Supplementary Material

Supplementary Tables

Supplementary Table 1. PERMANOVA analysis. Bray-Curtis dissimilarities between samples based on features. P-values <0.05 are shaded grey. R^2 values are recorded to two decimal places.

	R^2	p-value
Clinical group	1.15	1 x 10^{-3}
Sex	0.33	1 x 10^{-3}
Age	0.22	1 x 10^{-3}

Supplementary Table 2. PERMANOVA analysis of Bray-Curtis distances. P-values <0.05 are shaded grey. All pairwise comparisons between groups within groupings yielded q-values <0.05.

Grouping	P-value
CRC Adenoma Non-neoplastic Colonoscopy-normal Blood-negative	1x10^{-3}
Neoplasm (Adenoma and CRC) Non-neoplastic Colonoscopy-normal Blood-negative	1x10^{-3}
CRC Adenoma All controls (Non-neoplastic, Colonoscopy-normal, and Blood-negative)	1x10^{-3}
Neoplasm (Adenoma and CRC) All controls (Non-neoplastic, Colonoscopy-normal, and Blood-negative)	1x10^{-3}
CRC Adenoma Colonoscopy-controls (Non-neoplastic and Colonoscopy-normal)	1x10^{-3}
Neoplasm (Adenoma and CRC) Colonoscopy-controls (Non-neoplastic and Colonoscopy-normal)	1x10^{-3}

Supplementary Table 3. Pairwise Kruskal-Wallis analysis of Shannon diversity index. q values <0.05 are shaded grey. H values are recorded to two decimal places.

Group 1	Group 2	H	p-value	q-value
Adenoma	CRC	25.91	3.6 x 10^{-7}	7.2 x 10^{-7}
	Blood-negative	66.85	2.9 x 10^{-16}	1.5 x 10^{-15}
	Colonoscopy-normal	6.08	1.4 x 10^{-2}	1.9 x10^{-2}
	Non-neoplastic	0.80	3.7 x 10^{-1}	3.7 x10^{-1}
CRC	Blood-negative	5.87	1.5 x 10^{-2}	1.9 x10^{-2}
	Colonoscopy-normal	37.36	9.8 x 10^{-10}	2.5 x 10^{-9}
	Non-neoplastic	25.12	5.4 x 10^{-7}	9.0 x 10^{-7}
Blood-negative	Colonoscopy-normal	70.75	4.1 x 10^{-17}	4.1 x 10^{-16}
	Non-neoplastic	56.84	4.7 x 10^{-14}	1.6 x 10^{-13}
Colonoscopy-normal	Non-neoplastic	1.99	1.6 x 10^{-1}	1.8 x 10^{-1}

Supplementary Table 4. Number of samples available to Random Forest (RF) models. RF models were constructed using the following data: 'Clinical' = age and sex; **'**Bacteria' = relative abundance of genera; 'Bacteria & clinical' = relative abundance of genera, age and sex.

RF model			Number of samples					
			CRC	**Adenoma risk-group**			**Colonoscopy-Normal**	**Blood-negative**
				High	**Intermediate**	**Low**		
CRC vs blood-negative	**All models**	**Test**	217					243
		Validation	213					248
Neoplasm vs blood-negative	**Clinical**	**Test**	94	83	97	104		245
		Validation	91	108	90	88		246
	Bacteria	**Test**	84	99	102	101		250
		Validation	112	92	99	92		241
	Bacteria & clinical	**Test**	93	100	99	89		239
		Validation	94	91	93	91		252
CRC vs colonoscopy-normal	**All models**	**Test**	204				161	
		Validation	226				139	
Neoplasm vs colonoscopy-normal	**Clinical**	**Test**	104	89	97	91	151	
		Validation	83	102	94	104	149	
	Bacteria	**Test**	88	90	97	94	159	
		Validation	100	101	89	98	141	
	Bacteria & clinical	**Test**	92	99	95	105	143	
		Validation	89	92	101	95	157	

Supplementary Table 5. AUC results. RF models were constructed using the following data: ‘Clinical’ = age and sex; ‘Bacteria’ = relative abundance of genera; ‘Bacteria & clinical’ = relative abundance of genera, age and sex. ‘Neoplasm’ = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma.

<table>
<tr><th colspan="5">CRC vs blood-negative</th></tr>
<tr><th>RF Model</th><th>Test AUC</th><th>Validation AUC</th><th>p-value</th><th>Total AUC</th></tr>
<tr><td>Clinical</td><td>0.62
(0.57-0.68)</td><td>0.65
(0.60-0.70)</td><td>0.45</td><td>0.63 (0.60-0.67)</td></tr>
<tr><td>Bacteria</td><td>0.89
(0.86-0.91)</td><td>0.86
(0.82-0.89)</td><td>0.24</td><td>0.89 (0.87-0.91)</td></tr>
<tr><td>Bacteria & clinical</td><td>0.89
(0.87-0.92)</td><td>0.87
(0.83-0.90)</td><td>0.22</td><td>0.90 (0.88-0.92)</td></tr>
<tr><th colspan="2">Comparison of Validation AUC</th><th colspan="3">p-value</th></tr>
<tr><td colspan="2">‘Clinical’ compared with ‘Bacteria’</td><td colspan="3">2.1 x 10^{-11}</td></tr>
<tr><td colspan="2">‘Bacteria’ compared with ‘Bacteria & clinical’</td><td colspan="3">6.1 x 10^{-5}</td></tr>
<tr><th colspan="5">AUC for the ‘Bacteria’ total model restricted to the top 15 taxa</th></tr>
<tr><td colspan="5">0.88 (0.86-0.90)</td></tr>
</table>

<table>
<tr><th colspan="5">Neoplasm vs blood-negative</th></tr>
<tr><th>RF Model</th><th>Test AUC</th><th>Validation AUC</th><th>p-value</th><th>Total AUC</th></tr>
<tr><td>Clinical</td><td>0.61
(0.57-0.66)</td><td>0.64
(0.59-0.68)</td><td>0.48</td><td>0.63 (0.59-0.66)</td></tr>
<tr><td>Bacteria</td><td>0.82
(0.78-0.85)</td><td>0.78
(0.74-0.82)</td><td>0.12</td><td>0.84 (0.81-0.86)</td></tr>
<tr><td>Bacteria & clinical</td><td>0.81
(0.78-0.84)</td><td>0.84
(0.80-0.87)</td><td>0.31</td><td>0.85 (0.82-0.87)</td></tr>
<tr><th colspan="2">Comparison of Validation AUC</th><th colspan="3">p-value</th></tr>
<tr><td colspan="2">‘Clinical’ compared with ‘Bacteria’</td><td colspan="3">2.0 x 10^{-6}</td></tr>
<tr><td colspan="2">‘Bacteria’ compared with ‘Bacteria & clinical’</td><td colspan="3">3.0 x 10^{-2}</td></tr>
<tr><th colspan="5">AUC for the ‘Bacteria’ total model restricted to the top 15 taxa</th></tr>
<tr><td colspan="5">0.84 (0.82-0.86)</td></tr>
</table>

<table>
<tr><th colspan="5">CRC vs colonoscopy-normal</th></tr>
<tr><th>RF Model</th><th>Test AUC</th><th>Validation AUC</th><th>p-value</th><th>Total AUC</th></tr>
<tr><td>Clinical</td><td>0.57
(0.51-0.63)</td><td>0.61
(0.55-0.67)</td><td>0.40</td><td>0.59 (0.54-0.63)</td></tr>
<tr><td>Bacteria</td><td>0.77
(0.72-0.81)</td><td>0.79
(0.74-0.83)</td><td>0.59</td><td>0.78 (0.75-0.82)</td></tr>
<tr><td>Bacteria & clinical</td><td>0.77
(0.72-0.81)</td><td>0.79
(0.74-0.83)</td><td>0.48</td><td>0.78 (0.75-0.82)</td></tr>
<tr><th colspan="2">Comparison of Validation AUC</th><th colspan="3">p-value</th></tr>
<tr><td colspan="2">‘Clinical’ compared with ‘Bacteria’</td><td colspan="3">7.7 x 10^{-6}</td></tr>
<tr><td colspan="2">‘Bacteria’ compared with ‘Bacteria & clinical’</td><td colspan="3">3.6 x 10^{-2}</td></tr>
<tr><th colspan="5">AUC for the ‘Bacteria’ total model restricted to the top 15 taxa</th></tr>
<tr><td colspan="5">0.79 (0.75-0.82)</td></tr>
</table>

<table>
<tr><th colspan="5">Neoplasm vs colonoscopy-normal</th></tr>
<tr><th>RF Model</th><th>Test AUC</th><th>Validation AUC</th><th>p-value</th><th>Total AUC</th></tr>
<tr><td>Clinical</td><td>0.56
(0.51-0.62)</td><td>0.58
(0.53-0.63)</td><td>0.67</td><td>0.57 (0.53-0.60)</td></tr>
<tr><td>Bacteria</td><td>0.71
(0.66-0.75)</td><td>0.73
(0.68-0.77)</td><td>0.54</td><td>0.73 (0.70-0.76)</td></tr>
<tr><td>Bacteria & clinical</td><td>0.68
(0.63-0.73)</td><td>0.74
(0.69-0.78)</td><td>0.10</td><td>0.72 (0.69-0.75)</td></tr>
<tr><th colspan="2">Comparison of Validation AUC</th><th colspan="3">p-value</th></tr>
<tr><td colspan="2">‘Clinical’ compared with ‘Bacteria’</td><td colspan="3">5.2 x 10^{-5}</td></tr>
<tr><td colspan="2">‘Bacteria’ compared with ‘Bacteria & clinical’</td><td colspan="3">0.83</td></tr>
<tr><th colspan="5">AUC for the ‘Bacteria’ total model restricted to the top 15 taxa</th></tr>
<tr><td colspan="5">0.71 (0.67-0.74)</td></tr>
</table>

Supplementary Table 6. Confusion matrices. RF models were constructed using the following data: 'Clinical' = age and sex; 'Bacteria' = relative abundance of genera; 'Bacteria & clinical' = relative abundance of genera, age and sex. 'Neoplasm' = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma. PPV = positive predictive value. NPV = negative predictive value. FPR = false positive rate. FNR = false negative rate. Confusion matrices were created using the predict function of randomForest using the default vote proportion cutoff of 50%. It should be noted that disease prevalence within the study cohort differs from disease prevalence within the wider NHS Bowel Cancer Screening Programme population.

CRC vs blood-negative (total)									
Clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Blood-negative**	**CRC**							
Blood-negative	273	218	44%	68%	56%	57%	67%	44%	32%
CRC	137	293	32%						
Bacteria	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Blood-negative**	**CRC**							
Blood-negative	417	74	15%	76%	85%	82%	80%	15%	24%
CRC	104	326	24%						
Bacteria & clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Blood-negative**	**CRC**							
Blood-negative	416	75	15%	78%	85%	82%	82%	15%	22%
CRC	93	337	22%						

Neoplasm vs blood-negative (total)									
Clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Blood-negative**	**Neoplasm**							
Blood-negative	286	205	42%	67%	58%	71%	53%	42%	33%
Neoplasm	249	506	33%						
Bacteria	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Blood-negative**	**Neoplasm**							
Blood-negative	340	151	31%	83%	69%	81%	72%	31%	17%
Neoplasm	133	648	17%						
Bacteria & clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Blood-negative**	**Neoplasm**							
Blood-negative	346	145	30%	83%	70%	81%	74%	30%	17%
Neoplasm	124	626	17%						

CRC vs colonoscopy-normal (total)									
Clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Colonoscopy-normal**	**CRC**							
Colonoscopy-normal	135	165	55%	72%	45%	65%	53%	55%	28%
CRC	120	310	28%						
Bacteria	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Colonoscopy-normal**	**CRC**							
Colonoscopy-normal	182	118	39%	79%	61%	74%	67%	39%	21%
CRC	89	341	21%						
Bacteria & clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Colonoscopy-normal**	**CRC**							
Colonoscopy-normal	180	120	40%	79%	60%	74%	67%	40%	21%
CRC	89	341	21%						

Neoplasm vs colonoscopy-normal (total)									
Clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Colonoscopy-normal**	**Neoplasm**							
Colonoscopy-normal	145	155	52%	67%	48%	77%	37%	52%	33%
Neoplasm	251	513	33%						
Bacteria	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Colonoscopy-normal**	**Neoplasm**							
Colonoscopy-normal	137	163	54%	82%	46%	79%	50%	54%	18%
Neoplasm	137	620	18%						
Bacteria & clinical	**Predicted value**		**Error**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**FPR**	**FNR**
True value	**Colonoscopy-normal**	**Neoplasm**							
Colonoscopy-normal	137	163	54%	82%	46%	79%	50%	54%	18%
Neoplasm	139	629	18%						

Supplementary Table 7. Number of samples available to Random Forest (RF) models. 'Bacteria' = relative abundance of genera. 'Neoplasm' = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma. 'Colonoscopy-controls' = a group comprising an approximately equal ratio of non-neoplastic and colonoscopy-normal samples.

		Number of samples						
'Bacteria' RF model		**CRC**	**Adenoma risk-group**			**Colonoscopy-normal**	**Blood-negative**	**Non-neoplastic**
			High	**Intermediate**	**Low**			
CRC vs adenoma	**Test**	211	91	92	107			
	Validation	219	100	99	84			
Adenoma vs colonoscopy-normal	**Test**		88	88	100	160		
	Validation		103	103	91	140		
Adenoma vs blood-negative	**Test**		95	102	92		243	
	Validation		96	89	99		248	
CRC vs colonoscopy-controls	**Test**	217				148		150
	Validation	213				152		150
Adenoma vs colonoscopy-controls	**Test**		103	83	96	158		146
	Validation		88	108	95	142		154
Neoplasm vs colonoscopy-controls	**Test**	92	86	97	93	152		162
	Validation	99	105	94	98	148		138

Supplementary Table 8. AUC results. 'Bacteria' = relative abundance of genera. 'Neoplasm' = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma. 'Colonoscopy-controls' = a group comprising an approximately equal ratio of non-neoplastic and colonoscopy-normal samples.

'Bacteria' RF Model	**Test AUC**	**Validation AUC**	**p-value**	**Total AUC**
CRC vs adenoma	0.64 (0.59-0.69)	0.71 (0.66-0.76)	0.03	0.70 (0.67-0.74)
Adenoma vs colonoscopy-normal	0.70 (0.65-0.75)	0.72 (0.67-0.77)	0.55	0.72 (0.68-0.75)
Adenoma vs blood-negative	0.80 (0.76-0.84)	0.84 (0.80-0.87)	0.17	0.82 (0.79-0.84)
CRC vs colonoscopy-controls	0.71 (0.66-0.75)	0.76 (0.72-0.80)	0.06	0.74 (0.71-0.77)
Adenoma vs colonoscopy-controls	0.61 (0.57-0.66)	0.65 (0.61-0.70)	0.22	0.64 (0.61-0.67)
Neoplasm vs colonoscopy-controls	0.64 (0.60-0.69)	0.64 (0.60-0.68)	0.96	0.65 (0.62-0.68)

Supplementary Table 9. Confusion matrices. 'Bacteria' RF models were constructed using relative abundance of genera. 'Neoplasm' = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma. 'Colonoscopy-controls' = a group comprising an approximately equal ratio of non-neoplastic and colonoscopy-normal samples. PPV = positive predictive value. NPV = negative predictive value. FPR = false positive rate. FNR = false negative rate. Confusion matrices were created using the predict function of randomForest using the default vote proportion cutoff of 50%. It should be noted that disease prevalence within the study cohort differs from disease prevalence within the wider NHS Bowel Cancer Screening Programme population.

CRC vs adenoma (total RF model)	Predicted value		Error	Sensitivity	Specificity	PPV	NPV	FPR	FNR
True value	Adenoma	CRC							
Adenoma	454	119	21%	52%	79%	65%	69%	21%	48%
CRC	205	225	48%						

Adenoma vs colonoscopy-normal (total RF model)	Predicted value		Error	Sensitivity	Specificity	PPV	NPV	FPR	FNR
True value	Colonoscopy-normal	Adenoma							
Colonoscopy-normal	148	152	51%	79%	49%	75%	55%	51%	21%
Adenoma	123	450	21%						

Adenoma vs blood-negative (total RF model)	Predicted value		Error	Sensitivity	Specificity	PPV	NPV	FPR	FNR
True value	Blood-negative	Adenoma							
Blood-negative	363	128	26%	76%	74%	77%	73%	26%	24%
Adenoma	135	438	24%						

CRC vs colonoscopy-controls (total RF model)	Predicted value		Error	Sensitivity	Specificity	PPV	NPV	FPR	FNR
True value	Colonoscopy control	CRC							
Colonoscopy control	427	173	29%	61%	71%	60%	72%	29%	39%
CRC	167	263	39%						

Adenoma vs colonoscopy-controls (total RF model)	Predicted value		Error	Sensitivity	Specificity	PPV	NPV	FPR	FNR
True value	Colonoscopy control	Adenoma							
Colonoscopy control	368	232	39%	59%	61%	59%	61%	39%	41%
Adenoma	237	336	41%						

Neoplasm vs colonoscopy-controls (total RF model)	Predicted value		Error	Sensitivity	Specificity	PPV	NPV	FPR	FNR
True value	Colonoscopy control	Neoplasm							
Colonoscopy control	282	318	53%	73%	47%	64%	58%	53%	27%
Neoplasm	208	556	27%						

Supplementary Table 10.A. The 15 most important taxa from the bacteria “CRC vs adenoma” total RF model.

