



OPEN

Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma

Karla A. Lee^{1,20}, Andrew Maltez Thomas^{2,20}, Laura A. Bolte^{1b,3,20}, Johannes R. Björk³, Laura Kist de Ruijter^{1b,4}, Federica Armanini², Francesco Asnicar^{1b,2}, Aitor Blanco-Miguez^{1b,2}, Ruth Board⁵, Neus Calbet-Llopart^{1b,6,7}, Lisa Derosa⁸, Nathalie Dhomen⁹, Kelly Brooks⁹, Mark Harland¹⁰, Mark Harries^{11,12}, Emily R. Leeming¹, Paul Lorigan^{1b,13,14}, Paolo Manghi^{1b,2}, Richard Marais⁹, Julia Newton-Bishop¹⁰, Luigi Nezi^{1b,15}, Federica Pinto², Miriam Potrony^{1b,7,11}, Susana Puig^{1b,7,11}, Patricio Serra-Bellver¹³, Heather M. Shaw¹⁶, Sabrina Tamburini¹⁵, Sara Valpione^{9,13}, Amrita Vijay^{1,17}, Levi Waldron^{2,18}, Laurence Zitvogel^{1b,8}, Moreno Zolfo², Elisabeth G. E. de Vries^{1b,4}, Paul Nathan¹¹, Rudolf S. N. Fehrmann^{1b,4}, Véronique Bataille^{1,19}, Geke A. P. Hospers⁴, Tim D. Spector^{1b,1,21}✉, Rinse K. Weersma^{1b,3,21}✉ and Nicola Segata^{1b,2,15,21}✉

The composition of the gut microbiome has been associated with clinical responses to immune checkpoint inhibitor (ICI) treatment, but there is limited consensus on the specific microbiome characteristics linked to the clinical benefits of ICIs. We performed shotgun metagenomic sequencing of stool samples collected before ICI initiation from five observational cohorts recruiting ICI-naïve patients with advanced cutaneous melanoma ($n = 165$). Integrating the dataset with 147 metagenomic samples from previously published studies, we found that the gut microbiome has a relevant, but cohort-dependent, association with the response to ICIs. A machine learning analysis confirmed the link between the microbiome and overall response rates (ORRs) and progression-free survival (PFS) with ICIs but also revealed limited reproducibility of microbiome-based signatures across cohorts. Accordingly, a panel of species, including *Bifidobacterium pseudocatenulatum*, *Roseburia* spp. and *Akkermansia muciniphila*, associated with responders was identified, but no single species could be regarded as a fully consistent biomarker across studies. Overall, the role of the human gut microbiome in ICI response appears more complex than previously thought, extending beyond differing microbial species simply present or absent in responders and nonresponders. Future studies should adopt larger sample sizes and take into account the complex interplay of clinical factors with the gut microbiome over the treatment course.

Therapeutic targeting of immune checkpoints such as programmed death-ligand 1, programmed cell death protein 1 (PD-1) and cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) with ICIs has revolutionized the treatment of advanced melanoma. Several landmark randomized controlled trials have shown notable and durable survival benefits, resulting in changes to standard of care internationally^{1,2}. Presently, over 50% of patients

treated with a combination of PD-1 and CTLA-4 blockade are alive after five years¹. Despite these advances, fewer than half of the patients who receive a single-agent ICI respond to it, whereas a higher response to combined targeting of PD-1 and CTLA-4 is associated with frequent toxicity with immune-related adverse events^{3,4}.

The discovery of a link between the gut microbiome and response to ICIs, in melanoma and other tumors, highlighted the

¹Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. ²Department CIBIO, University of Trento, Trento, Italy.

³Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands.

⁴Department of Medical Oncology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands. ⁵Department of Oncology, Lancashire Teaching Hospitals NHS Trust, Preston, UK. ⁶Dermatology Department, Hospital Clínic Barcelona, Universitat de Barcelona, IDIBAPS, Barcelona, Spain. ⁷Centro de Investigación Biomédica en Red en Enfermedades Raras, Instituto de Salud Carlos III, Barcelona, Spain. ⁸U1015 INSERM, University Paris Saclay, Gustave Roussy Cancer Center and Oncobiome Network, Villejuif-Grand-Paris, France. ⁹Molecular Oncology Group, CRUK Manchester Institute, University of Manchester, Manchester, UK. ¹⁰Division of Haematology and Immunology, Institute of Medical Research at St. James's, University of Leeds, Leeds, UK. ¹¹Biochemical and Molecular Genetics Department, Hospital Clínic de Barcelona, IDIBAPS and University of Barcelona, Barcelona, Spain. ¹²Department of Medical Oncology, Guys Cancer Centre, Guys and St Thomas's NHS Trust, London, UK. ¹³The Christie NHS Foundation Trust, Manchester, UK. ¹⁴Division of Cancer Sciences, University of Manchester, Manchester, UK. ¹⁵European Institute of Oncology (Istituto Europeo di Oncologia, IRCCS), Milan, Italy. ¹⁶Department of Medical Oncology, Mount Vernon Cancer Centre, Northwood, UK. ¹⁷Rheumatology & Orthopaedics Division, School of Medicine, University of Nottingham, Nottingham, UK. ¹⁸Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA. ¹⁹Department of Dermatology, Mount Vernon Cancer Centre, Northwood, UK. ²⁰These authors contributed equally: Karla A. Lee, Andrew Maltez Thomas, Laura A. Bolte. ²¹These authors jointly supervised this work: Tim D. Spector, Rinse K. Weersma, Nicola Segata.

✉e-mail: tim.spector@kcl.ac.uk; r.k.weersma@umcg.nl; nicola.segata@unitn.it

gut microbiome as a potential biomarker of response^{5,6} as well as a therapeutic target^{7,8}. Although there is compelling evidence for specific gut microbial features associated with beneficial responses in mouse studies^{9,10}, little consensus exists on which microbiome characteristics are associated with treatment responses in the human setting. In one of the largest metagenomic studies to date, Routy et al.¹¹ found responders to harbor significantly higher relative abundances of *Akkermansia muciniphila*, *Alistipes* and generally more *Firmicutes* compared to nonresponders¹¹, whereas Gopalakrishnan et al.⁶ found a higher relative abundance of *Faecalibacterium prausnitzii* in responders compared with nonresponders. In addition, Matson et al.⁵ found that responsiveness to PD-1 therapy was defined by an increased relative abundance of a group of eight species driven by *Bifidobacterium longum*. Frankel et al. reported that microbiota differed by ICI regimen but that enrichment of *Bacteroides caccae* was common in responders treated with any ICI regimen¹². Several confounding factors may have contributed to this lack of consensus, such as collection and DNA extraction protocols, dietary and medication-use differences among countries, issues of sample size and statistical power, variability in microbiome signatures among responders and microbial signals that are functionally related but intrinsic to each cohort. Cohort effects ranging from population-specific characteristics to methodological choices in sample processing and analysis are important problems in microbiome studies^{13,14}. Thus, larger and diverse cohorts with metagenomic data and standardized metadata are needed to better elucidate the microbiome determinants of response to immunotherapy.

The Predicting Response to Immunotherapy for Melanoma with Gut Microbiome and Metabolomics (PRIMM) studies are two separate prospective observational cohort studies recruiting subjects in parallel in the United Kingdom (PRIMM-UK) and the Netherlands (PRIMM-NL) since 2018. These cohorts of previously ICI-naïve patients with advanced melanoma provide extensive biosamples, including stool, serum and peripheral blood mononuclear cells, before and during ICI treatment, with detailed clinical and dietary data collected at regular intervals longitudinally. To study the role of the gut microbiome in ICI response, we performed shotgun metagenomic sequencing of baseline stool samples from the PRIMM cohorts, as well as three additional cohorts of ICI-naïve patients with advanced cutaneous melanoma (originating from Barcelona, Leeds and Manchester). A total of 165 microbiome samples sequenced from the patients enrolled in this study were analyzed together with 147 samples from smaller publicly available datasets. This method provided the largest assessment of the potential of the gut microbiome as a biomarker of response to ICI and allowed for investigation of specific microbial species or functions associated with response.

Results

In total, we recruited 175 patients from five distinct cohorts across the United Kingdom, the Netherlands and Spain with unresectable stage III and stage IV cutaneous melanoma who received ICI treatment after collecting stool samples (Table 1). We initially focused on the two single prospectively recruited PRIMM cohorts with more than 50 samples each, both of which exceed the sample size of all previously published datasets. In these cohorts, comprehensive clinical data, standardized radiological response assessments (RECIST v1.1) and biological specimens were collected before treatment and longitudinally at each cycle of treatment. There were statistically relevant clinical differences between these two cohorts, including the proportion of subjects who had received previous systemic targeted therapy (40% within PRIMM-NL and 20% within PRIMM-UK, $P=0.03$, Fisher's exact test), the proportion with BRAF-mutated tumors (55% within PRIMM-NL and 31% within PRIMM-UK, $P=0.02$, Fisher's exact test), the proportion of patients receiving combination ICI (ipilimumab and nivolumab) (22% within

PRIMM-NL and 53% within PRIMM-UK, $P=0.003$, Fisher's exact test, Supplementary Table 1) and the proportion of patients with M1d disease (36% within PRIMM-NL and 9% within PRIMM-UK, $P=0.0005$, Fisher's exact test). For these reasons, we decided to analyze the cohorts separately. All samples were subjected to shotgun metagenomic sequencing at an average sequencing depth of 7.74 Gb, and the 165 samples that passed strict quality control were analyzed at the taxonomic and functional potential levels using bioBakery 3 (ref. 15).

Linking the gut microbiome with response in the PRIMM cohorts. First, we assessed the quantitative taxonomic composition of the microbiome in relation to ORR, defined as a complete/partial response or disease stability as assessed by RECIST v1.1 6 months following initiation of ICI treatment and consistent with the response definition from previous microbiome-ICI studies^{11,12,16}. Using this definition, we found a borderline statistically significant difference in the microbiome composition of responders compared to nonresponders in the PRIMM-UK cohort (permutational multivariate analysis of variance (PERMANOVA) $P=0.05$, Fig. 1a), but not in PRIMM-NL ($P=0.61$, Fig. 1a). Progression-free survival at 12 months (PFS12), defined as duration of a complete/partial response or disease stability as assessed by RECIST v1.1 12 months following initiation of ICI treatment, was available for all patients in PRIMM-NL and 98% of patients in PRIMM-UK and showed similar association patterns (Fig. 1a). Anthropometric factors (e.g., body mass index (BMI), age and gender), information on previous non-immunotherapy treatments, previous drug therapies (e.g., antibiotics, proton pump inhibitors (PPIs) and steroids) and dietary patterns from food frequency questionnaires (FFQs), which we converted into dietary indices (i.e., the modified Mediterranean diet score and plant-based diet index; Supplementary Table 2) were collected in these cohorts for the majority of patients and thus considered in a multivariate analysis (Fig. 1b). Consistently with the univariate analysis, we found that in the PRIMM-UK cohort, ORR and PFS12 were the two variables explaining the largest variance ($P=0.09$ and $P=0.018$, respectively), whereas in PRIMM-NL, we found the use of PPIs, gender, Eastern Cooperative Oncology Group performance status and previous antitumor therapy to significantly explain the variance (PERMANOVA $P<0.05$) potentially hiding any PFS12- and ORR-associated signatures. Alpha diversity was generally not associated with response to ICIs, with only ORR in PRIMM-NL reaching statistically higher Shannon diversity in responders, but not richness, after accounting for confounding factors (Extended Data Fig. 1). We thus noticed cohort-dependent variability in the association between microbiome composition and population characteristics, including differences in dietary patterns (Supplementary Fig. 1), that could in part explain the difference in the link between the microbiome and ORR/PFS12 in the two PRIMM cohorts.

Response-associated microbiome features may not be reflected at the whole-microbiome level by common beta diversity summary metrics. We thus used a Lasso-based machine learning framework^{14,17} to estimate the prediction ability of the combination of taxonomic and functional features of the microbiome to segregate responders and nonresponders (cross-validation setting with nested cross-validation for feature selection; Methods and Fig. 1c). When exploring the relative abundance of each detected microbial species in this framework, we found substantial microbiome prediction capability in PRIMM-NL when using PFS12 as the endpoint (area under the receiver operating characteristic curve (AUC-ROC) 0.64) and in PRIMM-UK when using ORR (AUC-ROC 0.78), but ORR in PRIMM-NL and PFS12 in PRIMM-UK achieved much lower prediction levels (0.53 and 0.57, respectively). We then looked at the predicted functional potential of the microbiome by estimating the presence and relative abundance of microbial gene families (specifically Kyoto Encyclopedia of Genes and Genomes (KEGG)

Table 1 | Patient characteristics at study entry, treatment details and outcomes by RECIST v1.1

	PRIMM-UK (<i>n</i> = 55)	PRIMM-NL (<i>n</i> = 55)	Manchester (<i>n</i> = 25)	Leeds (<i>n</i> = 18)	Barcelona (<i>n</i> = 12)	All cohorts (<i>n</i> = 165)
Gender, <i>n</i> (%)						
Male	36 (65%)	31 (56%)	16 (64%)	11 (61%)	6 (50%)	100 (61%)
Female	19 (35%)	24 (44%)	9 (36%)	7 (39%)	6 (50%)	65 (39%)
Age (yr) at stage IV diagnosis, median (range)	65 (19–94)	61 (25–85)	66 (36–87)	60 (35–88)	64 (37–88)	63 (19–94)
Metastatic stage						
Stage 3 unresectable	5 (9%)	1 (2%)	0 (0%)	1 (6%)	0 (0%)	7 (4%)
M1a	13 (24%)	7 (13%)	5 (20%)	3 (17%)	4 (33%)	32 (20%)
M1b	12 (22%)	9 (16%)	6 (24%)	5 (27%)	5 (42%)	37 (22%)
M1c	20 (36%)	18 (33%)	11 (44%)	6 (33%)	3 (25%)	58 (35%)
M1d	5 (9%)	20 (36%)	3 (12%)	3 (17%)	0 (0%)	31 (19%)
BRAF mutant, <i>n</i> (%)	17 (31%)	30 (55%)	3 (12%)	8 (44%)	3 (25%)	61 (37%)
Previous systemic antimelanoma therapy, <i>n</i> (%)	11 (20%)	22 (40%)	7 (28%)	0 (0%)	1 (8%)	41 (25%)
Performance status						
0	17 (31%)	36 (65%)	15 (60%)	16 (89%)	7 (58%)	91 (55%)
1	32 (58%)	12 (22%)	9 (36%)	2 (11%)	1 (8%)	56 (34%)
2	5 (9%)	4 (7%)	0 (0%)	0 (0%)	0 (0%)	9 (5%)
3	1 (2%)	1 (2%)	1 (4%)	0 (0%)	0 (0%)	3 (2%)
Unknown	0 (0%)	2 (4%)	0 (0%)	0 (0%)	4 (34%)	6 (4%)
BMI (kg m ⁻²), mean (range)	28.7 (18.8–47.7)	27.3 (18.8–40.7)	26.9 (19.0–35.8)	30.2 (21.6–38.6)	27.1 (21.0–36.1)	28 (18.8–47.7)
Antibiotic use within 3 months of ICI, <i>n</i> (%)	9 (16%)	11 (20%)	4 (16%)	2 (11%)	0 (0%)	26 (16%)
PPI use within 3 months of ICI, <i>n</i> (%)	14 (25%)	20 (36%)	4 (16%)	5 (28%)	1 (8%)	44 (27%)
Treatment details						
ICI used						
Ipilimumab and nivolumab	29 (53%)	12 (22%)	2 (8%)	10 (56%)	1 (8%)	54 (32%)
Pembrolizumab	18 (33%)	10 (18%)	13 (52%)	3 (17%)	9 (75%)	53 (37%)
Nivolumab	8 (14%)	32 (58%)	0 (0%)	5 (27%)	2 (17%)	48 (24%)
Ipilimumab	0 (0%)	1 (2%)	10 (40%)	0 (0%)	0 (0%)	11 (7%)
Outcomes following ICI						
RECIST v1.1 response at 6 months						
Complete response	8 (15%)	6 (11%)	4 (16%)	2 (12%)	2 (17%)	22 (13%)
Partial response	16 (29%)	16 (29%)	6 (24%)	4 (22%)	1 (8%)	43 (26%)
Stable disease	9 (16%)	11 (20%)	3 (12%)	4 (22%)	2 (17%)	29 (18%)
Progressive disease	22 (40%)	22 (40%)	12 (48%)	8 (44%)	7 (58%)	71 (43%)
PFS12, <i>n</i> (%)	28 (51%)	30 (55%)	13 (52%)	8 (44%)	7 (58%)	86 (52%)

BMI, body mass index; PFS12, PFS at 12 months; PPI, proton pump inhibitor.

ortholog families; Fig. 1c). Here, the AUC-ROC exceeded 0.59 for both datasets and endpoints (ORR and PFS12). When we only considered patients who did not progress between 6 and 12 months (Fig. 1c), the prediction capabilities were more consistent across cohorts (0.62 for PRIMM-NL and 0.71 for PRIMM-UK when using species; 0.68 for PRIMM-NL and 0.72 for PRIMM-UK when using gene families). Clinical metadata fields were not predictive for response when considered alone and did not provide any clear improvements over microbiome features (Extended Data Figs. 2 and 3), suggesting these associations are not substantially influenced by potential confounding factors. These findings show that the gut microbiome does have a relevant association with response

to ICIs but appears to be cohort dependent and is likely due to features of the microbial community that are not responsible for large shifts in the global microbiome composition.

