Identification and validation of a machine learning model of complete response to radiation in rectal cancer reveals immune infiltrate and TGFβ as key predictors

Enric Domingo,[a,m] Sanjay Rathee,[a,m] Andrew Blake,[a] Leslie Samuel,[b] Graeme Murray,[b] David Sebag-Montefiore,[c] Simon Gollins,[d] Nicholas West,[c] Rubina Begum,[e] Susan Richman,[c] Phil Quirke,[c] Keara Redmond,[f] Aikaterini Chatzipli,[g] Alessandro Barberis,[a] Sylvana Hassanieh,[a] Umair Mahmood,[a] Michael Youdell,[a] Ultan McDermott,[g] Viktor Koelzer,[h,i] Simon Leedham,[j] Ian Tomlinson,[a] Philip Dunne,[f] S:CORT consortium, Francesca M. Buffa,[a,k,n,*] and Timothy S. Maughan[a,l,n,**]

[a]Department of Oncology, Medical Sciences Division, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford, OX3 7DQ, UK
[b]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Foresterhill, Aberdeen, AB25 2ZD, UK
[c]Leeds Institute of Medical Research, University of Leeds, LS9 7TF, UK
[d]North Wales Cancer Treatment Centre, Besti Cadwaladr University Health Board, Bodelwyddan, Denbighshire, LL18 5UJ, UK
[e]Cancer Research & University College London Clinica Trial Unit, London, United Kingdom
[f]The Patrick G Johnston Centre for Cancer Research, Queens University Belfast, Belfast, BT7 9AE, UK
[g]Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1RQ, UK
[h]Department of Pathology and Molecular Pathology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[i]Department of Oncology and Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[j]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, UK
[k]Department of Computing Sciences, Bocconi University, Bocconi Institute for Data Science and Analytics (BIDSA), Milano, Italy
[l]Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, UK

Summary

Background It is uncertain which biological features underpin the response of rectal cancer (RC) to radiotherapy. No biomarker is currently in clinical use to select patients for treatment modifications.

Methods We identified two cohorts of patients (total N = 249) with RC treated with neoadjuvant radiotherapy (45Gy/25) plus fluoropyrimidine. This discovery set included 57 cases with pathological complete response (pCR) to chemoradiotherapy (23%). Pre-treatment cancer biopsies were assessed using transcriptome-wide mRNA expression and targeted DNA sequencing for copy number and driver mutations. Biological candidate and machine learning (ML) approaches were used to identify predictors of pCR to radiotherapy independent of tumour stage. Findings were assessed in 107 cases from an independent validation set (GSE87211).

Findings Three gene expression sets showed significant independent associations with pCR: Fibroblast-TGFβ Response Signature (F-TBRS) with radioresistance; and cytotoxic lymphocyte (CL) expression signature and consensus molecular subtype CMS1 with radiosensitivity. These associations were replicated in the validation cohort. In parallel, a gradient boosting machine model comprising the expression of 33 genes generated in the discovery cohort showed high performance in GSE87211 with 90% sensitivity, 86% specificity. Biological and ML signatures indicated similar mechanisms underlying radiation response, and showed better AUC and p-values than published transcriptomic signatures of radiation response in RC.

Interpretation RCs responding completely to chemoradiotherapy (CRT) have biological characteristics of immune response and absence of immune inhibitory TGFβ signalling. These tumours may be identified with a potential biomarker based on a 33 gene expression signature. This could help select patients likely to respond to treatment with a primary radiotherapy approach as for anal cancer. Conversely, those with predicted radioresistance may be candidates for clinical trials evaluating addition of immune-oncology agents and stromal TGFβ signalling inhibition.

*Corresponding author. Department of Oncology, Medical Sciences Division, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford, OX3 7DQ, UK.
**Corresponding author. Department of Oncology, Medical Sciences Division, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford, OX3 7DQ, UK.
    E-mail addresses: francesca.buffa@oncology.ox.ac.uk, francesca.buffa@unibocconi.it (F.M. Buffa), tim.maughan@oncology.ox.ac.uk, t.maughan@liverpool.ac.uk (T.S. Maughan).
[m]Co-first.
[n]Joint last authors.

Articles

Research in context

Evidence before this study

Patients with RC are usually treated with radiotherapy as part of neoadjuvant therapy although they show a wide range of responses. Currently there are no clinical or molecular biomarkers implemented in the clinic to select which patients benefit or not from such treatment. We searched PubMed with the terms "rectal", "radiotherapy", and "expression", for articles published up to Oct 2, 2020. Nine studies reported a gene expression RNA signature for response to CRT in primary RC using pre-treatment biopsies. However, they had important limitations, notably low statistical power, with the largest discovery set being only 77 cases, and a missing validation set in 3 cases. Furthermore, most studies had an heterogeneous clinical setting including diverse regimens of chemotherapy combined with conventionally fractionated radiotherapy.

Added value of this study

Our study used a discovery set of 249 cases, drawn from one large case series and the control arm of a national clinical trial.

Multi-omic profiling was undertaken by a multi-institution consortium using state of the art analytical tools. Advanced ML algorithms and methodology were used to detect previously defined gene sets associated with pCR. The findings were validated in an independent, external dataset of 107 cases and the predictive performance compared against all other published signatures. We reveal clear biological features (immune activity) underpinning response to radiation and define predictive biomarkers of pCR with excellent performance.

Implications of all the available evidence

We find that the stromal and immune cell compartments within tumours are key determinants of pCR to CRT in RC. Targeting these compartments provide the basis for novel studies using improved patient selection and combination treatments, which may be able to improve radiation response.

Introduction

Rectal Cancer (RC) accounts for 30% of the total incidence of colorectal cancer (CRC) and radiotherapy is used as part of the primary treatment of 40% patients with RC.[1] Pre-operative staging with MRI of the pelvis identifies patients with locally advanced RCs, in whom the surgical resection margin is threatened or involved, and in these people neoadjuvant treatment including chemoradiation or short course radiotherapy with pre-operative chemotherapy is indicated, with the intention of maximising the likelihood of achieving an R0 resection.[2] Neoadjuvant therapy also has an emerging role in the non-operative management of RC in patients who are reluctant or unfit to undergo radical surgery, or for those in whom pCR is observed in the 6–12 weeks following completion of CRT. In these patients a watch and wait policy may be undertaken and a recent meta-analysis shows increased local recurrence rates but long term outcomes comparable in clinical complete responders to total mesorectal excision.[3]

Prediction of response to CRT in RC has not been possible to date despite a number of small scale attempts to define predictive biomarkers from the transcriptome.[4–12] This is plausibly owing to a number of limitations in such studies such as different

chemotherapy combinations with radiotherapy, unmatched endpoints mostly based on different cut points to call response, lack of correction for stage, critically low statistical power, and absence of a validation cohort (Supplementary Table S1). Notably, reported biological insights from such studies are few and inconsistent. The lack of biomarkers to predict pCR to neoadjuvant radiotherapy is a clinical unmet need. Such a stratifier could enable a proportion of patients with high likelihood of achieving pCR to be offered definitive, radiation based treatment like in anal cancer, enabling a significant number of patients to be treated without surgery and avoid a stoma. In contrast, identification of those with poor response to CRT, who also have a higher rate of distant metastases,[13] could spare them the toxic effects and inconveniences of such treatment. It may also enable selection of patients for suitably modulated first line therapy designed to target the underlying biology driving radioresistance and metastasis.

In this study from the Stratification in COloRectal cancer (S:CORT) consortium, we aimed to identify the biological basis of complete response to radiotherapy in patients with RC and to derive a predictive biomarker. We have undertaken multi-omic profiling of 249 pre-radiotherapy biopsies from two cohorts of patients with

RC, comprising transcriptome-wide mRNA expression, mutations in 80 CRC driver genes and genome-wide copy number alterations. Using multiple regression and ML approaches we identify and validate biological predictors of pCR to radiotherapy in RC with a level of performance that suggests clinical utility.

