Supplementary Materials: Intestinal Microbiota Composition is Predictive of Radiotherapy-induced Acute Gastrointestinal Toxicity in Prostate Cancer Patients.

Authors:

Jacopo Iacovacci[1,§,*], Mara Serena Serafini[2,§], Barbara Avuzzi[3], Fabio Badenchini[4], Alessandro Cicchetti[1], Andrea Devecchi[2], Michela Dispinzieri[3], Valentina Doldi[5], Tommaso Giandini[6], Eliana Gioscio[1], Elisa Mancinelli[2], Barbara Noris Chiorda[3], Ester Orlandi[7], Federica Palorini[4], Luca Possenti[1], Miguel Reis Ferreira[8,9], Sergio Villa[3], Nadia Zaffaroni[5], Loris De Cecco[2,†], Riccardo Valdagni[4,3,10,†] and Tiziana Rancati[1,†]

* Correspondence: jacopo.iacovacci@istitutotumori.mi.it
§ Contributed equally
† Jointly supervised

Affiliations:

[1] Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy
[2] Unit of Experimental Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy
[3] Unit of Radiation Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy
[4] Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy
[5] Unit of Molecular Pharmacology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy
[6] Unit of Medical Physics, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy
[7] Radiation Oncology Clinical Department, National Center for Oncological Hadron Therapy (CNAO), Pavia, Italy
[8] King's College London, London, UK
[9] Guys and St Thomas NHS Foundation Trust, London, UK
[10] Department of Oncology and Hematology-Oncology, Università degli Studi di Milano, Milan Italy

Patient population

Supplementary Table S1 | MicroLearner PCa: patient population characteristics

Characteristic	Pooled cohort (n=215)	Discovery cohort (n=136)	Validation cohort (n=79)
Age (years): median (range)	71 (46-84)	70 (46-84)	71 (53-81)
Race: Caucasian	100%	100%	100%
BMI: median (range); missing	25.77 (17.96-47.60); 1	25.71 (17.96-47.60); 1	26.30 (19.92-37.5); 0
Diabetes: number (percentage)	24 (11.2%)	14 (10.3%)	10 (12.7%)
Diverticulitis: number (percentage)	12 (5.6%)	5 (3.7%)	7 (8.9%)
Bowel infarction: number (percentage)	3 (1.4%)	1 (0.7%)	2 (2.5%)
Cardiac condition: number (percentage)	43 (20%)	30 (22.1%)	13 (16.5%)
Autoimmune condition: number (percentage)	2 (0.9%)	2 (1.5%)	0 (0%)
Hypertension: number (percentage)	120 (55.8%)	74 (54.4%)	46 (58.2%)
PCa risk group*: number (percentage)			
Low	27 (12.6%)	19 (14%)	8 (10.1%)
Intermediate	84 (39.1%)	52 (38.2%)	32 (40.5%)
High	65 (30.2%)	41 (30.1%)	24 (30.4%)
Very high	5 (2.3%)	2 (1.5%)	3 (3.8%)
Metastatic	12 (5.6%)	9 (6.6%)	3 (3.8%)
Post-op	22 (10.2%)	13 (9.6%)	9 (11.4%)

* NCCN risk classification

Acronyms
BMI: Body Mass Index
PCa: Prostate Cancer
NCCN: National Comprehensive Cancer Network

Radiotherapy treatment

All patients received Volumetric Modulated Arc RT (VMAT). One hundred and forty-five patients received conventionally fractionated RT. Fifty-seven were post-prostatectomy patients (adjuvant setting 11 patients, salvage 46 patients) treated with a prescribed dose of 70-72 Gy. Eighty-eight radical patients had a prescribed dose of 74-78 Gy. Seventy radical patients received moderately hypofractionated RT (2.6 Gy/day) with a prescribed dose of 65-67.6 Gy. One hundred and seventy-two patients received whole pelvis RT (50 Gy) and 142 irradiation of seminal vesicles (66-70 Gy). Treatment details are reported in Supplementary Table S2.

Supplementary Table S2| MicroLearner PCa: patient treatment characteristics

Treatment factor	Pooled cohort (n=215)	Discovery cohort (n=136)	Validation cohort (n=79)
Radiotherapy: number (percentage)			
WPRT	172 (80%)	108 (79.4%)	64 (81%)
PORT	43 (20%)	28 (20.6%)	15 (19%)
Adjuvant	11 (5.1%)	8 (5.9%)	3 (3.8%)
Radical	151 (70.2%)	92 (67.6%)	59 (74.7%)
Salvage	47 (21.9%)	33 (24.3%)	14 (17.7%)
Oligometastatic	6 (2.8%)	3 (2.2%)	3 (3.8%)
SV irradiation	142 (66%)	85 (62.5%)	57 (72.2%)
HT neo/adjuvant: number (percentage)			
Antiandrogen	92 (42.8%)	60 (44.1%)	32 (40.5%)
LHRH	77 (35.8%)	47 (34.6%)	30 (38%)
BAT	3 (1.4%)	2 (1.5%)	1 (1.3%)
Antibiotics*: number (percentage)	1 (0.5%)	1 (0.7%)	0 (0%)
Appendectomy: number (percentage)	41 (19.1%)	26 (19.1%)	15 (19%)
Cholecystectomy: number (percentage)	13 (6%)	7 (5%)	6 (7.6%)
Prostatectomy: number (percentage)	64 (29.8%)	47 (34.6%)	14 (17.7%)

* taken less than 3 months previous to radiotherapy initiation

Acronyms
WPRT: Whole-pelvis radiotherapy
PORT: Prostate-only radiotherapy
SV: Seminal vesicles
HT: Hormone therapy

Diet and lifestyle factors analysis

Patients completed a questionnaire on life style with information on diet habit and physical activity before starting their treatment. A template of the diet and physical activity questionnaire, including the food frequency questionnaire (FFQ) section is reported as Supplementary Table 3. Specific food classes from FFQ were aggregated in the analysis of the diet composition (Supplementary Fig. 1). Supplementary Table 4 reports the statistic of diet and physical activity categories in the patient population, while Supplementary Table S5 reports the criteria used to define the categories.

Supplementary Table S3| Questionnaire for the evaluation of lifestyle and diet habit

Physical activity					
Do you practice regularly any physical activity? (e.g., gym exercise, walk, swimming, …)	• Yes • No				
If yes, how frequently?	• Less than once a week • Once a week • More than once a week				
Dietary habits					
Do you follow any specific diet? (e.g., vegetarian, vegan, zone, crudism, …)	• Yes • No				
If yes, specify					
Are you gluten intolerant?	• Yes • No				
Do you suffer any diet-related allergy or any other food intolerance?	• Yes • No				
If yes, specify					
Frequency of consumption of classes of food (for each food category, fill only one box indicating the number of times you consume the corresponding food)					
	times				
How often do you consume	year	month	week	day	never
Cereals and cereal-derived food? (e.g., pasta, rise, bread, pizza, spelt, barley, etc.)					
Cereal-based products? (e.g., cornflakes, biscuits, rusks, crackers, etc.)					
Fresh meat?					
Processed meat? (e.g., ham, salami, wurstel, etc.)					
Fish and other fisheries? (e.g., sea bass, octopus, shrimps)					
Milk and yogurt?					
Milk-derived products? (e.g., fresh or aged cheese)					
Vegetables? (e.g., lettuce, radish, spinach, Swiss chard, etc.)					
Legumes? (e.g., beans, lentils, peas, chickpeas)					
Eggs?					
Desserts? (e.g., cakes, ice creams, chocolate, etc.)					
Sweet drinks? (e.g., coke, sprite, sparkling orange juice, etc.)					
Alcohol? (e.g., wine, beer, liquors, etc.)					

Supplementary Table S4| Statistic of lifestyle factors in the patient population

Lifestyle factor	MicroLearner PCa (n=215)
Physical activity:	
number (percentage)	
Physically active	63 (29.3%)
Sedentary	117 (54.4%)
Undeclared	35 (16.3%)
Smoke:	
number (percentage)	
Smoker	104 (48.4%)
Non-smoker	111 (51.6%)
Diet:	
number (percentage)	
No diet regime	184 (85.6%)
Vegetarian*	7 (3.3%)
Undeclared	13 (6%)
Unspecified/other	11 (5.1%)
Intake of animal proteins & fats:	
number (percentage)	
High	21 (9.8%)
Not high	114 (53%)
No FFQ	80 (37.2%)
Intake of plants proteins & fibres:	
number (percentage)	
High	30 (14%)
Not high	105 (48.8%)
No FFQ	80 (37.2%)

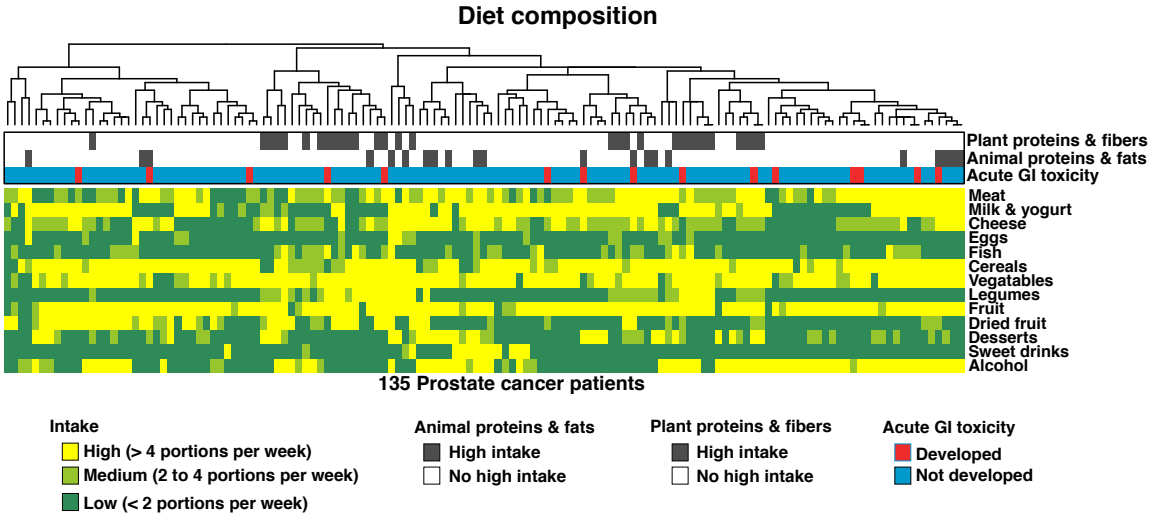
* includes pescatarian and vegan

Acronyms
PCa: Prostate Cancer
FFQ: Food Frequency Questionnaire

Supplementary Table S5| Rules used to define lifestyle-related factors from questionnaires