Limited reproducibility of response predictions across cohorts.

We next performed an integrated and cross-cohort analysis of the PRIMM cohorts together with three additional smaller cohorts from Barcelona, Leeds and Manchester, as well as five publicly available cohorts with consistent metagenomic and immunotherapy response data. The three cohorts we provide here in addition to PRIMM-UK and PRIMM-NL consist of a total of 55 samples (Table 1) and include patients who also collected stool samples and

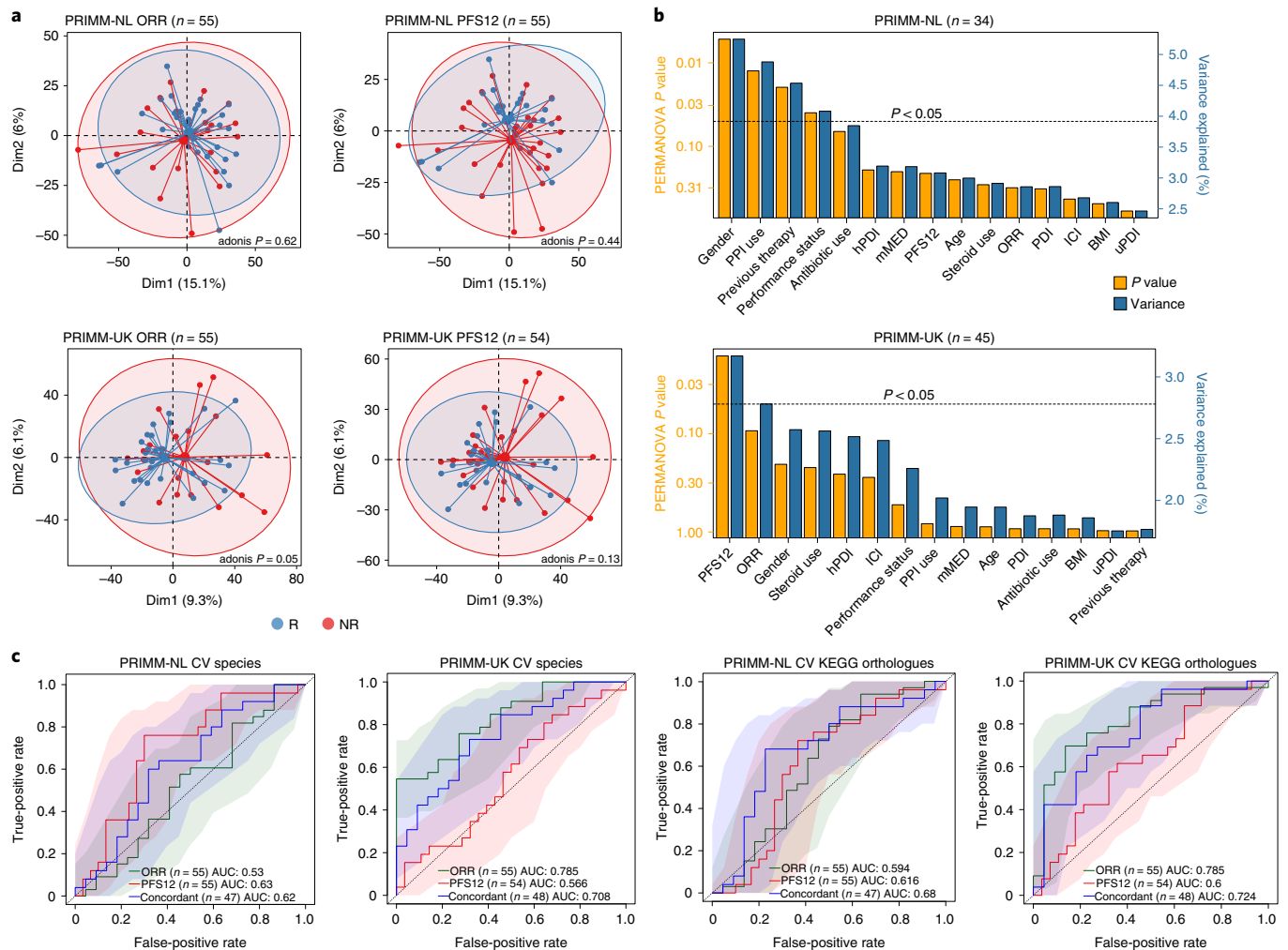


Fig. 1 | Association between the gut microbiome and response in the PRIMM-NL and PRIMM-UK cohorts. a, Response evaluated by ORR (Methods) is associated with the overall microbiome structure for PRIMM-UK ($P = 0.05$), but not for PRIMM-NL or PFS12, as represented visually using principal-component analysis (PCA) of species-level centered log-ratio-transformed relative abundances. P values were calculated using adonis and 999 permutations (Extended Data Figs. 2 and 3 show additional beta diversity analysis). Dim1, dimension 1; Dim2, dimension 2. **b**, Multivariate analysis showing the amount of inferred variance explained (R^2 , blue vertical bars) by each identified covariate and their respective P value (orange vertical bars) as determined by PERMANOVA on species-level centered log-ratio-transformed relative abundances. **c**, Machine learning association analysis between taxonomic (species abundance) and functional profiles (KEGG ortholog abundances) of the microbiome and response showed consistent associations with both response types (ORR or PFS12). The ‘concordant’ label includes only patients who did not progress between 6 and 12 months. AUC-ROC curves are computed using Lasso models trained using 100-repeated fivefold-stratified cross-validations. Shaded areas represent AUC-ROCs from each individual machine learning model. ICI indicates the use of a combination of ipilimumab and nivolumab or single agent. AUC, area under the curve; CV, cross-validation; hPDI, healthy plant-based diet index; PDI, plant-based diet index; uPDI, unhealthy plant-based diet index; mMED, modified Mediterranean diet score.

have metadata available as for the PRIMM cohorts. In this analysis, we also included four previously published cohorts with available metagenomic sequencing data and with ORR endpoints available^{5,6,12,16} and two with PFS12 endpoints available^{16,18}. As expected, the different datasets exhibited strong cohort-dependent effects in the microbial population, with ‘cohort’ explaining nearly ten times more variance than any other variable (Fig. 2a). Restricting the analysis to the datasets we sequenced, the batch effects were less strong, but even when batch-correction adjustment using an empirical Bayes framework¹⁹ on centered log-ratio-transformed species-level relative abundances was applied, ORR and PFS12 still only explained a limited fraction of the total variation in the microbiome (Fig. 2a) and were not associated with increased or decreased alpha diversity (Extended Data Fig. 1b,c).

Next, we assessed whether and how a microbiome-based machine learning model can predict ICI response in samples or whole cohorts not considered in the training of the classification model^{13,14,20} (Fig. 2b). We confirmed some single datasets had ORR-prediction levels higher than PRIMM-UK (e.g., Barcelona AUC-ROC 0.76) and PFS12-prediction levels higher than PRIMM-NL (WindTT_2020 AUC-ROC 0.90), but the higher AUC-ROC values usually corresponded to small cohorts on which the estimations are more variable when assessing cross-validation, and overall, the AUC-ROC values for the additional cohorts only occasionally exceeded 0.6 (Fig. 2b). The same sample-size limitations may also explain the AUC-ROC values obtained when a response-specific microbiome model was fitted on one cohort and then tested on a different one; such values reached 0.7 in only two cases and exceeded 0.6 in

around one-third of the cases (31.4%), with large variability among predictions. Interestingly, more consistent results, albeit still limited in their absolute values, were obtained when the model was fit on all but one cohort and applied on the left-out one (Fig. 2b). In this setting, which tried to alleviate cohort-specific effects by considering all but one cohort in the same model, we produced prediction values averaging 0.59 and 0.60 across the ORR and PFS12 endpoints, respectively, encompassing largely overlapping datasets. Functional characteristics of the microbiome profiled via classifications such as the enzyme category system²¹ or KEGG orthology (KO)²² achieved higher predictions of response compared to species relative abundances in several cases (Fig. 2c and Extended Data Fig. 4) but with higher variability and thus less cross-cohort consistency. These taxonomic and functional predictions (even when combined; Fig. 2c) were not dependent on the specific machine learning approach, as adopting random forest instead of Lasso produced similar results (Extended Data Fig. 5), and overall, they do not point to substantial cross-cohort reproducible links between the microbiome and ICI response. Nevertheless, they still highlight that profiling the microbiome at the species and gene-family levels in consistent cohorts can provide relevant indications of tumor response to ICIs.

Few reproducible biomarkers of response across cohorts. We then looked for microbial taxa or functions consistently associated with response to ICIs across the available cohorts in relation to ORR and PFS12 (Fig. 3, Extended Data Fig. 6 and Supplementary Tables 3 and 4). This analysis revealed two uncultivated *Roseburia* species associated with response; *Roseburia* sp. CAG:182 (also associated with response when using PFS12) and *Roseburia* sp. CAG:471 increased on average in responders consistently across all datasets with ORR endpoints available. Of note, *Roseburia* sp. CAG:182 was found to be in the panel of the five bacterial species most associated with favorable nutritional and cardiometabolic health markers in a recent large metagenomic population-based study²⁰, with a particularly strong inverse correlation with the inflammatory surrogate glycoprotein acetyl²³. With respect to those patients for whom PFS12 data were available, *Phascolarctobacterium succinatutens* and *Lactobacillus vaginalis* were both enriched in responders across all seven datasets, and each was detected as significant by three of the eight meta-analysis approaches using the same significance thresholds. *A. muciniphila* and *Dorea formicigenerans* were two species with high overall prevalence (65.8% and 85.9%, respectively) associated with ORR and PFS12 (by three and six meta-analysis methods, respectively). *Bacteroides clarus* instead exhibited higher relative abundances in nonresponders, both for ORR and PFS12, for all the seven datasets in which it can be detected and was significantly associated with nonresponders in PRIMM-NL after covariate adjustment (Fig. 3d). No single bacterium was a fully

consistent biomarker of response across all datasets; however, applying the same meta-analysis methods in the context of colorectal cancer as a methodological control confirmed strong and consistent biomarkers across cohorts, reinforcing previous meta-analyses^{13,14} (Extended Data Fig. 7 and Supplementary Table 5). We then focused on the panel of species associated with responders by at least three meta-analysis methods, which showed a high concordance with results from a Bayesian framework for compositional sequencing data (Pibble models; Supplementary Fig. 2). This panel contains taxa generally associated with healthy host conditions, including species with probiotic potential (*Bifidobacterium pseudocatenulatum*, *Lactobacillus ruminis* and *Turicibacter sanguinis*), as well as species involved in butyrate production (*Roseburia* spp., *Eubacterium hallii* and *Butyrivibrio synergistica*). *A. muciniphila* also belonged to this group, with significant association also in the PRIMM-NL study, supporting previous findings on its role in immunotherapy¹¹, whereas the presence of *Ruminococcus gnavus* as a biomarker of nonresponders in PRIMM-UK is in line with multiple reports of its association with poor cardiometabolic health²⁰ and several diseases^{24–26}. Expanding our analysis to gut microbiome members beyond bacteria and archaea, we found no association between response and the presence of *Blastocystis* (8.9% of individuals were *Blastocystis* positive, the highest prevalence of eukaryotic organisms found; Supplementary Table 6) or viruses (Supplementary Fig. 3).

When assessing treatment response with the predicted functional potential of the gut microbiome, we identified a number of KOs increased in responders (Supplementary Tables 7 and 8). These included a glycosyltransferase (*epsJ*) known to be essential for pellicle formation²⁷, with abundance contributions from *Intestinimonas butyriciproducens*, *Bifidobacterium angulatum* and *Bifidobacterium pseudoangulatum* and significant in six of the eight meta-analysis approaches used (Fig. 3c). We also detected an increased abundance of DNA adenine methylases, which are part of restriction-modification systems and can influence the expression of virulence genes²⁸, in responders using both ORR (Fig. 3c) and PFS12 (Extended Data Fig. 8), with contributions mostly from *I. butyriciproducens*. We also found a gluconate symporter to be increased in nonresponders, with abundance contributions mostly from *Ruminococcus torques*, *Escherichia coli* and *Klebsiella pneumoniae*. Gluconate is an important part of mucin 2 and is involved in binding of flagella²⁹.

