</i> , <i>LTBP1</i> , <i>LTBP2</i> , <i>PDGFRA</i> , <i>PDGFRB</i> |
| 6. Unidentified   | 14              |   | <i>DUSP1</i> , <i>EGR2</i> , <i>SERPINE1</i>   |
| 7 and 8. Immune response  | 103             | Immune response (42 genes) 2.0E-28<br>Positive regulation of immune system process (16 genes) 4.0E-9<br>Antigen processing and presentation via MHC class II (6 genes) 7.5E-5<br>Defence response (31 genes) 3.3E-17<br>Chemokine signalling pathway (9 genes) 2.2E-3<br>Lymphocyte activation (11 genes) 2.1E-5<br>Regulation of programmed cell death (14 genes) 2.1E-2   | Cytokines: <i>CCL3</i> , <i>CXCL5</i> , <i>CXCL9</i> , <i>CXCL10</i> , <i>CXCL11</i> , <i>SPP1</i> , <i>LTB</i><br>MHC class II: <i>HLA-DMB</i> , <i>HLA-DPA1</i> , <i>HLA-DRA</i> , <i>CD74</i><br>MHC class I: <i>HLA-F</i> , <i>TAP1</i> , <i>TAP2</i><br>Anti-apoptotic: <i>BCL2A1</i> , <i>CD74</i> , <i>BIRC3</i> , <i>IFI6</i> , <i>TNFAIP3</i> , <i>TNFAIP3</i><br>Apoptotic: <i>STAT1</i> , <i>XAF1</i><br>Interferon-induced proteins: <i>IFI30</i> , <i>IFI16</i> , <i>IFI44</i> , <i>IFI16</i> , <i>IFIH1</i> , <i>IFIT3</i>   |
| <i>Cluster-unassigned meta-genes with colon crypt cell markers (enterocytes/top of the crypt)</i> |                 |   |  |
| Meta-gene 105   | 6               | Top of the crypt genes  | <i>FAM55A</i> , <i>FAM55D</i> , <i>MUC12</i> and <i>CEACAM7</i> [59], <i>SLC26A2</i> [59], <i>SLC26A3</i> [59]   |
| Meta-gene 144   | 5               | Enterocytes, goblet cells markers   | <i>LOC644844</i> , <i>NGEF</i> , <i>HEPH</i> , <i>KRT20</i> [59], <i>MUC20</i> [59]  |
| <i>Cluster-unassigned meta-genes associated with chromosomal location O</i>                       |                 |   |  |
| Meta-gene 81  | 7               | Chromosome X (7 genes) 1.1E-8   | <i>CXorf15</i> , <i>EIF1AX</i> , <i>HDHD1A</i> , <i>MED14</i> , <i>PNPLA4</i> , <i>SCML1</i> , <i>SMC1A</i>  |
| Meta-gene 97  | 6               | Chromosome 20p (5 genes) 5.0E-11  | <i>CDC25B</i> , <i>CSNK2A1</i> , <i>MRPS26</i> , <i>PTPRA</i> , <i>RP5-1022P6.2</i> , <i>SNRNP</i>   |
| Meta-gene 84  | 7               | Chromosome 8 (7 genes) 5.4E-9   | <i>AGPAT5</i> , <i>FDF1</i> , <i>GTF2E2</i> , <i>LONRF1</i> , <i>MTUS1</i> , <i>VPS37A</i> , <i>ZNF395</i>   |
| <i>Other cluster-unassigned meta-genes</i>  |                 |   |  |
| Meta-gene 141   | 5               | EREG  | <i>AK3L1</i> , <i>ARID3A</i> , <i>EREG</i> , <i>LDLRAD3</i> , <i>ZBTB10</i>  |
| Meta-gene 112   | 6               | Lipid synthesis (4 genes) 5.0E-2  | <i>DHCR7</i> , <i>FASN</i> , <i>FGFBP1</i> , <i>HMGCS1</i> , <i>IDI1</i> , <i>PCSK9</i>  |
| Meta-gene 95  | 6               | Homeobox genes  | <i>HOXA10</i> , <i>HOXA11</i> , <i>HOXA13</i> , <i>HOXA5</i> , <i>HOXA7</i> , <i>HOXA9</i>   |
| Meta-gene 124   | 5               | Metallothioneins  | <i>MT1E</i> , <i>MT1F</i> , <i>MT1G</i> , <i>MT1M</i> , <i>MT1X</i>  |
| Meta-gene 131   | 5               | Disulphide bonds (5 genes) 1.7E-02  | <i>CXCL5</i> , <i>IL6</i> , <i>MMP1</i> , <i>MMP3</i> , <i>PTGS2</i>   |
| Meta-gene 143   | 5               | Unidentified  | <i>DUSP5</i> , <i>ERRF1</i> , <i>KLF6</i> , <i>MXD1</i> , <i>PLAUR</i>   |
| Meta-gene 80  | 7               | Regulation of RNA metabolic process (6 genes) 4.9E-2  | <i>ATF3</i> , <i>C8orf4</i> , <i>FOS</i> , <i>JUNB</i> , <i>NR4A1</i> , <i>SIK1</i> , <i>ZFP36</i>   |
| Meta-gene 71  | 8               | Gut development (3 genes) 3.5E-2  | <i>CCL11</i> , <i>CH25H</i> , <i>EDNRB</i> , <i>F2RL2</i> , <i>FOXF1</i> , <i>FOXF2</i> , <i>PCDH18</i> , <i>WNT5A</i>   |