Lifestyle class	Rule
Physical activity:	
Physically active	physical exercise > once p.w.
Sedentary	physical exercise <= once p.w.
Smoke:	
Smoker	if had any smoking history
Non-smoker	if never smoked
Intake of animal proteins & fats:	
High	>4 p.p.w. of meat & >4 p.p.w. of cheese
Not high	other cases
Intake of plant proteins & fibres:	
High	>2 p.p.w. of legumes & >2 p.p.w. of vegetables & >2 p.p.w. of fruit & not in the **High intake of animal proteins & fats** group
Not high	other cases

Acronyms
p.w.: per week
p.p.w.: portions per week



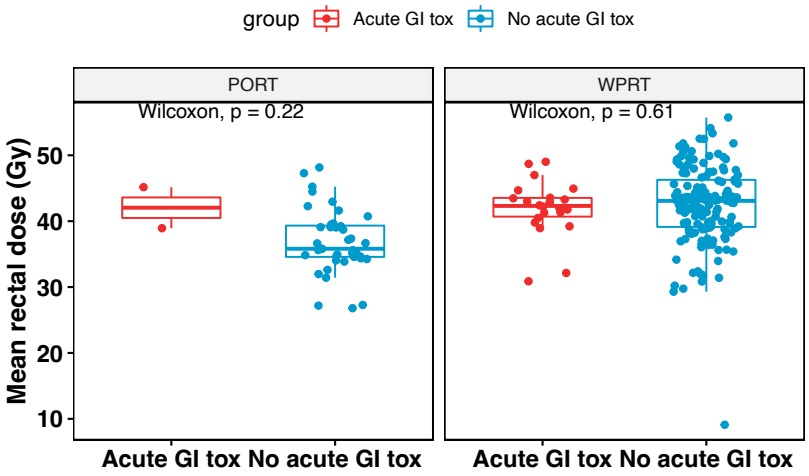
Supplementary Fig. S1 | Analysis of diet composition from patient-compiled food frequency questionnaires (FFQs)
Hierarchical clustering of patients (heatmap columns) based on per week intake of food categories (heatmap rows) using Hamming distance and Ward linkage after defining integer-valued intake scores (high-intake=2, medium-intake=1, low-intake=0); food categories "Cereals" and "Meat" include aggregated portions per week (sum) of "Cereals and cereal-derived food" and "Cereal-based products", and of "Fresh meat" and "Processed meat", respectively; "Milk-derived products" from FFQs was labelled "Cheese" for simplicity.

Radio-induced gastrointestinal toxicity

Supplementary Table S6| The statistic of GI toxicity grade by symptom assessed in the patient population (n=215) according to the CTCAE

	Any symptom	Proctitis	Diarrhoea	Flatulence	Rectal bleeding	Faecal incontinence
Maximum grade number (% pts)						
Grade 0	46 (21.4%)	109 (50.7%)	116 (54%)	114 (53%)	194 (90.2%)	213 (99%)
Grade 1	119 (55.3%)	76 (35.3%)	80 (37.2%)	96 (44.7%)	20 (9.3%)	1 (0.5%)
Grade 2	49 (22.8%)	30 (14%)	18 (8.4%)	5 (2.3%)	1 (0.5%)	1 (0.5%)
Grade 3	1 (0.5%)	0 (0%)	1 (0.5%)	NA	0 (0%)	0 (0%)

Acronyms
pts: patients



Supplementary Fig. S2 | Analysis of whole-pelvis versus prostate-only radiotherapy
Distribution of mean rectal dose in the toxicity groups stratified in the sub-populations of patients who received prostate only radiotherapy (PORT, n=43) and in whole-pelvis radiotherapy (WPRT, n=172). Despite the whole pelvis can be considered an additional risk factor for toxicity as patients receive more dose to the rectum (Wilcoxon's rank test p <0.001), no association between mean rectal dose and acute GI toxicity was observed in the separate WPRT and PORT sub-populations, respectively.

Supplementary Table S7| Association of potential risk/beneficial factors or drugs with the onset of acute gastrointestinal toxicity during radiotherapy

Potential risk factor	Factor size	Toxicity counts	Toxicity rate factor=TRUE	Toxicity rate factor=FALSE	Fisher's exact test (p greater)
Appendectomy	41	6	14.6%	10.3%	0.29
Cholecystectomy	13	2	15.4%	10.9%	0.44
Prostatectomy	64	6	9.4%	11.9%	0.78
Smoke	104	13	12.5%	9.9%	0.35
High Intake of animal proteins & fats	21	4	19%	9.6%	0.18
Diabetes*	24	5	20.8%	9.9%	0.11
Diverticulitis	12	2	17%	10.8%	0.4
Cardiac condition	43	3	7%	12.2%	0.9
Hypertension**	120	13	10.8%	11.6%	0.65
Potential beneficial factor	**Factor size**	**Toxicity counts**	**Toxicity rate factor=TRUE**	**Toxicity rate factor=FALSE**	**Fisher's exact test (p lower)**
Physically active	63	4	6.3%	(sedentary) 13.7%	0.1
Vegetarian diet	7	0	0%	(no diet regime) 11.4%	0.44
High intake of plants proteins & fibres	30	4	13.3%	10.5%	0.79
Drug class	**Factor size**	**Toxicity counts**	**Toxicity rate factor=TRUE**	**Toxicity rate factor=FALSE**	**Fisher's exact test (p lower)**
Anticoagulant	7	1	14.3%	11.1%	0.82
Antiplatelet	20	1	5%	11.8%	0.32
ACE inhibitor	43	2	4.7%	12.8%	0.1
Alphalitics	36	4	11%	11.2%	0.63
Beta blocker	40	5	12.5%	10.7%	0.72
Antihypertensives	60	6	10%	11.6%	0.47
Statins	58	2	3.4%	14%	**0.0194**
Oral hypoglycaemic agents	20	3	15%	10.8%	0.83
Probiotics***	10	2	20%	10.7%	0.91

* treated with oral hypoglycaemic agents; ** treated with ACE inhibitors and/or beta blockers and/or other antihypertensives; *** habitual intake of probiotic: Yogurt, 2 patients; Refluor (*Lactobacillus reuteri*), 2 patients; Enterolactis (*Lacticaseibacillus paracasei*), 1 patient; Serobioma (*Bifidobacterium lactis*, *Bifidumbacterium longum*, *Lactobacillus rhamnosus*), 1 patient; Bifidolactis (*Bifidobacterium lactis*), 1 patient; Codex (*Saccharomyces boulardii*), 1 patient;

Lactoflorene (*Lactobacillus acidophilus*, *Bifidobacterium lactis, Bacillus coagulans, Lactobacillus casei*), 1 patient; Dicoflor (*Lactobacillus rhamnosus*), 1 patient.

Cytokine measurement

Ten millilitres of EDTA blood samples (BD Vacutainer™ K2 EDTA-367525) were obtained at baseline. Samples were kept 4 °C for maximum one hour after sampling and then centrifuged for 20 minutes at 2200*g* at 4 °C. Plasma was collected (~4 mL) and centrifugated for 10 minutes at 2200 *g* at 4 °C. Supernatant was immediately stored at ≤−80 °C until analysis (stored in Nalgene CryoBox-50260909).

All analyses were carried out blind to patient and therapy factors. The amount CCL2, PDGF-BB, TGF-β1, TNF-α and TNFR1 was determined using commercially available ELISA kits (Quantikine®ELISA R&D Systems Inc., Minneapolis, MN, USA), according to manufacturer's protocols: Quantikine®ELISA Human TGF-β1 cod. DB100B R&D System, Quantikine®ELISA Human CCL2/MCP-1 cod. DCP00 R&D System, Quantikine®ELISA Human PDGF-BB cod. DBB00 R&D System, Quantikine®ELISA Human TNF-α cod. DTA00C R&D System and Quantikine®ELISA Human TNF RI cod. DRT100 R&D System.

Polycytokinic risk score

We developed a logistic stepwise backward regression model to predict the classes of "Acute GI toxicity" (Positive cases) and "No acute GI toxicity" (Negative cases) using as predictors the log-transformed values of concentration of the plasma cytokines measured. We used the *stepAIC* function from the R package MASS version 7.3-57[23] to identify the best predictors, so selecting CCL2, TNFR1 and TNF-α. The logit model based on the selected predictors was used to obtain a risk probability for each patient and the polycytokinic risk score was defined by standardising the probabilities across the patient population. Model details are reported in Supplementary Table S8.

Supplementary Table S8| Polycytokinic risk score (PRS) model

Population				
Sample size (n)	99			
Positive cases	13 (13.13%)			
Negative cases	86 (86.87%)			
Stepwise regression				
	Coefficient	**Std. Error**	**Wald Z**	**p-value**
log(CCL2)	-3.75434	2.077	3.2674	0.0707
log(TNFα)	1.45759	0.82512	3.1206	0.0773
log(TNFR1)	-2.01011	1.2036	2.7892	0.0949
Intercept	7.71938	3.99513	3.7334	0.0533
PRS model				
	Coefficient	**Std. Error**	**Wald Z**	**p-value**
PRS	0.7139	0.3009	2.37	0.0177
Intercept	-2.0381	0.3282	-6.21	<0.0001
	Odds ratio	**95% CI**	**AUC**	
PRS	2.042	[1.13 – 3.68]	0.76	
Classification (PRS cut-off=0)*				
	Predicted	**Positive (PRS>0)**	**Negative (PRS<0)**	**Accuracy**
True classes				
Positive		10	3	77%

Negative		30	56	65%
	Toxicity rate	25%	5%	
* un-normalized cut-off probability value=0.13				

Acronyms
PRS: polycytokinic risk score

Faecal sample collection and processing

Faecal specimens were collected using the OMNIgene•GUT stool devices (DNA Genotek Inc. Ottawa, ON, Canada) consisting of a tube with a preservation buffer to stabilise microbial DNA and a bearing steel bead. Patients were instructed to collect the faeces into the tube avoiding contaminations and to homogenize the sample by shaking. The samples were stored at room temperature and delivered to the centralised laboratory for metagenomics analyses.

