USE OF MULTIVARIATE ANALYSIS TO SUGGEST A NEW MOLECULAR CLASSIFICATION OF COLORECTAL CANCER

Short title: Molecular classification of colorectal cancer

Enric Domingo[1], Rajarajan Ramamoorthy[1,2], Dahmane Oukrif[3], Daniel Rosmarin[1,4], Michal Presz[4], Haitao Wang[4], Hannah Pulker[1,5], Helen Lockstone[6], Tarjei Hveem[7,8], Treena Cranston[5], Havard Danielsen[7,8], Marco Novelli[3], Brian Davidson[2], Zheng-Zhou Xu[9], Peter Molloy[9], Elaine Johnstone[4], Christopher Holmes[10], Rachel Midgley[4], David Kerr[11], Oliver Sieber[12], Ian Tomlinson[1]

1. Molecular and Population Genetics Laboratory and 6. Bioinformatics Core, Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK
2. Division of Surgery and Interventional Science, University College London, Royal Free Hospital, London NW3 2QG, UK
3. Department of Pathology, University College Hospital, London WC1E 6JJ, UK
4. Department of Oncology, University of Oxford, Old Road Campus Research Building, Oxford OX3 7DQ, UK
5. Oxford Medical Genetics Laboratories, Churchill Hospital, Oxford OX3 7LJ
7. Institute for Medical Informatics, Oslo University Hospital, Oslo, Norway
8. Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway
9. CSIRO Preventative Health Flagship, North Ryde, New South Wales, Australia
10. Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
11. Nuffield Department of Clinical Laboratory Sciences, University of Oxford, Oxford OX3 7DU, UK
12. Ludwig Colon Cancer Initiative Laboratory, Ludwig Institute for Cancer Research, Melbourne, Australia

Corresponding authors: Enric Domingo (enric@well.ox.ac.uk) and Ian Tomlinson (iant@well.ox.ac.uk), Molecular and Population Genetics, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK; Tel. +44 01865 287661, Fax: +44 01865 287501

Contributions: Experimental work by ED, RR, DO, DR, MP. HW, HP. Data analysis by ED, HL, TH, HD, CH, OS, IT. Samples and information from ZZX, PM, EJ, RM, DK. Supervision by TC, HD, MN, BD, OS, IT. Study designed and manuscript written by ED and IT.

ABSTRACT

Molecular classification of colorectal cancer (CRC) iscurrently based onmicrosatellite instability (MSI), *KRAS* or *BRAF* mutation and, occasionally, chromosomal instability (CIN). Whilst useful, these categories may not fully represent the underlying molecular sub-groups. We screened 906 stage II/III CRCs from the VICTOR clinical trial for somatic mutations. Multivariate analyses (logistic regression, clustering, Bayesian networks) identified the primary molecular associations. Positive associations occurred between: CINand *TP53*mutation;MSIand *BRAF*mutation; and*KRAS*and *PIK3CA*mutations. Negative associations occurred between: MSIand CIN; MSIand *NRAS* mutation; and *KRAS*mutation and each of *NRAS*, *TP53* and *BRAF* mutations.Some complex relationships were elucidated: *KRAS*and *TP53* mutations had both a direct negative association, and a weaker, confounding positive association *via TP53*-CIN-MSI-*BRAF-KRAS.* Our results suggesteda new molecular classification of CRCs: (1) MSI+ and/or *BRAF*-mutant; (2) CIN+ and/or *TP53*-mutant, with wildtype*KRAS* and *PIK3CA*; (3) *KRAS*-and/or *PIK3CA*-mutant, CIN*+,TP53*-wildtype; (4) *KRAS*- and/or *PIK3CA*-mutant, CIN-, *TP53*-wildtype; (5) *NRAS*-mutant; (6) no mutations; (7) others.As expected, Group 1 cancers weremostly proximal and poorly-differentiated, usually occurring in women. Unexpectedly, different types of CIN+ CRC were found: Group 2 cancers were usually distal and occurred in men, butGroup 3showed neither of these associations, but were of higher stage.By comparison, CIN+ cancers haveconventionally been associated with all three of these variables, because they have been tested *en masse*. Our classification also showed potentially improved prognostic capabilities, with Group 3, and possibly Group 1, independently predicting disease-free survival.