Table 2. Subtype-specific minimal gene set as identified by Elastic net

| Subtype               | Minimal gene sets specifying a subtype  |  |
|-----------------------|---|--|
|                       | Up-regulated from population mean   | Down-regulated from population mean  |
| A. Surface crypt-like | <i>ADTRP, B3GNT7, CLCA1, MUC2, NR3C2, PADI2, RETNLB, STYK1</i>  | <i>CHI3L1, FNDC1, TIMP3, SULF1</i>   |
| B. Lower crypt-like   | <i>CCDC113, CDHR1, FARP1, GPSM2, GRM8, HNF4A, IHH, KCNK5, KIAA0226L, MYRIP, PLAGL2, PRR15, QPRT, RNF43, RPS6KA3, SLC5A6, TP53RK, TSPAN6, VAV3, YAE1D1</i> | <i>ALOX5, BASP1, CREB3L1, CXCR4, EPB41L3, FSCN1, GFPT2, GPX8, ITPRIP, KCNMA1, KCTD12, MT1E, RARRES3, RNASE1, SGK1, SOCS3</i> |
| C. CIMP-H-like        | <i>ANP32E, EGLN3, IDO1, PLK2, RAB27B, RARRES3, RPL22L1, TFAP2A</i>  | <i>ATP9A, C10orf99, CXCL14, KIAA0226L</i>  |
| D. Mesenchymal        | <i>ANK2, BOC, C7, CRYAB, DCHS1, DDR2, GEM, PRICKLE1, TAGLN</i>  | <i>HOOK1, RBM47</i>  |
| E. Mixed              | <i>CEACAM6, CXCL5, HSD11B1, IL1B, IL6, MRPS31, PI15, RAP2A, UQCC</i>  | <i>AGR3, RAB27B, REG4</i>  |

it consisted of a number of genes significantly associated with CRC. The analysis of pairwise intra-gene module correlations in normal samples of both discovery and validation set identified as cancer-specific gene modules of chromosome 20q, several immune response, EMT/stroma and GDC gene modules, homeobox genes and gut development (see Supplementary material, Figure S1D). The relationship between subtypes and meta-genes is illustrated by the heat map (Figure 1A), in which the major molecular motifs and their role in subtype definition stand out. Table S4 (see Supplementary material) contains median subtype values per meta-gene and the results of differential meta-gene expression testing between subtypes. Subtypes are not determined by individual biological components but each of them contributes to the molecular identity of the subtypes. The EMT/stroma cluster stands out in subtypes A + B (low expression) and D + E (high expression), while subtype C notably contained a high expression of immunity-associated cluster. High expression of meta-genes representing upper colon crypt cells in subtypes A and B, correlated with serrated and papillary (A) and complex tubular (B) morphological patterns (see below). Given the enterocyte-like morphology and retained polarity of the neoplastic cells in these patterns, they are considered as well differentiated. Subtype C is associated with the mucinous phenotype. Interestingly, subtypes A and C show high expression of metallothioneins, subtypes C and E show high expression of the homeobox gene module, while subtypes E and B strongly express a gene module containing the *EREG* gene (Table 1). The high expression of chromosome 20q cluster in subtype B was correlated with a significantly higher copy number gain/amplification of all of 20q in this subtype (see Supplementary material, Figure S8). The low expression of lipid synthesis genes is striking for subtype D and low expression of the gut development gene module for subtype C. A refined picture of differences is given by a quantitative comparison of (meta-)gene expression between subtype pairs (see Supplementary material, Tables S4 and S5, Figure S4). For each subtype we also identified a minimum set of characteristic genes (Table 2; for more details, see Supplementary material, Supplementary methods and results).

### Normal colon mucosa in the context of subtypes

When applied to the 64 normal samples, the LDA classifier assigned them all to subtype A, with posterior probability > 0.99, supporting the observation that A is well differentiated and closest to normal colonic epithelium in terms of gene expression pattern. For validation, we analysed five public datasets comprising 205 profiles of normal/adenoma/carcinoma samples. Most of the normal and adenoma samples were classified by LDA as subtype A (74.5% of 51 and 69.0% of 71, respectively) or subtype B (28.2% and 21.6%, respectively), confirming subtype A as the most normal-like. The 80 carcinoma samples were distributed over all subtypes (26.2% A, 30.0% B, 11.3% C, 18.7% D and 13.8% E).

### Subtypes and patient survival

We assessed whether subtypes differ in survival, as a general read-out of biological significance, and then tested the association of each meta-gene with prognosis, using the complete discovery set of 1113 patients (Figure 1B-C see also Supplementary material, Table S6). Kaplan–Meier curves for RFS, OS, SAR, hazard ratios (HRs) and *p* values of pairwise differences between subtypes are shown in Figure 1C. The results indicate that subtypes C and D are associated with poor OS. For subtype D, this is primarily due to early relapse correlated with high expression of EMT genes and low expression of proliferation-associated genes. For subtype C it is the result of short SAR, correlated with low expression of GDC, top colon crypt, *EREG* and Chr 20q genes and high expression of meta-gene 126 (see Supplementary material, Table S1). For subtype E the trend towards poorer OS and RFS was not statistically significant, although borderline significant poorer SAR was found relative to subtype B. Subtypes A and B had better prognosis than D for all three endpoints, although for OS in subtype A this was not significant.

The analysis of clinical and molecular markers (below) showed that subtype C is enriched for MSI tumours and *BRAF* mutant tumours, the latter present also in subtype D. The literature indicates that MSI is associated with better RFS, while *BRAF* mutation is an indicator of worse SAR [27]. To analyse how these two contradictory components affect survival in

Table 3. Result of additive multivariate Cox proportional hazards model, with subtype, *BRAF* mutation, MSI and stage<sup>a</sup>

| Variable     | RFS HR | <i>p</i> | OS HR | <i>p</i> | SAR HR | <i>p</i> |
|--------------|--------|----------|-------|----------|--------|----------|
| A            | 0.906  | 0.760    | 1.381 | 0.390    | 1.726  | 0.180    |
| C            | 0.940  | 0.850    | 1.560 | 0.220    | 3.675  | 0.0022*  |
| D            | 1.688  | 0.0055*  | 2.161 | 0.0011*  | 1.906  | 0.014*   |
| E            | 1.506  | 0.210    | 2.201 | 0.035*   | 2.046  | 0.075    |
| <i>BRAFm</i> | 1.633  | 0.085    | 2.472 | 0.0034*  | 3.361  | 0.00072* |
| MSI          | 0.478  | 0.044*   | 0.275 | 0.004*   | 0.356  | 0.036*   |
| Stage 3      | 0.770  | 0.190    | 0.943 | 0.820    | 1.780  | 0.062*   |

<sup>a</sup>Baseline is subtype B, MSS, *BRAF* wt and Stage 2.

