

Integrative ensemble modelling of cetuximab sensitivity in colorectal cancer patient-derived xenografts

Received: 10 January 2023

Accepted: 3 October 2024

Published online: 11 November 2024

 Check for updates

Umberto Perron^{1,15,17}, Elena Grassi^{2,3,17}, Aikaterini Chatzipl^{4,5,17}, Marco Viviani^{2,3}, Emre Karakoc⁴, Lucia Trastulla^{1,6}, Lorenzo M. Brochier^{1,7}, Claudio Isella^{2,3}, Eugenia R. Zanella², Hagen Klett⁸, Ivan Molineris⁹, Julia Schueler⁸, Manel Esteller^{10,11,12,13}, Enzo Medico^{2,3}, Nathalie Conte¹⁴, Ultan McDermott^{4,16}, Livio Trusolino^{2,3,18} ✉, Andrea Bertotti^{2,3,18} ✉ & Francesco Iorio^{1,4,18} ✉

Patient-derived xenografts (PDXs) are tumour fragments engrafted into mice for preclinical studies. PDXs offer clear advantages over simpler in vitro cancer models - such as cancer cell lines (CCLs) and organoids - in terms of structural complexity, heterogeneity, and stromal interactions. Here, we characterise 231 colorectal cancer PDXs at the genomic, transcriptomic, and epigenetic levels, along with their response to cetuximab, an EGFR inhibitor used clinically for metastatic colorectal cancer. After evaluating the PDXs' quality, stability, and molecular concordance with publicly available patient cohorts, we present results from training, interpreting, and validating the integrative ensemble classifier CeSta. This model takes in input the PDXs' multi-omic characterisation and predicts their sensitivity to cetuximab treatment, achieving an area under the receiver operating characteristics curve > 0.88. Our study demonstrates that large PDX collections can be leveraged to train accurate, interpretable drug sensitivity models that: (1) better capture patient-derived therapeutic biomarkers compared to models trained on CCL data, (2) can be robustly validated across independent PDX cohorts, and (3) could contribute to the development of future therapeutic biomarkers.

Colorectal cancer (CRC) is a heterogeneous disease with distinctly variable molecular features and responses to therapy. It is among the most prevalent causes of cancer mortality worldwide, with more than 1.85 million cases and 850,000 annual deaths globally¹. Around 20% of newly diagnosed CRC patients have metastatic disease (mCRC) at presentation, with 25% later developing metastases²⁻⁴.

In recent years, several clinical trials⁵⁻⁷ have suggested that genome-based treatment selection leads to therapeutic benefits for patients, reduced exposure to ineffective therapies, and median survival for mCRC patients exceeding 30 months⁸. Specifically, ~50% of mCRC patients have KRAS-NRAS-BRAF wild-type (triple negative)

tumours and are routinely treated with cetuximab and panitumumab, monoclonal antibody inhibitors of the epithelial growth factor receptor EGFR in combination with chemotherapy as an alternative to surgery. This protocol extends median survival by 2 to 4 months, compared with chemotherapy alone¹. Unfortunately, the overall mCRC clinical trial success rate remains low: 32% of combined phase II and III clinical trials failed between 2013 and 2015, up from 23% in 2010⁹. This highlights the need for more robustly predictive markers of drug response for CRC patients.

Biomarkers of response to cetuximab and cetuximab plus chemotherapy, such as the triple negative signature mentioned above,

have been derived from clinical and molecular analysis of patients and patient-derived experimental models of CRC, including immortalised cancer cell lines, organoids, and patient-derived xenografts (PDX)^{2,10–13}. However, several other systematic therapeutic biomarkers discovery efforts conducted using in vitro models have confirmed limited clinical translatability^{9,14,15}. This is primarily due to the intrinsic limitations of such models, encompassing genetic, epigenetic, and transcriptomic changes resulting from their selective adaptation to artificial culture conditions^{16,17}. Furthermore, cancer cell lines do not maintain the complex heterogeneity of the tumour of derivation; they often lose or gain specific subclones and might miss relevant components of the human tumour stromal microenvironment^{18,19}.

Unlike cancer cell lines, PDXs have been shown to offer good retention of tumour complexity, mimicking (at least to a certain extent) stromal interactions. They are relatively easy to screen and characterise. Further, histopathological characterisation has confirmed a high degree of concordance between PDXs and corresponding parental tumours in terms of differentiation, mucus secretion, and stromal composition, as well as maintenance of primary intratumoral clonal heterogeneity^{2,3,20–22}.

These factors have contributed to PDXs playing a pivotal role in translational cancer research, furthering our understanding of tumour biology and drug response mechanisms in CRC^{23,24}. As a result, extensive multi-institutional efforts (such as EuroPDX²⁵) are now ongoing, aiming to establish and characterise extensive collections of PDX models at the molecular and histopathological level to ensure that they recapitulate the broadest possible diversity of clinical cases²⁶.

Using data derived from the multi-omics characterisation of CRC PDXs paired with their pharmacological/phenotypic features is a profitable strategy for training supervised machine learning models to predict drug response in CRC patients. In this case, the extent of training data availability is a critical determinant of the accuracy of a model, especially when considering high-dimensional multi-omics datasets. Machine learning models of drug response trained on large pooled pan-cancer cell line datasets ($N=329$) outperform models which only used cell lines ($N=28–68$) from a specific tissue²⁷. This suggests that, in some cases, data quantity can outweigh data specificity. Kurilov and colleagues have also noted that predicting PDX drug response using models trained on cell line data results in poor performance across three out of four examined cohorts, except for the erlotinib lung cancer cohort²⁷.

In summary, most of the pre-clinical studies of cetuximab response in CRC cohorts performed to date have been characterised by (1) relatively small sample sizes, (2) single platform profiling often aimed at characterising the status of few known CRC driver genes, (3) reliance on biological models which have proved to be suboptimal for translational purposes, or a combination of these factors. These aspects negatively influenced the studies' ability to capture the tumour ecosystem's complexity and inter-tumour heterogeneity's impact on drug response, ultimately contributing to the increasingly low success rate of early-stage CRC clinical trials.

Here, we present one of the largest thoroughly characterised CRC PDX collections to date ($N=231$), which closely recapitulates gold-standard CRC patient cohorts across three 'omics (genomics, transcriptomics, and methylomics) and results from training an ensemble classifier predicting the response of these models to cetuximab treatment, based on an integrative stacked architecture.

Our model outperforms other state-of-the-art (SOTA) predictive methods and the biomarker of cetuximab response currently used in the clinic, i.e., the KRAS-NRAS-BRAF mutational status, internally and when tested on an independent cohort of CRC PDXs.

Finally, we show that our model's predictions provide an extent of interpretability, highlighting potential biomarkers of cetuximab sensitivity.

Results

We selected 231 first-pass CRC PDXs (the IRCC-PDX collection), which were fully characterised across multiple omics (encompassing genomics, transcriptomics, methylomics), clinical metadata, and were screened with cetuximab, from a larger cohort of >600 xenografts (Fig. 1a). These tumour models were uniquely derived from surgical resections of CRC liver metastases performed at the Candiolo Cancer Institute (Candiolo, Torino, IT), the Mauriziano Umberto I Hospital (Torino, IT), the San Giovanni Battista Hospital (Torino, IT) and the Niguarda Hospital (Milano, IT) between 2008 and 2015.

The initial 'raw' multi-omics characterisation of IRCC-PDX consisted of the methylation status of 700,298 Illumina probes, 33,670 gene transcription levels from RNAseq, 1272 copy number (CN) alteration and driver variant features, and 45 clinical features covering patient demographics, primary tumour characteristics, and previous patient treatment for a total of 735,285 features (Fig. 1a). In line with the clinical definition of 'disease control', which denotes clinical benefit, we categorised as 'responders' those PDXs in which cetuximab induced tumour shrinkage objective response (OR), more than 50% tumour volume reduction compared with baseline tumour volumes or stable disease (SD), less than 50% tumour shrinkage and less than 35% increase in tumour volume².

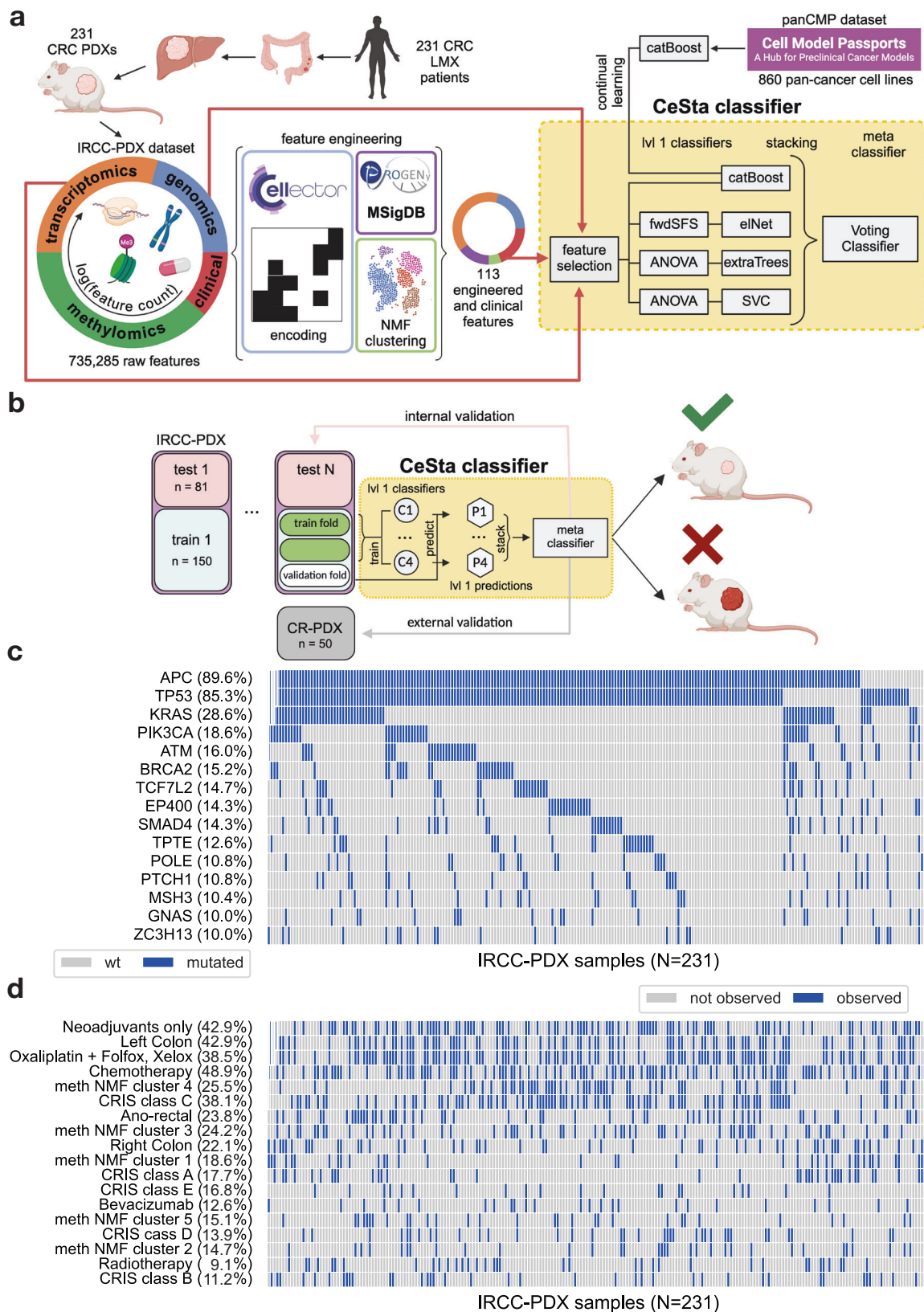
We performed several omic-specific feature engineering steps (Methods, Fig. 1a) before using this data with our integrative classifier (Fig. 1a, b). These aggregated some of the dimensions of the original 'raw' IRCC-PDX dataset (e.g., non-negative matrix factorisation clustering²⁸ of methylation features), introduced feature curation via prior knowledge of gene regulatory pathways, e.g., PROGENY²⁹ (11 features) and MSigDB³⁰ gene set analysis scores (50 features), generated potentially more informative collective feature-sets, e.g., CELLector³¹ genomic signatures (17 features), and retained 25 sufficiently curated clinical features for a final total number of 113 features across 231 PDXs (Supplementary Data 1).

Multi-omic characterisation of the IRCC-PDX collection

Previous comprehensive genetic characterisations of CRC models have shown that the frequency of common genetic mutations observed in PDXs is similar to that observed in primary tumours^{2,3,20,21,32,33}. Targeted sequencing of 116 genes in our PDX cohort identified 6426 driver mutations (Methods), with *APC* (observed in 90% of the IRCC-PDXs), *TP53* (85%), *KRAS* (29%), *PIK3CA* (19%), and *ATM* (16%) being the most frequently affected genes (Fig. 1c and Supplementary Fig. 1). In our PDX collection, mutational frequencies for *KRAS* and *BRAF* were lower than those reported for large CRC patient cohorts such as TCGA COAD/READ (<https://www.cancer.gov/tcga>) and MSK IMPACT³⁴ (<https://www.mskcc.org/msk-impact>). *KRAS*'s case is due to a pre-hoc enrichment of *KRAS* wild-type models for subsequent treatment with cetuximab (as *KRAS* mutant models were assumed to be cetuximab resistant a priori). In the case of *BRAF*, the lower frequency is ascribable to our PDXs being derived from metastatic samples. *BRAF* mutant tumours are frequently characterised by microsatellite instability (MSI). Because MSI CRCs have a better prognosis and rarely progress to metastasis³⁵, they are under-represented in our dataset. Indeed, after removing MSI samples, the frequency of *BRAF* mutated tumours in TCGA is reduced to 5.3%, which is comparable to that detected in our collection.

Aside from these exceptions, the mutational landscape of the IRCC-PDXs closely matched that of the previous CRC patient cohorts (Supplementary Fig. 2) and recapitulated known top frequently mutated CRC driver genes^{36,37}.

To further control our PDX models' ability to recapitulate characteristics of their tumour sample of origin, we investigated PDX mutational profile stability for a subset of more extended PDX lineages (i.e., those where targeted sequencing data was available beyond the first-passage; Supplementary Fig. 3). We observed a significant



agreement between all models belonging to a given lineage, regardless of their distance from their sample of origin in terms of passages, with few exceptions attributable to sequencing errors or clonal expansion (Supplementary Fig. 4).

CN alterations, derived from the same 116 genes in the targeted sequencing panel (Methods), affected some known CRC drivers,

including *EGFR* and *SMAD4*, and showed a positive correlation (Spearman $r = 0.87$ and 0.93 , respectively, for CN losses and gains) with CN alteration frequencies observed in TCGA COAD/READ samples (Supplementary Fig. 5 and Supplementary Fig. 6).