INTRODUCTION

Two main molecular types of colorectal carcinoma(CRC) have been described, based on the "molecular phenotypes" of chromosomal instability (CIN) and microsatellite instability (MSI or MIN). CIN is the more common and is generally detected by the presence of an abnormal chromosome complement or number (aneuploidy or polyploidy) [1]. MSI is the result of mismatch repair deficiency [2], resulting in an increased mutation rate, principally manifest as insertions and deletions in repetitive sequences. In terms of the somatic genetic pathways followed by MSI+ and CIN+ CRCs, there seems to be considerable functional overlap, but the specific mutations tend to differ: for example, MSI+ tumours tend to acquire mutations in *AXIN1*, *BRAF* and *BAX*, whereas CIN+ tumours have mutations in *APC*, *KRAS* and *TP53*[3]. In addition, CIN+ and MSI+ tumours are associated with different clinico-pathological features: the former tend to be well/moderately differentiatedand distal, and the latter poorly differentiated, proximal and more frequent in women [4].MSI is also an established marker of good prognosis. However, there remains considerable heterogeneity within the MSI+ or CIN+ groups. A third molecular phenotype, known as the CpG island methylator phenotype (CIMP), has also been described [5]. CIMP is characterised by a high degree of age-independent methylation in gene promoters and tends to overlap with MSI, in part because promoter methylation of the mismatch repair gene *MLH1* is the most usual alteration leading to MSI [6].

The very concept of somatic genetic pathways implies that some mutations are co-selected, presumably as a result of variation in the cancer cell's microenvironment, one component of which is the pre-existing mutations in that cell. However, whilst some consistent pairwise associations between mutations in CRC have been identified, it is far from clear as to which mutations are co-selected and which are secondarily associated *via* other genetic changes. We hypothesised that some of the considerable residual heterogeneity in the behaviour of CRCscould be explained by refining the established genetic pathways of tumorigenesis and by identifying new ones. However, most previous studies, including our own, have used insufficient samples and/or analysed too few genes for this objective to be achieved[7-12]. In this study, we analysed over 900 CRCs from the VICTOR clinical trial of stage II/III colorectal cancer. We profiled 11 somatic genetic alterationsand performed multivariate analysis using regression, clustering and Bayesian network approaches. This strategy allowed us to better characterise the existing pathways of colorectal tumorigenesis, to find additional, less common pathways and to propose primary molecular determinants of tumour behaviour.

MATERIALS AND METHODS

The VICTOR randomised trial of rofecoxib or placebo post-primary treatment recruited a total of 2,434 stage II/III CRC patients between 2002 and 2004 [13]. Formalin-fixed paraffin-embedded blocks were available for 965 of these patients. Haematoxylin and eosin (H&E)-stained sections were reviewed, and normal tissue and colorectal carcinoma within each section were identified. Samples from 59 patients were discarded because of lack of tumour, leaving 906 cancers (Table 1). All tumour samples were collected prior to non-surgical therapy, with the exception of 67 rectal carcinomas that had been treated with neo-adjuvant radiotherapy. Clinico-pathological variables at presentation were obtained from the trial database and treated as either binary (location (left *v* right), sex, stage (II v III)) or continuous (age, differentiation (well, moderate, poor)) as appropriate. Paired normal samples from 795 of these patients were obtained from additional blocks with only normal tissue (N=479), blood sample (N=244) or pure, clearly separate normal tissue in the same block as the tumour (N=72). For DNA extraction, all the normal blocks and carcinomas with >80% cancer cells were cut into scrolls. Other tumour blocks were cut into 10μm sections and needle-microdissected with an H&E section as a guide to ensure population purity. Tissues from scrolls and microdissections were digested with proteinase K and DNA was extracted with the DNeasy kit (Qiagen). DNA from blood was extracted with the Maxwell 16 Blood DNA Purification Kit (Promega). For ploidy analysis, all carcinoma blocks were cut into a further scroll. All research was performed according to the tenets of the Declaration of Helsinki and approved by Oxfordshire Research Ethics Committee B 05\Q1605\66.

Mutation screening, microsatellite instability, LOH and CIN analysis used standard, previously-described methods. These and the statistical methods are described in the Supplementary Data.

RESULTS

Baseline analysis of molecular associations

A set of 906 stage II/III cancers colorectal cancers from the VICTOR trial (Table 1) was analysed for CIN, MSI and almost all of the most common somatic mutations in colorectal cancer (*KRAS*, *NRAS*, *BRAF*, *PIK3CA*, *TP53* and *FBXW7*/*CDC4*). LOH analysis at 3 sites was performed in 795 tumours from which constitutional DNA was available, targeting chromosomes 5q near *APC*, 17p near *TP53* and 18q near *SMAD4*. A summary of the molecular findings is shown in Table 2. Overall, the frequencies of molecular alterations, the mutation spectra (Supplementary Figure 1) and the pairwise associations (Supplementary Tables 1 and 2), were in good agreement with those previously established in the literature[4, 6, 12, 14-19], and we shall not consider them further here.A new findings was that *FBXW7/CDC4* mutations were not associated with any other mutation or clinico-pathological variable. We also examined stage-specific associations, the only significant results being that Stage III cancers tended to be CIN+.