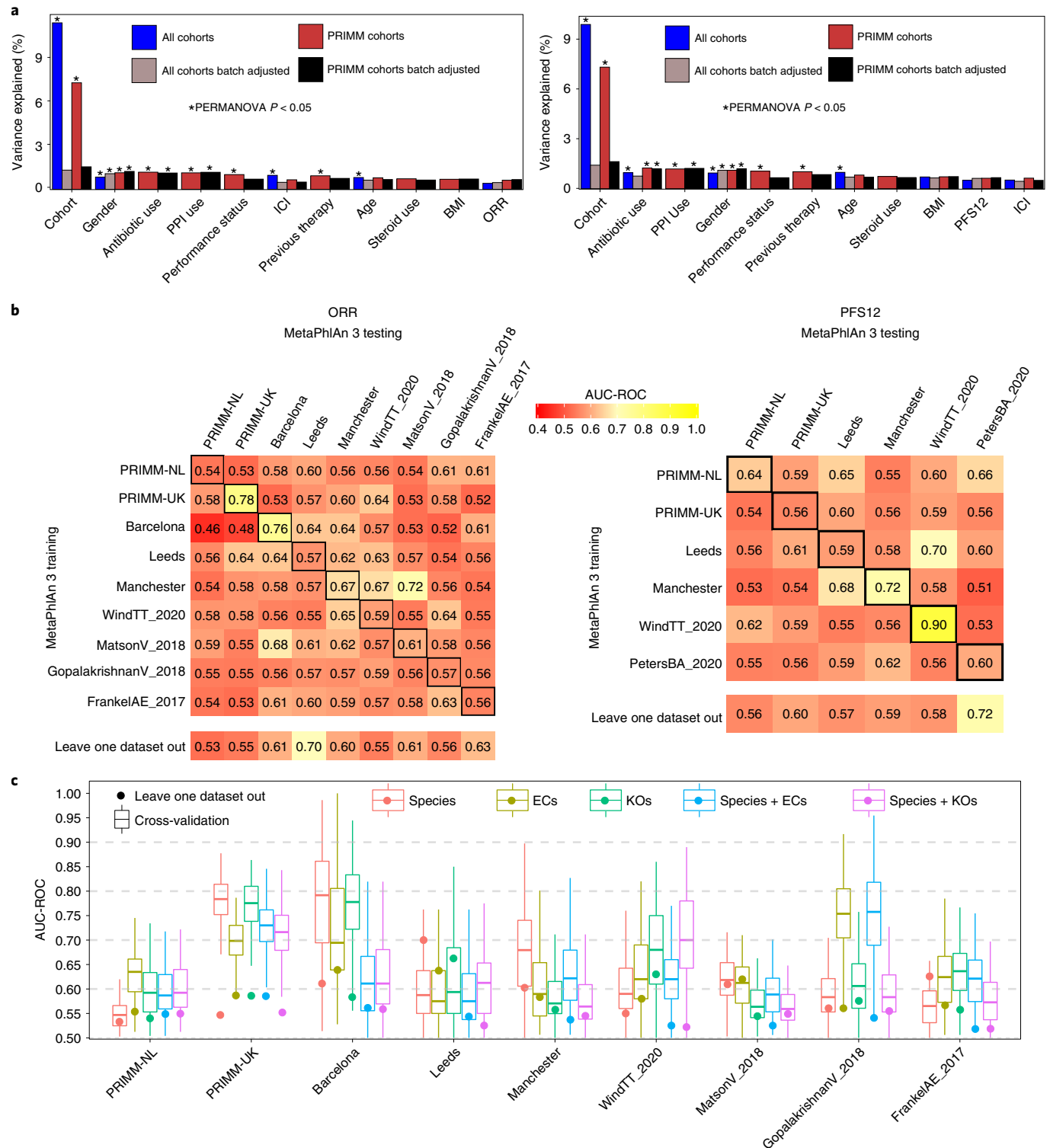
Clinical parameters linked to the microbiome. Patients with unresectable advanced melanoma have potential clinical confounding factors that can affect both the microbiome and might obfuscate the association between the pre-ICI microbiome composition and clinical response. A consistent panel of clinical information within our cohorts, ranging from PPI usage before treatment to

Fig. 2 | Integrated analysis of newly sequenced and publicly available datasets for cross-cohort response–microbiome association. **a**, Contribution of variables to the overall microbial community composition highlights the heterogeneity of the microbiome structure across cohorts that has a substantially higher effect than both anthropometric and clinical parameters. We either used all available cohorts or newly sequenced cohorts for which additional metadata were available. Batch-correction methods were applied to species-level abundances prior to distance calculations. The plot on the left uses ORR as the outcome variable, whereas the plot on the right adopts PFS12. **b**, Prediction matrix for microbiome-based prediction of response assessed via ORR (left matrix) and PFS12 (right matrix) within each cohort (values on the diagonal), across pairs of cohorts (one cohort used to train the model and the other for testing) and in the leave-one-cohort-out setting (training the model on all but one cohort and testing on the left-out cohort). We report the AUC-ROC values obtained from Lasso models on species-level relative abundances. Values on the diagonal refer to the median AUC-ROC values of 100-repeated fivefold-stratified cross-validations. Off-diagonal values refer to AUC-ROC values obtained by training the classifier on the cohort of the corresponding row and applying it to the cohort of the corresponding column. The leave-one-out row refers to the performances obtained by training the model using all but the cohort of the corresponding column and applying it to the cohort of the corresponding column. The same prediction matrix using functional microbiome profiles are available in Extended Data Fig. 4. **c**, ORR ($n=284$) cross-validation AUC-ROC values obtained from Lasso models trained using 100-repeated fivefold-stratified cross-validations (boxplots) and leave-one-dataset-out AUC-ROC values from Lasso models obtained by training the model using species-level relative abundances and all but the corresponding (circles). The lower and upper hinges of boxplots correspond to the 25th and 75th percentiles, respectively. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than 1.5 \times interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles. EC, enzyme category.

performance status and toxicity, allowed us to test whether and how such parameters were connected with microbiome composition and the microbiome signatures of response to ICIs. In line with previous reports^{30–32}, we found that PPIs had the strongest link with the microbiome in both PRIMM-NL and PRIMM-UK (Fig. 4a); however, PPI use was not associated with response or PFS12 (Extended Data Fig. 9) and so is not per se a factor biasing the microbiome–response associations. Moreover, ORR in PRIMM-UK reached the same level of prediction based on the taxonomic-based

microbiome model, which confirmed that the response–microbiome signature is independent of that of PPI use given the limited overlap in nonresponding patients who reported PPI use before ICI treatment ($P = 1$, Fisher's exact test). When looking at the functional potential, we also found that PPI use was the covariate best predicted by the gut microbiome in a leave-one-dataset-out setting for both PRIMM cohorts (Extended Data Fig. 9).

Performance status was also quite well inferred from the microbiome in PRIMM-UK and also in PRIMM-NL when



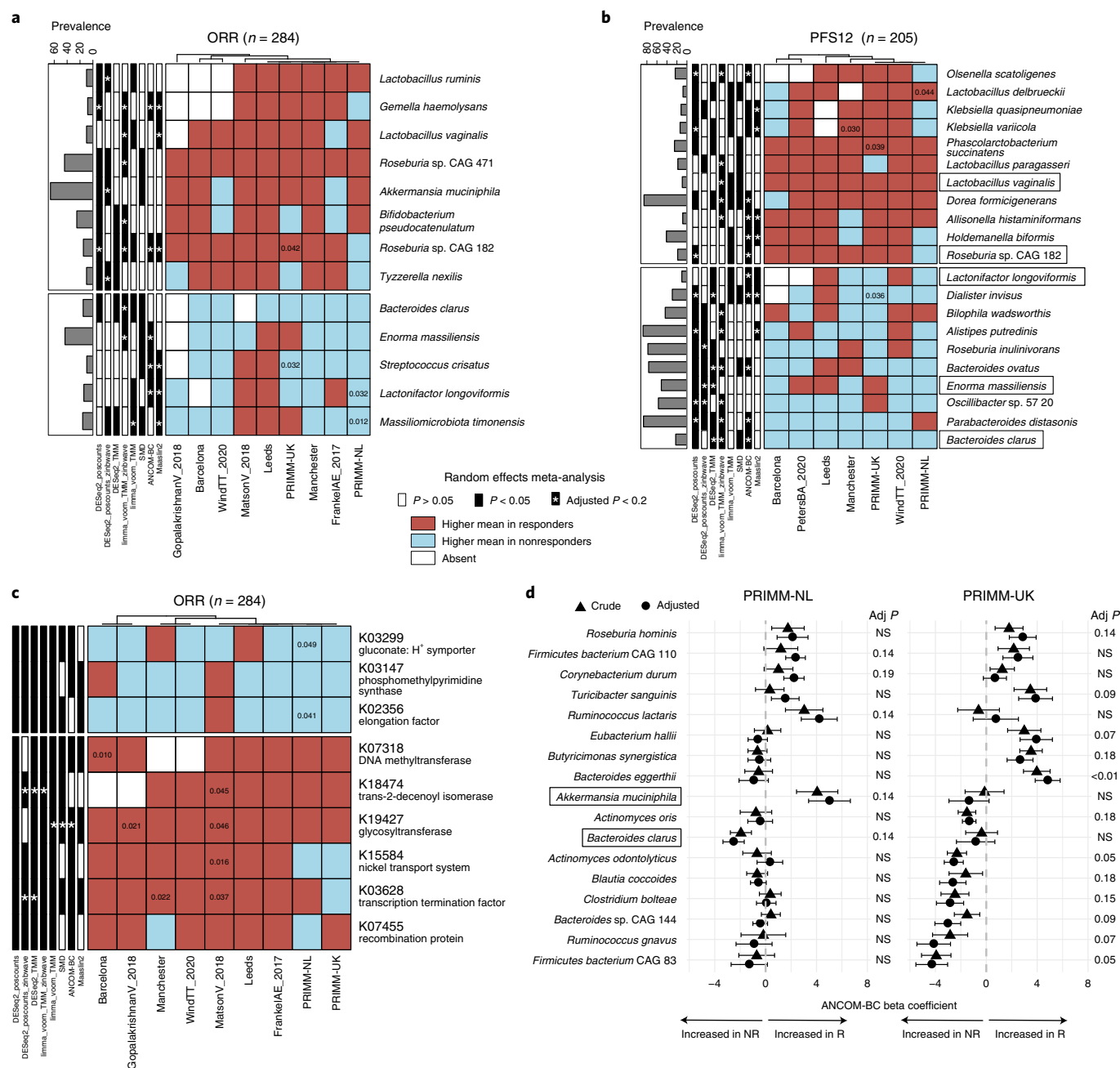


Fig. 3 | A panel of potential taxonomic and function microbiome biomarkers for response across cohorts. a, Species associated with ORR identified by a meta-analysis using different differential abundance methods. Species shown have random-effects model P values < 0.05 in at least three methods. Values inside the cells refer to unadjusted P values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of responders and nonresponders. The color of the cell was determined by comparing the mean relative abundance in responders to nonresponders; if the mean was higher in responders, then the cells were colored red; if it was higher in nonresponders, then it was colored blue. **b**, Species associated with PFS12 identified by a meta-analysis using different differential abundance methods. Species shown have random-effects model P values < 0.05 in at least three methods. Values inside the cells refer to unadjusted P values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of responders and nonresponders. **c**, KOs associated with response status identified by a meta-analysis using different differential abundance methods. The KEGG orthologues shown have random-effects model P values < 0.05 in at least six methods. Values inside the cells refer to unadjusted P values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of responders and nonresponders. **d**, Species associated with ORR in the two PRIMM cohorts (PRIMM-NL ($n = 47$) and PRIMM-UK ($n = 53$)) before and after adjusting for covariates that included PPI, antibiotic and steroid use; gender, performance status; previous therapy; age; and ICI. Species shown have covariate-adjusted multiple hypothesis testing-corrected $q < 0.2$ in one of the cohorts identified by ANCOM-BC. Symbols (circles and triangles) show the ANCOM-BC beta coefficient, and error bars represent standard error. Adj, adjusted; NR, nonresponders; R, responders; SMD, standardized mean differences.

considering the leave-one-dataset-out strategy (Fig. 4a and Extended Data Fig. 9). ICI-induced toxicity and colitis, for which preliminary immunological, genomic and microbiology

biomarkers are available³³, were less associated with the microbiome in our study, in which 43% and 36% of individuals experienced grade 3 or higher immune-related adverse events in

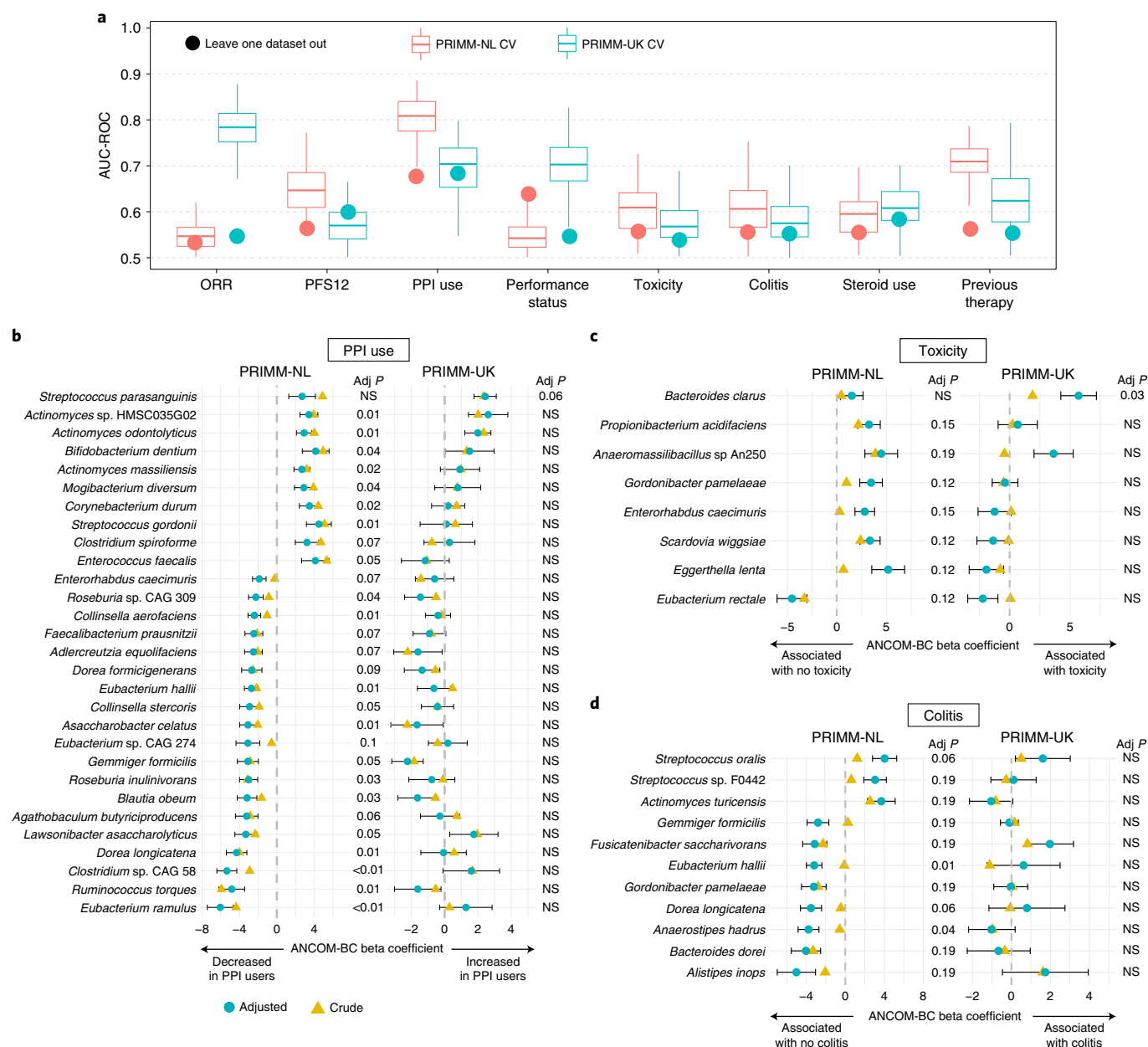


Fig. 4 | Covariate associations with the gut microbiome from the PRIMM cohorts. a, Cross-validation AUC-ROC values obtained from Lasso models trained using 100-repeated fivefold-stratified cross-validations (boxplots) and leave-one-dataset-out AUC-ROC values from Lasso models obtained by training the model using species-level relative abundances and all but the corresponding PRIMM cohort (circles). The lower and upper hinges of boxplots correspond to the 25th and 75th percentiles. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than 1.5 \times interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles (PRIMM-NL, $n = 55$; PRIMM-UK, $n = 55$). **b–d**, Species associated with PPI use (<3 months after the start of ICI), toxicity and colitis identified by ANCOM-BC with and without covariate adjustment (PRIMM-NL, $n = 47$; PRIMM-UK, $n = 53$). Covariates included in all models were ORR, performance status, previous therapy, age, ICIs (combination of ipilimumab and nivolumab or single agent), gender and antibiotic and steroid use. PPI use was also included as a covariate when analyzing colitis and toxicity. Species shown have covariate-adjusted multiple hypothesis testing-corrected $q < 0.2$ in one of the cohorts identified by ANCOM-BC. Symbols (circles and triangles) show the ANCOM-BC beta coefficient, and error bars represent standard error.

