A Low-Cost Multiplex Biomarker Assay Stratifies Colorectal Cancer Patient Samples into Clinically-Relevant Subtypes

Chanthirika Ragulan*[1,2], Katherine Eason*[1], Gift Nyamundanda[1,2], Yatish Patil[1,2], Pawan Poudel[1], Elisa Fontana[1,2], Maguy Del Rio[3], Koo Si-Lin[4], Tan Wah Siew[5], Pierre Martineau[3], David Cunningham[6,7], Iain Beehuat Tan[5,8,9] and Anguraj Sadanandam[1,2, &]

[1] Division of Molecular Pathology, The Institute of Cancer Research (ICR), London, United Kingdom

[2] Centre for Molecular Pathology, Royal Marsden Hospital (RMH), London, United Kingdom

[3] Institut de Recherche en Cancérologie de Montpellier (IRCM), INSERM U1194, Université de Montpellier, Institut régional du Cancer de Montpellier, Montpellier, France.

[4] National Cancer Centre Singapore, Singapore

[5] Singapore General Hospital, Singapore

[6] Department of Medicine, Royal Marsden Hospital (RMH), London, United Kingdom

[7] Division of Clinical Studies, The Institute of Cancer Research (ICR), London, United Kingdom

[8] Genome Institute of Singapore, Singapore

[9] Duke-NUS Medical School, Singapore

* These authors contributed equally to this work

[&] Correspondence to

Dr Anguraj Sadanandam, Institute of Cancer Research (ICR), 15 Cotswold Road, Sutton, Surrey, SM2 5NG, United Kingdom

Keywords: colorectal cancer subtypes, CRCAssigner subtypes, CMS subtypes, subtype diagnostic assay, multiplex and low-cost biomarker assay, nCounter platform, mutations, class prediction, biomarkers

Abstract

Objective: In order to personalise standard therapies based on molecular profiles, we previously classified colorectal cancers (CRCs) into five distinct subtypes (CRCAssigner) and later into four consensus molecular subtypes (CMS) with different prognoses and treatment responses. For clinical application, here we developed a low-cost multiplex biomarker assay.

Design: Three cohorts of untreated fresh frozen CRC samples (n=57) predominantly from primary tumours and profiled by microarray/RNA-Seq were analysed. A reduced 38-gene panel (CRCAssigner-38) was selected from the published 786-gene CRCAssigner signature (CRCAssigner-786) using an in-house gene selection approach. A customised NanoString Technologies' nCounter platform-based assay (NanoCRCAssigner) was developed for comparison with different classifiers (CMS subtypes), platforms (microarrays and RNA-Seq), and gene sets (CRCAssigner-38 and CRCAssigner-786).

Results: NanoCRCAssigner classified samples (n=48; except those showing a mixture of subtypes) into all five CRCAssigner subtypes with overall high concordance across platforms (>87%) and with CMS subtypes (81%) irrespective of variable tumour cellularity. The association of subtypes with their known molecular (microsatellite-instable and stemness), mutational (*KRAS*/*BRAF*), and clinical characteristics (including overall survival) further demonstrated assay validity. To reduce costs, we switched from the standard protocol to a low-cost protocol with a high Pearson correlation co-efficient (0.9) between protocols. Technical replicates were highly correlated (0.98).

Conclusion: Here we developed a low-cost and potentially clinically deployable NanoCRCAssigner assay to facilitate prospective validation of (CRCAssigner and potentially CMS) subtypes in clinical trials and beyond.

Summary "box"

What is already known about this subject?

- Colorectal cancer (CRC) is a heterogeneous disease.
- We previously identified 5 gene expression-based CRC subtypes (CRCAssigner; enterocyte, goblet-like, inflammatory, stem-like and transit amplifying) using a 786-gene signature (CRCAssigner-786) later reconciled into the 4 Consensus Molecular Subtypes (CMS1-4).
- These subtypes were identified by profiling samples using microarray and RNA-Seq platforms, which are expensive, time-consuming and impractical for routine clinical use.
- CRCAssigner subtypes have prognostic and potential predictive differences (to anti-EGFR and FOLFIRI; a combination of irinotecan, 5-fluorouracil and leucovorin).
- Previous analysis of randomised clinical trials assessing patient responses to oxaliplatin in addition to fluouracil-leucovorin indicated that CRCAssigner subtypes may predict responders (compared to CMS) using only the discovery cohort, but larger cohorts are warranted to validate this finding.
- Subtype-driven clinical trials require a validated low-cost assay suitable for routine clinical use.

What are the new findings?

- A reduced 38-gene signature (CRCAssigner-38) from CRCAssigner-786 gene set can be utilised to classify samples into the CRCAssigner subtypes with minimal misclassification error rate.
- The CRCAssigner subtypes can be assessed in fresh frozen samples using a customised CRCAssigner-38 signature-based assay (NanoCRCAssigner) by applying nCounter platform (NanoString Technologies) which is a cost-effective method and provides highly reproducible results.
- Subtype prediction with NanoCRCAssigner was highly concordant with subtypes predicted using the CMS classifier on microarray or RNA-Seq platforms.
- NanoCRCAssigner assay is potentially independent of tumour cellularity, predicts patient prognosis and is consistent with mutational and molecular profiles of CRC subtypes.

How might it impact on clinical practice in the foreseeable future?

- This study demonstrates how molecular CRCAssigner (for the first time) and CMS subtypes can be detected using a biomarker assay (NanoCRCAssigner) suitable for clinical validation.
- With further modification of the protocol to analyse formalin-fixed paraffin embedded (FFPE) samples (not within the focus of the current manuscript), this assay may facilitate patient stratification within clinical trials and the prospective assessment of potential subtype-specific treatments in the future using biopsy or surgical samples.

Introduction

Colorectal cancer (CRC) is the fourth leading cause of cancer-related deaths worldwide [1]. The median overall survival of metastatic (m)CRC patients with unresectable disease remains in the order of 24 months with standard chemotherapy options. The implementation of targeted therapies including anti-EGFR monoclonal antibodies, anti-angiogenic and more recently immunotherapy agents may extend the survival to up to 30 months in selected mCRC patients [2]. However, how to identify those patients who will benefit from different systemic drug options remains challenging. Additional predictive biomarkers are required to spare patients from unnecessary toxicities, improve patients' outcomes and increase cost-effectiveness of treatment.

In order to classify colorectal cancers into subgroups with distinct biology and to effectively match existing therapies to facilitate subtype-specific therapeutic development, we previously identified five distinctive gene expression subtypes and an associated 786-gene signature (CRCAssigner-786) [3]. Based on the gene expression similarities with different cell types of the normal colonic mucosa, we named the subtypes as goblet-like, enterocyte, stem-like, inflammatory and transit-amplifying (TA). We demonstrated significantly poorer disease-free survival (DFS) in untreated patients for the stem-like subtype, intermediate DFS for inflammatory and enterocyte and better DFS for goblet-like and TA [3]. Then, from two different datasets that included drug response information, we observed increased responses within the stem-like subtype to irinotecan, fluorouracil and leucovorin treatment combination (FOLFIRI) [4] and the TA subtypes to anti-EGFR monoclonal antibody (cetuximab) [5]. These treatment responses were further validated by other studies [6, 7].

Five other groups independently identified between 3 and 6 molecularly distinct CRC subtypes [8, 9, 10, 11, 12]. These and our findings were aggregated by a CRC Subtyping Consortium (CRCSC) into 4 consensus molecular subtypes (CMS): CMS1 (similar to inflammatory subtype); CMS2 (enterocyte and TA); CMS3 (goblet-like); and CMS4 (stem-like) and a "mixed" subtype representing either the existence of additional subtypes or the presence of multiple subtypes in a single sample [13].

