Article

A gut microbial signature for combination immune checkpoint blockade across cancer types

Ashray Gunjur [1,2] ✉, **Yan Shao** [1], **Timothy Rozday**[1], **Oliver Klein**[3,4,5], **Andre Mu**[1,6], **Bastiaan W. Haak**[1,7], **Ben Markman** [8,9,10], **Damien Kee**[4,11,12], **Matteo S. Carlino**[13,14], **Craig Underhill** [15,16], **Sophia Frentzas**[8], **Michael Michael**[11,17], **Bo Gao**[13], **Jodie Palmer**[3], **Jonathan Cebon**[3,4], **Andreas Behren** [3], **David J. Adams** [2,18] **& Trevor D. Lawley** [1,18] ✉

Immune checkpoint blockade (ICB) targeting programmed cell death protein 1 (PD-1) and cytotoxic T lymphocyte protein 4 (CTLA-4) can induce remarkable, yet unpredictable, responses across a variety of cancers. Studies suggest that there is a relationship between a cancer patient's gut microbiota composition and clinical response to ICB; however, defining microbiome-based biomarkers that generalize across cohorts has been challenging. This may relate to previous efforts quantifying microbiota to species (or higher taxonomic rank) abundances, whereas microbial functions are often strain specific. Here, we performed deep shotgun metagenomic sequencing of baseline fecal samples from a unique, richly annotated phase 2 trial cohort of patients with diverse rare cancers treated with combination ICB ($n = 106$ discovery cohort). We demonstrate that strain-resolved microbial abundances improve machine learning predictions of ICB response and 12-month progression-free survival relative to models built using species-rank quantifications or comprehensive pretreatment clinical factors. Through a meta-analysis of gut metagenomes from a further six comparable studies ($n = 364$ validation cohort), we found cross-cancer (and cross-country) validity of strain−response signatures, but only when the training and test cohorts used concordant ICB regimens (anti-PD-1 monotherapy or combination anti-PD-1 plus anti-CTLA-4). This suggests that future development of gut microbiome diagnostics or therapeutics should be tailored according to ICB treatment regimen rather than according to cancer type.

The past decade has seen an 'immuno-oncology revolution' largely driven by the rapid uptake of immune checkpoint blockade (ICB) agents targeting cytotoxic T lymphocyte protein 4 (CTLA-4), programmed cell death protein 1 (PD-1) or programmed death ligand 1 (PD-L1, the ligand of PD-1). Combination ICB (CICB) targeting both PD-1 and CTLA-4 has demonstrated synergistic antitumor activity preclinically[1] and is now an approved standard of care for patients with diverse cancers, including melanoma[2], clear-cell renal cell carcinoma[3], non-small cell lung cancer

(NSCLC)[4], mesothelioma[5] and hepatocellular carcinoma[6]. However, this success is tempered by the unpredictable nature of responses (seen in only 20–60% of patients across these cancer indications[7]) and the more frequent severe immune-related adverse effects experienced with CICB when compared to anti-PD-1 or anti-PD-L1 monotherapy[8]. Thus, despite the promise it offers, the judicious use of CICB is paramount. Additionally, predictive biomarkers for tumor response and/or toxicity would be highly valuable to guide patient management.

Currently approved tumor-agnostic biomarkers for PD-1 blockade include tumor mutational burden and mismatch repair deficiency[9]; however, both have limitations and rely on available, contemporaneous tumor tissue. A promising 'tumor-extrinsic' avenue for predicting ICB response and/or toxicity a priori is assessing a patient's baseline gut microbiome composition, referring to the community of microbiota (predominantly bacteria) resident within the gastrointestinal tract. Culture-free methods to taxonomically profile fecal microbiomes have progressed from low-resolution 16S rRNA gene sequencing to high-resolution shotgun metagenomics, with studies of clinical cohorts finding associations between baseline *Akkermansia muciniphila* (lung cancer)[10–13] and *Faecalibacterium prausnitzii* (melanoma)[14–16] fecal abundances and tumor responses among anti-PD-1 recipients. Unfortunately, previous meta-analyses across metagenomic studies have found limited reproducibility of these candidate microbial biomarkers for ICB response[17–20]. Although this poor reproducibility may be partly attributable to methodological or geographic differences between studies, we hypothesize that species-level taxonomic biomarkers may lack the precision necessary to capture the specific microbial traits associated with ICB response or nonresponse. For example, there is growing awareness of the diversity of intraspecies (strain) variation among commensal bacteria (such as *A. muciniphila* and *F. prausnitzii*), with diverging functional potentials and differing associations with host phenotypes[21,22].

Here, we performed deep shotgun metagenomic sequencing of baseline fecal samples from patients on the CA209-538 clinical trial of ipilimumab (anti-CTLA-4) and nivolumab (anti-PD-1) for 106 patients with diverse rare cancers (our discovery cohort). Using a bespoke, genome-resolved metagenomics approach, we discovered baseline subspecies (strain-level) gut microbial abundance signatures of response that reproduce between cancer subtypes and externally to published CICB cohorts despite marked cohort heterogeneity. Notably, we found that the predictiveness of signatures trained on CICB cohorts does not extend to anti-PD-1 monotherapy cohorts. This suggests that, although tumor agnostic, different microbiota–host relationships are relevant to distinct ICB regimens.

Results

Clinical characteristics of the CA209-538 cohort

The CA209-538 clinical trial, titled 'A phase 2 trial of ipilimumab and nivolumab for the treatment of rare cancers', is a prospective, multi-center clinical trial (NCT02923934) that enrolled 120 patients with histologically confirmed advanced rare solid-organ cancers across five Australian hospital networks (Methods). Notably, patients had diverse tumor histologies grouped into three prespecified cohorts: upper gastrointestinal and biliary cancers (UGB), neuroendocrine neoplasms (NEN) and rare gynecological tumors (GYN). Most patients (*n* = 108) had received prior systemic anticancer therapies (median of one line (range 0–6 lines)). All participants were treated on trial with combination nivolumab and ipilimumab for up to four doses (induction), followed by nivolumab maintenance for up to 2 years or until progressive disease (PD) or unacceptable toxicity (Fig. 1a). The prespecified secondary endpoint of the trial was to develop 'tumor-agnostic' biomarkers for CICB response by leveraging the unique clinical trial design of CA209-538, which included patients with diverse cancers, but with highly standardized clinical and experimental procedures. Therefore, a pretreatment fecal sample was collected from most (*n* = 106)

participants (Table 1). No major clinical differences were observed between microbiome-evaluable patients and those who were not sampled (Supplementary Table 1).

The clinical efficacy and safety outcomes for subgroups from CA209-538 have been published previously[23–26]. As expected, overall survival (OS) significantly differed by histology (Extended Data Fig. 1a); however, progression-free survival (PFS) was more consistent (Extended Data Fig. 1b). Notably, the percentage of patients with an objective response (complete response (CR) or partial response (PR)) was remarkably stable across histological cohorts (24–25%) (Fig. 1b), with the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 best overall response (BOR) being strongly associated with PFS and OS (Fig. 1c,d). Using univariable statistical testing, we found a strong positive monotonic association between albumin and BOR (Kendall *P* = 0.0056) and a negative monotonic association between neutrophil-to-lymphocyte ratio (NLR) and BOR (Kendall *P* = 0.0033) (Extended Data Fig. 1c). This was particularly driven by patients with rapid clinical progression (clinical PD (cPD)) having significantly lower albumin and higher NLR, both responses to inflammation shown to be strongly prognostic across cancer types and treatment settings[27,28].

Microbiome profiling of baseline fecal samples

To understand the composition of patient gut microbiomes, we performed deep shotgun metagenomic sequencing of the 106 available baseline fecal samples (median 20.4 million paired-end reads per sample). For precise taxonomic quantification, we used a genome-resolved approach of first assembling a study-specific strain reference database using metagenome-assembled genomes (MAGs), supplemented with relevant Genome Taxonomy Database (GTDB) species reference genomes (SRGs) (Methods). Ultimately, this database included 1,397 strain genomes covering 904 known species and additionally included 34 'new' strains that could be taxonomically classified only to the genus level. The Bowtie 2 alignment rates to our tailored strain reference library were high (median 88.4%), with a median of 10.2 million mapped paired-end reads (50%) passing stringent quality control and used for precise strain quantification (Supplementary Fig. 1 and Methods).

We first evaluated whether there were gross compositional differences based on the patients' BOR. Notably, we found a positive monotonic association between BOR and the fecal Shannon diversity index, a common alpha diversity metric (Fig. 1e). Associations between alpha diversity and cancer patient outcomes have been found in the setting of patients receiving hematopoietic cell transplant[29] or cervical cancer chemoradiation[30] but not in anti-PD-1 recipients with metastatic melanoma[16,18]; thus, such associations may be treatment regimen specific. We then assessed intersample beta diversity using the Aitchison distance and also found gross microbial compositional differences by BOR group (permutational multivariate analysis of variance (PERMANOVA) *P* = 0.0319) (Fig. 1f). Indeed, among the 23 pre-treatment clinical and technical metadata tested, BOR group was the metadata variable explaining the most microbial variance (Extended Data Fig. 1d). By contrast, patient PFS at 12 months (PFS12) or OS at 12 months was associated with little microbial variance. A PERMANOVA of baseline microbial variance versus a moving PFS threshold revealed a peak association at <4 months (Extended Data Fig. 1e), indicating that, in our cohort, patients with rapid progression had the most distinct gross baseline microbial compositions.

Strain–response signatures are valid across cancer types

Given the gross compositional differences, we hypothesized that specific strains may allow for prediction of CICB efficacy in our cohort. We assessed objective response versus progression (RvsP), defined as a RECIST BOR of CR or PR versus PD or cPD, as our primary endpoint. In doing so, we excluded patients with a BOR of stable disease (SD) (*n* = 29), given its ambiguity in a pan-cancer cohort, in which it may represent disease control or simply indolent cancer behavior. As a
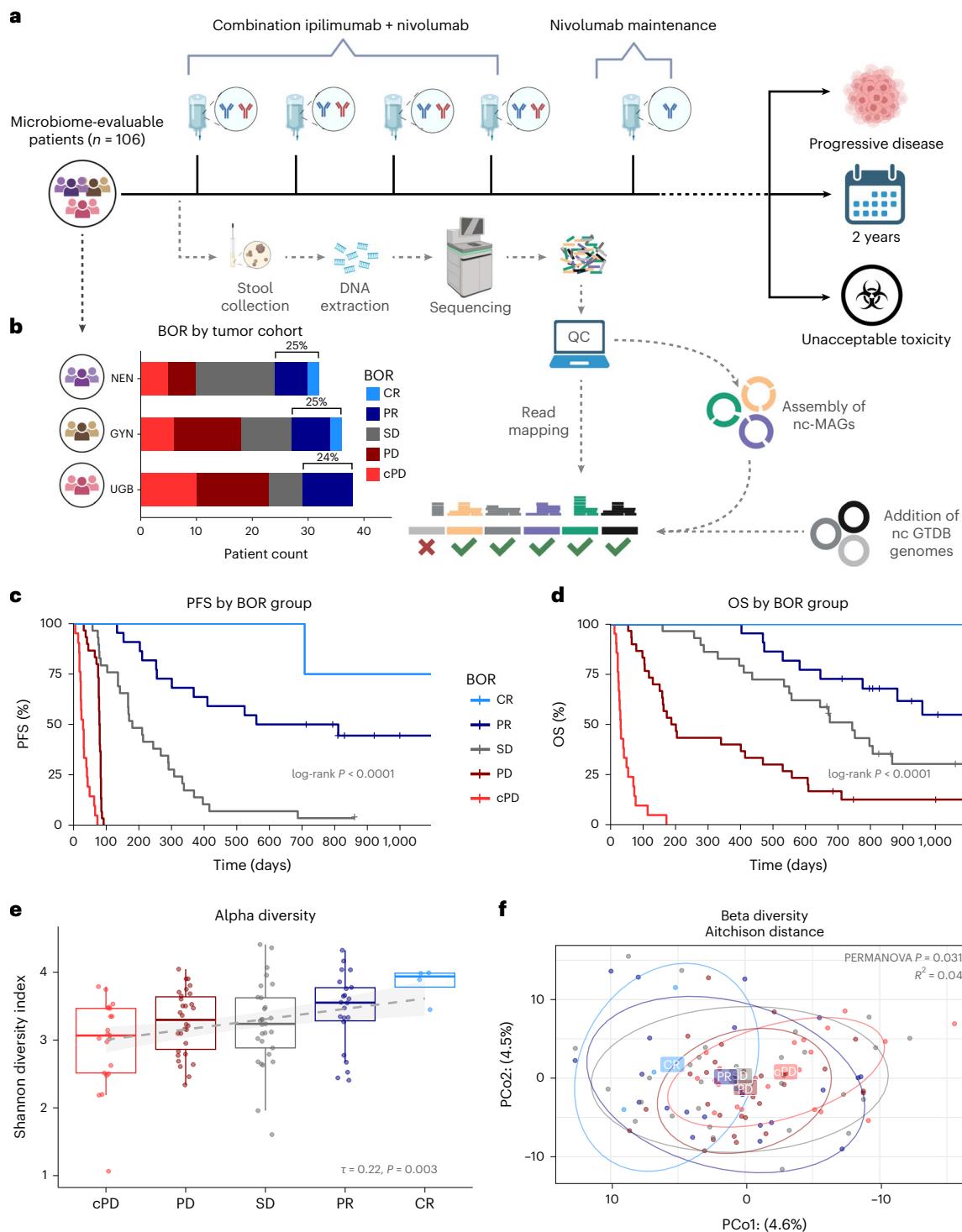
Fig. 1 | Clinical and gut microbiome compositional differences between responders and nonresponders. a, CA209-538 study and microbiome analysis schema (created using BioRender.com). Pretreatment fecal samples were collected from $n = 106$ trial participants and subjected to DNA extraction, shotgun metagenomic sequencing, and analysis using a genome-resolved metagenomics pipeline, involving quality control (QC), de novo assembly of near-complete MAGs (nc-MAGs) and precise read mapping. Further to the standard filters, reads mapping to genomes with <50% coverage breadth were removed. **b**, Bar plot of patient RECIST 1.1 BOR by histology cohort for microbiome-evaluable patients. The percentages of patients with an objective response (PR or CR) are indicated. **c**, Kaplan–Meier curve of PFS stratified by BOR category (cPD $n = 21$, PD $n = 30$, SD $n = 29$, PR $n = 22$, CR $n = 4$). Log-rank test $P = 2.1 \times 10^{-42}$. **d**, Kaplan–Meier curve of OS stratified by BOR category (cPD $n = 21$,

PD $n = 30$, SD $n = 29$, PR $n = 22$, CR $n = 4$). Log-rank test $P = 1.2 \times 10^{-34}$.
e, Boxplots of microbiome alpha diversity, as measured by the Shannon diversity index, across BOR categories (cPD $n = 21$, PD $n = 30$, SD $n = 29$, PR $n = 22$, CR $n = 4$). Boxplot center line indicates the median; box limits indicate the upper and lower quartiles; and whiskers indicate 1.5× the interquartile range. The linear model (line of best fit) for the Shannon diversity index and BOR (with shaded 95% confidence interval) is superimposed (in gray). Kendall τ and P values for the association between the Shannon diversity index and BOR are indicated.
f, Principal coordinate 1 (PCo1) versus 2 (PCo2) using the Aitchison distance of strain abundances, colored by patient BOR category. Ellipses depict 0.8 of each group's multivariate t distribution. PERMANOVA P value and R^2 using 9,999 permutations are indicated.