PRIMM-UK and PRIMM-NL, respectively (Methods). Within PRIMM-NL, *E. hallii* and *Anaerostipes hadrus* were both found to be significantly associated with the absence of colitis before ICI treatment, whereas no significant species emerged from the analyses of PRIMM-UK samples. *Bacteroides clarus* was significantly associated with the development of ICI-induced toxicity within PRIMM-UK and was also associated with nonresponse in

PRIMM-NL and the meta-analysis. *Eubacterium rectale* was significantly associated with the absence of toxicity in PRIMM-NL (and nonsignificantly in PRIMM-UK) and was associated with response when using PFS12 (Extended Data Fig. 8) in both PRIMM-UK and PRIMM-NL. Overall, several clinical parameters were found to be linked with the composition of the gut microbiome, but such associations appeared to be rather independent of ICI response.

Discussion

In this study, we present the largest metagenomic study to date to identify gut microbiome associations with ICI response by integrating published melanoma cohorts ($n=147$) with five new cohorts ($n=165$). This large set of real-world cohorts was meta-analyzed extending previous attempts³⁴ to verify whether the partially conflicting biomarkers of response to ICI found in single small datasets^{5,11,12,16,18} could be reconciled. Using ORR and PFS12 by RECIST v1.1 at 6 and 12 months, respectively, as endpoints, we confirmed both the presence of cohort-specific biomarkers and the absence of taxonomic or functional microbiome biomarkers that are consistent across all datasets. The lack of solid cross-study reproducible microbial biomarkers of ICI response is not attributable to analytical choices, as applying the same meta-analysis methods in the context of colorectal cancer confirmed the strong and consistent biomarkers across cohorts described elsewhere^{13,14}. The variability of the microbiome link to tumor response to ICIs was confirmed by machine learning analysis, which suggested that the microbiome is predictive of the response in some, but not all, cohorts. Importantly, the task of cross-cohort prediction to identify ICI responders and nonresponders in cohorts different from those used to develop the machine learning model clarified that it is still very difficult to overcome the limitations of population and microbiome heterogeneity to predict response reliably. Although we confirmed that the microbiome holds the potential to support the clinical practice for the treatment of patients with melanoma, as shown via fecal microbiota transplantation pilots^{7,8}, several limitations need to be overcome before we find robust microbial biomarkers.

There are multiple limitations of linking ICI and the gut microbiome. The rather small sample sizes of each individual cohort, despite the new additions in this work, are certainly one aspect, but it is inherently difficult to collect very large, properly annotated sample sizes in this clinical setting. Our work also highlighted previously overlooked limitations in defining response (as seen by differences in microbiome links to ORR and PFS12 in the same cohorts) and the vast number of potentially confounding factors for which it is difficult to adjust for. Moreover, the generally strong cohort microbiome effects even within multicenter studies appear to be much stronger in the analyzed melanoma ICI treatment trials than in other clinical settings. Current methodology to profile the gut microbiome can survey in-depth taxonomic and functional aspects of the microbial communities but generally lacks the ability to characterize the microbiome features at the immunological interface.

It is unclear why PFS12 was seen to have substantially higher microbiome prediction capability in PRIMM-NL, whereas ORR performed substantially better within the PRIMM-UK cohort. Nevertheless, aside from the large sample size of the combined cohorts, our study has multiple strengths; samples for this study came from three European countries and a number of publicly available datasets, representing a diverse and heterogeneous cohort of real-world patients, and we account for a variety of important confounders and prognostic factors that are often overlooked.

The gut microbiome has an important role to play in ICI response^{7,8}, as we also confirmed here, but this role appears likely to be more complex than previously reported in initial studies, extending beyond differential microbial abundances and encompassing complex interactions of the gut microbiome with clinical predictors and biological factors that may be specific to geographies in patients who benefit from treatment with ICIs. The gut microbiome is unique in each individual, even when considering identical twins^{20,35}, and a large fraction of this uniqueness is encoded at the level of single strains^{36–38}. Therefore, analytical methods are needed for deeper exploration of such individual-specific microbial diversity, but such resolution would also require sample sizes in the order of several thousand individuals to deal with the substantial increase

in the number of relevant microbial features. Understanding the functional output of the gut microbiota, immunological relations that exist between specific microbiota and the host at the strain level or even perhaps examination of the gut mucosa as distinct from stool will allow us further insight into the mechanisms by which the gut microbiome impacts treatment with ICIs. Links between the gut microbiome and ICI response can also be partially cancer specific and influenced by many factors that are inherently difficult to account for, and it will thus be important to extend analyses and meta-analysis beyond the cutaneous subtype of melanoma to include the uveal and mucosal subtypes, as well as other cancers for which ICI treatments are available. Even within the same cancer (sub)type, it is unlikely that the same microbiome features can reflect the uniqueness of the genetic and immune characteristics of each tumor, meaning that expectations on the existence of a universal, very accurate and highly reproducible link between the human microbiome and ICI response should be lowered. Continued efforts should thus be put into performing metagenomic investigations at substantially larger scales with improved representation of distinct populations while controlling for clinical covariates and ensuring that samples are collected and processed in the same manner and using the same techniques. This study improves the current understanding of the link between the gut microbiome and ICI response and sheds light on the complexities of microbiome science in human disease.

Online content

Any methods, additional references, Nature Research 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/s41591-022-01695-5>.

Received: 24 March 2021; Accepted: 13 January 2022;

Published online: 28 February 2022

References

- Larkin, J. et al. Five-year survival with combined nivolumab and ipilimumab in advanced melanoma. *N. Engl. J. Med.* **381**, 1535–1546 (2019).
- Ascierto, P. A. et al. Survival outcomes in patients with previously untreated BRAF wild-type advanced melanoma treated with nivolumab therapy: three-year follow-up of a randomized phase 3 trial. *JAMA Oncol.* **5**, 187–194 (2019).
- Larkin, J. et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N. Engl. J. Med.* **373**, 23–34 (2015).
- Amaria, R. N. et al. Neoadjuvant immune checkpoint blockade in high-risk resectable melanoma. *Nat. Med.* **24**, 1649–1654 (2018).
- Matson, V. et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).
- Gopalakrishnan, V. et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).
- Baruch, E. N. et al. Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* **37**, 602–609 (2020).
- McQuade, J. L. et al. Gut microbiome modulation via fecal microbiota transplant to augment immunotherapy in patients with melanoma or other cancers. *Curr. Oncol. Rep.* **22**, 74 (2020).
- Dubin, K. et al. Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. *Nat. Commun.* **7**, 10391 (2016).
- Vétizou, M. et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
- Routy, B. et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
- 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).
- Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).

14. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
15. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, e65088 (2021).
16. Wind, T. T. et al. Gut microbial species and metabolic pathways associated with response to treatment with immune checkpoint inhibitors in metastatic melanoma. *Melanoma Res.* **30**, 235–246 (2020).
17. 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).
18. Peters, B. A. et al. Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome Med.* **11**, 61 (2019).
19. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
20. Asnicar, F. et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* <https://doi.org/10.1038/s41591-020-01183-8> (2021).
21. Artimo, P. et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **40**, W597–W603 (2012).
22. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
23. Kettunen, J. et al. Biomarker glycoprotein acetyls is associated with the risk of a wide spectrum of incident diseases and stratifies mortality risk in angiography patients. *Circ. Genom. Precis. Med.* **11**, e002234 (2018).
24. Hall, A. B. et al. A novel *Ruminococcus gnavus* clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).
25. Valles-Colomer, M. et al. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632 (2019).
26. Ni, Y. H., Chua, H.-H., Chou, H.-C. C., Chiang, B.-L. & Liu, H.-H. Gut dysbiosis featured by abundant *Ruminococcus gnavus* heralds the manifestation of allergic diseases in infants. *Gastroenterology* **152**, S214 (2017).
27. Roux, D. et al. Identification of poly-*N*-acetylglucosamine as a major polysaccharide component of the *Bacillus subtilis* biofilm matrix. *J. Biol. Chem.* **290**, 19261–19272 (2015).
28. Heithoff, D. M., Sinsheimer, R. L., Low, D. A. & Mahan, M. J. An essential role for DNA adenine methylation in bacterial virulence. *Science* **284**, 967–970 (1999).
29. Troge, A. et al. More than a marine propeller: the flagellum of the probiotic *Escherichia coli* strain Nissle 1917 is the major adhesin mediating binding to human mucus. *Int. J. Med. Microbiol.* **302**, 304–314 (2012).
30. Imhann, F. et al. Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).
31. Macke, L., Schulz, C., Koletzko, L. & Malfertheiner, P. Systematic review: the effects of proton pump inhibitors on the microbiome of the digestive tract—evidence from next-generation sequencing studies. *Aliment. Pharmacol. Ther.* **51**, 505–526 (2020).
32. Llorente, C. et al. Gastric acid suppression promotes alcoholic liver disease by inducing overgrowth of intestinal *Enterococcus*. *Nat. Commun.* **8**, 837 (2017).
33. 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).
34. Limeta, A. et al. Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma. *JCI Insight* **5**, e140940 (2020).
35. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584.e3 (2016).
36. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
37. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
38. Karcher, N. et al. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biol.* **21**, 138 (2020).

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



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

© The Author(s) 2022

Methods

The prospective PRIMM cohorts and the other enrolled cohorts.

We prospectively collected fecal samples from patients receiving ICIs between August 2015 and January 2020 for patients with advanced cutaneous melanoma treated in the United Kingdom (PRIMM-UK study) and the Netherlands (PRIMM-NL study, made up of eligible patients from the COLIPI, POINTING and OncoLifeS studies). For the present study, we analyzed stool samples collected prior to treatment with ICIs (at baseline). PRIMM-UK (NCT03643289) is sponsored by East and North Hertfordshire NHS Trust with ethical approval from King's College London. OncoLifeS (Medisch Ethische Toetsingscommissie (METc) 2010/109), COLIPI (METc 2012/085, NCT02600143) and POINTING (METc 2018/350, NCT04193956) have all been approved by the METc of the University Medical Center Groningen in the Netherlands. OncoLifeS information is available on the Netherlands Trial Register (<https://www.trialregister.nl/trial/7839>). Patient samples within the Manchester cohort were collected with written full-informed patient consent under Manchester Cancer Research Centre Biobank ethics application 07/H1003/161 + 5 (updated in 18/NW/0092) and approval for the work under Manchester Cancer Research Centre Biobank Access Committee application 13_RIMA_01. Barcelona cohort samples were subjected to the ethical committee of Hospital Clínic de Barcelona approval (registry HCB/2015/1032). Data and samples from Leeds were collected in a study named "Developing a blood test of immunity in illness: a study examining the peripheral blood transcriptome in patients with cancer, autoimmune disease, immunodeficiency or iatrogenic immune suppression" (Research Ethics Committee (REC) reference 15/NW/0933). Informed written consent was obtained for collection of samples and data, sharing anonymized data and working with collaborators whether academic or commercial.

Samples collected within the three external cohorts of Barcelona, Manchester and Leeds were subject to similar ethical approvals. Patients who fulfilled the following criteria were eligible for these analyses: (i) histologically or cytologically confirmed nonresectable advanced (stage III or IV) cutaneous melanoma, (ii) treatment with ICIs (nivolumab, pembrolizumab or ipilimumab or a combination of ipilimumab and nivolumab) at the recommended dose as a first-line ICI and (iii) 18 years of age or older. Written informed consent was obtained from all patients. High-quality fecal samples were collected from these patients before initiation of ICI treatment ($n = 55$ for the UK cohort and $n = 55$ for the Dutch cohort). Additional patients were enrolled from cohorts outside the setting a prospective clinical trial and performed in Leeds ($n = 19$), Barcelona ($n = 12$) and Manchester ($n = 30$) between March 2015 and November 2019 but from whom fecal samples were collected at time points similar to those used in our included prospective studies. Written informed consent was obtained from all patients.

Sample and data collection. Baseline demographics, including sex, age, BMI, Eastern Cooperative Oncology Group performance status and medication use, were collected, along with tumor staging and previous anticancer therapy data. Demographic and dietary data were collected as part of a screening visit up to 14 days before ICI treatment began. All baseline antibiotic or PPI use within 3 months of commencing ICI treatment was documented. Tumor staging took place up to 1 month before the start of treatment. Routine blood hematology and biochemistry data, including white cell subsets, platelet count, lactate dehydrogenase and albumin, were recorded at baseline and with each treatment cycle. All baseline antibiotic or PPI use within 3 months of commencing ICI was documented.

Patients received oral and written instructions regarding the stool collection procedure. Patients within PRIMM-UK and PRIMM-NL were requested to collect approximately 3–5 ml plain feces using a collection kit that could be used at home and then store the sample in their freezer directly after collection. PRIMM-NL samples were transported to the hospital in a frozen, insulated cooling bag to prevent thawing. Due to the geographic dispersal of PRIMM-UK patients, samples were collected and placed in Thermo Fisher Scientific kits and sent by special post to the laboratory at King's College London. After arrival in the hospital, the samples were directly stored at -80°C . Plain stool samples from the Manchester cohort were either collected on site at the hospital and stored directly at -80°C within 4–6 h of collection or collected into sample containers and sent by special post to the laboratories of CRUK Manchester Institute and stored directly at -80°C upon arrival. Patients within the Barcelona cohort used the OMNIgene GUT collection kit (DNA Genotek). Fecal DNA was extracted from 1 to 14 days after sample collection using the PowerFecal DNA Isolation Kit (previously Mo Bio, currently Qiagen) and kept frozen until needed. Patients from Leeds also collected stool at home using the OMNIgene GUT collection kit (DNA Genotek), and samples were returned to the research nurse.

Radiological evaluation, consisting of a computed tomography (CT) scan of the thorax, abdomen and pelvis and magnetic resonance imaging of the brain, was performed at baseline (i.e., before the first dose of immunotherapy). A small number of patients had positron emission tomography scans with a CT component. Follow-up radiological evaluation was performed every 10–14 weeks as long as the patient received systemic therapy. Additional CT and/or magnetic resonance imaging scans were performed when there was suspicion of progression. If the first radiological evaluation after start of therapy was inconclusive, then a confirmatory scan was performed 4–12 weeks later.

Definition of response to therapy. Response to ICI was classified according to RECIST v1.1 criteria. On the basis of radiographic response, patients were classified as responders (CR, PR or SD) or nonresponders (PD). Clinical endpoints were defined as overall response rate and PFS (defined as the time from the first dose of an ICI to the first event; i.e., disease progression or death from any cause). All patients had toxicity during or after ICI treatment, as documented by Common Terminology Criteria for Adverse Events v5. Toxicity was deemed present at or above Common Terminology Criteria for Adverse Events grade 1, as was the case with colitis. In order to include late responders in our analysis, patients with progressive disease (PD) on the first radiological evaluation but a response at the second radiological evaluation compared to baseline were also labeled responders. Patients with PD on the first radiological evaluation that was confirmed on the next follow-up scan, or patients with PD on the first radiological evaluation who were unable to complete a confirmation scan due to clinical progression or death, were labelled nonresponders.

Dietary data collection. In 93 of the 110 (84.5%) prospectively recruited patients, dietary data were collected before treatment and within 14 days of commencing ICIs. Dietary intake was assessed through two different FFQs: the Dutch Healthy Diet-FFQ³⁹ and the EPIC-Norfolk FFQ^{40,41}. Food items were mapped to create one dataset. Four food-based scores were calculated to address relative dietary quality as confounder in the microbiome–response analysis: the modified Mediterranean diet score⁴², original plant-based diet index⁴³, healthy plant-based diet index⁴³ and unhealthy plant-based diet index⁴³. The composition of scores by food groups in the UK and Dutch cohorts is given in Supplementary Table 2.