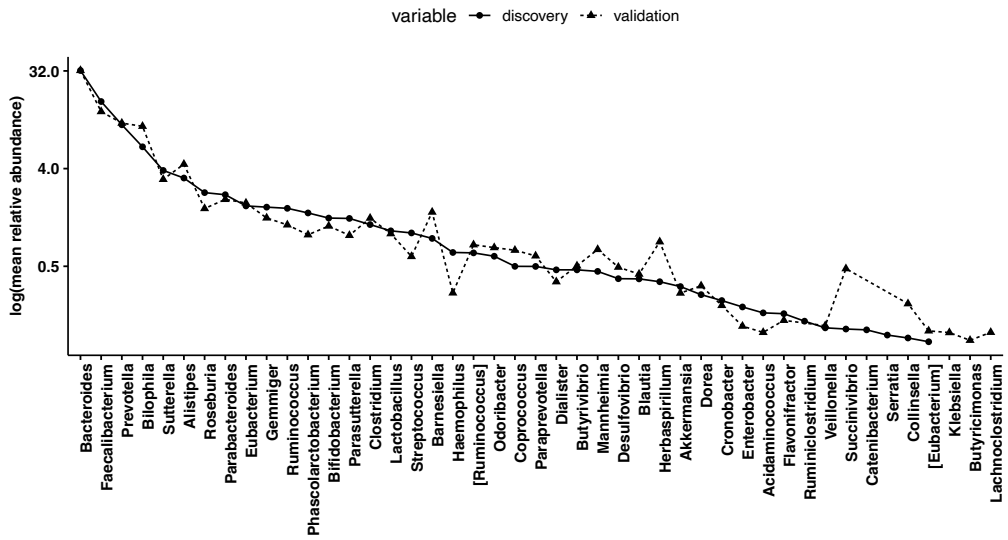
Upon arrival, DNA was extracted by QIAsymphony DSP Virus /Pathogen Midi kit (Qiagen) after mechanical lysis with silica beads on an automated QIAsymphony station (Qiagen). Microbial DNA was quantified using a Qubit fluorometer (ThermoFisher) and quality was assessed using a 4200 TapeStation (Agilent). Samples reaching good quality (DNA integrity number >7) both at RT baseline and at RT end were used for metagenomics profiling.

The NGS libraries were performed using 16S Metagenomics kit (ThermoFisher) following manufacturer's instructions. The 16S region was amplified with primer sets recognising V2, V4, V8 and V3, V6-7, V9 hypervariable regions in 2 separate PCR reactions. Fifty nanograms of amplicons were combined and processed for library prep using Ion Plus Fragment Library Kit and Ion Xpress Barcodes Adapters (ThermoFisher). After PCR amplification (1 cycle of 95°C for 5 min; 5 cycles of 95°C for 15sec, 58°C for 15 sec, 70°C for 1 min) and purification using 1.4 volumes of Agencourt AMPure beads (Beckman Coulter), libraries were eluted and their size and quantity were assessed with TapeStation.

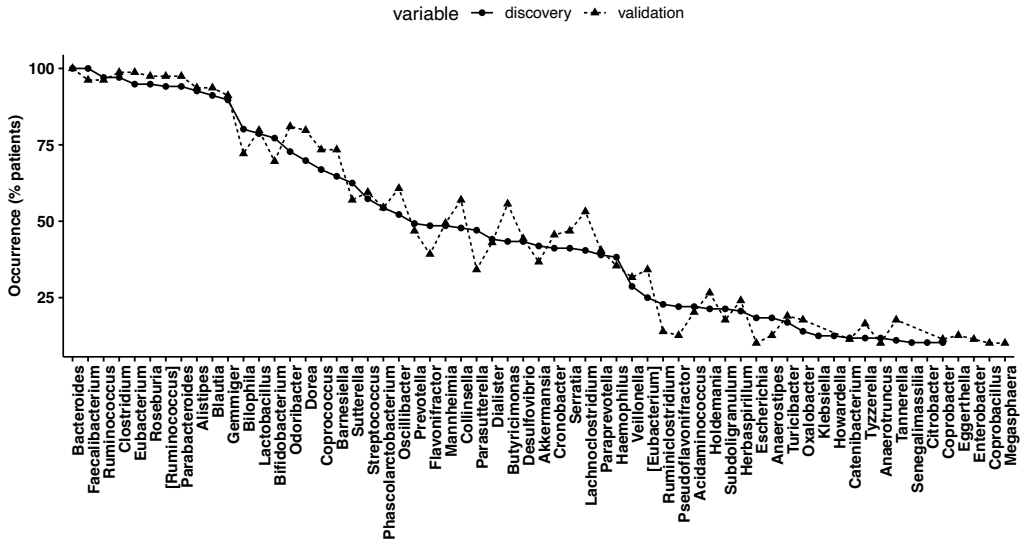
Sequencing was performed for 16S libraries by Ion S5 XL whereas base calling and demultiplexing was performed by Torrent Suite (ThermoFisher).

The ThermoFisher Ion Reporter Software metagenomics 16S analysis pipeline was used to generate operational taxonomic unit abundances from the 16S rRNA reads and to assign taxonomy at the genus level by clustering sequences at 97% similarity threshold.

Intestinal microbiota analysis



Supplementary Fig. S3 | Abundance-based ranking of intestinal microbiota genera
Binary logarithm of the population-averaged relative abundance of genera in the intestinal microbiota stratified by cohort (discovery and validation) for all the sequenced genera having average relative abundance >=1% in either of the cohorts.



Supplementary Fig. S4 | Occurrence-based ranking of intestinal microbiota genera
The fraction of patient population where genera were profiled, stratified by cohort (discovery and validation) for all the sequenced genera having occurred in at least 10% of patients in either of the cohorts.

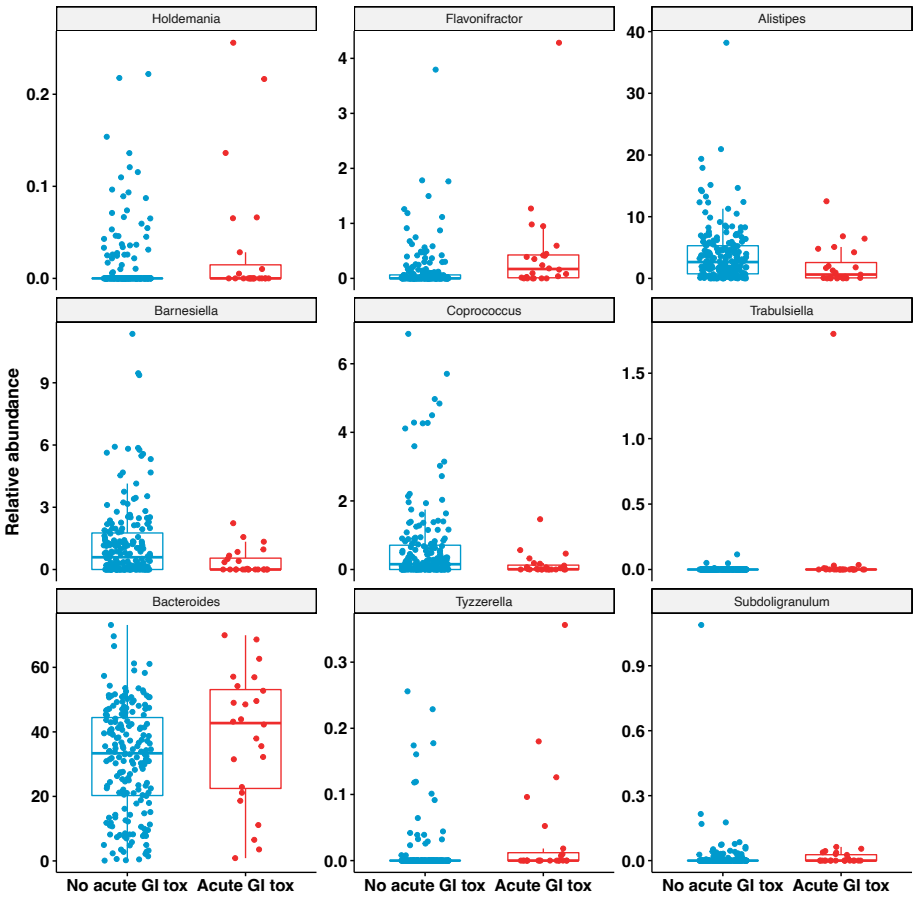
Supplementary Table S9 | Differential abundance analysis of genera between toxicity groups

variable* (cohort)	**Occurrence (pooled)**	**p (discovery)**	**adj. p (discovery)**	**p (validation)**	**adj. p (validation)**	**p (pooled)**	**adj. p (pooled)**
Genus							
Bacteroides	73.5	**0.0136**	1	0.994	1	**0.0327**	1
Eubacterium	72.6	0.795	1	0.239	1	0.766	1
Clostridium	72.6	0.223	1	0.942	1	0.246	1
[Ruminococcus]	71.6	0.733	1	0.287	1	0.787	1
Parabacteroides	71.6	0.082	1	0.814	1	0.216	1
Roseburia	71.6	0.103	1	0.801	1	0.234	1
Ruminococcus	70.7	0.591	1	1	1	0.627	1
Faecalibacterium	70.7	0.992	1	0.814	1	0.899	1
Alistipes	68.8	**0.00656**	0.878	0.655	1	**0.0183**	1
Blautia	68.8	0.564	1	0.49	1	0.932	1
Gemmiger	67	**0.0424**	1	0.814	1	0.0673	1
Odoribacter	59.5	0.821	1	0.903	1	0.67	1
Lactobacillus	58.6	0.393	1	0.17	1	0.918	1
Dorea	58.6	0.927	1	0.929	1	0.877	1
Barnesiella	54	0.0774	1	0.102	1	**0.0172**	1
Coprococcus	54	**0.0189**	1	0.5	1	**0.0205**	1
Bilophila	53	0.163	1	0.185	1	0.753	1
Bifidobacterium	51.2	0.853	1	0.239	1	0.632	1
Oscillibacter	44.7	0.648	1	0.764	1	0.483	1
Streptococcus	43.7	0.624	1	0.929	1	0.703	1
Sutterella	41.9	0.784	1	0.239	1	0.269	1
Collinsella	41.9	0.911	1	0.564	1	0.449	1
Butyricimonas	40.9	0.343	1	0.133	1	0.0711	1
Phascolarctobacterium	40	0.489	1	0.403	1	0.237	1
Lachnoclostridium	39.1	0.233	1	0.45	1	0.85	1
Mannheimia	36.3	0.601	1	0.916	1	0.637	1
Prevotella	34.4	0.987	1	0.789	1	0.976	1
Serratia	34.4	0.748	1	0.113	1	0.416	1
Cronobacter	33.5	0.0774	1	0.852	1	0.19	1
Desulfovibrio	32.6	0.472	1	0.903	1	0.732	1
Dialister	31.6	0.577	1	0.751	1	0.721	1
Paraprevotella	29.8	0.624	1	0.394	1	0.462	1
Flavonifractor	28.8	**0.000314**	**0.0427**	0.16	1	**8.30E-05**	**0.00954**
Akkermansia	27	0.261	1	0.48	1	0.185	1
Haemophilus	26	0.472	1	0.643	1	0.28	1
[Eubacterium]	25.1	0.239	1	0.542	1	0.106	1
Parasutterella	25.1	0.703	1	0.776	1	0.83	1
Veillonella	23.3	**0.00501**	0.676	0.245	1	0.0653	1