As described above, we also assessed CN profile stability along PDX lineages which extend beyond the first passage. We observed

Fig. 1 | Multi-omic Overview of the Colorectal Cancer PDX Cohort and Cetuximab Response Modelling Approach. **a** The left panel presents the IRCC patient derive xenografts (PDX) collection, from 231 unique colorectal cancer (CRC) liver metastasis (LMX) resections. This collection was characterised at a multi-omic level and assessed for cetuximab response. A schematic of the omic-specific feature engineering is also provided. The right panel outlines the CeSta classifier pipeline. Input features selected from the training set (Methods) using univariate tests (Fisher's exact, Mann-Whitney *U*-test) and multivariate linear models feed into three independent level 1 classifier pipelines: forward feature selection plus elastic net, ANOVA feature selection plus extra trees, and ANOVA feature selection plus support vector classifiers. A fourth classifier, a catBoost model, is pre-trained on pan-cancer data from the Cell Model Passport repository and fine-tuned using IRCC-PDX data. The predictions from these level 1 classifiers are stacked and inputted into a meta-classifier, which produces the final binary classification (cetuximab-responder/non-responder) using argmax-based soft voting. **b** CeSta nested cross-validation approach: 50 train/test splits are generated via stratified

sampling of the IRCC-PDX collection. CeSta is trained and tuned independently across these 50 splits. In each iteration, the training set is divided into three folds. Two folds are used in three rounds as the 'training fold', while the remaining fold serves as the 'validation fold'. Predictions from level-1 classifiers for the validation fold are stacked and input into the meta-classifier. After validation, first-level classifiers are fitted to the entire training set, and CeSta's performance is evaluated on the test set (pink rectangle, $N = 81$). CeSta is then trained on the entire IRCC-PDX dataset and tested on an independent CR-PDX dataset (grey rectangle, $N = 50$) for external validation. **c** Top frequently mutated genes in the IRCC-PDX cohort.

d Selection of multi-omic and clinical features across the IRCC-PDX collection, including CRIS expression cluster labels, methylation NMF cluster labels, primary sample anatomical location, and treatment backbone. Source data are provided as a Source Data file. Fig. 1AB has been Created in BioRender [lorio, F. (2024) BioRender.com/q01w468] and released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license (<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>).

solid intra-lineage CN consistency overall (median log₂R Pearson coefficient 0.927, Supplementary Fig. 7) and at the gene level (94% of driver genes are CN-stable within lineages, Supplementary Fig. 8), in line with previous reports³⁸.

We characterised our PDX collection's transcriptional landscape using two approaches to classify samples into subtypes: CMS^{39,40} and CRIS⁴. Results from these analyses were broadly consistent with TCGA COAD/READ and other CRC datasets where expression data is available (Fig. 1d and Supplementary Fig. 9).

To concisely represent our PDXs' epigenomics profiles, we grouped samples into five clusters obtained through non-negative matrix factorisation²⁸ (Methods). We observed that the samples belonging to one of these groups (cluster 1) were remarkably more hypermethylated over all measured CpG islands (median beta methylation level = 0.81, Kruskal-Wallis test, chi-squared = 289.47, $df = 4$, p value < 2.2×10^{-16} , effect size = 0.59, Supplementary Fig. 10). Consistent with our cluster definition, we also found cluster 1 to be highly enriched for the CpG island methylator phenotype (CIMP⁴¹) in 130 out of 146 PDXs (Supplementary Fig. 10). This heterogeneity of PDX methylation profiles resembled that observed in CRC patients, even though the percentage of IRCC-PDX samples classified as CIMP was slightly lower than that reported in TCGA COAD/READ (44% vs 58%, Supplementary Fig. 11). This is expected considering the low prevalence of MSI tumours—which are typically enriched for CIMP cases—within metastatic CRC cohorts such as ours³⁴.

Overall, our multi-omic overview of the PDX collection indicates that IRCC-PDX closely recapitulates the genomics, transcriptomics, and methylomics landscape of gold-standard human CRC cohorts, such as TCGA COAD/READ and MSK-IMPACT.

Exploratory single-omic analysis of the IRCC-PDX collection

To further investigate the molecular profiles of the PDXs in an unsupervised manner we conducted UMAP dimensionality reduction⁴² across individual omics and a density-based cluster analysis (via HDBSCAN)⁴³. This was followed by an enrichment analysis of covariates in the resulting clusters, as well as of differential drug response across them (Methods) to identify, respectively, specific molecular features and drug responses that discriminate between different subgroups (Supplementary Data 2).

By applying this approach to IRCC-PDX binary gene mutation features, we identify two clusters (Supplementary Fig. 12). The first cluster ($id = 0$) was almost entirely made up of *TP53* mutated (95%), and *KRAS* and *PIK3CA* wild-type PDXs (96% and 89%). The majority of models in this cluster belonged to CRIS type C (51%, chi-square statistic (χ^2) = 8.53, chi-square adjusted p value (cap) = 0.15) and were sensitive to cetuximab treatment (SD, OR, 65%, $\chi^2 = 10.34$, $cap = 0.06$). Almost all PDXs in the second cluster ($id = 1$) were also *TP53* mutant

(67%) but also *KRAS* mutant, differently from cluster 0 (98%, $\chi^2 = 122.43$, $cap = 2.85 \times 10^{-26}$). This cluster was also enriched for CRIS type A (43%, $\chi^2 = 26.06$, $cap = 1.49 \times 10^{-5}$) and non-responder (PD) models (75%, $\chi^2 = 11.66$, $cap = 0.03$) (Supplementary Data 2).

When applied to binary CNV features, the same approach identified four clusters (Supplementary Fig. 13 and Supplementary Data 2). One cluster ($id = 3$) exhibited a high CNV burden and CN gain count, with 19 out of 21 associated CNV (at a $cap < 0.1$) being a CN gain and present in 82% of the samples, among which top significant genes are *NFX1*, *ESRRA* and *MARK2*. Moreover, this cluster was predominantly composed of strong cetuximab responders (OR: 44%, $\chi^2 = 10.17$, $cap = 0.1$). Another cluster ($id = 0$) displayed instead a low count of CN events, and it was composed mainly of CRIS type A PDXs (67%, $\chi^2 = 29.93$, $cap = 7.49 \times 10^{-5}$). Finally, cluster 2 was characterised by losses in *FGD5*, *RAFI*, *XPC* and *SATB1* genes (all present in 79% of the samples, at a $cap < 0.1$), while cluster 1 showed mild enrichment of losses in *BCL2*, *SMAD4* and *MALT1* (30%, at a chi-square p Value < 0.01).

In addition, we identified 3 UMAP/HDBSCAN clusters from continuous gene-level RNAseq features (Supplementary Fig. 14). Among these, one cluster ($id = 1$) was almost exclusively composed of cetuximab non-responders (PD, 88%, $\chi^2 = 8.99$, chi-square p value = 0.002, $cap = 0.16$), as well as it was enriched for hypermethylated (NMF cluster 1, 76%, $\chi^2 = 50.91$, $cap = 5.81 \times 10^{-11}$) and CRIS type A PDXs (96%, $\chi^2 = 89.22$, $cap = 2.11 \times 10^{-19}$). Finally, we projected the methylation NMF cluster labels (Methods) onto 2D UMAP embeddings computed from probe-level methylation features and analysed their distribution across 5 obtained clusters (Supplementary Fig. 15). The first one ($id = 0$, exactly matching NMF cluster 1, $\chi^2 = 189$, $cap = 3.94 \times 10^{-41}$) was the most hypermethylated overall and largely made up of cetuximab non-responders (83%, $\chi^2 = 12.49$, $cap = 0.03$). Another one ($id = 3$, mostly matching NMF cluster 4, 96%, $\chi^2 = 154.43$, $cap = 1.4 \times 10^{-33}$) was enriched for CRIS type C (72%, $\chi^2 = 22.87$, $cap = 0.0001$) and OR PDXs (32%, $\chi^2 = 22.41$, $cap = 0.0001$).

Confirmation of established biomarkers of cetuximab sensitivity

Around half of the tumours in mCRC patients are wild-type for the *KRAS*-*NRAS*-*BRAF* genes (triple negative). These patients routinely receive anti-EGFR treatment with cetuximab or panitumumab in combination with chemotherapy as an alternative to surgery resulting in a median survival extension of 2 to 4 months, compared with chemotherapy alone¹. Retrospective analysis of triple-negative CRC patients from the CRYSTAL and FIRE3 trials has also highlighted that patients with left-sided tumours treated with anti-EGFR antibodies had better survival and treatment response than patients with right-sided tumours⁴⁴.

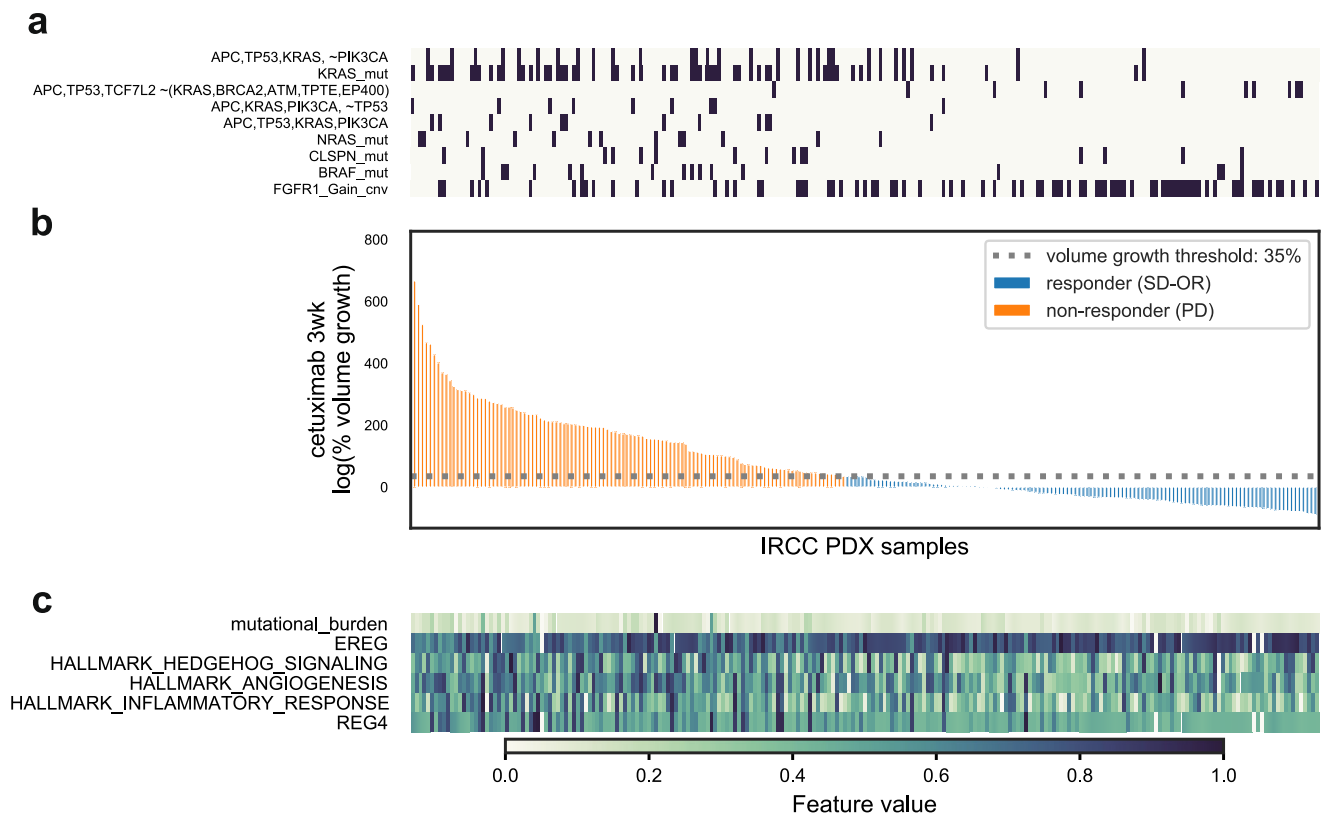


Fig. 2 | Overview of cetuximab response and biomarker candidates. **a** Mutation patterns of CRC driver genes and mutational signature features among those with the most significant impact on CeSta predictions (Fig. 4a) **b** cetuximab non-responders ('PD', volume growth > 35%, in orange) and responders ('SD-OR',

volume growth \leq 35%, in blue). **c** Selection of continuous features which best differentiate between PD and SD-OR PDX models. Source data are provided as a Source Data file.

Treatment intervention in our PDXs (Methods) closely matched that of cetuximab human trials such as PEAK^{7,45} and FIRE3⁵ as well as current clinical best practices^{46,47}.

Across our IRCC-PDX collection ($N = 231$), *KRAS* mutations were much more frequently observed in PDXs with a cetuximab non-responder phenotype (Fisher's exact test's odds ratio (FETo) = 0.12, p value (FETp) = 1.2×10^{-10} , 95% confidence interval (FETci) = [0.06,0.26], standardised residuals (FETSres) = 6.37) Fig. 2a and (Supplementary Data 1). *NRAS* (FETo = 0.06, FETp = 8.49×10^{-4} , FETci = [0.002,0.47], FETSres = 3.35) and *BRAF* (FETo = 0.27, FETp = 0.035, FETci = [0.06,0.95], FETSres = 2.31) mutations were noticeably more likely to occur in non-responder PDXs, though only 13 and 16 mutant PDXs were observed across IRCC-PDXs, respectively. However, overall mutational and CN alteration burden, defined as the total number of events per PDX and intended as coarse-grained proxies for tumour progression and genomic stability, did not appear to correlate with cetuximab sensitivity (Fig. 2b, c and Supplementary Data 1).

Finally, a right-sided localisation of the original tumour showed a moderate association with a non-responder phenotype (FETo = 0.42, FETp = 0.01, FETci = [0.20,0.84], FETSres = 2.66).

As previously mentioned, the *KRAS*-*NRAS*-*BRAF* triple negative signature is widely recognised as the best-established biomarker of cetuximab sensitivity. This association is being used both as a clinical discriminant for treatment and as an entry criterion for anti-EGFR trials, and it is clearly visible in our IRCC-PDX collection (FETo = 11.38, FETp = 4.91×10^{-16} , FETci = [5.80,23.40], FETSres = 7.99). These observations thus indicate that our IRCC-PDX collection recapitulates the best available marker of cetuximab sensitivity in patients.

To further explore how molecular characteristics of the IRCC-PDXs distribute with respect to their response to cetuximab, we

performed a differential expression analysis, comparing cetuximab responder versus non-responder PDXs. This yielded 230 upregulated and 1534 downregulated genes (at a negative binomial generalised linear model adjusted p value < 0.05 and $|\log$ fold change (logFC) | > 0.58, Supplementary Data 3). A functional enrichment analysis via preranked GSEA using the Hallmark gene-signature collections from the Molecular Signature Database (MsigDB)⁴⁸, unveiled, as expected, many significantly down-regulated gene sets (Supplementary Data 3). Among these, HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION and HALLMARK_INFLAMMATORY_RESPONSE were the most significantly enriched (NES = -1.92 and -1.82, respectively, with an adjusted p value = 0.001 for both, Supplementary Data 4 and Supplementary Fig. 16).

A single sample GSEA extended to Reactome pathways⁴⁹ confirmed Inflammatory processes as among those upregulated in non-responder samples (positive score in 63 out of 121 non-responder PDXs for REACTOME_INTERFERON_GAMMA_SIGNALING, Supplementary Data 5 and Supplementary Fig. 17). The other consistent signal for upregulated genes in resistant PDXs (positive score in 77 out of 121 non-responder PDXs for REACTOME_KERATINIZATION) was related to keratinisation, which has been previously found associated with a set of more aggressive tumours in this same cohort⁵⁰. The only Reactome pathway significantly upregulated in cetuximab responder PDXs was, unsurprisingly, REACTOME_SIGNALING_BY_EGFR_IN_CANCER (positive score in 42 out of 121 non-responder PDXs).

Consistently, a transcription factor binding site enrichment analysis yielded significant results only for promoters of the genes downregulated in cetuximab-sensitive PDXs, and involved transcription factors with known roles in inflammation, such as *STAT1/2* and *KLF15* (Supplementary Fig. 18).

A stacked classifier modelling cetuximab sensitivity

Results from the single-omic and differential analyses highlighted the need for a more elaborate integrative modelling approach. To predict whether a CRC PDX responds to cetuximab treatment in terms of tumour volume shrinkage², we considered its multi-omic characterisation and reduced the task to a binary classification problem. We selected and integrated multi-omic features into a stacked classifier pipeline⁵¹: the cetuximab Stacked classifier (CeSta, Fig. 1a). Stacking is a supervised ensemble learning technique which combines multiple weak classification models (level 1 classifiers, lv1) using a meta-classifier. This architecture improves upon individual classifiers' performance. It is well suited for a classification task such as ours, which is based on tabular data with relatively few examples (231) and a much larger number of features (35,053, Supplementary Data 1): a scenario where more complex models and deep neural networks fare poorly^{52,53}. A similar architecture has been successfully used to predict drug response in breast cancer patients from the multi-omic characterisation of their tumours⁵⁴.