DNA extraction and sequencing. DNA was isolated at King's College London using the Max Core protocol. Samples with a high-quality DNA profile were further processed. Sequencing libraries were prepared using the Illumina Nextera DNA Flex Library Prep Kit according to the manufacturer's protocols. Libraries were multiplexed using dual indexing and sequenced for 300-bp paired-end reads using the Illumina NovaSeq6000 platform according to the manufacturer's protocols. We obtained a total of 1,283 Gb with an average of 53,919,210 reads per sample before quality control and preprocessing.

The publicly available datasets considered. We download metagenomic data from four publicly available datasets (PetersBA_2020, GopalakrishnanV_2018, MatsonV_2018 and FrankelAE_2017) through the Sequence Read Archive using the accession numbers SRP197281 (ref. ¹⁸), ERP104610 (ref. ⁶), SRP116709 (ref. ⁵) and SRP115355 (ref. ¹²). Metagenomic data and metadata from WindTT_2020 (ref. ¹⁶) were provided by the authors of the study. These publicly available cohorts are shown in Supplementary Table 11. We excluded any samples taken after the start of ICI therapy, nonmetagenomic samples, nonfecal samples and samples with low sequencing depth (less than one million reads). We classified patients into responder and nonresponder groups according to RECIST 1.1 criteria; patients with complete or partial response, as well as stable disease at first evaluation, were classified as responders, whereas patients with PD were classified as nonresponders.

Metagenome quality control and preprocessing. Shotgun metagenomic sequencing was performed at the NGS Core Facility at University of Trento. The quality of all sequenced metagenomes was controlled using the preprocessing pipeline implemented in <https://github.com/SegataLab/preprocessing>. This preprocessing pipeline consists of three main stages: (1) initial quality control by removing low-quality reads (quality score $<Q20$), fragmented short reads (<75 bp) and reads with more than two ambiguous nucleotides; (2) contaminant DNA removal using Bowtie 2 (ref. ⁴⁴) and the sensitive local parameter, removing both the phiX174 Illumina spike-in and human-associated reads (hg19); and (3) sorting and splitting for the creation of standard forward, reverse and unpaired reads output files for each metagenome. Of 190 baseline samples collected across the five observational cohorts, 25 samples failed the metagenomic sequencing and preprocessing pipeline and resulted in metagenomes with <1 Gb pairs. The remaining 165 samples passed strict quality control and were processed into taxonomic and predicted pathway abundances.

Microbiome taxonomic and functional potential profiling. The metagenomic analysis was performed following the general guidelines⁴⁵ and relying on the bioBakery 3 environment^{15,46}. bioBakery 3 is a set of integrated and improved methods for taxonomic (MetaPhlAn 3.0), strain-level (StrainPhlAn 3.0 and PanPhlAn 3.0), functional (HUMANn 3.0) and phylogenetic (PhyloPhlAn 3.0) profiling of metagenomes. bioBakery 3 leverages a set of 99,200 high-quality and fully annotated reference microbial genomes spanning 16,800 species and the 87.3 million UniRef90 functional annotations available in UniProt as of January 2019. The taxonomic profiling and quantification of organisms' relative abundances of all metagenomic samples were quantified using MetaPhlAn 3.0 (ref. ⁴⁷) with default parameters. Functional potential analysis of the metagenomic samples was performed using HUMANn 3.0 (ref. ⁴⁸) with default parameters. In total, we identified 608 species, 510 pathways, 6,131 distinct KOs and 2,558 enzyme categories in 165 samples from the five different cohorts. Detection of *Blastocystis* in gut metagenomes was

performed as previously described⁴⁹. For viral profiling, we used 699 bacteriophages from Viral RefSeq⁵⁰ that were found in more than 20 samples within the unbinned fraction of the metagenomic assembled genomes described by Pasolli et al.³⁷ and from a set of 255 highly enriched viromes selected with the ViromeQC tool⁵¹. Sequences were clustered at 70% identity and further joined into 'viral groups' if they shared at least 90% similarity across clusters. In total, 128 groups of known viral bacteriophages were constructed. To calculate the depth of coverage of each viral group, we first mapped the raw reads of each sample against a nonredundant collection of representative sequences for each viral group. Then, a second mapping was performed on the best-matching sequence for each group and each sample. Depth of coverage was calculated with CMSeq (<https://github.com/SegataLab/cmseq>).

Statistical analysis. Multivariate survival analyses were performed using Cox regression models to determine hazard ratios and 95% confidence intervals for PFS and adjusting for other clinicopathologic features using the `cophx` function from the R survival package (v3.2-7). We conducted PERMANOVA using the function `adonis` from the `vegan` R package (v2.5-7) with the Atchinson distance on centered log ratios using both species and KO relative abundances. The *P* and *R*² values were determined by 10,000 permutations using all variables in the model. Batch correction was applied to centered log ratios using the `combat` function available in the Surrogate Variable Analysis (v3.38) R package. Univariate analysis was performed using both two-tailed Wilcoxon tests comparing differences in relative abundances (values inside the heatmap cells in Fig. 3a–c) and ANCOM-BC (v.1.0.1)⁵² with default parameters, a library size cutoff of 1,000 reads and no structural zero detection. Multivariate analysis was performed using ANCOM-BC, including the covariates of interest in the model formula. *P* values were corrected for multiple hypothesis testing using the Benjamin–Hochberg procedure, and a false discovery rate <0.2 was defined as the significance threshold. In addition, we also implemented a Bayesian multinomial logistic-normal linear regression model called Pibble from the R package `fido`⁵³, which allows for linking covariates to compositional overdispersed count data. We transformed the taxonomic relative abundances into count values for Pibble via logistic-normal distribution modeling⁵⁴. In Pibble, the regression coefficients are ranked to determine which microbial features change the most between conditions and are identical to the rankings of absolute differentials⁵⁵.

Meta-analysis. An ensemble of differential abundance methods and normalizations (eight in total) were used to estimate fold changes with their respective confidence intervals between responders and nonresponders and supplied to a random-effects model via the `rem_mv` function in the MetaVolcanoR R package (v.1.4.0) using the restricted maximum-likelihood estimator model. Random-effects *P* values obtained from each of these methods were corrected for multiple hypothesis testing using the Benjamin–Hochberg procedure. For methods requiring count data (i–v and vii), absolute raw counts were estimated from species-level MetaPhlAn 3 relative abundances by multiplying these values by the total number of reads for each sample. The following methods were based on an assessment of statistical methods to detect differentially abundant features in microbiome data⁵⁶ and are available at <https://github.com/mcargaro93/sc2meta>:

- i. DESeq2 (v.1.30.0)⁵⁷ with the poscounts estimator (DESeq2_poscounts);
- ii. DESeq2 with the poscounts estimator and a zero-inflated negative binomial model (DESeq2_poscounts_zb), and observational weights were computed using the `zinbwave` package (v.1.12.0)⁵⁸ and supplied to the DESeqDataSet class object to account for zero inflation;
- iii. DESeq2 with trimmed mean of *M* values (TMM; DESeq2_TMM), where normalization factors were calculated using TMM normalization, and a negative binomial generalized log-linear model was fit to the read counts of each feature using the `glmFit` function and the `edgeR` (v.3.32.0)⁵⁹ package;
- iv. limma (v3.46.0)⁶⁰ with TMM values (limma_voom_TMM) (the limma package includes a voom function that transforms previously normalized counts to log counts per million, estimates a mean–variance relationship and uses this to compute appropriate observational-level weights); and
- v. limma with TMM values and a zero-inflated negative binomial model (limma_voom_TMM_zb) (to adapt the limma-voom framework to zero inflation, `zinbwave` weights were multiplied by voom weights); Other methods not originally considered in this assessment were also included in the meta-analysis:
- vi. standardized mean differences (relative abundances were arcsine-square root transformed and followed the same procedure as in Thomas et al.¹⁴);
- vii. ANCOM-BC (v.1.0.1)⁵², which uses a linear regression framework in log scale and accounts for sampling fraction by introducing a sample-specific bias correction that is estimated from the observed data (we used the same parameters as described in the univariate/multivariate analysis); and
- viii. Maaslin2 (v.1.4.0)⁶¹, where logit-transformed relative abundances were normalized with total-sum scaling and supplied to the `maaslin2` function using the variable of interest as a fixed effect.

Machine learning analysis. Data preprocessing, model building and model evaluation were performed using the SIAMCAT⁶² R package (v.1.6.0). Species

relative abundances were filtered to remove markers with low overall abundance (1×10^{-4} maximum abundance cutoff), \log_{10} -transformed (after adding a pseudocount of 1×10^{-5} to avoid nonfinite values) and standardized as *z*-scores. Functional profiles, such as KEGG orthologous and level 4 enzyme category abundance profiles were preprocessed similarly but using 1×10^{-6} as the maximum abundance cutoff and 1×10^{-9} as a pseudocount during log transformation.

Cross-validation. A nested cross-validation procedure was applied to calculate within-cohort accuracy (cells on the diagonal in Fig. 2b) by splitting data into training and test sets for 100-times repeated, fivefold-stratified cross-validation (balancing class proportions across folds). For each split, an L1-regularized (Lasso) logistic regression model was trained on the training set, which was then used to predict the test set. The lambda parameter was selected for each model to maximize the AUC-ROC under the constraint that the model contained at least five nonzero coefficients.

Cross-study validation. Metagenomic classifiers were trained on a single cohort and their performance was externally assessed on all other cohorts (off-diagonal cells in Fig. 2b), which were normalized for comparability in the same way as the training dataset. All 500 models derived from the cross-validation on the training dataset (100-times-repeated fivefold cross-validation) were applied to the hold-out dataset, and median predictions were taken from all models.

Leave one dataset out. Data from one cohort were set aside as an external validation set, whereas data from the remaining cohorts were pooled as a single training set on which we implemented the same procedure as above for 100-times-repeated fivefold-stratified cross-validation.

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

Data availability

The metagenomes and main metadata relevant to the analyses are deposited in the European Nucleotide Archive under accession number [PRJEB43119](https://www.ebi.ac.uk/ena/record/PRJEB43119). The four publicly available datasets were downloaded through the Sequence Read Archive using the accession numbers SRP197281, [ERP104610](https://www.ebi.ac.uk/ena/record/ERP104610), SRP116709 and SRP115355. All MetaPhlAn 3 and HUMAnN 3 profiles are available within the latest version of curatedMetagenomicData (<https://bioconductor.org/packages/curatedMetagenomicData/>).

References

39. van Lee, L. et al. Evaluation of a screener to assess diet quality in the Netherlands. *Br. J. Nutr.* **115**, 517–526 (2016).
40. Bingham, S. A. et al. Nutritional methods in the European Prospective Investigation of Cancer in Norfolk. *Public Health Nutr.* **4**, 847–858 (2001).
41. Bingham, S. A. et al. Validation of dietary assessment methods in the UK arm of EPIC using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin C and carotenoids as biomarkers. *Int. J. Epidemiol.* **26**, S137–S151 (1997).
42. Trichopoulos, A., Costacou, T., Bamia, C. & Trichopoulos, D. Adherence to a Mediterranean diet and survival in a Greek population. *N. Engl. J. Med.* **348**, 2599–2608 (2003).
43. Satija, A. et al. Healthful and unhealthful plant-based diets and the risk of coronary heart disease in U.S. adults. *J. Am. Coll. Cardiol.* **70**, 411–422 (2017).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
45. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
46. McIver, L. J. et al. bioBakery: a metaomic analysis environment. *Bioinformatics* **34**, 1235–1237 (2018).
47. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
48. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
49. Beghini, F. et al. Large-scale comparative metagenomics of *Blastocystis*, a common member of the human gut microbiome. *ISME J.* **11**, 2848–2863 (2017).
50. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
51. Zolfo, M. et al. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).
52. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
53. Silverman, J. D., Roche, K., Holmes, Z. C., David, L. A. & Mukherjee, S. Bayesian multinomial logistic normal models through marginally latent Matrix-T processes. Preprint at <https://arxiv.org/abs/1903.11695> (2019).
54. Atchison, J. & Shen, S. M. Logistic-normal distributions: some properties and uses. *Biometrika* **67**, 261–272 (1980).

55. Morton, J. T. et al. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).
56. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol.* **21**, 191 (2020).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
58. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
59. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
60. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
61. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* **17**, e1009442 (2021).
62. Wirbel, J. et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 93 (2021).

Acknowledgements

This work was supported by the Seerave Foundation. The work was also supported by the European Research Council (ERC-STG project MetaPG to N.S.); MIUR 'Futuro in Ricerca' (grant RBFR13EWWI_001 to N.S.); the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project to N.S.); the National Cancer Institute of the National Institutes of Health (grant 1U01CA230551 to N.S. and L.W.); the Premio Internazionale Lombardia e Ricerca 2019 (to N.S.); Spanish Fondo de Investigaciones Sanitarias (grants PI15/00716 and PI15/00956) of the Instituto de Salud Carlos III, Spain, co-financed by European Development Regional Fund "A way to achieve Europe" to ERDF; the Wellcome Trust (grant 100282/Z/12/Z to N.C.-L.); Cancer Research UK (grants A27412 and A22902 to N.C.); and the Dutch Cancer Society (grant 10034 POINTING to E.G.E.d.V.). N.C.-L. is the recipient of a PhD fellowship (FPU17/05453) from Ministerio de Educación, Cultura y Deportes, Spain. The Leeds group was supported by the Medical Research Council (grant MR/MO19012/1). S.V. was supported by a Harry J. Lloyd Charitable Trust Career Development Award. We acknowledge the Seerave Foundation, which funded this work and many other projects investigating the gut microbiome in human health and disease. We thank the team of the NGS Core Facility at University of Trento for support in sample preparation and for metagenomic sequencing and the high-performance computing team at the University of Trento. We also thank Oncobiome and the Institut Gustave Roussy for their involvement and support of our initiative. The collection of samples from the Hospital Clínic in Barcelona was funded by grant PI15/00716 from Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III, Spain. We would like to thank Ranko Gacesa and the authors of the WindTTT_2020 study for kindly providing metagenomic data and metadata

for their cohort. Lastly, we would like to thank all the patients who selflessly took the time to collect samples for this project.

Author contributions

K.A.L., L.K.d.R., L.A.B., N.C.-L., P.S.-B. and M. Harland collected the clinical data. N.D., K.B., M.P., S.V., F. Armanini, F.P. and A.V. provided organizational and scientific assistance relating to sample collection and processed the samples. L.A.B. supervised the dietary aspects of this study, assisted by E.R.L. A.M.T. performed the analyses, assisted by J.B.R., F. Asnicar, A.B.-M., P.M. and M.Z. and supervised by N.S. K.A.L., A.M.T., L.A.B. and N.S. wrote the manuscript, which was reviewed and edited by all authors prior to submission. L.Z., L.D., L.N., S.T. and L.W. supported the interpretation of data, analysis and results. R.B., P.L., R.M., J.N.-B., H.M.S., M. Harries, E.G.E.d.V., P.N., R.S.N.F., S.P., V.B., G.A.P.H., T.D.S. and R.K.W. provided supervision over the studies and collection protocols within which samples were collected. All authors approved the final manuscript.

Competing interests

R.K.W. acted as a consultant for Takeda; received unrestricted research grants from Takeda, Johnson & Johnson, Tramedico and Ferring; and received speaker fees from MSD, AbbVie and Janssen Pharmaceuticals. E.R.L. is a consultant for ZOE Global. E.G.E.d.V. reports an advisory role at Daiichi Sankyo, NSABP and Sanofi (paid to University Medical Center Groningen) and research funding from Amgen, AstraZeneca, Bayer, Chugai Pharma, CytomX Therapeutics, G1 Therapeutics, Genentech, Nordic Nanovector, Radius Health, Regeneron, Roche, Servier and Synthon (paid to University Medical Center Groningen). S.P. received speaker fees from Almirall, BMS, ISDIN, La Roche Posay, Leo Pharma, Regeneron, Roche and Sanofi; acted as advisory board member of Almirall, ISDIN, La Roche Posay, Pfizer, Roche, Regeneron, Sanofi and Sun Pharma; and received research funding from Abbvie, AMGEN, ISDIN, La Roche Posay, Leo Pharma and Novartis. R.B. has received honoraria from, and sits on advisory boards of, Novartis, BMS and MSD. All other authors declare no competing interests.

