Supplementary Data

Supplementary Methods

Readout of molecular subtype classification

The three intrinsic molecular subtype classes were identified in a previous study by unsupervised analysis of full genome expression data on a cohort of 188 primary CRC.[1] It has been reported that a hierarchical method is not optimally suited for molecular subtype classification,[2] but requires a single sample based predictor or classifier (SSP) for reliable readout, especially on independent samples. Similarly as has been done for breast cancer subtype classification, e.g. BluePrint and PAM50,[3,4] we have developed such a classifier for accurate diagnostic readout of the subtypes in a CLIA/CAP certified laboratory.

Custom made Agilent full genome DNA microarrays were used for gene expression readout of the development cohort, while dedicated diagnostic arrays containing subtype related gene and normalization probes were used for the validation cohort. Molecular subtype classification by the developed SSP is based on three distinct gene expression signatures representative for the molecular subtypes. These three distinct gene signature were identified using the original hierarchical clusters on the 188 tumor samples[1] and by performing a pair-wise Student T-test analysis for the most differentially expression genes between the three clusters (i.e. A-type versus B-and C-types; B-type versus A- and C-types; and C-type versus A- and B-types). This T-test analysis was performed within a 10-fold cross validation procedure and genes were ranked according to their Student's T-test statistics. The top-ranked gene lists across the multiple cross validation iterations were combined, resulting in a final gene selection of genes commonly (>50%) used in the cross

validation iterations: 32 genes for A-type, 53 genes for B-type and 102 genes for C-type (**Table 2**)

Next, we have used these three gene signature to develop a centroid based SSP that scores each sample for its association with the three subtypes. Normalized gene expression levels of individual samples were compared to each of subtype specific gene centroids (Pearson correlation) and used to determine the most representative subtype (A-, B- or C-type). This final verdict of this 3-way classification system was based on the highest correlation score using the formula below was derived by optimizing the SSP classification accuracy against the original hierarchical clustering (optimal accuracy was 97%). The developed classification methods for the CRC molecular subtypes is very similar to the one used in the BluePrint breast cancer classification model.[3]

SubtypeClass = Max (*ScoreA*, *ScoreB*, *ScoreC*)

in which

ScoreA = $2^* r(S_A, C_A) – r(S_B, C_B) – r(S_C, C_C) – 1.3$

ScoreB = $2^* r(S_B, C_B) – 2^* r(S_A, C_A) – r(S_C, C_C) + 1,8$

ScoreC = $2^* r(S_C, C_C) – 2^* r(S_A, C_A) – r(S_B, C_B) – 0.5$

S_A = sample gene profile of A-type genes

S_B = sample gene profile of B-type genes

S_C = sample gene profile of C-type genes

C_A = A-type centroid (‘core’ A-type signature)

C_B = B-type centroid (‘core’ B-type signature)

C_C = C-type centroid (‘core’ C-type signature)

Mutational analysis (*BRAF, KRAS* and *PIK3CA*)

Mutations in *BRAF* V600, *KRAS* codons 12, 13 and 61, and *PIK3CA* exons 9 and 20 were assessed in cDNA by Sanger sequencing of PCR products using primers with M13 tails after RT-PCR (ServiceXS BV). V600E *BRAF* mutation were analyzed after amplification of exon 15 using primers 5'-tgatcaaacttatagatattgcacga (upstream) and 5'-tcatacagaacaattccaaatgc (downstream). *KRAS* whole coding region was analyzed using primers 5'-aggcctgctgaaaatgactg (upstream) and 5'-tggtgaatatcttcaaatgatttagt (downstream). For *PIK3CA* the primers used were 5'-ccacgcaggactgagtaaca (upstream) and 5'-ggccaatcttttacccaagca (downstream) for exon 9, and 5'-tgagcaagaggctttggagt (uptstream) and 5'-agtgtggaatccagagtgagc (downstream) for exon 20. The Mutation Surveyor Software (SoftGenetics LLC) was used for sequence analysis.