Table 1 | Baseline clinical characteristics by patient BOR category

Characteristics	Best response					P
	cPD (*n*=21)	PD (*n*=30)	SD (*n*=29)	PR (*n*=22)	CR (*n*=4)	
Histology cohort						
GYN	6 (28.6%)	12 (40.0%)	9 (31.0%)	7 (31.8%)	2 (50.0%)	0.18
NEN	5 (23.8%)	5 (16.7%)	14 (48.3%)	6 (27.3%)	2 (50.0%)	
UGB	10 (47.6%)	13 (43.3%)	6 (20.7%)	9 (40.9%)	0 (0%)	
No. of prior systemic therapies						
Mean (s.d.)	1.43 (0.811)	1.60 (1.22)	1.97 (1.61)	2.00 (1.23)	1.00 (0.82)	0.32
Median (min, max)	1 (0, 3)	1 (0, 5)	2 (0, 5)	2 (1, 5)	1 (0, 2)	
Measurable tumor (mm)						
Mean (s.d.)	115 (82.3)	72.6 (48.4)	84.1 (53.8)	90.0 (78.2)	70.0 (19.1)	0.49
Median (min, max)	108 (17.0, 344)	58.0 (24.0, 219)	77.0 (17.0, 220)	64.0 (11.0, 325)	77.5 (42.0, 83.0)	
Age (years)						
Mean (s.d.)	59.1 (13.9)	56.5 (15.1)	59.9 (14.3)	56.3 (12.5)	65.3 (9.29)	0.85
Median (min, max)	65.0 (20.0, 75.0)	62.5 (26.0, 75.0)	60.0 (22.0, 82.0)	53.5 (38.0, 74.0)	64.0 (57.0, 76.0)	
Sex						
Female	9 (42.9%)	22 (73.3%)	17 (58.6%)	17 (77.3%)	4 (100%)	0.048
Male	12 (57.1%)	8 (26.7%)	12 (41.4%)	5 (22.7%)	0 (0%)	
Site						
AUS	2 (9.5%)	8 (26.7%)	5 (17.2%)	7 (31.8%)	0 (0%)	0.34
BLA	2 (9.5%)	7 (23.3%)	4 (13.8%)	3 (13.6%)	0 (0%)	
BMO	4 (19.0%)	1 (3.3%)	3 (10.3%)	1 (4.5%)	0 (0%)	
MON	8 (38.1%)	6 (20.0%)	11 (37.9%)	3 (13.6%)	2 (50.0%)	
PMC	5 (23.8%)	8 (26.7%)	6 (20.7%)	8 (36.4%)	2 (50.0%)	
Season						
Autumn	9 (42.9%)	13 (43.3%)	12 (41.4%)	11 (50.0%)	1 (25.0%)	0.8
Spring	3 (14.3%)	2 (6.7%)	7 (24.1%)	2 (9.1%)	1 (25.0%)	
Summer	5 (23.8%)	10 (33.3%)	5 (17.2%)	5 (22.7%)	2 (50.0%)	
Winter	4 (19.0%)	5 (16.7%)	5 (17.2%)	4 (18.2%)	0 (0%)	
BMI (kg m^{-2})						
Mean (s.d.)	26.1 (5.74)	27.2 (5.30)	28.6 (6.23)	25.7 (5.05)	25.6 (3.49)	0.97
Median (min, max)	25.1 (19.1, 38.2)	28.2 (18.6, 37.0)	28.2 (18.9, 48.2)	25.0 (18.8, 35.3)	24.6 (22.8, 30.5)	
PPIs (<8 weeks)						
Yes	14 (66.7%)	9 (30.0%)	8 (27.6%)	7 (31.8%)	3 (75.0%)	0.017
No	7 (33.3%)	21 (70.0%)	21 (72.4%)	15 (68.2%)	1 (25.0%)	
Antibiotics (<8 weeks)						
Yes	3 (14.3%)	1 (3.3%)	3 (10.3%)	2 (9.1%)	0 (0%)	0.65
No	18 (85.7%)	29 (96.7%)	26 (89.7%)	20 (90.9%)	4 (100%)	
Platelets (×10^9 l^{-1})						
Mean (s.d.)	297 (134)	279 (81.1)	224 (97.8)	287 (118)	283 (50.1)	0.32
Median (min, max)	302 (87.0, 603)	273 (133, 575)	189 (62.0, 431)	276 (144, 559)	300 (211, 321)	
Albumin (g l^{-1})						
Mean (s.d.)	30.9 (5.66)	35.0 (5.34)	36.2 (4.34)	35.7 (3.47)	37.0 (2.16)	0.0056
Median (min, max)	32.0 (20.0, 38.0)	36.5 (20.0, 44.0)	37.0 (24.0, 44.0)	36.0 (29.0, 41.0)	36.5 (35.0, 40.0)	
NLR						
Mean (s.d.)	10.7 (15.8)	3.33 (1.99)	3.24 (2.18)	3.66 (2.57)	2.74 (0.144)	0.0033
Median (min, max)	5.27 (2.22, 66.0)	2.92 (0.970, 10.7)	2.72 (1.00, 10.0)	3.27 (0.960, 9.80)	2.70 (2.62, 2.95)	
LDH (U l^{-1})						
Mean (s.d.)	380 (208)	277 (155)	264 (143)	480 (696)	283 (59.2)	0.89
Median (min, max)	296 (149, 945)	215 (162, 898)	219 (128, 912)	302 (140, 3,440)	295 (202, 339)	

Clinical characteristics (metadata) are reported stratified by BOR category for microbiome-evaluable participants (*n*=106). Numerical metadata are summarized with both means and median values, and statistical associations with BOR (an ordinal variable increasing from cPD to CR) were computed using the Kendall τ test. Categorical metadata were analyzed using frequency tables, with statistical associations with BOR computed using the chi-squared test. AUS, Austin Hospital (Melbourne); BLA, Blacktown Hospital (Sydney); BMO, Border Medical Oncology (Albury); MON, Monash Hospital (Melbourne); PMC, Peter MacCallum Cancer Centre (Melbourne); BMI, body mass index; PPIs, proton-pump inhibitors; LDH, lactate dehydrogenase.

sensitivity analysis, we also evaluated PFS12, with responders and those with PFS12 largely overlapping given the durability of CICB efficacy (Extended Data Fig. 2a).

We used a supervised machine learning (ML) workflow (Fig. 2a). As input features (predictors), we tested the 15 potentially relevant clinical factors (Methods) and the microbial factors (centered log ratio (CLR)-transformed strain abundances) separately and combined to assess their relative and synergistic performance, respectively. In addition to strain-level rank, we tested microbial abundances aggregated to higher taxonomic ranks (species, genus and family levels) to determine the influence of taxonomic resolution on predictive performance. For each feature set, we performed a thorough random hyperparameter search across 1,000 iterations of a 20 times repeated fivefold cross-validation (Methods). For predictions, we used a random forest (RF) classifier, previously shown to generally outperform other classical ML algorithms for microbiome–host predictions[31].

Interestingly, we found that clinical factors alone were poorly predictive of RvsP (mean receiver operating characteristic (ROC) area under the curve (AUC) = 0.56) (Fig. 2b). This was despite the previously observed relationship between low blood albumin, high NLR and cPD, suggesting that these factors are more useful for delineating patients with the worst prognosis rather than distinguishing responders and nonresponders. Furthermore, it affirms the current difficulty of predicting clinical activity using routinely available factors and emphasizes the need for further technical innovation. In contrast, clinical factors were more predictive of PFS12 (AUC = 0.65; Extended Data Fig. 2b), inferring that these are more prognostic markers than predictors of antitumor activity.

When microbiome features were used, there was a positive monotonic association between the mean AUC score and taxonomic resolution for both endpoints (increasing from family to strain level) (Kendall $P = 1.1 \times 10^{-11}$ for RvsP, $P = 7.1 \times 10^{-15}$ for PFS12). In particular, strain-resolved abundances provided the best predictive performance (AUC = 0.73 for RvsP, AUC = 0.70 for PFS12), significantly outperforming the more common species-level abundances. Consistent with their poor standalone performance, clinical factors failed to augment microbiome predictors. Overall, these data suggest that microbial abundances, especially at strain-level resolution, are more valuable in predicting tumor response or landmark PFS than higher taxonomic aggregations or clinical features.

We subsequently focused on strain–RvsP classifiers, given their superior performance and larger incremental benefit over routine clinical factors. We were particularly interested in assessing the concordance of strain–RvsP predictions from the entire cohort ($n = 77$ evaluable) with actual patient BOR outcomes. Notably, despite being trained on binary RvsP, the predicted probabilities of patients were correctly ranked by their actual BOR category (Kendall $P < 2.2 \times 10^{-16}$), including (on average) central predictions for the SD group that were 'unseen' during model training ($n = 29$) (Extended Data Fig. 2c). Intrigued, we assessed whether RvsP predictions could distinguish a 'better' or 'worse' SD group. Indeed, we found a nonsignificant improvement in the OS of patients with SD with an above-median RvsP prediction, although this analysis was likely underpowered (log-rank $P = 0.17$; Extended Data Fig. 2d).

Finally, a key priority was to identify whether microbial signatures are tumor agnostic; that is, whether they generalize from one distinct tumor type to another. As our study naturally has three distinct cancer cohorts (GYN, NEN and UGB), we performed a leave-one-group-out cross-validation (training strain–RvsP classifiers using two groups and then testing on the left-out group). Notably, the mean AUC of the left-out group was consistently superior to that of a random model (overall mean AUC = 0.75) (Fig. 2c). Although the small sample size limits its interpretability, the particularly good performance for the UGB and GYN groups may reflect the specific relevance of the gut microbiome in these cancers.

Our ML analysis of our discovery cohort demonstrates that strain-level gut microbial predictors of CICB response may be relatively robust across diverse cancer types and are superior to ML predictors built using routine clinically available data. Furthermore, predictions trained on binary RvsP appear to capture the RECIST BOR biologically and may have utility for predicting the durability of SD.

Faecalibacterium strains are positively implicated

We next sought to understand which features (strain abundances) were most important in driving the strain–RvsP model predictions. To do this, we used the SHapley Additive exPlanations (SHAP) 'Tree-Explainer' algorithm[32] (Methods). We first noted that, although most strains contributed little to predictions, a few were disproportionately important (Extended Data Fig. 3a). Twenty-two strains were within half as impactful as the most important strain (a strain of *Faecalibacterium* sp900539885, an uncultured species), which we opted to focus on subsequently. Interestingly, these strains were neither rare (<5% prevalent) nor core (>50% prevalent) taxa within our cohort (Extended Data Fig. 3b).

To visualize the phylogenetic relationships of these 'top 22' strains in the context of all study-specific bacterial strains, we constructed an approximately maximum-likelihood phylogenetic tree using the GTDB toolkit (GTDB-tk) (Methods and Fig. 3a). This demonstrated that 20 of the 22 strains were gram positives, with most (18 of 20) belonging to the Firmicutes (Bacillota) phylum. The most 'beneficial' strains (that is, higher strain abundances shifted predictions toward 'response') clustered in one clade of the Ruminococcaceae family, with four being strains within the *Faecalibacterium* genus. Until recently, the National Center for Biotechnology Information taxonomy database recognized only one species within the genus *Faecalibacterium* (*F. prausnitzii*)[33], and its fecal abundance has been associated with good general health[34] and response to anti-PD-1 monotherapy in patients with melanoma[16] or hepatobiliary cancers[15]. However, more recent analyses have revealed considerable phylogenetic and functional diversity within the *F. prausnitzii* species complex[22]. In keeping with this, at the 98% genomic identity threshold, our custom strain reference library included $n = 35$ distinct *Faecalibacterium* strains (from $n = 13$ distinct species), with the most important (and prevalent) clustering near the *F. prausnitzii D* phylogenetic clade (Supplementary Fig. 2).

Conversely, 15 of the 22 strains appeared to have a negative association with response in our discovery cohort. As before, most were Firmicutes, with 6, 3 and 2 (of the 15) strains belonging to the Lachnospiraceae, Oscillospiraceae and Ruminococcaceae families, respectively. Notably, eight of these strains belonged to thus far uncultivated (and thus unnamed) species. The remaining four 'negative' strains belonged to the species *Bifidobacterium dentium*, *A. muciniphila B* and *Spyradocola merdavium*. It should be noted that *A. muciniphila B* is a distinct species from *A. muciniphila*; although the latter was positively implicated in anti-PD-1 efficacy in NSCLC[13] (also positive in our study but not within the top 22 strains), recent analyses have revealed that it is phylogenetically and phenotypically distinct from *A. muciniphila B* (known as *Akkermansia* SGB9228 by MetaPhlAn4 taxonomy)[21]. The juxtaposition of *Bifidobacterium longum* 1 and *B. dentium* 1 as positive and negative, respectively, also highlights how closely related taxa can have discordant relationships with host phenotypes. Indeed, while the species *B. longum* has been linked to positive health outcomes, such as protection from inflammatory bowel disease[35], protection from childhood malnutrition[36], and anti-PD-1 responses[37], *B. dentium* is a known oral opportunistic pathogen linked to tooth decay[38].

We next aimed to interrogate the genomes of the top 22 strains to understand functional potentials that may underpin their strong (positive and negative) response associations. We first evaluated them for virulence factors and found that they harbored none, suggesting that even the negative strains are not prototypical 'pathogens'. To look more broadly at strain functional potential, we queried the presence
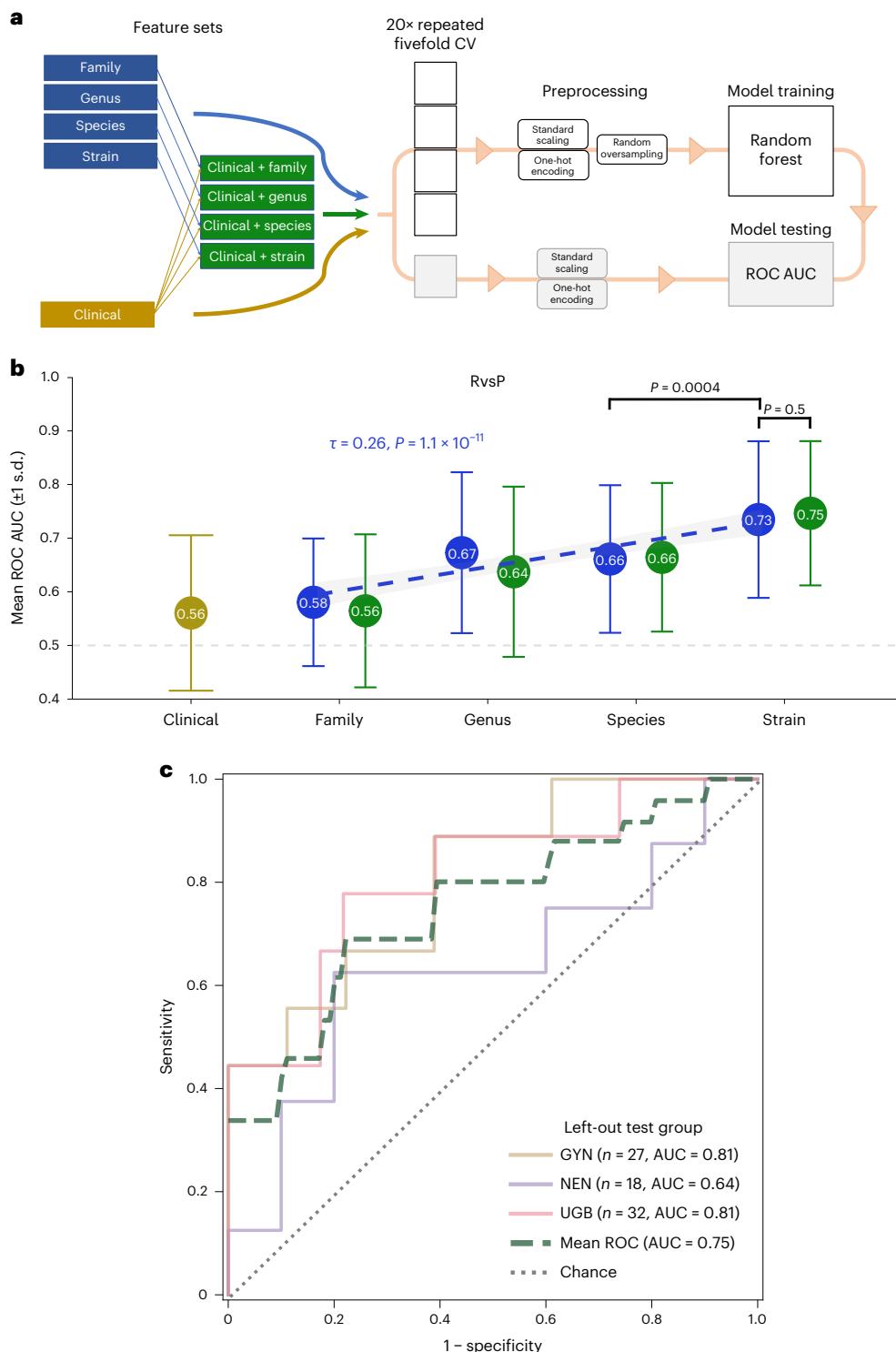
Fig. 2 | Strain-resolution gut microbial signatures outperform clinical predictors and cross-validate across tumor histology types. a, Schematic of the supervised ML framework. Input features (clinical, microbiome or combined) and the target variables (RvsP or PFS12) were split into five folds (four training folds, one testing fold). The process was repeated 20 times per iteration, with the AUC score used to select the best hyperparameters. CV, cross-validation. **b**, AUC scores for the best iteration of RvsP classifiers for each feature set combination during 20 times repeated fivefold cross-validation (100 folds each): clinical (yellow), microbiome (blue) and combined (green), at different taxonomic resolutions. Data represent the mean (circle) and s.d. (error bars) over the