CRC vs adenoma			
Taxa	**Mean decrease Gini**	**Taxa**	**Mean decrease Accuracy**
D_5__Fusobacterium	9.8	D_5__Fusobacterium	17.9
D_5__Parvimonas	6.5	D_5__Parvimonas	15.1
D_5__Ruminococcaceae.UCG.002	6.0	D_5__Peptostreptococcus	13.7
D_5__Peptostreptococcus	5.8	D_5__Porphyromonas	10.2
D_5__Porphyromonas	4.7	D_5__Ruminococcaceae.UCG.002	9.5
D_5__Family.XIII.AD3011.group	4.5	D_5__Gemella	9.4
D_5__Christensenellaceae.R.7.group	4.4	D_5__Family.XIII.AD3011.group	8.5
D_5__Coprococcus.1	4.4	D_2__Mollicutes.D_3__NB1.n.__.__	5.9
D_5__Roseburia	4.2	D_5__Ruminiclostridium.6	5.6
D_5__Alistipes	4.2	D_3__Clostridiales.D_4__Family.XIII.D_5__uncultured	5.5
D_5__.Ruminococcus..torques.group	4.1	D_5__Ruminococcaceae.UCG.013	5.5
D_5__Odoribacter	4.1	D_5__Ruminococcaceae.UCG.004	5.3
D_5__Ruminococcaceae.UCG.013	4.0	D_5__Coprococcus.1	4.8
D_5__Escherichia.Shigella	4.0	D_5__Alistipes	4.5
D_5__Dialister	3.9	D_5__Family.XIII.UCG.001	4.5

Supplementary Table 10.B. The 15 most important taxa from the bacteria “Adenoma vs colonoscopy-normal” total RF model.

Adenoma vs colonoscopy-normal			
Taxa	**Mean decrease Gini**	**Taxa**	**Mean decrease Accuracy**
D_5__Streptococcus	5.8	D_5__Lactobacillus	10.3
D_5__Lactobacillus	5.0	D_5__Streptococcus	10.2
D_5__Faecalibacterium	3.7	D_5__Akkermansia	7.7
D_5__Akkermansia	3.6	D_5__Faecalibacterium	6.8
D_5__Erysipelotrichaceae.UCG.003	3.5	D_5__Veillonella	6.1
D_5__Bifidobacterium	3.5	D_5__Ruminiclostridium.9	5.6
D_5__Subdoligranulum	3.4	D_5__Lachnospiraceae.ND3007.group	5.5
D_5__.Eubacterium..ventriosum.group	3.3	D_5__Erysipelotrichaceae.UCG.003	5.4
D_5__Ruminococcus.1	3.2	D_5__Rothia	5.1
D_5__Ruminiclostridium.9	3.2	D_5__Odoribacter	5.1
D_5__Coprobacter	3.2	D_5__Coprobacter	5.1
D_4__Rhodospirillaceae.D_5__uncultured	3.1	D_5__Subdoligranulum	4.9
D_5__Escherichia.Shigella	3.1	D_5__.Clostridium..innocuum.group	4.7
D_5__Lachnospiraceae.ND3007.group	3.1	D_5__Senegalimassilia	4.5
D_4__Lachnospiraceae.__	3.1	D_5__Enterococcus	4.5

Supplementary Table 10.C. The 15 most important taxa from the bacteria “Adenoma vs blood-negative” total RF model.

Adenoma vs blood-negative			
Taxa	**Mean decrease Gini**	**Taxa**	**Mean decrease Accuracy**
D_5__Faecalibacterium	12.3	D_5__Faecalibacterium	17.8
D_5__Coprococcus.3	8.9	D_5__Coprococcus.3	17.0
D_5__Ruminococcaceae.NK4A214.group	8.0	D_5__Ruminococcaceae.NK4A214.group	12.5
D_5__Ruminococcaceae.UCG.010	7.7	D_5__Peptostreptococcus	12.5
D_5__Ruminococcaceae.UCG.002	7.6	D_5__Ruminococcaceae.UCG.010	11.5
D_5__Ruminococcaceae.UCG.005	7.0	D_5__Ruminococcaceae.UCG.002	11.5
D_5__.Ruminococcus..torques.group	6.6	D_5__Ruminococcaceae.UCG.005	11.4
D_5__Escherichia.Shigella	6.5	D_5__Akkermansia	11.0
D_5__Akkermansia	6.4	D_5__Christensenellaceae.R.7.group	10.4
D_4__Clostridiales.vadinBB60.group.D_5__uncultured.bacterium	6.1	D_4__Christensenellaceae.D_5__uncultured	10.1
D_5__Christensenellaceae.R.7.group	6.0	D_5__Fusobacterium	10.0
D_5__Ruminococcaceae.UCG.014	6.0	D_4__Clostridiales.vadinBB60.group.D_5__uncultured.bacterium	10.0
D_2__Mollicutes.D_3__NB1.n.__.__	5.6	D_5__Anaerococcus	9.9
D_5__Alistipes	5.5	.D_2__Mollicutes.D_3__NB1.n.__.__	9.6
D_2__Mollicutes.__.__.__	5.5	D_2__Mollicutes.__.__.__	9.4

Supplementary Table 10.D. The 15 most important taxa from the bacteria “CRC vs colonoscopy-controls” total RF model. ‘Colonoscopy-controls’ = a group comprising an approximately equal ratio of non-neoplastic and colonoscopy-normal samples.

CRC vs colonoscopy-controls			
Taxa	**Mean decrease Gini**	**Taxa**	**Mean decrease Accuracy**
D_5__Fusobacterium	6.8	D_5__Fusobacterium	14.0
D_5__Streptococcus	6.7	D_5__Ruminococcaceae.UCG.013	11.8
D_5__Ruminococcaceae.UCG.013	6.1	D_5__Peptostreptococcus	10.8
D_5__Lactobacillus	6.1	D_5__Parvimonas	10.7
D_5__Odoribacter	5.0	D_5__Lactobacillus	9.0
D_5__Ruminococcaceae.UCG.002	4.8	D_5__Ruminococcaceae.UCG.002	8.2
D_5__Peptostreptococcus	4.5	D_5__Rothia	8.0
D_5__.Eubacterium..coprostanoligenes.group	4.5	D_2__Mollicutes.D_3__NB1.n.__.__	7.7
D_5__Family.XIII.AD3011.group	4.4	D_5__Gemella	7.6
D_5__Alistipes	4.3	D_5__Streptococcus	7.6
D_4__Ruminococcaceae.D_5__uncultured	4.3	D_5__Odoribacter	7.5
D_5__Erysipelotrichaceae.UCG.003	4.3	D_5__Family.XIII.AD3011.group	7.3
D_4__Rhodospirillaceae.D_5__uncultured	4.2	D_5__Ruminococcaceae.UCG.014	6.4
D_4__Enterobacteriaceae.__	4.2	D_5__Alistipes	5.9
D_5__Faecalibacterium	4.1	D_4__Ruminococcaceae.D_5__uncultured	5.8

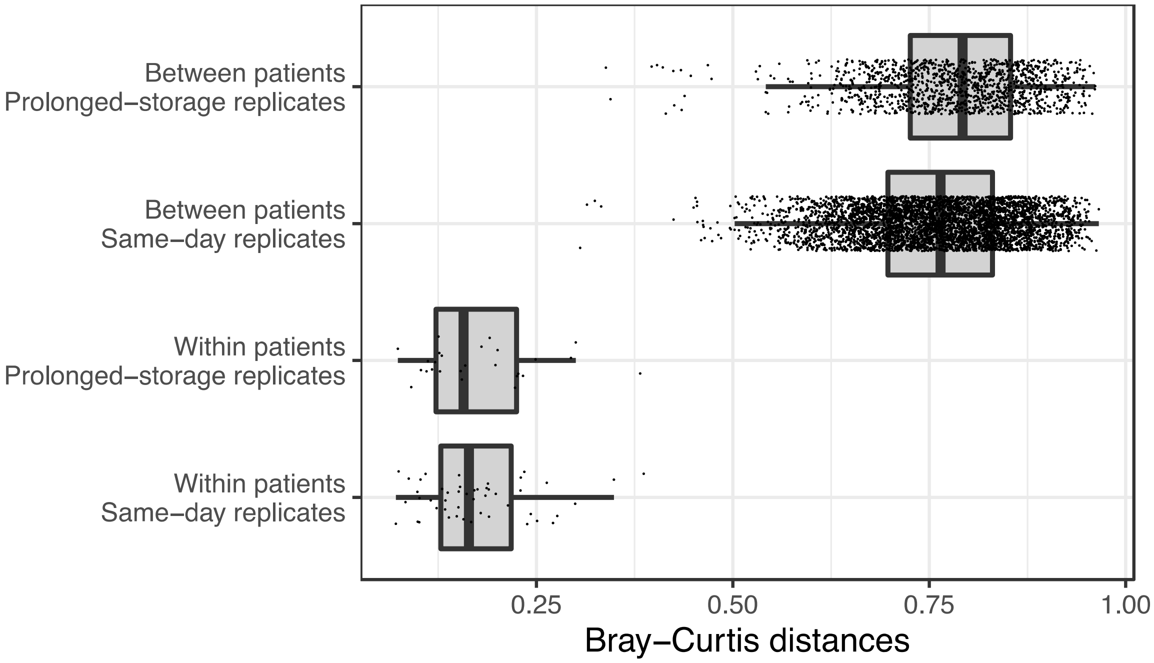
Supplementary Table 10.E. The 15 most important taxa from the bacteria "Adenoma vs colonoscopy-controls" total RF model. 'Colonoscopy-controls' = a group comprising an approximately equal ratio of non-neoplastic and colonoscopy-normal samples.

Adenoma vs colonoscopy-controls			
Taxa	**Mean decrease Gini**	**Taxa**	**Mean decrease Accuracy**
D_5__Streptococcus	7.7	D_5__Enterococcus	5.5
D_5__Erysipelotrichaceae.UCG.003	7.2	D_5__Erysipelotrichaceae.UCG.003	5.5
D_4__Lachnospiraceae.__	6.7	D_5__.Eubacterium..ventriosum.group	5.1
D_5__Lachnospiraceae.ND3007.group	6.6	D_5__Ruminiclostridium.5	4.7
D_5__Subdoligranulum	6.5	D_5__Lactobacillus	4.7
D_5__Faecalibacterium	6.4	D_5__Prevotella.7	4.7
D_5__Lactobacillus	6.3	D_5__Lachnospiraceae.ND3007.group	4.4
D_5__.Eubacterium..ventriosum.group	6.3	D_5__Ruminococcus.1	4.3
D_5__Butyricicoccus	6.2	D_5__.Eubacterium..coprostanoligenes.group	4.3
D_5__Roseburia	6.0	D_4__Ruminococcaceae.D_5__uncultured	4.2
D_5__Ruminiclostridium.5	5.9	D_4__Lachnospiraceae.__	4.1
D_5__.Ruminococcus..torques.group	5.8	D_5__Streptococcus	4.1
D_5__Escherichia.Shigella	5.8	D_4__Ruminococcaceae.__	3.9
D_5__.Eubacterium..coprostanoligenes.group	5.8	D_5__Ruminococcaceae.NK4A214.group	3.8
D_4__Enterobacteriaceae.__	5.7	D_5__Peptoniphilus	3.8

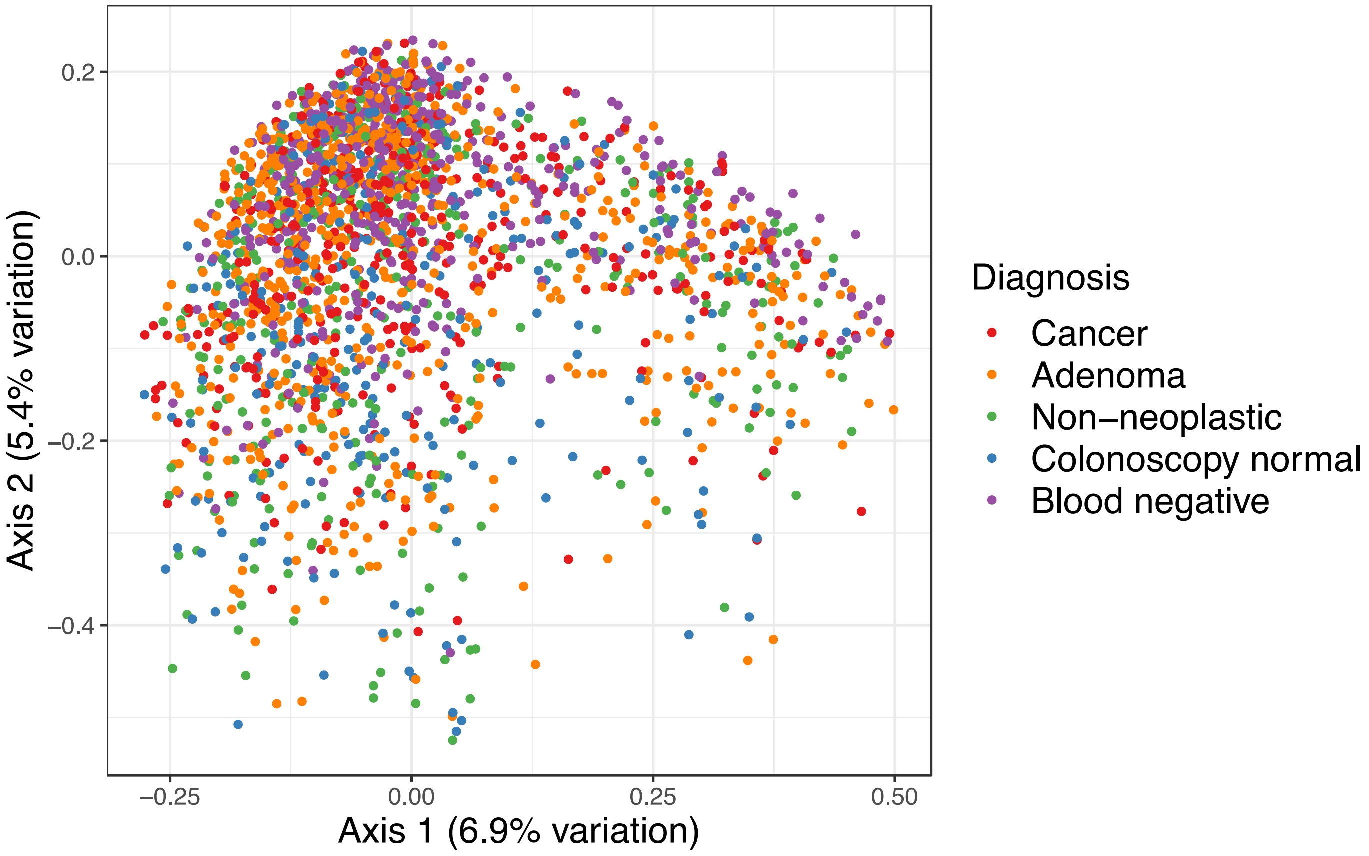
Supplementary Table 10.F. The 15 most important taxa from the bacteria "Neoplasm vs colonoscopy-controls" total RF model. 'Neoplasm' = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma. 'Colonoscopy-controls' = a group comprising an approximately equal ratio of non-neoplastic and colonoscopy-normal samples.