Recently, the exploratory clinical applicability of our CRCAssigner-786 subtypes was demonstrated when secondary analysis of a randomised clinical trial assessing patient benefit from the addition of oxaliplatin to fluorouracil-leucovorin in early-stage disease revealed that

benefits were highly enriched in the enterocyte subtype compared to the other subtypes in the discovery cohort. Although this finding did not reach the same level of significance in the validation cohort, the trend was identical [14]. Other subtype classifications, including CMS [13] were not able to isolate patients who may potentially benefit from adjuvant oxaliplatin [14]. Notably, while the enterocyte and TA subtypes are merged into CMS2 of the consensus classification, only the enterocyte subset of CMS2 was observed to benefit from adjuvant oxaliplatin. It is, of course, plausible that individual classification systems may demonstrate specific therapeutic and prognostic relevance beyond CMS. Therefore, in this study we have applied both our CRCAssigner and CMS subtype classifications for assay development and comparison.

Although biologically appealing, translating these findings into routine clinical practice remains challenging. This is mainly because of the lack of a fit-for-purpose assay, able to classify patient samples into subtypes in a timely fashion to maintain a clinically relevant turnaround time, with reasonable costs and from the commonly available tissue samples. The majority of the previous classifiers were developed from microarray/RNAseq gene expression profiles. Technologies such as microarrays and RNA-Seq are expensive and time consuming, require dedicated bioinformatics expertise, and have total turnaround times that are too long to be clinically applicable. Also, they rely on pre-amplification of RNA, with consequent impact on accuracy and reproducibility. We previously demonstrated proof-of-concept assays using immunohistochemistry and quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) methods [3]. Nevertheless, these methods may suffer from reproducibility issues. Hence, we applied nCounter platform (NanoString Technologies) to develop clinically-relevant biomarker assay for CRC subtype classification.

The nCounter platform has previously been exploited to develop the Food and Drug Administration (FDA)-approved Prosigna® Breast Cancer Prognostic Gene Signature Assay [15] to predict risk of recurrence in patients treated with adjuvant hormonal therapy, as well as assays to predict medulloblastoma [16] and lymphoma [17] subtypes. This platform measures gene expression in the form of discrete counts of barcoded mRNAs, and requires no amplification step, eliminating a potential source of bias. In the present study, we evaluated the suitability of this technology as a platform for a gene expression-based assay for our CRCAssigner subtypes (for the first time) using a low-cost protocol to subtype CRCs in three different cohorts with different clinical and mutational characteristics. The results were then

compared to the CMS subtype classifications and other platforms. A summary of the classifiers utilised in this study is given in Table 1.

Table 1. Summary of the classifiers utilised in this study. An overview is provided of each classifier's derivation, platform, and the publication in which it was first introduced. The relationships of the classifiers to each other is also provided.

Classifier	**Description**	**Platform**	**Original publication**
CRCAssigner-786	• 786 genes defining the 5 CRCAssigner subtypes (enterocyte, goblet-like, inflammatory, TA and stem-like) on the microarray platform. • Derived from gene expression profiles of primary CRC samples (n = 387). • Subtype-specific drug responses have been demonstrated for stem-like (FOLFIRI) [3] and a subset of TA tumours (cetuximab) [3], and suggested in the enterocyte subtype (oxaliplatin) [14].	Microarray/ RNA-Seq	Sadanandam *et. al.* 2013 [3]
CMS	• 693 genes defining the 4 CMS subtypes (CMS1-4), here applied to the microarray/RNA-Seq platforms. • Derived from reconciling the CRCAssigner subtypes with 5 additional sets of gene expression subtypes [8, 9, 10, 11, 12].	Microarray/ RNA-Seq	Guinney *et. al.* 2015 [13]
CRCAssigner-38	• A subset of the CRCAssigner-786 genes. 38 genes able to classify samples into the CRCAssigner subtypes with minimal misclassification errors on the microarray and RNA-Seq platforms. • Derived from a subset of the samples used to train the CRCAssigner-786 classifier that were highly typical of their subtype (see *Materials & Methods*).	Microarray/ RNA-Seq	Introduced here
NanoCRCAssigner	• A custom nCounter assay (NanoString Technologies) for CRCAssigner-38 genes. Gene expression is profiled using a highly reproducible and low-cost nCounter assay named as *NanoCRCAssigner*.	nCounter Custom Gene Expression Assay (NanoString Technologies). Quantifies mRNA counts of 47 genes selected from CRCAssigner-786 (including all CRCAssigner-38 genes) plus 3 additional genes not used for subtyping.	Introduced here

Materials & Methods

Patient cohorts

Three different cohorts were studied. The first included 17 metastatic (stage IV) CRC samples (Montpellier cohort) from patients enrolled in the single centre REGP prospective study at the Institut du Cancer de Montpellier, which was part of a previously published study [4]. These patients had histopathologically confirmed colon adenocarcinoma and no prior chemotherapy. The microarray data for these samples are previously published and available at GEO Omnibus (GSE62080). A new cohort (Singapore; SG) included 23 CRC patients (20 Chinese, 2 Malay and 1 Indian) was also selected. For the SG cohort, tumour tissues were obtained from patients who have given written informed consent to participate in an ongoing observational CRC cohort study from the National Cancer Centre of Singapore and Singapore General Hospital (SingHealth Institutional Review Board: 2013/110/B). These patients were untreated prior to surgery. RNA-Seq data for these patients are deposited with GEO accession GSE101588. As an additional cohort (OriGene) is commercially available total RNA from 16 primary CRC tumours and 1 liver metastasis purchased from OriGene (Rockville, MD, USA) (microarray data available at GSE101472). All these patients had histopathological confirmation of their diagnosis of colon adenocarcinoma and most had percentage tumour cellularity available; additionally the cohort encompassed all stages and grades of the disease.

NanoString nCounter gene expression assays - standard and low-cost protocols

nCounter® Max Analysis System (NanoString Technologies, Seattle, CA, USA) was used to perform the nCounter platform analysis using either standard or low-cost (Elements chemistry) protocols as per the manufacturer's instructions (Figure 1a). For standard protocol, custom CodeSets (pre-built capture probes tagged to biotin labels and reporter probes tagged to fluorescently colour coded molecular barcodes) for selected sets of genes were designed and built by NanoString Technologies. For the low-cost protocol, nCounter Elements™ TagSets reagents containing nCounter Elements TagSet (NanoString Technologies) and custom-designed target-specific oligonucleotide probe pairs (reporter/Probe A and capture/Probe B probes; from Integrated DNA Technologies, Inc., Leuven, Belgium) were obtained. The TagSets consist of fluorescently-labelled specific Reporter Tags to resolve and count individual nucleic acid target sequences during data collection (Probe A), and biotinylated universal Capture Tags to capture the hybridised target

nucleic acid sequence to the streptavidin-coated imaging surface (Probe B). In order to assemble the nCounter Elements™ TagSets and probe pairs to be ready for hybridisation, master stocks and working stocks of Probe A (at final hybridising concentration of 20 pM) and Probe B (at final hybridising concentration of 100 pM) concentrations were prepared. The Probe A and TagSets were first mixed to create a partial master mix followed by Probe B to create a complete master mix, all at room temperature.