100 folds. The linear model (line of best fit) for the AUC score and taxonomic rank of microbiome-only feature sets (with shaded 95% confidence interval) is superimposed. Kendall τ and P values for the association between the AUC score and taxonomic rank of microbiome-only feature sets are indicated. The Mann–Whitney *U* test P value for comparing the AUCs of specific pairwise feature sets (depicted by calipers) is also indicated. **c**, ROC curves for the strain–RvsP classifiers retrained using leave-one-histology-cohort-out cross-validation. Model training and testing were repeated 100 times, with predictions averaged to account for model stochasticity.

or absence of metabolic pathways using the tool gapseq (Methods). As expected, we observed clustering of metabolic potential by phylogeny; however, the two negative Ruminococcaceae (strains of the *Ruthenibacterium lactatiformans* and *Avimicrobium caecorum* species) were quite distinct from the five 'positive' strains (Extended Data Fig. 3c).

We hypothesized that specific metabolic functions may distinguish these negative and positive Ruminococcaceae. One metabolite of particular interest was butyrate, given that it has been implicated in anticancer cytotoxic T cell activation preclinically[39–41], and fecal butyrate has been positively associated with ICB efficacy in clinical cohorts[42,43]. Additionally, although butyrate-producing potential has previously been broadly ascribed to Ruminococcaceae, more recent analyses have revealed marked strain-level variation within this family[44]. Indeed, the acetyl-CoA butyrate pathway (which dominates among Firmicutes bacteria) was complete in all (five of five) positive but no negative (none of two) top 22 Ruminococcaceae (Fig. 3b). In contrast, taking a 'strain-agnostic' approach of quantifying the abundance of the acetyl-CoA butyrate terminal enzymes (*but* + *buk*) in metagenomic samples did not reveal a significant enrichment in responders (Fig. 3c), highlighting the need for strain-aware approaches to develop context-specific functional hypotheses.

Microbial signatures may be ICB regimen specific

To evaluate the external generalizability of our strain–RvsP signature, we reanalyzed all comparable shotgun metagenomic cohorts (Methods and Supplementary Fig. 3). We included cohorts that analyzed baseline (±15 days of ICB commencement) fecal samples, performed Illumina paired-end shotgun metagenomic sequencing, and provided either RECIST BOR (five studies) or pathological response (one study) metadata. Including our discovery cohort (CA209-538 cohort), the seven studies recruited participants from 11 cities across five countries (United States, United Kingdom, Netherlands, Spain and Australia) (Fig. 4a) and represent *n* = 470 total patients (*n* = 383 after excluding patients with a BOR of SD). Quality-controlled reads were mapped to the same reference library to estimate abundances for the same 1,397 strains. Although we were mindful that the reference library derived from the CA209-538 cohort might not represent all bacterial strains in external studies, we were reassured by both the high overall Bowtie 2 alignment rates (median 79.2–87.6% across external studies) and the high proportion of quality-controlled reads used for abundance estimation after stringent filtering (median 50.7–62.1% across external studies) (Extended Data Fig. 4a).

A summary of the key characteristics of the included studies is provided in Table 2. Given that all external studies evaluated patients with melanoma, known to be particularly amenable to ICB, it is not surprising that their objective response rates trended higher than those in our study that evaluated patients with diverse rare cancer types (38–84% versus 25%; Fig. 4a). This highlights that tumor type is an important variable in determining ICB response but does not preclude the existence of universal gut microbiota that may enhance or detract from an individual's likelihood of showing an antitumor ICB response.

A PERMANOVA of individual metadata variables revealed that the leading sources of microbial variance across the meta-cohort were study site (city) (9.3%) and DNA extraction kit (8.0%) (Extended Data Fig. 4b). However, these two factors were also strongly associated with one another (chi-squared test *P* < 2.2 × 10⁻¹⁶), with distinct studies recruiting participants from specific cities but also using distinct DNA extraction kits (Extended Data Fig. 4c,d). Although it would be desirable to 'correct' for DNA extraction kit (which has a well-described influence on downstream microbial quantifications[45]), this would likely also mitigate the true biological variance caused by patient geography[46] (which is important when evaluating the cross-country validity of a biomarker). Furthermore, a recent reanalysis of an intratumoral microbiome meta-analysis raised concerns that statistical batch correction may artificially inflate cross-cohort predictions due to data leakage[47]. Therefore, to evaluate the performance of our strain–RvsP classifier as robustly as possible, we opted not to adjust abundances beyond CLR transformation.

Given their distinct mechanisms of action, we were particularly interested in differentially evaluating performance on CICB and anti-PD-1 monotherapy cohorts. Of the six external studies, two comprised only anti-PD-1 recipients, two comprised only CICB recipients and two comprised both and were split based on regimen, creating eight external validation cohorts (four CICB, four anti-PD-1). Notably, there was a marked difference in the performance of the CA209-538 strain–RvsP signature between these groups, with overall modest external generalizability to CICB cohorts (mean AUC = 0.65; Fig. 4b) but no generalizability to anti-PD-1 cohorts (mean AUC = 0.51; Fig. 4c).

Intrigued, we sought to use our meta-cohort to evaluate whether this difference could also be seen more generally. We thus trained and tested strain–RvsP RF classifiers using all strain abundances and every pairwise combination of cohorts (nine cohorts, keeping 2017_Frankel and 2022_Lee split by ICB regimen) and evaluated AUCs. Consistent with our previous observation, we found that the predictive performance was better when training and testing on 'concordant' cohorts—that is, when the training and test cohorts received the same ICB regimen—rather than 'discordant' cohorts (Fig. 4d). Importantly, this was also true for strain–RvsP signatures trained on anti-PD-1 monotherapy cohorts. Taken together, the results showed a significant improvement in the cross-study strain–RvsP predictive performance in concordant rather than discordant regimen cohorts (Mann–Whitney *U* test *P* = 2.8 × 10⁻⁷).

Discussion

In this study, we used strain-resolved metagenomic classification to discover a signature of 22 gut microbial strains associated with response to combination ipilimumab (anti-CTLA-4) plus nivolumab (anti-PD-1) in a phase 2 trial cohort of Australian patients with diverse rare cancers (*n* = 106). To our knowledge, this represents the largest gut microbiome study of patients treated with CICB published to date. Using supervised ML, we demonstrate the value that precise, strain-level gut microbial quantifications provide in predicting clinical response or PFS12, exceeding the value of routinely available clinical information or that of higher taxonomic rank abundances. Furthermore, we show the external generalizability of strain-level response signatures across cancer histology types and countries, both within the trial (comparing across the predetermined histology cohorts) and externally (to metastatic melanoma cohorts from other industrialized countries). This was despite a

Fig. 3 | Firmicutes bacteria dominate the gut microbiome strain–response signature. a, Phylogenetic tree of bacterial strains in our custom reference library (*n* = 1,391 strains, excluding *n* = 6 archaea), highlighting the top 22 strains (labels are colored by impact (that is, feature importance) on RvsP predictions). Four main phyla are shown by the colored ring, with the Ruminococcaceae, Oscillospiraceae and Lachnospiraceae families highlighted. The scale for phylogenetic distance is shown in the center of the tree. **b**, Phylogenetic tree of the top 22 strains, with the tips colored by strain impact and sized by strain prevalence. The adjacent heat map depicts the presence or absence of genes within the primary butyrate-producing (acetyl-CoA) pathway. Full enzyme (encoding gene) names: acetyl-CoA acetyltransferase (*thl*), β-hydroxybutyryl-CoA dehydrogenase (*bhbd*), crotonase (*cro*), butyryl-CoA dehydrogenase (*bcd*), and the alternative terminal enzymes butyryl-CoA:acetate CoA transferase (*but*) and butyrate kinase (*buk*). **c**, Boxplots of the sample-wise abundance of butyrate acetyl-CoA terminal enzymes (*but* + *buk*), split by patient response (progression (P) *n* = 51, response (R) *n* = 26). Boxplot center line indicates the median; box limits indicate the upper and lower quartiles; and whiskers indicate 1.5× the interquartile range. Abundance is normalized as reads per million (RPM). *P* value by the Mann–Whitney *U* test is indicated.

strong heterogeneity in microbiome composition across cohorts, likely influenced by divergent fecal collection and DNA extraction methods. Finally, we observed a striking difference in the cross-study performance of response classifiers trained and tested on concordant versus discordant ICB cohorts, implying that different microbial relationships likely underlie these distinct treatment regimens.
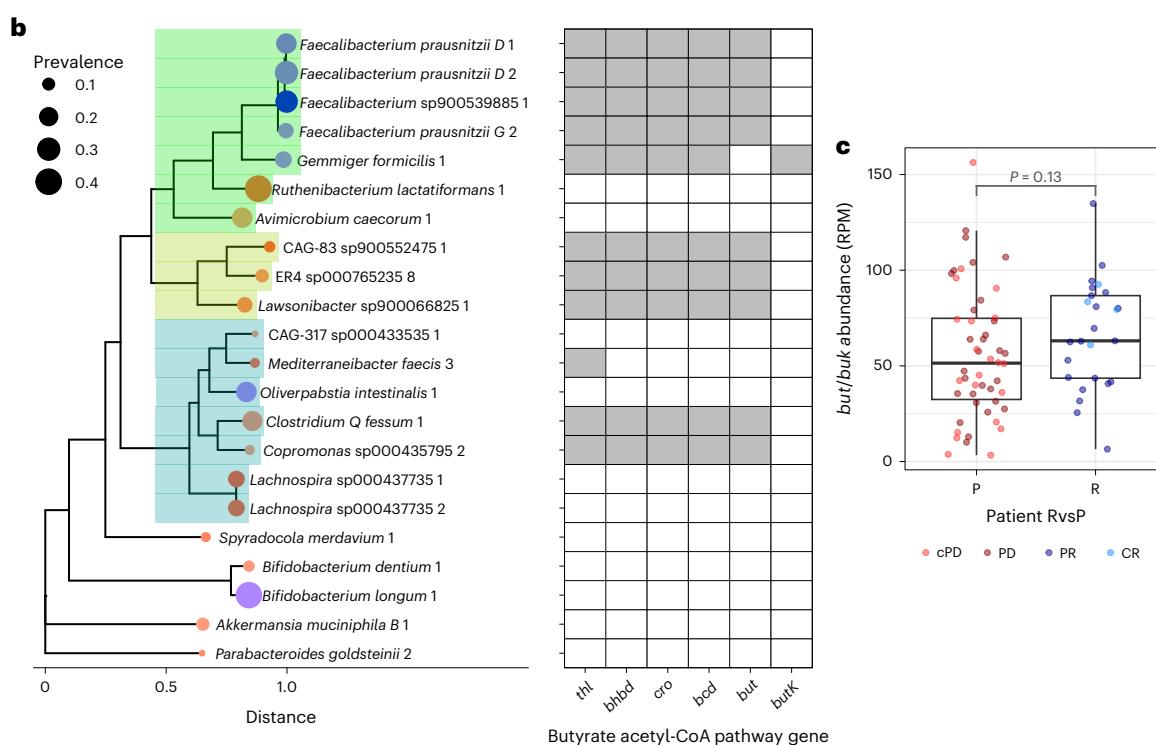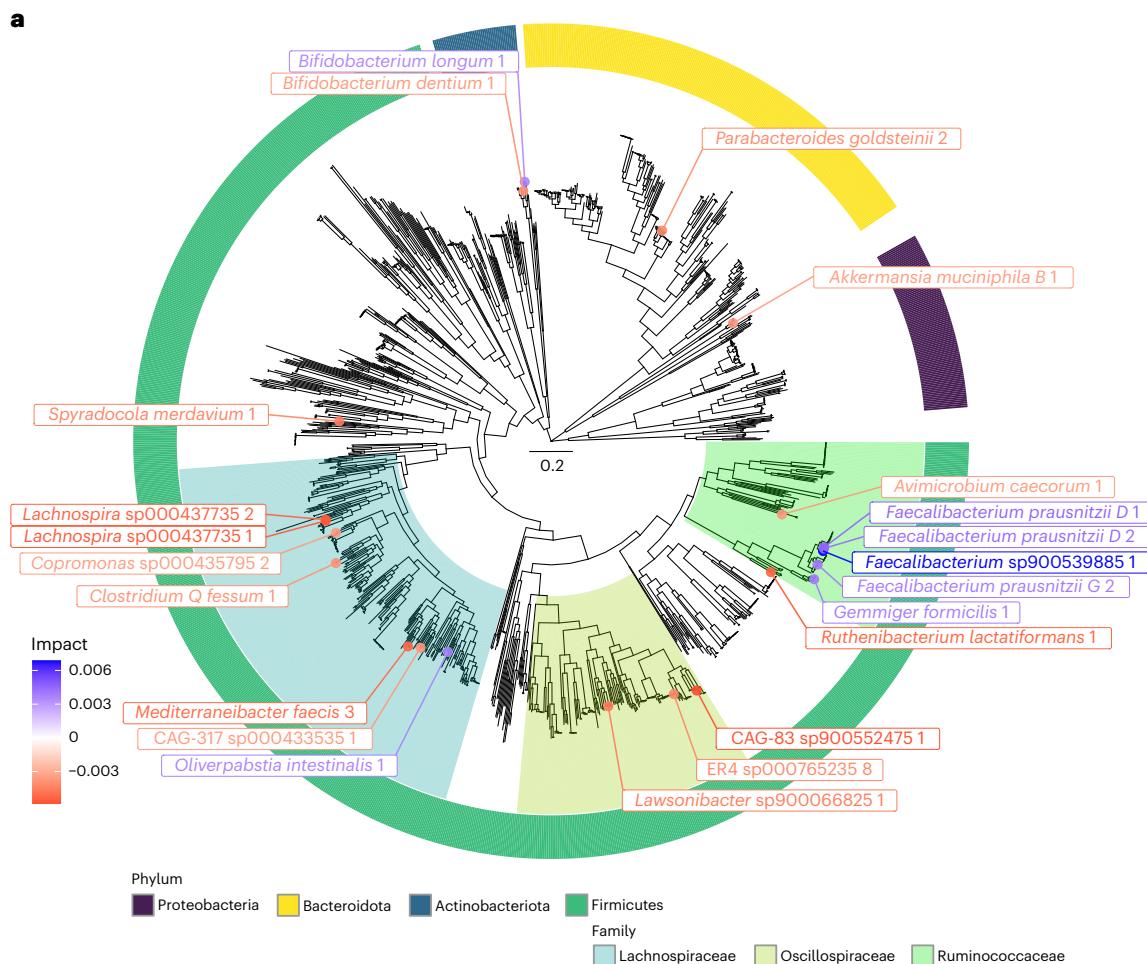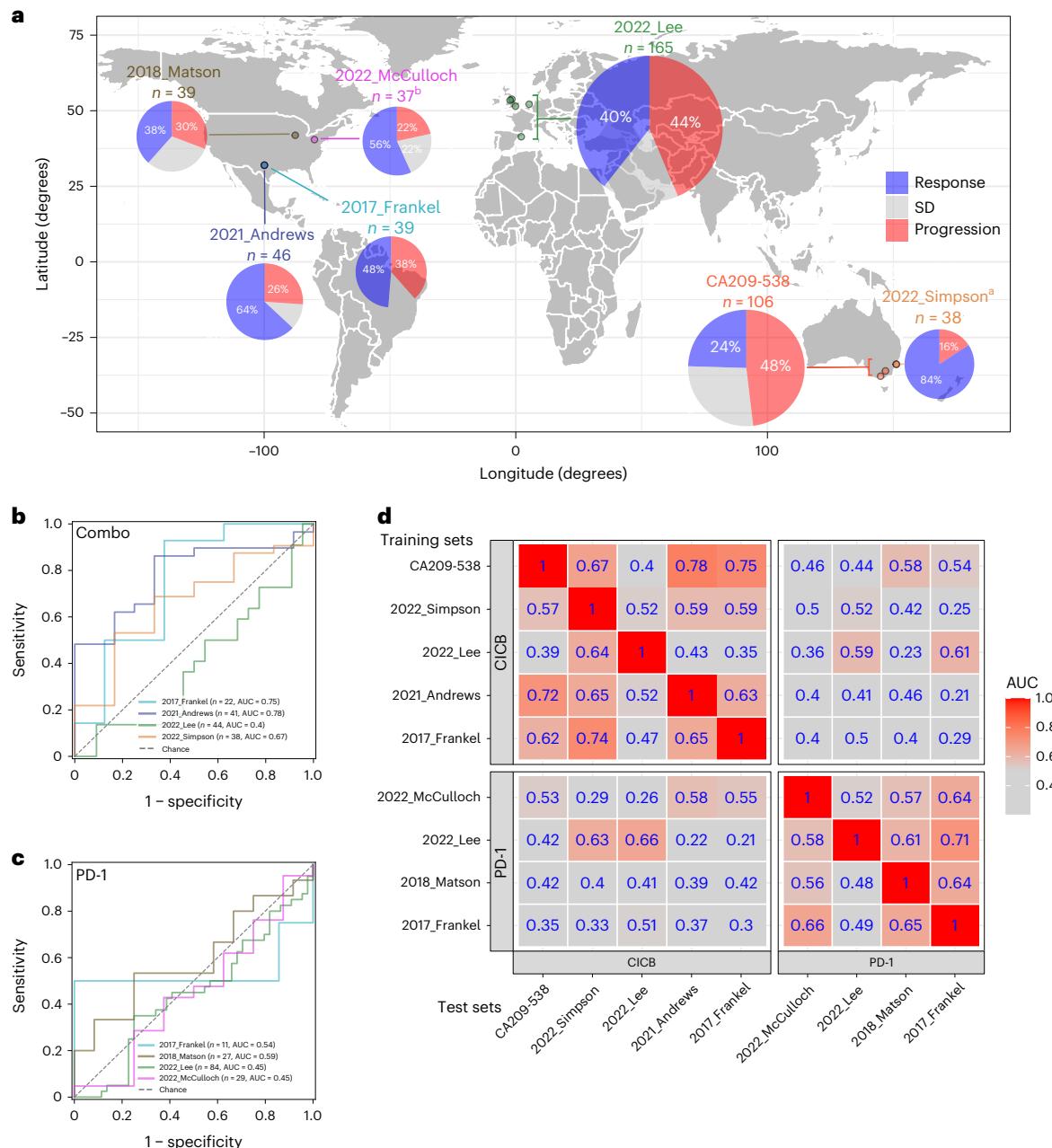
Fig. 4 | Meta-analysis reveals that gut microbiome strain–response signatures are ICB regimen specific. a, World map showing the studies included in our meta-analysis. Bordered circles depict the coordinates of recruiting sites (cities). Pie charts depict the proportion of patients with tumor response, progression or SD. The area of the pie charts depicts the sample size. [a]2022_Simpson studied neoadjuvant ipilimumab + nivolumab for stage III melanoma and thus used pathological response criteria (International Neoadjuvant Melanoma Consortium criteria); all other studies used the RECIST 1.1 criteria.