180 (21% in total) CRCs were "double-negative" (MSI-CIN-). In general, the double-negative tumours resembled MSI-CIN+ cancers more than MSI+CIN- cancers, and we therefore compared MSI-CIN- with MSI-CIN+ tumours. The MSI-CIN- cancers presented at an earlier stage (102/180 *v* 244/557 Stage II respectively, p=0.003, q=0.01), had lower frequencies of *TP53* mutation and 17p LOH, and showed a borderline association with *KRAS* mutation (Supplementary Tables2 and 3). Logistic regression-based multivariate analysis showed that the MSI-CIN- tumours remained associated only with lower stage and lack of *TP53* mutation (Supplementary Table 3).

Twenty-three cancers (3%) were "double-positive" (MSI+CIN+). Compared with all other cancers, the double-positive tumours tended to be right-sided (14/22 *v* 278/804, p=0.007, q=0.03). No molecular alteration, including *TP53* mutation and 17p LOH, was significantly associated with this small group of tumours, although they had a relatively high frequency of *BRAF* mutation (Supplementary Table 2). Overall, double-positive cancers appeared to resemble MSI+CIN- cancers most closely.

Searching for the primary inter-molecular associations

In order to define new genetic pathways of colorectal carcinogenesis, we sought evidence for the primary drivers of the associations between somatic mutations (*KRAS*, *NRAS*, *BRAF*, *PIK3CA*, *TP53*, *FBXW7*), CIN, MSI and the 3 sites of LOH (Supplementary Tables 2 and3). Approximately half of the

associations found by pairwise analysis were no longer significant, suggesting that they were secondary to a primary association.

Since logistic regression analysis to identify the primary determinants of associations can be sensitive to small, chance associations or missing data when multiple highly correlated events occur, we additionally performed a Bayesian network analysis to detect primary associations among selected molecular variables (*KRAS, BRAF, NRAS, PIK3CA, TP53,* MSI, CIN) that had been successfully typed in the full data set. *FBXW7* was excluded owing to its lack of any association and the LOH events were omitted owing to their strong associations with each other, CINand *TP53* mutation. The network analysis took the form of a probabilistic model that represents the conditional dependencies between random variables via a directed acyclic graph. We found that the primary positive associations were between (i) CIN and *TP53,* (ii) MSI and *BRAF,* and (iii) *KRAS* and *PIK3CA.* The primary negative associations were between MSI and both CIN and *NRAS*, and between *KRAS* and each of *BRAF, NRAS* and *TP53.* These associations are shownin a simple, graphical form in Figure 1A*.* In almost all of these cases, the network analysis found the same primary associations as the logistic regression analysis (Figure 1A, Supplementary Table 3). Exceptions were the associations between *PIK3CA* mutation and MSI and between *PIK3CA* mutation and *TP53* that were only present in the logistic regression analysis (Figure 1A, Supplementary Table 3).

The network analysis additionally detected associations that were formally absent from the logistic regression, because those variables were dropped from the latter owing to excessive co-variation. A case in point was the primary negative associations between *NRAS* mutation and both *KRAS* mutation and MSI. The network analysis also showed the underlying reasons why some associations that were significant in the pairwise analysis were no longer significant in the multivariate logistic regression analysis. Examples included the negative associations between CIN and *BRAF* (which was secondary to associations with MSI) and between MSI and *TP53* (which was secondary to associations with CIN) (Figure 1A).

Of particular interest in the network analysis was the detection of association loops (Figure 1A). The loops between MSI and *NRAS* suggested two independent negative associations, one direct and the other indirect *via* CIN, *TP53* and *KRAS.* Most intriguing was the loop between MSI and *KRAS*. Here, there existed not only a direct negative association between *TP53* and *KRAS* mutations, but also a weaker, indirect positive association via CIN, MSI, *KRAS* and *BRAF*. The positive association found by multivariate analysis was consistent with data in the literature from pairwise association

testing, but the direct negative association was less well supported. We therefore performed a meta-analysis of other published studies [7, 9, 17, 20-25]. and this confirmed an overall pairwise negative *KRAS*-*TP53* mutation association (Supplementary Table 4).

The logistic regression and network analyses had provided strong evidence to show whichof the reported inter-molecular associations in CRC were primary and which were secondary (indirect). We undertook one further analysis to support our findings by performing unsupervised hierarchical clustering of the same alterations used in the network analysis. The cluster analysis fully supported the other two methods, indicating the four basic groups as: *KRAS* and/or *PIK3CA*; CIN and/or *TP53*; MSI and/or *BRAF*; and *NRAS* (Figure 1B, vertical axis). Clustering of the tumours (Figure 1B, horizontal axis) showed them to be separated in two main groups where the main discriminating molecular variable was *KRAS* mutation: all the mutants were present in the second cluster and none of the first cluster were mutant. Most *NRAS* and *PIK3CA* mutants were in the first and second cluster respectively. CIN and *TP53* mutations were found in both clusters but overrepresented in the first one. MSI and *BRAF* mutations were present as sub-clusters within both of the two main clusters.