Neoplasm vs colonoscopy-controls			
Taxa	**Mean decrease Gini**	**Taxa**	**Mean decrease Accuracy**
D_5__Streptococcus	8.8	D_5__Streptococcus	7.6
D_5__Lactobacillus	7.3	D_5__Erysipelotrichaceae.UCG.003	7.2
D_5__Erysipelotrichaceae.UCG.003	7.1	D_4__Enterobacteriaceae.__	6.6
D_5__Roseburia	6.6	D_5__Lactobacillus	6.5
D_5__Butyricicoccus	6.3	D_5__Rothia	6.1
D_5__Subdoligranulum	6.3	D_5__Enterococcus	5.8
D_5__.Eubacterium..coprostanoligenes.group	6.1	D_4__Ruminococcaceae.D_5__uncultured	5.8
D_5__Ruminococcaceae.UCG.013	6.1	D_5__Veillonella	5.6
D_5__Escherichia.Shigella	6.1	D_5__Roseburia	5.3
D_5__Lachnospiraceae.ND3007.group	6.0	D_5__Ruminococcaceae.UCG.013	5.2
D_4__Ruminococcaceae.D_5__uncultured	6.0	D_5__Subdoligranulum	5.0
D_5__Faecalibacterium	6.0	D_5__.Clostridium..innocuum.group	4.6
D_5__.Ruminococcus..torques.group	5.9	D_5__Lachnospiraceae.NK4A136.group	4.6
D_4__Lachnospiraceae.__	5.9	D_5__Lachnospiraceae.ND3007.group	4.5
D_5__Alistipes	5.8	D_5__.Eubacterium..fissicatena.group	4.4

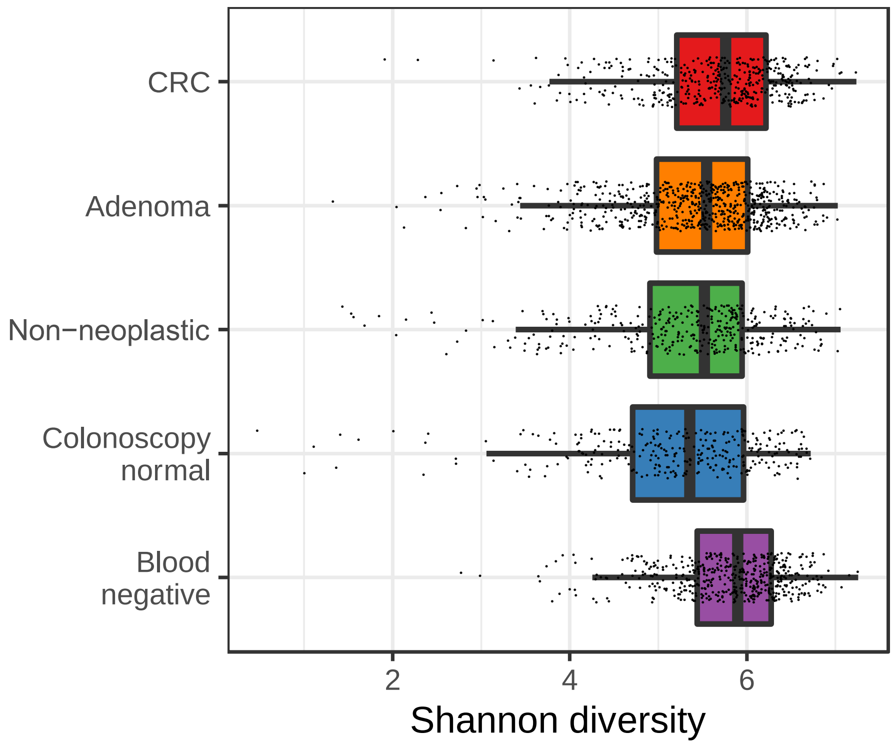
Supplementary Figures



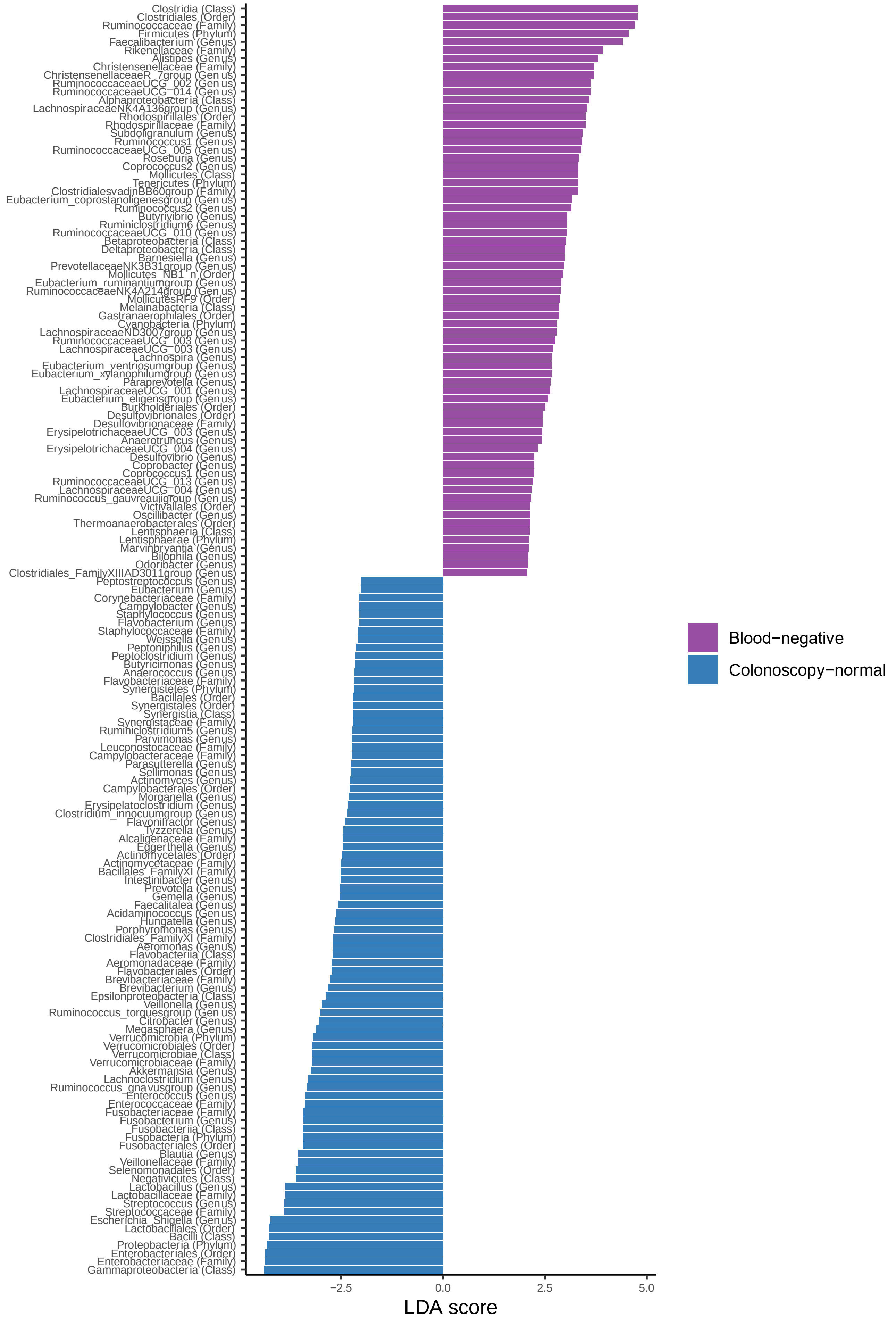
Supplementary Figure 1. Distribution of Bray-Curtis distances within and between DNA extraction replicates. The lower two boxplots depict the range of Bray-Curtis distances *within* pairs of replicates; the range is similar for samples extracted simultaneously or after a period of storage at ambient temperature. The upper two boxplots depict the range of Bray-Curtis distances *between* all of the samples within each group respectively; the ranges are larger, as is to be expected when comparing samples from different participants.



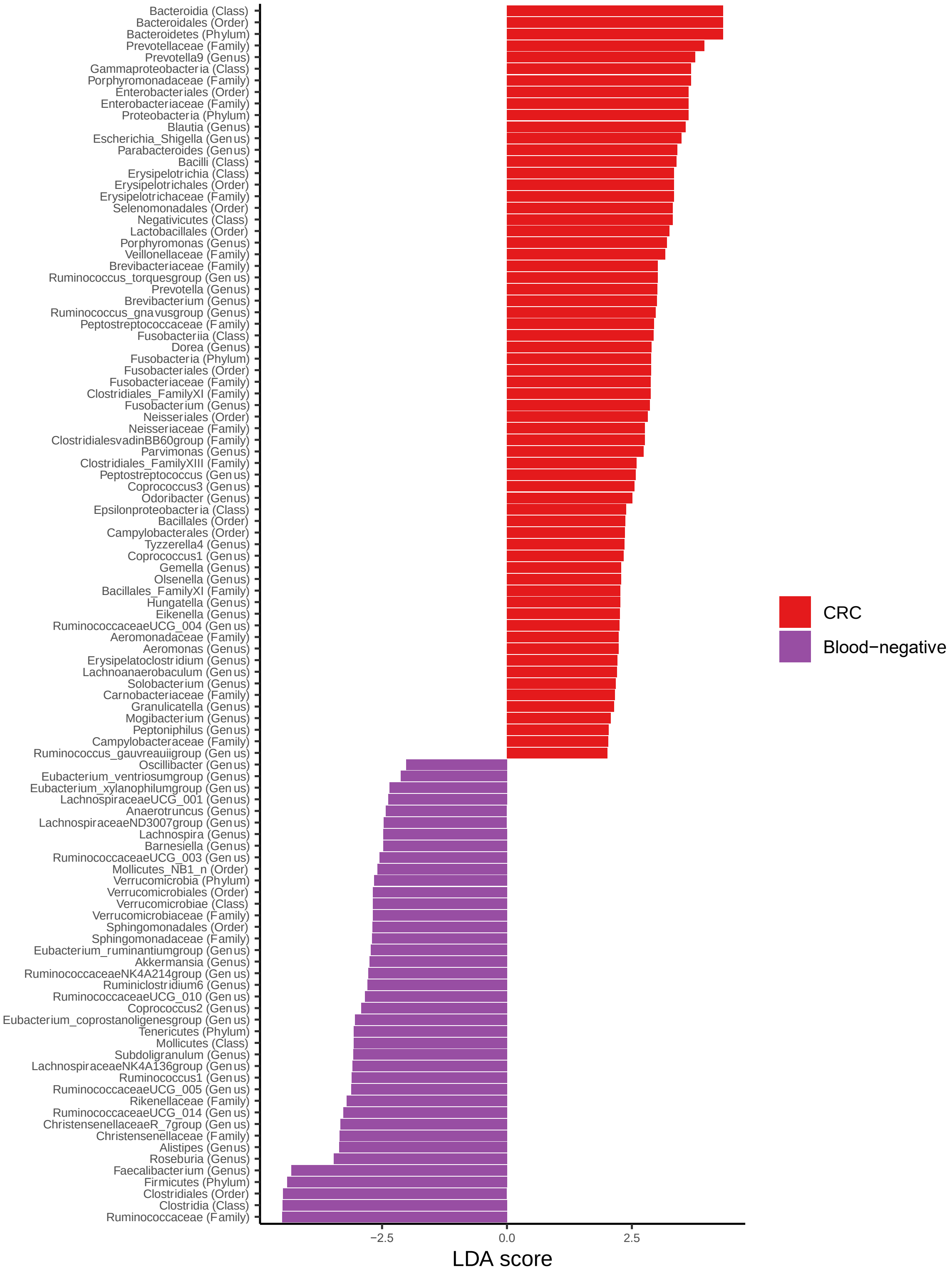
Supplementary Figure 2A. Principle coordinate analysis (PCoA) of Bray-Curtis distances between all samples.



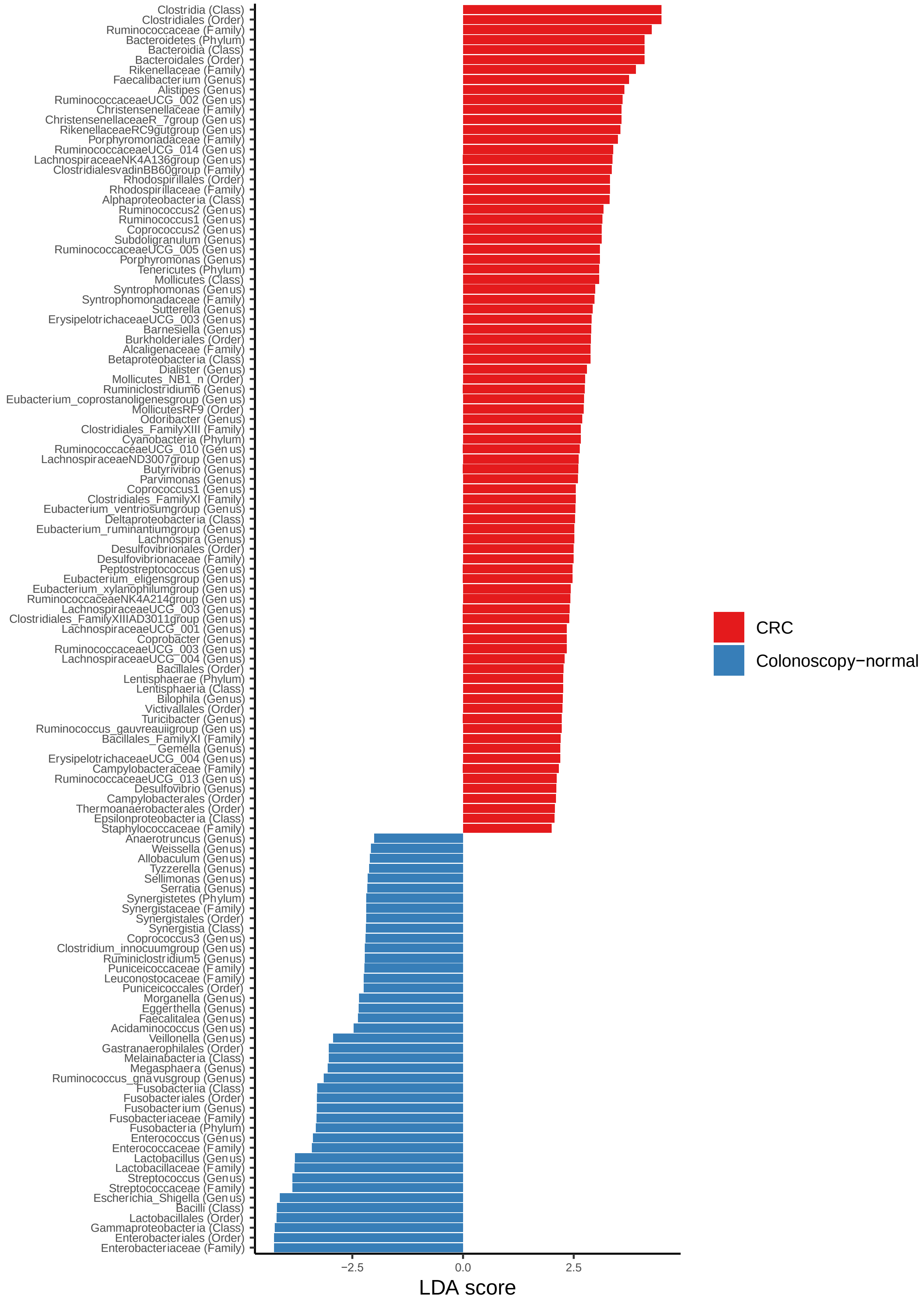
Supplementary Figure 2B. Distribution of Shannon diversity indices.



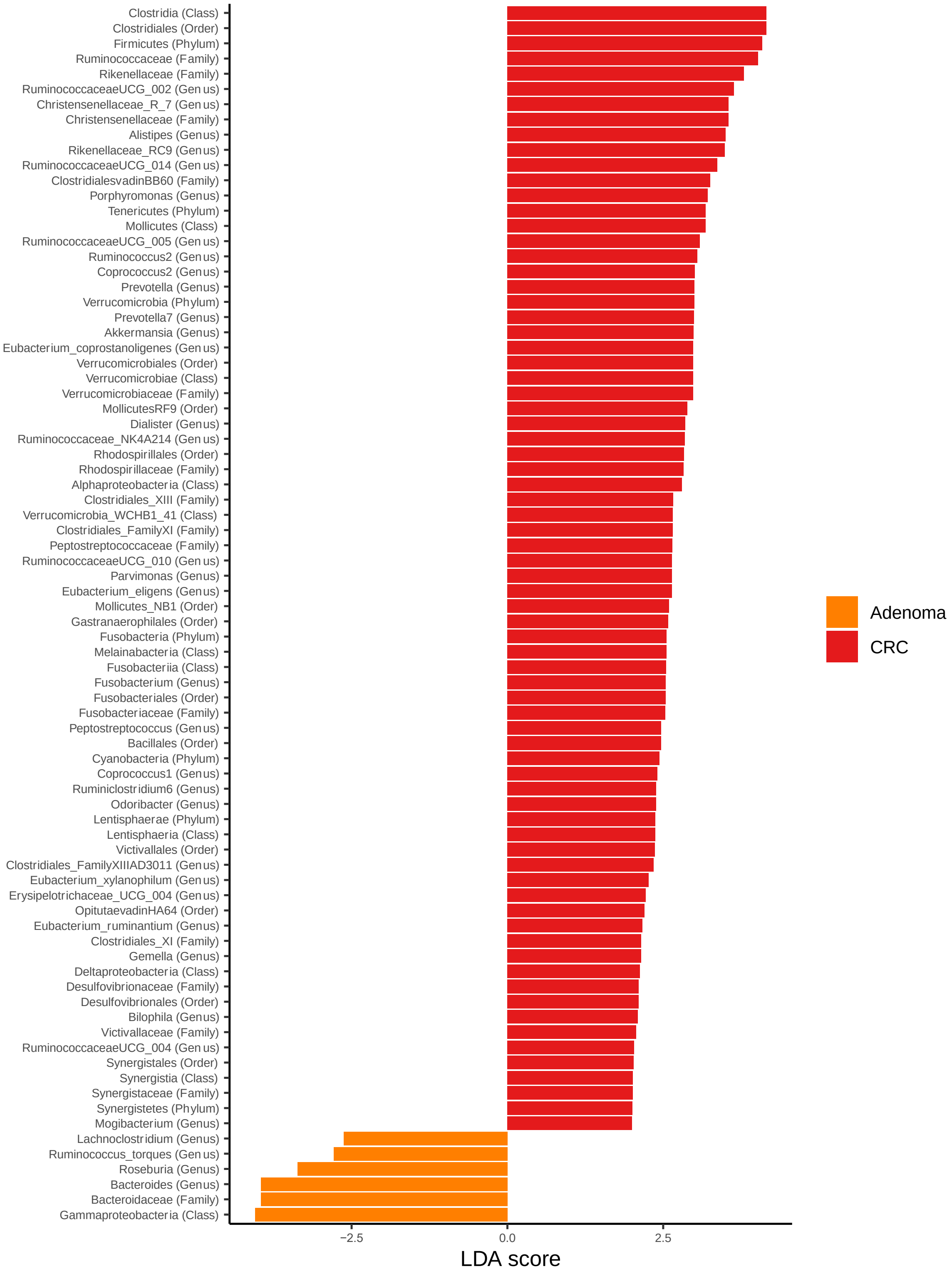
Supplementary Figure 3A. LEfSe plot of taxa enriched in blood-negative compared with colonoscopy-normal samples.