For both standard and low-cost protocols, 100 ng of total mRNA from fresh frozen tumour tissues was diluted with RNase-free water at 20 ng/uL. The hybridisation reactions were prepared according to manufacturers' instructions for either 18 h at 65 °C using Standard CodeSets reagents for standard protocol or for 20 h at 67 °C using Elements™ TagSets reagents for the low-cost protocol. Hybridised samples were pipetted using the nCounter Prep Station and immobilised on to the sample cartridge for data quantification and collection using nCounter Digital Analyzer (NanoString Technologies). The nCounter Prep Station and Digital Analyzer together constitute the nCounter® Max Analysis System. The summary of the protocol is shown in Figure 1a. For the PanCancer Progression Panel (NanoString Technologies), the same protocol as for standard protocol was used.

Processing and quality control of nCounter data

Data quality was checked using nSolver analysis software v3.0 (NanoString Technologies). Firstly, counts were corrected to background noise using geometric means of 8 negative control probes followed by the correction using geometric means of 6 internal positive control spike-ins in each lane/sample to correct potential sources of variations across the samples. These negative and positive probes were built-in in both standard and low-cost protocols. Only those housekeeping genes with raw molecular counts greater than 50 and those selected by geNorm algorithm (part of the nSolver analysis software) were retained for further analysis. Variations due to RNA input volume were corrected by normalising to the expression of geNorm selected housekeeping genes. The normalised final count data were $\log_2$ transformed for further analysis. Those genes having zeros in more than 80% of the samples (representing potential technical error) were removed. Data generated from PanCancer Progression were also analysed for the biological pathways using the nCounter Advanced Analysis plugin (v1.0.84) for the nSolver analysis software.

Comparison of standard and low-cost protocols

For comparison of standard and low-cost protocols, housekeeping genes that were common between these two datasets were used for normalisation. Data from standard and low-cost protocols were row (gene)-median centred across samples separately, before being combined to perform hierarchical clustering.

Gene Selection for nCounter assay

The following genes were initially selected for inclusion in the CRCAssigner subtype custom nCounter assay (NanoCRCAssigner) based on our previous report [3]: a) the seven genes from the CRCAssigner-7 signature; b) among the top 10 highest scoring genes for each subtype from predictive analysis of microarrays (PAM; [18]) centroids; c) five genes that are specific to TA (cetuximab-sensitive and –resistant) sub-subtypes; and d) those representing the characteristics of certain subtypes such as WNT signalling and specific inflammatory genes such as *AXIN2* and *IFIT3*. These 50 genes were further reduced to robust 38 (CRCAssigner-38) genes using an in-house developed computational pipeline bioinformatics tools – *idSample* and *intPredict* (see *Supplementary Materials & Methods* for more information).

Assigning subtypes to samples

CRCAssigner subtypes were assigned by already published single-sample prediction (SSP) tool [13] by performing Pearson correlation of gene-wise median-centred expression profiles for each sample with centroids for the subtypes. The subtype with the highest correlation was then assigned to that sample. Samples were marked as having "undetermined" subtype if the sample's correlation with the subtype centroid was correlation co-efficient (R^2) $\leq$ 0.15, or labelled as "mixed" if the correlation was high for multiple subtype centroids (R^2 difference between first and second highest R^2 subtypes $\leq$ 0.06), in line with the published CMS classifier [13].

CMS subtypes were determined from microarray or RNA-Seq data using the CMSclassifier R package (v1.0.0) and the *classifyCMS* function using SSP tool. Those samples labelled as "not available (NA)" by *classifyCMS* were further explored and assigned as "undetermined if the minimum, median and maximum correlation to the 20 *classifyCMS* centroids (four subtypes $\times$ five sample types) were all less than 0.15 and a mixed subtype if any of those three had a second subtype within 0.06 of the first.

Subtype concordance and significance

Subtype concordance between two different platforms was calculated as the percentage of samples that showed same subtype (not including mixed and undertermined samples) in both. Subtypes were deemed to be concordant between the CRCAssigner and the CMS subtypes based on the following equivalence: CMS1 = Inflammatory; CMS2 = Enterocyte and TA; CMS3 = Goblet-like; CMS4 = Stem-like [13]. Fisher's exact test between different platforms or classifiers was performed to assess statistical significance of their subtype concordance.

See Supplementary Methods & Materials for more information.

Results and Discussion

Development of nCounter assay for CRC subtyping.

In order to evaluate the applicability of the CRCAssigner and CMS subtype classifications in the clinic, we initially developed a custom nCounter assay using a 50-gene panel, including 47 genes selected from the CRCAssigner-786 signature and an additional 3 genes [3] (Table S1; see *Materials & Methods*). Initially, we applied a standard protocol (in which biotin labels and molecular barcodes are directly attached to the mRNA probes; Figure 1a) from the manufacturer and tested the performance of the custom nCounter assay using primary tumour RNA obtained from 22 histopathologically confirmed colorectal cancers from two different cohorts (Montpellier and OriGene; see *Materials & Methods*; Table S2a). The distribution of samples across principal subspace using principal component analysis (PCA; Figure S1a) showed no batch effect between the two cohorts of samples. In addition, hierarchical clustering analysis using nCounter profiles clustered these 22 samples into different groups that potentially represent different subtypes (Figure 1b).

Selection of alternative cost-effective method

Next, we evaluated if a more economical method employing a "low-cost" protocol (Elements chemistry; custom unique probes are attached to the biotin labels and molecular barcodes separately; approximately 35% less cost compared to standard protocol; Figure 1a) from NanoString Technologies can deliver similar classification performance compared to the standard protocol-based assay. The results from the low-cost protocol-based profiling (Table S2b) showed similar distribution of samples (n=22) across principal subspace using PCA (Figure S1b) and clustered into potential subtypes (Figure S1c) in a similar fashion to the standard protocol.

We merged both the standard and low-cost protocols' gene expression data after normalising (gene-wise median centring) each dataset and performed PCA (Figure S1d) to assess potential batch effect. Figure 1c shows the clustering of same samples between protocols. Measurements showed high correlation (R^2 = 0.90, $p<0.001$; Figure 1d) between these different protocols. This demonstrates that we can successfully replicate results from standard protocol using low-cost protocol for a more cost-effective assay. Therefore, we adopted low-cost protocol for our assay (Figure 1a).

Reproducibility of NanoCRCAssigner assay between different batches

For the purposes of a clinical biomarker assay for CRC subtyping, it is imperative that results are highly reproducible between batches (e.g. at different time points pre and post-treatment). To test this aspect of our platform, we performed our assay on five of the above samples twice, in separate batches of maximum 40 weeks apart; Table S2c-d). Figure 1e and Figure S1e shows the clustering of replicate samples together with negligible batch effect. Next, when we compared the expression of each gene in each sample across the two replicates, we achieved high concordance between the assays, with a Pearson's R^2 of 0.98 ($p < 0.001$; Figure 1f). This establishes the high reproducibility of our assay over non-negligible periods of time. Hence, in the future, we can use this assay to test the state of subtypes using matched pre- and post-treatment biopsies or surgical materials.

Selection of robust set of genes for the assay

Successful clinical biomarker assays should be able to classify samples into subtypes with high concordance, and this requires a robust set of genes. Hence, we tested the robustness of our selected 47 CRCAssigner genes (out of 50-gene panel) using a reduced set of our published training dataset (n=192; Figure S2a and Table S3a) [3] and our in-laboratory developed *intPredict* bioinformatics tool, which contains a pipeline of supervised class prediction methods (*see Supplementary Information*, Figure S2b and Table S3b-c). Our tool identified 38 robust genes (CRCAssigner-38) out of 47 CRCAssigner genes with lowest misclassification error rate (MCR) (0.01%) (Figures 1g and S2c-d; Table S3b). In order to further effectively classify our samples into CRCAssigner subtypes using the selected CRCAssigner-38 genes, we applied a newly derived CRCAssigner-38 gene classification metric (gene centroids using Prediction Analysis of Microarrays statistical method; see *Supplementary Information* and *Materials & Methods*; Figures S2e-f and Table S3d) with only 1.6% MCR (with leave-out cross validation analysis) to classify samples profiled on our nCounter assay (NanoCRCAssigner assay) and compared to the other gene profiling platforms and classifiers in three different CRC cohorts.