Identifying groups of colorectal cancers based on shared molecular changes

We sought to identify groups of CRCs that would form the basis of our proposed molecular classification.The data suggested that this classification should be based not only on MSI, but also on *NRAS* mutation, on the negative association between *TP53* and *KRAS* mutations, and on CIN (Figure 1, Supplementary Table3). Based on the primary positive and negative inter-molecular associations described in the previous section, we chose CRC groupscharacterised by:

(1) MSI and/or *BRAF* mutation;

(2) CIN and/or *TP53* mutation, with wildtype*KRAS* and *PIK3CA*;

(3) *KRAS* and/or *PIK3CA* mutation with CIN,but without *TP53* mutation;

(4) *KRAS* and/or *PIK3CA* mutation without CIN or *TP53* mutation; and

(5) *NRAS* mutation.

These groups encompassed over 80% of all the CRCs studied. In addition,we proposed two further groups: Group 6with no detectable mutations; and a “miscellaneous” Group 7. Rationale for the grouping is provided in Figure 2.

Association between CRC groups and clinico-pathological variables

Sinceit wehoped that the 7 CRC groups would have clinical relevance, we used multiple logistic regression analysis to testthe groups for associations with clinico-pathological variables (gender, age, tumour location, stage, grade, trial randomisation arm, and treatment with chemotherapy or radiotherapy). Each group was tested in turn against all others, incorporating all variables in a reverse stepwise analysis (Table 3). Group 1 essentially included the MSI+ group of tumours and, as expected, these cancers were strongly associated with proximal location, poor differentiation, and female gender.

Most CIN+ CRCs fell into Groups 2 and 3. Although CIN is classically associated with distal location, male gender and higher stage, our two main sets of CIN+ cancers (Groups 2 and 3) showed distinct differences in their associations (Table 3). Group 2 cancers showed a strong tendency to be distally located and to occur in men, but had no association with stage. Group 3 cancers, on the other hand, were not associated with gender or location, but tended to be stage III tumours. Group 6 cancers were predominantly found in men, for reasons that are unclear. The other three Groups showed no association with the clinico-pathological variables, although we note that some groups, such as Group 5 (*NRAS*-mutants), were small.

Association between cancer groups and disease-free survival

We tested whether our molecular groups of CRC were independent predictors of five-year disease-free survival in the VICTOR study. All analyses were conditioned on the full set of clinico-pathological variables. We initially analysed Group as a categorical variable in the regression model, and then tested each group against all other groups. Very similar results were obtained in each case. We found that Group 3 – *KRAS*- and/or *PIK3CA*-mutant; MSI-; CIN+; *TP53*-wildtype – patients had poor survival, whereas all the other cancer groups showed no significant differences in survival (Hazard Ratio=1.59, 95% CI 1.13-2.24, P=0.008, Cox proportional hazards, Group 3 *v* all other groups; Figure 3).The only other independent predictor of survival was stage (Hazard Ratio=1.99, P=9.0×10^{-6}, stage III *v* stage II. Although Group 1 was not an independent predictor of survival, we wondered whether this negative result was caused by some *BRAF*-mutant, MSI- cancers having poor survival[26]. We therefore included MSI status in the survival model, since this variable has consistently been associated with better outcome in several studies. Group 3 (Hazard Ratio=1.48, 95% CI 1.05-2.09,

P=0.027) and stage (Hazard Ratio=1.98, P=1.2x10^{-5}) remained independent predictors of poor prognosis, although MSI was a borderline significant indicator of good prognosis (Hazard Ratio=0.59, P=0.054).

DISCUSSION

Colorectal carcinogenesis follows a multistep model in which sequential molecular alterations occur throughout tumour progression [27]. Here, we have analysed most of the common somatic mutations in a large CRC patient set with high-quality clinical trial data. We have undertaken multivariate regression, cluster and Bayesian network analyses that have consistently identified the primary positive and negative associations between molecular changes, thus showing other associations to be indirect. The identification of loops in the pairwise association relationships (Figure 1A) was of particular interest. One such loopprovidedthe basis for explaining the previously-postulated negative association between *TP53* and *KRAS* mutation, that we have now confirmed [7-9, 17, 20-25, 28]. One reason why the *KRAS*-*TP53* association had not found wide acceptance is that it has seemed contrary to the well-established indirect association linking *TP53*-CIN-MSI-*BRAF*-*KRAS*. Our data show that there is indeed an indirect positive association between *TP53* and *KRAS* mutations that follows this route, but that this is outweighed by a direct negative association. One potential explanation for the negative association lies in the transcription of genes such as *CDKN1A* (p21) by mutant K-ras through p53-dependent and -independent mechanisms. Effects on cell cycle arrest, senescence and apoptosis might sometimes be sub-optimal for tumour growth if both genes are mutated.