[b]For this study, only the subset of patients (*n* = 37) with stool collected within 15 days of the start of ICB therapy was included in the meta-analysis. **b**, ROC curve of strain–RvsP classifiers trained on the discovery cohort (CA209-538) and tested on external CICB cohorts separately. **c**, ROC curve of strain–RvsP classifiers trained on the discovery cohort (CA209-538) and tested on external anti-PD-1 monotherapy cohorts separately. **d**, Heat map denoting the AUC scores for strain–RvsP classifiers trained on one dataset (column) and tested on another (rows). Panels are faceted by ICB regimen (CICB or anti-PD-1 monotherapy).

Given the success of combination anti-PD-1 and anti-CTLA-4 ICB across diverse cancers, there is great interest in defining tumor-agnostic pretreatment biomarkers, including through using gut microbial abundance signatures. A recent review by Thomas et al.[20] defined cross-cancer ICB response ('Gut OncoMicrobiome Signature') implemented using species-level abundances. This study differs, first, in using strain-level signatures and, second, by deliberately splitting cohorts into those receiving anti-PD-1 monotherapy and those receiving anti-PD-1 plus anti-CTLA-4 CICB. Of note, although Thomas et al. found good left-out performance for the exclusively anti-PD-1-treated

NSCLC and renal cell carcinoma cohorts, performance was poor among left-out melanoma cohorts, potentially due to patients receiving monotherapy and those receiving CICB being admixed.

Although the external performance of the CA209-538 strain–response signature fell short of what is required for clinical use, its performance was remarkably better in CICB (AUC = 0.67, 0.40, 0.78 and 0.75) than anti-PD-1 (AUC = 0.46, 0.44, 0.58 and 0.54) melanoma cohorts from other industrialized countries. Consistent with this, strain–response signatures trained on external cohorts were also superior when tested on concordant rather than discordant regimen

Table 2 | Characteristics of studies included in the meta-analysis

Characteristics	Study						
	CA209-538 (*n*=106)	2022_Simpson[43] (*n*=38)	2021_McCulloch[19] (*n*=37)[a]	2022_Lee[18] (*n*=165)	2021_Andrews[59] (*n*=46)	2018_Matson[37] (*n*=39)	2017_Frankel[60] (*n*=39)
Country	Australia	Australia	USA	UK, Netherlands, Spain	USA	USA	USA
Cancer type (%)	UGB (36%), GYN (34%), NEN (30%)	MEL (100%)	MEL (100%)	MEL (100%)	MEL (100%)	MEL (100%)	MEL (100%)
ICB regimen (%)	CICB (100%)	CICB (100%)	Anti-PD-1 (100%)	CICB (33%), anti-PD-1 (61%), anti-CTLA-4 (7%)	CICB (100%)	Anti-PD-1 (100%)	CICB (62%), anti-PD-1 (36%), anti-CTLA-4 (3%)
Response criteria	RECIST 1.1	Pathological (INMC)	RECIST 1.1	RECIST 1.1	RECIST 1.1	RECIST 1.1	RECIST 1.1
Response							
CR	3.8%	PathR: 84%	5.4%	13%	11%	5.1%	13%
PR	21%		51%	26%	52%	33%	36%
SD	27%		22%	17%	11%	31%	13%
PD	28%	Non-pathR: 16%	22%	42%	26%	31%	39%
cPD	20%		0%	1.8%	0%	0%	0%
Stool collection kit	OMR-200	EasySampler	EasySampler	LO—TF kits, MA—plain tube, NL—plain tube, LD—OMR-200, BL—OMR-200	OMR-200	EasySampler	NR
DNA extraction kit	FastDNA soil	FastDNA feces	PowerSoil	LO—TF MagMAX, MA—TF MagMAX, LD—TF MagMAX, NL—TF MagMAX, BL—PowerFecal	PowerSoil	PowerFecal	Other
Sequencer (bases per read)	NovaSeq (2×151)	NovaSeq (2×151)	NovaSeq (2×151)	NovaSeq (2×151)	NextSeq (2×151)	NextSeq (2×151)	HiSeq (2×100)
Clean PE reads (millions)							
Minimum	9.10	5.05	2.72	4.38	12.8	19.3	18.1
Median	20.4	22.5	30.5	20.7	40.0	35.6	45.4
Maximum	53.8	34.5	72.2	104	69.8	77.7	59.9

The clinical and technical characteristics of the studies included in the meta-analysis are summarized. Published studies are denoted by 'year_author'. USA, United States of America; UK, United Kingdom; MEL, melanoma; INMC, International Neoadjuvant Melanoma Consortium; LO, London; MA, Manchester; NL, Netherlands; LD, Leeds; BL, Barcelona; TF, Thermo Fisher Scientific; PathR, pathologic response; NR, not reported; PE, paired-end. [a]Of the original 'Pittsburgh early cohort' (*n*=63), *n*=37 had their analyzed stool sample collected between day −15 and day 15 of starting ICB and were therefore deemed eligible.

cohorts. Thus, we believe that this work makes a strong case for distinct microbial consortia underpinning response or nonresponse to each regimen. This is biologically plausible, given that we know that CICB has a distinct mechanism of action compared to anti-PD-1 monotherapy[48] and distinct baseline tumor immune microenvironment signatures[49]. Furthermore, the addition of anti-CTLA-4 has a profound effect on gut barrier permeability[50,51], potentially changing the influence of the gut microbiome on ICB response. Nevertheless, the poor generalizability of the CA209-538 strain–RvsP signature to anti-PD-1 cohorts is still intriguing, given the similarity in key positive strains and those species or genera previously associated with response. For example, *Faecalibacterium* has been linked to the efficacy of anti-PD-1 monotherapy in patients with melanoma[16] or hepatobiliary[15] cancers, and *B. longum* has been linked to anti-PD-1 efficacy in patients with melanoma[37] and NSCLC[52]. Therefore, we postulate that the distinction may lie in the negative taxa, with many of the top negative strains in our signature being members of the Lachnospiraceae family (previously broadly associated with anti-PD-1 response in melanoma cohorts[19]). This is also conceptually consistent with the observation of more discrepancies in the pretreatment tumor immunotranscriptomic landscape of anti-PD-1 and CICB nonresponders compared to responders[49].

This work has several limitations that should be addressed in the future. First, despite our relatively large discovery cohort and meta-analysis, the individual cohort and total sample sizes are still small, limiting the statistical power of our signature. Future meta-analyses will benefit from larger, more geographically diverse cohorts, ideally with standardized, best-practice approaches to fecal collection and DNA extraction methods[45]. Moreover, although we used a state-of-the-art bioinformatics pipeline to generate and quality control MAGs to represent study-specific strains (many of which are new or uncultivated), they still potentially harbor errors (such as fragmentation, assembly breaks and contamination)[53]. Although not possible due to the collection medium used in this study, our group has previously demonstrated large-scale fecal strain-culturing methods[54], which, when coupled with whole-genome sequencing, have allowed us to build comprehensive, context-specific genome reference libraries that improve the accuracy of reference-based metagenomic taxonomic classification[55]. Finally, such patient-specific culturing is necessary to perform in vitro and in vivo testing of microbial strains or consortia to derive precise mechanistic insights into their associations with response or nonresponse to ICB and to confirm the direction of causality.

Until then, we believe that this work provides a number of readily implementable insights to help future research and development in this

field. First, it highlights the added value of strain resolution in developing gut microbial ICB biomarkers. There is now ample evidence that intraspecies variation of gut microbiota can substantially change their effect on hosts, first described for enteric pathogens (for example, *Escherichia coli*[56]) but more recently demonstrated for immunomodulatory commensals[57,58], providing further conceptual support for this notion. Second, it suggests that strain signatures may be generalizable across cancer types and geographic locations, supporting investment in developing 'pan-cancer' gut microbial diagnostics and/or therapeutic ICB adjuncts. Lastly, the distinct performance of CICB and anti-PD-1 gut microbial signatures suggests that we should disaggregate these regimens in future analyses to define the relationships between gut microbiota and ICB more precisely in a regimen-specific fashion and, eventually, to use this information in personalizing the care of cancer patients.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-02823-z.

References

1. Curran, M. A., Montalvo, W., Yagita, H. & Allison, J. P. PD-1 and CTLA-4 combination blockade expands infiltrating T cells and reduces regulatory T and myeloid cells within B16 melanoma tumors. *Proc. Natl Acad. Sci. USA* **107**, 4275–4280 (2010).
2. Wolchok, J. D. et al. Overall survival with combined nivolumab and ipilimumab in advanced melanoma. *N. Engl. J. Med.* **377**, 1345–1356 (2017).
3. Motzer, R. J. et al. Nivolumab plus ipilimumab versus sunitinib in advanced renal-cell carcinoma. *N. Engl. J. Med.* **378**, 1277–1290 (2018).
4. Hellmann, M. D. et al. Nivolumab plus ipilimumab in advanced non-small-cell lung cancer. *N. Engl. J. Med.* **381**, 2020–2031 (2019).
5. Baas, P. et al. First-line nivolumab plus ipilimumab in unresectable malignant pleural mesothelioma (CheckMate 743): a multicentre, randomised, open-label, phase 3 trial. *Lancet* **397**, 375–386 (2021).
6. Yau, T. et al. Efficacy and safety of nivolumab plus ipilimumab in patients with advanced hepatocellular carcinoma previously treated with sorafenib: the CheckMate 040 randomized clinical trial. *JAMA Oncol.* **6**, e204564 (2020).
7. Morad, G., Helmink, B. A., Sharma, P. & Wargo, J. A. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell* **184**, 5309–5337 (2021).
8. de Ávila Machado, M. A. et al. Real-world analyses of therapy discontinuation of checkpoint inhibitors in metastatic melanoma patients. *Sci. Rep.* **10**, 14607 (2020).
9. Wang, Y. et al. FDA-approved and emerging next generation predictive biomarkers for immune checkpoint inhibitors in cancer patients. *Front. Oncol.* **11**, 683419 (2021).
10. Routy, B. et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
11. Zheng, Y. et al. Gut microbiome affects the response to anti-PD-1 immunotherapy in patients with hepatocellular carcinoma. *J. Immunother. Cancer* **7**, 193 (2019).
12. Derosa, L. et al. Gut bacteria composition drives primary resistance to cancer immunotherapy in renal cell carcinoma patients. *Eur. Urol.* **78**, 195–206 (2020).
13. Derosa, L. et al. Intestinal *Akkermansia muciniphila* predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer. *Nat. Med.* **28**, 315–324 (2022).
14. Peters, B. A. et al. Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome Med.* **11**, 61 (2019).
15. Mao, J. et al. Gut microbiome is associated with the clinical response to anti-PD-1 based immunotherapy in hepatobiliary cancers. *J. Immunother. Cancer* **9**, e003334 (2021).
16. Spencer, C. N. et al. Dietary fiber and probiotics influence the gut microbiome and melanoma immunotherapy response. *Science* https://doi.org/10.1126/science.aaz7015 (2021).
17. Limeta, A., Ji, B., Levin, M., Gatto, F. & Nielsen, J. Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma. *JCI Insight* **5**, e140940 (2020).
18. Lee, K. A. et al. Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat. Med.* **28**, 535–544 (2022).
19. McCulloch, J. A. et al. Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. *Nat. Med.* **28**, 545–556 (2022).
20. Thomas, A. M. et al. Gut OncoMicrobiome Signatures (GOMS) as next-generation biomarkers for cancer immunotherapy. *Nat. Rev. Clin. Oncol.* **20**, 583–603 (2023).
21. Karcher, N. et al. Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol.* **22**, 209 (2021).
22. De Filippis, F., Pasolli, E. & Ercolini, D. Newly explored *Faecalibacterium* diversity is connected to age, lifestyle, geography, and disease. *Curr. Biol.* **30**, 4932–4943 (2020).
23. Klein, O. et al. Evaluation of combination nivolumab and ipilimumab immunotherapy in patients with advanced biliary tract cancers: subgroup analysis of a phase 2 nonrandomized clinical trial. *JAMA Oncol.* **6**, 1405–1409 (2020).
24. Klein, O. et al. Immunotherapy of ipilimumab and nivolumab in patients with advanced neuroendocrine tumors: a subgroup analysis of the CA209-538 clinical trial for rare cancers. *Clin. Cancer Res.* **26**, 4454–4459 (2020).
25. Klein, O. et al. Combination immunotherapy with nivolumab and ipilimumab in patients with rare gynecological malignancies: results of the CA209-538 clinical trial. *J. Immunother. Cancer* **9**, e003156 (2021).
26. Klein, O. et al. Combination immunotherapy with ipilimumab and nivolumab in patients with advanced adrenocortical carcinoma: a subgroup analysis of CA209-538. *Oncoimmunology* **10**, 1908771 (2021).
27. Gupta, D. & Lis, C. G. Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature. *Nutr. J.* **9**, 69 (2010).
28. Cupp, M. A. et al. Neutrophil to lymphocyte ratio and cancer prognosis: an umbrella review of systematic reviews and meta-analyses of observational studies. *BMC Med.* **18**, 360 (2020).
29. Peled, J. U. et al. Microbiota as predictor of mortality in allogeneic hematopoietic-cell transplantation. *N. Engl. J. Med.* **382**, 822–834 (2020).
30. Sims, T. T. et al. Gut microbiome diversity is an independent predictor of survival in cervical cancer patients receiving chemoradiation. *Commun. Biol.* **4**, 237 (2021).
31. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
32. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