Methods

Clinical cohorts

Two independent cohorts of patients with RC were included in our discovery set: 125 cases from a sequential cohort from the Aberdeen area in the UK ('Grampian cohort') and 124 cases from the control arm of the UK multicentre Aristotle clinical trial which compared the efficacy of standard CRT with or without irinotecan (ISRCTN09351447). Patients identified received 'standard CRT' comprising pelvic irradiation (45–50.4Gy in 25 fractions over 5 weeks) with capecitabine 900 mg/m^2 bd days Monday to Friday, throughout radiotherapy (Supplementary Figure S1a and b, Supplementary Methods) based on a threatened or involved circumferential rectal fascia on pre-treatment MRI scan. Sex was included as a covariable. Validation testing was performed on 107 cases from a publicly available RC cohort (GSE87211)[14] with transcriptomic data from pre-treatment biopsies, selected to have been treated with similar neoadjuvant regimen (50.4Gy in 28 fractions with infusional 5-fluorouracil alone) expected to have equivalent biological effects. Full details for the 3 cohorts are available in Supplementary Methods. The primary endpoint in all cases was pCR after CRT as assessed by specialist pathologist (GIM for Grampian and NPW for Aristotle).

Profiling

All clinical samples assessed were taken from Formalin Fixed Paraffin Embedded (FFPE) RC biopsies from patients before commencement of CRT given as neoadjuvant therapy for RC management. After assessment of haematoxylin and eosin-stained sections, areas of tumour were macrodissected and nucleic acids were extracted using standard protocols. A specifically designed panel of RNA baits (Agilent SureSelect) enabled capture and sequencing of all exons of 80 CRC driver genes, 66 regions of recurrent copy number gains/losses, 960 reference SNPs distributed across the genome (allowing low resolution copy number estimations) and 123 regions for microsatellite instability (MSI). Samples showing neutral copy number calls in >20% of the length of all chromosomes combined were classified negative for Chromosomal Instability (CIN), otherwise as positive.[15] RNA expression profiling used the 24,441 genes/110,961 probesets ALMAC/Affymetrix XCEL microarrays (Supplementary Methods).

Statistics

Twenty pre-defined, hypothesis-based candidate gene sets, pathways or molecular classifiers for response to radiotherapy were selected by an expert panel of researchers and tested for association with pCR using logistic regression analyses (Supplementary Table S2, Supplementary Methods).

In a second, hypothesis-free analysis, our aim was to build a machine learning (ML) model to predict pCR from the transcriptomic data. A pipeline consisting of five steps was established and applied using 12 different modelling approaches: Quality check, Pre-processing, including correction of class imbalance, differentially expressed genes (DEG) selection, Decision making genes (DMG) selection, Training to build a model (ML) from the discovery cohort as previously described.[16] The best performing ML model was tested on the full GSE87211 validation cohort after appropriate batch correction to derive sensitivity, specificity, and area under curve. Additional models were also tested. The complete methodology is provided in Supplementary Methods.

Additionally, a literature search identified 9 reported signatures for prediction of response to radiotherapy in RC (Supplementary Table S1) which were also tested in the validation cohort.

Ethics

The S:CORT consortium including this specific analysis was reviewed and approved by the South Cambs Research Ethics committee (REC ref 15/EE/0241; IRAS reference 169363). All patients provided written informed consent for further research to be undertaken on samples.

Role of funders

The Stratification in Colorectal Cancer Consortium (S:CORT) was funded by the Medical Research Council and Cancer Research UK. The funders played no role in the study design, data collection, data analyses performed, interpretation or writing of the report.