Our CeSta pipeline implements a late integration approach to prevent high-dimensional 'omics (transcriptomics, methylomics) from overwhelming those with fewer features (typically genomics) by dominating the feature selection phase (Fig. 1a). We used a nested cross-validation approach for model tuning, training, and validation, based on generating 50 train/test split replicates of our IRCC-PDX dataset (with 150 and 81 PDXs, respectively, for the training set and test set) assembled via stratified sampling (Fig. 1b). On each of these 50 training sets, our classifier pipeline performed a custom single omic feature selection step which reduced the initial input of 113 engineered and clinical features plus 34,940 raw transcriptomics and genomics features (Fig. 1a and Supplementary Data 1) to a smaller subset, with the size of the latter being amongst the hyperparameters tuned independently, across data splits (Fig. 1a and Supplementary Fig. 19, Methods). We used these pre-selected IRCC-PDX features as the input to 4 different lv1 classifier pipelines: (1) model-based forward feature selection, followed by elastic net logistic regression, (2) ANOVA-based feature selection, followed by either support vector classifier (SVC) or (3) extraTrees classifiers, and (4) a catBoost classifier pre-trained on a set of 55 multi-omic features from a collection of 860 pan-cancer cell lines from the Cell Model Passports (panCMP⁵⁵), then refined on the same set of 55 features from the IRCC-PDX (continual learning, Methods). The lv1 predicted probabilities were then stacked and combined using a soft voting classifier which outputs a binary classification of cetuximab sensitivity (Fig. 1a, b, Methods).

Candidate biomarkers of cetuximab sensitivity

Our CeSta pipeline selects the most informative biomarkers of cetuximab sensitivity across training examples sampled from the IRCC-PDX collection by combining univariate statistical tests (Fisher's exact, Mann-Whitney U test), percent lift, and logit (statsmodels v0.13.2 logit⁵⁶) models (Fig. 1a, b and Supplementary Data 6, and Methods). Here and in Fig. 2a, we provide an overview of some of CeSta's top features (i.e., as ranked by their impact on CeSta's predictions) and their relationship with cetuximab sensitivity. The latter represents our binary target variable, with 'responder' PDXs defined as those that grew in volume by 35% or less at three weeks after treatment (a proxy of disease control, as mentioned above) (Fig. 2b, Methods).

Among the considered genomics features, beyond the KRAS-NRAS-BRAF triple negative signature, *CLSPN* (percent lift: -0.675), *PTEN* (percent lift: -0.594), and *PIK3CA* (percent lift: -0.654) mutations were also more frequently observed in non-responder PDXs. Additionally, a few other driver gene mutations (e.g., *EGFR* (percent lift: -0.721) and *MET* (percent lift: -0.702)) were noticeably more likely to occur in non-responder PDXs, although they were rare overall (21 and 8 observations in IRCC-PDX, respectively). Only mutations in *KRAS* (logit

p-value (logit *p*) = 0.002), *BRAF* (logit *p* = 0.037), *PTEN* (logit *p* = 0.049), and *NRAS* (logit *p* = 0.03) were found to be associated with cetuximab resistance via single-omic multivariate logit regression. Our CeSta approach combines these metrics (univariate and multivariate *p* values, percent lift) into an aggregated feature selection score (Methods) which allows us to detect both well-supported and rare candidate markers. CELLector subgroups 7 (*APC*, *TP53*, *KRAS*, *PIK3CA* mutated), 16 (*TP53* wild-type; *APC*, *KRAS*, *PIK3CA* mutated), and 5 (*APC*, *TP53*, *KRAS* mutated; *PIK3CA* wild-type) were significantly associated with a non-responder phenotype (FETo = 13.46, 9.51, 4.12, FETp = 0.002, 0.014, 0.001, FETci = [1.89,587.67], [1.24,427.84], [1.61,11.98], FETSres = 3.17, 2.55, respectively).

In contrast, subgroup 12 (*APC*, *TCF7L2*, and *TP53* mutated; *KRAS*, *BRCA2*, *ATM*, *TPTE*, *EP400* wild-type) was more likely to contain responder PDXs (FETo = 9.57, FETp = 0.011, FETci = [1.32, 421.45], FETSres = 2.59). However, this is a rare occurrence, with only 8 PDXs presenting this signature across IRCC-PDX. Subgroups 7,16 and 5 were also significantly associated with cetuximab resistance after multivariate logit regression (logit *p* = 2×10^{-6} , 3×10^{-6} and 3×10^{-4} , respectively).

Finally, *FGFR1* CN gains (FETo = 3.19, FETp = 5.98×10^{-5} , FETci = [1.74, 5.98], FETSres = 4.04) were more frequently observed in responder PDXs. Although *ERBB2* and *MET* amplification events (i.e., >2 copies gained) were rare (5 and 3 examples in IRCC-PDX, respectively), they were more frequent in non-responders (percent lift: -1 for both). These genomic signatures agree with previous surveys of CRC poor-prognosis driver alterations^{36,57}, suggesting at least a partial overlap between markers of CRC progression and those of cetuximab resistance in PDX.

As transcriptomics features (Fig. 2c), while *EGFR* (Mann-Whitney U-test *p* value (MWU *p*) = 4) and *EGF* (MWU *p* = 17) were not differentially expressed in cetuximab responders versus non-responders PDXs, *REG4* (MWU *p* = 0.001) and *EREG* (MWU *p* = 7×10^{-5}) were instead significantly upregulated in resistant and sensitive cases, respectively. *REG4* (Regenerating Islet-Derived Protein 4) is a C-type lectin-like mitogenic protein known to stimulate EGFR signalling and promote migration and invasion in CRC⁵⁸. High *REG4* expression is associated with poor prognosis and low recurrence-free survival in CRC patients⁵⁹ and, more specifically, with cetuximab resistance¹² in CRC organoids and PDX models. A suggested mechanistic explanation points to FZD and LRP5/6, both upstream components of the Wnt/ β -catenin pathway, which are involved in the *REG4*-mediated promotion of stemness induced by *KRAS* mutation in CRC with *APC* loss⁶⁰. *EREG* (epiregulin) is a member of the EGF family and an EGFR ligand; it is thus involved in inflammation, cell proliferation, and cancer progression. *EREG* activity has been associated with cetuximab sensitivity in preclinical models and patients^{61,62}, and it has been suggested that, in an inflammatory environment, *EREG* can promote stemness and cancer cell proliferation by stimulating ERK signalling through *EGFR* activation in a variety of cancer types⁶³⁻⁶⁵.

We also observed high PROGENy²⁹ EGFR pathway expression scores associated with a non-responder phenotype (MWU *p* = 0.002, percent lift: -1.879), whereas, as mentioned above, EGFR expression as an individual feature was not. We observed a similar pattern for *KRAS*: It was not differentially expressed across responders versus non-responders PDXs (MWU *p* = 0.23) but high MSigDB^{48,66} HALLMARK_KRAS_SIGNALING_UP gene set ssGSEA scores were associated with non-responder PDXs (MWU *p* = 0.001, percent lift: -10.688). These observations suggest that engineering aggregated expression features using ssGSEA and PROGENy scores might be more informative than individual gene expression features for cetuximab sensitivity prediction. However, it is also important to note that feature aggregation might introduce additional complexity. PROGENy signals for EGFR could be partly driven by downstream ERK-mediated signals, which are hard to disentangle from *KRAS*-triggered inputs. This may explain why both

EGFR and KRAS signatures are associated with resistance to EGFR blockage.

Finally, we observed that higher MSigDB gene set ssGSEA scores for angiogenesis (percent lift: -2.168), inflammatory response (percent lift: -3.7), UV and DNA damage response (percent lift: -6.63), and Hedgehog (Hh) signalling (percent lift: -5.44), were all associated with non-responder PDXs (MWU $p < 0.01$ for all). The Hh hallmark score is fascinating as it might corroborate the evidence that Hh pathway activity correlates with reduced response to cetuximab⁶⁷.

When considering methylation features (Fig. 2c), NMF cluster 1, the most hypermethylated, was enriched for non-responders and MSI-like PDXs (FETp = 2×10^{-4} , percent lift: -0.796). Cluster 4, the second-most hypo methylated, was enriched for responder PDXs (FETp = 3×10^{-4} , percent lift: 2.299).

Across all omics, both categorical (Fig. 2a) and continuous features (Fig. 2b) were either too sparse or too noisy to be adequate predictors of cetuximab response when considered individually. This highlights the effectiveness of an integrative model which combines the most informative features across 'omic boundaries.

Validation of the CeSta classifier

We set out to internally assess CeSta's performance on our IRCC-PDX collection using a holdout shuffle approach, followed by testing the null hypothesis that results generated by different classifiers are equivalent⁶⁸.

We started by generating 50 train/test set split (150 and 81 PDXs, respectively) replicates from our IRCC-PDX dataset. We used a nested cross-validation approach to tune and train 50 independent CeSta replicates (Fig. 1b, 'internal validation'). To provide a realistic and stringent benchmark, we evaluated many baseline cetuximab sensitivity classifiers of varying complexity (Fig. 3a, b and Supplementary Fig. 20). Here, we present results from a performance comparison of our CeSta classifier against three of the best-performing baseline classifiers. These build on the SOTA clinical predictor of cetuximab sensitivity: the KRAS-NRAS-BRAF triple negative marker^{46,47} and whether the original tumour is located in the left portion of the patient's colon⁴⁴. These features were combined into a cetuximab sensitivity classifier using either (1) a rule-based approach entirely analogous to the clinical criterion for cetuximab treatment (i.e., PDXs with the triple negative marker were predicted as responders to cetuximab, Fig. 3a, 'tripleNegRule' and 'tripleNegRightRule') or (2) an elastic net penalised logistic regression model (Fig. 3a, 'elNet baseline') taking in input the four features above as possible regressors (Methods). As for CeSta, we tuned and trained 50 independent replicates of this latter baseline classifier over the 50 split replicates we previously generated.

CeSta outperformed all baseline models (mean F1: 0.941, Mann-Whitney post-hoc test p value: < 0.001) on this internal validation setup (Fig. 3). Interestingly, the elNet baseline performance, measured via F1 score (i.e., the harmonic mean of precision and recall), fully matched the triple negative rule-based classifier, indicating that the elNet model can recapitulate the clinical decision criterion. Figure 3b shows that CeSta outperforms this same elNet baseline classifier for the vast majority of replicate splits (mean AUROC = 0.821 versus 0.780, Mann-Whitney post-hoc test p value: < 0.001), with an average of 0.04 increase in ROC AUC, computed using the ROC AUC variance formula first proposed by Delong and colleagues⁶⁹⁻⁷¹.

Based on our finding, we evaluated whether our CeSta classifier would outperform the clinical SOTA baseline classifier on an independent cohort of CRC PDX models (Fig. 1b, 'external validation'). This external validation cohort (from now on CR-PDX), consisting of 50 CRC xenografts, was collected and characterised at the genomic, transcriptomic and clinical levels at Charles River Discovery Research Services and included samples from European patients (Supplementary Data 7, Methods).

We tuned and trained a single instance of our CeSta pipeline as well as a single instance of the baseline model over the entire IRCC-PDX collection ($N = 231$). We then compared their predictive performance on the never-before-seen CR-PDX set ($N = 50$) using the same set of multi-omic engineered features we described previously for IRCC-PDX (Supplementary Data 7, Methods). Similar to what we observed in the internal validation phase, our CeSta classifier outperformed the clinical baseline classifier (AUROC = 0.88 and 0.78, respectively), with an improvement of 0.1 ROC AUC (Fig. 3c and Supplementary Data 8). More specifically, our CeSta pipeline correctly predicted three additional KRAS-NRAS-BRAF triple-negative PDXs as cetuximab non-responders and one additional non-triple-negative as a responder; on top of matching biomarkers correctly predicted by the baseline classifier (Fig. 3d, e and Supplementary Data 8). The three triple-negative non-responders (relative tumour volumes at 3 weeks: 125%, 485% and 1380%) have a lower than average (VST 7.4 vs 9.9) *EREG* one and higher than average (VST 13.9 vs 9.4) *REG4* expression, highlighting how the transcriptional features identified by CeSta help its correct predictions, despite the genetic features, for those samples. Interestingly they are two CRIS-A and one CRIS-B, subtypes generally refractory to anti-EGFR therapy. The non triple-negative sample has a canonical BRAF mutation (V600E) that usually determines resistance to cetuximab, but CeSta correctly identified it as a responder (-18.4% relative tumour volume at 3 weeks), in this case the relevant feature that steered its prediction in the right direction most probably is the *FGFR1* amplification (log2 0.18).

Further external validations of CeSta on patients' data are currently unfeasible, due to the lack of datasets from the characterisation of cohorts that are unselected for *KRAS* mutational status, treated with cetuximab monotherapy and with multi-omics data available. However, we tested whether any of the predictive transcriptional features identified by CeSta differentiate between cetuximab responder and non-responder patients. This analysis was conducted using gene expression data obtained from a limited single-omic CRC patient dataset, accompanied by cetuximab response data and encompassing 43 non-responder patients and 25 responders patients⁶². When comparing ssGSEA scores (Supplementary Data 4) computed across the two groups for the three Hallmark pathways identified as predictive by CeSta and with the highest percent-lift (as previously discussed), the one related to inflammatory processes showed a statistically significant difference (Wilcoxon test p value = 0.02, Supplementary Fig. 21). While these results confirm the association between higher inflammation marker expression and lack of sensitivity to cetuximab in patients, a larger, multimodal dataset would be ideal for performing a more rigorous validation of the CeSta predictive features in human patients in the future.

Explanation of the CeSta classifier

Post hoc explanations approximate the behaviour of a classifier by modelling relationships between feature values and the classifier's predictions. Here, we relied on SHapley Additive exPlanations (SHAP⁷²) to define local feature importance and their impact on the CeSta classifier's predictions. SHAP is a game theoretic approach through which values representing a feature's average marginal contributions over all possible feature coalitions are computed.