Subtyping using NanoCRCAssigner assay

In order to confirm that our NanoCRCAssigner assay can stratify patients into clinically relevant groups, we first profiled a retrospective cohort of 17 primary tumours from stage IV patients [4, 6] (Figure 2a, *Materials & Methods*) using our 47-gene custom nCounter panel (Figure S3a-b; Table S4a-b). We performed NanoCRCAssigner assay and assigned subtypes

to each sample by correlating the expression with the CRCAssigner-38 gene centroids (see *Materials & Methods*). Figure 2b shows the expression of our CRCAssigner-38 signature (NanoCRCAssigner assay) in these samples as measured on the nCounter platform. All the CRCAssigner subtypes were present in this cohort, and all samples were successfully classified, with none showing mixed (similar to that published for CMS classification [13]) or undetermined (those that cannot be classified confidently; Figure 2c; see *Materials & Methods*) subtype characteristics. We observed a non-uniform distribution of the subtypes with enterocyte contributing 41.2% of all the samples followed by stem-like (23.5%) and goblet-like (17.6%) subtypes. However, inflammatory (11.8%) and TA (5.9%) subtypes were low in numbers in this cohort of samples (Figure 2c). This clearly suggests that our NanoCRCAssigner assay confidently predicts all the five CRCAssigner subtypes.

Comparing subtyping concordance between NanoCRCAssigner assay and microarray platform

In order to compare the NanoCRCAssigner assay performance to the original technology platform from which the subtypes and signatures were derived, we also classified the samples using Affymetrix Human Genome U133 Plus 2.0 (HG-U133 Plus2) microarray profiles [4] (Figure S3c). Similar to that of the NanoCRCAssigner assay, we used the CRCAssigner-38 centroids and classified the samples' microarray profiles into all the five subtypes with 35.2% as enterocyte, 23.5% as stem-like, 17.6% as goblet-like, 11.8% as inflammatory and 5.9% as TA (Figure S3d). There was only one sample that was defined as a mixed subtype (5.9% of the 17 samples) expressing both inflammatory and enterocyte genes. The expression of enterocyte genes was consistent with the classification of the sample as enterocyte on the NanoCRCAssigner platform. The fact that the sample was classified as a mixture of subtypes by the CRCAssigner-38 may be attributed to platform-specific effects. Overall, the NanoCRCAssigner assay showed perfect 100% concordance with the microarray-based CRCAssigner-38 classification after excluding samples with mixed classification (due to challenges in comparing these mixed subtypes to others) (Figure 2d).

Similarly, we applied the original CRCAssigner-786 centroids to the microarray data to classify samples into the five subtypes. The CRCAssigner-786 classification yielded 23.5% enterocyte, 29.4% stem-like, 23.5% goblet-like, 11.8% inflammatory and 5.9% TA samples (Figure S3e). There was one mixed sample (5.9% of 17 samples; Figures 2b and S3e) that was different from that called mixed by CRCAssigner-38. This sample did not show either

subtype in the mixture matching with the NanoCRCAssigner subtype (enterocyte); instead, it was classified as TA/stem-like. Given that different gene sets were used for the CRCAssigner-786 and CRCAssigner-38/NanoCRCAssigner classifiers, this discordance suggests that changing the number of genes used for classification can affect the subtype allocated. Nevertheless, robust and concise gene sets are important for consistent classification in a clinical setting.

Irrespective of the different number of genes profiled on the different platforms, the NanoCRCAssigner assay showed 87.5% concordance with the CRCassigner-786 subtypes (Figure 2d). Again, the 12.5% discordance may be attributable to noisy genes present in the CRCassigner-786 signature, as we observed a saturation of the MCR beyond 38 genes using multiple class prediction methods (Figure S2d). Overall, the NanoCRCAssigner assay and CRCAssigner-786 classification perform well with low discordance.

In order to statistically validate these findings, we performed Fisher's exact test using these classifications, excluding mixed or undetermined samples. We found that NanoCRCAssigner assay was significantly (false discovery rate; FDR<0.001; Figure 2e) associated with both the CRCAssigner-38 and -786 classification systems. This again statistically validates the high concordance between NanoCRCAssigner classification and different gene- and platform-based CRCAssigner classifications, further confirming the robustness of our NanoCRCAssigner subtypes. In sum, all the three classifications from the two different platforms identified all the five subtypes and the CRCAssigner-38 gene classifier predicted robust subtypes irrespective of platform differences.

Comparison of the CMS classification with NanoCRCAssigner subtypes and survival analysis

We assessed if our NanoCRCAssigner assay would be useful to classify samples according to the CMS subtypes. We classified the Montpellier cohort of 17 microarray gene expression profiles into CMS subtypes using the published CMS classifier [13]. We successfully classified the samples into all of the CMS subtypes: 47.1% were CMS2 (enterocyte and TA); 11.8% each of CMS3 (goblet-like) and CMS4 (stem-like); and 5.9% of CMS1 (inflammatory). However, we found 17.6% mixed and 5.9% undetermined samples (Figure 2f, *Materials & Methods*). Thus, NanoCRCAssigner showed fair concordance (84.6%) with the CMS classifier excluding the mixed and undetermined samples (Figure 2d). We further

performed pairwise Fisher's exact test between the CMS and NanoCRCAssigner subtypes (Figure 2e), which confirmed borderline significant association to the CMS classification (FDR=0.07). This suggests that NanoCRCAssigner may be applied as a surrogate to predict CMS subtypes with low discordance (only 15.4%) in addition to CRCAssigner subtypes.

Although it is challenging to assess predictive power of an assay in a small Montpellier cohort, we sought to understand the potential of the assay to predict prognosis. Therefore, we performed Kaplan-Meier survival analysis and log-rank test for overall survival (OS) using the NanoCRCAssigner assay and 15 non-mixed samples. The CRCAssigner subtypes identified using NanoCRCAssigner showed borderline significance of p=0.08 (log-rank test; Figure 2g). Similar results were obtained using CRCAssigner-38 and CRCAssigner-786 (Figures S3f-g). On the other hand, the CMS subtypes showed no significance in OS (Figure S3h) potentially due to small sample size (n=13, excluding mixed and undetermined samples). Overall, the NanoCRCAssigner assay has the potential to predict patient prognosis. This prognostic prediction of the NanoCRCAssigner assay requires further validation with a large cohort of samples.

Performance of NanoCRCAssigner assay in a multi-stage Asian cohort

We then utilised an additional cohort of 23 CRC samples from an Asian population comprising all disease stages (SG cohort; *Materials & Methods*) to further evaluate our subtyping assay using the NanoCRCAssigner assay (Figure 3a; Table S4c-d). To our knowledge, these are the first set of Asian CRC samples profiled for molecular subtyping using nCounter platform (Figures S4a-c). Interestingly, the NanoCRCAssigner assay identified all the five CRCAssigner subtypes representing that these subtypes are also present in the Asian population. In this data set, again the enterocyte subtype showed the highest prevalence with 30.4%. This was followed by goblet-like (17.4%) and the three other subtypes at 8.7% prevalence (Figures 3b-c). Interestingly, one of the samples (sample number - 1017) with goblet-like subtype characterisation also showed mucinous characteristics by pathological evaluation (Table S4e) representing that our assay predicts subtypes based on their characteristics. 26.1% of samples were mixed subtype in the SG cohort (Figure 3c), an increase with respect to the previous set, suggesting that there may be a different proportion of mixed samples in the Asian population compared to the Western population. There was no visible trend in association between stages and subtypes (Figure S4d; Table S4e).