We confirmed that a small group of double-positive (MSI+CIN+) CRCs exists, as does a larger group of double-negative (MSI-CIN-) cancers. The latter were generally similar to MSI-CIN+ lesions, but had a much lower frequency of *TP53* mutations and lower stage, although relatively high frequencies of *KRAS* and *PIK3CA* mutations. Whilst these findings support *TP53* mutations in some way causing or being permissive for CIN, we found no quantitative differences in ploidy between *TP53*-mutant and *TP53*-wildtype CIN+ tumours (data not shown), suggesting that an alternative (epi)mutation to *TP53* inactivation may exist in the latter case.

Uniquely among the molecular changes, *FBXW7/CDC4* mutations occurred randomly in CRC irrespective of clinico-pathological or other molecular features. Very few other studies have

addressed the issue of *FBXW7*'s position in the pathways of colorectal tumorigenesis, apart from functional studies linking its mutation to CIN [29], but we find no evidence of an association with CIN here. One possible explanation for the absence of associations with *FBXW7* is that it mutates early in tumorigenesis, consistent with two studies that have found *FBXW7* mutations in colorectal adenomas [30, 31]. Fitting (epi)mutations into genetic pathways is likely to become increasingly difficult as next-generation sequencing discovers more low-frequency mutations like *FBXW7* that drive tumorigenesis but do not obviously fit into any specific molecular pathway.

We have proposed a molecular classification of CRC into 7groups (Figure 2) based solely on primary positive and negative associations between molecular changes (MSI, CIN, *TP53* and the type of Ras pathway mutation). This classification essentially retained the MSI+ set of cancers as a separate Group, but with the addition of *BRAF*-mutant MSI- cancers. We remain open-minded as to whether the latter cancers are distinct from their MSI+ counterparts, as some have suggested. Our classification split the CIN+ cancers into two groups and established a further *KRAS*- and/or *PIK3CA*-mutant, CIN-, *TP53*-wildtype group. Despite our classification being entirely independent of clinico-pathological variables, it showed interesting associations with clinical variables. Our data suggested that the often-reported associations between CIN and gender, tumour location, and stage actually comprised two separate associations: (i) between gender, distal location and Group 2 tumours (CIN+ and/or *TP53*- mutant with wildtype*KRAS* and *PIK3CA*); and (ii) between stage III and Group 3 tumours (*KRAS*- and/or *PIK3CA*-mutant, CIN+, *TP53*-wildtype). These findings cannot readily be explained by differences in sample size and hence statistical power (Table 3). Group 3, moreover, was an independentpredictor of survival, having worse prognosis than any of the other groups. Interestingly, *TP53* mutation is not an established prognostic marker in CRC, despite its strong association with stage, and our finding that CIN+ *TP53*-wildtype cancers have poorest prognosis is consistent with this. Furthermore, Hutchins *et al* [30]had previously found poorer disease-free survival in *KRAS*-mutant stage II/III CRCs from the QUASAR trial, although they only performed a univariate analysis. There was additionally some evidence in our data that MSI might additionally be an independent survival predictor.

Overall, our results show that sufficiently large and homogeneous sample sets and methods based on multivariate and cluster analysis can allow the genetic pathways of cancer to be teased apartfor a relatively well-characterised tumour such as CRC.However, this is a complex task that may become even more difficult as next-generation sequencing discovers more low-frequency mutations like *FBXW7* that drive tumorigenesis but do not obviously fit into any specific molecular

pathway.However, such studies may provide additional insights through different analyses, one example being the identification of hypermutant, yet MSI-, CRCs in a recent large, landmark exome sequencing study [31]. Our proposed groups of CRC requirereplication and refinement, but our data already suggest that a finer-scale molecular classification of CRCs is both possible and desirable, and we expect that a more complex, validated classification of CRC will emerge gradually in the next few years.

FUNDING

This work was supported by the Oxford Partnership Comprehensive Biomedical Research Centre with funding from the Department of Health's NIHR Biomedical Research Centres funding scheme and by Cancer Research UK. The Wellcome Trust Centre for Human Genetics receives core funding from Wellcome Trust grant 090532/Z/09/Z. DO is supported by the UCL Experimental Cancer Medicine Centre which is funded by Cancer Research UK and the NIHR. None of the funders had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

None of the authors declares any conflict of interest.

ACKNOWLEDGMENTS

We thank all individuals who participated in this study and colleagues who undertook sample collection. We are grateful to Mike Churchman for technical advice.

Word count: 3087

TABLES

Table 1. Clinico-pathological profiles of the tumours used in the study.

	Number of tumours (%)
Stage	
II	445 (49)
III	461 (51)
Gender	
Male	581 (64)
Female	325 (36)
Site	
Proximal/Right	306 (35)
Distal/Left	577 (65)
No info	23
Age (years)	
Mean (SD)	64.1 (10)
Median (Range)	64.6 (24.6-86.3)
Differentiation	
Well	73 (8)
Moderate	725 (82)
Poor	88 (10)
No info	20

Table 2. Overall frequencies of the molecular alterations analysed.