Our CeSta classifier leverages additional informative genomic (e.g., *FGFR1* amplification) and transcriptomics (e.g., *EREG* and *REG4* expression; angiogenesis, inflammation, and Hh signalling ssGSEA scores) features (Fig. 4a) to improve upon the clinical baseline classifier (Fig. 3b,c) while retaining the latter's top predictive features, namely the KRAS-NRAS-BRAF signature. As shown in the CeSta SHAP waterfall plot in Fig. 4b, we observed high Hh signalling, high angiogenesis ssGSEA scores, and the *KRAS*, *APC*, *TP53* mutation signatures being predictive of cetuximab resistance. In the same panel, high *EREG*

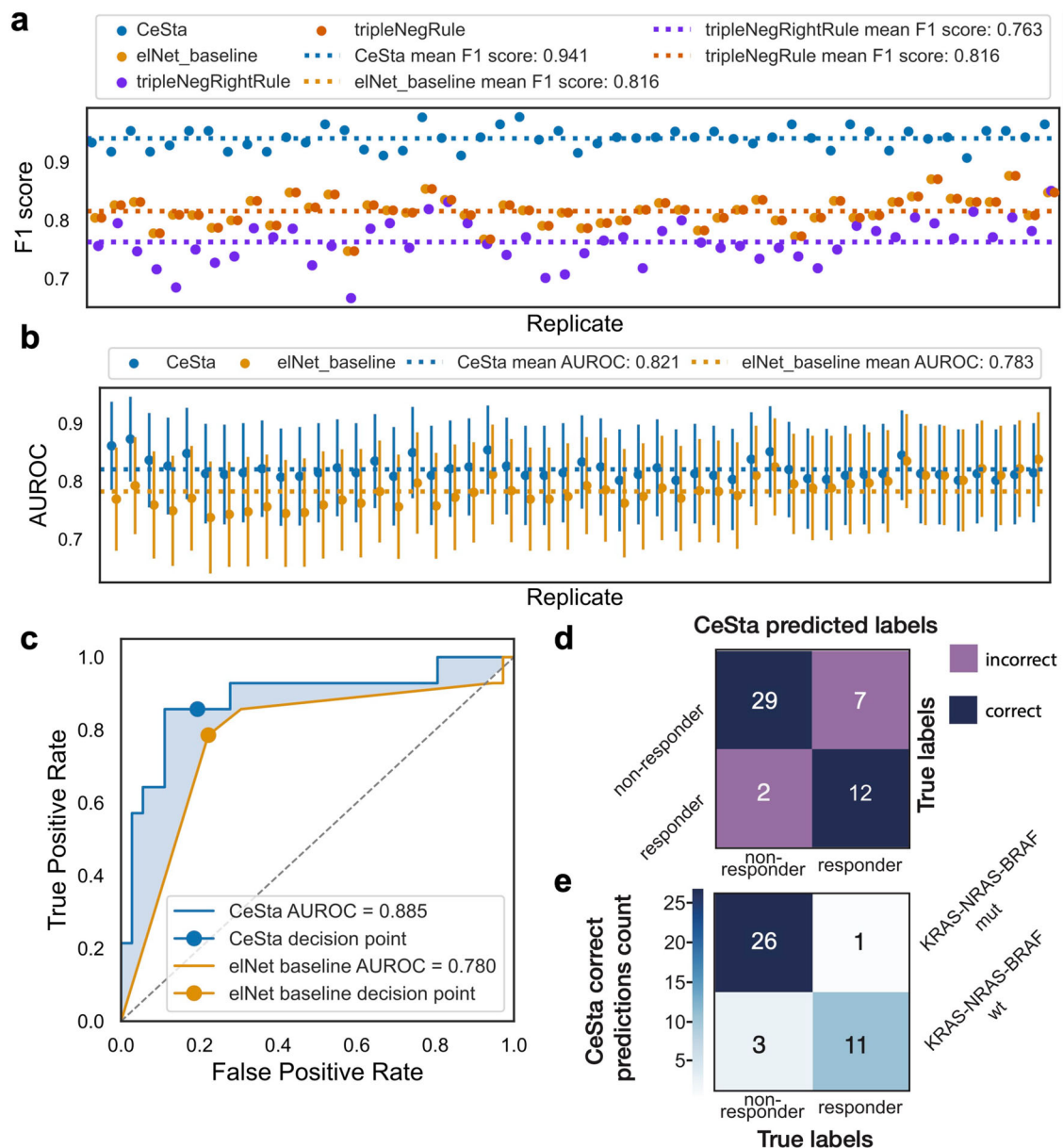


Fig. 3 | CeSta outperforms the state-of-the-art baseline classifier on IRCC-PDX and CR-PDX. **a** Classification performances quantified through F1 scores (harmonic mean of precision and recall) across 50 train/test IRCC-PDX split replicates (x-axis) for the stacked classifier ('CeSta', in blue), an elastic net penalised logistic model ('elNet baseline', in tan) which uses state-of-the-art clinical features for cetuximab sensitivity in CRC (KRAS, NRAS, BRAF mutational status, right colon tumour location), a rule-based classifier using the KRAS-BRAF-NRAS triple negative clinical signature (tripleNegRule, in orange) as a binary predictor, and another rule-based classifier which uses both the aforementioned triple-negative signature and the 'right colon' feature (tripleNegRightRule, in green). **b** Area under the receiver-operating-characteristic curve (AUROC) values and error bars, obtained via DeLong's method, indicating 95% confidence intervals^{69,70} across 50 IRCC-PDX of $n = 150$ and 81 train/test split replicates (x-axis), for CeSta (in blue) and the elastic net penalised logistic model ('elNet baseline', in tan) described in (a). **c** AUROC (DeLong's method) computed over the external validation CR-PDX

dataset for CeSta (in blue) and the elNet baseline classifier ('elNet baseline', in tan) after a single instance of both models is trained and tuned over the entire IRCC-PDX dataset. The shaded area between the CeSta and elNet baseline ROC curves represents the improvement in AUROC. Decision point coordinates correspond to the false-positive and true positive rates obtained from the corresponding classifier's predictions. Here, rule-based classifier decision points overlap with the elNet baseline's. **d** Confusion matrix from a comparison of CeSta classifier outcomes (same validation setup as c) and PDXs actual cetuximab response over the external validation CR-PDX dataset. Correct predictions are on the diagonal highlighted in blue, incorrect predictions off the diagonal are highlighted in purple. **e** CeSta correct prediction counts (same validation setup as c) over the CR-PDX external validation set grouped by PDX cetuximab sensitivity (x-axis) and PDX KRAS-NRAS-BRAF triple-negative status (y-axis). CeSta correctly predicts additional triple-negative non-responders (3) and triple-positive responders (1), which all baseline classifiers miss. Source data are provided as a Source Data file.

expression and, more noisily, low *REG4* expression and *FGFR1* amplification appeared to influence the model towards a 'responsive' prediction. Further, stacking our four lvl1 classifiers resulted in a slight performance increase over the best-performing lvl 1 classifier (i.e., the ANOVA SVC pipeline) taken on its own, albeit with substantial AUROC confidence interval overlap (Fig. 4e).

To further characterise the outstanding CeSta features, we considered the 275 genes from the engineered transcriptomic features with the highest SHAP values and characterised their coded proteins in terms of physical interaction. To this aim, we performed a protein-protein interaction network analysis, through String-db⁷³, observing a larger number of interactions than expected by chance (135 vs 23, p

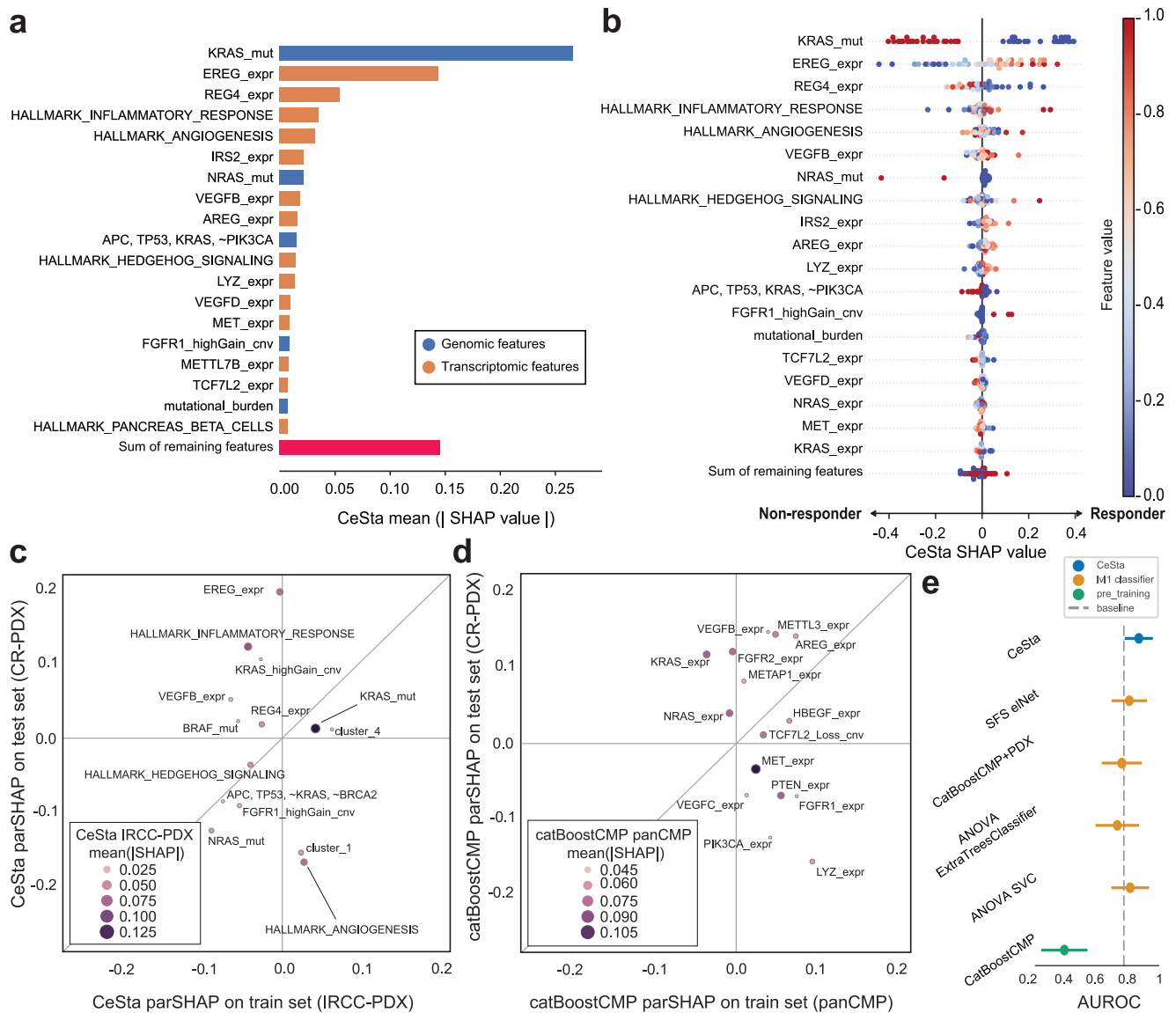


Fig. 4 | CeSta leverages informative features and combines weaker classifiers.

a Feature importance as determined by CeSta, represented by the mean absolute SHAP value (x-axis) for the top significant features (y-axis). **b** Top significant features' impact on CeSta output using SHAP values (x-axis) across all 50 PDXs in the CR-PDX validation set (scatter dots). The most important features in (a) have the greatest impact on model outcomes, with a clear separation between positive and negative effects. **c** Performance of CeSta's top features on IRCC PDXs and the external cohort. The relationship between a feature's SHAP values and cetuximab sensitivity on the train set (full IRCC PDX set, x-axis) and test set (CR PDX set), after removing other features' effects (partial correlation, parSHAP). Dot size and colour indicate a feature's mean absolute SHAP value on the training set. Dots closer to the diagonal indicate consistent performance across train and test sets. Key features like KRAS mutation and EREG expression align closely with the diagonal, indicating

a good fit or slight underfitting. **d** Underperformance of CMP-trained features on the external cohort. The relationship between CatBoostCMP feature SHAP values and cetuximab sensitivity on the train (panCMP set) and test (CR-PDX) sets, after removing other features' effects. Dot size and colour represent a feature's impact on model prediction. Many top features of this model fall in the lower right quadrant, indicating overfitting. **e** AUROC confidence intervals (CI, 95%) for CeSta (blue), three level 1 classifiers (orange), the catBoost model trained on the panCMP dataset (green), and the same catBoost model retrained on the IRCC-PDX dataset. CeSta shows a slight performance improvement over the best level 1 classifier, with overlapping CIs. The cell-line-trained CatBoost classifier poorly predicts cetuximab sensitivity in PDXs, but retraining improves its performance. Source data are provided as a Source Data file.

value $< 10^{-16}$) and confirmed the relevance of subnetworks involved in the transcriptional control (RELA, NFKB1-2, IRF1) and execution (CXCL9-11-10 chemokines) of the inflammatory response (Supplementary Fig. S22). This analysis also underscored interactions involving the MAPK pathway (i.e., between RAF1 and NRAS), transcriptional regulators of WNT signalling (TCF7L2 and TLE1/3) and adhesion-dependent growth control (LICAM, and FGFR1).

We also detected very low collinearity among the top CeSta features' values, with the largest anticorrelation between the Hh signalling ssGSEA score and EREG expression (Pearson's $r = -0.2$). In contrast,

high EREG expression was associated with both increased angiogenesis (Pearson's $r = 0.4$) and high inflammatory response (Pearson's $r = 0.3$) ssGSEA scores (Supplementary Fig. S23).

Comparison of cetuximab response in cell lines and PDX models

PDX models are thought to recapitulate inter and intra-tumour heterogeneity observed in patients more faithfully than immortalised cell lines. They provide at least some stromal microenvironment interactions and are more likely to follow pathways of drug sensitivity or resistance found in primary human tumours⁷⁴. However, 2d cell line

models are undeniably cheaper as well as simpler to screen and characterise, an advantage that has enabled the generation of large multi-omics cell line datasets^{55,75,76} and aided systematic drug and functional genetic screening efforts^{14,75}.

Several methods have been proposed to align gene-expression and other omics datasets from different model collections and patients. These methods include anchoring on common genes⁷⁷ or employing batch correction methods originally developed for single-cell data, thus harmonising all the features into a shared space across datasets/model-collections (such as *Celligner*⁷⁸). To avoid information leakage we opted instead for a supervised continual learning approach.

Particularly, we investigated whether a cetuximab sensitivity classifier trained (1) on a large pan-cancer multi-omic dataset (panCMP, $N = 860$) of 2d cell line models derived from the CMP dataset⁵⁵, or (2) on a small CRC-specific subset of the same panCMP cell-line dataset (CRC-CMP, $N = 44$) would compare favourably against (1) the classifier itself, retrained on the IRCC-PDX dataset ($N = 231$) or (2) the classifier itself, retrained on a randomly selected subsample of IRCC-PDX, with the same size as the colorectal 2d cell-line dataset (subIRCC-PDX, $N = 44$).

We observed that a panCMP-trained boosting classifier *catBoost*⁷⁹ performed very poorly in predicting PDX sensitivity to cetuximab (Fig. 4). This poor performance persisted even when considering the inclusion of the cell line tissue of origin as a categorical covariate (Fig. S24). When this *catBoost* model was further trained on the IRCC-PDX dataset (continual learning, Methods), its performance on the CR-PDX validation set became comparable to that of the other IRCC-PDX trained *lv1* classifiers. We observed a similar result when we traded several examples for tissue specificity in the cell-line dataset and compared a CRC-CMP-trained classifier against itself after retraining on subIRCC-PDX (Supplementary Fig. 25).

We evaluated the partial correlation between a feature's SHAP values and the target variable (parSHAP) to investigate further these differences in model performance across different training datasets. In this case, a positive parSHAP suggests that the classifier has identified and successfully exploited an informative feature for its current classification task. Given that our CeSta classifier performed just as well on the internal and external validations, it was not surprising to see matching parSHAP across CeSta SHAP values and cetuximab response in IRCC-PDX and CR-PDX (Fig. 4c and Supplementary Data 9) for most features, and particularly for those with the most significant impact on model prediction (Fig. 4a, b and Supplementary Data 9). On the other hand, several of the panCMP *catBoost* classifier's top features (*VEGFBC*, *PTEN*, *MET*, *PIK3CA* and *LYZ* expression, *TCF7L2* loss) did not perform as well on CR-PDX, compared to the cell lines training dataset (Fig. 4d, e), that is: their SHAP values' partial correlation with the target variable was lower across CR-PDX. This suggests that cell-line-trained models of cetuximab response struggle to predict PDX cetuximab sensitivity, primarily due to differences in the relationship between expression features and the target variable. These transcriptional differences between cell lines and PDXs might be due to the intense selection pressure imposed during cell line establishment, which makes available 2d models only partially representative of the general patient population¹⁹.

Discussion

We have described and made available multi-omic characterisation and drug screening data for one of the largest CRC PDX collections to date. This dataset recapitulates typical CRC alteration patterns observed in patient trials and gold-standard primary cohorts across all examined 'omics, and offers a combination of complete cetuximab response labels as well as dense multi-omic features. The cohort provides a realistic, stable platform for cetuximab sensitivity biomarker discovery and drug response modelling.