Then, we compared the 17 SG samples that were matched between platforms (nCounter and RNA-Seq) and classifier (CMS subtypes). Consistent with the observations from Montpellier cohort, NanoCRCAssigner subtypes showed 100% concordance and significant association (FDR <0.001; Fisher's exact test) with both CRCAssigner-38 and CRCAssigner-786 subtypes when mixed subtypes were not included. Also, there was a high concordance of 88.9% between NanoCRCAssigner and CMS subtypes (Figures S4f-j). In all 6 samples classified as mixed by NanoCRCAssigner, one of the subtypes in the mixture was also identified as the only subtype by CRCAssigner-38 or CRCAssigner-786. Overall, our NanoCRCAssigner assay identifies subtypes in Asian CRC patients independent of the disease stages.

Evaluating associations between NanoCRCAssigner assay subtypes and mutational and microsatellite instability (MSI) profiles

Previously, it has been reported that the CMS1/inflammatory subtype is associated with MSI and *BRAF* mutations, whereas CMS3 is highly associated with *KRAS* mutations [13]. To further validate the NanoCRCAssigner subtypes, we compared these with the mutational (*BRAF* and *KRAS*) and MSI status (Figure 3d) of cancers in the Asian cohort (*n=11*). All (100%) the inflammatory subtype CRCs were associated with MSI. Interestingly, one of the two *BRAF* mutant tumours was associated with the inflammatory (CMS1) subtype and MSI status. Similarly, all three goblet-like subtype (CMS3; 100%) tumours were associated with *KRAS* mutation. There were three other *KRAS* mutant tumours associated with the enterocyte or stem-like subtypes, representing *KRAS* mutant tumours are also associated with other subtypes as previously reported [13]. This analysis with our NanoCRCAssigner assay corresponds with known associations of subtypes with mutational and MSI profiles with this small SG data set of 17 samples. However, additional large data sets and NanoCRCAssigner assays are warranted to validate these observations.

Effect of tumour cellularity on NanoCRCAssigner assay

Next, we sought to test if tumour cellularity affects our NanoCRCAssigner assay. For this purpose, we used a commercially purchased cohort of 17 primary tumour samples (OriGene, Rockville, MD, USA; *Materials & Methods*) spanning all stages of CRC, with variable tumour cellularity ranging from 50% to 85% (Figures 4a and S5a-b; Table S4f-g). One sample had extremely low cellularity of 10% and one additional sample was from a liver metastasis. We assigned subtypes to these 17 RNA samples using the different classifiers as

shown for the previous cohorts. Since the *CA1* gene could not be mapped from the probes in the Affymetrix GeneChip Human Transcriptome Array (HTA) array, we reduced our CRCAssigner-38 to 37 genes for comparison (Figure S5c-d). For consistent evaluation across platforms, we maintained the 37 genes for all assays in this cohort. The NanoCRCAssigner assay predicted all five subtypes subtypes together with 17.6% mixed subtypes (Figures 4b-c). Unlike the other cohorts, OriGene showed the highest proportion of TA (29.4%) followed by enterocyte (23.5%), goblet-like (11.8%), inflammatory (11.8%), and stem-like (5.9%; Figure 4c). This suggests that NanoCRCAssigner assay is not biased towards any particular subtype.

The distribution of subtypes in OriGene cohort according to CRCAssigner-38, CRCAssigner-786 and CMS is shown in Figures S5e-g. The NanoCRCAssigner classifier had 78.5% concordance with both the CRCAssigner classifications and 70% concordance with CMS classification in the OriGene cohort (Figure S5h). It is possible that profiling samples using the HTA microarray introduced platform-specific effects given that the CRCAssigner-786 and CRCAssigner-38/NanoCRCAssigner gene centroids were derived from samples profiled on the HGU-133 Plus 2.0 platform, however, Fisher's exact test showed significant association (FDR<0.05) between all 4 classifiers (Figure S5i). Overall, this again validates the performance of our NanoCRCAssigner assay compared to another gene expression profile platform.

Furthermore, we postulated that if cellularity affects our assay, the low cellularity samples should be either qualified as "undetermined" or "mixed subtype" samples. Interestingly, none of the samples were classified as having undetermined subtype, regardless of cellularity. On the other hand, the three mixed samples were among those having high cellularity (50-75%) (Figures 4d-e). These results suggest that our NanoCRCAssigner may not be affected by cellularity due to the selection of robust gene sets.

Overall concordance and distribution of subtypes across all cohorts

To assess the stability of individual subtypes across the various platforms, gene sets and data sets, we plotted Figure 5a using all significantly classified and non-mixed samples for all three cohorts (n = 39; Table S5). Three of the assays are shown (CRCAssigner-786, CRCAssigner-38 and NanoCRCAssigner) along with the 5 CRCAssigner subtypes. While the goblet-like subtype was consistent across all the different classifications, there were four

samples that had different subtypes across classifications (shown in grey in Figure 5a). Two of the samples were classified as either enterocyte or stem-like, one each as either enterocyte or inflammatory and either TA or inflammatory. Three of these four samples were from the OriGene cohort (Figure 4b), potentially due to platform-specific effects as discussed previously. However, overall concordance between platforms was good (Figure 5a), with 35/39 non-mixed/undetermined samples (89.7%) showing the same subtype across all 3 assays.

We also sought to confirm that platform and gene set differences did not bias the distribution of subtypes assigned to the samples. Figure 5b shows the proportion of subtypes according to each classifier across the three cohorts, excluding mixed or undetermined samples (n = 39). P-values are the result of proportion tests and show that there is no significant difference in the distribution of each subtype across the three CRCAssigner-based classifiers.

Characteristics of the identified subtypes

In order to further confirm if the classified subtypes represent the molecular characteristics of their original published phenotypes, we performed analysis using NanoString Technologies' PanCancer Progression Panel that represents mainly epithelial, mesenchymal and extracellular matrix genes (Table S6a-b; Figure S6a-b). As expected, the stem-like samples had increased expression of these genes. We observed specific pathways associated with epithelial to mesenchymal transition (EMT), stem cell, metastatic response, extracellular matrix (ECM) structure and receptor interaction, cellular differentiation, collagen family and others (Figures 5c and S6c-d). In addition, the expression of the CRCAssigner-786 genes is shown in Figure 5d alongside the NanoCRCAssigner subtypes of the samples. Overall, these analyses demonstrate that the subtypes identified by NanoCRCAssigner assay represent the molecular characteristics of these subtypes in 51 samples (with matched gene expression profiles from other platforms) from three independent cohorts.

Conclusion

In summary, we developed a low-cost protocol-based biomarker assay (NanoCRCAssigner) using CRCAssigner-38 robust gene signatures that classifies CRC samples into five CRCAssigner subtypes along with mixed subtypes for the first time to our knowledge. The subtype prediction by NanoCRCAssigner assay is highly concordant with multiple platforms and may serve as a surrogate for CMS subtypes. Also, this assay represents the clinical, mutational and molecular characteristics of the CRC subtypes. Since multiple CRC clinical trials require low-cost, reproducible and rapid clinically implementable assays to prospectively validate CRC subtypes for subtype-specific studies, our NanoCRCAssigner assay may potentially facilitate this process in the clinic after further development using FFPE samples, which requires further modification of the protocol that is currently on-going and not a focus of the current manuscript.