33. Duncan, S. H., Hold, G. L., Harmsen, H. J. M., Stewart, C. S. & Flint, H. J. Growth requirements and fermentation products of *Fusobacterium prausnitzii,* and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* **52**, 2141–2146 (2002).

34. Miquel, S. et al. *Faecalibacterium prausnitzii* and human intestinal health. *Curr. Opin. Microbiol.* **16**, 255–261 (2013).

35. Yao, S., Zhao, Z., Wang, W. & Liu, X. *Bifidobacterium longum*: protection against inflammatory bowel disease. *J. Immunol. Res.* **2021**, 8030297 (2021).

36. Barratt, M. J. et al. *Bifidobacterium infantis* treatment promotes weight gain in Bangladeshi infants with severe acute malnutrition. *Sci. Transl. Med.* **14**, eabk1107 (2022).

37. Matson, V. et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).

38. Ventura, M. et al. The *Bifidobacterium dentium* Bd1 genome sequence reflects its genetic adaptation to the human oral cavity. *PLoS Genet.* **5**, e1000785 (2009).

39. Bachem, A. et al. Microbiota-derived short-chain fatty acids promote the memory potential of antigen-activated CD8$^+$ T cells. *Immunity* **51**, 285–297 (2019).

40. He, Y. et al. Gut microbial metabolites facilitate anticancer therapy efficacy by modulating cytotoxic CD8$^+$ T cell immunity. *Cell Metab.* **33**, 988–1000 (2021).

41. Zhang, S.-L. et al. Pectin supplement significantly enhanced the anti-PD-1 efficacy in tumor-bearing mice humanized with gut microbiota from patients with colorectal cancer. *Theranostics* **11**, 4155–4170 (2021).

42. Nomura, M. et al. Association of short-chain fatty acids in the gut microbiome with clinical response to treatment with nivolumab or pembrolizumab in patients with solid cancer tumors. *JAMA Netw. Open* **3**, e202895 (2020).

43. Simpson, R. C. et al. Diet-driven microbial ecology underpins associations between cancer immunotherapy outcomes and the gut microbiome. *Nat. Med.* **28**, 2344–2352 (2022).

44. Kircher, B. et al. Predicting butyrate- and propionate-forming bacteria of gut microbiota from sequencing data. *Gut Microbes* **14**, 2149019 (2022).

45. Costea, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).

46. He, Y. et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).

47. Gihawi, A. et al. Major data analysis errors invalidate cancer microbiome findings. *mBio* **14**, e0160723 (2023).

48. Wei, S. C. et al. Combination anti-CTLA-4 plus anti-PD-1 checkpoint blockade utilizes cellular mechanisms partially distinct from monotherapies. *Proc. Natl Acad. Sci. USA* **116**, 22699–22709 (2019).

49. Gide, T. N. et al. Distinct immune cell populations define response to anti-PD-1 monotherapy and anti-PD-1/anti-CTLA-4 combined therapy. *Cancer Cell* **35**, 238–255 (2019).

50. Coutzac, C. et al. Systemic short chain fatty acids limit antitumor effect of CTLA-4 blockade in hosts with cancer. *Nat. Commun.* **11**, 2168 (2020).

51. Mager, L. F. et al. Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy. *Science* **369**, 1481–1489 (2020).

52. Jin, Y. et al. The diversity of gut microbiome is associated with favorable responses to anti-programmed death 1 immunotherapy in Chinese patients with NSCLC. *J. Thorac. Oncol.* **14**, 1378–1389 (2019).

53. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).

54. Browne, H. P. et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).

55. Forster, S. C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).

56. Figler, H. M. & Dudley, E. G. The interplay of *Escherichia coli* O157:H7 and commensal *E. coli*: the importance of strain-level identification. *Expert Rev. Gastroenterol. Hepatol.* **10**, 415–417 (2016).

57. Yang, C. et al. Fecal IgA levels are determined by strain-level differences in *Bacteroides ovatus* and are modifiable by gut microbiota manipulation. *Cell Host Microbe* **27**, 467–475.e6 (2020).

58. Geva-Zatorsky, N. et al. Mining the human gut microbiota for immunomodulatory organisms. *Cell* **168**, 928–943 (2017).

59. Andrews, M. C. et al. Gut microbiota signatures are associated with toxicity to combined CTLA-4 and PD-1 blockade. *Nat. Med.* **27**, 1432–1441 (2021).

60. Frankel, A. E. et al. Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients. *Neoplasia* **19**, 848–855 (2017).

¹Host–Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, UK. ²Experimental Cancer Genetics, Wellcome Sanger Institute, Hinxton, UK. ³Olivia Newton-John Cancer Research Institute, La Trobe University School of Cancer Medicine, Melbourne, Victoria, Australia. ⁴Department of Medical Oncology, Austin Health, Melbourne, Victoria, Australia. ⁵Central Clinical School, Monash University, Melbourne, Victoria, Australia. ⁶European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. ⁷Center for Experimental and Molecular Medicine, Amsterdam UMC, Amsterdam, Netherlands. ⁸Department of Medical Oncology, Monash Health, Melbourne, Victoria, Australia. ⁹Department of Medical Oncology, Alfred Health, Melbourne, Victoria, Australia. ¹⁰School of Clinical Sciences, Monash University, Melbourne, Victoria, Australia. ¹¹Department of Medical Oncology, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. ¹²Rare Cancer Laboratory, Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria, Australia. ¹³Department of Medical Oncology, Blacktown and Westmead Hospitals, Sydney, New South Wales, Australia. ¹⁴Melanoma Institute of Australia, University of Sydney, Sydney, New South Wales, Australia. ¹⁵Border Medical Oncology and Haematology Research Unit,

Albury–Wodonga Regional Cancer Centre, Albury–Wodonga, New South Wales, Australia. [16]Rural Medical School, University of New South Wales, Albury, New South Wales, Australia. [17]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. [18]These authors jointly supervised this work: David J. Adams, Trevor D. Lawley. ✉e-mail: ag35@sanger.ac.uk; tl2@sanger.ac.uk

Methods

CA209-538: clinical trial procedures

CA209-538, titled 'A phase 2 trial of ipilimumab and nivolumab for the treatment of rare cancers', is an investigator-initiated, prospective, multicenter, single-arm clinical trial (NCT02923934). The study was approved by the Austin Health (Melbourne, Australia) Human Research Ethics Committee (approval: HREC/16/Austin/152).

Between October 2017 and February 2020, 120 adult patients with rare cancers were recruited across five sites in southeastern Australia (Austin Health, Peter MacCallum Cancer Centre, Monash Health, Blacktown Hospital and Albury Wodonga Health/Border Medical Oncology). Patients were recruited into three prespecified 'histology cohorts' of approximately equal sizes: (1) UGB, comprising cholangiocarcinomas, gallbladder cancers, duodenal cancers and gastrointestinal stromal tumors; (2) NEN, including neuroendocrine tumors or carcinoma of any primary organ (except small cell lung carcinoma) or adrenocortical carcinoma; and (3) GYN, comprising diverse histologies including carcinosarcoma, low-grade serous carcinoma and clear-cell carcinoma of gynecological organs.

Patients were eligible if they had a histologically confirmed diagnosis of a target rare cancer (UGB, NEN or GYN cancers) that was advanced or metastatic, an Eastern Cooperative Oncology Group (ECOG) performance status of 0–1, a measurable tumor lesion per RECIST 1.1 criteria[61] and screening blood laboratory values largely within normal limits. Prior systemic therapy or radiotherapy was permitted if completed at least 4 or 2 weeks, respectively, of the first administration of the study drugs and all related adverse events had stabilized or returned to baseline. The exclusion criteria included active central nervous system metastases (brain or leptomeningeal); prior CICB (monotherapy was permitted); prior malignancy active in the previous 3 years; active, known or suspected autoimmune conditions; and requirement for systemic corticosteroids >10 mg prednisolone daily or equivalent. Participants provided fully informed written consent, including for the collection and analysis of biospecimens (including fecal samples) and sharing of anonymized data as part of research collaborations. The data cutoff was May 7, 2022, providing a minimum of 26 months of follow-up for all participants.

All patients were intended to be treated with CICB in the form of nivolumab 3 mg kg^{-1} and ipilimumab 1 mg kg^{-1} three weekly for four doses (induction), followed by nivolumab monotherapy maintenance (3 mg kg^{-1} two weekly or 480 mg four weekly after a protocol amendment) for up to 2 years or until PD or unacceptable toxicity. The trial's prespecified primary endpoint was to determine the clinical efficacy of CICB in patients with rare cancers using the RECIST 1.1 BOR[61]. In brief, BOR was determined at data cutoff and defined as the investigator-assessed RECIST 1.1 best response designation at any on-trial time point until the date of objectively determined progression per RECIST 1.1 or the date of subsequent anticancer therapy commencement. For participants without documented progression or subsequent therapy, all available response designations contributed to their BOR assessment. The trial's minimum duration criterion for the determination of SD was 9 weeks.

For the assessment of radiographic response, all patients were intended to undergo whole-body cross-sectional imaging with computed tomography or magnetic resonance imaging at baseline (within 28 days before registration), 12 weeks, 18 weeks and then 12 weekly thereafter (±1 week). Patients with rapid disease-related clinical deterioration who were thus unable to undergo restaging imaging at the first restaging time point were deemed to have cPD. PFS and OS were determined from the date of first treatment; the efficacy and safety outcomes for various trial subcohorts have been reported previously[23–26]. Given the accumulating evidence of 'pseudoprogression' in a minority of ICB recipients[62], under the trial protocol, ICB therapy could extend beyond RECIST 1.1-defined PD if there was investigator-assessed clinical benefit and good participant tolerance of the study drugs until there was evidence of a further 10% or greater increase in target lesion dimensions or further new disease sites.

Other clinical metadata. Detailed information on tumor characteristics, demographic factors, blood laboratory values and concomitant medications was collected by the site investigators into an electronic case report form. For this analysis, we included the following 15 clinical metadata variables, as we hypothesized their potential relevance to treatment response and/or gut microbial compositions based on our literature review: patient age (years, at time of trial commencement), sex, body mass index, ECOG performance status, histology cohort (based on the pathology report), extent of measurable tumor (based on the sum of RECIST target lesion diameters calculated using the computed tomography scan at trial screening), study site, season of fecal sample collection, antibiotic use, proton-pump inhibitor use, chemotherapy use, blood NLR, platelet count, albumin levels and lactate dehydrogenase levels (Supplementary Table 3). Only one participant had received prior ICB monotherapy (a NEN cohort patient treated with anti-PD-1 therapy ceased 20 months before trial treatment); given that only one patient was involved, this was not included as a clinical variable. Antibiotic, proton-pump inhibitor and chemotherapy use was defined as their recorded use within the 8 weeks before cycle 1 of study treatment, given the evidence of antibiotic perturbations of gut microbial compositions lasting this duration[63]. The different antibiotics used were amoxicillin, amoxicillin plus clavulanic acid, ampicillin, azithromycin, cefalexin, cefazolin, ceftriaxone, clindamycin, co-trimoxazole, doxycycline, flucloxacillin, gentamicin, norfloxacin, penicillin, piperacillin plus tazobactam and metronidazole. As only 9 of the 106 microbiome-evaluable patients had used any antibiotics in this 8-week period, they were not further subcategorized based on class or antimicrobial coverage. The different proton-pump inhibitors used were esomeprazole, pantoprazole, rabeprazole and omeprazole.

Fecal sample collection. The collection of fecal samples was added to the study protocol in version 5 (July 24, 2017). Participants were trained and provided OMR-200 'OMNigene GUT kits' (DNA Genotek) to collect a fecal sample immediately before treatment (from day −7 to day 0 relative to cycle 1 of trial treatment). OMR-200 kits are designed to stabilize DNA and have been shown to enhance DNA quantifications and stability across storage temperatures relative to nonpreservative alternatives[64]. Fecal samples were express-shipped to the Olivia Newton-John Cancer Research Institute, where they were then frozen at −80 °C for long-term storage. DNA was extracted using the FastDNA kit (MP Biomedicals), including a negative control using ultrapure water. DNA samples were shipped to the Wellcome Sanger Institute on dry ice for shotgun metagenomic sequencing.

Fecal shotgun metagenomic sequencing and analysis

DNA sequencing and quality control. DNA samples were quantified using a Qubit fluorometer, and whole metagenome libraries were deeply sequenced on a single run of the NovaSeq 6000 S4 platform (2 × 150-bp paired-end reads), generating a median of 20,477,028 raw paired-end reads per sample (interquartile range 19,244,530–22,056,539 paired-end reads). Raw sequencing data were first human decontaminated by the Wellcome Sanger Institute core sequencing team by removing read pairs in which one or both aligned to the GRCh37 human genome assembly using bwa (v0.7.17; 'aln' then 'sampe' commands)[65]. These data were further quality controlled using the metaWRAP (v1.2)[66] 'reads_qc' pipeline, which first trimmed low-quality bases using trim-galore (v0.6.7)[67] (default parameters) and then performed a second pass of human decontamination with BMTagger (v3.101)[68] using the GRCh38 human genome assembly. Finally, a median of 20,359,318 clean paired-end reads per sample (interquartile range 19,014,843–21,771,873) were available for further analysis.