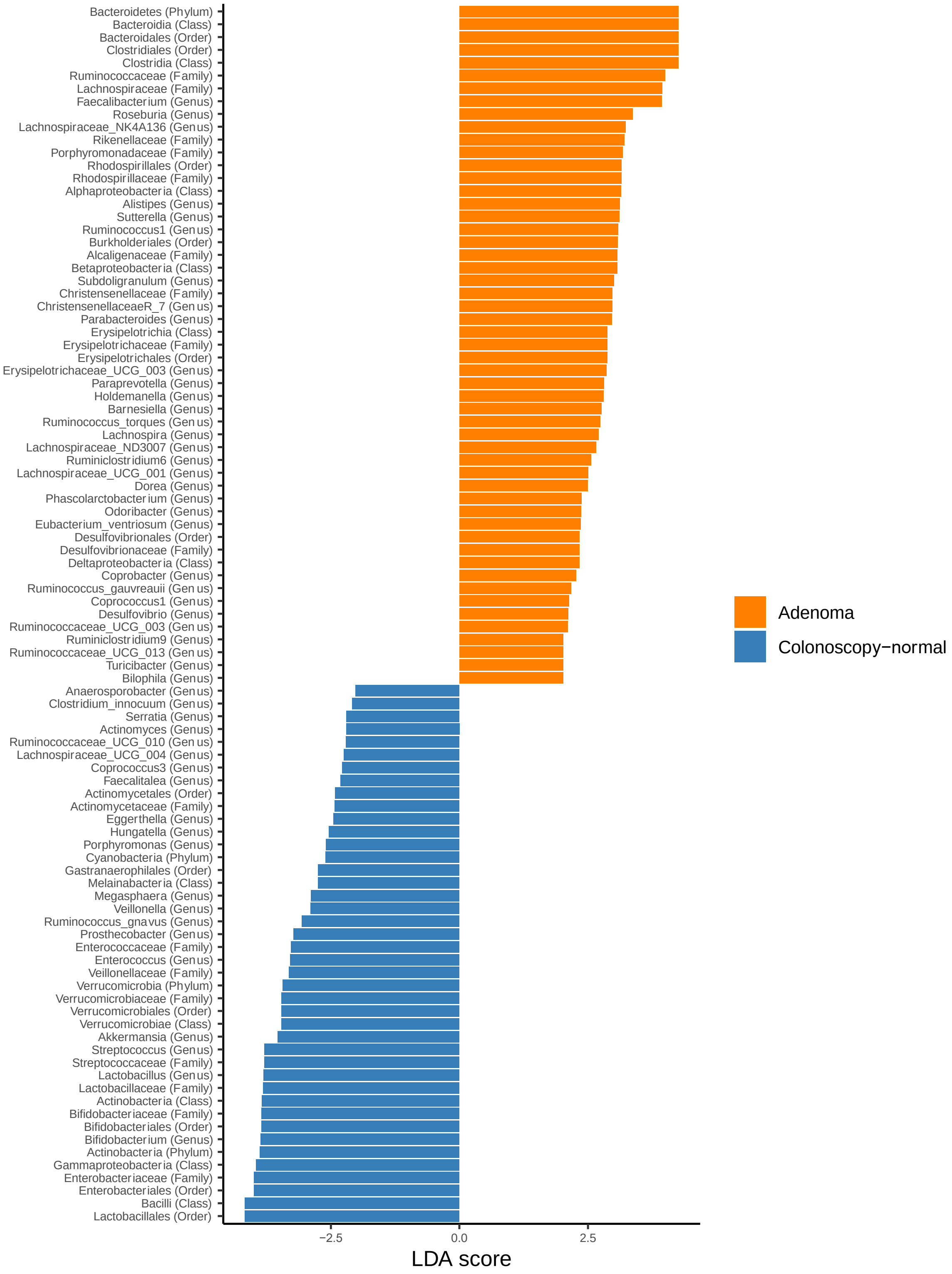
Supplementary Figure 3B. LEfSe plot of taxa enriched in CRC compared with blood-negative samples.



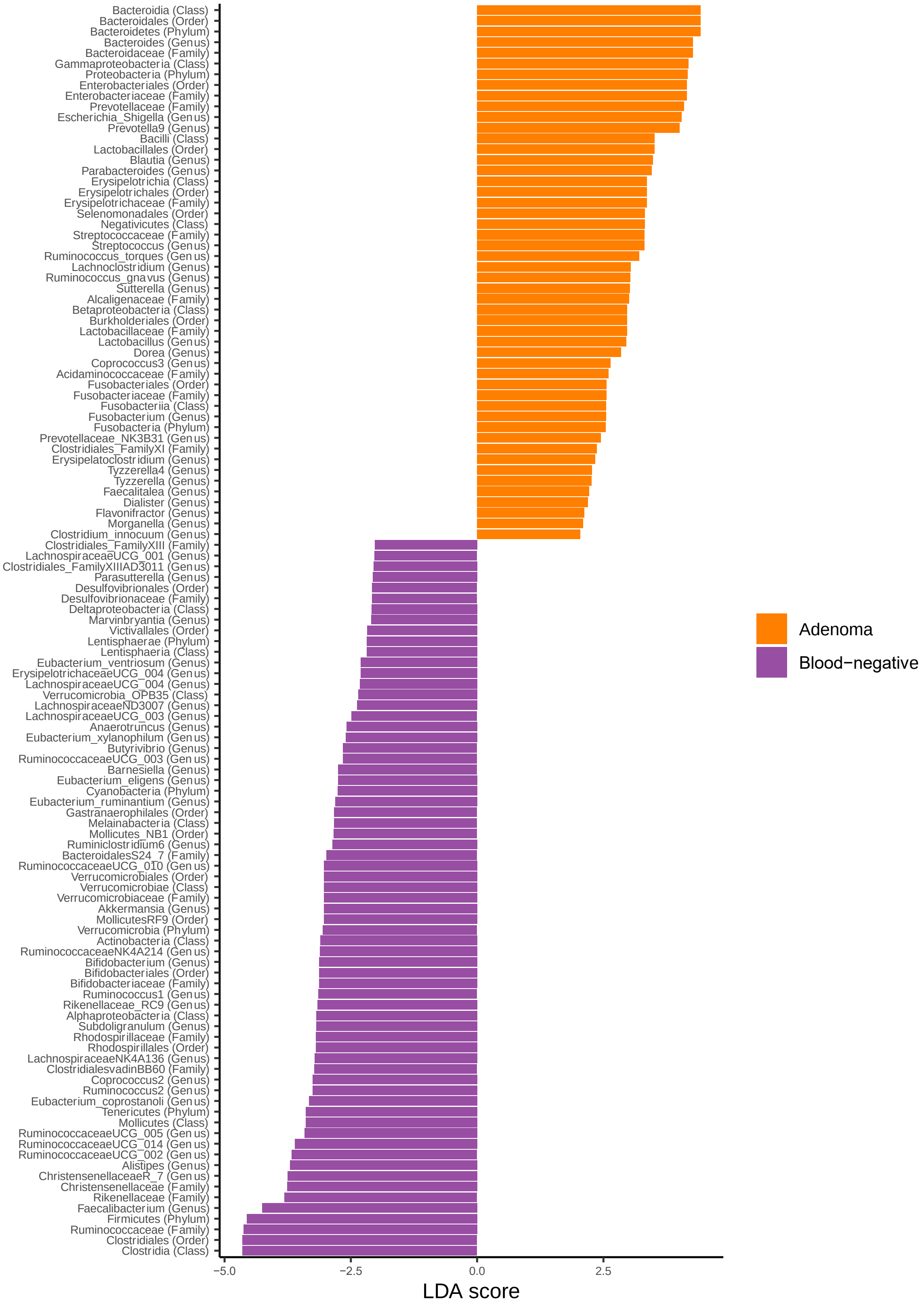
Supplementary Figure 3C. LEfSe plot of taxa enriched in CRC compared with colonoscopy-normal samples.



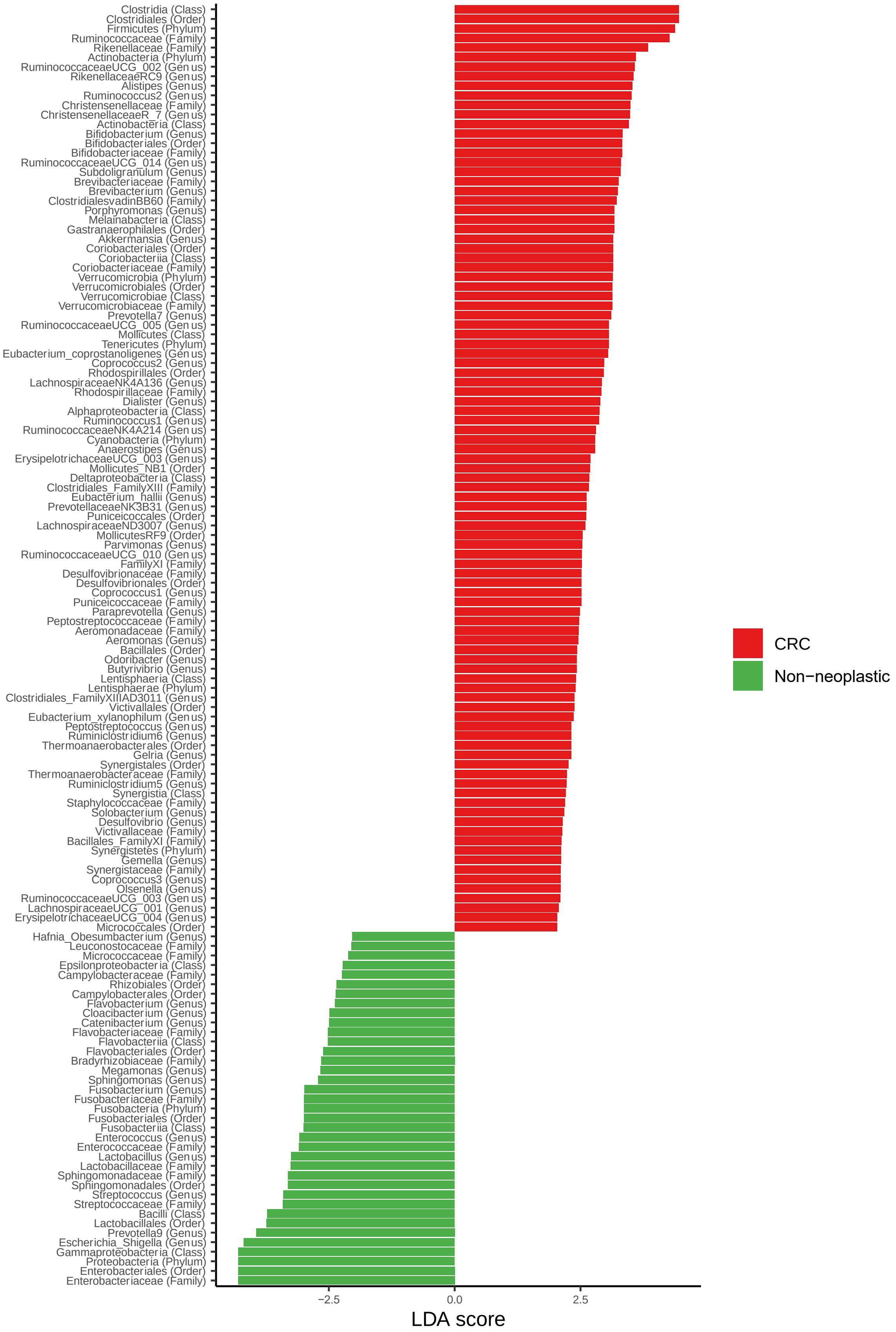
Supplementary Figure 3D. LEfSe plot of taxa enriched in adenoma compared with CRC samples.



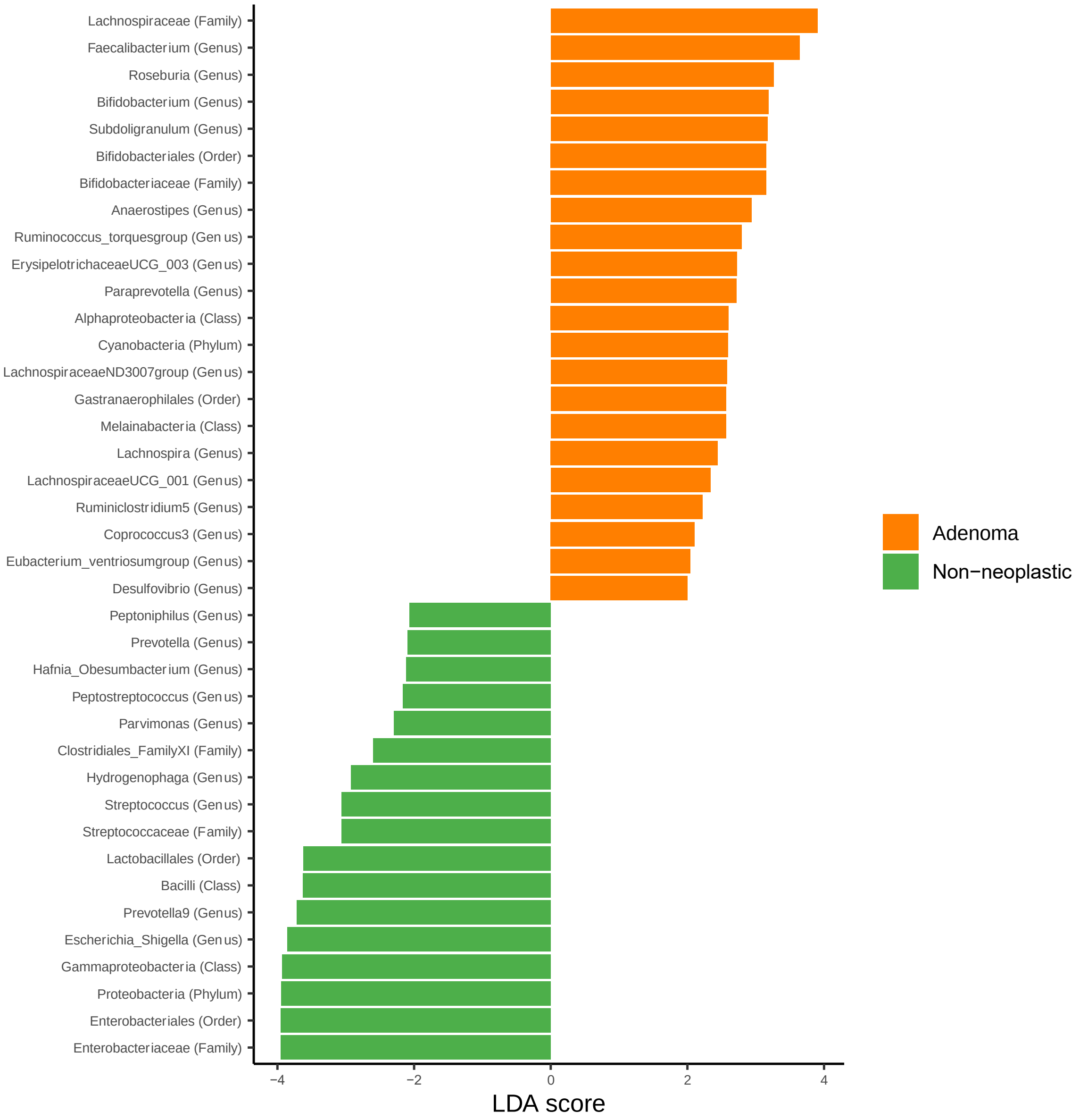
Supplementary Figure 3E. LEfSe plot of taxa enriched in adenoma compared with colonoscopy-normal samples.