Building on this PDX collection, we developed CeSta, a multi-omic ensemble classifier of cetuximab sensitivity based on a stacked ensemble architecture. CeSta identifies and leverages transcriptional markers and predicts cetuximab responses (in an internal holdout shuffle validation as well as when tested on an external independent dataset) more accurately than other state-of-the-art classifiers and outperforms the criteria currently used in the clinic to address CRC patients to cetuximab treatment, specifically the KRAS-NRAS-BRAF triple negative genotype. Among the predictive transcriptional markers identified by CeSta are *EREG* expression (which is higher in responsive models) and *REG4*, Hh signalling, angiogenesis, and inflammation gene set cumulative expression scores (all more represented in resistant models). Some of these response predictors confirm previous findings in independent datasets. *EREG* has been documented to positively correlate with response to cetuximab in mCRC patients⁶²; *EREG* is an EGFR ligand, so it is conceivable that high *EREG* abundance leads to sustained EGFR signalling and sensitisation to EGFR inhibition. *REG4* is a poor-prognosis biomarker in CRC, possibly due to its ability to promote cancer cell stemness⁸⁰, and its expression was found to correlate with resistance to cetuximab and other ERBB family inhibitors in CRC PDX models¹². Interestingly, some elements of the Hh pathway proved to be upregulated in a CRC cell line in which resistance to cetuximab was attained by prolonged drug treatment⁶⁷. Our results provide translational significance for this observation by extending its reach to patient tumours on a population scale. We report that the expression of genes associated with hallmarks of angiogenesis and inflammation correlate with poor response to cetuximab and, in the case of inflammation, this applies also to clinical samples from patients. These transcriptional biomarkers might therefore be viable candidates for inclusion into an improved companion diagnostic for cetuximab sensitivity using clinical-grade gene expression technologies, such as Nanostring.

The identified features show a weaker association with cetuximab response in 2d CRC models than PDXs, corroborating our observation of poorer predictive performance for models trained on cancer cell line datasets. While this evidence supports the accuracy of PDX models for biomarker discovery, it is fair to acknowledge that PDXs only partly recapitulate the complexity of human tumours. In particular, the human stroma is rapidly substituted by murine counterparts during PDX serial passaging, and human immune components are not substituted by host populations owing to the severe immunodeficiency of mouse recipients⁸¹. Accordingly, all molecular data that contributed to CeSta development and application were limited to the exploration of cancer cell-intrinsic traits. However, the fact that CeSta predictive ability was maintained when analysing bulk transcriptomic datasets from patients' samples⁶² indicates that CeSta, although unfit to capture stromal and immune characteristics, has clinical applicability.

Collectively, our results highlight the value of extensive, cancer type-specific, and well-characterised PDX collections for drug screening, drug sensitivity modelling and mechanism of action discovery, and motivate future efforts to increase resource dimensions and improve analytical approaches as a means to further enhance the informative power and translational potential of PDX-based research.

Methods

Ethics statement

Tumour samples were obtained from 570 patients with CRC who underwent surgical resection of liver metastases at the Candiolo Cancer Institute (Candiolo, Torino, Italy), Ospedale Mauriziano Umberto I (Torino), Città della Salute e della Scienza di Torino—Presidio Molinette (Torino), and Grande Ospedale Metropolitano Niguarda (Milano, Italy) during the period 2008–2015. Informed consent for research use, including for the collection of sex and age information, was obtained in written form from all patients at the enrolling institution before tissue banking. Donor patient sex was

determined by self-report. Study approval was obtained from the Review Board of the Fondazione del Piemonte per l'Oncologia FPO–IRCCS (PROFILING protocol No. 001-IRCC-00IIS-10, version 11.0, updated July 13, 2022). Tumour tissue (hepatic metastasis) not required for diagnosis was used to generate PDXs. Animal procedures were approved by the Candiolo Cancer Institute Institutional Animal Care and Use Committee (IACUC) and by the Italian Ministry of Health (authorization 816/2016-PR) and were compliant with all relevant ethical regulations.

Genomic data collection

Illumina PairEnd pre-capture libraries were synthesised from double-stranded DNA according to Illumina's protocol (Illumina Inc.). Genomic DNA quality was validated and for each sample, 200 ng were used for library preparation. DNA was sheared into 300 base-pair fragments (Iug DNA in 100 µl volume) using the E210 Covaris plate system (Covaris, Inc. Woburn, MA). The fragmentation settings used are Intensity of 4200 Cycles per Burst, for 120 s. Sequencing libraries were amplified using the 'bridge-amplification' process by Illumina HiSeq pair read cluster generation kits (TruSeq PE Cluster Kit v2.5, Illumina) and were hybridised to custom RNA baits for the Agilent SureSelect® protocol. Paired-end, 75 bp sequence reads were generated using Illumina HiSeq 2000®. The sample mean sequencing coverage was ~700X if the lost coverage because of duplicated and off-target reads is considered. Reads were aligned to the reference human genome (NCBI build37) using BWA-aln 0.5.9⁹², and sequencing data were archived in bam files stored at the European Genome-Phenome Archive (<https://www.ebi.ac.uk/ega/> at the EBI) with accession number EGAD00001003334 (cram files are in EGAD00001003334, the study accession number is EGAS00001001171).

555 samples were sequenced using a custom-designed targeted colon cancer panel (SureSelect, Agilent, UK) consisting of all coding exons of 116 genes, 22 genes recurrently amplified/deleted, 51 CN regions, 121 MSI regions and 2 gene fusions (RSPO2 and 3). Samples were fragmented to an average insert size of 150 bp and subjected to Illumina DNA sequencing library preparation using Bravo automated liquid handling platform.

Sequencing was performed on an Illumina HiSeq2000 machine using the 75-bp paired-end protocol targeting 1 Gb sequence per sample. Data quality was checked for 95% target coverage at 100x and mutation analysis was performed using an in-house algorithm. Sequencing reads were aligned to the NCBI 37 human genome build using the BWA algorithm⁸² with Smith-Waterman correction and PCR duplicates were removed. Base substitutions, small insertions or deletions, and breakpoints were identified by comparison against an unmatched control using established bioinformatic algorithms: CaVEMan (<https://github.com/cancerit/CaVEMan/>) for mutations, Pindel (<https://github.com/genome/pindel>) to detect insertions and deletions, and CNVKit (<https://github.com/etal/cnvkit>) for CN detection.

We used an unmatched blood sample sequenced to an equivalent depth as control. To account for the absence of matched control, a bespoke variant selection pipeline was developed. To enrich for high-confidence somatic variants, we performed further filtering by removing known somatic polymorphisms using human variation databases—Ensembl GRCh37, 1000 genomes release 2.2.2 and ESP6500—and whether the same polymorphism was observed recurrently in 93 normal DNA samples sequenced using the same protocol and depth.

Cancer genes (CGs) are genes for which we can observe evidence of positive selection. Several statistical approaches have been developed to categorise the likelihood of a given gene in a specific tumour type to undergo a mutation at a high enough frequency for this to be indicative of a positive selection process. The majority of these methods rely on a comparison of non-synonymous (dN) and

synonymous (dS) mutations in each gene and factor in additional covariates. We have elected to use as the foundation of our set of colorectal CGs two recent statistical approaches developed using large TCGA datasets^{83,84}.

6426 driver variants across 113 genes were identified using the statistically significant single-codon hotspots from Chang et al.³⁶ and the intOGen⁸⁵ framework. These variants were combined to generate a reference set of driver variants, annotated based on their origin (Intogen driver only, Chang driver only, or common to both), their hotspot status, and whether they were known drivers for CRC. The final set of driver variants was used for annotating the PDX variants.

To assign segment log2R to individual genes we used coordinates overlap (BEDtools v2.29.2⁸⁶, <https://github.com/arq5x/bedtools2>) between them and gene coordinates (TSS-TES) obtained from GENCODE (version 34, <https://www.encodegenes.org>) for a set of 568 intOGen driver genes.

TCGA COAD/READ copy number calling

Segmented CN variation (CNV) data from TCGA-COAD and TCGA-READ (-1200 samples) on 02/09/2020 was downloaded via the Genomic Data Commons Data Portal (GDC, <https://portal.gdc.cancer.gov/repository>) using the TCGAbiolinks R package (v2.20.0⁸⁷).

The GDC CNV pipeline uses Affymetrix SNP 6.0 array data (harmonised to GRCh38) to identify genomic regions that are repeated and infer the CN of these repeats. This pipeline uses the DNACopy R-package⁸⁸ to perform a circular binary segmentation (CBS) analysis. CBS translates noisy intensity measurements into chromosomal regions of equal CN. The final output files are segmented into genomic regions with the estimated CN for each region. The GDC further transforms these CN values into segment mean values, which are equal to $\log_2(\text{copy-number}/2)$. Diploid regions will have a segment mean of zero, amplified regions will have positive values, and deletions will have negative values⁸⁹. Masked CN segments are generated using the same method except that a filtering step is performed that removes the Y chromosome and probe sets that were previously indicated to be associated with frequent germline copy-number variation.

Robust CNV events across our patient cohort were identified by searching for matches in the combined TCGA COAD/READ data, using GISTIC2.0⁹⁰ (<ftp.broadinstitute.org/pub/GISTIC2.0>) and ADMIRE v1.2⁹¹ (<https://ccb.nki.nl/software/admire/>).

GISTIC2.0 was applied using the recommended 'GISTIC2 Command Line Parameters' listed in the GDC CN segmentation documentation at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/#copy-number-segmentation. Here the 'segmentation file' corresponds to the masked segmented CN variation downloaded from TCGA COAD/READ, the 'marker file' contains the aforementioned probe coordinates filtered for 'freqcnv == FALSE' as per the GDC reference files (<https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files>), and the 'reference gene file' is the GRCh38 reference provided alongside GISTIC2.0.

ADMIRE1.2 was applied using the same parameter configuration shown in the example use case provided at <https://ccb.nki.nl/software/admire/readme.txt> with the 'segmented CNA' file again corresponding to the combined COAD/READ data, and the 'marker file' containing the filtered probe coordinates.

The output of these two analyses identifies CNV events spanning multiple segments from different samples across the patient cohort. We then merged these results by computing the union of all (fully or partially) overlapping ADMIRE or GISTIC segments, and included all non-overlapping segments from either tool resulting in a set of 2382 events. From this combined output, we extracted event and segment coordinates and mapped both to 552 known cancer driver genes in the intOGen catalogue⁸⁵ (02/02/2020 release, <https://www.intogen.org/download?file=IntOGen-Cohorts-20191112.zip>) using BEDtools v2.29.2⁸⁶ (<https://github.com/arq5x/bedtools2>) (Supplementary Fig. 27).

CNV event frequencies are denoted as the number of equivalent CNV events in TCGA samples divided by the number of COAD/READ samples.

Comparing driver gene SNPs in TCGA COAD/READ and PDXs

Frequencies of somatic alteration for TCGA samples was obtained from cBioPortal, selecting the Colorectal Adenocarcinoma TCGA, PanCancer atlas (https://www.cbioportal.org/study/summary?id=coadread_tcg_pan_can_atlas_2018) dataset.

Comparing copy number variation events in TCGA COAD/READ and PDXs

We first binned PDX segment log₂R values into three categories ('Loss', 'Neutral', 'Gain'), using the same GISTIC log₂R thresholds we applied to the TCGA COAD/READ data (-0.2, 1) [using the same threshold as in TCGA data here might be too strict for PDX sequencing data where there's less non-tumour tissue contamination as murine cells/DNA are filtered out]. We then computed gene-specific CNV event frequencies by counting the number of PDX samples with CN altered segments mapping to each gene, divided by the number of PDX samples.

We then computed the Spearman correlation coefficient for the TCGA and PDX gene-wise CNV event (here only 'Loss', 'Gain') frequencies.

Assessing PDX copy number stability within lineages

We grouped 91 PDX samples, according to their genealogy, into 13 multi-passage lineages and retrieved gene-specific log₂R data for 569 genes from the analysis described in the previous sections. We then computed the Pearson correlation across all gene log₂R for each pair of PDX samples and labelled each Pearson coefficient according to whether the two samples belonged to the same lineage or to different ones.

Assessing PDX mutational stability within lineages

We analysed somatic mutations along multi-passage PDX lineages using the same set of 91 PDX samples grouped into 13 lineages as described above. To rule out false positive calls for putative WT samples in lineages with apparent inconsistencies (Fig. S4), we further checked the coverage and absolute number of reads supporting each individual SNVs and only found single mutated reads in three WT samples with coverages < 400X.

Gene expression data collection and differential expression analysis

RNA was extracted using miRNeasy Mini Kit (Qiagen), according to the manufacturer's protocol. The quantification and quality analysis of RNA was performed on a Bioanalyzer 2100 (Agilent), using RNA 6000 Nano Kit (Agilent). Total RNA was processed for RNA-seq analysis with the TruSeq RNA Library Prep Kit v2 (Illumina) following manufacturer's instructions. Sequencing was then performed on Illumina Nextseq 500 at Biodiversa SRL, obtaining single end 151 bp reads, aiming at 20 M reads.

Read counts were obtained using an automated pipeline (<https://github.com/molinerisLab/StromaDistiller>), that uses a hybrid genome composed of both human and mouse sequences to exploit the aligner ability to distinguish between human derived reads, representing the tumour component, and mouse ones, representing the murine host contaminating RNA material.

Reads were aligned using STAR⁹² (version 2.7.1a, parameters--outSAMunmapped Within--outFilterMultimapNmax 10--outFilterMultimapScoreRange 3--outFilterMismatchNmax 999--outFilterMismatchNoverLmax 0.04) versus this hybrid genome (GRCh38.p10 plus GRCh38.p5hg38 with GENCODE version 27 and mouse GRCh38 with GENCODE version 16, indexed with standard

parameters and including annotation information from the GENCODE 27 plus m16 comprehensive annotation).

Aligned reads were sorted using sambamba⁹³ (version 0.6.6) and only non-ribosomal reads were retained using split_bam.py⁹⁴ (version 2.6.4) and rRNA coordinates obtained from the GENCODE annotation and repeatmasker track downloaded from UCSC genome browser hg38 and mm9.

featureCounts (<https://rdrr.io/bioc/Rsubread/man/featureCounts.html>, version 1.6.3) was run with the appropriate strandness parameter (-s 2) to count the non-multi-mapping reads falling on exons and reporting gene level information (-t exon -g gene_name) using combined GENCODE basic gene annotation (27 plus m16).

Sequencing data was available for 480 samples, but different filtering criteria lead to 470 QC passing samples. These criteria include: (1) ≥ 15 M total reads, (2) $\geq 60\%$ reads assigned to genes by feature counts, (3) $\geq 30\%$ reads assigned to human genes over the total of assigned reads.

These filters let us retain only samples with at least 5 M human reads.

To remove samples with lymphomatous characteristics⁴, 2 criteria were applied: (i) Principal Component analysis of expression data (samples with PC2 ≥ 30 were discarded): (ii) computation of a sample-level score for a leucocyte expression signature⁹⁵, averaging FPKM values for all the signature genes (samples with an average leucocyte signature ≥ 48 were discarded). Positivity for either criterion flagged samples as lymphomatous and excluded them from analysis.

Gene-level variance stabilised expression (VST) and robust fpkm values for 33,670 genes were obtained using DESeq2⁹⁶ (version 1.26.0), tmm using edgeR⁹⁷ (version 3.28.1) using only read counts from human genes.

CRIS and CMS subtyping was obtained for each individual tumour averaging the VST values for replicates, when available, using the R package CMScaller³⁹ (v2.0.1, FDR = 0.05 and RNAseq = TRUE) and the R package CRISclassifier⁴ (v1.0.0, FDR < 0.2).

Differential expression analysis to compare responders and non responders was run with DESeq2⁹⁶ with the formula 'batch + response', where batch indicates which one of the five different sequencing batches the sample belongs to ssGSEA scores were calculated using GSVA⁹⁸ (version 1.34.0) on tmm values with kcdf = 'Gaussian' and method = 'gsva'.

The obtained DEGs were used to perform GSEA enrichment analysis with R libraries ClusterProfiler^{99,100} (v3.14.3), DOSE¹⁰¹ (v3.12.0), msigdb³⁰ (v7.4.1) and enrichplot (<https://yulab-smu.top/biomedical-knowledge-mining-book/>) (v1.6.1). Protein-protein interactions network analysis was performed with STRING (string-db.org), the 20th April 2023 (<https://version-11-5.string-db.org/cgi/network?networkId=bjPsEA2nP3WX>).