Additional information

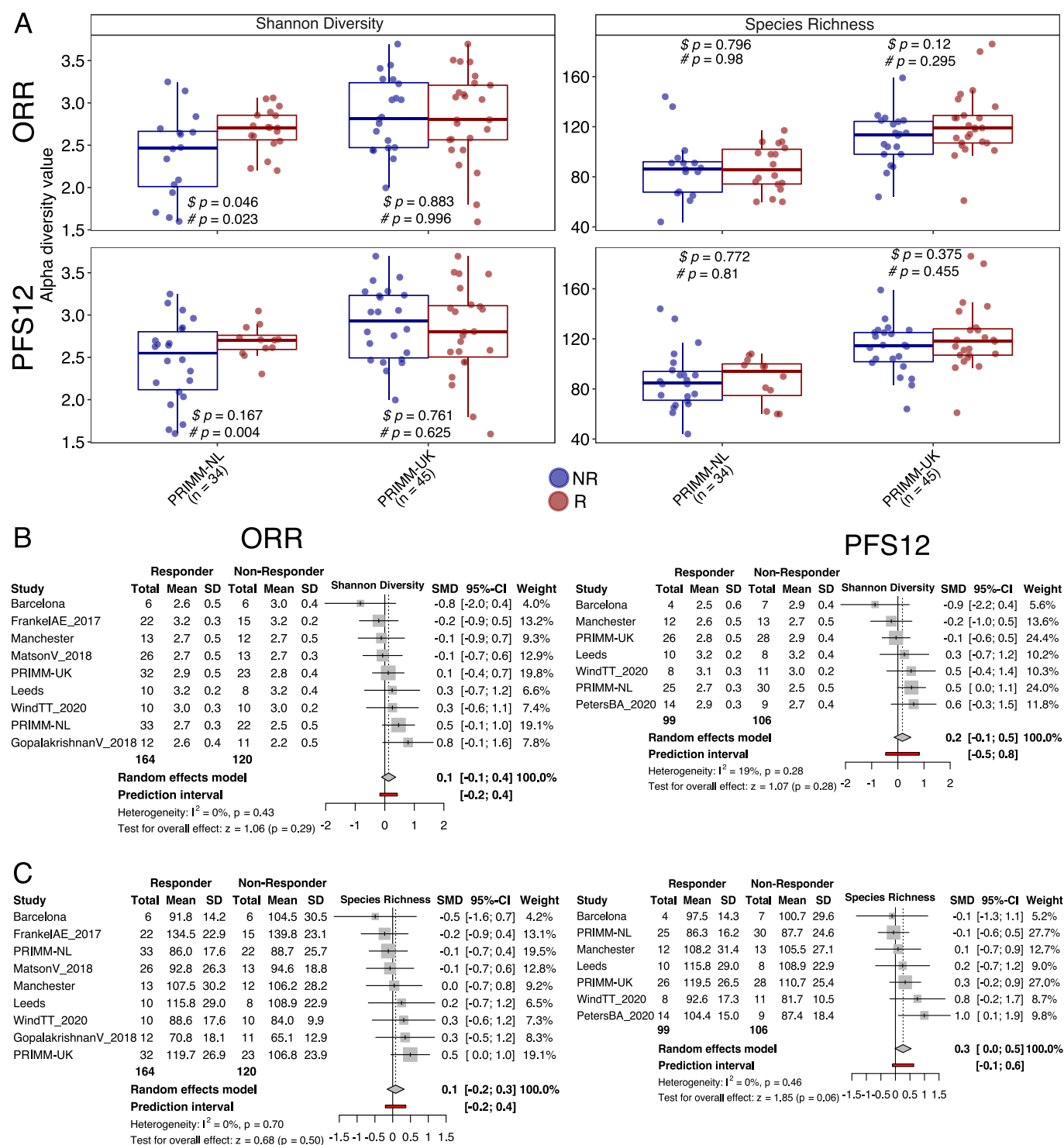
Extended data is available for this paper at <https://doi.org/10.1038/s41591-022-01695-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-022-01695-5>.

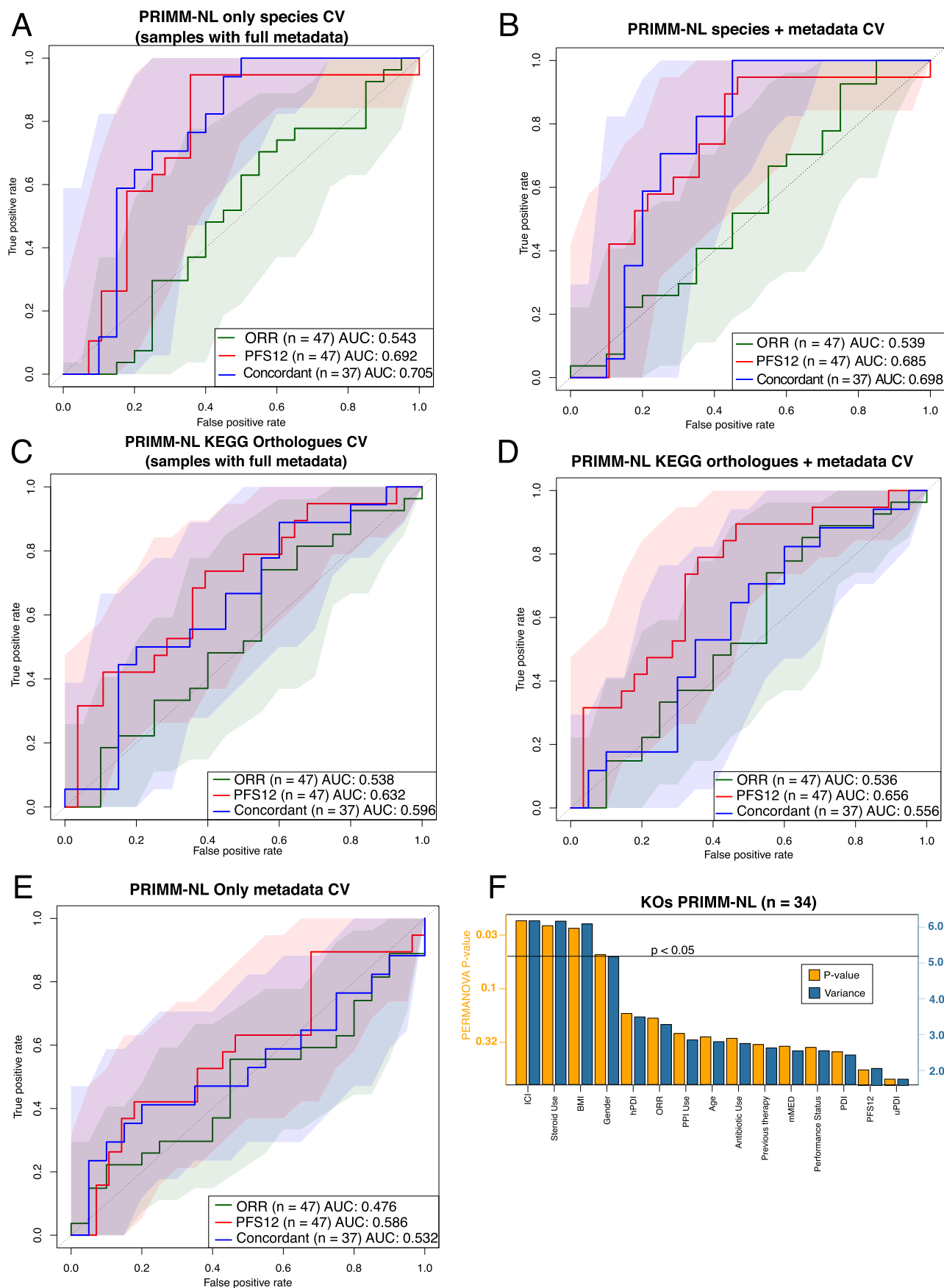
Correspondence and requests for materials should be addressed to Tim D. Spector, Rinse K. Weersma or Nicola Segata.

Peer review information *Nature Medicine* thanks David Liu, Ami Bhatt and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Javier Carmona and Saheli Sadanand are the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

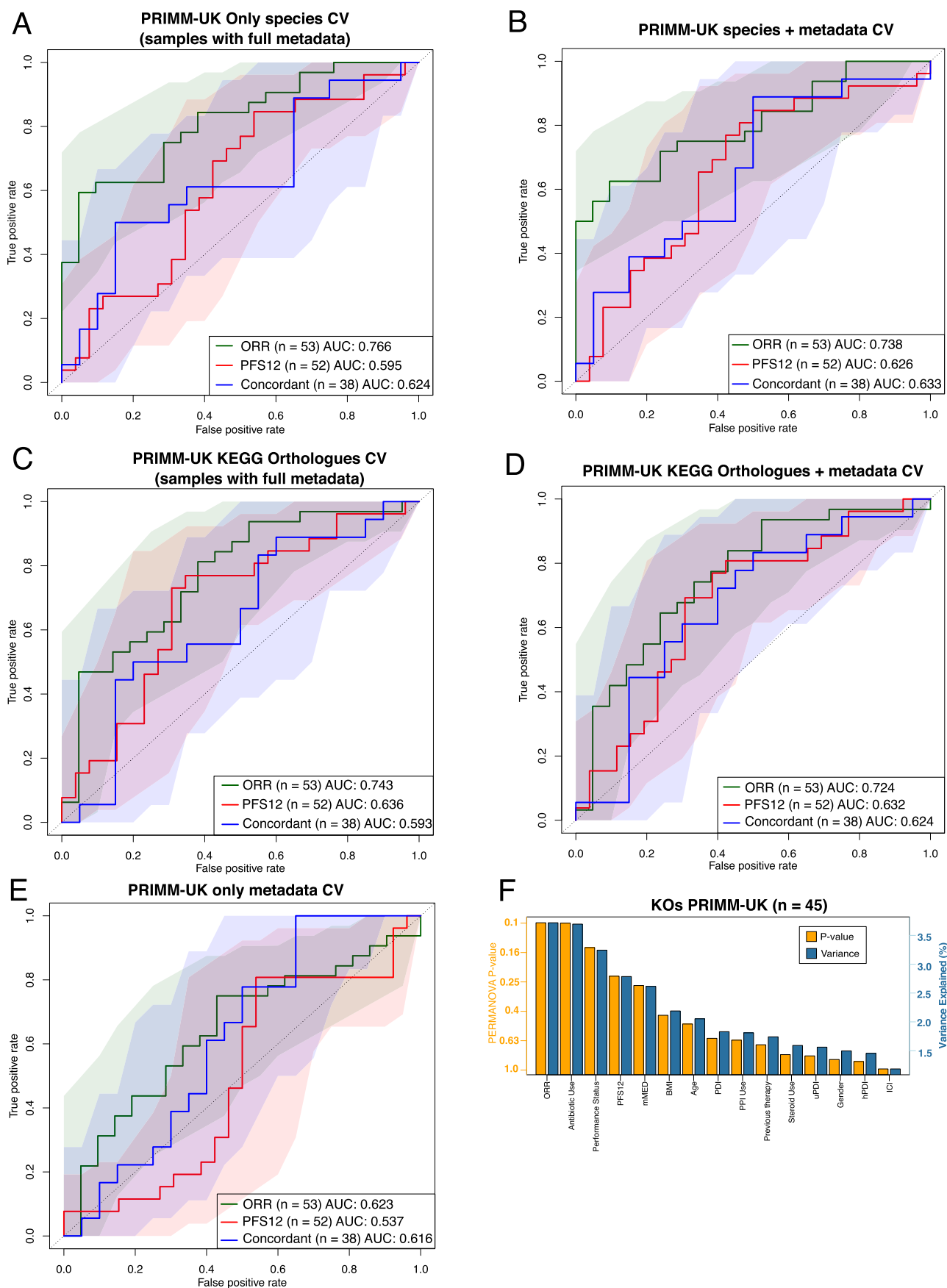


Extended Data Fig. 1 | Associations between alpha diversity and response. (a) Alpha diversity measures in the two PRIMM cohorts. # p refers to P values calculated using limma linear models including PPI, antibiotic and steroid use, gender, performance status, previous therapy, age, plant-based diet index, unhealthy plant-based diet index, mediterranean diet score and ICI in the model. \$ p refers to p values calculated using the Wilcoxon Rank-Sum Test. The lower and upper hinges of boxplots correspond to the 25th and 75th percentiles, respectively. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than $\times 1.5$ interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles. Meta-analysis results obtained using standardized mean differences between responders and nonresponders using (b) Shannon diversity and (c) species richness. The centre of the error bars (grey shaded squares) represent the standardized mean difference for each cohort.



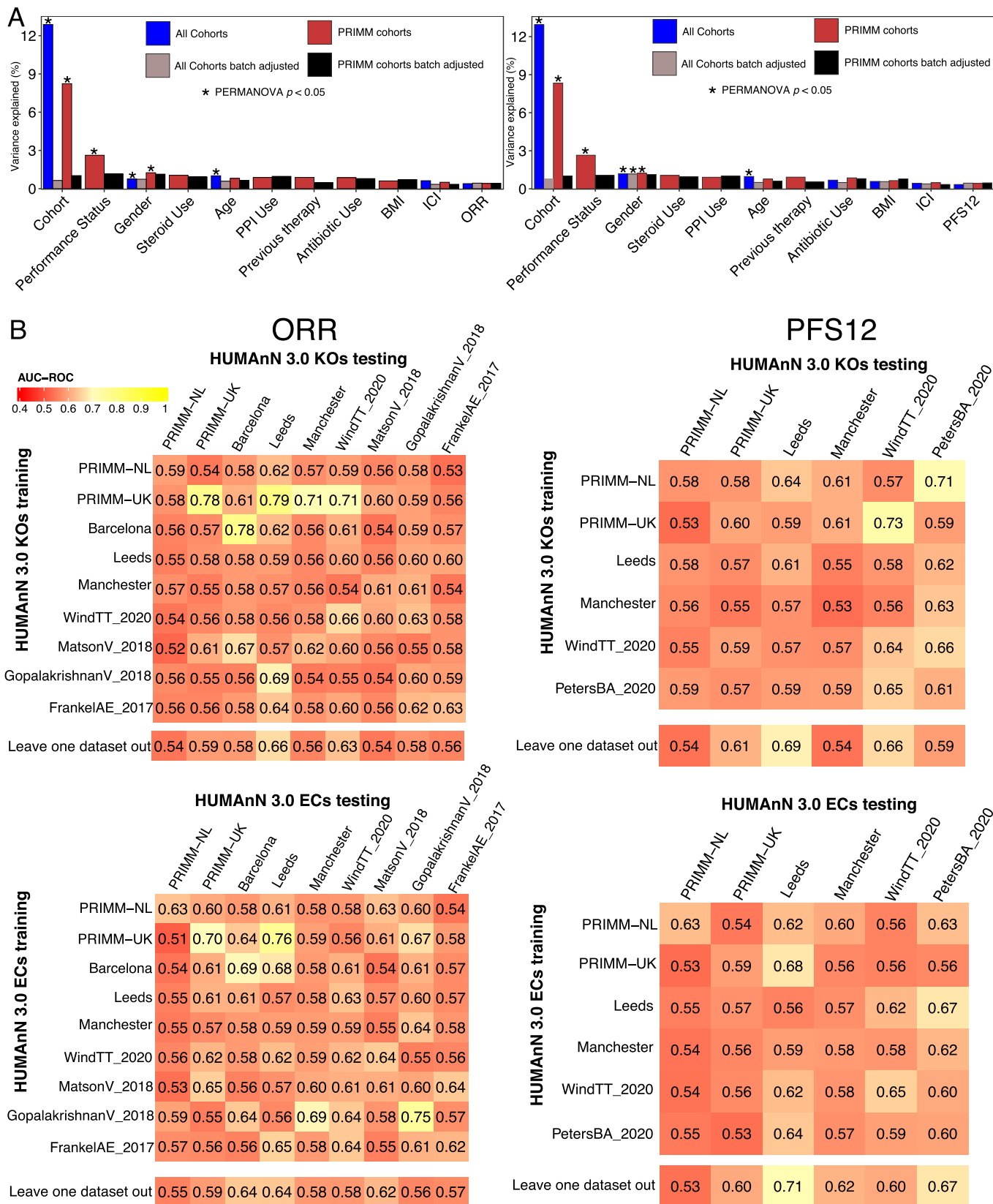
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Machine learning association analysis of the microbiome and response in PRIMM-NL. (a–e) Machine learning association analysis of the microbiome and response using either metadata alone or in combination with taxonomic (species abundance) and functional profiles (KEGG orthologs' abundances) in PRIMM-NL. AUC-ROC curves are computed using LASSO models trained using 100-repeated fivefold-stratified cross-validations. Metadata used in the models included: age, gender, performance status, PPI use, antibiotic use, steroid use, ICI and previous therapy. Shaded areas represent AUC-ROCs from each individual machine learning model. (f) Multivariate analysis showing the amount of inferred variance explained (R^2 , blue vertical bars) by each identified covariate and their respective p value (orange vertical bars) as determined by PERMANOVA on KEGG clr-transformed relative abundances.