MAG assembly. Quality-controlled paired-end reads were first assembled individually with SPAdes (v3.14) using option '-meta' (refs. [69,70]). Unassembled reads were then recovered by mapping raw reads back to metaSPAdes-assembled contigs using bwa 'mem' (v0.7.17)[65], followed by reassembly with MEGAHIT (v1.2.4)[71] using default parameters. Subsequently, the sample-wise metaSPAdes and MEGAHIT assemblies were combined and sorted, with short contigs (<1,500 bp) removed. The resulting assemblies were then independently binned with MetaBAT 2 (v2.13)[72], MaxBin2 (v2.2.4)[73] and CONCOCT (v0.4)[74] using default parameters and a minimum contig length threshold of 1,500 bp (option '--minContig 1500'). The depth of contig coverage required for the binning was inferred by mapping the raw reads back to their assemblies with bwa-mem and then calculating the corresponding read depths for each contig with samtools (v1.5)[75] ('samtools view -Sbu' followed by 'samtools sort'), together with the 'jgi_summarize_bam_contig_depths' function in MetaBAT 2.

Thereafter, individual bin sets produced by the three binning programs were consolidated into a refined bin set consisting of the best version of each bin based on the most optimal genome completion and contamination metrics among all seven versions of hybridized bin sets (MetaBAT 2, MaxBin2, CONCOCT, MetaBAT 2 + MaxBin2, MetaBAT 2 + CONCOCT, MaxBin2 + CONCOCT, MetaBAT 2 + MaxBin2 + CONCOCT), as estimated by CheckM (v1.1.2)[76] using the metaWRAP (v1.2) 'bin_refinement' pipeline[66]. Finally, the final bin sets were further improved by performing reassembly with SPAdes in '--careful' mode after both strict and permissive mapping of raw reads and keeping the bin sets with the best CheckM metrics. In total, 4,277 MAGs with ≥50% completion and ≤5% contamination were generated. These were then further quality controlled, now for ≥90% completeness and ≤5% contamination using CheckM2 (v0.1.3)[77] and for strain-level contamination using GUNC (v1.0.5)[78] to finally identify 2,209 quality-controlled nc-MAGs consistent with the MIMAG (minimum information about a MAG) criteria[79]. Finally, study-specific MAGs were taxonomically classified (using GTDB r207 taxonomy) with GTDB-tk (v2.1)[80], pplacer (v1.1)[81] and fastANI (v1.3)[82].

Generation of a custom, MAG-informed reference database. As the recovery of MAGs may be challenging for some (for example, low abundance or difficult to assemble) strains, we sought to supplement our study-specific strain genome reference database with SRGs from GTDB r207 (62,291 bacterial and 3,412 archaeal genomes) to create a 'hybrid' reference library. To identify a relevant shortlist of GTDB SRGs, we first mapped quality-controlled reads from our study to the full GTDB r207 SRG database with Bowtie 2 (v2.3.5)[83] and inStrain (v1.3.0)[84] (using default settings in '--database' mode). After further filtering of reads mapped to <0.5 SRG breadth, we determined that n = 1,076 SRGs were present. We combined these SRGs with the study-specific nc-MAGs (total 3,285) and used dRep (v2.0.0)[85] to dereplicate the combined genome set to 98% identity using the settings '-comp 90 --con 5 --S_algorithm fastANI --S_ani 0.98 --cov_thresh 0.50 --multiround_primary_clustering --greedy_secondary_clustering'. An absolute nucleotide identity (ANI) threshold of 98% was chosen as a compromise between offering subspecies (strain-level) resolution for read classification while still mitigating 'read stealing' due to overly similar reference genomes (as detailed in the inStrain documentation). Ultimately, n = 1,397 genomes were selected using dRep and formed our 'hybrid' custom strain reference database. Of these, just over half were study-specific nc-MAGs (714, 51%), whereas the remainder were either near-complete isolate (423, 30%) genomes or nc-MAG (260, 19%) SRGs. Using GTDB-tk, we could classify 1,363 of the 1,397 genomes to 904 separate GTDB r207 species clusters (898 bacteria, 6 archaea), with the remaining 34 (32 bacteria, 2 archaea) representing completely new species. For the 904 'known' species, 705 species had 1 strain, whereas 199 species had 2–21 strains each. The species with n = 21 distinct strains by 98% ANI delimitation was *Ruminococcus D bicirculans* (Supplementary Table 16).

Read mapping to a custom strain database. We first used Bowtie 2 to generate a mapping index and then to align reads to our custom reference database. We then used the inStrain profile, now with settings '--min_read_ani 0.95 --min_genome_coverage 1', to perform more precise quality control of the mapped reads. InStrain uses information on paired-end read orientation, mapQ score, insert size and ANI value to filter read mappings stringently, resulting in high-confidence quantifications.

To enhance our confidence about read mappings further, we removed reads mapped with <0.5 genome breadth coverage, as low genome breadth might indicate mapping to mobile genetic elements or mismapping. For our discovery cohort, a median of 50% (range 39–73%) of quality-controlled reads were ultimately used for abundance estimation of strains within each sample (Supplementary Fig. 1).

We finally used Decontam (v1.16.0)[86] to screen for potential contaminants. Reassuringly, after the above steps, no bacteria were identified in our negative control sample for the discovery cohort. Based on the 'frequency' method (inverse correlation between the abundance of strains and the DNA concentration of submitted samples), one strain was identified as a potential contaminant in over 10% of samples from our discovery cohort (CA209-538 cohort) and was thus removed (*Pseudomonas E* sp002874965; Supplementary Fig. 4).

Downstream analysis of taxonomic abundances. Most downstream microbiome analyses were performed in the R (v4.1.0) environment, using 'phyloseq' (v1.12.0)[87], 'microbiome' (v1.12.0)[88] and 'vegan' (v2.6.4). Specifically, alpha diversity was computed using the Shannon diversity index on strain relative abundances (each sample's sum abundances transformed to a sum of 1). As we found no association between the Shannon diversity index and clean paired-end reads in our discovery cohort (Pearson $R = 0.068$, $P = 0.49$), we did not perform rarefaction. Beta diversity was calculated using the strain Aitchison distance, a measure of Euclidean distance of CLR-transformed abundances, computed using $\log(a/\text{gma})$, where a is the species relative abundance and gma is the sample geometric mean relative abundance (with a small pseudocount of one-half the minimum nonzero abundance added to all values to account for zeros). As CLR abundances may better account for the inherent compositionality of microbial abundance data[89], CLR-transformed feature abundances were exclusively used for the supervised ML analyses.

Generation and visualization of phylogenetic trees. For whole bacterial kingdom genome sets, approximately maximum-likelihood phylogenetic trees were constructed using GTDB-tk (v2.1.0)[80] (aligning 120 ubiquitous bacterial genes) and FastTree (v2.1.0)[90] using the WAG model (Fig. 3a,b). For the tree of *Faecalibacterium* genomes, pairwise whole-genome ANI distances were computed using FastANI[82] (many-to-many mode), which was converted into a distance matrix and then to a Newick-format tree using rapidNJ (v2.3.3)[91] (Supplementary Fig. 2). Trees were visualized using the R package ggtree (v3.2.1)[92].

Functional annotation. To evaluate the presence of virulence factor genes, we used abricate (v1.0.1)[93] to screen relevant strain genomes against the VFDB (Virulence Factor Database)[94]. To profile strain metabolic potential broadly, we used gapseq (v1.2)[95] using the 'gapseq find' command with default settings. Briefly, this involved performing a homology search of genomes (using TBLASTN (https://doi.org/10.1186/1471-2105-10-421)) for 28,768 reactions from 2,910 metabolic pathways (curated from MetaCyc and manually). Metabolic pathways were deemed present if ≥80% complete (lowered to ≥67% if 'key' reactions were present).

To evaluate butyrate production potential specifically, we used a previously validated multilevel approach involving hidden Markov models (HMMs)[44,96]. Briefly, we used a published database of 1,716

genomes and 19,284 genes to build HMM profiles (using HMMER v3.2.1; http://hmmer.org/) for the six genes encoding the acetyl-CoA butyrate-producing pathway (responsible for butyrate production through carbohydrate degradation). These genes are acetyl-CoA acetyltransferase (*thl*), β-hydroxybutyryl-CoA dehydrogenase (*bhbd*), crotonase (*cro*), butyryl-CoA dehydrogenase (*bcd*), and the alternative terminal enzymes butyryl-CoA:acetate CoA transferase (*but*) and butyrate kinase (*butk*). We then used these models to screen the strain genomes for the presence of these respective genes. As an orthogonal approach, we also mapped cleaned sample paired-end reads to the above genes' sequences using Bowtie 2 and then used inStrain 'quick profile' to count mappings to estimate their sample-wise gene abundance (normalized per million reads) agnostic of source strain. The output is available in Supplementary Tables 13 (strain_top22_acetyl-coa_pwy) and 14 (sample_acetylcoa_pwy).

Supervised ML analysis

Supervised ML analyses were performed in the Python 3 environment using the packages sklearn (v1.1.1)[97], imblearn (v0.9.1)[98] and their dependencies. The supervised ML pipeline involved a preprocessing step before model training and testing, performed separately for each training and testing instance to ensure no data leakage. This involved standard-scaling numerical features (computed using the formula $z = (x − u)/s$, where x is the feature value (for example, the CLR-transformed strain abundances), u is the mean of the fold samples and s is the s.d. of the fold samples) and one-hot encoding categorical features. Subsequently, only before classifier training (but not testing), classes of the target variable (RvsP or PFS12) were balanced with random oversampling with replacement.

We chose to use RF as our classifier, given its, on average, superior performance using microbial feature sets in previous benchmarking studies[31]. RF uses bootstrapped data to create an ensemble of decision trees (each trained on a subset of features), with the ultimate classification based on consensus; thus, it is feature scale invariant and able to capture nonlinearities and is also interpretable using TreeSHAP (described subsequently).

Our hyperparameter tuning procedure involved a random hyperparameter search over a broad array of options with 1,000 separate combinations tested, aiming to maximize the ROC AUC averaged over 20 times repeated fivefold cross-validation (that is, 100 separate models trained and tested (splits), for each 1,000 iterations, for each feature and classifier combination).

ROC AUC is a popular classifier performance metric that evaluates the discriminative performance across all potential decision thresholds, thus allowing for a head-to-head comparison of differently calibrated classifiers[99]. Ultimately, the best hyperparameter combination (based on mean AUC) was selected and referred to as the 'tuned' pipeline. The optimal hyperparameters and AUC scores for all 100 splits for all full feature sets are listed in Supplementary Table 9 (hyperparam_tuning_all).

To evaluate model performance, we used cross-validation (for example, leave-one-histotype-out cross-validation) or completely separate training and test cohorts (for example, training a model using one study cohort and then testing the fitted model on another cohort). Whenever evaluating model performance, training and testing procedures were repeated 100 times, and the resultant predictions were averaged to account for the stochasticity of our RF pipeline. As with hyperparameter tuning, ROC AUC was our metric of choice for gauging model performance.

Feature importances were evaluated with the 'shap' package using the TreeExplainer() function. Based on the foundation of game theory, TreeExplainer computes the influence of each feature (strain abundance) in determining the RF classifier's local (per-sample) prediction. Therefore, we computed global feature importances (cohort-wide average of the absolute TreeExplainer scores) and imputed the importance 'direction' (that is, positive or negative influence on response prediction) by constructing a simple linear model between the feature values and SHAP values. We repeated this procedure 1,000 times to account robustly for the RF pipeline's stochasticity. The global feature importance and s.d. values of all features are listed in Supplementary Table 12 (strain_importance).

Literature review and meta-analysis of relevant published datasets

We sought to identify all published clinical datasets that met the following criteria:

1. Evaluated baseline fecal microbiota from patients with cancer who were about to commence only ICB (anti-PD-1, anti-CTLA-4 or CICB) therapy. 'Baseline' samples were defined as those collected between day −15 and day 15 relative to the start of ICB to ensure that the profile reflected the patient's gut microbial context immediately before treatment and that the gut microbial profile had not been already affected by ICB therapy (for example, anti-CTLA-4 appears to modify gut barrier integrity[51] and thus could feasibly change microbial compositions).
2. Used short-read, paired-end shotgun metagenomic sequencing (to allow us to standardize and maintain stringency in our bioinformatic pipeline and quality control steps).
3. Reported tumor response. To be pragmatic, we accepted radiographic (using RECIST 1.1) or pathological response. However, we excluded studies that reported only PFS12 or where response was binned with SD.

To find all such datasets, we performed a structured PubMed database search combining the following three search strings that used both MeSH (Medical Subject Headings) terms and title and/or abstract keywords:

'neoplasms'[MeSH Major Topic] OR 'cancer'[Title/Abstract] OR 'malignancy'[Title/Abstract] OR 'tumor'[Title/Abstract]

OR

'immune checkpoint inhibitors'[MeSH Terms] OR 'pembrolizumab'[Title/Abstract] OR 'nivolumab'[Title/Abstract] OR 'atezolizumab'[Title/Abstract] OR 'avelumab'[Title/Abstract] OR 'durvalumab'[Title/Abstract] OR 'cemiplimab'[Title/Abstract] OR 'dostarlimab'[Title/Abstract] OR 'ipilimumab'[Title/Abstract] OR 'tremilimumab'[Title/Abstract] OR 'immunotherapy'[Title/Abstract] OR 'immune checkpoint'[Title/Abstract]

OR

'microbiota'[MeSH Terms] OR 'metagenome'[MeSH Terms] OR 'metagenomics'[MeSH Terms] OR 'microbiome'[Title/Abstract] OR 'microbiota'[Title/Abstract]

In total, this search yielded 1,181 records up to December 31, 2022. Titles and abstracts were manually reviewed to identify a total of 28 unique studies meeting eligibility criterion 1. A manual bibliography search yielded a further three studies meeting eligibility criterion 1 (Supplementary Table 17 (lit_review)). Of these, 19 studies used shotgun metagenomics, and 13 studies made these raw data available. Three studies were excluded as the shotgun metagenomic data were single end (Ion Torrent). Finally, of the remaining ten studies, four were excluded as they did not report response, yielding six studies that could be included in our meta-analysis (see Supplementary Fig. 3 for a PRISMA-style flowchart). Metadata for each cohort were curated from the corresponding publication tables or relevant sequencing repositories (for example, Sequence Read Archive, European Nucleotide Archive).

Shotgun metagenomic sequencing data for the six evaluable cohorts were downloaded and analyzed using a uniform bioinformatic procedure (as described earlier), including FASTQ file quality control and human DNA decontamination, and then read mapping to an identical custom strain database (generated from CA209-538 MAGs) using identical settings of Bowtie 2 and inStrain. Despite a wide range in the number of quality-controlled paired-end reads per sample, in general, all were deeply sequenced (Table 2). Subsequent downstream analysis of gut microbial profiles and supervised ML analyses were performed using identical methods to those previously described.

Statistical analysis

Statistical tests are cited in the text. In general, nonparametric statistical tests were preferred (all were two-sided). To determine associations between an ordinal and a numeric variable (for example, BOR versus a numeric metadata variable), we used the Kendall τ test. For associations between a binary and a numeric variable, the Mann–Whitney U (also known as the Wilcoxon rank-sum) test was used. For associations between a nonordinal categorical variable and a numeric variable, the Kruskal–Wallis test was used. The threshold for significance was set as a two-tailed P value of <0.05. Data were processed and visualized using the R packages 'tidyverse' (v2.0.0)[100], 'ggpubr' (v0.6.0), 'survival' (v3.5.5)[101], 'survminer' (v0.4.9) and 'table1' (v1.4.3) and the Python packages 'numpy' (v1.23.3)[102], 'pandas' (v1.4.3) and 'matplotlib' (v3.5.1)[103]. For all boxplots, the center line indicates the median, box limits indicate the upper and lower quartiles, and whiskers indicate 1.5× the interquartile range.