Results

Clinical and molecular profiles

Most selected samples (87%) from both Grampian (N = 125) and Aristotle (N = 124) were successfully profiled for both transcriptome and targeted NGS, with remaining cases profiled for one of them (Fig. 1a, Supplementary Figure S1a and b). The frequency of pCR in Grampian, Aristotle and GSE87211 cohorts was 26%, 19% and 21% respectively. Both pretreatment T and N stage showed some variation (Fig. 1b, Supplementary Table S3) while the frequencies of the main molecular profiles were within expected ranges, albeit different for some genetic variables such as *APC*
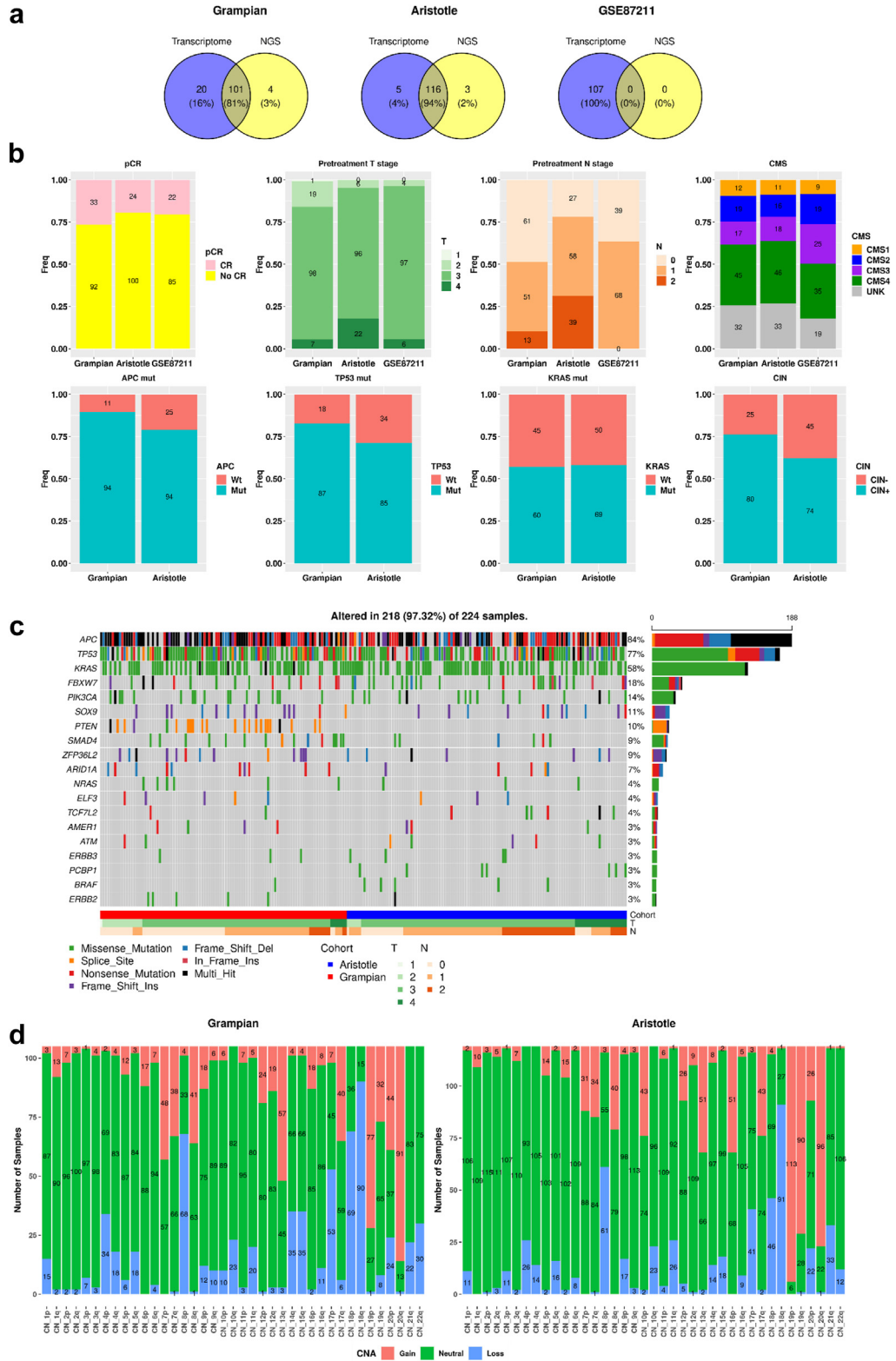
Fig. 1: Clinical and molecular data in Grampian, Aristotle and GSE87211. a. Overlap of samples profiled for RNA and DNA platforms. b. Comparison of the main clinical and molecular profiles by cohort. c. Most common driver mutations by gene. d. Copy number alterations by chromosome arm. (note RNA data only available from GSE87211).
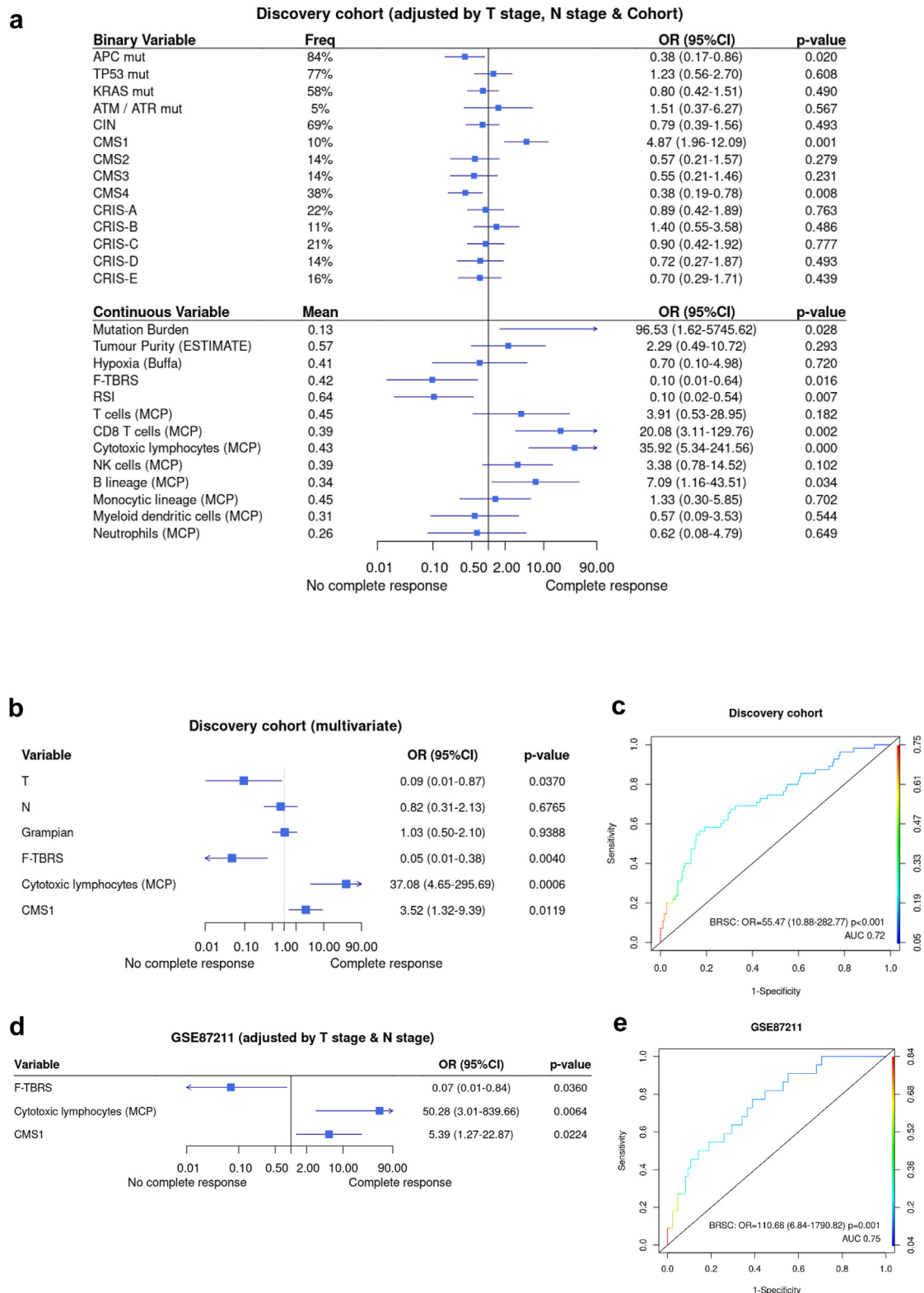
a

Discovery cohort (adjusted by T stage, N stage & Cohort)

Binary Variable	Freq		OR (95%CI)	p-value
APC mut	84%		0.38 (0.17-0.86)	0.020
TP53 mut	77%		1.23 (0.56-2.70)	0.608
KRAS mut	58%		0.80 (0.42-1.51)	0.490
ATM / ATR mut	5%		1.51 (0.37-6.27)	0.567
CIN	69%		0.79 (0.39-1.56)	0.493
CMS1	10%		4.87 (1.96-12.09)	0.001
CMS2	14%		0.57 (0.21-1.57)	0.279
CMS3	14%		0.55 (0.21-1.46)	0.231
CMS4	38%		0.38 (0.19-0.78)	0.008
CRIS-A	22%		0.89 (0.42-1.89)	0.763
CRIS-B	11%		1.40 (0.55-3.58)	0.486
CRIS-C	21%		0.90 (0.42-1.92)	0.777
CRIS-D	14%		0.72 (0.27-1.87)	0.493
CRIS-E	16%		0.70 (0.29-1.71)	0.439
Continuous Variable	**Mean**		**OR (95%CI)**	**p-value**
Mutation Burden	0.13		96.53 (1.62-5745.62)	0.028
Tumour Purity (ESTIMATE)	0.57		2.29 (0.49-10.72)	0.293
Hypoxia (Buffa)	0.41		0.70 (0.10-4.98)	0.720
F-TBRS	0.42		0.10 (0.01-0.64)	0.016
RSI	0.64		0.10 (0.02-0.54)	0.007
T cells (MCP)	0.45		3.91 (0.53-28.95)	0.182
CD8 T cells (MCP)	0.39		20.08 (3.11-129.76)	0.002
Cytotoxic lymphocytes (MCP)	0.43		35.92 (5.34-241.56)	0.000
NK cells (MCP)	0.39		3.38 (0.78-14.52)	0.102
B lineage (MCP)	0.34		7.09 (1.16-43.51)	0.034
Monocytic lineage (MCP)	0.45		1.33 (0.30-5.85)	0.702
Myeloid dendritic cells (MCP)	0.31		0.57 (0.09-3.53)	0.544
Neutrophils (MCP)	0.26		0.62 (0.08-4.79)	0.649

0.01 0.10 0.50 2.00 10.00 90.00
No complete response   Complete response

b

Discovery cohort (multivariate)

Variable	OR (95%CI)	p-value
T	0.09 (0.01-0.87)	0.0370
N	0.82 (0.31-2.13)	0.6765
Grampian	1.03 (0.50-2.10)	0.9388
F-TBRS	0.05 (0.01-0.38)	0.0040
Cytotoxic lymphocytes (MCP)	37.08 (4.65-295.69)	0.0006
CMS1	3.52 (1.32-9.39)	0.0119

0.01 0.10 1.00 10.00 90.00
No complete response   Complete response

c

Discovery cohort

BRSC: OR=55.47 (10.88-282.77) p<0.001
AUC 0.72

d

GSE87211 (adjusted by T stage & N stage)

Variable	OR (95%CI)	p-value
F-TBRS	0.07 (0.01-0.84)	0.0360
Cytotoxic lymphocytes (MCP)	50.28 (3.01-839.66)	0.0064
CMS1	5.39 (1.27-22.87)	0.0224

0.01 0.10 0.50 2.00 10.00 90.00
No complete response   Complete response

e

GSE87211

BRSC: OR=110.66 (6.84-1790.82) p=0.001
AUC 0.75

Fig. 2: Candidate analysis. a. Univariable regression adjusted for T and N stage for prediction of pCR in candidate biological features in discovery cohort (Grampian and Aristotle combined). b. Multivariable model adjusted for T and N stage after stepwise backwards regression in

and *TP53* mutations and CIN that were more common in Grampian than Aristotle (Fig. 1c and d, Supplementary Figure S2a Supplementary Table S3). Summaries of the most common driver mutations, copy number alterations at chromosome level and transcriptome signatures are shown in Fig. 1 and Supplementary Figure S2.