ACKNOWLEDGEMENTS

We thank Prof. Mitch Dowsett, Dr. Richard Buus, Dr. Maggie Cheang, Dr. Nicola Valeri, Dr. George Vlachogiannis and Dr. Andrea Lampis from the ICR for their advice on the use of nCounter platform, and Prof. Paul Workman for helpful comments on the manuscript. We acknowledge NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and the ICR.

AUTHOR CONTRIBUTIONS

A.S. conceived the idea. C.R. optimised, developed and performed nCounter assays, and performed microarray experiments. K.E. performed bioinformatics analysis and subtyping of all nCounter, microarray and RNA-Seq data. G.N. developed class prediction models and the *R* package. Y.P. performed RNA-Seq data analysis and data integration. E.F. helped with clinical interpretation. P.P. assisted with data analysis. M.D.R., and P.M. isolated and provided RNA for Montpellier cohort of samples and their associated clinical information and microarray data; and K.S., T.W.S., and I.B.T., isolated and provided RNA for SG cohort and their associated RNA-Seq data and clinical information. D.C. helped with the clinical interpretation of the data. C.R., K.E., and A.S. interpreted the results and wrote the manuscript.

ADDITIONAL INFORMATION

Previously published GEO Omnibus data sets were analysed for gene set selection (GSE14333 and GSE13294) and microarray-based subtyping of the Montpellier cohort (GSE62080). nCounter data for all cohorts (GSE101479 – standard protocol and GSE101481 – low-cost protocol) and microarray/RNA-Seq data for OriGene (GSE101472) and Singapore (GSE101588) cohorts are deposited under the SuperSeries with accession number GSE101651.

COMPETING FINANCIAL INTERESTS

A.S. has ownership interest (including patents) as a patent inventor for a patent entitled "Colorectal cancer classification with differential prognosis and personalized therapeutic responses" (patent number PCT/IB2013/060416).

ETHICAL APPROVAL AND INFORMED CONSENT

The SG cohort of this study was approved by the SingHealth Institutional Review Board: 2013/110/B.

References

1 Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer 2013.

2 Van Cutsem E, Cervantes A, Adam R, Sobrero A, Van Krieken JH, Aderka D, *et al.* ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. Annals of Oncology 2016:mdw235.

3 Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nature Medicine 2013;**19**:619-25.

4 Del Rio M, Molina F, Bascoul-Mollevi C, Copois V, Bibeau F, Chalbos P, *et al.* Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. Journal of Clinical Oncology 2007;**25**:773-80.

5 Khambata-Ford S, Garrett CR, Meropol NJ, Basik M, Harbison CT, Wu S, *et al.* Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. Journal of Clinical Oncology 2007;**25**:3230-7.

6 Del Rio M, Mollevi C, Bibeau F, Vie N, Selves J, Emile J-F, *et al.* Molecular subtypes of metastatic colorectal cancer are associated with patient response to irinotecan-based therapies. European Journal of Cancer 2017;**76**:68-75.

7 Medico E, Russo M, Picco G, Cancelliere C, Valtorta E, Corti G, *et al.* The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. Nature Communications 2015;**6**:7002.

8 De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LPMH, *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nature Medicine 2013;**19**:614-8.

9 Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Medicine 2013;**10**.

10 Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, *et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. Journal of Pathology 2013;**231**:63-76.

11 Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, *et al.* Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Medical Genomics 2012;**5**:1-15.

12 Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. International Journal of Cancer 2013;**134**:552-62.

13 Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, *et al.* The consensus molecular subtypes of colorectal cancer. Nature Medicine 2015;**21**:1350-6.

14 Song N, Pogue-Geile KL, Gavin PG, Yothers G, Rim Kim S, Johnson NL, *et al.* Clinical outcome from oxaliplatin treatment in stage II/III colon cancer according to intrinsic subtypes: Secondary analysis of NASBP C-07/NRG oncology randomized clinical trial. JAMA Onc 2016;**2**:1162-9.

15 Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics 2015;**8**:54.

16 Northcott PA, Shih DJH, Remke M, Cho YJ, Kool M, Hawkins C, *et al.* Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. Acta Neuropathologica 2012;**123**:615-26.

17 Scott DW, Wright GW, Williams PM, Lih C-J, Walsh W, Jaffe ES, *et al.* Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. Blood 2014;**123**:1214-7.

18 Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences 2002;**99**:6567-72.

Legends

Figure 1. Assessment of different protocols and replicates of nCounter assay. a. Flowchart showing the major steps of the NanoCRCAssigner assay protocols. Specifically, this flowchart demonstrates the difference between standard and low-cost protocols. Though low-cost protocol has additional steps, it substantially reduces the cost without significantly increasing the time of the assay. **b-c.** Heatmap of expression levels of the selected 47 CRCAssigner genes (and 3 additional genes) for 22 samples from the OriGene and Montpellier cohorts as measured on a custom nCounter panel using b) standard protocol and c) both standard and low-cost protocols. **d.** A scatter plot of gene expression measurements for all genes in all samples between the standard and low-cost protocols. Each point is coloured by the gene's weight (PAM score) in the CRCAssigner-786 centroids. **e.** Heatmap of expression levels of selected 47 CRCAssigner genes (and 3 additional genes) from 5 technical replicate samples assayed using low-cost protocol with a maximum interval of 40 weeks. **f.** Scatter plot of gene expression measurements for all genes in all samples between technical replicates (median centered within data sets before correlation to remove batch effects). Each point is coloured by the gene's weight (PAM score) in the CRCAssigner-786 centroids. **g.** Heatmap showing expression of CRCAssigner-38 gene signature in the 192 samples chosen from our previous publication (Sadanandam et al. 2013) to train gene selection methods for the NanoCRCAssigner assay. The top bar indicates the CRCAssigner-786 subtype of each sample as determined in our previous publication.

Figure 2. Montpellier cohort - NanoCRCAssigner assay, its comparison with other platforms and CMS classifier and survival analysis. a. A summary of the Montpellier cohort showing patient characteristics, sample size and available HG-U133 Plus 2.0 microarray data. **b.** Heatmap showing the expression of CRCAssigner-38 genes in Montpellier cohort as measured using NanoCRCAssigner assay. The subtypes as assigned using NanoCRCAssigner assay (nCounter platform), CRCAssigner-38 (microarray platform), CRCAssigner-786 and CMS classifications are shown on the top bars. **c**. Pie chart showing the proportion of different subtypes (including mixed and undetermined samples) from NanoCRCAssigner classification. **d-e.** Comparisons between NanoCRCAssigner and other classifications including CRCAssigner-38, CRCAssigner-786 and CMS showing d) percent concordance and e) statistical significance (Fisher's exact test). **f**. Pie chart showing the proportion of different subtypes (including mixed and undetermined samples) from CMS classification. **g.** Kaplan-Meier curve showing overall survival of patients stratified by

NanoCRCAssigner subtype. N38 – NanoCRCAssigner, C38 – CRCAssigner-38, C786 – CRCAssigner-786.

Figure 3. SG cohort - NanoCRCAssigner assay and subtype association with MSI and mutations. a. A summary of the SG cohort showing patient characteristics, sample size and available RNA-Seq data. **b.** Heatmap showing the subtypes as assigned using NanoCRCAssigner assay (nCounter platform), CRCAssigner-38 (microarray platform), CRCAssigner-786 and CMS classifications. **c.** Pie chart showing the proportion of different subtypes (including mixed samples) from NanoCRCAssigner classification. **d.** Heatmap showing MSI and mutational (*KRAS* and *BRAF*) status of subtypes from NanoCRCAssigner and CMS classifications.