\*Variables significant in the model.

Hazard ratios (HR) for relapse-free survival (RFS), overall survival (OS) and survival after relapse (SAR).

subtypes, we built a multivariate Cox proportional hazard model with subtype, stage, *BRAF* and MSI (Table 3; see also Supplementary material, Table S6). Subtype C remained significantly associated with poor SAR, even after the adjustment for *BRAF*, MSI and stage, but not with RFS. Subtypes B and D remained significantly prognostic for RFS, OS and SAR. No equivalent survival data were available for the datasets in the validation series, hence these observations could not be validated.

#### Colorectal stem cell and Wnt signatures within subtypes

We investigated the association of subtypes with Wnt [28–32], putative colon cancer stem cell (CSC) [33–35] signatures, and two signatures specific for upper and lower colon crypt compartments [36], using gene set enrichment analysis (Figure 2; see also Supplementary material, Table S7). Subtypes B and E highly expressed canonical Wnt signalling target signatures. Subtypes A and D and also normal samples, however, showed low expression of these signatures. This was in concordance with the differences in  $\beta$ -catenin nuclear immunoreactivity at the invasion front (IF; see Supplementary material, Figure S9 and Supplementary methods and results). Subtypes B and E showed the highest percentages, while subtypes A and D showed significantly lower percentages of the  $\beta$ -catenin-positive nuclei. Subtype C exhibited almost no  $\beta$ -catenin nuclear immunoreactivity at the IF. We analysed CSC signatures derived from low colon crypt compartment cells that had been identified either by a Wnt reporter construct TOP GFP or by high surface expression of *EphB2*. Subtypes D and E expressed both TOP GFP and *EphB2*-derived CSC signatures, while subtype B mainly expressed only the TOP GFP signature (Figure 2).

#### Subtypes complement clinical and molecular markers

An important goal of this study was to assess how our molecular subtypes complement known clinical variables and molecular markers. We found that MSI, *BRAF* mutation status, site, mucinous histology and expression of p53 were significantly associated with various subtypes (Figure 3), but not tumour stage,

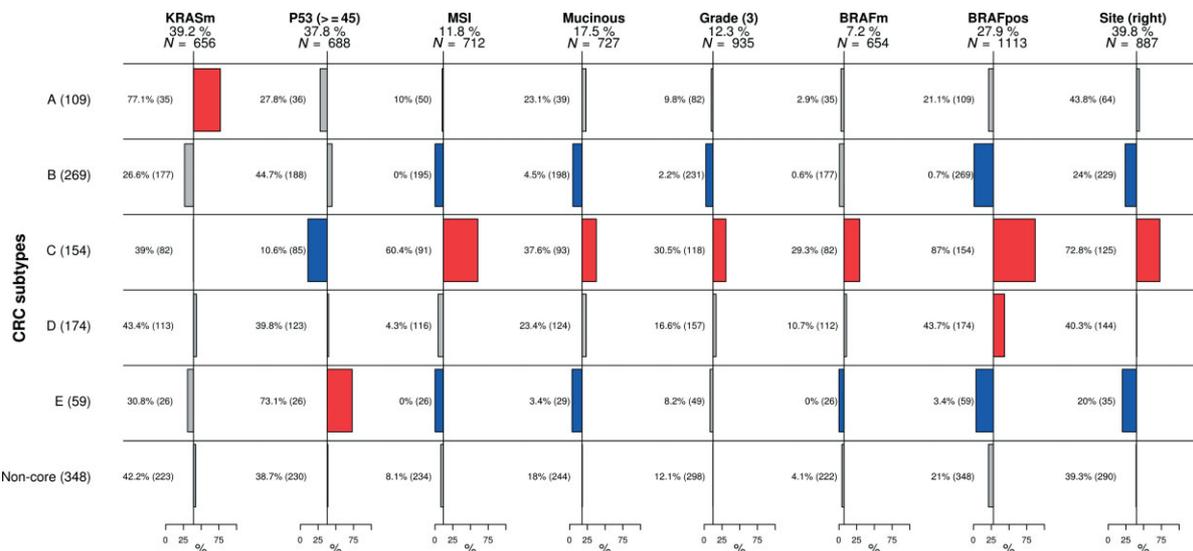
age, gender, *SMAD4* or *PIK3CA* mutations (see Supplementary material, Figure S5A). Subtype D was not significantly enriched for any of the tested variables except for the *BRAF* mutated signature and possibly represents a mixture of tumours that have the EMT/stroma signature in common. *KRAS* mutants occurred in all subtypes (see Supplementary material, Figure S5C), supporting the emerging notion that *KRAS*-mutated CRC are substantially heterogeneous [5,6,37], the oncogenic role of *KRAS* varying per specific mutation and the molecular background of the tumour in which it occurs [38]. Subtype C expressed the *BRAF* mutant signature we identified earlier [6] (87.0%), a CIMP-H signature ([11], Figure 2), and its characteristics (enrichment for MSI, right side and mucinous histology) corresponded with those of the previously reported CIMP-H phenotype [9,11,39,40] and hypermutated tumours [13]. Regarding the latter, subtype C had a similar low frequency of copy number variations (see Supplementary material, Figure S7). The distribution of MSI status, stage, age, gender, grade and site over the subtypes in the validation set followed the same patterns established in the discovery set [cf Figures 3 and S5B (see Supplementary material)]. A classification tree, trained with a combination of available clinical and molecular markers, did not identify our subtypes (see Supplementary material, Figure S5D), indicating that gene expression patterns reveal a layer of heterogeneity that goes beyond conventional CRC classification approaches.