Hypothesis-based analysis

To obtain robust predictive models, Grampian and Aristotle cohorts were merged and relabelled as Discovery cohort (Supplementary Figure S5). Then, logistic regression adjusted by cohort and baseline T and N stage was performed to identify predictors of pCR on 20 prespecified biological candidates (Supplementary Table S2) from our multiomic data. Nine features were significant (Fig. 2a): five associated positively with pCR were all indicators of immune cell abundance (CMS1, mutation burden, CD8 Tcells, cytotoxic lymphocytes and B lineage (by MCP)) and 4 negatively associated (APC mutation, CMS4, F-TBRS and RSI (radiosensitivity index)). After multivariable backward stepwise regression, only three remained independently significant F-TBRS (OR 0.05, p 0.004), cytotoxic lymphocytes (OR 37.08, p 0.0006) and CMS1 (OR 3.52, p 0.012) (Fig. 2b). This final model, which we have labelled as Biological RadioSensitivity Classifier (BRSC), in a ROC curve predicted pCR with AUC of 72% in the discovery set (Fig. 2c). BRSC performance was reproducible in the validation set in prediction of pCR (p = 0.001) showing AUC of 75% (Fig. 2d and e). Its addition to a model with T and N stage was highly significant (p = 0.0008, likelihood ratio test). Meta-analysis of these three variables (CMS1, cytotoxic lymphocytes and F-TBRS) in the three tested datasets showed no significant heterogeneity between cohorts (Supplementary Figure S3).

Analysis of genetic variables

We interrogated other multiomic profiles not tested as candidates. None of the mutations or copy number changes were significant (Supplementary Figure S4).

ML analysis

Using multiple ML methods (Supplementary Methods, Fig. 3a, Supplementary Figure S6) we aimed to develop a locked ML model as recommended by FDA.[16] We compared their performance in cross-validation, and identified in the discovery cohort the gradient boosting machine as providing the optimal geneset based on highest K-fold cross validation accuracy (84%) with

lowest number of genes required (33) (Fig. 3b, Supplementary Table S4). This locked model applied to all GSE87211 showed 89% accuracy and 89% AUC (Fig. 3c and d) with 89% sensitivity and 86% specificity (Fig. 3e), showing excellent ability to differentiate between pCRs and non-pCRs. Similar results were found when using a subset of GSE87211 balanced for pCR (Supplementary Figure S7b–d). We labelled this new geneset as RadioSensitive Signature (RSS). Considering the difference between the training and validation cohorts, batch correction was necessary (Supplementary Figure S7). Based on our research we identified the SVA tool (Combat function) as the most relevant algorithm for batch correction.[17] This requires some processing of the validation data. In particular, to avoid degrading the pCR signal during the batch correction, pCR data has to be disclosed to the batch correction function to properly merge the discovery and GSE87211 cohorts (full discussion on this is provided in Supplementary Methods). However RSS was also tested for validation without standardisation of the analytical platform by using a version of the validation transcriptome built fully agnostically from outcome. This procedure is suboptimal from a batch correction perspective, but it is an important validation that RSS was still predictive in a setting not disclosing the pCR data (64% accuracy, 74% AUC, p = 0.04) (Supplementary Figure S7e–g).

Finally, we performed a randomized permutation experiment to further test our geneset. An empirical null was generated by randomly picking 1000 gene sets (each containing 33 genes). The AUC for the discovery (Supplementary Figure S7h) & validation cohort (Supplementary Figure S7i and j) estimated using this empirical null was found significant (p-value <0.001, Predictor p-values in linear models).

BRSC and RSS across grades of clinical response

We have used pCR as a valuable clinical endpoint with a clear biological meaning. We asessed whether RSS might also provide additional granular information for broader response to radiotherapy than specifically for pCR. We checked the distribution of RSS scores across 4 levels of pathological response in the single cohort (Grampian) with these data available and across pathological T stage after treatment in the whole Discovery cohort and GSE87211 (Supplementary Figure S8a–c). pCRs and yT0 cases showed higher levels of RSS but all the other categories showed comparable, lower levels. The same was observed for BRSC scores (Supplementary Figure S8d–f). These results suggest

discovery cohort. c. ROC curve applying the 3 variables (F-TBRS, Cytotoxic lymphocytes and CMS1) as one compound variable in discovery cohort. d. Univariable model of the 3 variables (F-TBRS, Cytotoxic lymphocytes and CMS1) adjusted for T and N stage in GSE87211. e. ROC curve applying the 3 BRSC variables combined in GSE87211.
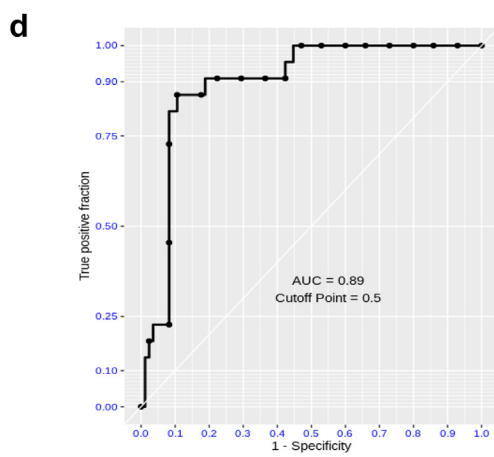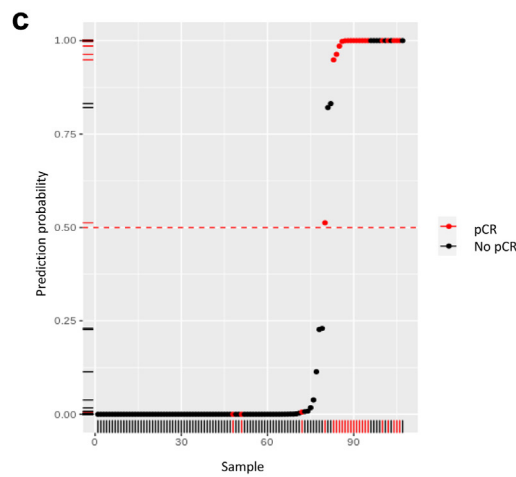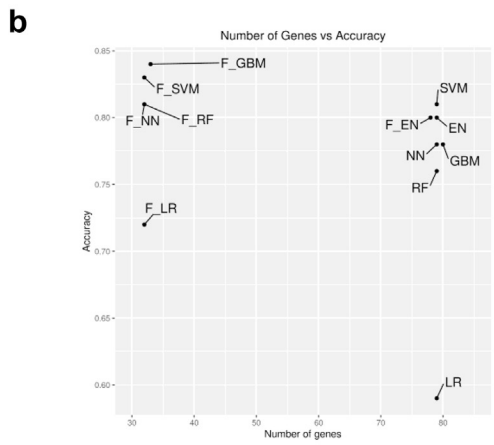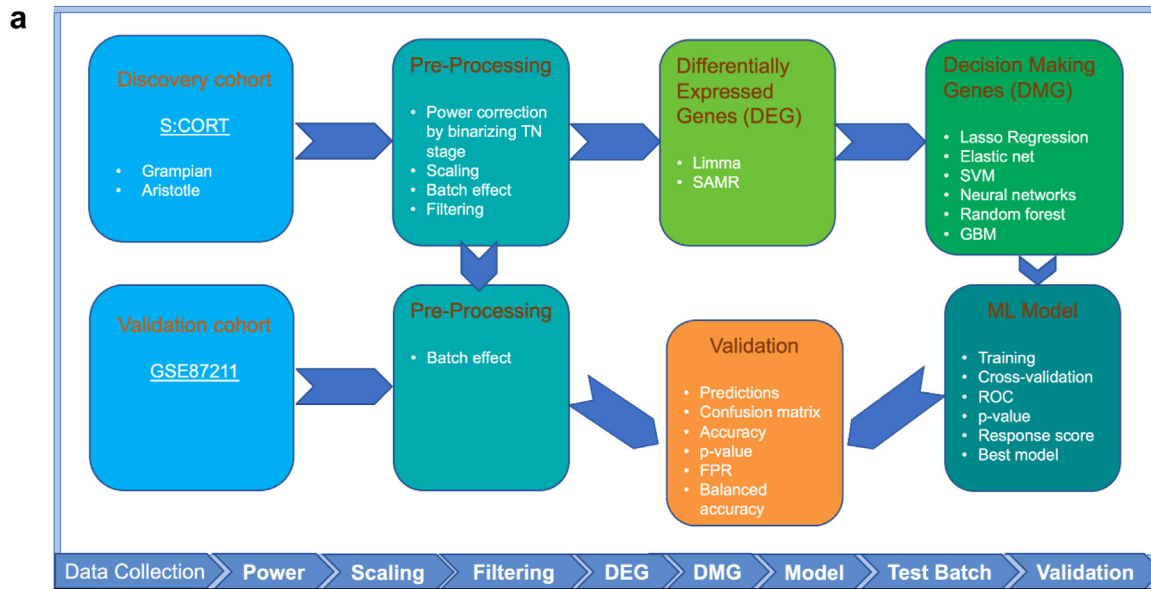
Fig. 3: Machine learning. a. Analytical pipeline used to derive and validate a new signature to predict pCR. b. 10-fold cross validation accuracy and number of genes for each ML method tested. The one with highest accuracy was selected. EN: elastic net, GBM: gradient boosting machine,