Molecular alteration	**Alteration frequency (%)**
CIN	586/861 (68)
MSI	119/892 (13)
KRAS	304/898 (34)
BRAF	91/903 (10)
NRAS	32/869 (4)
PIK3CA	104/896 (12)
TP53	329/753 (44)
FBXW7 (CDC4)	35/750 (5)
5q LOH	160/566 (28)
18q LOH	334/694 (48)
17p LOH	344/666 (52)

Table 3. Associations between proposed molecular groups of colorectal cancers and clinico-pathological variables.

The ods ratios (ORs) and P values are derived from unconditional logistic regression (Group N v all other Groups except Miscellaneous) and are shown for all variables with an association. Clinico-pathological variables with no association (P>0.05) with any Group are omitted. Significant associations are shown in bold. WT=wildtype, WD=well differentiated, MD=moderately differentiated, PD-poorly differentiated, F=female, M=male. Group 7 is not tested.

Outcome variable	**Description of Group**	**Stage (II v III)**	**Location (proximal v distal)**	**Differentiation (grade, WD, MD, PD)**	**Gender (F v M)**
Group 1	MSI and/or BRAF-mutant	OR=0.74, P=0.12	**OR=0.18, P=3x10^{-18}**	**OR=2.08, P=0.001**	**OR=2.06, P=2x10^{-4}**
Group 2	CIN+ and/or TP53- mutant with WT KRAS and PIK3CA	OR=0.95, P=0.74	**OR=3.09, p=4x10^{-11}**	OR=1.03, P=0.85	**OR=0.62, P=0.003**
Group 3	KRAS- and/or PIK3CA-mutant, CIN+, TP53-WT	**OR=1.58, P=0.02**	OR=0.98, P=0.91	OR=0.74, P=0.21	OR=1.22, P=0.3
Group 4	KRAS- and/or PIK3CA-mutant, CIN-, TP53-WT	OR=0.85, P=0.53	OR=1.23, P=0.44	OR=0.86, P=0.6	OR=0.76, P=0.29
Group 5	NRAS-mutant	OR=1.03, P=0.93	OR=0.96, P=0.93	OR=0.89, P=0.78	OR=0.93, P=0.86
Group 6	No mutations	OR=1.04, P=0.8	OR=1.11, P=0.56	OR=0.69, P=0.07	**OR=0.24, P=0.003**

FIGURE LEGENDS

Figure 1.Multivariate analysis of 7 molecular alterations in 705 tumours.

A. Bayesian network analysis. Edges represent conditional dependencies and nodes that are not connected represent variables which are conditionally independent of each other. Positive and negative associations have been marked as + and – respectively. Note that the direction of the arrows is not an indication of causality. **B.** Unsupervised hierarchical cluster analysis by tumour (horizontal) and mutation (vertical). Samples with and without alterations are marked in red and blue respectively.

Figure 2. Proposed molecular classification of colorectal cancers.

The proposed grouping is based on the following process. Initially, we utilised the near-complete lack of overlap between MSI+ and CIN+ cancers to identify 2 groups. Owing to the observed strong, primary associations, we then provisionally added *BRAF*-mutant tumours to the MSI+ groupand *TP53*-mutant cancers to the CIN+ group. We retained the MSI+ and/or *BRAF*-mutant tumour group, irrespective of other genetic changes, since *TP53* and *KNRAS* mutations were uncommon in these cancers. We then formed a group of *NRAS*-mutant tumours irrespective of their other genetic changes, since *NRAS* mutations were not positively associated with any other molecular variable. We next provisionally added a *KRAS*- and/or *PIK3CA*-mutant (but *TP53*-wildtype) group, owing to the negative association between *KRAS* and *TP53.* However, since we found no negative association between *KRAS* and CIN, we sub-divided the *KRAS*- and/or *PIK3CA*-mutant group into a CIN+ and CIN- group, leaving the great majority of *TP53*-mutant cancers in a CIN+, *KRAS*-wildtype and *PIK3CA*-wildtype group. In total, this classification encompasses 736/906 (81%) cancers. In addition, 78 double-negative cancers had no detected changes in *KRAS, PIK3CA, NRAS, TP53* or *BRAF*. The remaining 92 cancers had a variety of “atypical” mutation combinations.

Figure 3. Kaplan-Meier survival curves for disease-free survival (yrs) of the proposed colorectal cancer groups.

The 5 mutant CRC groups analysed are numbered as shown in the text. Group 3 is KRAS-and/or PIK3CA-mutant MSI- CIN+ TP53wildtypeBRAFwildtype. A sixth group with no detected mutations is also shown. The seventh residual “miscellaneous” group was not analysed. Note the inferior survival only of Group 3.