#### Histological characteristics of subtypes

To study whether or not our molecular subtypes are associated with histological patterns, we examined haematoxylin and eosin (H&E)-stained paraffin sections of a randomly selected subset of each subtype (23, 31, 31, 29 and 19 cases for subtypes A, B, C, D and E, respectively). In attempting to match histological morphotypes to molecular subtypes, architectural patterns were used, as illustrated in Figure 4A, rather than the recognized WHO classification of CRCs [1]. Not surprisingly, given intratumour heterogeneity, none of the tumours had a single pattern. However, the prevalent patterns showed appreciable differences between the subgroups (Figure 4B, C; see also Supplementary material, Figure S6). In subtype A, the serrated pattern was most frequent, followed by the papillary pattern; in



**Figure 2.** Subtypes and biological motifs. Subtype-specific fingerprints of biological motifs, represented either as mean values of gene set enrichment scores of gene sets from corresponding gene modules (EMT/stroma, immune, secretory cells, proliferation, GDC, chromosome 20q, top of the crypt – meta105 and meta144) or composed gene set enrichment scores of particular signatures (canonical Wnt targets, CSC-TopGFP, CSC-EphB2, colon crypt bottom and CIMP-H). The gene set enrichment scores represent whether the genes from the gene set show statistically significant enrichment between the down-regulated (negative scores, light blue area) or up regulated (positive scores) genes of a given subtype; details of score calculation can be found in the Supplementary material (Supplementary methods and results and Table S7.).



**Figure 3.** Clinical and mutational characterization of subtypes. Columns represent variables and rows subtypes. Horizontal bar plots represent proportions of the corresponding variable in each of the subtypes and non-core samples. Non-core samples were tested as one group to ensure that they did not share a common characteristic that would set them apart. Numbers in brackets adjacent to subtype name represent overall number of samples in the subtype. Under the title of each variable we denote the percentage representing baseline proportion in the population, with available information, and *N* denotes the number of patients for which the information on the respective feature was available. Bars in red represent significant enrichment and bars in blue significant depletion of a feature in the subtype in comparison to baseline, at the 5% significance level. Adjacent to each bar is the percentage of samples in the subtype with the specific feature and in brackets the overall number of samples in the subtype with the information available. We can read that, for instance, subtype C, comprising 154 samples, is enriched for microsatellite-unstable (MSI) tumours, where 60.4% of 91 samples with available information are MSI.

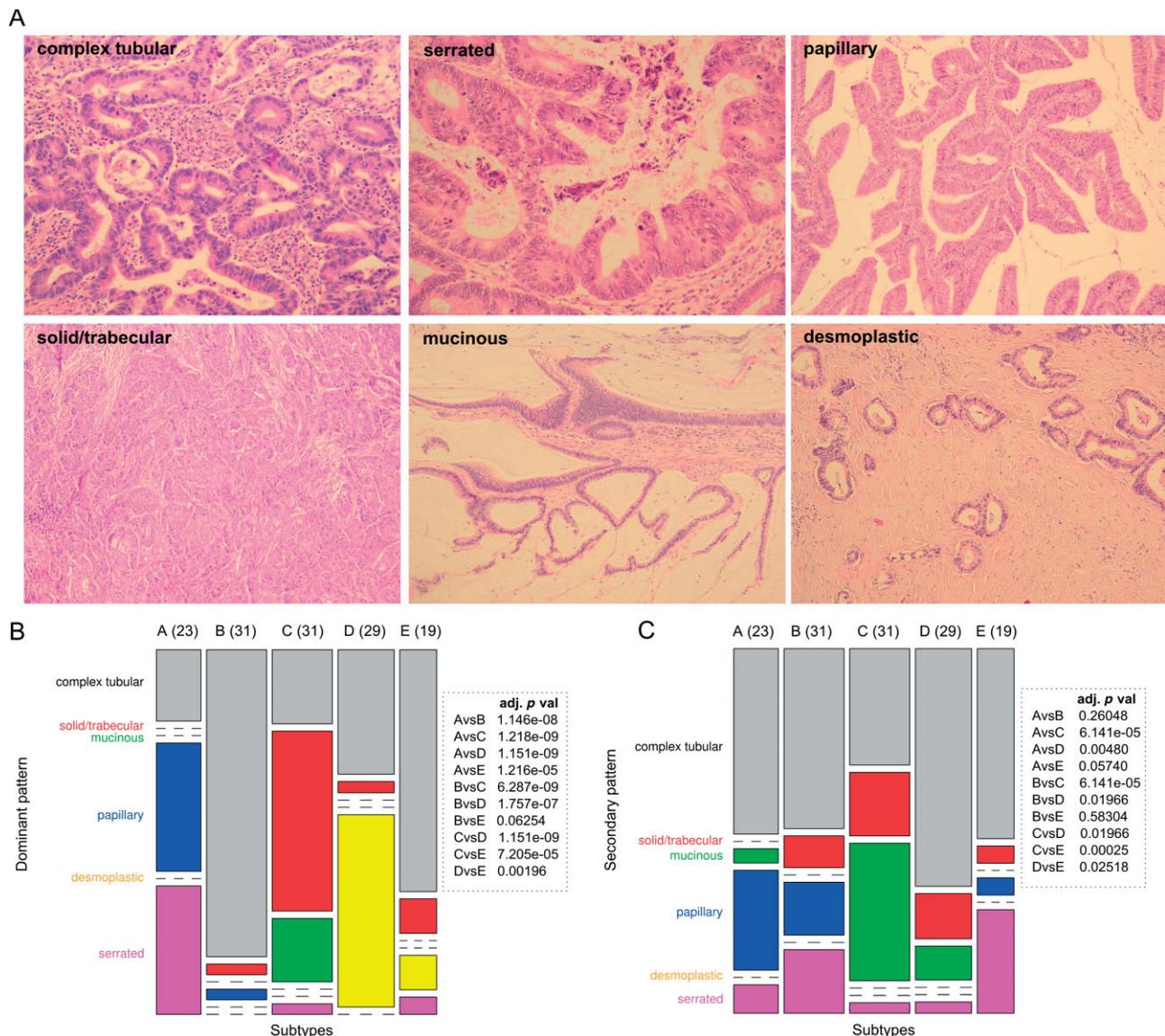


Figure 4. Morphological CRC patterns. (A) morphological CRC patterns scored in subtypes. (B, C) Distribution of dominant (B) and secondary (C) histological patterns in subtypes. Columns represent subtypes and widths are proportional to subtype frequency (numbers of samples in each subtype); rows represent dominant (B) or secondary (C) patterns and heights are proportional to pattern frequency. Boxes show adjusted  $p$  values of pairwise statistical testing of morphological pattern distribution between subtypes.

subtypes B and E, complex tubular dominated; in subtype C the solid pattern dominated, with mucinous as the second; most striking was the presence of a strong stromal reaction in subtype D.