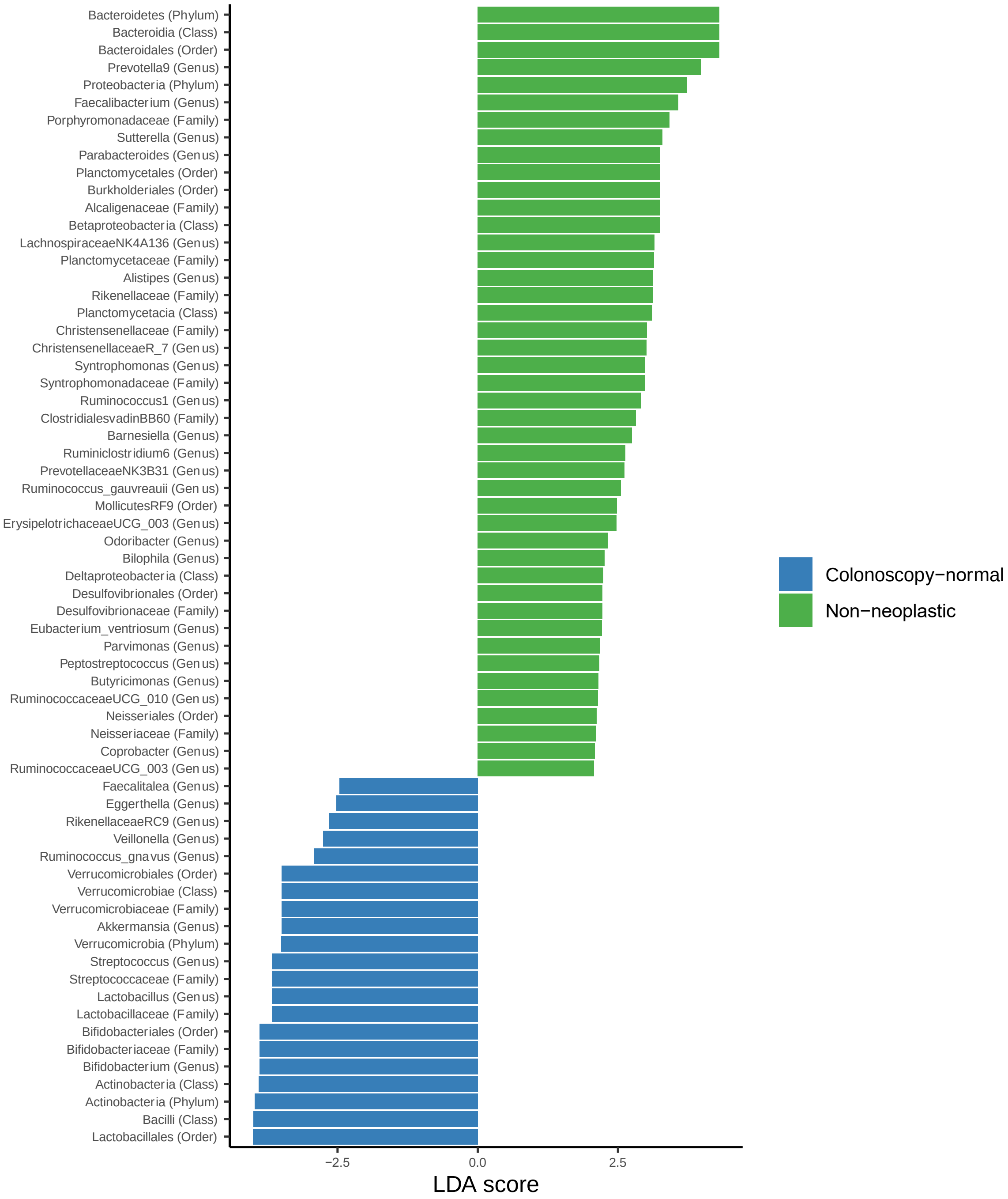
Supplementary Figure 3F. LEfSe plot of taxa enriched in adenoma compared with blood-negative samples.



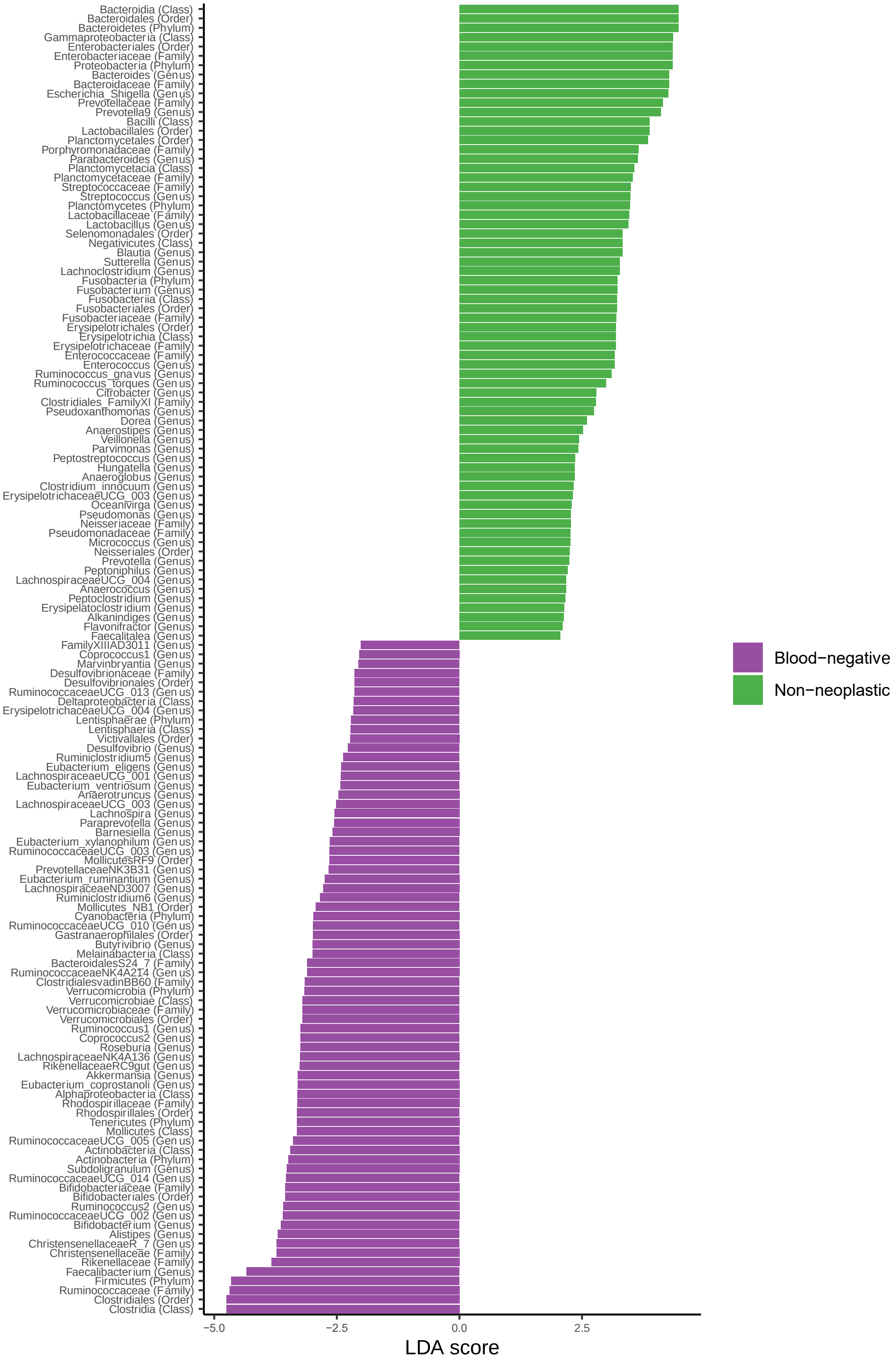
Supplementary Figure 3G. LEfSe plot of taxa enriched in CRC compared with non-neoplastic samples.



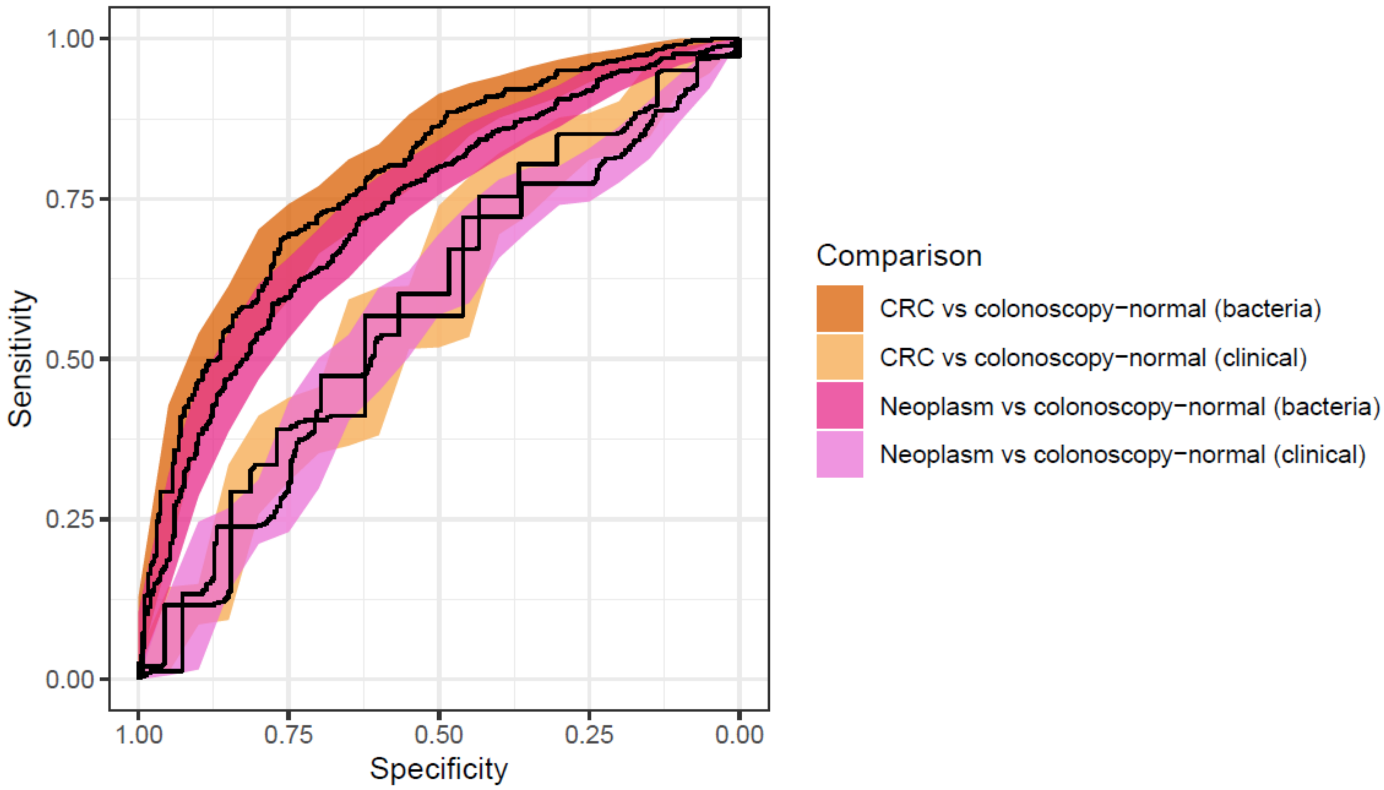
Supplementary Figure 3H. LEfSe plot of taxa enriched in adenoma compared with non-neoplastic samples.



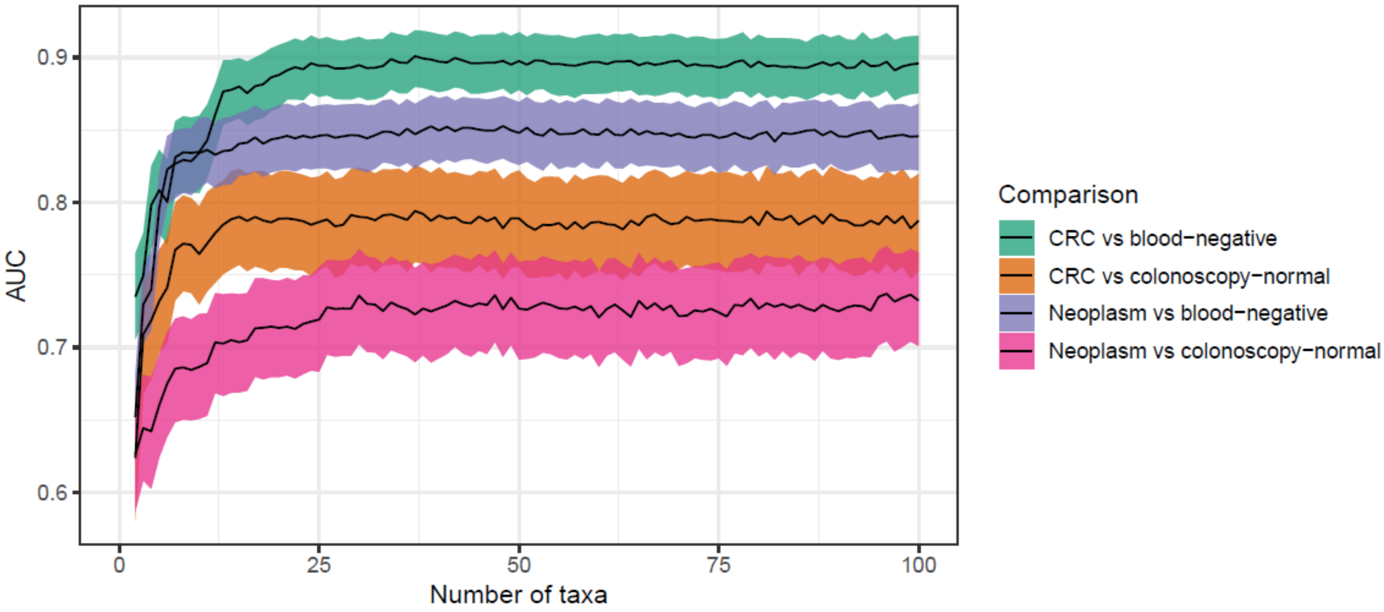
Supplementary Figure 3I. LEfSe plot of taxa enriched in colonoscopy-normal compared with non-neoplastic samples.



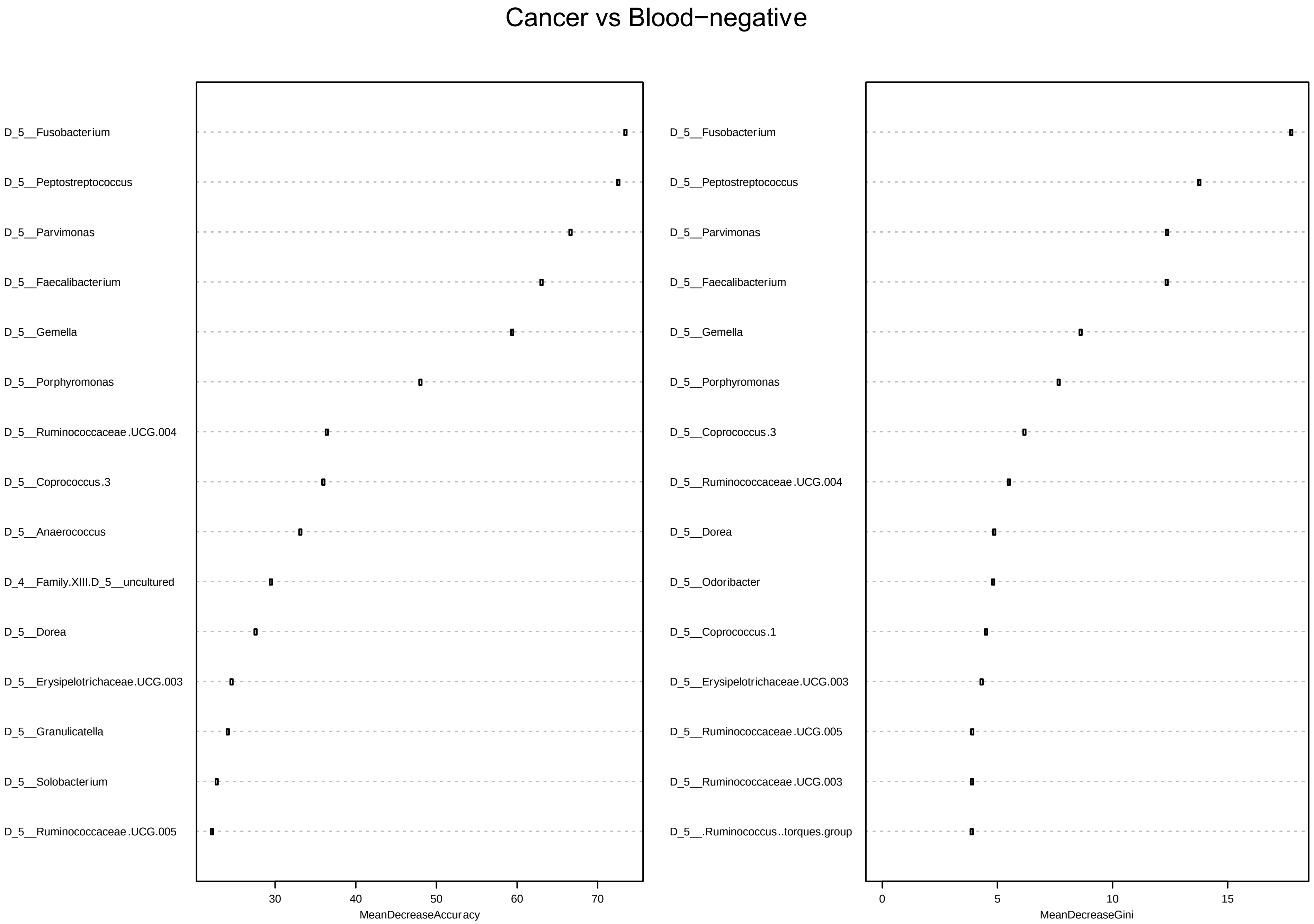
Supplementary Figure 3J. LEfSe plot of taxa enriched in blood-negative compared with non-neoplastic samples.