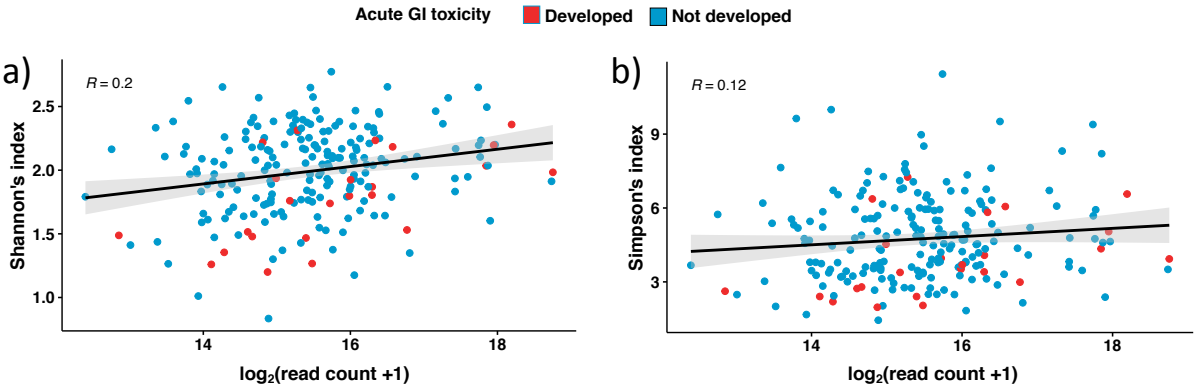
Holdemania	19.5	**0.0438**	1	0.232	1	**0.0165**	1
Herbaspirillum	17.7	0.79	1	0.981	1	0.888	1
Acidaminococcus	14.9	0.4	1	0.295	1	0.123	1
Turicibacter	14	0.0372	1	0.865	1	0.118	1
Tannerella	13	0.648	1	0.727	1	0.915	1
Oxalobacter	13	0.519	1	0.929	1	0.437	1
Subdoligranulum	13	0.106	1	0.165	1	**0.0305**	1
Tyzzerella	12.1	0.0907	1	0.175	1	**0.0339**	1
Ruminiclostridium	10.2	0.568	1	0.394	1	0.996	1
Eggerthella	9.3	0.144	1	0.667	1	0.18	1
Anaerostipes	9.3	0.933	1	0.586	1	0.627	1
Pseudoflavonifractor	9.3	0.322	1	0.48	1	0.825	1
Catenibacterium	8.37	0.258	1	0.789	1	0.135	1
Enterobacter	8.37	0.382	1	0.703	1	0.424	1
Coprobacter	8.37	0.497	1	0.839	1	0.662	1
Escherichia	7.44	0.267	1	0.801	1	0.319	1
Coprobacillus	7.44	0.573	1	0.553	1	0.346	1
Megasphaera	7.44	0.601	1	0.51	1	0.904	1
Anaerotruncus	7.44	0.997	1	0.367	1	0.588	1
Victivallis	6.51	0.412	1	0.643	1	0.6	1
Photobacterium	6.51	0.987	1	0.679	1	0.913	1
Butyrivibrio	6.51	0.718	1	0.929	1	0.576	1
Pseudobutyrivibrio	5.58	0.811	1	0.376	1	0.938	1
Slackia	5.58	0.537	1	0.839	1	0.83	1
Salmonella	5.58	0.0581	1	0.691	1	0.194	1
Howardella	5.58	0.869	1	0.5	1	0.882	1
Olsenella	5.58	0.826	1	0.994	1	0.803	1
Leclercia	5.58	0.906	1	0.929	1	0.99	1
Granulicatella	5.58	0.837	1	0.609	1	0.817	1
Senegalimassilia	4.65	0.703	1	0.789	1	0.591	1
Klebsiella	4.65	0.382	1	0.632	1	0.729	1
Trabulsiella	4.65	**0.0152**	1	0.342	1	**0.00871**	0.993
Gordonibacter	3.72	0.546	1	0.703	1	0.839	1
Fusobacterium	3.72	0.917	1	0.852	1	0.935	1
Enterococcus	3.72	0.759	1	0.727	1	0.979	1
Citrobacter	3.72	0.21	1	0.994	1	0.453	1
Escherichia/Shigella	3.72	0.97	1	0.789	1	0.814	1
Cloacibacillus	2.79	0.895	1	0.412	1	0.546	1
Lachnobacterium	2.79	0.653	1	0.865	1	0.572	1
Megamonas	2.79	0.371	1	0.968	1	0.455	1
Pyramidobacter	2.79	0.774	1	0.367	1	0.453	1
Mitsuokella	2.79	0.541	1	0.394	1	0.976	1

Synergistes	2.79	0.853	1	0.376	1	0.79	1
Rothia	2.79	0.869	1	0.789	1	0.904	1
Providencia	2.79	0.439	1	0.839	1	0.622	1
Kluyvera	2.79	0.568	1	0.826	1	0.809	1
Adlercreutzia	1.86	0.769	1	0.903	1	0.938	1
Paraeggerthella	1.86	0.863	1	0.942	1	0.777	1
Parasporobacterium	1.86	0.644	1	0.942	1	0.615	1
Acetivibrio	1.86	0.658	1	0.852	1	0.569	1
Aggregatibacter	1.86	0.917	1	0.929	1	0.814	1
Gilliamella	1.86	0.8	1	0.994	1	0.766	1
Morganella	1.86	0.115	1	0.903	1	0.237	1
Oscillospira	1.86	0.059	1	0.942	1	0.153	1
Pseudomonas	1.86	0.774	1	0.942	1	0.716	1
Gibbsiella	1.86	0.316	1	0.929	1	0.494	1
Campylobacter	1.86	0.949	1	0.929	1	0.904	1
Succinivibrio	1.86	0.326	1	0.703	1	0.174	1
Actinomyces	1.86	0.954	1	0.903	1	0.874	1
Shigella	0.93	0.842	1	0.955	1	0.985	1
Anaerofustis	0.93	0.559	1	0.955	1	0.719	1
Neisseria	0.93	0.779	1	0.994	1	0.75	1
Gemella	0.93	0.842	1	1	1	0.798	1
Pseudoalteromonas	0.93	0.949	1	1	1	0.893	1
Dielma	0.93	0.485	1	1	1	0.617	1
Atopobium	0.93	0.805	1	0.955	1	0.769	1
Butyricicoccus	0.93	0.764	1	1	1	0.924	1
Denitrobacterium	0.93	0.911	1	0.955	1	0.833	1
Lactococcus	0.93	0.577	1	1	1	0.708	1
Bulleidia	0.93	0.733	1	1	1	0.822	1
Weissella	0.93	0.713	1	0.994	1	0.698	1
Pantoea	0.93	0.506	1	0.994	1	0.65	1
Yersinia	0.93	0.748	1	0.955	1	0.866	1
Erwinia	0.93	0.443	1	0.955	1	0.612	1
Robinsoniella	0.93	0.774	1	0.994	1	0.885	1
Allisonella	0.93	0.506	1	1	1	0.632	1
Rikenella	0.93	0.582	1	0.852	1	0.322	1

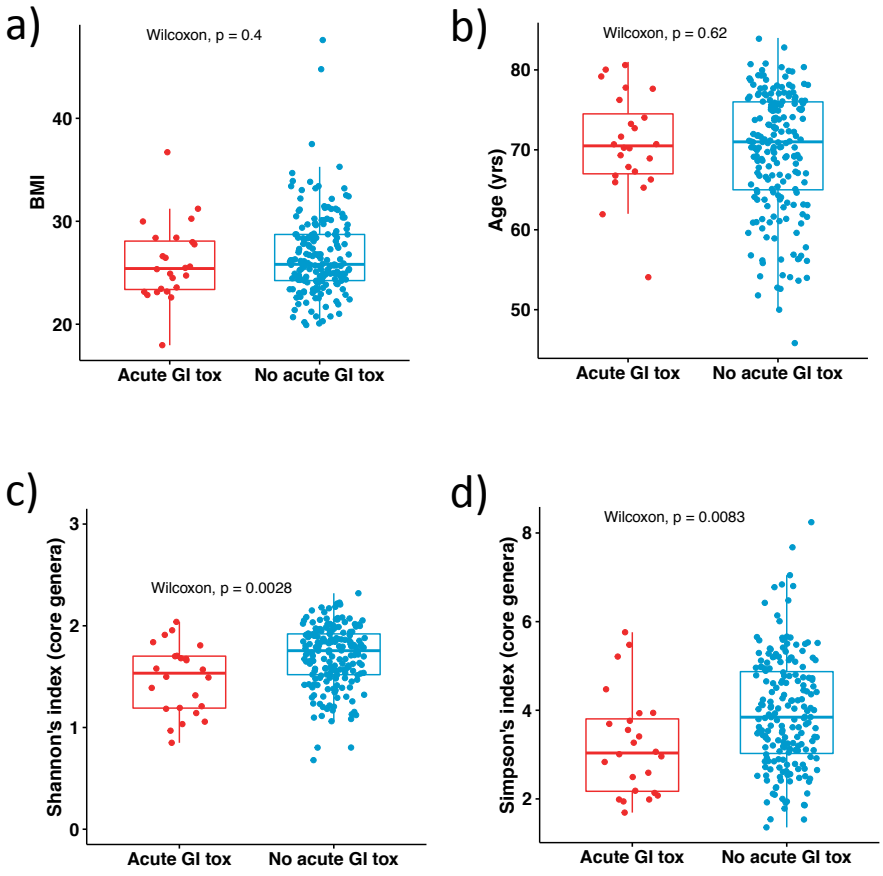
*Occurrence is the percentage of patients in which the genus was detected; p values obtained using Wilcoxon's rank test on the clr-transformed abundances and adjusted with Benjamini-Hochberg procedure.