## Discussion

Our approach, using gene modules on a large panel of samples, allowed us to identify five main CRC gene expression subtypes (Table 4). It is relevant to note that subtyping can be performed on FFPE tissues, an important prerequisite for wide clinical applications. An example is the hypermutated group identified in the TCGA study by whole exome sequencing [13], but according to our data also by gene expression profiling on routinely processed tissues (CIMP-H-like subtype).

The combination of gene expression, clinical, mutational, survival and morphological data contributes new insight into the heterogeneity of CRC. While the validation confirmed the robustness of our findings across different platforms (ALMAC versus Affymetrix), sample preparation methods (FFPE versus fresh-frozen) and dataset collections, larger datasets are necessary to assess and characterize the relevance of lower frequency subtypes (eg F, or further segregation of B into B1 and B2). Our data indicate that several major biological processes are key determinants of a complex subtype structure of CRC. Therefore our subtypes defined by gene expression do not substitute but complement groups defined by current clinicopathological variables and molecular markers. Notably, morphological subclassification of CRC has clearly reached its limits, given the often striking intratumour

Table 4. Summary of subtype characteristics

| Subtype               | CRC markers and mutations |      |      |     | Histopathology               |  | IHC                            |  | Median survival (months) |      |      |       | Clinical |   | Gene expression  |  |
|-----------------------|---------------------------|------|------|-----|------------------------------|--|--------------------------------|--|--------------------------|------|------|-------|----------|---|--|--|
|                       | MSI                       | BRAF | KRAS | P53 | Dominant                     |  | Nuclear $\beta$ -catenin at IF |  | OS                       | RFS  | SAR  | Site  | Grade    | Up-regulated  | Down-regulated   |  |
| A: Surface crypt-like | -                         | +    |      |     | Papillary or serrated        |  | -                              |  | NA                       | NA   | 28.9 |       |          | Top colon crypt, secretory cell, metallothioneins     | EMT/stroma, Wnt, CSC, Chr20q, proliferation                      |  |
| B: Lower crypt-like   | -                         | -    |      |     | Complex tubular              |  | +                              |  | NA                       | NA   | 50.4 | Left  | 2        | Top colon crypt, proliferation, Wnt                   | EMT/stroma, immune, secretory cell                               |  |
| C: CIMP-H-like        | +                         | +    | -    | -   | Solid/trabecular or mucinous |  | -                              |  | NA                       | NA   | 6.9  | Right | 3        | Proliferation, immune, metallothioneins               | GDC, top colon crypt, Chr20q                                     |  |
| D: Mesenchymal        | -                         | -    |      |     | Desmoplastic                 |  | -                              |  | NA                       | 79.5 | 19.8 |       |          | EMT/stroma, CSC, immune                               | Proliferation, secretory cell, top colon crypt, GDC, Wnt, Chr20q |  |
| E: Mixed              | -                         | -    | +    | +   | Complex tubular              |  | +                              |  | NA                       | NA   | 19.6 | Left  |          | EMT/stroma, immune, top colon crypt, Chr20q, GDC, CSC | Secretory cell   |  |

+, significantly enriched; -, significantly depleted; IF, invasion front; NA, not attained; no value, no significant enrichment in comparison to population baseline.

heterogeneity, which made us use a (primary and secondary) architectural pattern approach rather than the canonized histological subtypes (WHO). Profiling of microdissected patterns within a single tumour might reveal molecular mechanisms responsible for these morphotypes. This additional heterogeneity within the subtypes may reflect tumour polyclonality, similar to breast cancer [41]. Ultimately, aggregating clinical, pathological and further detailed molecular characteristics (including CNV, miRNA and methylation) will contribute to a more detailed perception of CRC heterogeneity and it is likely that more subtypes will emerge. This, however, would need more detailed molecular annotation of larger clinically well documented CRCs.

A striking association was found between the stromal subtype D and the EMT signature. The previously discovered EMT [12] also emerged from our analysis as the largest cluster of meta-genes associated with poor RFS (subtype D). Our histological assessment suggests that the EMT signature is the reflection of a strong mesenchymal stromal reaction, and this histological characteristic deserves to be tested for its capacity to predict resistance to therapy, in view of its strong association with poor survival. Studies requiring high tumour cell content as sample inclusion criteria (eg [13]) could miss this poor prognosis subtype. Identification of this subtype in cell lines or xenograft models is less straightforward and would benefit from the analysis of gene expression patterns between microdissected tumour and stromal cells.

EMT, however important, only partly explains CRC heterogeneity, as even subtypes with similar expression of EMT-associated genes (A–C or D–E) differ in survival, mutational, clinical and gene expression characteristics. Additional biological components, such as differentiation, immune response, proliferation, chromosome 20q or cluster of genes deregulated in CRCs, are important co-determinants that underpin a need for further subdivision of CRCs. The findings from the analysis of CSC and WNT signatures support the recently suggested hypothesis that the colon stem cell signature under the condition of silenced canonical WNT targets is associated with higher risk of recurrence (subtype D) [33]. This is consistent with subtype D showing a significantly lower percentage of  $\beta$ -catenin-positive nuclei than subtype B, with its Wnt-associated gene expression and better survival.