Reporting summary

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

Data availability

All CA209-538 fecal shotgun metagenomic sequencing data (after first-pass human decontamination) have been deposited to the European Nucleotide Archive (study accession no. ERP134027). The 1,397 quality-controlled (near-complete) study-specific genomes used as the custom reference database have been deposited to Zenodo (https://doi.org/10.5281/zenodo.10450122). CA209-538 clinical metadata and strain abundance data necessary to replicate our analyses are provided in the Supplementary Tables. The six publicly available shotgun metagenomics datasets were downloaded using the following accession numbers: EGAS00001006982 (2022_Simpson), PRJEB43119 (2022_Lee), PRJNA762360 (2022_McCulloch), EGAD00001006734 (2021_Andrews), PRJNA399742 (2018_Matson) and PRJNA397906 (2017_Frankel). Permission to access the 2021_Andrews raw sequencing dataset was kindly provided by J. Wargo and The University of Texas M.D. Anderson Cancer Center. Permission to access the 2022_Simpson raw sequencing data was kindly provided by G. Long and the Melanoma Institute of Australia. Associated sample-level clinical metadata for external datasets were collected from their relevant publications, the relevant sequencing repository or an associated GitHub repository.

Code availability

No unique software or computational code was created for this study. The relevant code to replicate our supervised machine learning analyses of CA209-538 data, using the data in Supplementary Table 8 (metadata_and_clr_abundances), is available at https://github.com/agunjur/cancer_microbiome_CICB/.

References

61. Eisenhauer, E. A. et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).

62. Park, H. J. et al. Incidence of pseudoprogression during immune checkpoint inhibitor therapy for solid tumors: a systematic review and meta-analysis. *Radiology* **297**, 87–96 (2020).

63. Palleja, A. et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat. Microbiol.* **3**, 1255–1265 (2018).

64. Maghini, D. G. et al. Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01754-3 (2023).

65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

66. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

67. Krueger, F. et al. FelixKrueger/TrimGalore: a wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. *GitHub* github.com/FelixKrueger/TrimGalore (2023).

68. The Bioconda Team. Package recipe 'bmtagger'—Bioconda documentation. bioconda.github.io/recipes/bmtagger/README.html (2016).

69. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

70. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

71. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

72. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

73. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

74. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

75. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

76. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

77. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

78. Orakov, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).

79. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

80. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

81. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).

82. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

83. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

84. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol*. **39**, 727–736 (2021).

85. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. **11**, 2864–2868 (2017).

86. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).

87. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).

88. Lahti, L. & Shetty, S. microbiome: microbiome analytics. https://doi.org/10.18129/B9.bioc.microbiome (2022).

89. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol*. **8**, 2224 (2017).

90. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

91. Simonsen, M., Mailund, T. & Pedersen, C. N. S. Rapid neighbour-joining. in *Algorithms in Bioinformatics* (eds. Crandall, K. A. & Lagergren, J.) 113–122 (Springer, 2008); https://doi.org/10.1007/978-3-540-87361-7_10

92. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol*. **8**, 28–36 (2017).

93. Seemann, T. & Grüning, B. tseemann/abricate. *GitHub* github.com/tseemann/abricate (2023).

94. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Res*. **44**, D694–D697 (2016).

95. Zimmermann, J., Kaleta, C. & Waschina, S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol*. **22**, 81 (2021).

96. Vital, M., Karch, A. & Pieper, D. H. Colonic butyrate-producing communities in humans: an overview using omics data. *mSystems* **2**, e00130-17 (2017).

97. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. **12**, 2825–2830 (2011).

98. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res*. **18**, 559–563 (2017).

99. Zou, K. H., O'Malley, A. J. & Mauri, L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* **115**, 654–657 (2007).

100. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw*. **4**, 1686 (2019).

101. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model* (Springer Science & Business Media, 2013).

102. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).

103. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng*. **9**, 90–95 (2007).

Acknowledgements

Author contributions

A.G., A.B., O.K., D.J.A. and T.D.L. conceived this study. A.G., Y.S., T.R., A.M. and B.W.H. analyzed the data. O.K. and J.C. were coordinating principal investigators of the CA209-538 trial and obtained its funding. A.G., O.K., J.C., B.M., D.K., M.S.C., C.U., S.F., M.M. and B.G. were clinical investigators of the CA209-538 trial, contributing to patient recruitment, care, clinical data collection and interpretation. A.B. oversaw fecal biospecimen collection, storage and initial processing. J.P. managed clinical data related to the CA209-538 trial and assisted with their interpretation. A.G., D.J.A. and T.D.L. wrote the original manuscript draft. All authors wrote, critically reviewed and approved the final version of the manuscript. D.J.A. and T.D.L. jointly provided overall supervision.

Competing interests

A.G. has received a speaker honorarium from Microbiotica Limited. B.M. has served on advisory boards for Amgen, Bristol Myers Squibb (BMS), Merck, Beigene and AstraZeneca (AZ). M.S.C. has served on advisory boards or as a consultant for Amgen, BMS, Eisai, Ideaya, Merck, Sharp & Dohme (MSD), Nektar, Novartis, Oncosec, Pierre-Fabre, Qbiotics, Regeneron, Roche, Merck, Moderna and Sanofi and received honoraria from BMS, MSD and Novartis. D.K. has served on advisory boards for BMS, MSD and Novartis. C.U. has served in a consulting/advisory role for Merck Serano and AZ and a speakers' bureau role for IQvia and AZ. His institution has received research funding from Akeso Biopharma, Arcus Biosciences, Atridia, BeyondSpring Pharmaceuticals, Boehringer Ingelheim, Decephera and Novotech. S.F. has received financial support from Amgen, MSD and AZ; honoraria for advisory boards from Akesobio, Ambrax and MSD; and institutional sponsorship/trials and research activities from Akesobio, Ambrax, Amgen, Axelia, AZ, Aulos, BeiGene, Cullinan, Daiichi Sankyo, Edison Oncology, Genentech, MSD, Takeda, HaiHe Biopharma, Vivace and WellMarker Bio. D.J.A. is a paid consultant for Ono Therapeutics and Microbiotica Limited and receives research support from AZ, OpenTargets and BMS. T.D.L. is cofounder and chief scientific officer at Microbiotica Limited. All other authors declare no competing interests.

Additional information

Extended Data Fig. 1 | Clinical and gut microbiome characteristics of the CA209-538 clinical trial cohort. a, Kaplan-Meier curve of progression-free survival stratified by histology (UGB n = 38, NEN n = 32, GYN n = 36). Log-rank test p-value for OS duration across groups printed. Chi-squared test p-value shown for proportion of OS12 per group printed. **b**, Kaplan-Meier curve of progression-free survival stratified by histology (UGB n = 38, NEN n = 32, GYN n = 36). Log-rank test p-value for PFS duration across groups printed. Chi-squared test of independence p-value shown for proportion of PFS12 per group printed. **c**, Boxplots of patient baseline blood albumin (g/L) and NLR levels (log-transformed) by BOR category (cPD n = 21, PD n = 30, SD n = 29, PR n = 22, CR n = 4). Boxplot centre line= median; box limits= upper and lower quartiles; whiskers=1.5x interquartile range. Linear model line-of-best-fit for respective variables (albumin and NLR) versus BOR (with shaded 95% confidence interval) superimposed (in grey). Kendall τ and p-value for association between respective variables (albumin and NLR) and BOR printed. Pairwise Mann-Whitney U test p-values for cPD vs other groups summarized (*: p < 0.05, **: p < 0.01, ***: p < 0.001). Exact p-values as follows: Albumin: cPD vs PD p = 0.0076, cPD vs SD p = 0.00079, cPD vs PR p = 0.0039, cPD vs CR p = 0.034; NLR: cPD vs PD p = 0.00093, cPD vs SD p = 0.00042, cPD vs PR p = 0.0075, cPD vs CR p = 0.025. **d**, Proportion of explained variance (R^2) of microbial composition by

each available clinical and technical metadata variable. Calculated using PERMANOVA on inter-sample Aitchison distance (9999 permutations). Metadata variables coloured by category (blood, exposome, patient, technical or tumour). PERMANOVA p-values summarized (*: p < 0.05). Exact p-values available in Supplementary Table 6 ('ca209-538_permanova'). **e**, Analysis of baseline microbial variance by moving PFS cut-off (1-monthly intervals, from 1-24 months). Top panel show microbial variance between groups formed by cut-off (inverse PERMANOVA p-value, 999 permutations) using Aitchison distance. Bottom panel shows proportion of progression-free-survivors at respective threshold (that is the proportion in each group). Dashed line with * indicates p = 0.05 threshold. Exact p-values available in Supplementary Tables 7 ('moving_pfs_permanova'). Acronyms: UGB = upper gastrointestinal & biliary, NEN = neuro-endocrine neoplasms, GYN = gynaecological, PFS = progression-free survival, OS = overall survival, BOR = best overall response, CR = complete response, PR = partial response, SD = stable disease, PD = progressive disease, cPD = clinical progressive disease, chemo = chemotherapy, PPI = proton-pump inhibitor, BMI = body-mass index, LDH = lactate dehydrogenase, NLR = neutrophil:lymphocyte ratio, ECOG = eastern cooperative oncology group, PERMANOVA = permutational multivariate analysis of variance.

a

n=31

P vs non-PFS12

n=7   R vs PFS12   n=5

R (n = 26)   P (n = 51)
PFS12 (n = 24)   non-PFS12 (n = 82)

b PFS12

τ = 0.29, p = 7.1e−15

p=1.4e-11
p=0.5

mean ROC AUC (± 1 sd)

.65
.53 .55 .57 .56 .56 .57 .70 .71

clinical family genus species strain
feature set

c Strain-RvsP predictions vs. actual BOR

τ= 0.64, p< 2.2*10-16

Strain-RvsP predicted probability (%)

cPD PD SD PR CR

d SD patients: survival by strain-RvsP prediction

prediction
above median
below median

Survival proportion

p= 0.17

Number of SD patients at risk

Prediction
above median 14 11 5 0
below median 15 10 2 0

Days

Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Sensitivity analyses of gut microbial strain-efficacy classifiers. a, Comparison of the RvsP and PFS12 binary endpoints. Venn diagrams show the overlap between the 'negative and 'positive' outcome populations (P/non-PFS12 and R/PFS12 respectively). Size of circles (area) in proportion to population size, with set differences labelled. **b**, AUC scores for the best iteration of PFS12 classifiers for each feature-set combination during 20-repeated 5-fold cross-validation (100 folds each): clinical (yellow), microbiome (blue) and combined (green), at different taxonomic resolutions. Mean (circle) and standard deviation (error bars) over the 100 folds. Linear model line-of-best-fit for AUC score and taxonomic rank of microbiome-only feature sets (with shaded 95% confidence interval) superimposed. Kendall τ and p-value for association between AUC score and taxonomic rank of microbiome-only feature sets printed. Mann-Whitney U p-value for comparison of AUCs of specific pairwise feature-sets (depicted by callipers) printed. **c**, Patient's predicted RvsP (using strain-RvsP RF classifiers trained on the full evaluable cohort) vs. actual