Supplementary Fig. S5 | Genera showing a trend of association with acute GI toxicity
The distributions of the relative abundance of those genera that showed a trend for association with acute GI toxicity endpoint ($p<0.05$ in pooled cohort, see Supplementary Table S9) is shown, stratified by toxicity group; only *Flavonifractor* (top-middle) was tested significantly differentially abundant between toxicity groups after adjustment of p values (Benjamini-Hochberg procedure).



Supplementary Fig. S6 | Alpha diversity measures
Shannon's index (a) and Simpson's index (b) show no significant correlation with the total number of taxonomy-assigned read counts at the genus level across the MicroLearner PCa patient population (n=215).

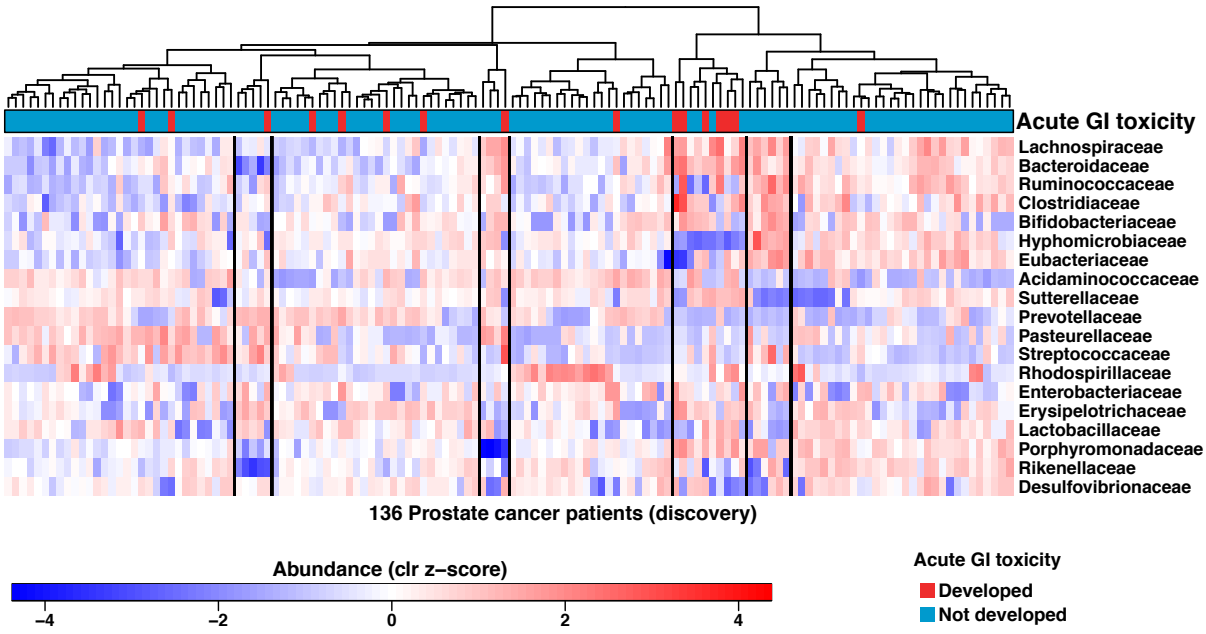


Supplementary Fig. S7 | Association between diversity of baseline microbiota and development of RT-induced early acute GI toxicity
(a) BMI and **(b)** age values across the pooled cohort (n=215) stratified by toxicity group. **(c)** Shannon's index and **(d)** Simpson's index measured in the pooled cohort by restricting to the relative abundance of genera in the core microbiota.

Microbiota-based clusters of patients

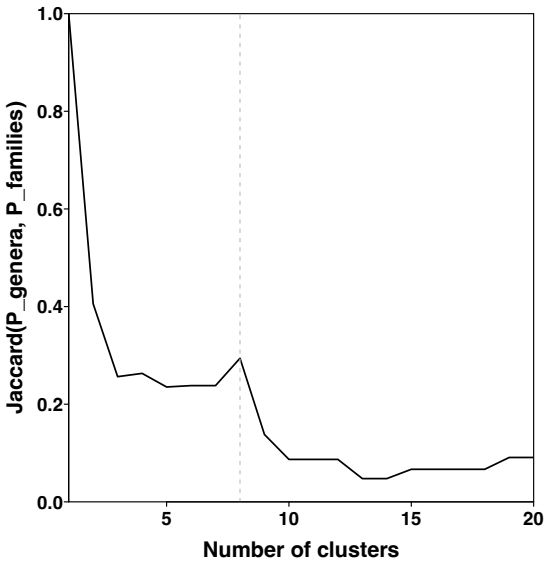
Hierarchical clustering using Euclidean distance and Ward linkage method was performed on the core clr abundance profiles standardised across the discovery population (column-wise normalization, heatmap Fig. 3a). The core included all genera having relative abundance ≥2% in ≥10% of the discovery samples and the count data of the core sub-composition were clr-transformed after imputation of zeros via Geometric Bayesian Multiplicative replacement method previous to clustering.

The optimal number of clusters was defined as the one maximising the Jaccard similarity index between the partition of toxicity events obtained by using core bacterial genera and the one obtained by using core bacterial families, defined as the families having relative abundance ≥2% in ≥10% of the discovery patients (Supplementary Fig. S8-S9).



Supplementary Fig. S8 | Patient clustering via core bacterial families

Heatmap showing the normalised abundance profiles (standardised clr-transformed relative abundance values) of the core microbiota bacterial families (rows) from patients in the MicroLearner PCa discovery cohort (columns); core families were defined as those families occurring in at least 10% of the cohort with relative abundance >=2%; hierarchical clustering was performed using Euclidean distance and Ward linkage method; the partition in 8 clusters (black vertical lines) is the one maximising the Jaccard similarity (in terms of toxicity events) with the partition obtained by clustering patients according to their core genera profiles.



Supplementary Fig. S9 | Optimal number of microbiota clusters

The Jaccard similarity index between the microbiota-based partition of toxicity events obtained by clustering patients according to the abundance profiles of core bacterial genera and the microbiota-based partition of toxicity events obtained by clustering patients according to the abundance profiles of core bacterial families, was measured in function of the number of clusters in the partitions. The optimal number of clusters (n=8) was defined as the non-trivial integer that maximised the similarity index in the range [3, 20].

Supplementary Table S10 | Enrichment analysis of microbiota clusters for patient characteristics and lifestyle factors (discovery cohort n=136)

Microbiota cluster ID	1	2	3	4	5	6	7	8
Cluster size (n)	12	23	6	26	4	24	31	10
(w.t. profiled cytokines)	10	15	2	21	4	20	20	7
(w.t. FFQ)	11	22	4	24	4	21	22	8
Microbiota risk class	mod.	mod.	mod.	low	mod.	mod.	mod.	high
Fisher's exact test counts p greater p lower								
Acute GI toxicity	1	2	1	0	1	4	1	6
(n=16)	0.7921	0.7984	0.5352	1	0.3975	0.3028	0.9880	**0.0002**
	0.5740	0.4670	0.8528	**0.0267**	0.9320	0.8767	0.078	1
age >= 65 years	9	17	4	21	2	20	20	9
(n=102)	0.6519	0.6628	0.8351	0.3149	0.9518	0.2218	0.9588	0.2333
	0.6183	0.5406	0.4683	0.8432	0.2601	0.9073	0.099	0.9499
BMI >= 30	1	4	0	4	1	3	10	2
(n=25)	0.9222	0.6526	1	0.7579	0.5607	0.8692	**0.0261**	0.5816
	0.3110	0.5798	0.2881	0.4525	0.8460	0.3096	0.9924	0.7301
Appendectomy	1	2	1	8	2	3	7	2
(n=26)	0.9307	0.9642	0.7274	0.0841	0.165	0.8882	0.3728	0.6057
	0.2879	0.1323	0.6775	0.9704	0.9779	0.2757	0.7961	0.7076
Cholecystectomy	1	1	0	0	1	0	3	0
(n=6)	0.4318	0.6785	1	1	0.1669	1	0.1315	1
	0.9120	0.7323	0.7588	0.2726	0.9906	0.3044	0.9758	0.6267
Prostatectomy	4	8	3	9	1	3	15	4
(n=47)	0.6499	0.5789	0.3419	0.5823	0.8209	0.9983	0.0533	0.4753
	0.5985	0.6103	0.8924	0.5985	0.5706	**0.0087**	0.9789	0.7685
Diabetes	4	0	0	3	1	2	2	2
(n=14)	**0.0219**	1	1	0.5256	0.3558	0.7516	0.8766	0.2741
	0.9971	0.0645	0.5144	0.7352	0.5352	0.5352	0.3369	0.9321
Physically active	5	4	3	8	2	7	10	2
(n=41)	0.2731	0.9619	0.2549	0.556	0.3503	0.6332	0.4661	0.8629
	0.8902	0.11	0.9328	0.6299	0.9183	0.5598	0.7004	0.3716
Smoker	4	13	3	13	2	7	17	7
(n=66)	0.9215	0.2701	0.6314	0.5199	0.6678	0.9906	0.2758	0.1395
	0.2126	0.8577	0.6873	0.65	0.7122	**0.03**	0.8424	0.9608
Vegetarian diet	1	1	1	0	0	0	0	0
(n=3)	0.2437	0.429	0.1275	1	1	1	1	1
	0.9795	0.926	0.9952	0.5263	0.9137	0.5558	0.4571	0.7938
High intake of	0	3	1	3	2	4	4	1
animal proteins & fats	1	0.7130	0.4955	0.7746	0.1131	0.4171	0.4584	0.7523
(n=18)	0.1425	0.5416	0.8868	0.4613	0.9884	0.8004	0.7684	0.6397
High intake of	3	4	0	7	0	6	2	0
plant proteins & fibers	0.3465	0.6443	1	0.1287	1	0.1736	0.9566	1
(n=12)*	0.8715	0.5930	0.4259	0.953	0.4894	0.9343	0.1558	0.1752
PRS>0	4	6	1	12	1	6	5	5
(n=40)**	0.6363	0.6206	0.6472	0.0663	0.8791	0.9077	0.9689	0.092
	0.628	0.6034	0.8392	0.9772	0.4663	0.2114	0.0923	0.9835
*cluster size (w.t. profiled cytokines) **cluster size (w.t. FFQ)								