Kinome mutation frequency analysis

DNA fragment libraries were prepared using the TruSeq DNA Sample Preparation Kit (Illumina) and were hybridized to the SureSelect Human Kinome bait library according to the manufacturer's protocol (Agilent). Captured DNA samples were sequenced on a HiSeq 2000 (Illumina) using a 55 bp paired-end protocol. Sequence reads were aligned to the human genome [GRCh37/hg19] and unique pairs were used for variant calling. Candidate variants were identified using SAMtools and the following inclusion criteria were applied: Minimum coverage = 10; minimum variant count = 5; a variant must be detected on both strands. Variants were assessed using the Ensembl variant effect predictor (v62) to define those that were likely to impact protein coding sequences and to filter out germline polymorphisms. Matched germline DNA was sequenced for 19 of the 73 tumor samples and an additional 56 normal samples (10 from matching adjacent normal colon, 23 from

breast tissue and 23 blood samples) were used to improve the removal of germline SNPs and sequencing errors. In this study we have focussed on mutation load; a full analysis of the sequence alterations is the subject of another study. More information about sequence read alignment and variant calling and be found in Ref 5.

MSI assessment (by hospital)

MSI-status for patients in the development cohort was defined by immunohistochemical staining of FFPE slides for the markers MLH1 and PMS2. MSI-status in patients from the Spanish hospital was determined by PCR amplification of six microsatellite DNA regions (D21S415, D21S1235, D12S95, D4S2948, SIT2, and BAT26) from paired normal and tumor tissues as described previously.[6] MSI multiplex analysis including 5 microsatellite DNA regions (BAT-25, BAT-26, NR-21, NR-24, Mono-27) was performed in 80 tumor samples accordingly to the local standard methodology (MSI Analysis System, Version 1.2, Promega). A tumor with only normal markers was defined as microsatellite stable (MSS). For all patients from the German hospital. Genomic DNA of tumor and corresponding normal colon mucosa were analyzed for microsatellite instability using the Qiagen® Type-it Microsatellite PCR Kit (Qiagen GmbH, Hilden, Germany). Two mononucleotide and three dinucleotide Bethesda markers (BAT25, BAT26, D2S123, D5S346, and D17S250) were investigated. A tumor with five normal markers was defined as microsatellite stable (MSS). Irregularity in one marker was defined as low grade microsatellite instability (MSI-L), irregularity in two or more markers was defined as high grade microsatellite instability (MSI-H).7 Patients with low grade MSI (MSI-L) were classified as MSS for all analysis.

MSI signature readout

The development of the MSI/dMMR signature is described in more detail in Ref 8. Shortly, expression measurements were normalized (lowess normalization) and log-ratios were used for identification of genes that were associated with the tumors' MSI status (based on two-sided Student's t-test). We used a 10-fold cross validation (CV10) procedure that was repeated a thousand times to determine classification performance and for robust gene selection. During each CV10 round genes were ranked by p-value. The 64 genes with highest frequency of appearance within the top ranking genes in each of the 1000 CV loops were selected as the final set with the strongest MSI association. The 64 gene set was used to construct a nearest centroid based classification method (cosine correlation), a MSI gene signature index for the individual samples was defined as the difference of the two correlations. Samples were classified within the MSI group if their index exceeded a pre-defined optimized threshold. This threshold was determined to reach a maximal overall accuracy (sum of sensitivity and specificity).

Loboda EMT gene expression readout

Previously, Loboda *et al*. reported an epithelial-to-mesenchymal (EMT) gene signature to identify colon cancer samples with mesenchymal features.[9] Ninety-six of their top 100 genes that were previously reported to be strongly associated with the EMT program (genes correlated with principle component 1 (PC1), see Figure 2 and Suppl Fig 12 in Ref 9) were available for readout on the Agilent full genome platform (see table below). Each of the individual genes was analyzed for differential expression across the three molecular subtypes (A-, B- and C) using an analysis of variance (ANOVA) model on 188 tumor samples.

In addition to the single-gene based analysis, we have constructed a centroid based classifier to determine a gene profile readout of the Loboda EMT profile. The mesenchymal centroid across the 96 genes was constructed based on the differential expression provided in Ref 9: all 43 epithelial associated genes had a value of -1 in the mesenchymal centroid, and all 53 mesenchymal associated genes had a value of +1. Next, the samples' EMT index a calculated by pearson correlation of its 96-gene profile with the mesenchymal centroid in which a higher index represented a more mesenchymal phenotype

Table of the Lobode EMT associated genes used in this study:

Gene	Epit/Mes	Gene	Epit/Mes	Gene	Epit/Mes	Gene	Epit/Mes
ACVR1	M	ETV5	M	LCN2	E	SH2D3C	M
ARMCX1	M	EVI1	E	LYPD5	E	SHH	E
ASPN	M	FBLN1	M	MAL2	E	SLC39A6	M
AXL	M	FBLN5	M	MAP3K3	M	SMAD1	M
CD24	E	FGF1	M	MEOX2	M	SMAD3	M
CD44	E	FGFR1	M	MET	E	SNAI2	M
CDH1	E	FLRT2	M	MGP	M	SOX9	E
CDH2	M	FN1	M	MMP15	E	SPARC	M
CDON	M	FOXA2	E	MMRN2	M	SPINT1	E
CDX1	E	FOXC1	M	MRAS	M	SRPX	M
CDX2	E	FOXC2	M	MSN	M	STX2	M
CEACAM1	E	FOXD2	E	MST1R	E	TAGLN	M
CLDN4	E	GLI2	M	NFIC	M	TCF4	M
CLDN7	E	GLI3	M	PKP3	E	TGFBR1	M
CLDN9	E	GLIS2	M	PRSS8	E	TIAM1	M
CRB3	E	HTRA1	M	RBM35A	E	TMPRSS4	E
CTGF	M	ISX	E	RBM35B	E	TNS4	E
DSC2	E	ITGB4	E	RECK	M	TWIST1	M
DZIP1	M	JUP	E	RNF11	M	TWIST2	M
ECM2	M	KAZALD1	E	RSL1D1	E	VEGFB	M
ELF3	E	KIAA0152	E	S100P	E	VIM	M
EPHA3	M	KRT19	E	SDC1	E	WASF3	M
EPHB3	E	KRT8	E	SFN	E	WISP1	M
ETS2	E	LAMB2	M	SFRP1	M	ZFPM2	M

Supplementary Tables and Figures

Supplementary Table 1: Patient characteristic in the development cohort, validation cohort, and the subset with kinome mutational analysis.

Variable		Development cohort n= 188		Validation cohort n= 543		Kinome Set n= 73	
Age ≥ 70	No	113	60.1%	311	57.3%	41	56.2%
	Yes	75	39.9%	232	42.7%	32	43.8%
Mean Age (range)		66.6 (21 - 91)		67.2 (33 - 94)		69 (51 - 89)	
Gender	Male	84	44.7%	317	58.4%	33	45.2%
	Female	104	55.3%	226	41.6%	40	54.8%
Hospital Country	Austria			24	4.4%		
	Germany	5	2.7%	232	42.7%	5	6.8%
	Italy			29	5.3%		
	Netherlands	176	93.6%			36	49.3%
	Spain			258	47.5%	32	43.8%
	United Kingdom	7	3.7%				
Stage	I	24	12.8%			9	12.3%
	II	100	53.2%	320	58.9%	35	47.9%
	III	56	29.8%	223	41.1%	29	39.7%
	IV	8	4.3%				
Localisation	Left	92	49.5%	296	55.6%	40	55.6%
	Right	77	41.4%	202	38.0%	28	38.9%
	Rectum	17	9.1%	34	6.4%	4	5.6%
	not available	*2*	-	*11*	-	*1*	-
Grade	low	11	6.0%	100	18.4%	9	12.3%
	intermediate	141	77.5%	318	58.6%	50	68.5%
	high	30	16.5%	125	23.0%	14	19.2%
	not available	*6*	-				
LN > 12	No	141	75.4%	107	19.7%	32	44.4%
	Yes	46	24.6%	435	80.3%	40	55.6%
	not available	*1*	-	*1*	-	*1*	
Mean LN assessed (range)		9.1 (0 - 31)		20.9 (3 - 72)		13.6 (1 - 32)	
pT	1	4	2.1%			1	1.4%
	2	22	11.7%	15	2.8%	10	13.7%
	3	149	79.3%	450	82.9%	58	79.5%
	4	13	6.9%	78	14.4%	4	5.5%
DM	No	137	72.9%	433	80.0%	62	86.1%
	Yes	51	27.1%	108	20.0%	10	13.9%
	not available			*2*	-	*1*	-
Mean DM follow up time (years)		5.6 (0 - 22.5)		6.0 (0 - 15.2)		5.6 (0.6 - 10.9)	
Rec	No	130	69.1%	407	75.0%	60	82.2%
	Yes	58	30.9%	136	25.0%	13	17.8%
Mean Rec follow up time (years)		5.5 (0 - 22.5)		5.9 (0 - 15.2)		5.5 (0.6 - 10.9)	
Death	No	106	56.4%	385	70.9%	53	72.6%
	Yes	82	43.6%	158	29.1%	20	27.4%
Mean Survival follow up time (years)		6.0 (0.2 - 22.5)		6.5 (0 - 15.2)		5.8 (0.6 - 10.9)	
Chemotherapy	No	148	83.6%	290	53.7%	42	60.9%
	Yes	29	16.4%	250	46.3%	27	39.1%
	not available	*11*	-	*3*	-	*4*	-