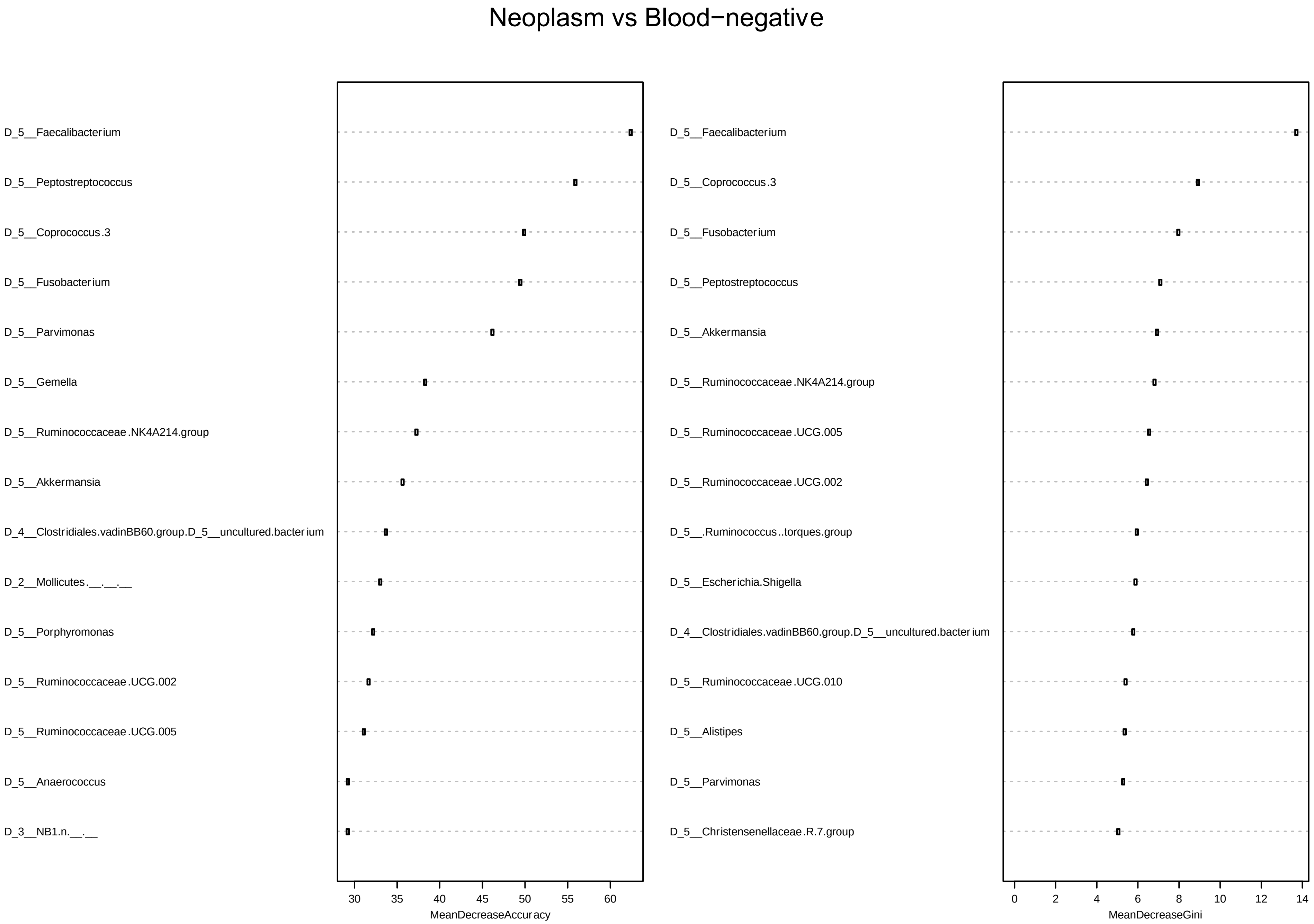
Supplementary Figure 4A. Comparison of ROC curves for RF models featuring 'colonoscopy-normal' as a comparison. These ROC curves represent the performance of the 'total' RF models. Shading represents the 95% CI. Clinical = age & sex. Neoplasm = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma samples. The performance of the 'bacteria' RF models is significantly superior to that of the 'clinical' models, created for comparison.



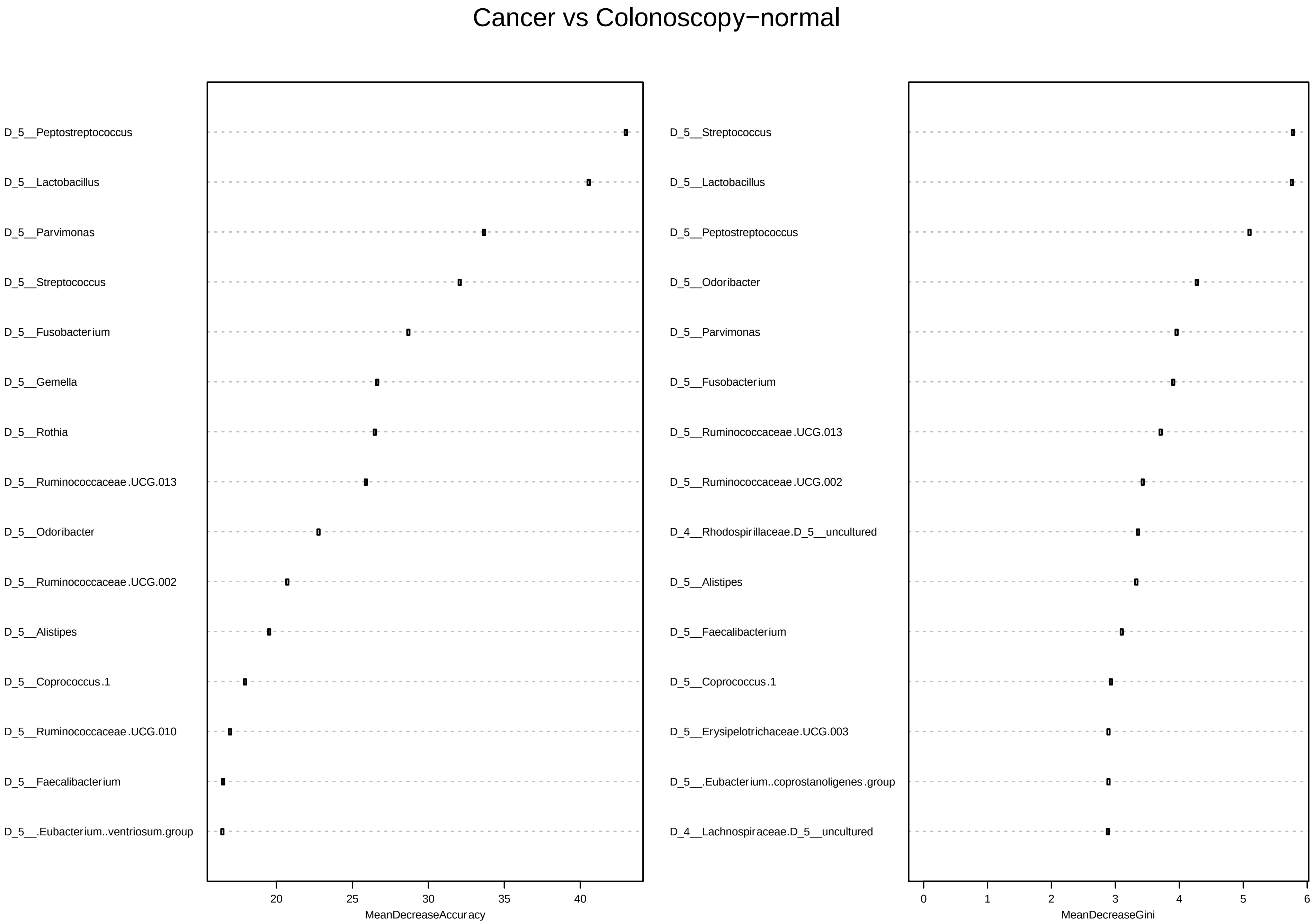
Supplementary Figure 4B. Improvement in RF model performance as the number of taxa available to the models increases. Genus-level bacteria only 'total' RF models were built using an increasing number of taxa of decreasing importance. Shading represents the 95% CI of the AUC. Neoplasm = a group comprising an approximately equal ratio of CRC, low-risk adenoma, intermediate-risk adenoma and high-risk adenoma samples. For each model, the AUC plateaus at approximately 15 taxa.



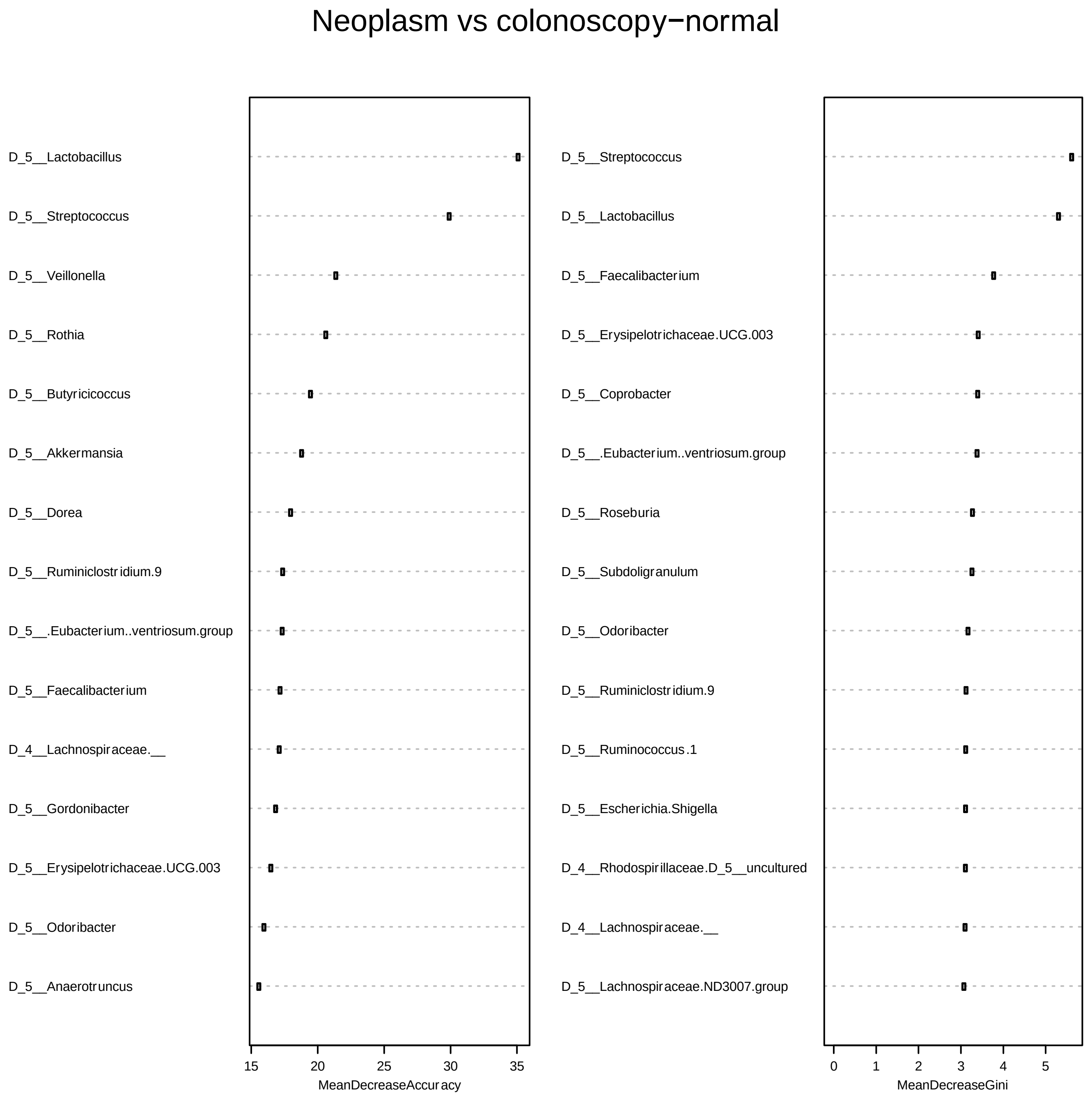
Supplementary Figure 4C. The 15 most important taxa from the bacteria “CRC vs blood-negative” total RF model.



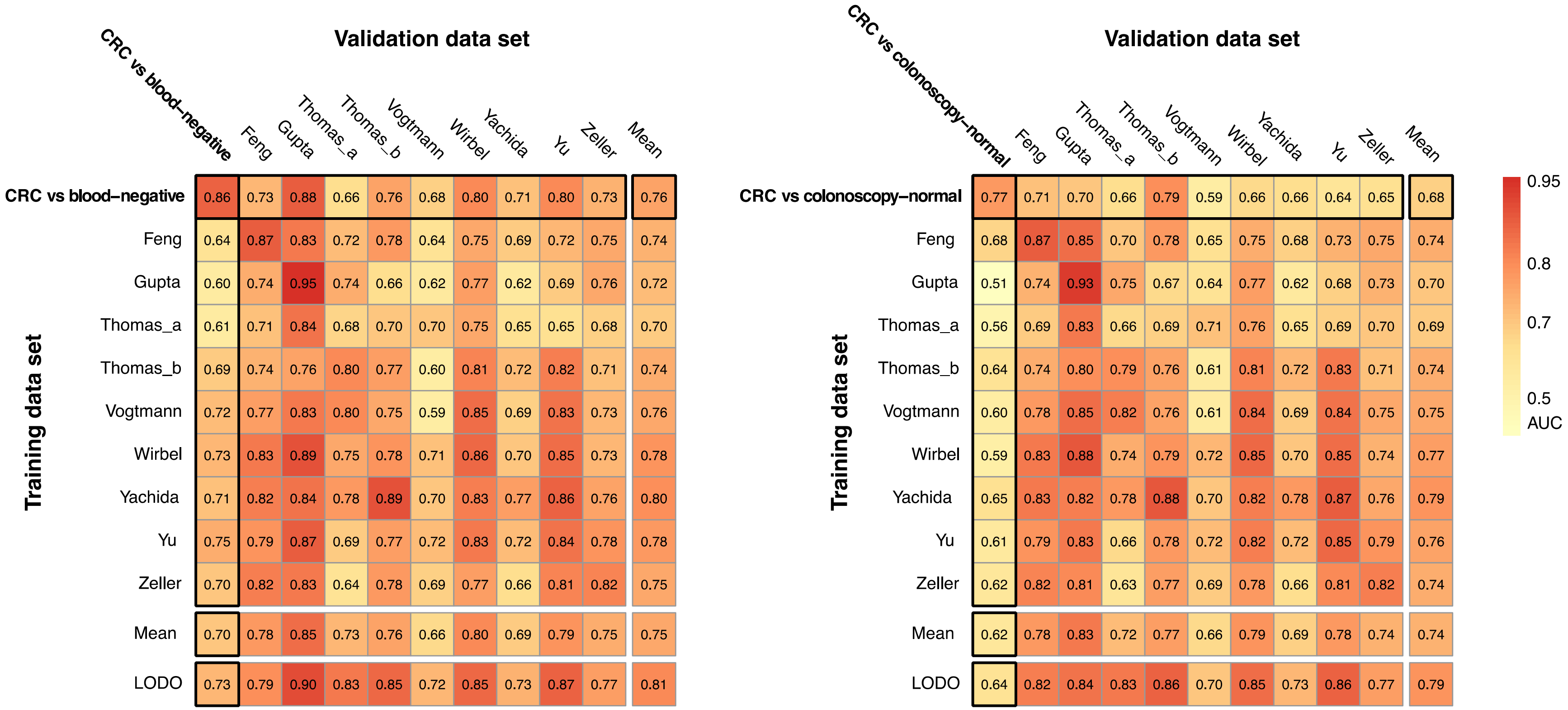
Supplementary Figure 4D. The 15 most important taxa from the bacteria "Neoplasm vs blood-negative" total RF model.



Supplementary Figure 4E. The 15 most important taxa from the bacteria "CRC vs colonoscopy-normal" total RF model.



Supplementary Figure 4F. The 15 most important taxa from the bacteria “Neoplasm vs colonoscopy-normal” total RF model.



Training data set	CRC vs blood–negative	Feng	Gupta	Thomas_a	Thomas_b	Vogtmann	Wirbel	Yachida	Yu	Zeller	Mean
CRC vs blood–negative	0.86	0.73	0.88	0.66	0.76	0.68	0.80	0.71	0.80	0.73	0.76
Feng	0.64	0.87	0.83	0.72	0.78	0.64	0.75	0.69	0.72	0.75	0.74
Gupta	0.60	0.74	0.95	0.74	0.66	0.62	0.77	0.62	0.69	0.76	0.72
Thomas_a	0.61	0.71	0.84	0.68	0.70	0.70	0.75	0.65	0.65	0.68	0.70
Thomas_b	0.69	0.74	0.76	0.80	0.77	0.60	0.81	0.72	0.82	0.71	0.74
Vogtmann	0.72	0.77	0.83	0.80	0.75	0.59	0.85	0.69	0.83	0.73	0.76
Wirbel	0.73	0.83	0.89	0.75	0.78	0.71	0.86	0.70	0.85	0.73	0.78
Yachida	0.71	0.82	0.84	0.78	0.89	0.70	0.83	0.77	0.86	0.76	0.80
Yu	0.75	0.79	0.87	0.69	0.77	0.72	0.83	0.72	0.84	0.78	0.78
Zeller	0.70	0.82	0.83	0.64	0.78	0.69	0.77	0.66	0.81	0.82	0.75
Mean	0.70	0.78	0.85	0.73	0.76	0.66	0.80	0.69	0.79	0.75	0.75
LODO	0.73	0.79	0.90	0.83	0.85	0.72	0.85	0.73	0.87	0.77	0.81

Training data set	CRC vs colonoscopy–normal	Feng	Gupta	Thomas_a	Thomas_b	Vogtmann	Wirbel	Yachida	Yu	Zeller	Mean
CRC vs colonoscopy–normal	0.77	0.71	0.70	0.66	0.79	0.59	0.66	0.66	0.64	0.65	0.68
Feng	0.68	0.87	0.85	0.70	0.78	0.65	0.75	0.68	0.73	0.75	0.74
Gupta	0.51	0.74	0.93	0.75	0.67	0.64	0.77	0.62	0.68	0.73	0.70
Thomas_a	0.56	0.69	0.83	0.66	0.69	0.71	0.76	0.65	0.69	0.70	0.69
Thomas_b	0.64	0.74	0.80	0.79	0.76	0.61	0.81	0.72	0.83	0.71	0.74
Vogtmann	0.60	0.78	0.85	0.82	0.76	0.61	0.84	0.69	0.84	0.75	0.75
Wirbel	0.59	0.83	0.88	0.74	0.79	0.72	0.85	0.70	0.85	0.74	0.77
Yachida	0.65	0.83	0.82	0.78	0.88	0.70	0.82	0.78	0.87	0.76	0.79
Yu	0.61	0.79	0.83	0.66	0.78	0.72	0.82	0.72	0.85	0.79	0.76
Zeller	0.62	0.82	0.81	0.63	0.77	0.69	0.78	0.66	0.81	0.82	0.74
Mean	0.62	0.78	0.83	0.72	0.77	0.66	0.79	0.69	0.78	0.74	0.74
LODO	0.64	0.82	0.84	0.83	0.86	0.70	0.85	0.73	0.86	0.77	0.79

Supplementary Figure 5A. Model performance compared with external metagenomic datasets. Performance of the bacteria “CRC vs blood-negative” and “CRC vs colonoscopy-normal” total RF models, compared to models built using external faecal metagenomic datasets. The matrices display cross-prediction AUCs. LODO (leave-one-dataset-out) denotes AUC generated by training a model using all but the dataset of the associated column and testing it using the dataset of that column.