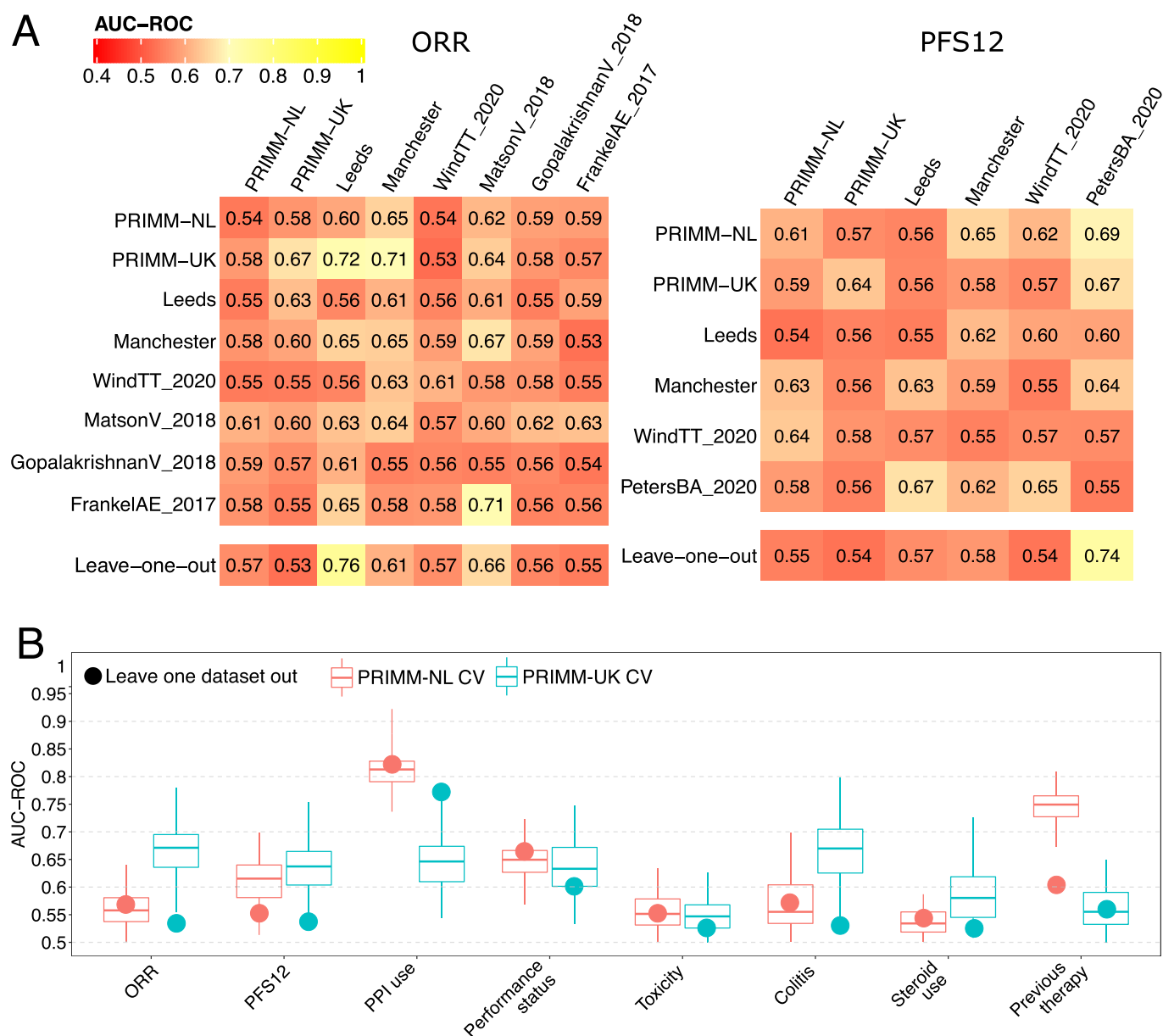
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Machine learning association analysis of the microbiome and response in PRIMM-UK. (a–e) Machine learning association analysis of the microbiome and response using either metadata alone or in combination with taxonomic (species abundance) and functional profiles (KEGG orthologs' abundances) in PRIMM-UK. AUC-ROC curves are computed using LASSO models trained using 100-repeated fivefold stratified cross-validations. Metadata used in the models included: age, gender, performance status, PPI use, antibiotic use, steroid use, ICI and previous therapy. Shaded areas represent AUC-ROCs from each individual machine learning model. (f) Multivariate analysis showing the amount of inferred variance explained (R^2 , blue vertical bars) by each identified covariate and their respective p value (orange vertical bars) as determined by PERMANOVA on KEGG clr-transformed relative abundances.



Extended Data Fig. 4 | See next page for caption.

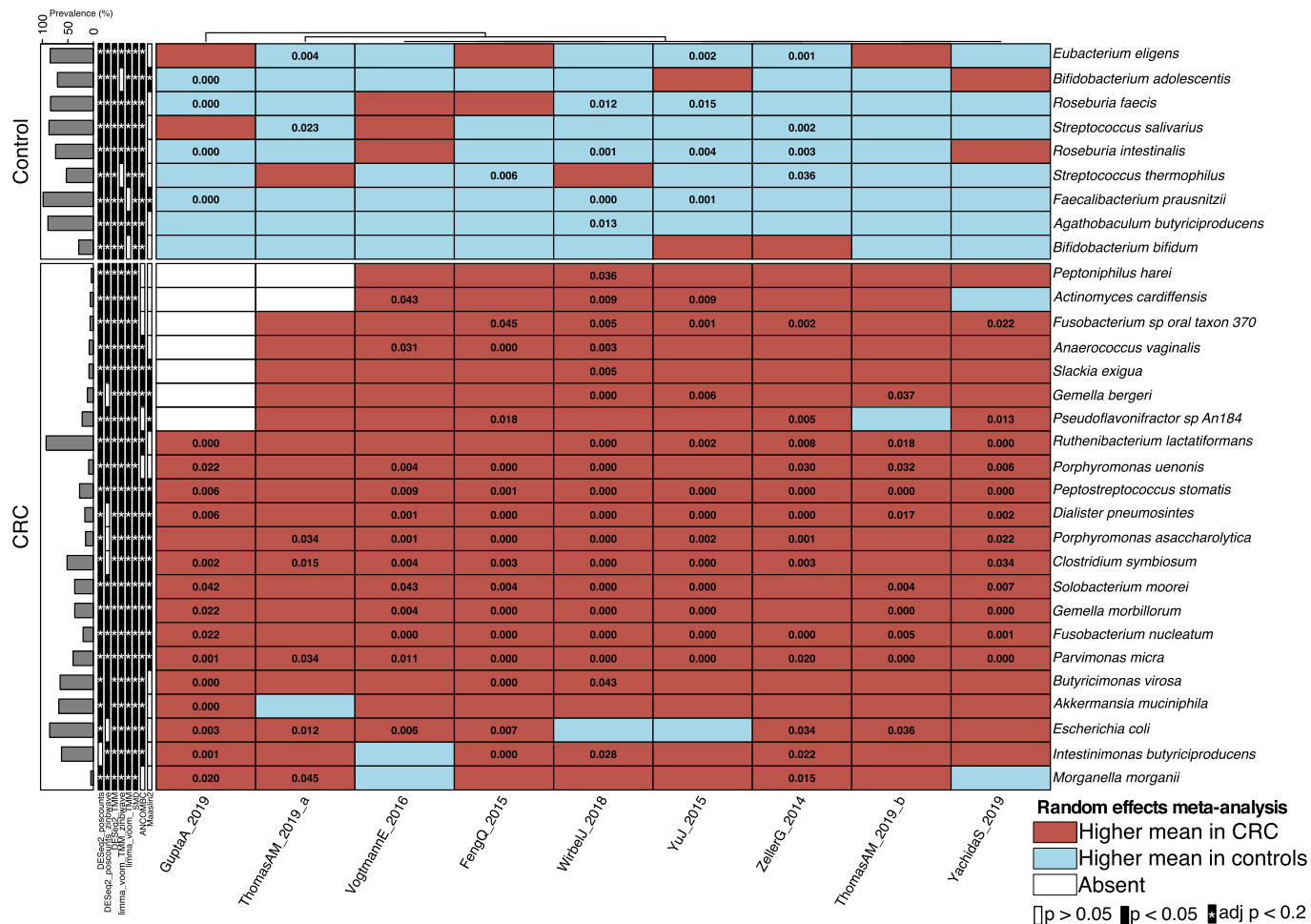
Extended Data Fig. 4 | Cross-cohort response-microbiome associations at the functional level. (a) Contribution of variables to the overall microbial community composition. Batch-correction methods were applied to KEGG abundances prior to distance calculations. The plot on the left uses ORR as the outcome variable, whereas the plot on the right adopts PFS12. **(b)** Prediction matrix for microbiome-based prediction of response assessed via ORR (left matrix) and PFS12 (right matrix) within each single cohort (values on the diagonal), across pairs of cohorts (one cohort used to train the model and the other for testing), and in leave-one-cohort-out setting (training the model on all but one cohort and testing on the left-out cohort). We report the AUC-ROC values obtained from LASSO models on KEGG relative abundances (top) and level 4 enzyme categories (bottom). Values on the diagonal refer to the median AUC-ROC values of 100-repeated fivefold stratified cross-validations. Off-diagonal values refer to AUC-ROC values obtained by training the classifier on the cohort of the corresponding row and applying it to the cohort of the corresponding column. The leave-one-out row refers to the performances obtained by training the model using all but the cohort of the corresponding column and applying it to the cohort of the corresponding column.



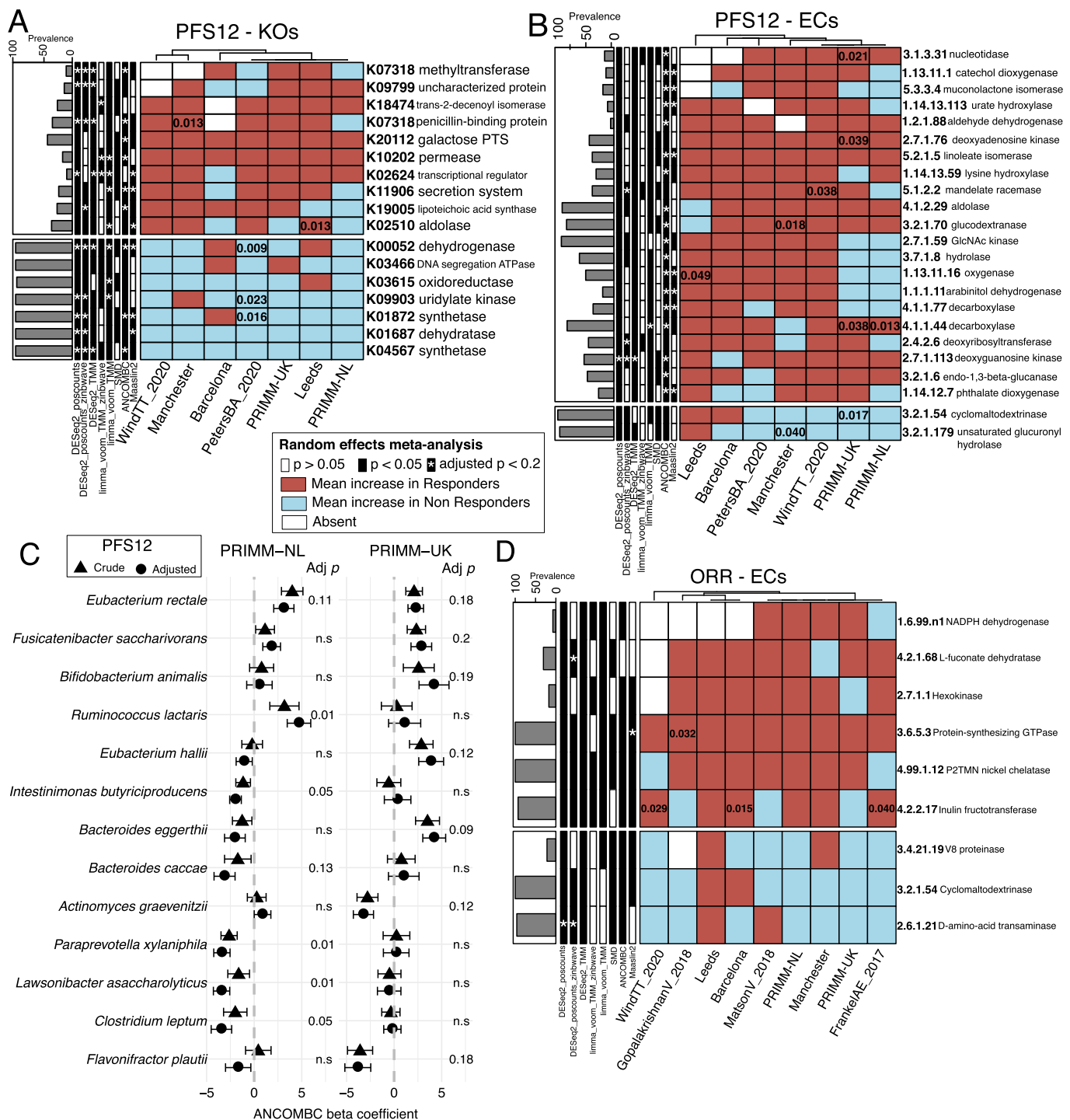
Extended Data Fig. 5 | Machine learning association analysis using random forest. (a) Prediction matrix for microbiome-based prediction of response assessed via ORR (left matrix) and PFS12 (right matrix) within each single cohort (values on the diagonal), across pairs of cohorts (one cohort used to train the model and the other for testing), and in leave-one-cohort-out setting (training the model on all but one cohort and testing on the left-out cohort). We report the AUC-ROC values obtained from Random Forest models on species-level relative abundances. Values on the diagonal refer to the median AUC-ROC values of 100-repeated fivefold stratified cross-validations. Off-diagonal values refer to AUC-ROC values obtained by training the classifier on the cohort of the corresponding row and applying it to the cohort of the corresponding column. The leave-one-out row refers to the performances obtained by training the model using all but the cohort of the corresponding column and applying it to the cohort of the corresponding column. (b) Cross-validation AUC-ROC values obtained from Random Forest models trained using 100-repeated fivefold stratified cross-validations (boxplots) and leave-one-dataset-out AUC-ROC values from Random Forest models obtained by training the model using species-level relative abundances and all but the corresponding PRIMM cohort (circles). PRIMM-NL ($n = 55$) and PRIMM-UK ($n = 55$). The lower and upper hinges of boxplots correspond to the 25th and 75th percentiles, respectively. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than $\times 1.5$ interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles.