BOR outcome (cPD n = 21, PD n = 30, SD n = 29, PR n = 22, CR n = 4). Boxplot centre line= median; box limits= upper and lower quartiles; whiskers= 1.5x interquartile range. Kendall rank correlation τ and p-value for association between predicted RvsP and actual BOR printed. **d**, Kaplan-Meier overall survival curves for those patients with a best overall response (BOR) of stable disease (n = 29), stratified by those with above median (blue) and below median (red) strain-RvsP RF classifier predictions. Bottom panel shows number of patients at risk at each marked interval. P-value by log-rank test printed. Acronyms: P= progressors (RECIST progressive disease (PD) or clinical progressive disease (cPD)), R= responders (RECIST complete response (CR) or partial response (PR)), GYN= gynaecological, NEN= neuro-endocrine neoplasm, UGB= upper gastrointestinal & biliary, ROC= receiver operating characteristic, AUC= area under curve, OS= overall survival, SD= stable disease, RvsP= response versus progression, cPD= clinical progressive disease, PD= progressive disease, SD= stable disease, PR= partial response, CR = complete response.
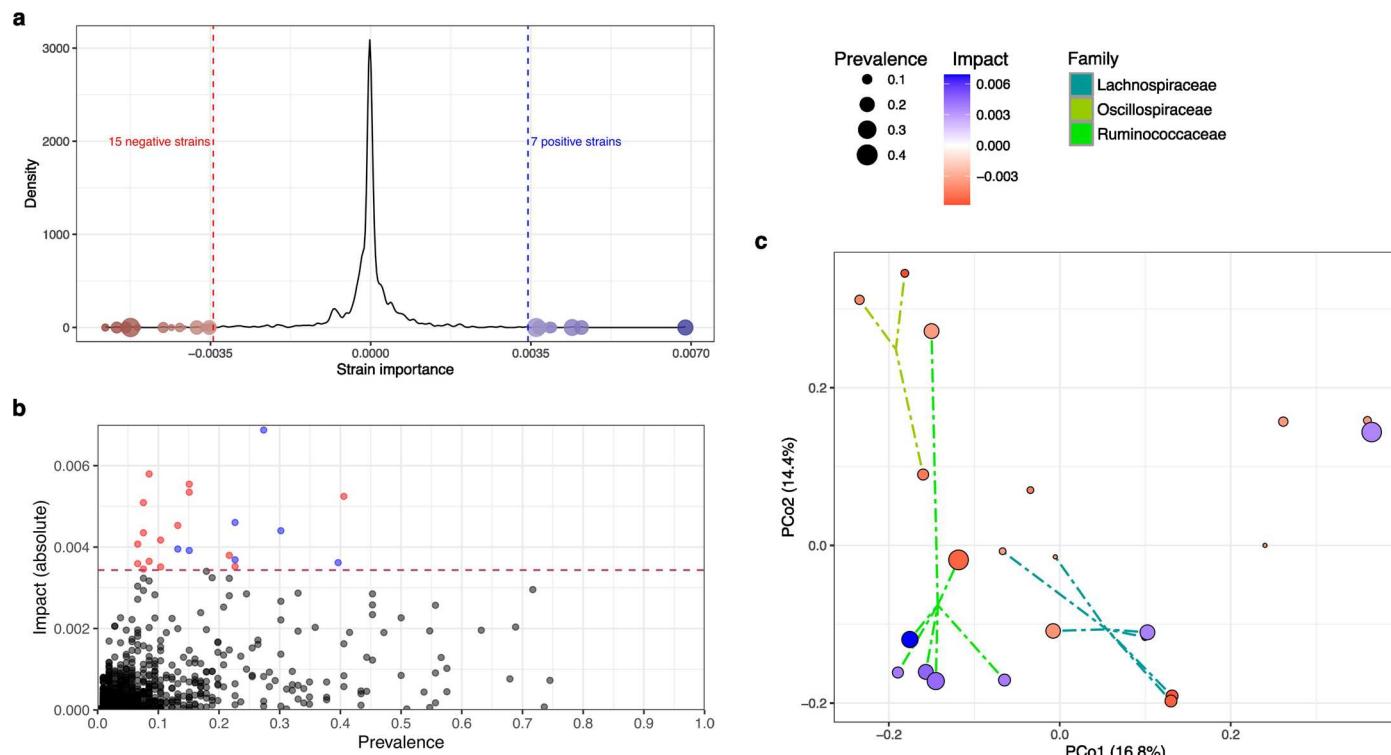
Extended Data Fig. 3 | Identification and metabolic-potential profiling of the top 22 response predictive strains. a, Kernel density plot of impact (feature importance) of strains in the strain-RvsP classifier. The top 22 strains with absolute impact within half maximal value shown (coloured by importance, and size by prevalence). **b**, Strain impact (absolute) versus prevalence in the CA209-538 cohort. Top 22 strains coloured (blue and red for positive and negative associations with response, respectively), with importance threshold depicted (red dashed line). **c**, Plot of principal coordinate 1 vs 2 using Jaccard dissimilarity of metabolic pathway presence/absence for top 22 strain genomes. Points (individual strains) coloured by impact on RvsP, and size by prevalence. Acronyms: PCo= principle coordinate.
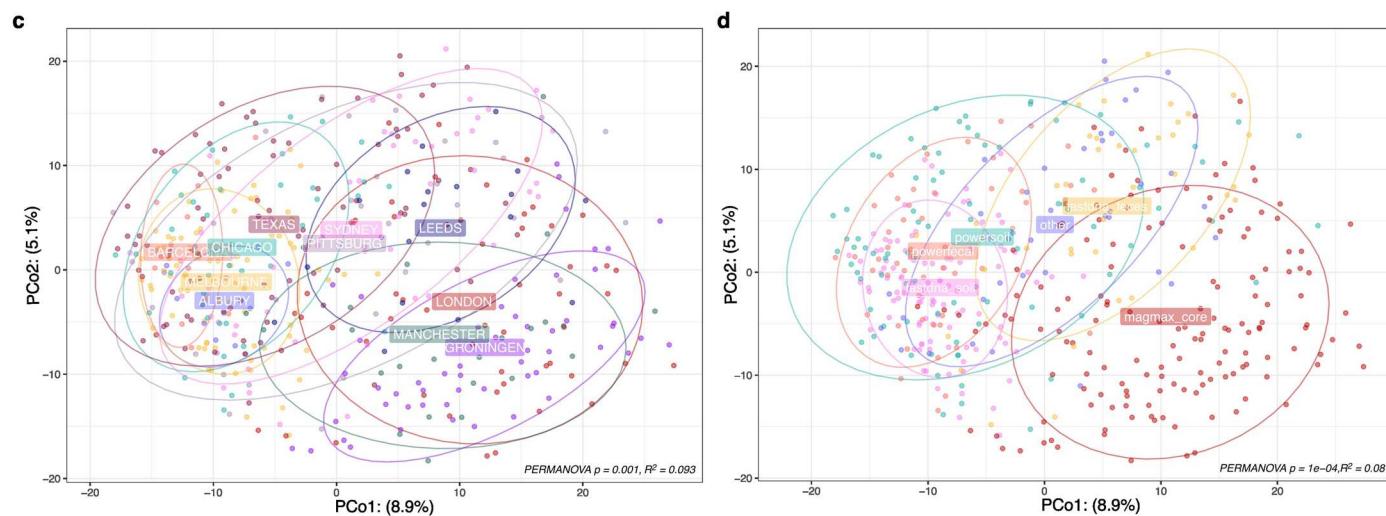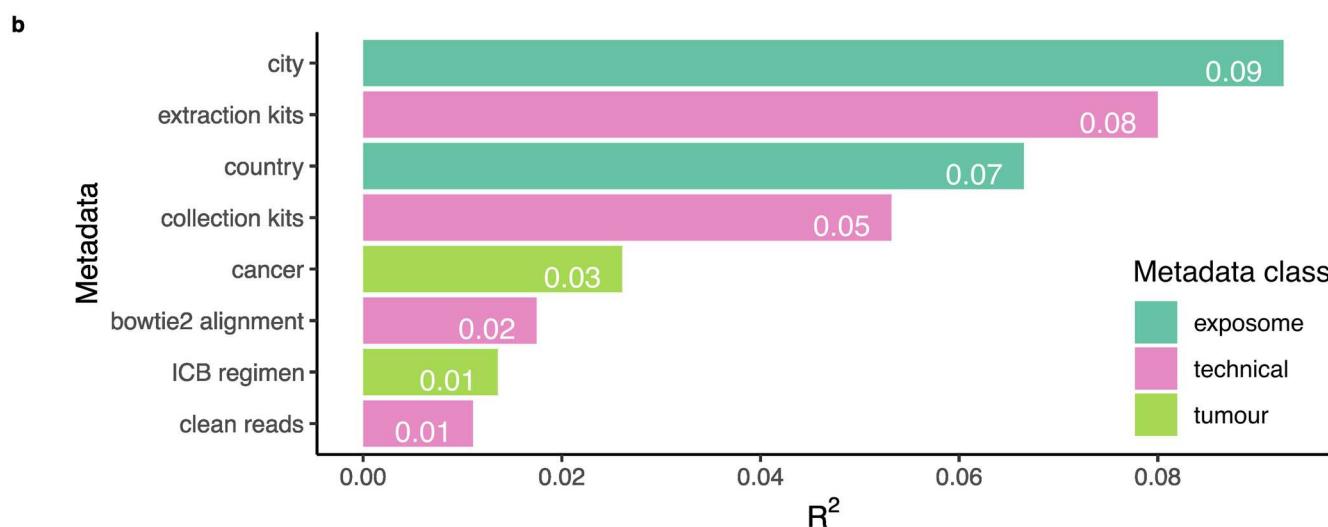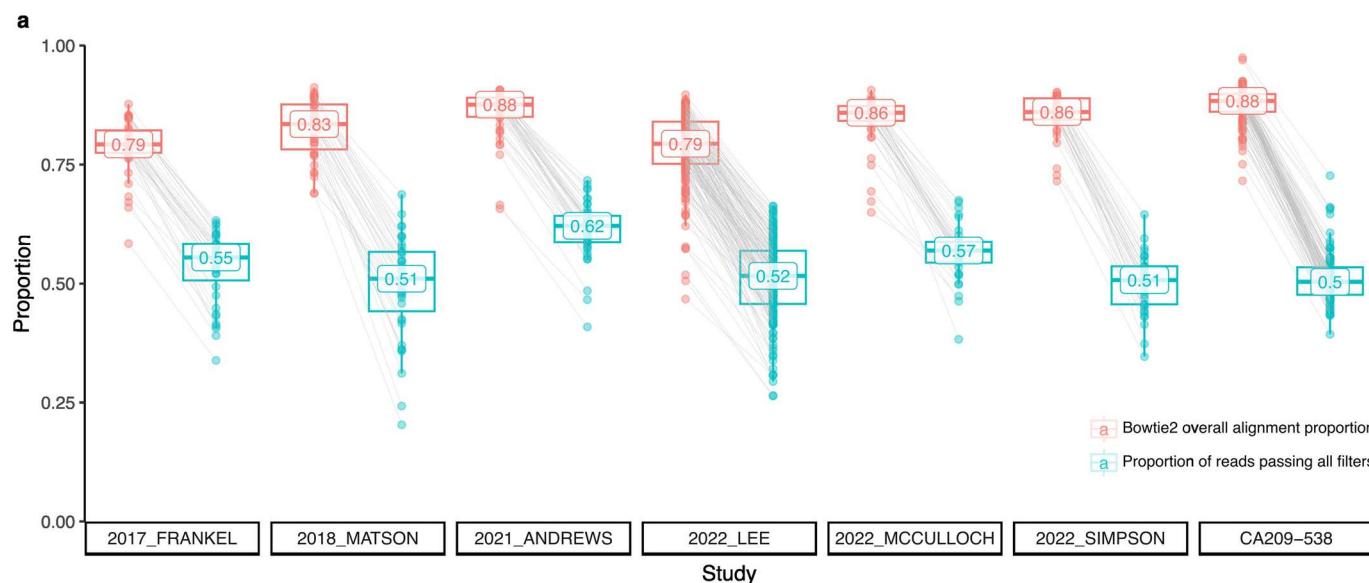
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Heterogeneity of baseline gut microbial compositions across meta-analysis cohorts. a, Proportion of quality-controlled paired-end reads aligned by Bowtie 2 (red), and ultimately used for abundance estimation after stringent filtering (cyan). Organised by study (2017_FRANKEL n = 39, 2018_MATSON n = 39, 2021_ANDREWS n = 46, LEE n = 165, 2022_MCCULLOCH n = 37, 2022_2022_SIMPSON n = 38, CA209-538 n = 106). Boxplot central line= median, box limits= upper and lower quartiles, and whiskers= 1.5x interquartile range. Median printed within each boxplot. **b**, Proportion of explained variance (R^2) of microbial composition by metadata variables (grouped into 'exposome', 'technical' and 'tumour' categories). R^2 values (printed on bar) calculated using PERMANOVA (9999 permutations). **c**, PCA plot of samples by CLR-transformed abundances (Aitchison's distance), with points coloured by sample city (the variable explaining the most variance). Ellipses depict 0.8 of each group's multivariate t-distribution. PERMANOVA p-value and R^2 using 9999 permutations printed. **d**, PCA plot of samples by CLR-transformed abundances (Aitchison's distance), with points coloured by extraction kit (the variable explaining the second-most variance). Ellipses depict 0.8 of each group's multivariate t-distribution. PERMANOVA p-value and R^2 using 9999 permutations printed. Acronyms: ICB= immune checkpoint blockade, PCo= principle coordinate.

nature portfolio

Corresponding author(s):	Dr Ashray Gunjur Dr Trevor D. Lawley
Last updated by author(s):	05 Jan 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specific software was used for data collection.
Data analysis	All software used for data analyses is publicly available, and is cited in the Methods section where possible. These software are: command-line tools: bwa v0.7.17, trim-galore v0.6.7, BMTagger v3.101, SPAdes v3.14, MEGAHIT v1.24, MetaBAT 2 v2.13, MaxBin2 v2.2.4, CONCOCT v0.4, samtools v1.5, CheckM v1.1.2, MetaWRAP v1.2, CheckM2 v0.1.3, GUNC v1.0.5, GTDB-tk v2.1, pplacer v1.1, fastANI v1.3, bowtie2 v2.3.5, inStrain v1.3.0, dRep v2.0.0, Gapseq v1.2, HMMER v3.2.1, FastTree v2.1.0, rapidNJ v2.3.3, abricate v1.0.1. Python 3: scikit-learn v1.1.1, imblearn v0.9.1, shap v0.41.0, numpy v1.23.3, pandas v1.4.3, matplotlib v3.5.1. R v4.1.0: decontam v1.16.0, tidyverse v2.0.0, ggpubr v0.6.0, survival v3.5.5, survminer v0.4.9, table1 v1.4.3, phyloseq v1.12.0, microbiome v1.12.0, vegan v2.6.4, ggtree v3.2.1. Code to replicate our supervised machine learning analysis of CA209-538 cohort data is deposited at https://github.com/agunjur/cancer_microbiome_CICB .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

All manuscripts must include a <u>data availability statement</u>. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our <u>policy</u>

All CA209-538 faecal shotgun metagenomic sequencing data (after first-pass human decontamination) has been deposited to the European Nucleotide Archive (study accession number ERP134027). The 1397 quality-controlled (near-complete) study-specific genomes used as the custom reference database have been deposited to zenodo (https://doi.org/10.5281/zenodo.10450122). CA209-538 clinical metadata and strain abundance data necessary to replicate our analyses is provided as the supplementary tables. The six publicly available shotgun metagenomics datasets were downloaded using the following accession numbers: 2022_SIMPSON: EGAS00001006982, 2022_LEE: PRJEB43119, 2022_MCCULLOCH: PRJNA762360, 2021_ANDREWS: EGAD00001006734, 2018_MATSON: PRJNA399742, 2017_FRANKEL: PRJNA397906. Permission to access the 2021_ANDREWS raw sequencing dataset for academic use was kindly provided by Dr Jennifer Wargo and The University of Texas M.D. Anderson Cancer Center. Permission to access the 2022_SIMPSON raw sequencing data was kindly provided by Professor Georgina Long and the Melanoma Institute of Australia. Associated clinical metadata for external datasets was collected from their relevant publications, the relevant sequencing repository or an associated github repository.

Research involving human participants, their data, or biological material

Reporting on sex and gender	CA209-538 clinical trial participant's self-reported sex was assessed by CA209-538 clinical investigators and recorded into the electronic case report form (eCRF). Sex is reported as a clinical variable in Table 1, Ext Table 1 and Ext Table 2, included as a metadata variable in the CA209-538 PERMANOVA analyses (Ext Fig 1d), and as a clinical feature in the supervised machine learning analyses (Fig 2b, Ext Fig 2b).
Reporting on race, ethnicity, or other socially relevant groupings	Race / ethnicity was not recorded or analysed.
Population characteristics	All CA209-538 participants were adults with advanced rare cancers falling into 3 histological cohorts: upper gastrointestinal / biliary tract (UGB), rare gynaecological (GYN) or neuro-endocrine neoplasms (NEN). All patients were adults (median age (years) 60 [range 20-82] and n=81 (68%) were female sex by self-report. Faecal samples were collected from most patients (n=106 'microbiome evaluable'). Patient-level metadata for microbiome-evaluable patients, including age, sex, body-mass index, ECOG performance status and study site is available in Supplementary table 3. More details on trial inclusion/exclusion criteria are available at https://classic.clinicaltrials.gov/ct2/show/NCT02923934.
Recruitment	CA209-538 participants were screened for eligibility based on protocol inclusion criteria at 5 clinical sites across two states in Australia (3 sites in Victoria: Monash Health, Austin Health, Peter MacCallum Cancer Centre; 2 sites in New South Wales: Blacktown Hospital, Border Medical Oncology Unit). This involved referring medical practitioners sending detailed referrals to site principle investigators, who subsequently reviewed patients to confirm eligibility, willingness to participate, and sign trial informed consent. To aid recruitment, the clinical trial was advertised broadly, including via the 'Cancer Council Victoria: Victorian Cancer Trials Link' (https://trials.cancervic.org.au/details.aspx?ID=vctl_nct02923934). Patient geography and knowledge of the study may have biased study participation.
Ethics oversight	CA209-538 was approved across the 5 clinical sites by the Austin Health Human Research Ethics Committee (reference: HREC/16/Austin/152).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The primary objective of CA209-538 was to evaluate the clinical efficacy (by RECIST 1.1 response criteria) of ipilimumab and nivolumab in rare cancers. At the time of its design there was limited/no available data to estimate response rates of combination anti-PD-1 plus anti-CTLA-4 blockade in patients with these selected rare cancers, with CA209-538 designed to address this gap. Therefore, no statistical sample size or power calculation could be performed a priori.

Data exclusions	No data were excluded intentionally. A minority (n=14) of trial participants were unable to provide a stool specimen immediately prior to commencement of trial therapy. Statistical analyses of microbiome-evaluable (n=106) vs missing (n=14) patients is presented in Ext. table 1. There was a higher proportion of non-evaluable patients from one site (BLA, n=7), but no other suggestions of bias. All n=106 evaluable samples produced high-quality metagenomic sequencing data and were included in our analysis.
Replication	No technical replicates of metagenomic sequencing was performed, however PERMANOVA analysis suggests technical variables such as DNA plate were little contributors to microbial variance (Ext Fig 1d).
Randomization	Not applicable as CA209-538 was designed as a single-arm study to evaluate the efficacy of combination immune checkpoint blockade across rare cancers (representing novel indications), as above.
Blinding	Not applicable as CA209-538 is a single-arm study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	ClinicalTrials.gov Identifier: NCT02923934
Study protocol	The clinical outcomes for CA209-538 histological subgroups have been reported and published previously. Version 8 of the study protocol is included in the Supplementary materials with this submission.
Data collection	CA209-538 participants were recruited between October 2017 and February 2020 across 5 clinical sites in Australia 5 clinical sites across two states in Australia (3 sites in Victoria: Monash Health, Austin Health, Peter MacCallum Cancer Centre; 2 sites in New South Wales: Blacktown Hospital, Border Medical Oncology Unit). Clinical sites were hospital outpatient settings. Site clinical trial investigators recorded de-identified patient information into an eCRF.
Outcomes	The pre-defined primary outcome of CA209-538 was to evaluate the clinical efficacy of ipilimumab and nivolumab in patients with advanced rare cancer types, as determined using RECIST 1.1 'clinical benefit' (complete response + partial response + stable disease). The pre-defined secondary outcome of CA209-538 clinical trial was to identify whether a common predictive biomarker or immune signature can be identified in responding patients that can occur irrespective of tumour type. Samples collected include baseline whole blood, serum, peripheral blood mononuclear cells, archival formalin-fixed paraffin embedded tumour, and faecal samples. Specific methodology to define this 'common predictive biomarker' was not prespecified, and specific performance measures were not pre-defined.