Acronyms

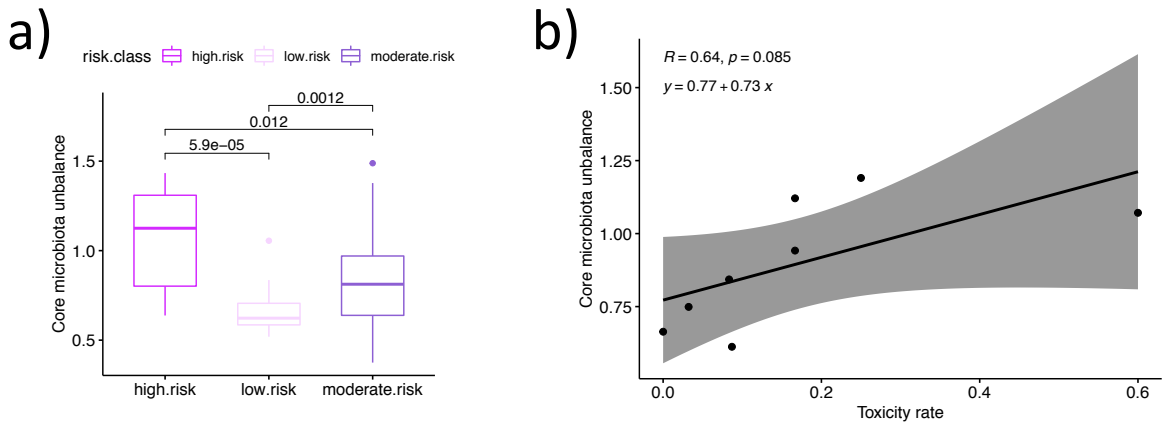
BMI: body mass index; PRS: polycytokinic risk score; FFQ: food frequency questionnaire); w.t.: with; mod.: moderate.

Supplementary Table S11 | Association between microbiota risk classes and variables associated with patient, microbiota and treatment (discovery cohort n=136)

	Wilcoxon's rank test		
	microbiota high risk vs microbiota low risk	microbiota high risk vs microbiota mod. risk	microbiota low risk vs microbiota mod. risk
Mean rectal dose	p=0.201	p=0.799	p=**0.016**
Number of core genera	p=**0.005**	p=0.058	p=0.081
Simpson's index	p=**0.019**	p=**0.014**	p=0.921
Shannon's index	p=0.031	p=0.034	p=0.658
Age	p=0.944	p=0.921	p=0.882
BMI	p=0.374	p=0.120	p=0.877
PRS	p=0.533	p=**0.022**	p=0.054

Acronyms
BMI: body mass index; PRS: polycytokinic risk score; mod.: moderate.

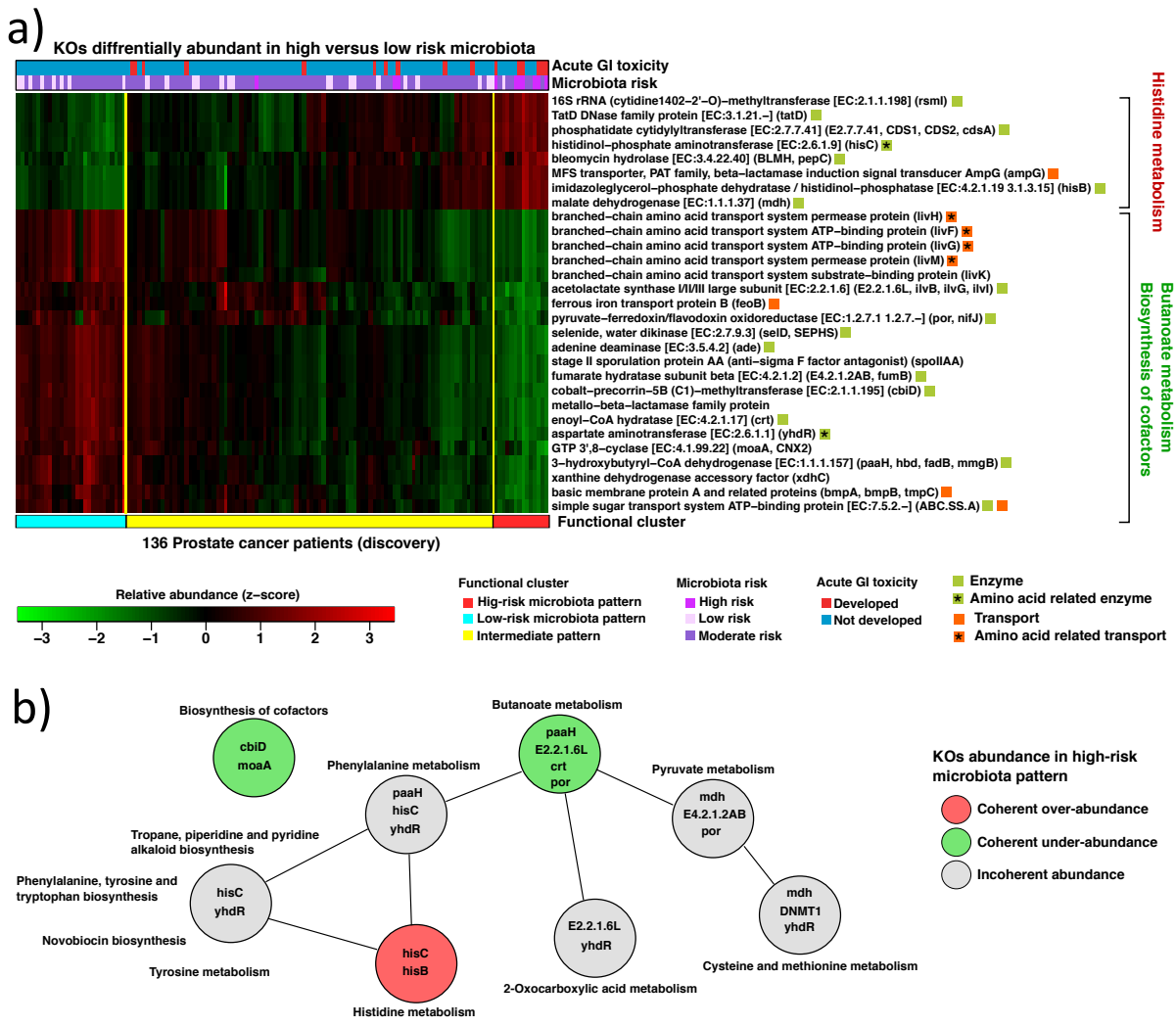
Core microbiota unbalance



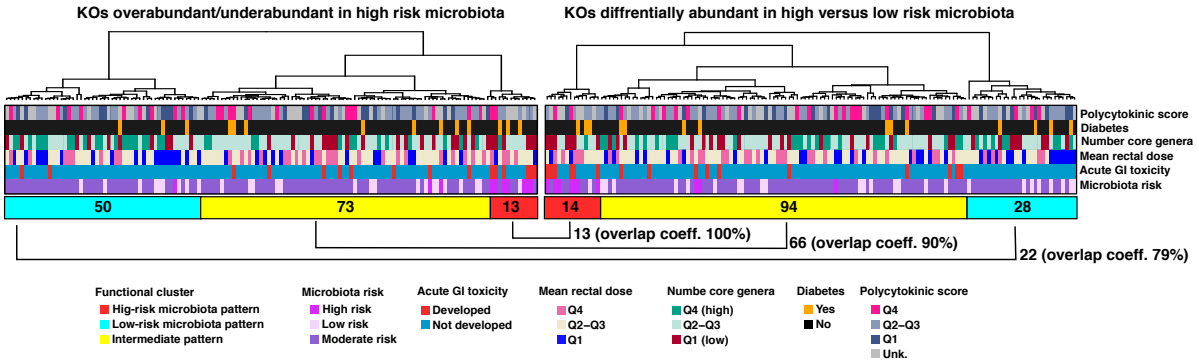
Supplementary Fig. S10 | Analysis of unbalance of the core microbiota profiles
(a) Statistic of the core microbiota unbalance score, corresponding to the L1-norm of the normalised clr-transformed abundance profiles of the core microbiota genera is shown for the MicroLearner discovery cohort (n=136) stratified by microbiota risk class. **(b)** The within-cluster average value of the core microbiota unbalance score is shown in function of the toxicity rate for the 8 microbiota clusters identified in the MicroLearner discovery cohort.

Functional imputation analysis
Functional imputed metagenomics analysis was performed using PICRUSt version 1.1.4[27]. We referenced the core metagenome profiles to the most updated reference collection of OTUs from Greengenes (gg_13_5_otus)[28]. Because the OTU assignment by the Ion Reporter metagenomic pipeline is based on ThermoFisher-curated Greengenes references, this choice maximised the coherence between inferred functional profiles and taxonomic abundance profiles. The profiles were normalised by the known or predicted 16S copy number abundance and a final metagenome was predicted and mapped to KEGG Orthologs (KOs). The inferred KO IDs annotations based on the Integrated Microbial Genomes and Microbiomes[29] were manually verified and curated against the most recent version of KEGG (v.104, 2022/10) and of the Transport Classification database[30] and annotated by KEGG modules or Brites.

The quality of the inferred metagenome was checked by calculating the average Nearest Sequenced Taxon Index (NSTI)[27] across samples, that quantifies the average over the samples of the average substitutions per site separating each OTU from the reference bacterial genome. We obtained NSTI=0.071+-0.028, consistent with values expected from faecal samples.