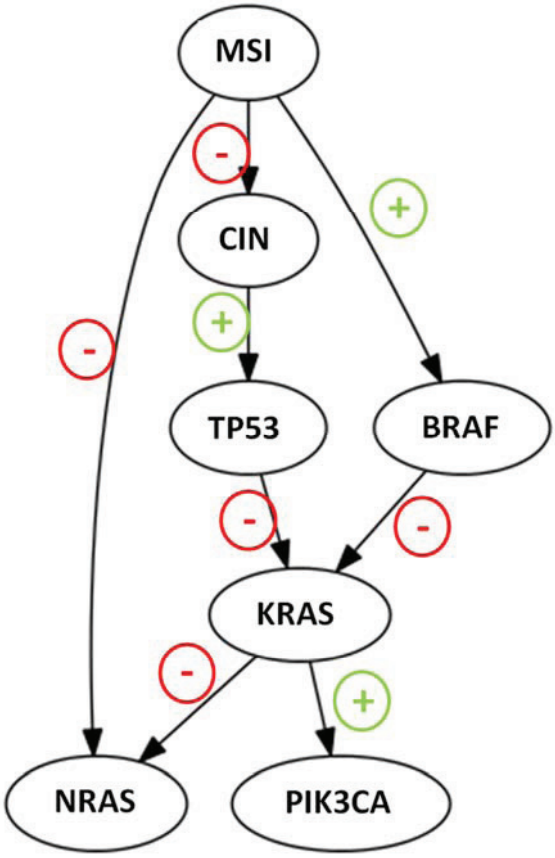
REFERENCES

1. Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. *Nature* 1997; **386**: 623-627.
2. Ionov Y, Peinado MA, Malkhosyan S, *et al*. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* 1993; **363**: 558-561.
3. Walther A, Johnstone E, Swanton C, *et al*. Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer* 2009; **9**: 489-499.
4. Sinicrope FA, Rego RL, Halling KC, *et al.* Prognostic impact of microsatellite instability and DNA ploidy in human colon carcinoma patients. *Gastroenterology* 2006; **131**: 729-737.
5. Toyota M, Ahuja N, Ohe-Toyota M, *et al*. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci USA* 1999; **96**: 8681-8686.
6. Weisenberger DJ, Siegmund KD, Campan M, *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *NatGenet* 2006; **38**: 787-793.
7. Peinado M, Fernandezrenart M, Capella G, *et al.* Mutations in the p53 suppressor gene do not correlate with C-k-ras oncogene mutations in colorectal-cancer. *Int J Oncol* 1993; **2**: 123-134.
8. Smith G, Carey FA, Beattie J, *et al.* Mutations in APC, Kirsten-ras, and p53 - alternative genetic pathways to colorectal cancer. *Proc Natl Acad Sci USA* 2002; **99**: 9433-9438.
9. Iacopetta B. TP53 mutation in colorectal cancer. *Hum Mut* 2003; **21**: 271-276.
10. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* 2007; **50**: 113-130.
11. Ostwald C, Linnebacher M, Weirich V, Prall F. Chromosomally and microsatellite stable colorectal carcinomas without the CpG island methylator phenotype in a molecular classification. *Int J Oncol* 2009; **35**: 321-327.
12. Rowan A, Halford S, Gaasenbeek M, *et al.* Refining molecular analysis in the pathways of colorectal carcinogenesis. *Clin Gastro Hepatol* 2005; **3**: 1115-1123.
13. Midgley RS, McConkey CC, Johnstone EC, *et al.* Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *J Clin Oncol* 2010; **28**: 4575-4580.
14. Samuels Y, Wang Z, Bardelli A, *et al.* High frequency of mutations of the PIK3CA gene in human cancers. *Science* 2004; **304**: 554.

15. Segditsas S, Sieber OM, Rowan A, *et al.* Promoter hypermethylation leads to decreased APC mRNA expression in familial polyposis and sporadic colorectal tumours, but does not substitute for truncating mutations. *Exper Mol Pathol* 2008; **85**: 201-206.
16. Vaughn CP, Zobell SD, Furtado LV, *et al*. Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer. *Genes Chrom Cancer*; **50**: 307-312.
17. Wood LD, Parsons DW, Jones S, *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* 2007; **318**: 1108-1113.
18. Esteller M, Hamilton SR, Burger PC, *et al*. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer Res* 1999; **59**: 793-797.
19. Jones AM, Douglas EJ, Halford SE, *et al.* Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene* 2005; **24**: 118-129.
20. Calistri D, Rengucci C, Seymour I, *et al.* Mutation analysis of p53, K-ras, and BRAF genes in colorectal cancer progression. *J Cell Physiol* 2005; **204**: 484-488.
21. De Bruijn MT, Raats DAE, Tol J, *et al.* Combined KRAS and TP53 mutation status is not predictive in CAPOX-treated metastatic colorectal cancer. *Anticancer Res*; **31**: 1379-1385.
22. Giaretti W, Venesio T, Sciutto A, *et al*. Near-diploid and near-triploid human sporadic colorectal adenocarcinomas differ for KRAS2 and TP53 mutational status. *Genes Chrom Cancer* 2003; **37**: 207-213.
23. Reid JF, Gariboldi M, Sokolova V, *et al.* Integrative approach for prioritizing cancer genes in sporadic colon cancer. *Genes Chrom Cancer* 2009; **48**: 953-962.
24. Samowitz WS, Slattery ML, Sweeney C, *et al*. APC mutations and other genetic and epigenetic changes in colon cancer. *Mol Can Res* 2007; **5**: 165-170.
25. Suehiro Y, Wong CW, Chirieac LR, *et al.* Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. *Clin Can Res* 2008; **14**: 2560-2569.
26. Samowitz WS, Sweeney C, Herrick J, *et al.* Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. *Cancer Res* 2005; **65**: 6063-6069.
27. Vogelstein B, Fearon ER, Hamilton SR, *et al.* Genetic alterations during colorectal tumor development. *New Eng J Med* 1988; **319**: 525-532.
28. Bass AJ, Lawrence MS, Brace LE, *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *NatGenet*; **43**: 964-968.
29. Rajagopalan H, Jallepalli PV, Rago C, *et al.* Inactivation of hCDC4 can cause chromosomal instability. *Nature* 2004; **428**: 77-81.