Supplementary Figure 5B. Model performance compared with external metagenomic datasets. Performance of the bacteria "CRC vs blood-negative" total RF model, compared to models built using external faecal metagenomic datasets. For each test/validation pair of cohorts, confusion matrices were created using the predict function of randomForest using the default vote proportion cutoff of 50%. Sensitivity was calculated as the proportion of CRC samples called as CRC within the validation dataset, based on the test dataset RF model. Specificity was calculated as the proportion of control samples called as control. For the self-validation comparisons, the mean sensitivity and specificity of the 20 repetitions was recorded.

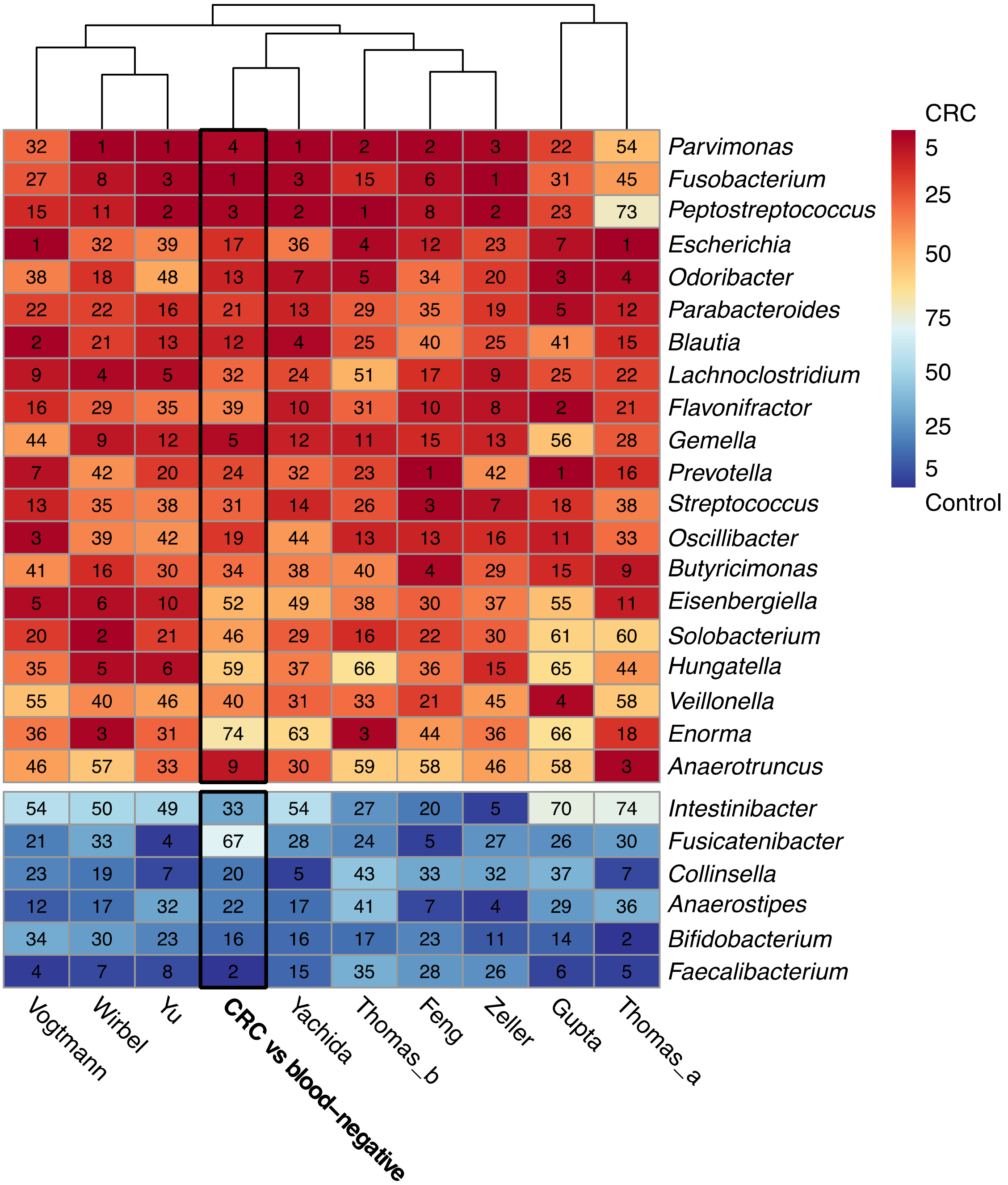
Sensitivity	CRC vs blood-negative	Feng	Gupta	Thomas_a	Thomas_b	Vogtmann	Wirbel	Yachida	Yu	Zeller
CRC vs blood-negative	0.72	0.91	0.97	0.76	0.94	0.83	0.77	0.96	0.95	0.94
Feng	1.00	0.77	0.70	0.76	0.78	0.90	0.78	0.67	0.79	0.79
Gupta	0.86	0.80	0.78	0.92	0.81	1.00	0.87	0.90	0.92	0.96
Thomas_a	0.75	0.37	0.80	0.69	0.44	0.79	0.53	0.55	0.49	0.60
Thomas_b	0.39	0.52	0.33	0.28	0.64	0.29	0.47	0.39	0.53	0.66
Vogtmann	0.82	0.61	0.47	0.52	0.63	0.60	0.58	0.66	0.73	0.72
Wirbel	0.72	0.85	0.53	0.44	0.78	0.77	0.68	0.76	0.91	0.87
Yachida	0.65	0.61	0.43	0.32	0.72	0.62	0.57	0.60	0.65	0.72
Yu	0.75	0.74	0.73	0.44	0.59	0.62	0.53	0.65	0.80	0.77
Zeller	0.72	0.52	1.00	1.00	0.53	0.71	0.70	0.82	0.85	0.67

Specificity	CRC vs blood-negative	Feng	Gupta	Thomas_a	Thomas_b	Vogtmann	Wirbel	Yachida	Yu	Zeller
CRC vs blood-negative	0.84	0.23	0.27	0.43	0.14	0.27	0.69	0.14	0.32	0.11
Feng	0.00	0.86	0.80	0.43	0.61	0.23	0.55	0.57	0.40	0.46
Gupta	0.22	0.44	0.89	0.43	0.21	0.08	0.35	0.12	0.25	0.15
Thomas_a	0.40	0.80	0.77	0.62	0.82	0.56	0.83	0.68	0.70	0.66
Thomas_b	0.96	0.82	0.97	1.00	0.74	0.90	0.97	0.91	0.94	0.69
Vogtmann	0.37	0.85	0.93	0.86	0.64	0.53	0.95	0.59	0.75	0.57
Wirbel	0.56	0.59	0.97	0.81	0.50	0.48	0.81	0.48	0.49	0.44
Yachida	0.65	0.90	0.93	0.95	0.93	0.81	0.95	0.80	0.91	0.67
Yu	0.59	0.67	0.97	0.76	0.71	0.77	0.95	0.63	0.74	0.66
Zeller	0.48	0.92	0.03	0.00	0.89	0.44	0.66	0.33	0.49	0.81

Supplementary Figure 5C. Model performance compared with external metagenomic datasets. Performance of the bacteria "CRC vs colonoscopy-normal" total RF model, compared to models built using external faecal metagenomic datasets. For each test/validation pair of cohorts, confusion matrices were created using the predict function of randomForest using the default vote proportion cutoff of 50%. Sensitivity was calculated as the proportion of CRC samples called as CRC within the validation dataset, based on the test dataset RF model. Specificity was calculated as the proportion of control samples called as control. For the self-validation comparisons, the mean sensitivity and specificity of the 20 repetitions was recorded.

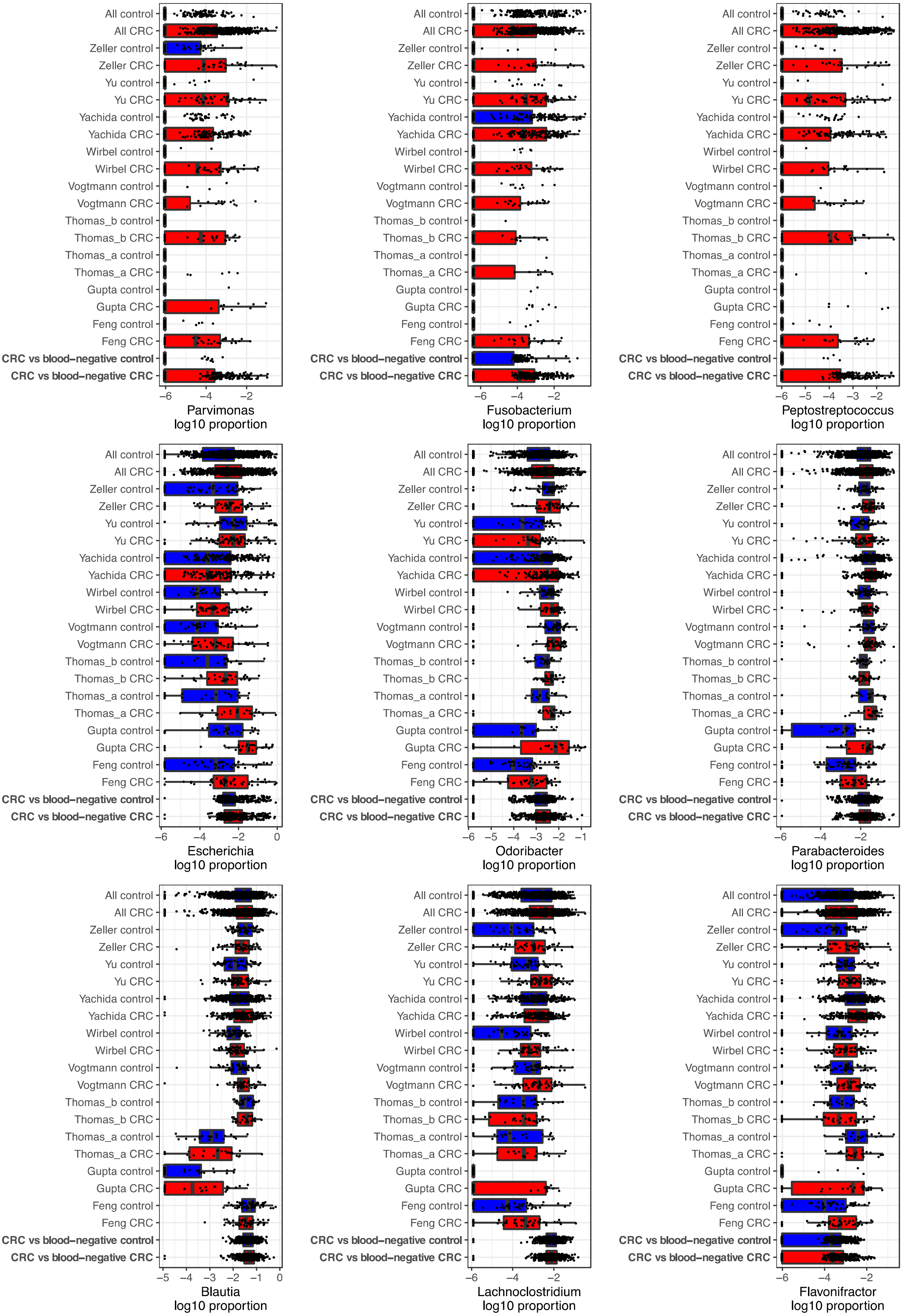
Sensitivity	CRC vs colonoscopy-normal	Feng	Gupta	Thomas_a	Thomas_b	Vogtmann	Wirbel	Yachida	Yu	Zeller
CRC vs colonoscopy-normal	0.76	0.33	0.07	0.24	0.66	0.42	0.52	0.33	0.40	0.55
Feng	1.00	0.72	0.80	0.76	0.78	0.94	0.80	0.67	0.80	0.77
Gupta	0.87	0.85	0.71	0.92	0.81	1.00	0.92	0.90	0.93	0.96
Thomas_a	0.78	0.33	0.70	0.65	0.41	0.79	0.52	0.62	0.53	0.68
Thomas_b	0.39	0.61	0.37	0.32	0.63	0.29	0.50	0.39	0.57	0.66
Vogtmann	0.81	0.57	0.57	0.56	0.66	0.53	0.60	0.61	0.75	0.75
Wirbel	0.69	0.85	0.53	0.48	0.81	0.83	0.68	0.78	0.89	0.87
Yachida	0.62	0.59	0.53	0.36	0.66	0.54	0.57	0.60	0.71	0.72
Yu	0.77	0.70	0.60	0.44	0.63	0.60	0.55	0.64	0.79	0.77
Zeller	0.72	0.50	1.00	1.00	0.53	0.71	0.73	0.83	0.88	0.68

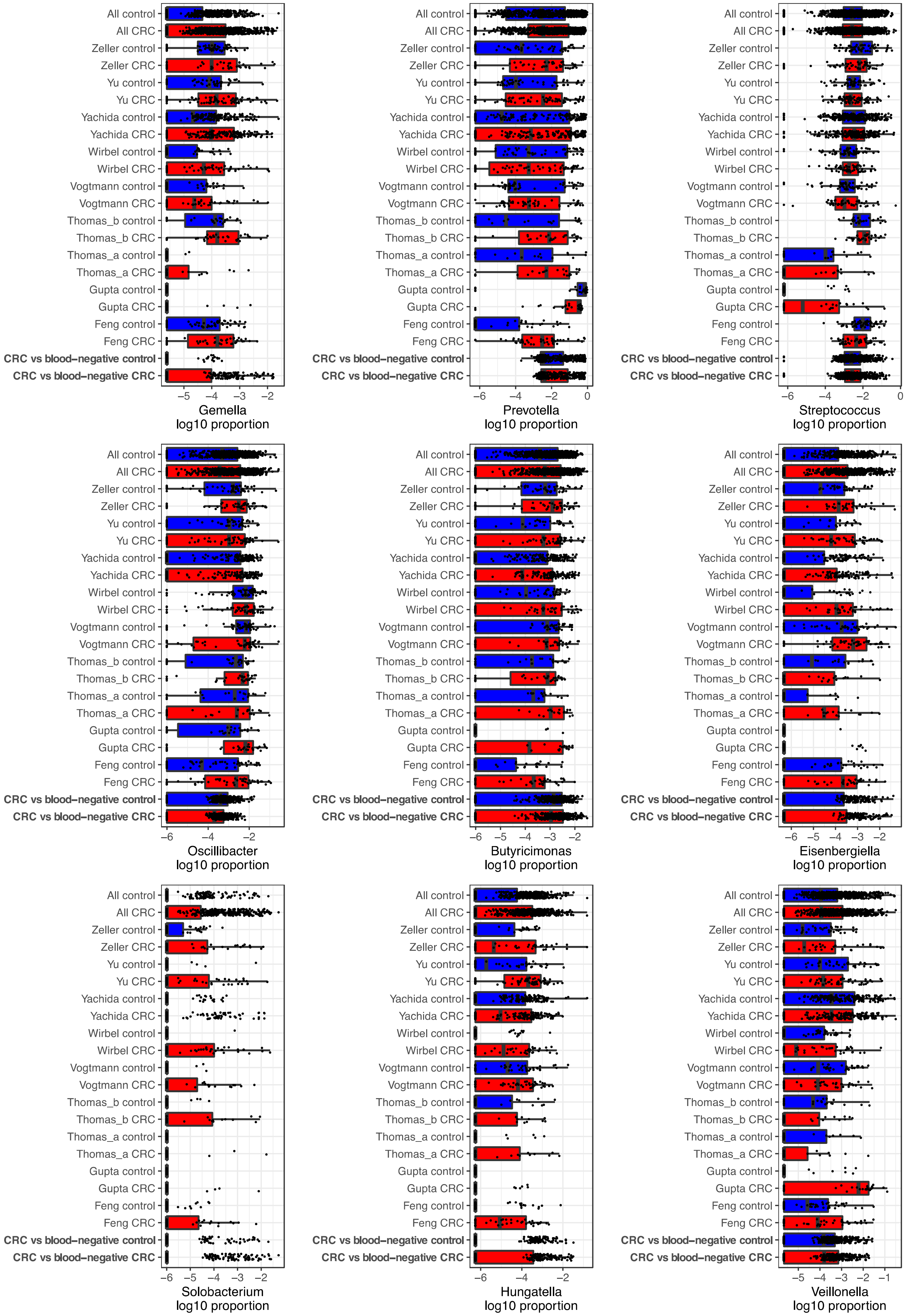
Specificity	CRC vs colonoscopy-normal	Feng	Gupta	Thomas_a	Thomas_b	Vogtmann	Wirbel	Yachida	Yu	Zeller
CRC vs colonoscopy-normal	0.64	0.98	1.00	1.00	0.89	0.73	0.72	0.86	0.77	0.74
Feng	0.03	0.89	0.83	0.43	0.57	0.25	0.55	0.55	0.40	0.48
Gupta	0.15	0.34	0.86	0.48	0.14	0.06	0.23	0.11	0.13	0.11
Thomas_a	0.33	0.79	0.77	0.60	0.75	0.52	0.85	0.65	0.72	0.57
Thomas_b	0.90	0.77	1.00	1.00	0.77	0.92	0.98	0.90	0.98	0.66
Vogtmann	0.22	0.90	0.93	0.86	0.68	0.64	0.92	0.65	0.74	0.62
Wirbel	0.34	0.57	0.97	0.76	0.50	0.48	0.84	0.45	0.47	0.44
Yachida	0.55	0.95	0.90	0.95	0.93	0.83	0.95	0.80	0.92	0.70
Yu	0.35	0.67	0.83	0.62	0.71	0.77	0.95	0.63	0.74	0.64
Zeller	0.38	0.92	0.07	0.00	0.86	0.44	0.69	0.32	0.49	0.81