the inherent biology uncovered might be specific to pCR rather than good response to radiotherapy.

Biological assessment of RSS

In order to understand the biology associated with RSS, we first performed gene set enrichment analysis (GSEA) of the RSS model predictions on GSE87211 using hallmark genesets. We compared "RSS positive" (eg predicted to be pCRs) against "RSS negative" (eg predicted to be not-pCRs). The same analysis was also performed with real pCR status for a fair comparison (Fig. 4a). We identified 15 hallmarks in common between both analyses, all of them in the same direction, while four hallmarks were identified only by RSS predicted status and also four only by pCR status. These results suggest strong concordance of RSS prediction with actual pCR status also at the level of biological pathways. In agreement with our previous BRSC results, two of the shared hallmarks showing stronger enrichment in pCRs were immune-related (interferon alpha response and interferon gamma response) while a hallmark tightly associated with stroma (epithelial mesenchymal transition) was strongly enriched in not-pCRs cases. Some other interesting hallmarks linked in the literature with radiotherapy were found such as hypoxia and oxidative phosphorylation.

We then aimed at comparing RSS with our BRSC model that underlines the biology driving pCR. The association of each individual RSS gene with each BRSC biological variable was analysed by regression and meta-analysed. As expected, the discovery set showed strong association for all 3 BRSC features without heterogeneity in the models. GSE87211 showed the expected trends although in this smaller cohort with more heterogeneity, two of the variables did not reach significance. The same analysis was then performed on RNAseq data from TCGA CRC resections. In this third large set all 3 BRSC features were significantly associated with RSS genes, albeit with high levels of heterogeneity (Fig. 4b and Supplementary Figure S9). These analyses suggest the 33 RSS genes broadly overlap with all 3 BRSC features.

Our biological analyses suggest RSS is strongly associated with the tumour microenvironment. Accordingly, we aimed at testing whether the expression of these 33 genes may potentially be driven by the composition of different cell types in the profiled tissue. Using an unrelated transcriptome cohort composed of four different cell types separated by FACS from six CRC cases,[18] we identified 14 RSS genes strongly expressed in specific cell types (epithelial, leucocytes, endothelial or fibroblasts) and 12 genes that were not (Supplementary Figure S10).

Comparison with similar published signatures

We compared RSS and BRSC with similar signatures reported in the literature to predict outcome to radiotherapy in RC (Supplementary Table S1). We first found very few overlapping genes between any of the signatures (Fig. 5a, Supplementary Figure S11a). We then looked for correlations in GSE87211 which was not used for training any of the signatures. RSS was only correlated with BRSC but not with any of the published signatures (Fig. 5b). We then compared how samples would have been ranked according to each signature. None of them seem to show similar ranks to RSS (Fig. 5c) but most of them seemed to show an inverse association for BRSC (Supplementary Figure S11b). Finally, we tested the prediction ability for pCR using each signature score. Only Palma and Park were statistically significant but at much lower level than BRSC and RSS (p = 0.047, p = 0.038, p = 0.002 and p < 0.001 respectively). Notably, better performance of our two models was clearly evidenced by better ORs and AUCs (Fig. 5d). In summary, most of these published signatures mildly associate with the same biology we have identified but they all clearly underperfom to predict pCR compared to RSS.

Discussion

CRC biology remains a challenge to the research and clinical communities. The 15% of MSI cases shows the importance of the interaction between epithelial cell biology and immune response as key determinant of outcome. The development of the CMS classification has helped to stratify the remaining 85% of CRC with proficient mismatch repair.[19] While MSI cases are allocated with some other cases with immune activation to CMS1, the key observation from the CMS and prior gene expression classification systems has been the identification of CMS4 with its increase in fibrotic stroma and prevalence of TGFβ signalling, poor prognosis and poor response to chemotherapy. Recently, we reported that image based consensus molecular subtypes correlate with radiation response with increased pCRs in imCMS1, but reduced pCRs in imCMS4.[20]

This study has shown that the predictors associated with pCR to radiotherapy are based on contributions from the tumour microenvironment, notably stromal cells (TGFβ signalling) and immune response (CMS1, cytotoxic lymphocytes). Increasing evidence points to the immune response as being a critical determinant of