Promoter sequences of the differentially expressed genes were defined as regions 1500 bp upstream and 500 bp downstream the TSS, using the same annotations for transcripts that were used with featureCounts. Motifs enrichments were obtained with HOMER (version 4.8).

Microarray data from Khambata-Ford⁶² (GSE5851) was obtained using Gene Expression Omnibus (GEO) query¹⁰² (version 2.54.1), applying a log transformation and selecting the most variable probe when multiple ones were available for the same gene, then ssGSEA scores were obtained as previously described.

GitHub repositories: <https://github.com/molinerisLab/StromaDistiller>, https://github.com/vodkatad/RNASeq_biod_metadata and https://github.com/vodkatad/biodiversa_DE.

The fastq files for all sequenced samples are stored at the European Genome-Phenome Archive (<https://www.ebi.ac.uk/ega/> at the EBI) with accession number EGAS00001006492.

Methylation data collection

Methylation profiles for 568 CRC samples were obtained using Illumina MethylationEPIC bead chip, which measures methylation status at about 850,000 sites using hybridisation on two different probes after bisulfite treatment on DNA. These samples comprise tissue from the original patient, either primary tumours or metastases, or both in some cases, and the corresponding engrafted tumours in mice (PDXs).

Raw data have been processed using the minfi package (<https://bioconductor.org/packages/release/bioc/html/minfi.html>, version 1.32.0). Data preprocessing was performed following the best practices outlined by Bioconductor minfi vignette and documentation, and Hinoue et al.⁴¹ (<https://www.bioconductor.org/packages/devel/workflows/vignettes/methylationArrayAnalysis/inst/doc/methylationArrayAnalysis.html>).

Background noise was removed using the minfi function *preprocessNoob()*, which implements the noob background subtraction method with dye-bias normalisation. Samples and probes that did not pass the quality control were then excluded from further analyses.

For samples, minfi provides a simple quality control plot that represents the log median intensity in both the methylated (M) and unmethylated (U) channels. By adopting the default median intensity cutoff of 10.5, six samples with lower values were removed from the dataset.

We then filtered the probes, based on their detection *p* value (det-Pval), which is indicative of the quality of the signal. By filtering out all those probes of which det-Pval was higher than 0.01 in at least one sample, we removed 64,361 probes. We also removed all the probes mapping on X and Y chromosomes (19,627), to remove gender bias, and those probes that are known to bind to common SNPs (30,435). Moreover, using the list originally published by Chen et al.¹⁰³, we removed 43,177 probes that have been demonstrated to map to multiple places in the genome.

To work with a coherent set of probes for all the samples, in particular xenografts, we decided to apply one last probes filter, removing all those probes known to specifically map on murine genome as well, in order to remove possible methylation signal coming from the murine infiltrate, with the same rationale followed for microarray data⁹⁵. To do this, we combined two lists of murine-specific probes, obtained from Needham et al.¹⁰⁴ and Gujar et al.¹⁰⁵, which resulted in removal of other 22,537 probes.

We combined the hg19 annotation package (IlluminaHumanMethylationEPICanno.ilm10b2.hg19 version 0.6.0), with the *liftOver()* function from the rtracklayer package¹⁰⁶ (version 1.46.0) and the imported file hg19ToHg38.over.chain.gz (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/>) in order to convert the remaining 700,298 probes' coordinates from hg19 to hg38.

Moreover, as done for expression data (See Gene expression data collection), we removed samples with clear lymphomatous characteristics. Specifically for methylation, samples with PC2 \geq 500 were almost always flagged by H&E analysis when it was available, therefore we considered all of them to be lymphomatous.

To identify groups of samples sharing similar methylation profiles, Beta values were used to run non-negative matrix factorisation algorithms in R (<https://www.rdocumentation.org/packages/NMF/>, version 0.22.0). *k* = 5 was identified as the best parameter by the cophenetic correlation coefficient (bootstrapping arguments: *rank* = 2.6, *nrun* = 100, *seed* = 42, *options* = 'p70'). We therefore selected 5 as the number of classes used to characterise the methylation landscape of our samples. We finally converted the five groups five engineered features from methylation data via one-hot encoding (Supplementary Data 1).

The idat files for all samples are available at the GEO with accession number GSE208713.

Clinical data collection

Since the patients whose tumours are included in our biobank were not enrolled in a specific clinical trial and underwent surgery in

different hospitals, our clinical data collection is based on personal communications with the Surgery Departments. This is the main reason behind the sparseness of the data.

Measuring cetuximab response in PDX models

After surgical removal from patients, each metastatic CRC specimen was fragmented; some fragments were frozen for molecular analyses, and two fragments were implanted in two 5-week-old female NOD-SCID mice. After engraftment and tumour mass formation, the tumours were passaged and expanded for two generations until production of 2 cohorts, each consisting of six 5-week-old male or female NOD-SCID mice. When tumours reached an average volume of 400 mm³, mice were randomised for treatment: 6 mice were treated with placebo and 6 mice were treated with cetuximab (20 mg/kg/twice-weekly i.p.).

Tumour size was evaluated once-weekly by calliper measurements and the approximate volume of the mass was calculated using the formula $4/3\pi \cdot (d/2)^2 \cdot D/2$, where *d* is the minor tumour axis and *D* is the major tumour axis. The maximum tumour diameter allowed by the IACUC and the Italian Ministry of Health (20 mm) was not exceeded. Sex was not factored into the study design because large-scale studies on cetuximab have demonstrated no significant differences in response between male and female mice^{2,20}. For assessing PDX models response to therapy, we used averaged volume measurements at 3 weeks after treatment normalised to the tumourgraft volume at the time of cetuximab treatment initiation. 231 tumour grafts were classified as follows: (1) 'OR' models with a decrease of at least 50% in tumour volume (2) 'progressive disease' (PD) models with at least a 35% increase in tumour volume, and (3) 'SD' for the ones in between².

Finally, to obtain a balanced dataset, we elected to combine the 'SD' and 'OR' classes into a single 'SD-OR' (i.e., treatment responder) class, turning our cetuximab response modelling task into a binary classification problem.

All animal procedures were approved by the Ethical Commission of the Candiolo Cancer Institute and by the Italian Ministry of Health (authorization 806/2016-PR). All animal procedures for the CR PDX data set were executed in an AAALAC-accredited animal facility and approved by the Committee on the Ethics of Animal Experiments of the regional council (Permit Numbers: G-13/13 & G18/12).

No statistical methods were employed to predetermine sample size. Sample sizes were guided by our prior experience with various PDX models^{2,20} and aligned with PDX minimal information standards¹⁰⁷. Tumour-bearing mice were randomised prior to treatment using the Laboratory Assistant Suite¹⁰⁸ by alternately assigning them to different treatment groups. Investigators were not blinded to group allocation during the experiments or outcome assessment.

Genomic feature engineering

To reduce data sparsity, we reshaped our mutational annotations into a binary matrix -- with columns (110 in total) corresponding to genes and rows (231 in total) corresponding to PDX models, where a value of 1 indicates that one or more SNVs mapping to a given gene have been observed in a given PDX model. We also generated additional mutational features: a 'mutational burden' feature containing the sum of all mutated genes for each PDX, and a set of 'multiple mutations' features, indicating the number of unique SNPs hosted by a given gene in a PDX model. Finally, we filtered out any binary feature which was observed in fewer than 5 PDXs across our IRCC-PDX collection. To obtain a compact representation of relevant co-occurrent or mutually exclusive mutations, we developed an extended version of the CELLector methodology³¹ that partitioned the PDX mutation landscape recursively finding subgroups defined by the most recurrent combinations of genomic events (mutations or CN alterations). Briefly, the original version of CELLector (from now on referred to as hierarchical), recursively applies the Eclat algorithm¹⁰⁹ on a population described by

a binary event matrix (BEM), with each column representing a genomic feature and 0/1 possible entries indicating the absence/presence of that feature in a sample. In the hierarchical version of CELLector, the genomic background of a population is represented as a binary tree whose topology is defined by the most frequently observed combination of genomic features (referred as signature) together with the fraction of samples for which those mutations occur and hence satisfy the signature rule (sequence of presence/absence of specific features). In particular, CELLector first identifies the root as the genomic feature with largest support, i.e., number of patients in which that feature is observed, and then defines two sibling nodes. The left child corresponds to the subset of samples satisfying the parent feature and the feature with greatest support among the samples in the parent node. The right child corresponds to the complementary population of the parent node, composed of samples not satisfying that feature, and among those the feature with greatest support. This algorithm is applied recursively until no sub-population satisfying a certain signature rule of at least a *minGlobSupp* percentage of samples is identified, with *minGlobSupp* being a hyperparameter defined apriori. This hierarchical structure outputs K recursive signature rules that can be converted into a partition of $K + 1$ groups as follows.

Starting from CELLector hierarchical binary tree,

1. For each node starting from the root, we define with U the set of samples satisfying that node rule defined as the corresponding signature S .
2. If the considered node has a left child ($U_l \subset U$) associated to feature F_l , we defined with $U_{rm} := U_l$ the set of samples to be removed from U .
3. If U_l has additionally a right child U_r defined by feature F_r , U_{rm} is updated with $U_{rm} := U_{rm} \cup U_r$.
4. If U_r has another right child $U_{r,r}$ defined by signature $F_{r,r}$, the update is repeated as $U_{rm} := U_{rm} \cup U_{r,r}$ and this step is performed recursively until the considered node has no right child.
5. The new set of samples is defined as $U_n = U \setminus U_{rm}$ and corresponding signature rule representing the group is defined as $S, \sim F_l, \sim F_r, \sim F_{r,r}, \dots$

If the condition in step 2. is not satisfied, the group is directly defined as samples in node U and satisfying signature S rule. Once every node in the hierarchical binary tree was considered, the last group was defined as the remaining samples that did not satisfying any hierarchical signature rule. The signature defining this group is created as the negation of the root node and all the recursive right childers, as described before. Note that the newly created groups could be composed of a fraction of patients lower than the *minGlobSupp*.

We applied the partitioned version of CELLector (V2.0.0) to the somatic mutation PDX space in BEM format with *minGlobSupp* fixed at 0.02.

Similarly to what we describe for above for mutation features, we discretise each of our 1162 gene-level log₂ features into four categories ('Loss', 'Neutral', 'Gain', 'High Gain'), using, in addition to the GISTIC log₂R thresholds for 'Loss' and 'Gain' (-0.2, 1), an additional threshold at 2, above which a gene is considered to be involved in a 'High Gain' event in which more than 1 additional copy is gained.

This 'High Gain' category is added to help capture any association between driver gene high-order CN gain and cetuximab sensitivity.

We then reshape these categorical CN annotations into a binary matrix with columns corresponding to individual CNV events involving a given gene (e.g., 'CD12_Gain') and rows corresponding to PDX models. We then remove features which have the same value in 85% or more of our training PDX models.

Transcriptomic feature engineering

To reduce RNAseq data dimensionality from an initial input of 33,668 gene-level expression features, as well as to include state-of-the-art

knowledge of cancer signalling pathways and transcription factor activity, we computed (1) GSVA scores³⁸ (<http://www.biomedcentral.com/1471-2105/14/7>) using the GSVA R package (version 1.34.0, R 3.6.3, `kcdf = 'gaussian'`) on *tmm* expression levels and the MSigDB Hallmark gene sets³⁰ as well as (2) PROGENy scores computed using the progeny R package²⁹. Both sets of scores were computed separately for each train/test replicate (see following sections) to avoid any information leakage. This step yields to a total of 66 aggregated transcriptomic features (Supplementary Data 1). Finally, we considered that many PROGENy and Hallmarks gene set are partially overlapping: for example PROGENy's 'NFkB' set corresponds to Reactome's 'TAK1 activates NFkB by phosphorylation and activation of IKKs complex' and 'RIP-mediated NFkB activation via ZBP1', and thus it shares 8 of its 48 genes with PROGENy's 'TNFa' set (Reactome's 'TNF signalling'). To avoid excessive collinearity between scores based on overlapping gene sets, we first computed the Pearson correlation coefficient (PCC) for all pairs of engineered transcriptomic features over all instances in the training set, and considered as 'collinear' all pairs with a PCC larger than 7. Here, for each pair of collinear features, we discard the one with the higher Mann-Whitney U test p value between responder and non-responder PDXs in each training split.

Clinical feature engineering

We consolidated our clinical data by: (1) dropping any features with more than 40% missing values, (2) dropping redundant or inconsistent features ('OXALIPLATIN-based treatments', 'N', 'T', 'N of other metastatic resections before collected metastasis', 'M', 'Site M', 'Site of primary', 'Site of primary DICOT'), (3) converting 'Stage at first diagnosis' annotations to an integer score and retaining only the highest score for a given PDX model where multiple annotations are present, (4) converting the 'Lymph node density' annotations to a numerical score corresponding to the ratio of positive lymph nodes over the total lymph node count, (5) encoding all treatment backbone annotations as categorical features, (6) one-hot-encoding all sample anatomical location annotations. This yielded 25 features covering patient, previous treatment, and tumour metadata (Supplementary Data 1).

Single-omic exploration of IRCC-PDX data

We conducted UMAP dimensionality reduction⁴² across individual omics using `umap-learn` (v 0.5.2, <https://github.com/lmcinnes/umap>) followed by density-based HDBSCAN⁴³ (v0.8.29-1, <https://github.com/scikit-learn-contrib/hdbSCAN>) of the resulting 2d UMAP embeddings. We then assessed omic feature distribution across clusters for each omic and highlighted (Supplementary Fig. 12-15) those which differentiate best between clusters using a Kruskal-Wallis test (*W. H. Kruskal & W. W. Wallis, 'Use of Ranks in One-Criterion Variance Analysis', Journal of the American Statistical Association, Vol. 47, Issue 260, pp. 583-621, 1952*) as implemented in `scipy` v1.11.1 (Supplementary Data 2). Further, we computed a 'one vs all' χ^2 test of independence for each categorical feature and each cluster or a Mann-Whitney U test for each continuous feature and each cluster (Supplementary Data 2), using `scikit-learn` v1.02 or `scipy` v1.11.1. This procedure was replicated to compute enrichment analysis of covariates -- that is sample annotations that were not included in the UMAP input, including NMF methylation cluster labels, CRIS classes -- in the the above mentioned UMAP + HDBSCAN clusters, as well as of differential drug response across them (Supplementary Data 2). P values were corrected via Bonferroni across all feature and clusters, divided per omic.

Model architecture

For our cetuximab response model we selected a stacking classifier architecture. Stacking is an ensemble learning technique which combines the individual contributions of multiple classification models (level-I-classifiers) via a meta-classifier. Here, we use a soft voting classifier which outputs the final binary class labels (cetuximab non-

responder; cetuximab responder) based on the argmax of the sums of the predicted probabilities from the level-1-classifiers (scikit-learn VotingClassifier^{10,11}, v1.02).

Our CeSta classifier pipeline uses a late integration approach to prevent high-dimensional 'omics (transcriptomics, methylomics) from overwhelming smaller omics by dominating the selected feature set. We perform an initial round of single-omic supervised feature selection whose output is then piped into each of the four lvl 1 classifiers described below (Fig. 1a).

This selection step ranks features according to the product of (1) a feature rank based on the Fisher's exact statistic (scipy v1.9^{10,11}) for binary features or Mann-Whitney *U*-test statistic (scipy v1.9) for continuous features, (2) a feature rank based on percent lift, and (3) a feature rank based on logit model (statsmodels v0.13.2 logit) coefficients. A set of top *K* features is then selected from this ranked list, with *K* being one of CeSta's hyperparameters. This selection process is applied exclusively to the training set in each train, test split replicate during the internal validation (Fig. 1b and below) to avoid any information leakage. Model explanation in Fig. 4 shows feature importance and corresponding statistics and metrics obtained when training CeSta over the entire IRCC-PDX set as per the CeSta instance used for external validation on CR-PDX sample (Fig. 1b).