MSI tumours represent a subclass in most unsupervised analyses and can be recognized at the gene expression level [42]. The more recent gene expression studies [14,15] suggest that MSI and *BRAF* share distinct gene expression patterns. Subtype C was enriched for both MSI and *BRAF* mutants and had one of the best outcomes for RFS, but the worse outcome in SAR, in concordance with previously reported results [43]. Subtype C retained its poor SAR prognostic value, even in the population of MSS and *BRAF* wild-type patients. Our data suggest that subtype C represents tumours with a common biology and a gene expression pattern

that might best characterize a group of tumours resistant to chemotherapy, once metastatic. In this sense, our work not only agrees with the current known markers (*BRAF* mutation status and MSI) but clearly adds new insight, putting together these previously unrelated clusters into one biologically meaningful group. This observation is in line with recently published work [6].

Our observations show that gene expression profiling contributes substantially to our insight into CRC heterogeneity in confirming and complementing data from sequencing, CNV and promoter methylation analysis. Our subtypes can be further functionally interrogated for driving oncogenes/events by *in vitro* functional screens. High-risk subtypes D and C might contribute to therapeutic decision making in either adjuvant or metastatic settings. Retrospective analysis of clinical trial series may identify drug sensitivity associated with particular subtypes, and might open new treatment optimization strategies to be tested in clinical trials with stratified cohorts, similar to the I-SPY2 trial for breast cancer [44].

In conclusion, our unsupervised approach using gene modules resulted in the identification of distinct molecularly defined CRC subtypes, which adds a new layer of complexity to CRC heterogeneity and opens new opportunities for understanding the disease. The challenge is now to assimilate conventional and these new molecular approaches into a comprehensive consensus classification, which might then be used in further clinical studies for patient stratification and experimental studies to further elucidate mechanisms involved in the development and progression of CRC.

## Acknowledgements

We thank all the clinicians who enrolled patients and participated in the PETACC-3 trial (see Appendix at: <http://jco.ascopubs.org/content/27/19/3117.long>), in particular the coordinators D Cunningham, R Labianca and E Van Cutsem. We thank F Schütz, T Sengstag and A Wirapati for providing analytical tools for clustering and gene set enrichment analysis. This work was supported in part by the Swiss National Science Foundation (SNF; <http://www.snf.ch/E/>; Grant No. 320030\_135421, to AR and MD); Krebsforschung Schweiz (KFS; <http://www.krebsforschung.ch/>; Grant No. 02697-08-2010, to AR and MD); Fondation Medic (to EB and MD); and the Swiss National Centre for Competence in Research Molecular Biology (<http://www.nccr-oncology.ch/>; to VP and MD). ST is a Senior Clinical Investigator of the Fund for Scientific Research Flanders (FWO-Vlaanderen) and has received research grants from the Belgian National Cancer Plan (Nationaal Kankerplan). We thank the TCGA research network for the availability of data.

## Author contributions

EB and MD designed the study; YP, FTB, ST, JGH and SW conceived and carried out microarray experiments; YP and FTB performed histopathological experiments and  $\beta$ -catenin scoring; EB, VP, GD, NL and AFN analysed the data, EB, VP, ST, FTB, KOS, NL, JGH, SW, MD and AR performed data interpretation; EB, NL and KOS performed the literature search; and EB generated figures and conceived the first manuscript draft. All authors were involved in writing the paper and had final approval of the submitted and published versions.

## References

1. Bosman FT, World Health Organization, International Agency for Research on Cancer. *WHO Classification of Tumours of the Digestive System*, 4th edn. International Agency for Research on Cancer (IARC): Lyons, 2010.
2. Tejpar S, Saridaki Z, Delorenzi M, *et al.* Microsatellite instability, prognosis and drug sensitivity of stage II and III colorectal cancer: more complexity to the puzzle. *J Natl Cancer Inst* 2011; **103**: 841–844.
3. Sinicrope FA, Sargent DJ. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clin Cancer Res* 2012; **18**: 1506–1512.
4. Vecchione L, Jacobs B, Normanno N, *et al.* EGFR-targeted therapy. *Exp Cell Res* 2011; **317**: 2765–2771.
5. Martini M, Vecchione L, Siena S, *et al.* Targeted therapies: how personal should we go? *Nat Rev Clin Oncol* 2011; **9**: 87–97.
6. Popovici V, Budinska E, Tejpar S, *et al.* Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J Clin Oncol* 2012; **30**: 1288–1295.
7. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* 2007; **50**: 113–130.
8. Shen L, Toyota M, Kondo Y, *et al.* Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA* 2007; **104**: 18654–18659.
9. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. *J Mol Diagn* 2008; **10**: 13–27.
10. Furlan D, Carnevali IW, Bernasconi B, *et al.* Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. *Mod Pathol* 2011; **24**: 126–137.
11. Hinoue T, Weisenberger DJ, Lange CP, *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012; **22**: 271–282.
12. Loboda A, Nebozhyn MV, Watters JW, *et al.* EMT is the dominant program in human colon cancer. *BMC Med Genom* 2011; **4**: 9.
13. TCGA CGAN. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330–337.
14. Salazar R, Roepman P, Capella G, *et al.* Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29**: 17–24.
15. Perez Villamil B, Romera Lopez A, Hernandez Prieto S, *et al.* Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer* 2012; **12**: 260.
16. Wirapati P, Sotiriou C, Kunkel S, *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified

- understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008; **10**: R65.
17. Farmer P, Bonnefoi H, Becette V, et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005; **24**: 4660–4671.
  18. Shedden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; **14**: 822–827.
  19. Xie T, G DA, Lamb JR, et al. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PLoS One* 2012; **7**: e42001.
  20. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008; **24**: 719–720.
  21. Monti S, Tamayo P, Mesirov J, et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003; **52**: 91–118.
  22. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer: New York, 2009.
  23. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; **67**: 301–320.
  24. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3**: Article 3.
  25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
  26. Arnholt AT. BSDA: Basic statistics and data analysis. R package v 1.01, 2012; <http://cran.r-project.org/web/packages/BSDA/index.html>
  27. Tejpar S, Bertagnolli M, Bosman F, et al. Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery. *Oncologist* 2010; **15**: 390–404.
  28. Mokry M, Hatzis P, de Bruijn E, et al. Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PLoS One* 2010; **5**: e15092.
  29. Hatzis P, van der Flier LG, van Driel MA, et al. Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol* 2008; **28**: 2732–2744.
  30. Van der Flier LG, Sabates-Bellver J, Oving I, et al. The intestinal Wnt/TCF signature. *Gastroenterology* 2007; **132**: 628–632.
  31. Sansom OJ, Reed KR, Hayes AJ, et al. Loss of APC in vivo immediately perturbs Wnt signaling, differentiation, and migration. *Genes Dev* 2004; **18**: 1385–1390.
  32. Fevr T, Robine S, Louvard D, et al. Wnt/ $\beta$ -catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Mol Cell Biol* 2007; **27**: 7551–7559.
  33. de Sousa EMF, Colak S, Buikhuisen J, et al. Methylation of cancer stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. *Cell Stem Cell* 2011; **9**: 476–485.
  34. Merlos-Suarez A, Barriga FM, Jung P, et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* 2011; **8**: 511–524.
  35. van der Flier LG, van Gijn ME, Hatzis P, et al. Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* 2009; **136**: 903–912.
  36. Kosinski C, Stange DE, Xu C, et al. Indian hedgehog regulates intestinal stem cell fate through epithelial–mesenchymal interactions during development. *Gastroenterology* 2010; **139**: 893–903.
  37. Faris JE, Ryan DP. Trees, forests, and other implications of a *BRAF* mutant gene signature in patients with *BRAF* wild-type disease. *J Clin Oncol* 2012; **30**: 1255–1257.
  38. Singh A, Sweeney MF, Yu M, et al. TAK1 inhibition promotes apoptosis in *KRAS*-dependent colon cancers. *Cell* 2012; **148**: 639–650.
  39. Tanaka H, Deng G, Matsuzaki K, et al. *BRAF* mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer. *Int J Cancer* 2006; **118**: 2765–2771.
  40. Hawkins N, Norrie M, Cheong K, et al. CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. *Gastroenterology* 2002; **122**: 1376–1387.
  41. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.
  42. Tian S, Roepman P, Popovici V, et al. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol* 2012; **228**: 586–595.
  43. Dahlin AM, Palmqvist R, Henriksson ML, et al. The role of the CpG island methylator phenotype in colorectal cancer prognosis depends on microsatellite instability screening status. *Clin Cancer Res* 2010; **16**: 1845–1855.
  44. Barker AD, Sigman CC, Kelloff GJ, et al. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther* 2009; **86**: 97–100.
  45. Moskaluk CA, Zhang H, Powell SM, et al. Cdx2 protein expression in normal and malignant human tissues: an immunohistochemical survey using tissue microarrays. *Mod Pathol* 2003; **16**: 913–919.
  46. van den Brink GR, Bleuming SA, Hardwick JC, et al. Indian Hedgehog is an antagonist of Wnt signaling in colonic epithelial cell differentiation. *Nat Genet* 2004; **36**: 277–282.
  47. Liu JY, Seno H, Miletic AV, et al. Vav proteins are necessary for correct differentiation of mouse cecal and colonic enterocytes. *J Cell Sci* 2009; **122**: 324–334.
  48. Zheng H, Ying H, Wiedemeyer R, et al. *PLAGL2* regulates Wnt signaling to impede differentiation in neural stem cells and gliomas. *Cancer Cell* 2010; **17**: 497–509.
  49. Hirose H, Ishii H, Mimori K, et al. The significance of *PITX2* overexpression in human colorectal cancer. *Ann Surg Oncol* 2011; **18**: 3005–3012.
  50. Kontos CK, Papadopoulos IN, Fragoulis EG, et al. Quantitative expression analysis and prognostic significance of L-DOPA decarboxylase in colorectal adenocarcinoma. *Br J Cancer* 2010; **102**: 1384–1390.
  51. Bhatavdekar J, Patel D, Ghosh N, et al. Interrelationship of prolactin and its receptor in carcinoma of colon and rectum: a preliminary report. *J Surg Oncol* 1994; **55**: 246–249.
  52. Gaber A, Johansson M, Stenman UH, et al. High expression of tumour-associated trypsin inhibitor correlates with liver metastasis and poor prognosis in colorectal cancer. *Br J Cancer* 2009; **100**: 1540–1548.
  53. Kawakami K, Ooyama A, Ruszkiewicz A, et al. Low expression of gamma-glutamyl hydrolase mRNA in primary colorectal cancer with the CpG island methylator phenotype. *Br J Cancer* 2008; **98**: 1555–1561.
  54. Chen Y, Tang Y, Guo C, et al. Nuclear receptors in the multidrug resistance through the regulation of drug-metabolizing enzymes and drug transporters. *Biochem Pharmacol* 2012; **83**: 1112–1126.
  55. Park SW, Zhen G, Verhaeghe C, et al. The protein disulfide isomerase *AGR2* is essential for production of intestinal mucus. *Proc Natl Acad Sci USA* 2009; **106**: 6950–6955.
  56. Noah TK, Kazanjian A, Whitsett J, et al. SAM pointed domain ETS factor (*SPDEF*) regulates terminal differentiation and maturation of intestinal goblet cells. *Exp Cell Res* **316**: 452–465.