LR: lasso regression, NN: neural net; RF: random Forest; SVM: support-vector machine. The prefix F- refers to functional. c. RSS predictive scores from Gradient boosting machine model on GSE87211. d. ROC curve for Gradient boosting machine model on GSE87211. e. Confusion matrix for the new RSS signature in GSE87211.
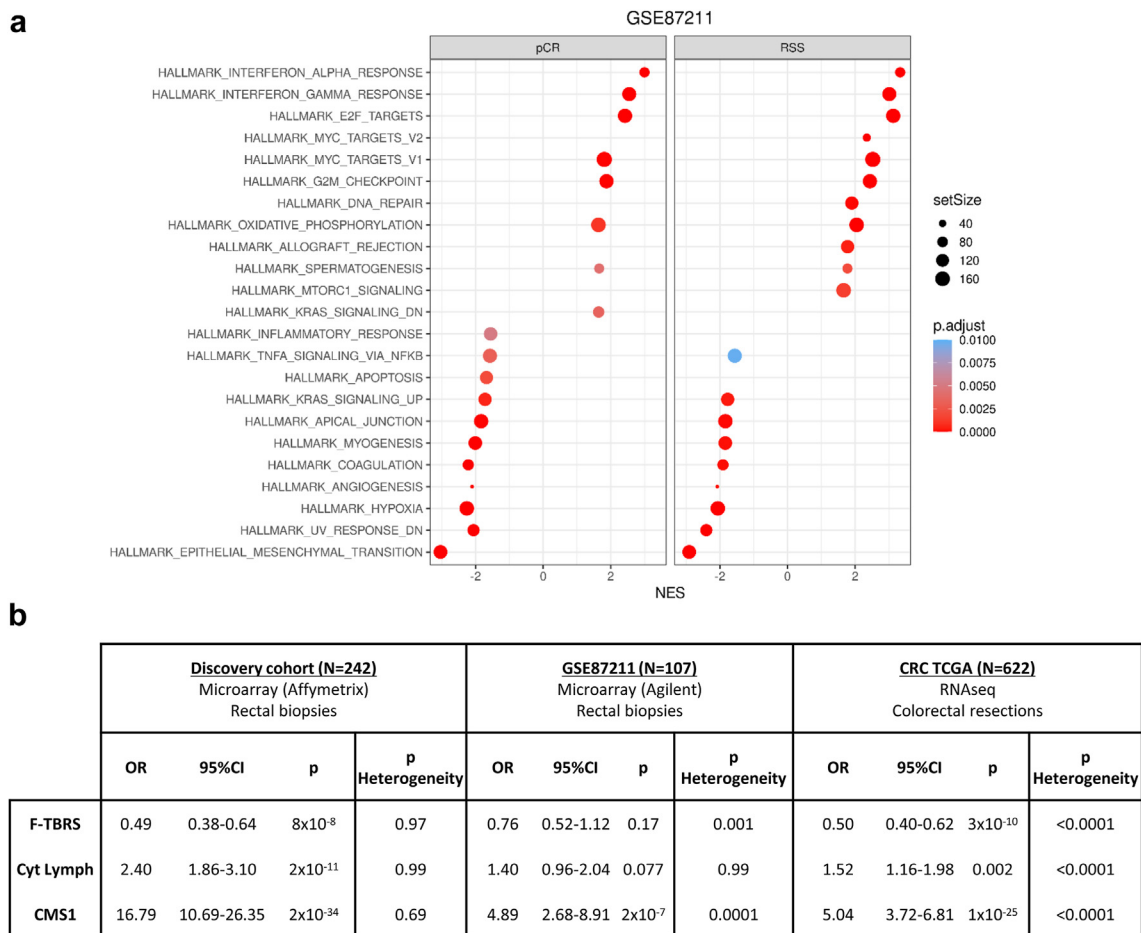
a

b

	Discovery cohort (N=242) Microarray (Affymetrix) Rectal biopsies				GSE87211 (N=107) Microarray (Agilent) Rectal biopsies				CRC TCGA (N=622) RNAseq Colorectal resections			
	OR	95%CI	p	p Heterogeneity	OR	95%CI	p	p Heterogeneity	OR	95%CI	p	p Heterogeneity
F-TBRS	0.49	0.38-0.64	8×10^{-8}	0.97	0.76	0.52-1.12	0.17	0.001	0.50	0.40-0.62	3×10^{-10}	<0.0001
Cyt Lymph	2.40	1.86-3.10	2×10^{-11}	0.99	1.40	0.96-2.04	0.077	0.99	1.52	1.16-1.98	0.002	<0.0001
CMS1	16.79	10.69-26.35	2×10^{-34}	0.69	4.89	2.68-8.91	2×10^{-7}	0.0001	5.04	3.72-6.81	1×10^{-25}	<0.0001

Fig. 4: **RSS biology** a. GSEA for cancer hallmarks from differential expression analyses based on pCR and RSS prediction in GSE87211. b. Association of BRSC molecular features and RSS genes by meta-analysis for all genes in discovery cohort, GSE87211 and TCGA separately (see Supplementary Figure S8 for full analysis).

radiation sensitivity. Using 97 rectal biopsies from patients treated with heterogeneous neoadjuvant radiotherapy regimens, estimations of immune infiltrate were found associated with increased response and prolonged disease free survival.[21] In a separate report, gene expression from over 10,000 pancancer patients showed high RSI[22] had distinct enrichment of interferon-associated signalling pathways and immune cell infiltrates.[23] The importance of an IL-1 mediated response from inflammatory cancer associated fibroblasts has been shown to mediate radiation resistance in RC.[24] Consistently, here we report radiation response associates with TGFβ and immune markers. TGFβ is a pleiotropic cytokine derived from many cell types but especially the myofibroblastic stromal cells, which is present at a high level in stroma-rich CRC,[19] is released from its extracellular matrix trap by radiotherapy,[20] reduces intrinsic radiosensitivity through DNA repair pathway switching[25] and has multiple inhibitory effects

on the immune response.[26] These predictive features are plausible biologically and targetable, enabling design of combination therapy through inhibition of TGFβ or related stromal cytokine pathways, such as IL-1, and enhancing the immune cell engagement through immune checkpoint inhibition and other related approaches. Interestingly, this biology overlaps extensively with that driving the development of metastasis in the poor prognosis CMS4 subtype. Batlle and colleagues have shown in an autochthonous tumour model of CMS4 (APTK) which is highly metastatic, inhibition of TGFβ and with PD-1 inhibition is the most potent approach to inhibiting the development of metastases and enhancing survival.[27] This approach has been shown in early clinical trials using the TGFBR1 inhibitor vactosertib in metastatic CRC in combination with pembrolizumab.[28]

We complemented this hypothesis-based approach with a hypothesis-free, ML method undertaken
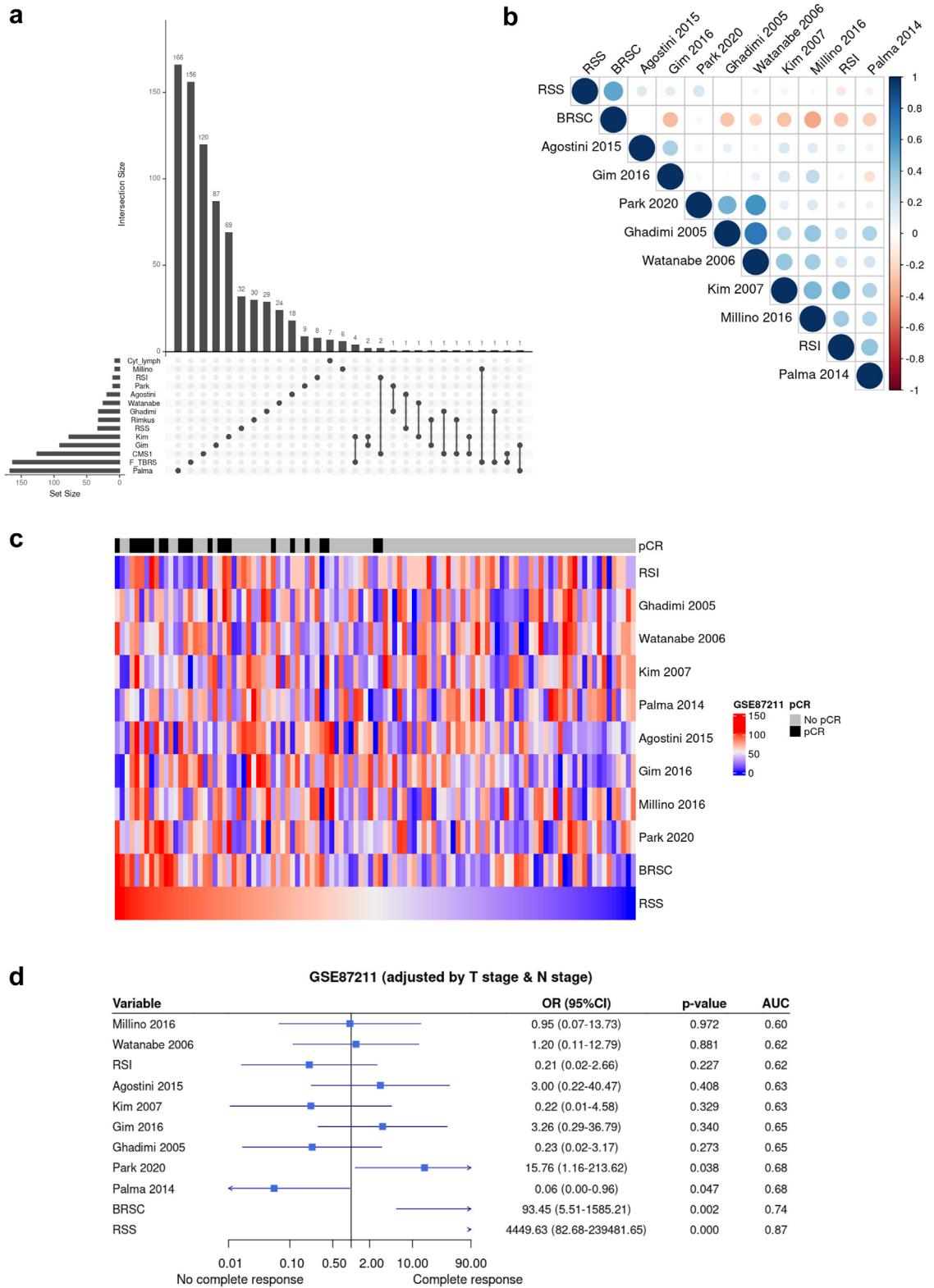
Fig. 5: Comparison of our signatures with published ones. a. Number of overlapping entrez genes in published signatures and ours. b. Pearson correlation across different signatures in GSE87211. c. Heatmap of samples ranked by each signature sorted by biological score of RSS in GSE87211. d. Prediction to pCR in GSE87211 by each signature sorted by AUC.