Supplementary Fig. S11 | Functional imputation analysis of core microbiota
(a) Heatmap showing the average relative contribution of core genera (columns) to selected KEGG Orthologs (KOs, rows) measured across the MicroLearner discovery cohort (n=136); selected functional terms include 13 KOs that are characteristic of the high-risk microbiota cluster and 29 KOs that are significantly different in terms of relative abundance between the high- and the low-risk core microbiota compositions; KO labels are coloured according to annotation by KEGG modules/Brites related to different aspects of metabolism and transport.



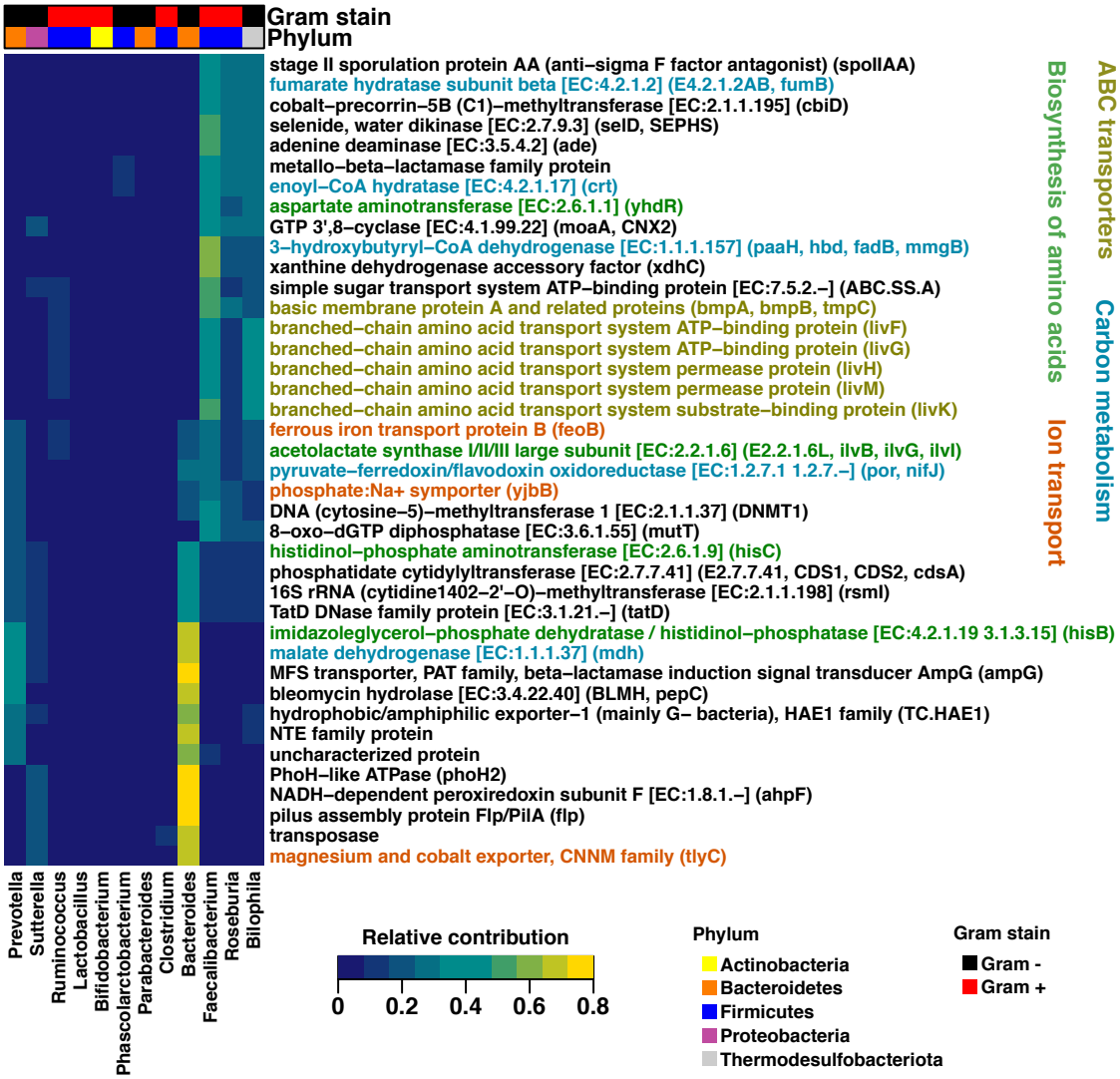
Supplementary Fig. S12 | Comparison of functional partitions of discovery cohort

The partitions of the discovery cohort obtained by clustering patients according to KEGG Ortholog genes (KOs) associated with high-risk microbiota (left) and with KOs differentially abundant in high- versus low-risk microbiota (right) are compared; in the two partitions, the functional clusters that contain most of the high-risk microbiota patents (red) have 100% overlap coefficients; the functional partition on the right can also cluster apart low-risk microbiota patients with more accuracy (cyan cluster, size 28 with 0 toxicity counts).

Supplementary Table S12 | KEGG Orthologs differentially abundant between the high-risk microbiota cluster and the low-risk microbiota cluster

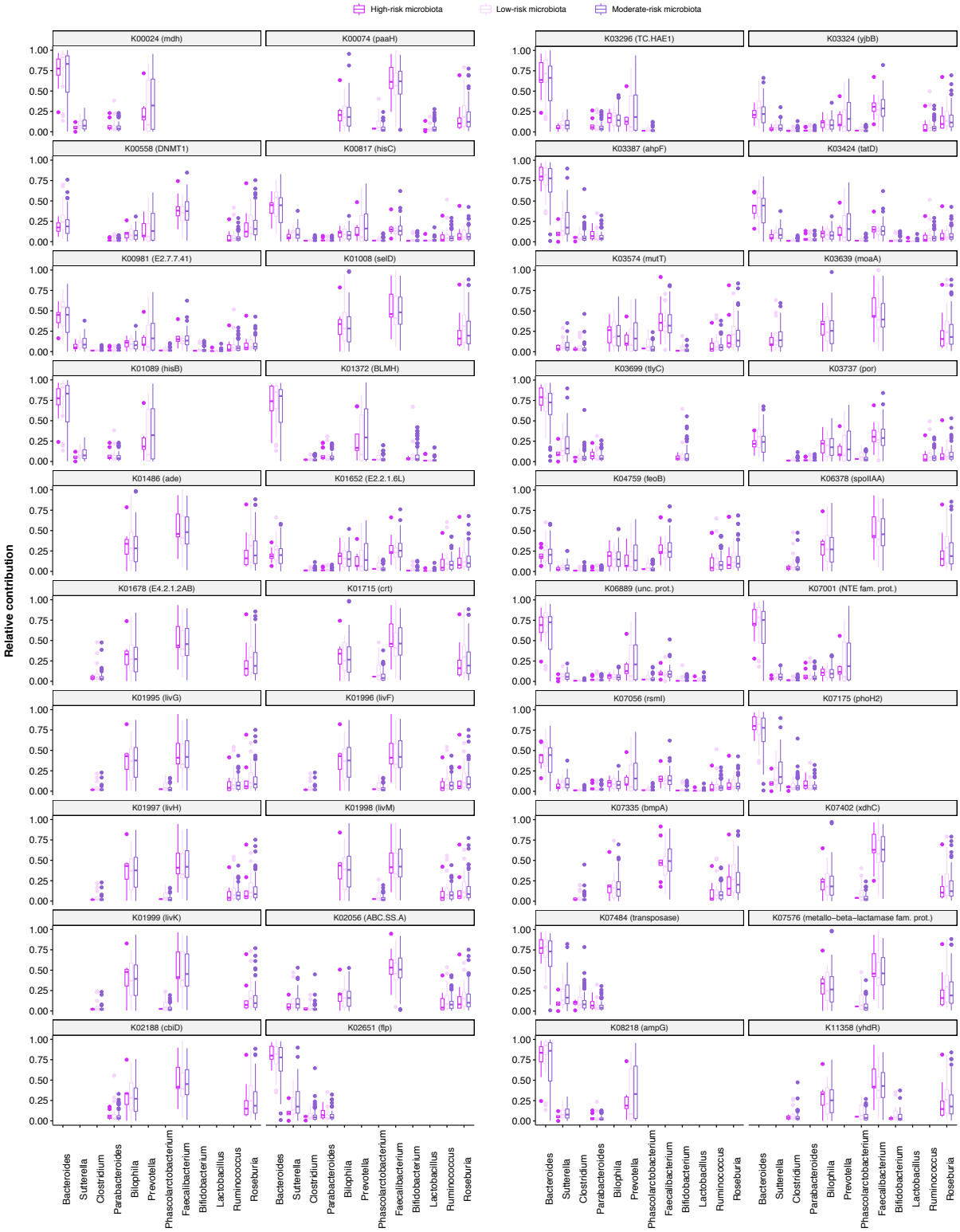
KEGG identifier (name)	logFC	p	adj. p
K00817 (hisC)	1.63	2.41E-06	0.0016
K01715 (paaH)	-1.71	4.08E-06	0.0016
K07576	-1.71	4.08E-06	0.0016
K01486 (ade)	-1.75	4.33E-06	0.0016
K01008 (selD)	-1.75	4.33E-06	0.0016
K04759 (feoB)	-1.62	5.16E-06	0.0016
K02188 (cbiD)	-1.69	5.57E-06	0.0016
K01998 (livM)	-1.64	7.46E-06	0.0016
K03424 (tatD)	1.56	9.19E-06	0.0016
K07056 (rsmI)	1.56	9.19E-06	0.0016
K01678 (fumB)	-1.64	1.06E-05	0.0016
K06378 (spoIIAA)	-1.64	1.06E-05	0.0016
K01996 (livF)	-1.59	1.06E-05	0.0016
K01997 (livH)	-1.59	1.06E-05	0.0016
K01995 (livG)	-1.59	1.06E-05	0.0016
K00981 (cdsA)	1.52	1.39E-05	0.0019
K03737 (por)	-1.54	1.42E-05	0.0019
K01999 (livK)	-1.54	1.68E-05	0.002
K01372 (pepC)	1.6	2.35E-05	0.0023
K01652 (ilvB)	-1.53	2.73E-05	0.0023
K07402 (xdhC)	-1.65	3.58E-05	0.0023
K07335 (bmpA)	-1.57	4.19E-05	0.0024
K11358 (yhdR)	-1.50	4.93E-05	0.0026
K00024 (mdh)	1.54	4.97E-05	0.0026
K01089 (hisB)	1.54	4.97E-05	0.0026
K03639 (moaA)	-1.52	5.88E-05	0.0029
K08218 (ampG)	1.51	6.04E-05	0.0029
K02056 (ABC.SS.A)	-1.5	7.47E-05	0.0031
K00074 (paaH)	-1.53	1.4E-04	0.0036

Acronyms
FC: fold change
KO: KEGG Orthologs



Supplementary Fig. S13 | Analysis of microbiota functional traits different in high- versus low- risk microbiota

(a) Heatmap showing the microbiota functional profiles (standardised relative abundance values) of 29 selected KEGG Orthologs (KOs, rows) imputed from 16S core microbiota abundance profiles of the patients in the MicroLearner PCa discovery cohort (n=136, columns) before RT initiation; the KOs selected are significantly differentially abundant between high- and low-risk microbiota classes of patients; hierarchical clustering was used to define 3 functional clusters of patients (yellow vertical lines) and to reveal the functional patterns associated with the composition of the high- and of the low-risk microbiota classes with respect to the onset of acute GI toxicity during RT; KEGG modules containing selected KOs consistently overrepresented/underrepresented in the high-risk microbiota pattern are reported in red/blue on the heatmap right side. **(b)** Network describing the relations between KEGG modules to which the selected KOs are annotated to; nodes describe biological modules and they are linked if they share at least one KO among the selected ones; nodes are coloured according to whether they include functional traits coherently/incoherently over- or under-abundant in the high-risk microbiota functional pattern.