We used four distinct level-1-classifier pipelines (Fig. 1a): (1) a model-based (scikit-learn KNeighborsClassifier) forward feature selection, followed by elastic net penalised logistic regression (scikit-learn LogisticRegression with 'penalty' set to 'elasticnet'), (2) ANOVA feature selection (scikit-learn f_classif), followed either by a support vector classifier (scikit-learn SVC) or (3) an extra trees classifier (scikit-learn ExtraTreesClassifier), and (4) a CatBoost classifier (catBoost 1.0.5⁷⁹) trained on a common subset of features from CMP, then on IRCC PDX (continual learning).

Each level-1-classifier was trained (or re-trained in the case of CatBoost, see following sections) on a dataset of features selected (see above) from our 5 'omic data sources (mutation, CNV, expression, methylation, clinical). Finally, level-1-classifier prediction probabilities were stacked and taken as input by our meta-classifier (see above) which, in turn, gave in output a final binary prediction.

Model training, tuning, and validation

We generated 50 train, test split (150/81 PDXs) holdout shuffle replicates by performing stratified sampling from our IRCC-PDX dataset. The latter consisted of 231 fully characterised (targeted sequencing, RNAseq, methylation assay, clinical metadata) PDX models which were labelled as cetuximab responders or non-responders according to tumour volume variation after treatment, as described above.

For the internal validation analysis, we used a nested cross-validation approach (inspired by mlexend's StackingCVClassifier¹²) to tune and train 50 independent CeSta replicates, one per each train, test split. Each training set replicate was further split into 3 folds, and in 3 successive rounds, 2 folds were used (in turn) to fit the level-1-classifiers. In each round, the level-1-classifiers were then applied to the remaining 1 subset not used for model fitting in each iteration. The resulting predictions were then stacked and provided -- as input data -- to the meta-classifier. After comparing the meta-classifier's prediction on the validation fold to the corresponding true labels, the first-level classifiers were fit to the entire training set replicate (Fig. 1a, b).

This model training process was performed using a hyperparameter combination suggested by Optuna¹³ across 200 trials, while maximising the average of the area under the ROC curve (ROC AUC) computed over 3 training folds. Tuned parameter include: the number of top features selected during the first selection step, 'colsample_bylevel', 'depth', 'boosting_type', 'boosting_type', 'bootstrap_type' for the CatBoost classifier; number of sequentially-selected features, elastic net 'C', 'l1_ratio' for the Logistic elastic net classifier pipeline; number of ANOVA-selected features, 'C' and 'kernel' for the SVC

classifier pipeline; number of ANOVA-selected features, 'n_estimators' for the ExtraTrees classifier pipeline. This hyperparameter space search was performed, independently, for each model replicate.

Finally, we validated each of our 50 CeSta pipelines by predicting each PDX model in their respective test set as a cetuximab 'responder' or 'non-responder', and computing the resulting ROC AUC and ROC AUC 95 confidence interval (using DeLong's method) by comparing predicted and true labels.

For the external validation analysis, the same tuning, training, and validation process was repeated using the entire IRCC-PDX dataset as a training set (*N* = 231), and the CR-PDX dataset as a test set (*N* = 50) (Fig. 1b, c).

Performance baselines

To provide a realistic benchmark for CeSta performance, we define and train a number of alternative, multi-omic cetuximab sensitivity predictors. The latter are all trained, tuned, and validated using a set of 30 holdout shuffle replicates, analogous to the setup we use for CeSta internal validation in Fig. 1b.

'tripleNegRule' is a rule-based classifier based on the KRAS-NRAS-BRAF mutational signature: it will output a 'non-responder' prediction if any of these three genes is mutated in the current PDX example.

'tripleNegRightRule' is a rule-based classifier based on the KRAS-NRAS-BRAF mutational signature and the 'right colon' marker (i.e., whether the original tumour was located in the right portion of the patient's colon). This decision strategy originates from a retrospective analysis of triple negative patients from the CRYSTAL and FIRE-3 trials where right-sided tumours had significantly poorer prognosis and lower response to cetuximab treatment (Tejpar et al. ⁴⁴).

tripleNegRightRule will output a 'non-responder' prediction if either (1) any of KRAS, NRAS, BRAF is mutated or (2) the original tumour was right-sided.

'elNet_baseline' is an Elastic-Net net penalised logistic regression classifier (scikit-learn LogisticRegression with penalty set to 'elasticnet') based on four binary features encoding the mutational status of KRAS, BRAF, NRAS (i.e., the 'triple negative' CRC signature), and whether the primary tumour is located in the Right Colon. This corresponds to the state-of-the-art clinical signature for cetuximab sensitivity in CRC, as we discuss in Introduction and Results.

'rawLElasticnet' is an Elastic-Net net penalised logistic regression classifier which uses our full set of raw (non aggregated) features, that is: 110 binary gene mutational status features, 33,668 variance-normalised gene-level RNAseq data, and 1162 binary CNV events.

'MixOmics sPLS-DA' uses mixOmics¹⁴ multivariate integration approach, based on Partial Least Squares (PLS) regression and discriminant analysis, in which the most informative features (i.e., those that best discriminate between cetuximab responsive and non-responsive PDXs) from different 'omics are selected with the constraint of correlation between their first PLS components. More specifically, here we follow the multi-omic classification case study illustrated in <http://mixomics.org/methods/spls/>. We (1) perform LASSO feature selection (glmnet v4.2, <https://www.rdocumentation.org/packages/glmnet>) for methylation (700,298 probe-level features) and expression (33,668 gene level features), (2) use a sparse partial least-squares discriminant analysis model (sPLS DA) for single-omic dimensionality reduction, (3) followed by a DIABLO model for horizontal multiple 'omics integration. We optimise both the number of PLS components and the number of selected features for each omic and each component via 3-fold cross-validation on each training set replicate.

Finally, we validate these benchmark classifiers on each test set replicate, as described for our CeSta classifier in Fig. 3b by labelling each PDX model as a cetuximab 'responder' or 'non-responder', and computing the resulting ROC AUC by comparing predicted and true labels, again using DeLong's method for computing the ROC AUC 0.95 confidence interval where possible.

Cell line multiomic data source

The Cell Model Passport portal⁵⁵ (<https://cellmodelpassports.sanger.ac.uk/>) catalogues and curates multi-omic data for cancer cell line and organoid models. When combined with the Genomics of Drug Sensitivity in Cancer dataset (<https://www.sanger.ac.uk/tool/gdsc-genomics-drug-sensitivity-cancer/>), it provides genomics, transcriptomics, and cetuximab response data for 860 unique cancer cell line models (panCMP dataset). Here, we repeat the same data preprocessing and feature engineering steps we performed for the IRCC-PDX dataset, with the exception of the NMF-based clustering of methylation probes as this omic is missing from the CMP collection. Further, as cell line cetuximab response is quantified as IC50 values, rather than tumour volume change, here we dichotomise our target variable using the median IC50 for all cell lines in the panCMP dataset with lines falling below this threshold being labelled as ‘responders’.

For the purpose of comparing the predictive performance of a model trained on cell line data against one trained on PDX data, we generate a panCMP training set which includes a subset of 860 examples and their multi omic characterisation (Data and Code Availability). These features correspond to the subset available in both the aforementioned panCMP dataset, our IRCC-PDX dataset, and the CR-PDX dataset. We then train and tune a catBoost classifier pipeline (see above for pipeline architecture, hyperparameters) over this panCMP training set using an 8-fold cross-validation approach across 50 Optuna trials. This cell-line trained ‘base model’ is then provided, as a starting point for continual learning, to a second round of training (using the ‘init_model’ flag) over either an IRCC PDX train set split for internal validation, or the entire IRCC-PDX dataset for external validation on the CR-PDX dataset (Fig. 1b).

From the panCMP dataset, we can further subset 44 colorectal cell lines (CRC-CMP), which are characterised with the same subset of features as in the panCMP dataset. This context-specific dataset can also be used to train a catBoost ‘base model’ which we then feed into a second catBoost classifier trained on IRCC-PDX.

External validation: Charles River dataset

An independent CRC PDX cohort¹⁵ (https://www.cancermodels.org/search?filters=data_source%3ACRL+AND+dataset_available%3Acopy+number+alteration%2Cexpression%2Cmutation+AND+model_type%3APDX+AND+primary_site%3Acolon) has been assembled and characterised by our collaborators at Charles River Discovery Research Services (CR). We use 50 CRC LMX, first-pass PDX models corresponding to 50 unique patient samples characterised using a partially overlapping set of multi-omics features as in the IRCC PDX cohort. For missing features (e.g., methylation NMF cluster labels, some clinical annotations, some CNV events) we impute their values for this CR-PDX cohort using the mode for categorical features and the median for continuous features. We then use this CR-PDX dataset as a fully independent validation cohort to compare our stacked classifier’s performance against that of baseline models after training on the entire IRCC-PDX dataset.

Post-hoc model explanation

As a cross-model proxy for feature importance, for each feature, we calculate the mean of the absolute SHAP values (<https://github.com/slundberg/shap> v0.4) across all instances in the test set. We consider the absolute values as we do not want positive and negative values to offset each other. Features that have large mean absolute SHAP values are those that more significantly impact model predictions.

We are also interested in assessing, for a given classifier, which features perform equally well across different datasets (i.e., panCMP, IRCC-PDX, CR-PDX). To do so, we start by evaluating the relationship between a feature’s SHAP values and the target variable. A positive correlation here indicates that the model has identified and it is successfully exploiting an informative feature for its current classification

task. Given that SHAP values are additive, with the model’s prediction being the sum of all feature SHAPs, it makes sense to remove the effect of other features’ contribution by computing the partial correlation between each feature and the target after removing the effect of all other features (i.e., controlling variables). Specifically, here we use pingouin¹⁶ (v0.5.1) and its partial_corr function specifying, in turn, all features but one as x-covariates.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data generated in this study have been deposited in the European Genome-Phenome Archive (EGA) database under the accession numbers [EGAS0000100171](https://ega-archive.org/studies/EGAS0000100171) (targeted DNA sequencing) and [EGAS00001006492](https://ega-archive.org/studies/EGAS00001006492) (RNAseq), and are accessible upon request via the EGA portal, as required for personally identifiable data. In compliance with legal requirements to safeguard patient privacy, access to the raw data stored in the EGA is managed by a Data Access Committee (DAC) overseen by E.G., L.T. and the Data Sharing office at Sanger. Researchers can request access by reaching out to the EGA, which will inform the DAC of the request. The DAC will approve access within roughly two weeks and decide the duration for which access will be granted. The raw methylation data generated in this study have been deposited in the GEO database under the accession numbers [GSE208713](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE208713) (methylation data). The Khambata-Ford dataset is available on GEO under the accession number [GSE5851](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5851). Intermediate data needed to fully replicate the results in Figs. 1, 2 is available at https://bitbucket.org/uperron/ircc-pdx_exploration. Intermediate data, models and code needed to fully replicate CeSta and the results in Figs. 3, 4 is available at https://bitbucket.org/uperron/cesta_pdx. The CR-PDX data is available on CancerModels.org. The sample list can be retrieved, matching sample identifiers as detailed in Supplementary Data 1, at: https://www.cancermodels.org/search?filters=data_source%3ACRL+AND+dataset_available%3Acopy+number+alteration%2Cexpression%2Cmutation+AND+model_type%3APDX+AND+primary_site%3Acolon. The associated multi-omic data can be obtained on request by registering a free account on criver.com at https://compendium.criver.com/search?m%5Bids%5D%5B0%5D=10715&m%5Bs%5D=1&f%5Bg%5D%5Bht%5D%5Bs%5D=&f%5Bg%5D%5Bmu%5D%5Ba%5D=&f%5Bg%5D%5Bmu%5D%5Be%5D=&f%5Bg%5D%5Bcn%5D%5Bmi%5D=0&f%5Bg%5D%5Bcn%5D%5Bmx%5D=0&f%5Bp%5D%5Bage%5D%5Bmin%5D=&f%5Bp%5D%5Bage%5D%5Bmax%5D=&f%5Bp%5D%5Borigintypes%5D=&f%5Bp%5D%5Bgenders%5D=&f%5Bp%5D%5Bspecies_population_ids%5D=&f%5Bp%5D%5Bdifferentiations%5D=&g%5Baccessions%5D=all, after registering a free user-account. Processed RNAseq data for the IRCC-PDX and the CR-PDX collections is available on figShare at <https://figshare.com/s/35d13c7e7cf8f4759334>¹⁷. The cell-line multi-omic and drug response data used in this study can be accessed on the CellModelPassports⁵⁵ and the Genomics of Drug Sensitivity in Cancer¹⁸ data portals, respectively at: <https://cellmodelpassports.sanger.ac.uk/downloads> and https://www.cancerrxgene.org/downloads/drug_data. Source data are provided with this paper.

Code availability

The StromaDistiller Code¹⁹ and the RNASeq_biod_metadata code¹²⁰ implementing a computational pipeline tracking counts and metadata across different sequencing batches for xenografts/organoids RNAseq is available at <https://github.com/molinerisLab/StromaDistiller> and https://github.com/vodkatad/RNASeq_biod_metadata. The biodiversa_DE Code¹²¹ performing differential expression analysis with DESeq2 and various enrichment analyses on the results is available at https://github.com/vodkatad/biodiversa_DE. CELLector v2.0.0¹²² is available at <https://github.com/francescojm/CELLector>. Additional code performing multi-omic data preprocessing, normalisation, and integration and

reproducing the results depicted in Figs. 1, 2, 3, 4 is available at https://bitbucket.org/uperron/pdx_multiomics_integration_preproc and https://bitbucket.org/uperron/ircc_pdx_exploration and https://bitbucket.org/uperron/cesta_pdx, respectively.