LN: Number of lymph node assessed
pT: pathological assessment of primary tumor
DM: Event Distant Metastasis
REC: Event Recurrance (local, regional or distant)

Note: Percentages might not add up to 100 due to rounding

Supplementary Table 2. Frequency of mutually exclusive *BRAF* and *KRAS* activating mutations

The mutually exclusive activating *BRAF* and *KRAS* mutations were combined. Samples for which the mutation status of one or both genes was unknown were excluded from this analysis (*not available*).

	A-type	B-type	C-type
BRAF or KRAS activating mutation	103	91	33
BRAF and KRAS wildtype/other	48	226	37
not available	*0*	*5*	*0*

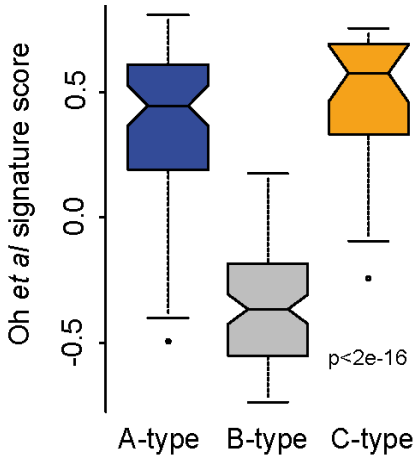
Supplementary Table 3. Prognostic value of CRC molecular subtype classification

Multivariate analysis in the validation cohort. Microsatellite instable (MSI) versus stable (MSS) status was based on hospital assessment, or if not available, using a previously described MSI gene signature.[8] NCCN risk assessment is accordingly to 2012.[10]

Variable	Hazard ratio	P-value
Molecular subtypes (A- *vs* C-type)	**0.175**	**0.0028**
Stage (II *vs* III)	0.298	0.086
Gender (male *vs* female)	0.872	0.78
Microsatellite (MSI *vs* MSS)	0.453	0.22
BRAF (mutant *vs* wildtype)	0.634	0.58
NCCN guidelines (low-risk *vs* high-risk)	0.892	0.87

Supplementary Figure 2: Readout of the signature by Oh *et al* is associated with B-type classification

The previously described gene signature by Oh *et al.*[11] was translated into a nearest centroid based classification methods that was in line with the procedure used for single sample predictor (SSP) bases readout of the colon intrinsic subtypes. Correlation based (Pearson) scores of the Oh *et al.* gene set were determined using 104 of the 114 gene that could be matched to the Agilent array. Correlation scores were calculated for all 188 samples with full genome. Oh *et al.* scores showed a highly significant association with the ABC classification (Anova p<2.2e-16), with B-type samples having a low score and A- and C-types having a high score.



Supplementary References

1. 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(1):17-24
2. Mackay A, Weigelt B, Grigoriadis A, et al. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *J Natl Cancer Inst*. 2011;103(8):662-673.
3. Krijgsman O, Roepman P, Zwart W, Carroll et al. A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response. *Breast Cancer Res Treat.* 2012;133(1):37-47.
4. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167.
5. Rigaill GJ, Cadot S, Kluin RJ, et al. A regression model for estimating DNA copy number data applied to capturesequencing data. *Bioinformatics*. 2012;28(18):2357-2365.
6. Gonzalez-Garcia I, Moreno V, Navarro M, et al. Standardized approach for microsatellite instability detection in colorectal carcinomas. *J Natl Cancer Inst*. 2000;92(7):544-549.
7. Nardon E, Glavač D, Benhattar J, et al. A multicenter study to validate the reproducibility of MSI testing with a panel of 5 quasimonomorphic mononucleotide repeats. *Diagn Mol Pathol*. 2010;19(4):236-342.
8. Tian S, Roepman P, Popovici V, et al. A robust genomic signature for detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol* 2012;228(4):586-595.
9. Loboda A, Nebozhyn MV, Watters JW, et al. EMT is the dominant program in human colon cancer. *BMC Med Genomics*. 2011;4:9.

10. National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology- Colorectal Cancer Screening. Version 2. 2012. Available from: http://www.nccn.org/professionals/physician_gls/pdf/colorectal_screening.pdf
11. Oh SC, Park YY, Park ES, et al. Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. Gut. 2012;61:1291-8