Extended Data Fig. 6 | Taxonomic overview of species associations with response. Cladogram showing species associated with Responders (red) and nonresponders (blue) using ORR and identified by a minimum of 2 meta-analysis methods. The height of the outer bar plots reflects the number of meta-analysis methods supporting the association.

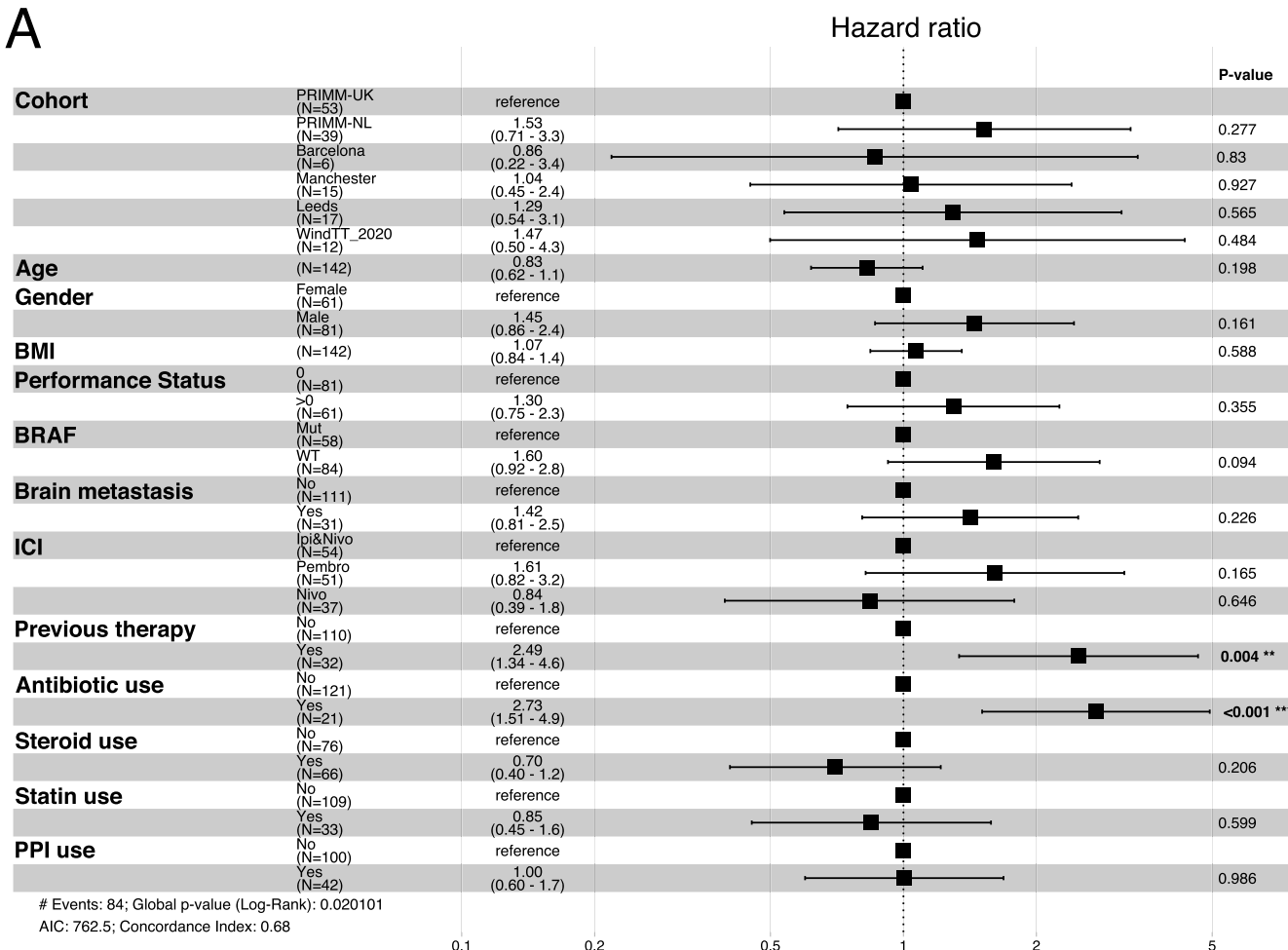


Extended Data Fig. 7 | Reproducible biomarkers for colorectal cancer across cohorts. Species associated with CRC identified by a meta-analysis using different differential abundance methods. Species shown have random-effects model p values < 0.05 in at least 6 methods out of 8 methods. Values inside the cells refer to unadjusted p values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of patients with CRC and controls.

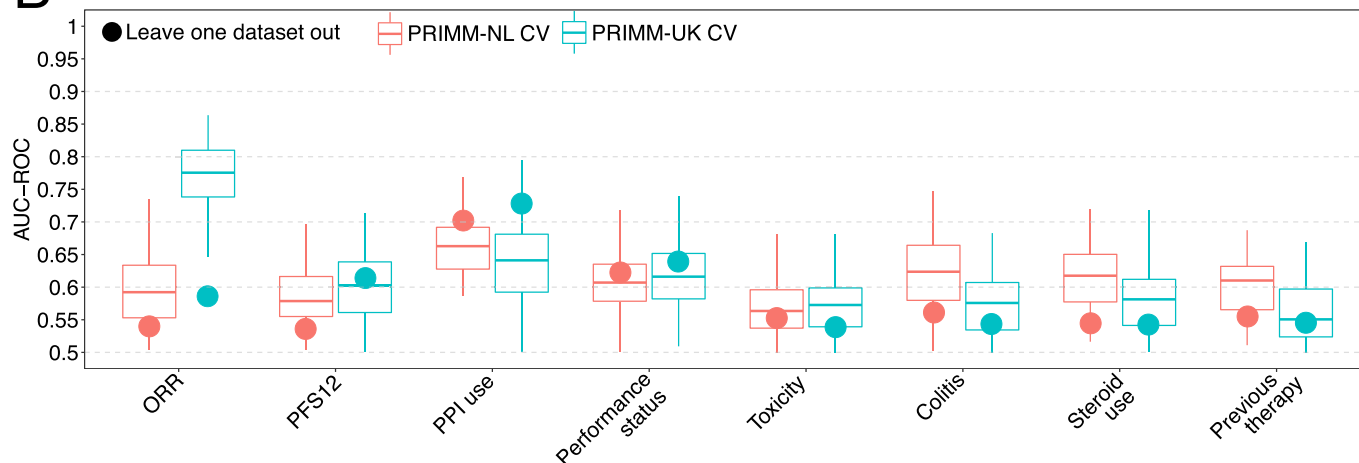


Extended Data Fig. 8 | Microbiome biomarkers of response across cohorts. (a) KEGG orthologues associated with PFS12 identified by a meta-analysis using different differential abundance methods. KEGGs shown have random-effects model p values < 0.05 in at least 6 methods out of 8 methods. Values inside the cells refer to unadjusted p values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of responders and nonresponders. (b) Level 4 enzyme categories associated with PFS12 identified by a meta-analysis using different differential abundance methods. ECs shown have random-effects model p values < 0.05 in at least 6 methods out of 8 methods. Values inside the cells refer to unadjusted p values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of responders and nonresponders. (c) Species associated with PFS12 in the two PRIMM-cohorts before and after adjusting for confounders that included PPI, antibiotic and steroid use, gender, performance status, previous therapy, age and ICI. PRIMM-NL (n = 47) and PRIMM-UK (n = 52). Species shown have covariate-adjusted multiple hypothesis testing-corrected q < 0.2 in one of the cohorts identified by ANCOM-BC. Symbols (circles and triangles) show the ANCOM-BC beta coefficient and error lines represent the standard error. (d) Level 4 enzyme categories associated with ORR identified by a meta-analysis using different differential abundance methods. ECs shown have random-effects model p values < 0.05 in at least 6 methods out of 8 methods. Values inside the cells refer to unadjusted p values < 0.05 obtained by two-tailed Wilcoxon tests on differences in the relative abundance of responders and nonresponders.

A



B



Extended Data Fig. 9 | Clinical parameters associated with response and the microbiome. (a) Forest plot showing Cox logistic regression multivariate analysis of progression-free survival. Error lines represent the 95% confidence interval of the hazard ratio. **(b)** Cross-validation AUC-ROC values obtained from LASSO models trained using 100-repeated fivefold stratified cross-validations (boxplots) and leave-one-dataset-out AUC-ROC values from LASSO models obtained by training the model using KEGG relative abundances and all but the corresponding PRIMM cohort (circles). PRIMM-NL (n=55) and PRIMM-UK (n=55). The lower and upper hinges of boxplots correspond to the 25th and 75th percentiles, respectively. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than $\times 1.5$ interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles.

Reporting Summary

Nature Research 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 Research 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 Data from questionnaires, clinical visits and laboratory data was entered using comma delimited files and excel spreadsheets.

Data analysis

Bowtie 2 (v 2.3.4.3)
 MetaPhlAn v.3.0 (<https://github.com/biobakery/MetaPhlAn>)
 HUMAnN v.3.0 (<https://github.com/biobakery/humann>)
 vegan (v2.5-7) R package
 Surrogate Variable Analysis (v3.38) R package
 ANCOM-BC (v.1.0.1) R package
 MetaVolcanoR (v.1.4.0) R package
 DESeq2 (v.1.30.0) R package
 Zinbwave (v.1.12.0) R package
 limma (v3.46.0) R package
 edgeR (v.3.32.0) R package
 Maaslin2 (v.1.4.0) R package
 SIAMCAT (v.1.6.0) R package
 Survival (v.3.2-7) R package
 CMSeq (<https://github.com/SegataLab/cmseq>)

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 Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The metagenomes and main metadata relevant to the analyses are deposited in the European Nucleotide Archive under accession no. PRJEB43119. The 4 publicly available datasets were downloaded through the Sequence Read Archive using the accession numbers SRP197281, ERP104610, SRP116709 and SRP115355. All MetaPhlAn 3 and HUMAnN 3 profiles are available within the latest version of curatedMetagenomicData (<https://bioconductor.org/packages/curatedMetagenomicData/>).

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 PRIMM studies are two separate prospective observational cohort studies recruiting subjects in parallel in the U.K. (PRIMM-UK) and Netherlands (PRIMM-NL) since 2018. To study the role of the gut microbiome in ICI response, we performed shotgun metagenomic sequencing of baseline stool samples from the PRIMM cohorts, as well as three additional cohorts of ICI-naïve patients with advanced cutaneous melanoma (originating from Barcelona, Leeds, and Manchester). A total of 165 microbiome samples sequenced from the patients enrolled in this study, were analyzed together with 147 samples from smaller publicly-available datasets. This provided the largest possible assessment of the potential of the gut microbiome as a biomarker of response to ICI, and allowed identification of specific microbial species or functions associated with response.
Data exclusions	We excluded samples of participants with non-metastasized and resectable Stage III melanoma who received ICI's as adjuvant treatment. Moreover patients who were not immunotherapy-naïve were excluded.
Replication	We performed a cross-cohort meta-analysis, adding three smaller cohorts and five previously published datasets with consistent metagenomic and response data to the PRIMM cohorts. The analyses showed limited reproducibility of microbiome-based response predictions across these cohorts as outlined in detail in the results and discussion (page 6 ff; page 15 ff). The limited reproducibility may in part result from heterogeneity across studies as we found a strong cohort-dependent effect in the microbial population. It is likely not attributable to analytical choices, as applying the same meta-analysis methods in the context of colorectal cancer confirmed the strong and consistent biomarkers across cohorts (Thomas et al. Nat. Med. 2019). Our study shows that the role of the gut microbiome for ICI-response is more complex than previously thought and extends beyond absence or presence of microbial species or functions in responders versus non-responders.
Randomization	n/a
Blinding	There was no control or placebo arm therefore blinding was not applicable

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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Covariates adjusted for in the analyses included PPI, antibiotic and steroid use, gender, performance status, previous therapy, age, ICI agent (combination Ipilimumab/Nivolumab or single-agent). There were two statistically relevant clinical differences between the 2 prospective PRIMM-cohorts: the proportion of subjects who had received previous systemic targeted therapy (40% within PRIMM-NL and 20% within PRIMM-UK, $P=0.03$, Fisher's exact test) and the proportion with BRAF-mutated tumors (55% within PRIMM-NL and 31% within PRIMM-UK, $P=0.02$, Fisher's exact test, Table 1).
Recruitment	The PRIMM studies are two separate prospective observational cohort studies recruiting subjects in parallel in the U.K. (PRIMM-UK) and Netherlands (PRIMM-NL) since 2018. Patients who fulfilled the following criteria were eligible for the analyses: (i) histologically or cytologically confirmed non resectable advanced (stage III or IV) cutaneous melanoma (ii) treatment with ICI (nivolumab, pembrolizumab, ipilimumab or a combination of ipilimumab and nivolumab) at recommended dose as first-line ICI, (iii) 18 year of age or older. High quality fecal samples were collected from these patients before initiation of ICI ($n=55$ for UK cohort, $n=55$ for Dutch cohort). Additional patients were enrolled from cohorts outside the setting of a prospective clinical trial and performed in Leeds ($n=19$), Barcelona ($n=12$) and in Manchester ($n=30$) between March 2015 and November 2019, but from whom fecal samples were collected at similar timepoints to those collected in our included prospective studies. Written informed consent was obtained from all patients.
Ethics oversight	King's College London (KCL); Medical Ethical Committee of the University Medical Center Groningen (METc UMCG); Manchester Cancer Research Centre (MCRC) Biobank Ethics and MCRC Biobank Access Committee; Ethical committee of Hospital Clinic of Barcelona.

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

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	NCT03643289; NCT04193956; https://www.trialregister.nl/trial/7839 ; MCRC 07/H1003/161+5 and MCRC 13_RIMA_01; HCB/2015/1032; REC Ref 15/NW/0933.
Study protocol	https://www.clinicaltrials.gov
Data collection	The PRIMM studies are two separate prospective observational cohort studies recruiting subjects in parallel in the U.K. (PRIMM-UK) and Netherlands (PRIMM-NL) since 2018. Additional patients were enrolled from cohorts outside the setting of a prospective clinical trial and performed in Leeds ($n=19$), Barcelona ($n=12$) and in Manchester ($n=30$) between March 2015 and November 2019, but from whom fecal samples were collected at similar time points to those collected in our included prospective studies.
Outcomes	<p>Clinical endpoints were defined as objective response rate (ORR) and progression free survival (PFS) at 6 and 12 months.</p> <p>Response to ICI was classified according to RECIST v1.1 criteria. On the basis of radiographic response, patients were classified as Responders (CR, PR, or SD) or Non-responders (PD). In order to include late responders in our analysis, patients with progressive disease (PD) on the first radiological evaluation but a response at the second radiological evaluation compared to baseline were also labelled responders. Patients with PD on the first radiological evaluation that was confirmed on the next follow-up scan, or patients with PD on the first radiological evaluation that were unable to complete a confirmation scan due to clinical progression or death were labelled non-responders.</p> <p>PFS was defined as the time from first dose of ICI to first event i.e. disease progression or death from any cause.</p>