Figure 4. Effect of tumour cellularity and probe differences on NanoCRCAssigner, CMS and other platform-based subtypes. a. A summary of OriGene cohort showing patient characteristics, sample size and available HTA microarray data. **b.** Heatmap showing the subtypes as assigned using NanoCRCAssigner assay (nCounter platform), CRCAssigner-38 (microarray platform), CRCAssigner-786 and CMS classifications. **c.** Pie chart showing the proportion of different subtypes (including mixed samples) from CRCAssigner classification. **d.** Heatmap showing tumour cellularity, stage and grade associated with subtypes from CRCAssigner and CMS classifications. **e.** Histogram showing the distribution of tumour cellularity.

Figure 5. Subtype-specific cross-cohort, pathway and CRCAssigner-786 gene signature analysis to assess subtype stability. a. Chord plot illustrating the tendency of samples to be classified as the same subtype between the three CRCAssigner-based assays. Samples from all three cohorts which had no mixed or undetermined subtype calls were included (n=39). Each arc connects the classification of a sample in two different assays, and each sample is represented by three arcs (connecting NanoCRCAssigner assay (nCounter platform), CRCAssigner-38 and CRCAssigner-786 (microarray platform) subtypes). Samples with the same subtype in all three assays are coloured by their subtype. Samples that had discordant classification between the assays are coloured grey. **b.** Distribution of subtypes according to each classifier. Samples that were mixed or undetermined subtype were excluded for each classifier. Results of statistical tests of proportion between the three CRCAssigner-based classifiers are shown on the left-hand side. **c-d.** Heatmap of c) nCounter PanCancer

Progression Panel-based gene expression profiles from the Montpellier and OriGene cohorts of samples (n=34) and d) batch-corrected RNA-Seq/microarray gene expression profiles from all the three cohorts (n=51). The first top bar indicates the NanoCRCAssigner subtype of the samples and the second top bar indicates the cohorts. Genes are grouped according to functional annotations provided by NanoString Technologies or subtype gene signatures.

Figure 1

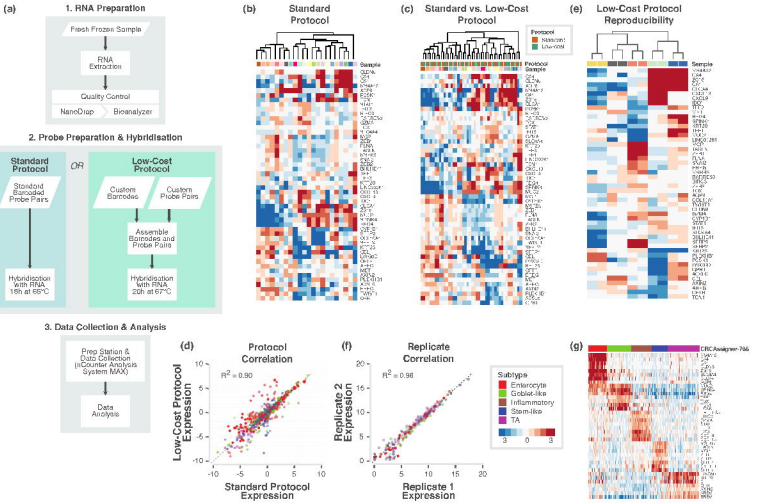

(a)
1. RNA Preparation
Fresh Frozen Sample
RNA Extraction
Quality Control
NanoDrop
Bioanalyzer
2. Probe Preparation & Hybridisation
Standard Protocol
OR
Low-Cost Protocol
Stoindard Barcoded Probe Pairs
Custom Barcodes
Custom Probe Pairs
Assemble Barcodes and Probe Pairs
Hybridisation with RNA, 18h at 65°C
Hybridisation with RNA, 20h at 67°C
3. Data Collection & Analysis
Prep Station & Data Collection (nCounter Analysis System MAX)
Data Analysis
(b)
Standard Protocol
Sample
(c)
Standard vs. Low-Cost Protocol
Protocol
Standard
Low-cost
Protocol
Sample
(e)
Low-Cost Protocol Reproducibility
Sample
(d)
Protocol Correlation
R2 = 0.90
Low-Cost Protocol Expression
Standard Protocol Expression
(f)
Replicate Correlation
R2 = 0.98
Replicate 2 Expression
Replicate 1 Expression
Subtype
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
(g)
CRC Analyser-766

Figure 2

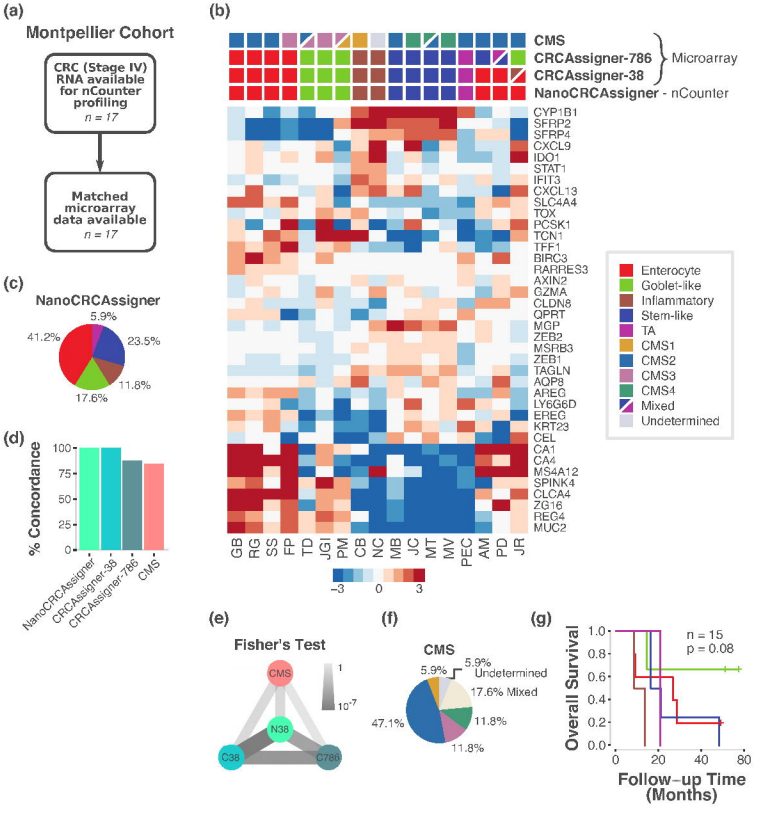

(a)
Montpellier Cohort
CRC (Stage IV) RNA available for nCounter profiling
n = 17
Matched microarray data available
n = 17
(b)
CMS
CRCAssigner-786
CRCAssigner-38
Microarray
NanoCRCAssigner - nCounter
CYP1B1
SFRP2
SFRP4
CXCL9
IDO1
STAT1
IFIT3
CXCL13
SLC4A4
TOX
PCSK1
TCN1
TFF1
BIRC3
RARRES3
AXIN2
GZMA
CLDN8
QPRT
MGP
ZEB2
MSRB3
ZEB1
TAGLN
AQP8
AREG
LY6G6D
EREG
KRT23
CEL
CA1
CA4
MS4A12
SPINK4
CLCA4
ZG16
REG4
MUC2
GB
RG
SS
FP
TD
JGI
PM
CB
NC
MB
JC
MT
MV
PEC
AM
PD
JR
−3
0
3
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
CMS1
CMS2
CMS3
CMS4
Mixed
Undetermined
(c)
NanoCRCAssigner
5.9%
23.5%
11.8%
17.6%
41.2%
(d)
% Concordance
100
75
50
25
0
NanoCRCAssigner
CRCAssigner-38
CRCAssigner-786
CMS
(e)
Fisher's Test
CMS
N38
C38
C786
1
10^-7
(f)
CMS
5.9%
5.9%
Undetermined
17.6% Mixed
47.1%
11.8%
11.8%
(g)
Overall Survival
1.0
0.8
0.6
0.4
0.2
0.0
n = 15
p = 0.08
0
20
40
60
80
Follow-up Time
(Months)