30. Hutchins G, Southward K, Handley K, *et al.* Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *J Clin Oncol* 2011; **29**: 1261-1270.

31. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330-337.


MSI
-
CIN
+
+
-
TP53
BRAF
-
-
KRAS
-
+
NRAS
PIK3CA

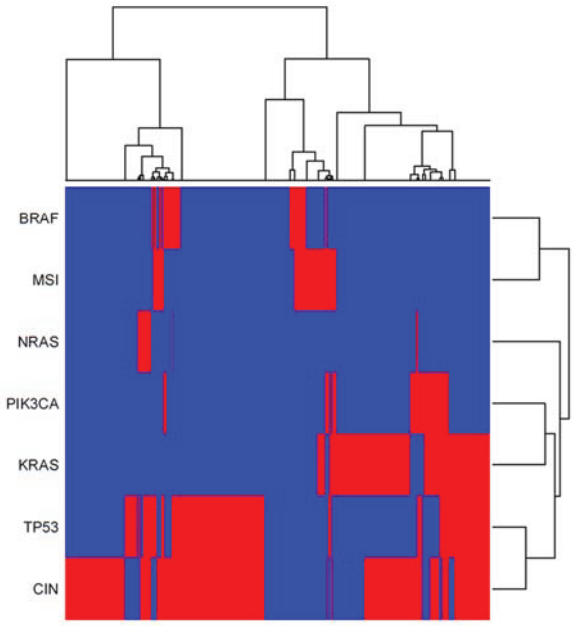
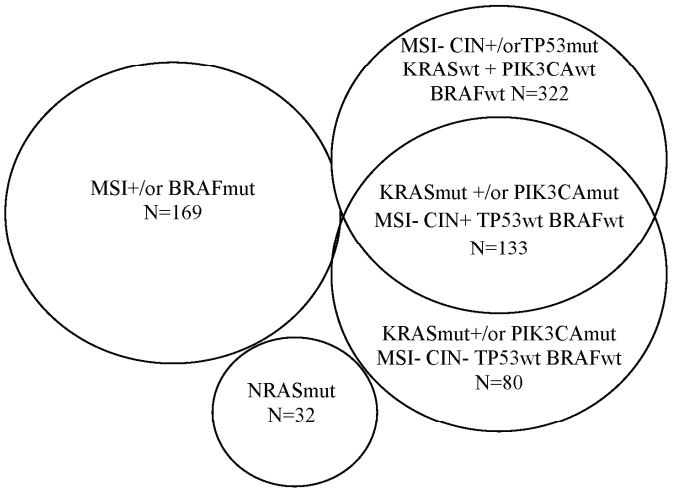
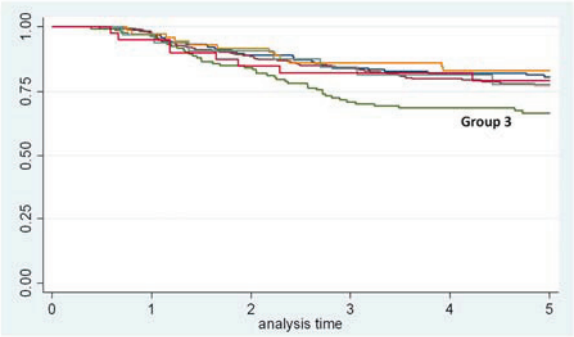

BRAF
MSI
NRAS
PIK3CA
KRAS
TP53
CIN

Figure 2.


MSI- CIN+/orTP53mut
KRASwt + PIK3CAwt
BRAFwt N=322
MSI+/or BRAFmut
N=169
KRASmut +/or PIK3CAmut
MSI- CIN+ TP53wt BRAFwt
N=133
KRASmut+/or PIK3CAmut
MSI- CIN- TP53wt BRAFwt
N=80
NRASmut
N=32


1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
analysis time
Group 3