57. Steppan CM, Brown EJ, Wright CM, *et al.* A family of tissue-specific resistin-like molecules. *Proc Natl Acad Sci USA* 2001; **98**: 502–506.
58. Heiskala K, Giles-Komar J, Heiskala M, *et al.* High expression of RELP (Reg IV) in neoplastic goblet cells of appendiceal mucinous cystadenoma and pseudomyxoma peritonei. *Virchows Arch* 2006; **448**: 295–300.
59. Dalerba P, Kalisky T, Sahoo D, *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011; **29**: 1120–1127.
60. \*R Development Core Team. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.
61. \*Gentleman RC, Carey VJ, Bates DM, *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**: R80.
62. \*Therneau T. *A Package for Survival Analysis in S*. R package version 2.36–14, 2012.
63. \*Bolstad BM, Collin F, Simpson KM, *et al.* Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol* 2004; **60**: 25–58.
64. \*Venables WNR, Ripley BD. *Modern Applied Statistics with S*, 4th edn. Springer: New York, 2002.
65. \*Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Statist* 2006; **15**: 651–674.
66. \*Van Cutsem E, Labianca R, Bodoky G, *et al.* Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J Clin Oncol* 2009; **27**: 3117–3125.
67. \*Jorissen RN, Gibbs P, Christie M, *et al.* Metastasis-associated gene expression changes predict poor outcomes in patients with Duke's stage B and C colorectal cancer. *Clin Cancer Res* 2009; **15**: 7642–7651.
68. \*IGC. Expression Project for Oncology, 2008 [cited; available from: <http://www.intgen.org/expo/>]
69. \*Smith JJ, Deane NG, Wu F, *et al.* Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 2010; **138**: 958–968.
70. \*Skrzypczak M, Goryca K, Rubel T, *et al.* Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One* 2010; **5**: e13091.
71. \*Hong Y, Ho KS, Eu KW, *et al.* A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res* 2007; **13**: 1107–1114.
72. \*Gyorffy B, Molnar B, Lage H, *et al.* Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLoS One* 2009; **4**: e5645.
73. \*Galamb O, Sipos F, Solymosi N, *et al.* Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 2835–2845.
74. \*Galamb O, Spisak S, Sipos F, *et al.* Reversal of gene expression changes in the colorectal normal–adenoma pathway by NS398 selective COX2 inhibitor. *Br J Cancer* 2010; **102**: 765–773.
75. \*Koinuma K, Yamashita Y, Liu W, *et al.* Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene* 2006; **25**: 139–146.
76. \*Jorissen RN, Lipton L, Gibbs P, *et al.* DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* 2008; **14**: 8061–8069.
77. \*Grone J, Lenze D, Jurinovic V, *et al.* Molecular profiles and clinical outcome of stage UICC II colon cancer patients. *Int J Colorectal Dis* 2011; **26**: 847–858.
78. \*Birnbaum DJ, Laibe S, Ferrari A, *et al.* Expression profiles in stage II colon cancer according to APC gene status. *Transl Oncol* 2012; **5**: 72–76.
79. \*Giancarlo R, Scaturro D, Utro F. Computational cluster validation for microarray data analysis: experimental assessment of Cleft, Consensus Clustering, Figure of Merit, *Gap Statistics and Model Explorer*. *BMC Bioinform* 2008; **9**: 462.
80. \*Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: x1–13.

\*Cited only in the Supplementary material.

## SUPPLEMENTARY MATERIAL ON THE INTERNET

The following supplementary material may be found in the online version of this article:

Supplementary methods and results (contains a further table and two further figures)

**Figure S1.** (A) Consensus clustering and similarity dendrogram of samples. (B) Subtype projection in the four-dimensional space of LDA axes. (C) Heat map matrix of pairwise meta-gene Fisher Z-transformed Pearson pairwise correlations. (D) Box plots of intra gene module pairwise gene–gene Pearson correlations in normal samples in both discovery and validation sets

**Figure S2.** Validation of meta-gene expression pattern of subtypes represented by heat maps

**Figure S3.** (A) Heat map representing validation of gene expression patterns of subtypes. (B) Pairwise Fisher Z-transformed correlations of meta-genes in validation set. (C) Box plots representing medians of pairwise gene–gene Pearson correlations in the validation datasets

**Figure S4.** Expression of top five down- and top five up regulated genes from all pairwise comparisons between subtypes

**Figure S5.** (A) Other clinical and mutational markers tested and found non-significant between subtypes. (B) Clinical variables tested in the clusters of the validation test. (C) Distribution of significant clinical and mutational markers across subtypes. (D) Classification tree trained on clinical variables

**Figure S6.** Graphs of joined distribution of dominant versus secondary patterns in each of the subtypes

**Figure S7.** Heat map of CNV profiles of 154 samples from the discovery set, randomly ordered inside each of the subtypes

**Figure S8.** Result of hypothesis testing of median log-scale copy number estimates of chromosome 20 of subtype B versus all other subtypes

**Figure S9.** Distribution of  $\beta$ -catenin immunoreactivity of the invasion front counts between subtypes

**Table S1.** Detailed description of gene module members and detailed results of meta-gene expression tests pairwise between subtypes and of subtypes to meta-gene medians

**Table S2.** Multiclass linear discriminant (LDA) subtype assignment of samples from validation set

**Table S3.** Correlations of subtype-specific gene expression profiles (1 versus all moderated *t* test statistics) when accounting for subtype F in the training set

**Table S4.** Detailed results of meta-gene expression tests pairwise between subtypes and of subtypes to meta-gene medians

**Table S5.** Detailed results of pairwise comparisons of differentially expressed gene between subtypes

**Table S6.** Detailed results of Cox proportional hazards models for RFS, OS and SAR for subtype, stage, MSI and *BRAF* and for meta-genes

**Table S7.** Results of GSEA comparison of enrichment tested signatures in individual subtypes and normal samples

## 100 Years ago in the *Journal of Pathology*...

**The technique of cultivating adult animal tissues *in vitro*, and the characteristics of such cultivations**

Albert J. Walton

**Experiments on hæmolytic icterus**

J. W. M'Nee

**Congenital aneurysm in a young rabbit**

W. Henwood Harvey

**To view these articles, and more, please visit:**

**[www.thejournalofpathology.com](http://www.thejournalofpathology.com)**

Click 'ALL ISSUES (1892 - 2011)', to read articles going right back to Volume 1, Issue 1.

**The Journal of Pathology**  
***Understanding Disease***