Supplementary Fig. S14 | Relative contribution of core microbiota genera to KEGG Ortholog genes of interest.

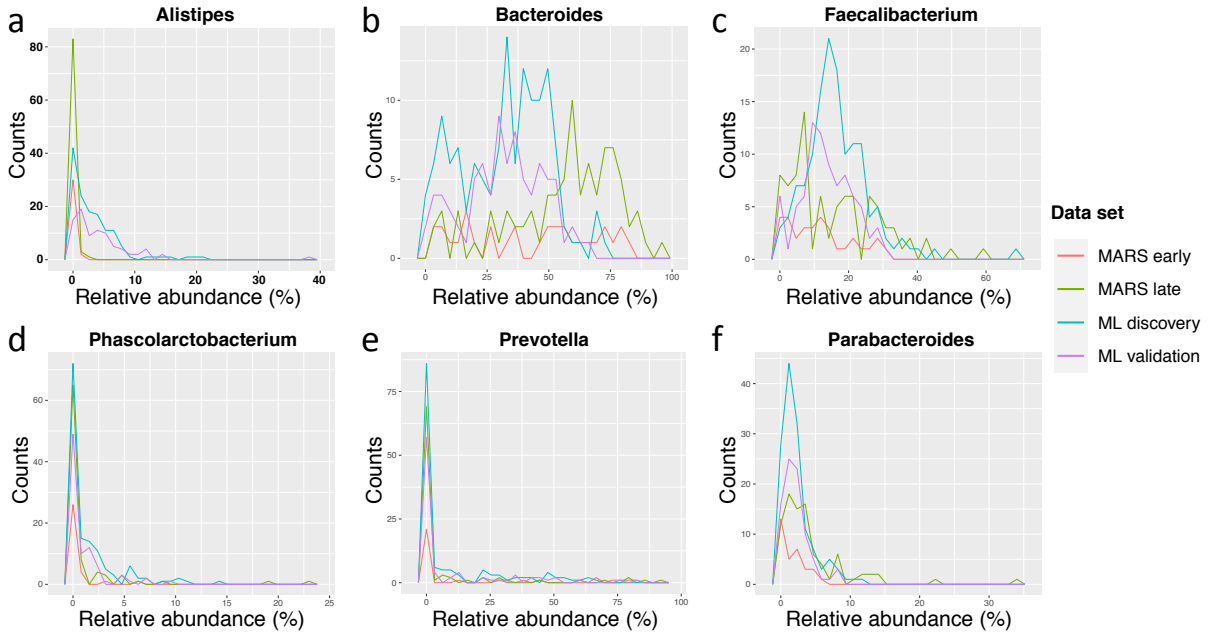
The distributions across patients of the relative contribution of core microbiota genera to the KEGG Orthologs associated with high-risk microbiota or differentially abundant in high- versus low-risk microbiota are shown, stratified by microbiota risk classes.

Machine-learning for microbiota-based prediction of acute gastrointestinal toxicity

Supplementary Table S13 | Genera differentially abundant between the high-risk microbiota cluster and each of the other microbiota clusters

Genus	logFC*	p	adj. p	microbiota cluster (ID sequence)
Gemmiger	-1.4	6.063365e-07	8.246176e-05	6
Faecalibacterium	-1.93	3.318977e-04	2.256904e-02	6
Gemmiger	-1.44	2.351736e-07	1.818781e-05	4
Alistipes	-1.98	2.674678e-07	1.818781e-05	4
Faecalibacterium	-1.89	8.400547e-06	3.380068e-04	4
Phascolarctobacterium	1.35	9.941377e-06	3.380068e-04	4
Prevotella	-5.86	1.928824e-22	2.623200e-20	2
Bacteroides	4.48	3.699207e-10	2.515461e-08	2
Parabacteroides	1.7	1.390334e-03	4.727136e-02	2
Alistipes	-2.04	4.075588e-07	0.000055428	5 (low-risk)
Gemmiger	-1.15	2.097678e-05	0.001426421	5 (low-risk)
Faecalibacterium	-1.77	7.856912e-05	0.003561800	5 (low-risk)
Bilophila	1.95	2.493194e-06	0.0003390744	8
Faecalibacterium	-2.06	1.195243e-05	0.0008127654	8
Prevotella	-3.74	3.637121e-07	4.946484e-05	1
Alistipes	-1.44	8.462327e-07	5.754382e-05	1
Faecalibacterium	-1.78	6.070532e-06	2.578632e-04	1
Bacteroides	1.73	7.584212e-06	2.578632e-04	1

*FC = fold change of high-risk cluster versus cluster specified in column 5



Supplementary Fig. S15 | Analysis of MICLIDE tree predictors across data sets

The distribution of relative abundance of predictors is shown across MicroLearner (ML) discovery and validation cohorts and across MARS early (n=32) and late cohorts (n=87); while the total number of counts is cohort-dependent, the peaks and the shapes of the distributions appears to be similar, suggesting that the MICLIDE tree predictors are detected consistently across data sets and that they are robust with respect to sequencing technology and cohort geography.

Supplementary Table S14 | Confusion matrix of the MICLIDE tree from prediction on the training set

MICLIDE tree training				
MicroLearner PCa (discovery cohort, n=136)				
	Predicted risk	**high**	**low**	**moderate**
Risk group				
high (n=10)		80%	10%	10%
low (n=26)		0%	54%	46%
moderate (100)		0%	6%	94%

Supplementary Table S15 | Prediction table of the MICLIDE tree from prediction on test sets

MICLIDE tree validation				
MicroLearner PCa (validation cohort, n=79)				
	Predicted risk	**high**	**low**	**moderate**
Toxicity endpoint				
Acute GI tox (n=8)		3	1	4
No acute GI tox (n=71)		2	18	51
	Toxicity rate	60%	5%	7%
MARS				
	Predicted risk	**high**	**low**	**moderate**
Toxicity endpoint				
Acute GI tox (n=8)		2	0	6
No acute GI tox (n=24)		4	0	20
	Toxicity rate	33%	not applicable	23%

Multivariable logistic models of acute gastrointestinal toxicity

We developed different multivariable logistic models (MLMs) to check whether the interplay of potential risk factors could reveal any predictive potential and outperform the microbiota-based prediction.

We considered a first model with cytokines, BMI, age and mean rectal dose as predictors (n=99, patients with cytokines data available). We performed logistic multivariable regression using the lrm function of the 'rms' R package. The model characteristics are summarised in Supplementary Table S16.

Supplementary Table S16 | MLM of acute toxicity including cytokines and clinical risk factors

predictor	**coefficient estimate**	**standard error**	**p-value**
log(CCL2)	-3.30	2.24	0.14
log(PDGF_BB)	-0.63	0.86	0.46
log(TGFB1)	-0.1	0.56	0.86
log(TNFR1)	-1.76	1.3	0.18
log(TNFa)	1.42	0.85	0.09
BMI	-0.07	0.1	0.48
Age	0.04	0.04	0.38
Mean rectal dose	0.03	0.07	0.63

No coefficient was found to be significantly different from zero ($p<0.05$).
When we included in the model the Shannon's index for the alpha diversity of the core microbiota as a predictor, we found that the Shannon's index was the only feature with a coefficient significantly different from zero (Supplementary Table 17).

Supplementary Table S17 | MLM of acute toxicity including cytokines, clinical factors, and microbiota diversity

predictor	coefficient estimate	standard error	p-value
log(CCL2)	-4.25	2.48	0.09
log(PDGF_BB)	-0.74	0.94	0.43
log(TGFB1)	-0.19	0.57	0.75
log(TNFR1)	-2.28	1.89	0.23
log(TNFa)	1.71	0.94	0.07
BMI	-0.03	0.11	0.79
Age	0.06	0.05	0.22
Mean rectal dose	0.03	0.07	0.72
Shannon's index	-2.45	1.16	**0.04**

By running a stepwise backward regression starting from the latter set of predictors we found that the best predictors were the set of cytokines already identified for the definition of the polycytokinic risk score together with the Shannon's index (Supplementary Table 18), confirming that the polycytokinic risk defined has predictive potential but lack specificity.

Supplementary Table S18 | MLM of acute toxicity including cytokines and microbiota diversity

predictor	coefficient estimate	standard error	p-value
log(CCL2)	-4.15	2.2	0.06
log(TNFR1)	-2.03	1.67	0.23
log(TNFa)	1.63	0.91	0.07
Shannon's index	-2.17	1.06	**0.04**

Finally, we developed a model enforcing as predictors the clinical risk factors and the microbiota diversity excluding cytokines (n=136, all patients from the discovery cohort). Again, we found that the only coefficient significantly different from zero was the Shannon's index coefficient (Supplementary Table 19).

Supplementary Table S19 | MLM of acute toxicity including clinical risk factors and microbiota diversity

predictor	coefficient estimate	standard error	p-value
BMI	-0.03	0.09	0.689
Age	0.03	0.04	0.496
Average rectal dose	-0.01	0.05	0.817
Shannon's index	-2.65	0.89	**0.003**