separately using the same datasets. The availability of more complex and personalised statistical tools for gene expression data analysis coupled with an unprecedented large dataset enabled us to derive robust biomarkers. A dataset without proper pre-processing to handle batch effect, causal inference, class imbalance and outliers can lead to misleading results.[29] Given the validation set was processed in an entirely different laboratory setting, we performed batch correction. This requires disclosure of clinical variables, including response, in order to avoid degrading the biological signal. However, we also performed an analysis not revealing the response data during batch correction, which is suboptimal from a batch correction perspective, but avoids disclosing the response data. We also used a set of pre-processing techniques including correction of class imbalance using downsampling to balance classes in the discovery cohort and multiple iterations using differing not-pCR cases to minimise biases in our discovery cohort. After addressing most of the known biases, DEG analysis was done on 100 subsets of the discovery cohort with two tools finding a robust geneset of 80 DEGs in all iterations. Finally, we used six well known ML algorithms trained on all 80 DEGs or selected features to make prediction models and internal validation was undertaken using 10-fold cross-validation. This identified a RSS of 33 genes which is independent of T and N stage and highly predictive of pCR with high robustness and accuracy in an independent dataset with optimal pre-processing, which remained significant when omitting pCR from the combat step. GSEA analysis of samples with either high RSS scores or with actual pCR, showed interferon alpha response and interferon gamma response were highly enriched in responding patients while a hallmark tightly associated with stroma (epithelial mesenchymal transition) was strongly enriched in not-pCRs cases showing the two methods had independently shown the same biological features.

The ML approach based on rigorous statistics has therefore provided us with the basis for a biomarker which could be tested prospectively to identify patients with RC who are highly likely to achieve pCR to standard dose radiotherapy coupled with fluoropyrimidines. The 15–25% of patients who have the radiosensitive biology could be identified at diagnosis and counselled that a primary radiotherapy based approach to management is possible similar to that currently undertaken in anal cancer, with planned close follow up on a 'watch and wait' basis as is currently performed in those found to have a complete clinical response after neoadjuvant CRT. Conversely, it identifies patients unlikely to obtain pCR who may be considered for trials targeting the stromal inhibitory mechanisms (TGFβ) and enhancing the immune response in combination with radiation. Further research may also test whether RSS may also be useful in earlier stage disease for patients currently not considered for CRT who may achieve an organ preserving strategy (Fig. 6, Supplementary Figure S12).

Only recently RNA profiling has been successfully undertaken using optimised 3' RNAseq from FFPE tissue. In our study we used the Almac Xcel array which is optimised for FFPE samples and we were able to achieve good yields from 98% of rectal biopsies. In collaboration with the Wellcome Trust Sanger Institute we also developed a reduced input to the sequencing process enabling 50 ng DNA to be used rather than the previous minimum standard of 200 ng. The use of FFPE obviously makes collection of cases much more straightforward and also makes the results more readily implementable in the clinical workflow as fresh frozen samples are not required.

The main strengths of our study compared to previous efforts are larger discovery and validation sets, strong sample selection resulting in clear clinical homogeneity, adjustment of the potential confounders T and N baseline stage, assessment of the associated biology, additional use of mutation and copy number data, and the use of a clear endpoint both clinically and
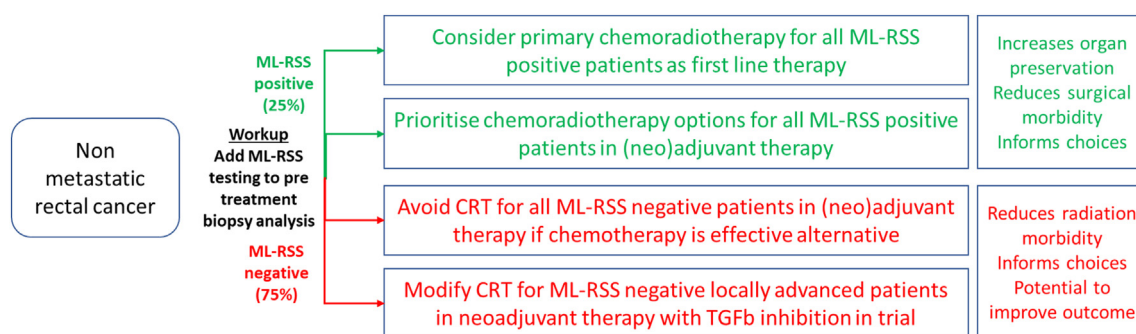

Fig. 6: Clinical implementation of RSS. NCCN guidelines define clinical options for patients with rectal cancer. Adding a certified test based on the 33 gene signature described here would inform decision making, resulting in increased organ preservation, reduced treatment morbidity, and has potential to improve outcomes.

biologically such as pCR. Although the latter may have resulted in some loss of signal from good responders not reaching pCR, our biomarkers showed similar low levels across responses and yT stage within non-pCRs. This may be expected since they were derived specifically for pCR. Overall, our results provide a comprehensive picture that was missing in this field.

Our research is limited to analysis of patients treated with long course CRT. Total noadjuvant chemotherapy (TNT) is an emerging standard of care based on randomised trials[30] and meta-analyses[31,32] showing improved outcomes. Recently, we have assessed predictors of response with varying chemoradiotherapy regimens and show that the addition of oxaliplatin to CRT seems to reduce the negative predictive effect of stromal biology on pCR.[33] Further research, including acquisition of transcriptomics data in the relevant cohorts, will be needed to fully determine whether our findings may apply to patients treated with TNT, and also to stage I and IV patients as they were mostly missing in our cohorts. It is also unclear whether the same level of signal may be shown in patients with regimens other than with addition of fluoropyrimidines or even in different tumour types. RSS needs further development using a standardised analytical laboratory process to avoid the requirement for batch correction and to be further validated in independent datasets using that locked down protocol to become a clinically validated tool. To our knowledge this is the first time that a predictive model has been used to predict individual patient outcome in an entirely independent dataset. Finally, while our two biomarkers do not show heterogeneity in our three curated cohorts, we also show early evidence that the signals from some of the individual RSS genes may come from different cell lineages. More studies are needed to properly detail the involvement of each RSS gene in different cell types and why their combination results in stable signals to predict pCR in rectal biopsies. Such analyses may also consider that while we identify three biological features independently associated with pCR in multivariable analysis (F-TBRS, CL, CMS1), six others were also found in univariable models and it can not be ruled out that any of them or others not profiled in this study may have a relevant role.