References

1. Biller, L. H. & Schrag, D. Diagnosis and treatment of metastatic colorectal cancer: a review. *JAMA* **325**, 669–685 (2021).
2. Bertotti, A. et al. A molecularly annotated platform of patient-derived xenografts ('xenopatients') identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discov.* **1**, 508–523 (2011).
3. Burgenske, D. M. et al. Establishment of genetically diverse patient-derived xenografts of colorectal cancer. *Am. J. Cancer Res.* **4**, 824–837 (2014).
4. Isella, C. et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat. Commun.* **8**, 15107 (2017).
5. Stintzing, S. et al. FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab for metastatic colorectal cancer (FIRE-3): a post-hoc analysis of tumour dynamics in the final RAS wild-type subgroup of this randomised open-label phase 3 trial. *Lancet Oncol.* **17**, 1426–1434 (2016).
6. Heinemann, V. et al. FOLFIRI plus cetuximab or bevacizumab for advanced colorectal cancer: final survival and per-protocol analysis of FIRE-3, a randomised clinical trial. *Br. J. Cancer* **124**, 587–594 (2021).
7. Schwartzberg, L. S. et al. PEAK: a randomized, multicenter phase II study of panitumumab plus modified fluorouracil, leucovorin, and oxaliplatin (mFOLFOX6) or bevacizumab plus mFOLFOX6 in patients with previously untreated, unresectable, wild-type KRAS exon 2 metastatic colorectal cancer. *J. Clin. Oncol.* **32**, 2240–2247 (2014).
8. Van Cutsem, E. et al. ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Ann. Oncol.* **27**, 1386–1422 (2016).
9. Harrison, R. K. Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug Discov.* **15**, 817–818 (2016).
10. Yonesaka, K. et al. Activation of ERBB2 signaling causes resistance to the EGFR-directed therapeutic antibody cetuximab. *Sci. Transl. Med.* **3**, 99ra86 (2011).
11. van de Wetering, M. et al. Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* **161**, 933–945 (2015).
12. Schütte, M. et al. Molecular dissection of colorectal cancer in pre-clinical models identifies biomarkers predicting sensitivity to EGFR inhibitors. *Nat. Commun.* **8**, 14262 (2017).
13. Lupo, B. et al. Colorectal cancer residual disease at maximal response to EGFR blockade displays a druggable Paneth cell-like phenotype. *Sci. Transl. Med.* **12**, eaax8313 (2020).
14. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
15. Tsherniak, A., Vazquez, F., Montgomery, P. G. & Weir, B. A. Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
16. Ledford, H. US cancer institute overhauls cell lines: veteran cells to be replaced by human tumours grown in mice. *Nature* **530**, 391 (2016).
17. Santaballa, A. et al. SEOM clinical guideline in ovarian cancer (2016). *Clin. Transl. Oncol.* **18**, 1206–1212 (2016).
18. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
19. Trastulla, L., Noorbakhsh, J., Vazquez, F., McFarland, J. & Iorio, F. Computational estimation of quality and clinical relevance of cancer cell lines. *Mol. Syst. Biol.* **18**, e11017 (2022).
20. Bertotti, A. et al. The genomic landscape of response to EGFR blockade in colorectal cancer. *Nature* **526**, 263–267 (2015).
21. Ziemke, E. K. et al. Sensitivity of KRAS-mutant colorectal cancers to combination therapy that cotargets MEK and CDK4/6. *Clin. Cancer Res.* **22**, 405–414 (2016).
22. Julien, S. et al. Characterization of a large panel of patient-derived tumor xenografts representing the clinical heterogeneity of human colorectal cancer. *Clin. Cancer Res.* **18**, 5314–5328 (2012).
23. Cho, Y. B. et al. Colorectal cancer patient-derived xenografted tumors maintain characteristic features of the original tumors. *J. Surg. Res.* **187**, 502–509 (2014).
24. Cayrefourcq, L. et al. Establishment and characterization of a cell line from human circulating colon cancer cells. *Cancer Res.* **75**, 892–901 (2015).
25. Dudová, Z. et al. The EuroPDX data portal: an open platform for patient-derived cancer xenograft data sharing and visualization. *BMC Genom.* **23**, 156 (2022).
26. Byrne, A. T. et al. Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nat. Rev. Cancer* **17**, 254–268 (2017).
27. Kurilov, R., Haibe-Kains, B. & Brors, B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci. Rep.* **10**, 2849 (2020).
28. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinforma.* **11**, 367 (2010).
29. Schubert, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).
30. Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
31. Najgebauer, H. et al. CELLector: genomics-guided selection of cancer in vitro models. *Cell Syst.* **10**, 424–432.e6 (2020).
32. Tignanelli, C. J., Herrera Loeza, S. G. & Yeh, J. J. KRAS and PIK3CA mutation frequencies in patient-derived xenograft models of pancreatic and colorectal cancer are reflective of patient tumors and stable across passages. *Am. Surg.* **80**, 873–877 (2014).
33. Julien, S., Merino-Trigo, A., Lacroix, L., Pocard, M. Characterization of a large panel of patient-derived tumor xenografts representing the clinical heterogeneity of human colorectal cancer colorectal patient. *Clin. Cancer Drugs* **18**, 5314–5328 (2012).
34. Cheng, D. T. et al. Memorial Sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
35. Battaglin, F., Naseem, M., Lenz, H.-J. & Salem, M. E. Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives. *Clin. Adv. Hematol. Oncol.* **16**, 735–745 (2018).
36. Chang, M. T. et al. Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* **8**, 174–183 (2018).
37. Normanno, N. et al. Implications for KRAS status and EGFR-targeted therapies in metastatic CRC. *Nat. Rev. Clin. Oncol.* **6**, 519–527 (2009).
38. Woo, X. Y. et al. Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nat. Genet.* **53**, 86–99 (2021).
39. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci. Rep.* **7**, 16618 (2017).
40. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
41. Hinoue, T. et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22**, 271–282 (2012).
42. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4314> (2018).
43. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Proc.*

- Advances In Knowledge Discovery And Data Mining* (eds. Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G.) 160–172 (Springer Berlin Heidelberg, 2013).
44. Tejpar, S. et al. Prognostic and predictive relevance of primary tumor location in patients with RAS wild-type metastatic colorectal cancer: retrospective analyses of the CRYSTAL and FIRE-3 trials. *JAMA Oncol.* **3**, 194–201 (2017).
 45. Vogel, C. L. et al. First and subsequent cycle use of pegfilgrastim prevents febrile neutropenia in patients with breast cancer: a multicenter, double-blind, placebo-controlled phase III study. *J. Clin. Oncol.* **23**, 1178–1184 (2005).
 46. Yoshino, T. et al. Pan-Asian adapted ESMO consensus guidelines for the management of patients with metastatic colorectal cancer: a JSMO-ESMO initiative endorsed by CSCO, KACO, MOS, SSO and TOS. *Ann. Oncol.* **29**, 44–70 (2018).
 47. Benson, A. B. et al. Colon cancer, version 2.2021, NCCN clinical practice guidelines in oncology. *J. Natl Compr. Cancer Netw.* **19**, 329–359 (2021).
 48. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
 49. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
 50. Pernice, S. et al. CONNECTOR, fitting and clustering of longitudinal data to reveal a new risk stratification system. *Bioinformatics* **39**, btad201 (2023).
 51. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).
 52. Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).
 53. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *NIPS'22 Conference Proceedings*. 507–520 (NIPS, 2022).
 54. Sammut, S.-J. et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
 55. van der Meer, D. et al. Cell model passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.* **47**, D923–D929 (2019).
 56. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In *Proc. Python in Science Conference*. <https://doi.org/10.25080/majora-92bf1922-011> (2010).
 57. Ried, T. et al. The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. *Mol. Asp. Med.* **69**, 48–61 (2019).
 58. Rafa, L. et al. REG4 acts as a mitogenic, motility and pro-invasive factor for colon cancer cells. *Int. J. Oncol.* **36**, 689–698 (2010).
 59. Kang, G., Oh, I., Pyo, J., Kang, D. & Son, B. Clinicopathological significance and prognostic implications of REG4 immunohistochemical expression in colorectal cancer. *Medicina* **57**, 938 (2021).
 60. Hwang, J.-H. et al. A mutant KRAS-induced factor REG4 promotes cancer stem cell properties via Wnt/ β -catenin signaling. *Int. J. Cancer* **146**, 2877–2890 (2020).
 61. Jonker, D. J. et al. Epiregulin gene expression as a biomarker of benefit from cetuximab in the treatment of advanced colorectal cancer. *Br. J. Cancer* **110**, 648–655 (2014).
 62. Khambata-Ford, S. et al. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J. Clin. Oncol.* **25**, 3230–3237 (2007).
 63. Kumbrink, J. et al. p130Cas is correlated with EREG expression and a prognostic factor depending on colorectal cancer stage and localization reducing FOLFIRI efficacy. *Int. J. Mol. Sci.* **22**, 12364 (2021).
 64. Liu, S. et al. EREG-driven oncogenesis of head and neck squamous cell carcinoma exhibits higher sensitivity to Erlotinib therapy. *Theranostics* **10**, 10589–10605 (2020).
 65. Zhang, Y. et al. Epiregulin increases stemness-associated genes expression and promotes chemoresistance of non-small cell lung cancer via ERK signaling. *Stem Cell Res. Ther.* **13**, 197 (2022).
 66. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
 67. Park, S. H. et al. Sonic hedgehog pathway activation is associated with cetuximab resistance and EPHB3 receptor induction in colorectal cancer. *Theranostics* **9**, 2235–2251 (2019).
 68. Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 2: comparing methods. *J. Comput. Aided Mol. Des.* **30**, 103–126 (2016).
 69. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
 70. Sun, X. & Xu, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal. Process. Lett.* **21**, 1389–1393 (2014).
 71. Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals. *J. Comput. Aided Mol. Des.* **28**, 887–918 (2014).
 72. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *NIPS'17 Conference Proceedings*. 4768–4777 (NIPS, 2017).
 73. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
 74. Rivera, M. et al. Patient-derived xenograft (PDX) models of colorectal carcinoma (CRC) as a platform for chemosensitivity and biomarker analysis in personalized medicine. *Neoplasia* **23**, 21–35 (2021).
 75. Behan, F. M. et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).
 76. Ghandi, M. et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
 77. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
 78. Warren, A. et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nat. Commun.* **12**, 22 (2021).
 79. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulín, A. CatBoost: unbiased boosting with categorical features. In *NIPS'18 Conference Proceedings*. 6639–6649 (NIPS, 2017).
 80. Bishnupuri, K. S., Sainathan, S. K., Ciorba, M. A., Houchen, C. W. & Dieckgraefe, B. K. Reg4 interacts with CD44 to regulate proliferation and stemness of colorectal and pancreatic cancer cells. *Mol. Cancer Res.* **20**, 387–399 (2022).
 81. Zanella, E. R., Grassi, E. & Trusolino, L. Towards precision oncology with patient-derived xenografts. *Nat. Rev. Clin. Oncol.* **19**, 719–732 (2022).
 82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
 83. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **174**, 1034–1035 (2018).
 84. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
 85. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
 86. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
 87. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
 88. Seshan, V. E. & Olshen, A. B. DNACopy: a package for analyzing DNA copy data. *Bioconductor Vignette* (2014).

89. Beroukhi, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
90. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
91. van Dyk, E., Reinders, M. J. T. & Wessels, L. F. A. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. *Nucleic Acids Res.* **41**, e100 (2013).
92. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
93. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
94. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
95. Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
96. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
97. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
98. Hänzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7 (2013).
99. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
100. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
101. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).
102. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
103. Chen, Y.-A. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
104. Needhamsen, M. et al. Usability of human Infinium MethylationEPIC BeadChip for mouse DNA methylation studies. *BMC Bioinform.* **18**, 486 (2017).
105. Gujar, H., Liang, J. W., Wong, N. C. & Mozhui, K. Profiling DNA methylation differences between inbred mouse strains on the Illumina Human Infinium MethylationEPIC microarray. *PLoS ONE* **13**, e0193496 (2018).
106. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
107. Meehan, T. F. et al. PDX-MI: minimal information for patient-derived tumor xenograft models. *Cancer Res.* **77**, e62–e66 (2017).
108. Baralis, E., Bertotti, A., Fiori, A. & Grand, A. LAS: a software platform to support oncological data management. *J. Med. Syst.* **36**, S81–S90 (2012).
109. Kaur, M. & Grag, U. ECLAT algorithm for frequent itemsets generation. *Int. J. Comput. Syst. Softw. Eng.* **10**, 23263–23279 (2015).
110. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. In *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases* (2013).
111. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
112. Raschka, S. MLxtend: providing machine learning and data science utilities and extensions to Python’s scientific computing stack. *J. Open Source Softw.* **3**, 638 (2018).
113. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, 2019).
114. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: an R package for omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
115. Conte, N. et al. PDX Finder: a portal for patient-derived tumor xenograft model discovery. *Nucleic Acids Res.* **47**, D1073–D1079 (2019).
116. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).
117. Perron, U. et al. Data Package for Perron et al. (Nature Communications, 2024). figshare <https://doi.org/10.6084/M9.FIGSHARE.27103786> (2024).
118. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
119. Molineris, I. *molinerisLab/StromaDistiller: v1.0.0*. (Zenodo, 2024). <https://doi.org/10.5281/ZENODO.13682494>.
120. Grassi, E. *vodkatad/RNASeq_biod_metadata: CeSta Paper First Revision*. (Zenodo, 2024). <https://doi.org/10.5281/ZENODO.13682838>.
121. Grassi, E. *vodkatad/biodiversa_DE: CeSta Paper First Revision*. (Zenodo, 2024). <https://doi.org/10.5281/ZENODO.13640241>.
122. Trastulla, L. & Iorio, F. *francescojm/CELLector: v2.0.0*. (Zenodo, 2024). <https://doi.org/10.5281/ZENODO.13629554>.

Acknowledgements

This work was conducted with funding from AIRC, Associazione Italiana per la Ricerca sul Cancro, Investigator Grants 20697 (to AB), 22802 (to LT) and 28772 (to FI); AIRC 5×1000 grant 21091 (to AB, EM, and LT); AIRC/CRUK/FC AECC Accelerator Award 22795 (to LT). This work has been partially funded by European Union (European Research Council Consolidator - ERC) Consolidator Grants 724748 BEAT (to AB) and 101125051 DepSHOCK (to FI), H2020 grant agreement no. 754923 COLOSSUS (to LT); H2020 INFRAIA grant agreement no. 731105 EDiReX (to AB and EM). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. In addition, this work has been partially funded by Fondazione Piemontese per la Ricerca sul Cancro-ONLUS, 5×1000 Ministero della Salute 2016 (to LT) and 2022 (to EM and LT); Italian Ministry of Health, GR-2016–02362726 (to CI); Italian Ministry of University and Research, National Recovery and Resilience Plan, project PNC0000001 (to LT) and Ricerca Finalizzata 2021 Giovani Ricercatori, ID. GR-2021-12375316 (to EG). AB and LT are members of the EurOPDX Consortium. LMB PhD fellowship is funded by Nerviano Medical Sciences.

Author contributions

U.P., E.G., A.C., N.C., U.M., L.Tru., A.B. and F.I. conceived the project and scope. E.R.Z. derived and characterised the IRCC-PDX collection, collected metadata from the original patients, coordinated by E.M., M.E., L.Tru. and A.B., U.P., E.G., A.C., M.V., E.K., L.M.B., L.Tra., C.I. and I.M. processed and analysed the IRCC-PDX data. U.M., L.Tru., A.B. and F.I. supervised IRCC-PDX data processing and analysis. H.K. and J.S. derived, characterised, and analysed the CR-PDX collection. U.P. designed and implemented the CeSta pipeline. U.P., E.G. and A.C. drafted the manuscript and designed the figures. U.P., E.G., A.C., L.Tru., A.B. and F.I. edited and revised manuscript and figures. L.Tru., A.B. and

F.I. supervised the study. All authors discussed the results and contributed to the final manuscript.

Competing interests

FI receives funding from Open Targets, a public-private initiative involving academia and industry, and from Nerviano Medical Sciences and performs consultancy for the joint Cancer Research Horizon— AstraZeneca Functional Genomics Centre and for Mosaic T.X., L.T. has received research grants from Menarini, Merck KGaA, Merus, Pfizer, Servier and Symphogen. U.P. is a consultant for Omniscope Inc. H.K. and J.S. are employee of Charles River. U.M. is an employee and holder of company stock of AstraZeneca. All the other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-53163-y>.

Correspondence and requests for materials should be addressed to Livio Trusolino, Andrea Bertotti or Francesco Iorio.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

¹Human Technopole, Milano, Italy. ²Candiolo Cancer Institute FPO IRCCS, Candiolo, Torino, Italy. ³Department of Oncology, University of Torino, Candiolo, Torino, Italy. ⁴Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. ⁵Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ⁶Open Targets, Wellcome Genome Campus, Hinxton, UK. ⁷Nerviano Medical Sciences, Milan, Nerviano, Italy. ⁸Charles River Germany GmbH, Freiburg, Germany. ⁹Department of Life Sciences and Systems Biology, University of Torino, Torino, Italy. ¹⁰Josep Carreras Leukemia Research Institute (IJC), Badalona, Barcelona, Catalonia, Spain. ¹¹Centro de Investigacion Biomedica en Red Cancer (CIBERONC), Madrid, Spain. ¹²Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. ¹³Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Barcelona, Catalonia, Spain. ¹⁴European Molecular Biology Laboratory European Bioinformatics Institute, Cambridge, UK. ¹⁵Present address: Omniscope España, Barcelona, Spain. ¹⁶Present address: AstraZeneca Oncology R&D, Cambridge, UK. ¹⁷These authors contributed equally: Umberto Perron, Elena Grassi, Aikaterini Chatzipli. ¹⁸These authors jointly supervised this work: Livio Trusolino, Andrea Bertotti, Francesco Iorio. ✉ e-mail: livio.trusolino@ircc.it; andrea.bertotti@ircc.it; francesco.iorio@fht.org