Figure 3

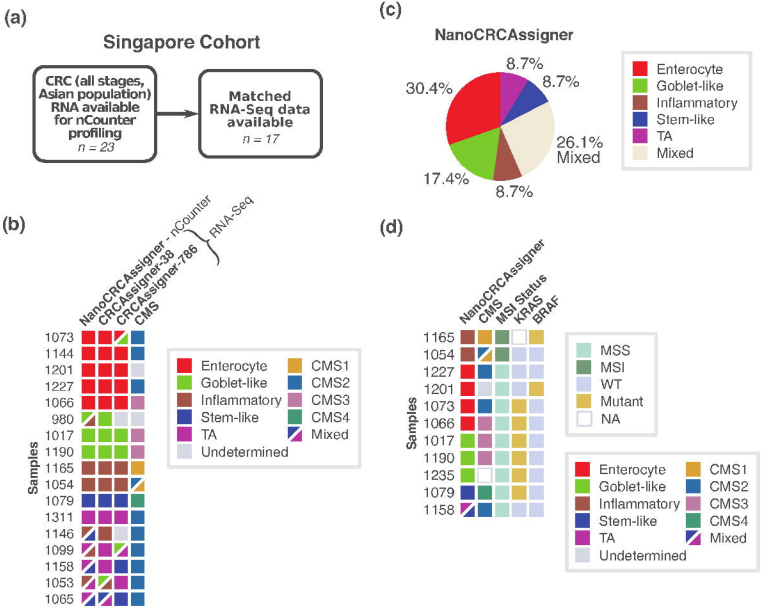

(a)
Singapore Cohort
CRC (all stages, Asian population) RNA available for nCounter profiling
n = 23
Matched RNA-Seq data available
n = 17
(b)
NanoCRCAssigner
CRCAssigner-39
CRCAssigner-786
CMS
nCounter
RNA-Seq
Samples
1073
1144
1201
1227
1066
980
1017
1190
1165
1054
1079
1311
1146
1099
1158
1053
1065
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
Undetermined
CMS1
CMS2
CMS3
CMS4
Mixed
(c)
NanoCRCAssigner
8.7%
8.7%
30.4%
26.1%
Mixed
17.4%
8.7%
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
Mixed
(d)
NanoCRCAssigner
CMS
MSI Status
KRAS
BRAF
Samples
1165
1054
1227
1201
1073
1066
1017
1190
1235
1079
1158
MSS
MSI
WT
Mutant
NA
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
Undetermined
CMS1
CMS2
CMS3
CMS4
Mixed

Figure 4

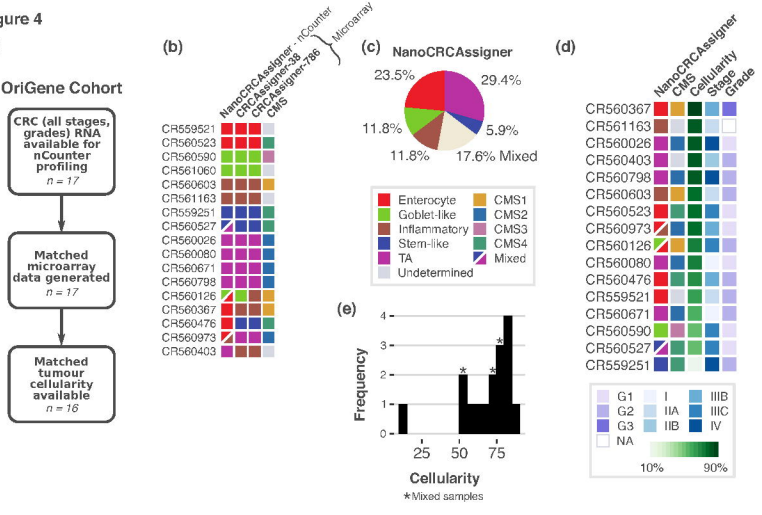
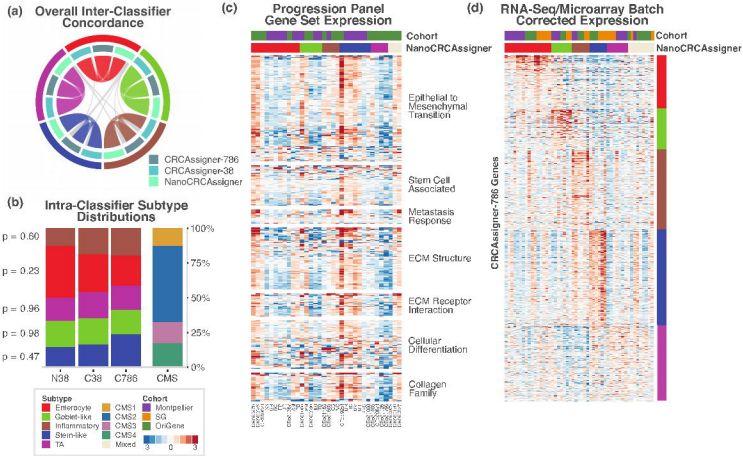

(a)
OriGene Cohort
CRC (all stages, grades) RNA available for nCounter profiling
n = 17
Matched microarray data generated
n = 17
Matched tumour cellularity available
n = 16
(b)
NanoCRCAssigner - nCounter
CRCAssigner-38
CRCAssigner-786
Microarray
CMS
CR559521
CR560523
CR560590
CR561060
CR560603
CR561163
CR559251
CR560527
CR560026
CR560080
CR560671
CR560798
CR560126
CR560367
CR560476
CR560973
CR560403
(c)
NanoCRCAssigner
23.5%
29.4%
5.9%
17.6% Mixed
11.8%
11.8%
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
Undetermined
CMS1
CMS2
CMS3
CMS4
Mixed
(d)
NanoCRCAssigner
CMS
Cellularity
Stage
Grade
CR560367
CR561163
CR560026
CR560403
CR560798
CR560603
CR560523
CR560973
CR560126
CR560080
CR560476
CR559521
CR560671
CR560590
CR560527
CR559251
G1
G2
G3
NA
I
IIA
IIB
IIIB
IIIC
IV
10%
90%
(e)
Frequency
0
1
2
3
4
25
50
75
Cellularity
*Mixed samples

Figure 5


(a)
Overall Inter-Classifier Concordance
CRCAssigner-786
CRCAssigner-38
NanoCRCAssigner
(b)
Intra-Classifier Subtype Distributions
p = 0.60
p = 0.23
p = 0.96
p = 0.98
p = 0.47
100%
75%
50%
25%
0%
N38
C38
C786
CMS
Subtype
Enterocyte
Goblet-like
Inflammatory
Stem-like
TA
CMS1
CMS2
CMS3
CMS4
Mixed
Cohort
Montpellier
SG
OnGene
3
0
3
(c)
Progression Panel Gene Set Expression
Cohort
NanoCRCAssigner
Epithelial to Mesenchymal Transition
Stem Cell Associated
Metastasis Response
ECM Structure
ECM Receptor Interaction
Cellular Differentiation
Collagen Family
(d)
RNA-Seq/Microarray Batch Corrected Expression
Cohort
NanoCRCAssigner
CRCAssigner-786 Genes