In summary, RCs that respond completely to radiotherapy have the biological characteristics of immune activation as identified by CMS1 and cytotoxic lymphocytic infiltration and an absence of immune inhibitory TGFβ signalling. These tumours may be identified by measurement of expression of a 33 geneset that merits further research for validation and development as a potential new biomarker. This could open the door to selection of responsive patients for treatment of patients with RC with a primary radiotherapy based approach as for anal cancer. Conversely those with predicted radioresistance would be candidates for clinical trials evaluating addition of immune-oncology agents and stromal signalling inhibition to overcome these drivers of immune exclusion and radioresistance.

Patients and/or the public were involved in the design and conduct of this work through the S:CORT consortium.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105228.

References

1 Chemotherapy, radiotherapy and tumour resections in England: 2013-2021. https://digital.nhs.uk/data-and-information/publications/statistical/chemotherapy-radiotherapy-and-surgical-tumour-resections-in-england/2013-to-2021.

2 Benson AB, Venook AP, Al-Hawary MM, et al. NCCN guidelines insights: rectal cancer, version 6.2020. *J Natl Compr Canc Netw.* 2020;18(7):806–815.

3 Yu G, Lu W, Jiao Z, Qiao J, Ma S, Liu X. A meta-analysis of the watch-and-wait strategy versus total mesorectal excision for rectal cancer exhibiting complete clinical response after neoadjuvant chemoradiotherapy. *World J Surg Oncol.* 2021;19(1):305.

4 Ghadimi BM, Grade M, Difilippantonio MJ, et al. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *J Clin Oncol.* 2005;23(9):1826–1838.

5 Gim J, Cho YB, Hong HK, et al. Predicting multi-class responses to preoperative chemoradiotherapy in rectal cancer patients. *Radiat Oncol.* 2016;11:50.

6 Kim IJ, Lim SB, Kang HC, et al. Microarray gene expression profiling for predicting complete response to preoperative chemoradiotherapy in patients with advanced rectal cancer. *Dis Colon Rectum.* 2007;50(9):1342–1353.

7 Park IJ, Yu YS, Mustafa B, et al. A nine-gene signature for predicting the response to preoperative chemoradiotherapy in patients with locally advanced rectal cancer. *Cancers (Basel).* 2020;12(4):800.

8 Millino C, Maretto I, Pacchioni B, et al. Gene and MicroRNA expression are predictive of tumor response in rectal adenocarcinoma patients treated with preoperative chemoradiotherapy. *J Cell Physiol.* 2017;232(2):426–435.

9 Watanabe T, Komuro Y, Kiyomatsu T, et al. Prediction of sensitivity of rectal cancer cells in response to preoperative radiotherapy by DNA microarray analysis of gene expression profiles. *Cancer Res.* 2006;66(7):3370–3374.

10 Rimkus C, Friederichs J, Boulesteix AL, et al. Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. *Clin Gastroenterol Hepatol.* 2008;6(1):53–61.

11 Palma P, Cano C, Conde-Muiño R, et al. Expression profiling of rectal tumors defines response to neoadjuvant treatment related genes. *PLoS One.* 2014;9(11):e112189.

12 Agostini M, Zangrando A, Pastrello C, et al. A functional biological network centered on XRCC3: a new possible marker of chemo-radiotherapy resistance in rectal cancer patients. *Cancer Biol Ther.* 2015;16(8):1160–1171.

13 Fokas E, Fietkau R, Hartmann A, et al. Neoadjuvant rectal score as individual-level surrogate for disease-free survival in rectal cancer in the CAO/ARO/AIO-04 randomized phase III trial. *Ann Oncol.* 2018;29(7):1521–1527.

14 Hu Y, Gaedcke J, Emons G, et al. Colorectal cancer susceptibility loci as predictive markers of rectal cancer prognosis after surgery. *Genes Chromosomes Cancer.* 2018;57(3):140–149.

15 Burrell RA, McClelland SE, Endesfelder D, et al. Replication stress links structural and numerical cancer chromosomal instability. *Nature.* 2013;494(7438):492–496.

16 Barberis A, Aerts HJWL, Buffa FM. Robustness and reproducibility for AI learning in biomedical sciences: RENOIR. *Sci Rep.* 2024;14:1933. https://doi.org/10.1038/s41598-024-51381-4.

17 Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf.

18 Fisher NC, Byrne RM, Leslie H, et al. Biological misinterpretation of transcriptional signatures in tumor samples can unknowingly undermine mechanistic understanding and faithful alignment with preclinical data. *Clin Cancer Res.* 2022;28(18):4056–4069.

19 Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med.* 2015;21(11):1350–1356.

20 Lafarge MW, Domingo E, Sirinukunwattana K, et al. Image-based consensus molecular subtyping in rectal cancer biopsies and response to neoadjuvant chemoradiotherapy. *NPJ Precis Oncol.* 2024;8(1):89. https://doi.org/10.1038/s41698-024-00580-3. PMID: 38594327; PMCID: PMC11003957.

21 Chatila WK, Kim JK, Walch H, et al. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. *Nat Med.* 2022;28(8):1646–1655.

22 Chatila WK, Kim JK, Walch H, et al. Systems biology modeling of the radiation sensitivity network: a biomarker discovery platform. *Int J Radiat Oncol Biol Phys.* 2009;75(2):497–505.

23 Grass GD, Alfonso JCL, Welsh E, et al. The radiosensitivity index gene signature identifies distinct tumor immune microenvironment characteristics associated with susceptibility to radiation therapy. *Int J Radiat Oncol Biol Phys.* 2022;113(3):635–647.

24 Nicolas AM, Pesic M, Engel E, et al. Inflammatory fibroblasts mediate resistance to neoadjuvant therapy in rectal cancer. *Cancer Cell.* 2022;40(2):168–184.e13.

25 Liu Q, Lopez K, Murnane J, Humphrey T, Barcellos-Hoff MH. Misrepair in context: TGFβ regulation of DNA repair. *Front Oncol.* 2019;9:799.

26 Batlle E, Massague J. Transforming growth factor-beta signaling in immunity and cancer. *Immunity.* 2019;50(4):924–940.

27 Tauriello DVF, Palomo-Ponce S, Stork D, et al. TGFbeta drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature.* 2018;554(7693):538–543.

28 Kim TW, Lee KW, Ahn JB, Hong YS. Efficacy and safety of vactosertib and pembrolizumab combination in patients with previously treated microsatellite stable metastatic colorectal cancer. *J Clin Oncol.* 2021;39(15_suppl):3573.

29 Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform.* 2020;2(3):lqaa078.

30 Bahadoer RR, Dijkstra EA, van Etten B, et al. Short-course radiotherapy followed by chemotherapy before total mesorectal excision (TME) versus preoperative chemoradiotherapy, TME, and optional adjuvant chemotherapy in locally advanced rectal cancer (RAPIDO): a randomised, open-label, phase 3 trial. *Lancet Oncol.* 2021;22(1):29–42.

31 Liu S, Jiang T, Xiao L, et al. Total neoadjuvant therapy (TNT) versus standard neoadjuvant chemoradiotherapy for locally advanced rectal cancer: a systematic review and meta-analysis. *Oncologist.* 2021;26(9):e1555–e1566.

32 Kasi A, Abbasi S, Handa S, et al. Total neoadjuvant therapy vs standard therapy in locally advanced rectal cancer: a systematic review and meta-analysis. *JAMA Netw Open.* 2020;3(12):e2030097.

33 Mahmood U, Blake A, Rathee S, et al. Stratification to neoadjuvant radiotherapy in rectal cancer by regimen and transcriptional signatures. *Cancer Research Comms.* 2024. https://doi.org/10.1158/2767-9764.CRC-23-0502. In press.