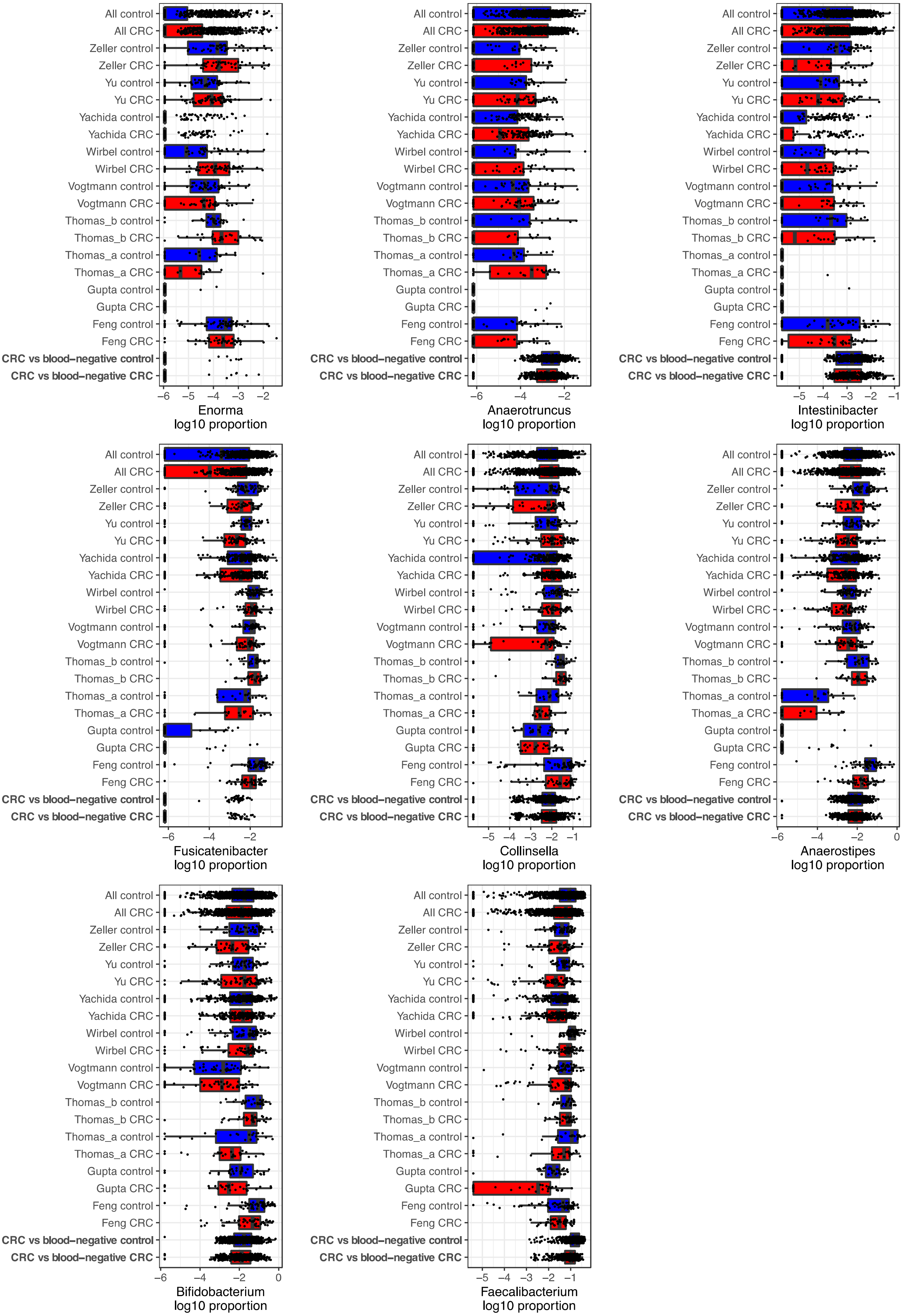


	Vogtmann	Wirbel	Yu	CRC vs blood–negative	Yachida	Thomas_b	Feng	Zeller	Gupta	Thomas_a
Parvimonas	32	1	1	4	1	2	2	3	22	54
Fusobacterium	27	8	3	1	3	15	6	1	31	45
Peptostreptococcus	15	11	2	3	2	1	8	2	23	73
Escherichia	1	32	39	17	36	4	12	23	7	1
Odoribacter	38	18	48	13	7	5	34	20	3	4
Parabacteroides	22	22	16	21	13	29	35	19	5	12
Blautia	2	21	13	12	4	25	40	25	41	15
Lachnoclostridium	9	4	5	32	24	51	17	9	25	22
Flavonifractor	16	29	35	39	10	31	10	8	2	21
Gemella	44	9	12	5	12	11	15	13	56	28
Prevotella	7	42	20	24	32	23	1	42	1	16
Streptococcus	13	35	38	31	14	26	3	7	18	38
Oscillibacter	3	39	42	19	44	13	13	16	11	33
Butyricimonas	41	16	30	34	38	40	4	29	15	9
Eisenbergiella	5	6	10	52	49	38	30	37	55	11
Solobacterium	20	2	21	46	29	16	22	30	61	60
Hungatella	35	5	6	59	37	66	36	15	65	44
Veillonella	55	40	46	40	31	33	21	45	4	58
Enorma	36	3	31	74	63	3	44	36	66	18
Anaerotruncus	46	57	33	9	30	59	58	46	58	3
Intestinibacter	54	50	49	33	54	27	20	5	70	74
Fusicatenibacter	21	33	4	67	28	24	5	27	26	30
Collinsella	23	19	7	20	5	43	33	32	37	7
Anaerostipes	12	17	32	22	17	41	7	4	29	36
Bifidobacterium	34	30	23	16	16	17	23	11	14	2
Faecalibacterium	4	7	8	2	15	35	28	26	6	5

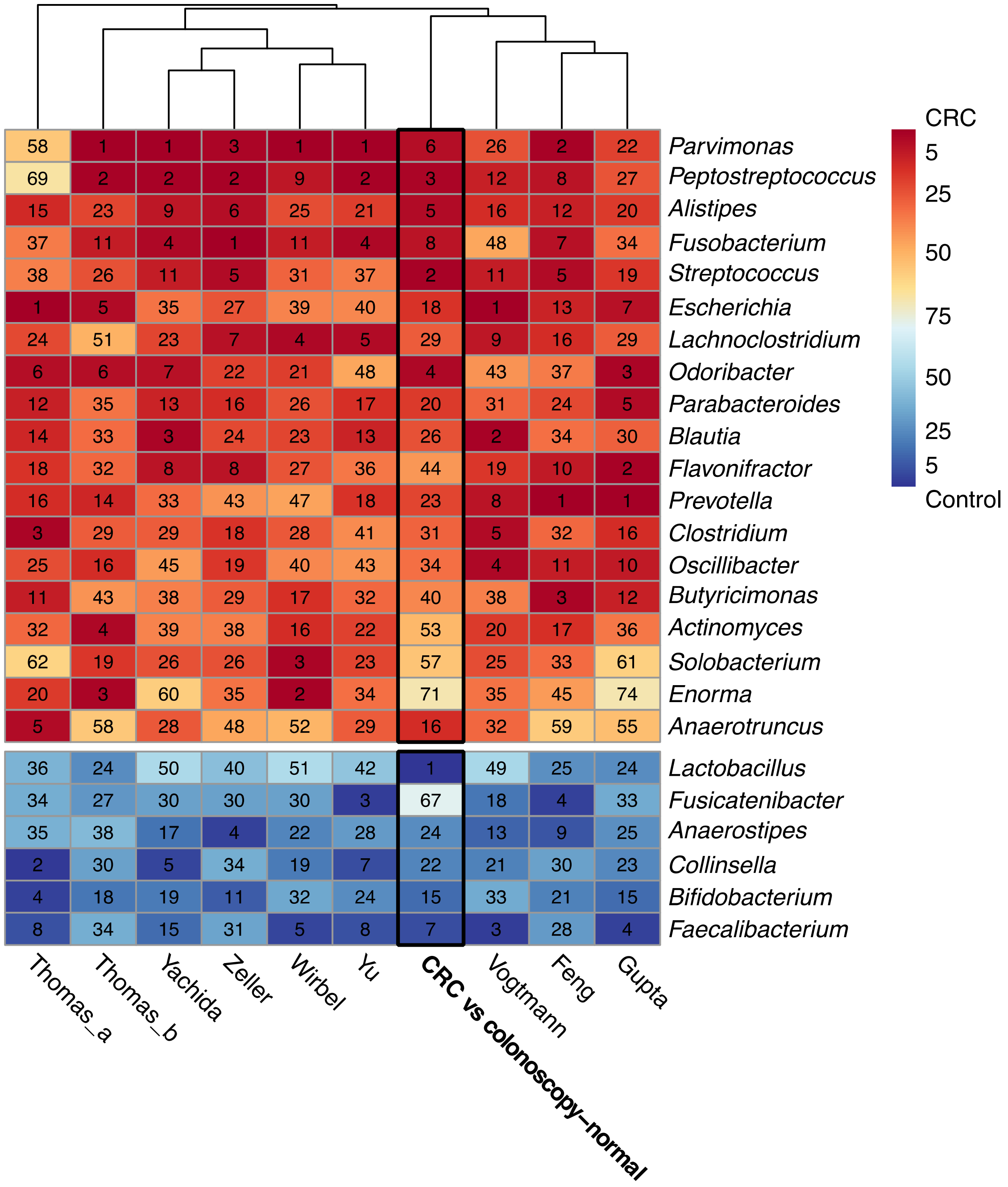
Supplementary Figure 5D. Comparison of RF feature ranking across datasets. The bacteria "CRC vs blood-negative" total RF model compared to external faecal metagenomic datasets. Importance of each genus for cross-validation performance using gini values. Only genera in the top five in at least one dataset are displayed.


All control
All CRC
Zeller control
Zeller CRC
Yu control
Yu CRC
Yachida control
Yachida CRC
Wirbel control
Wirbel CRC
Vogtmann control
Vogtmann CRC
Thomas_b control
Thomas_b CRC
Thomas_a control
Thomas_a CRC
Gupta control
Gupta CRC
Feng control
Feng CRC
CRC vs blood−negative control
CRC vs blood−negative CRC
−6 −4 −2
Parvimonas
log10 proportion
−6 −4 −2
Fusobacterium
log10 proportion
−6 −5 −4 −3 −2
Peptostreptococcus
log10 proportion
−6 −4 −2 0
Escherichia
log10 proportion
−6 −5 −4 −3 −2 −1
Odoribacter
log10 proportion
−6 −4 −2
Parabacteroides
log10 proportion
−5 −4 −3 −2 −1 0
Blautia
log10 proportion
−6 −4 −2
Lachnoclostridium
log10 proportion
−6 −4 −2
Flavonifractor
log10 proportion


All control
All CRC
Zeller control
Zeller CRC
Yu control
Yu CRC
Yachida control
Yachida CRC
Wirbel control
Wirbel CRC
Vogtmann control
Vogtmann CRC
Thomas_b control
Thomas_b CRC
Thomas_a control
Thomas_a CRC
Gupta control
Gupta CRC
Feng control
Feng CRC
CRC vs blood−negative control
CRC vs blood−negative CRC
−5 −4 −3 −2
Gemella
log10 proportion
−6 −4 −2 0
Prevotella
log10 proportion
−6 −4 −2 0
Streptococcus
log10 proportion
−6 −4 −2
Oscillibacter
log10 proportion
−6 −5 −4 −3 −2
Butyricimonas
log10 proportion
−6 −5 −4 −3 −2
Eisenbergiella
log10 proportion
−6 −5 −4 −3 −2
Solobacterium
log10 proportion
−6 −4 −2
Hungatella
log10 proportion
−5 −4 −3 −2 −1
Veillonella
log10 proportion

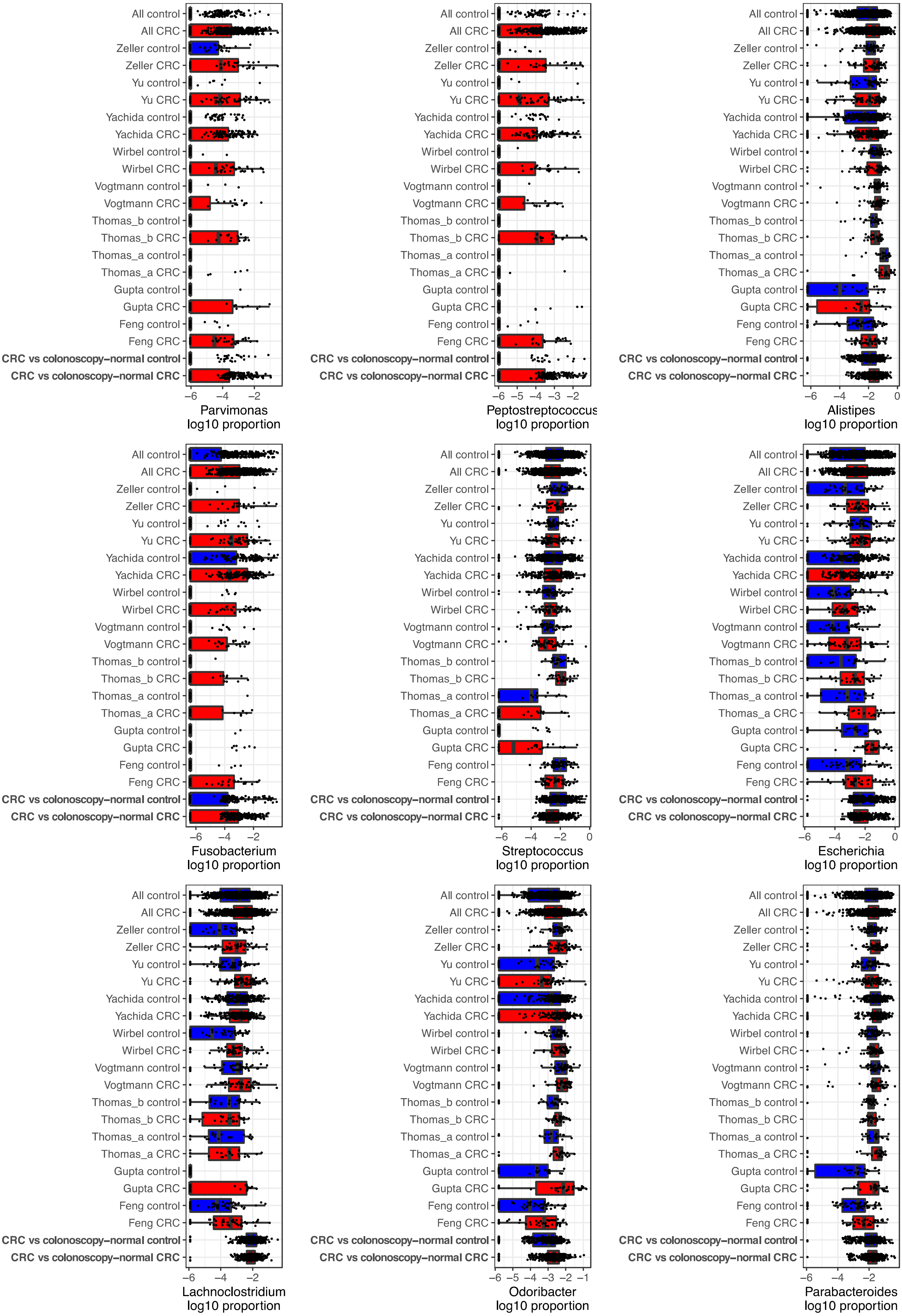


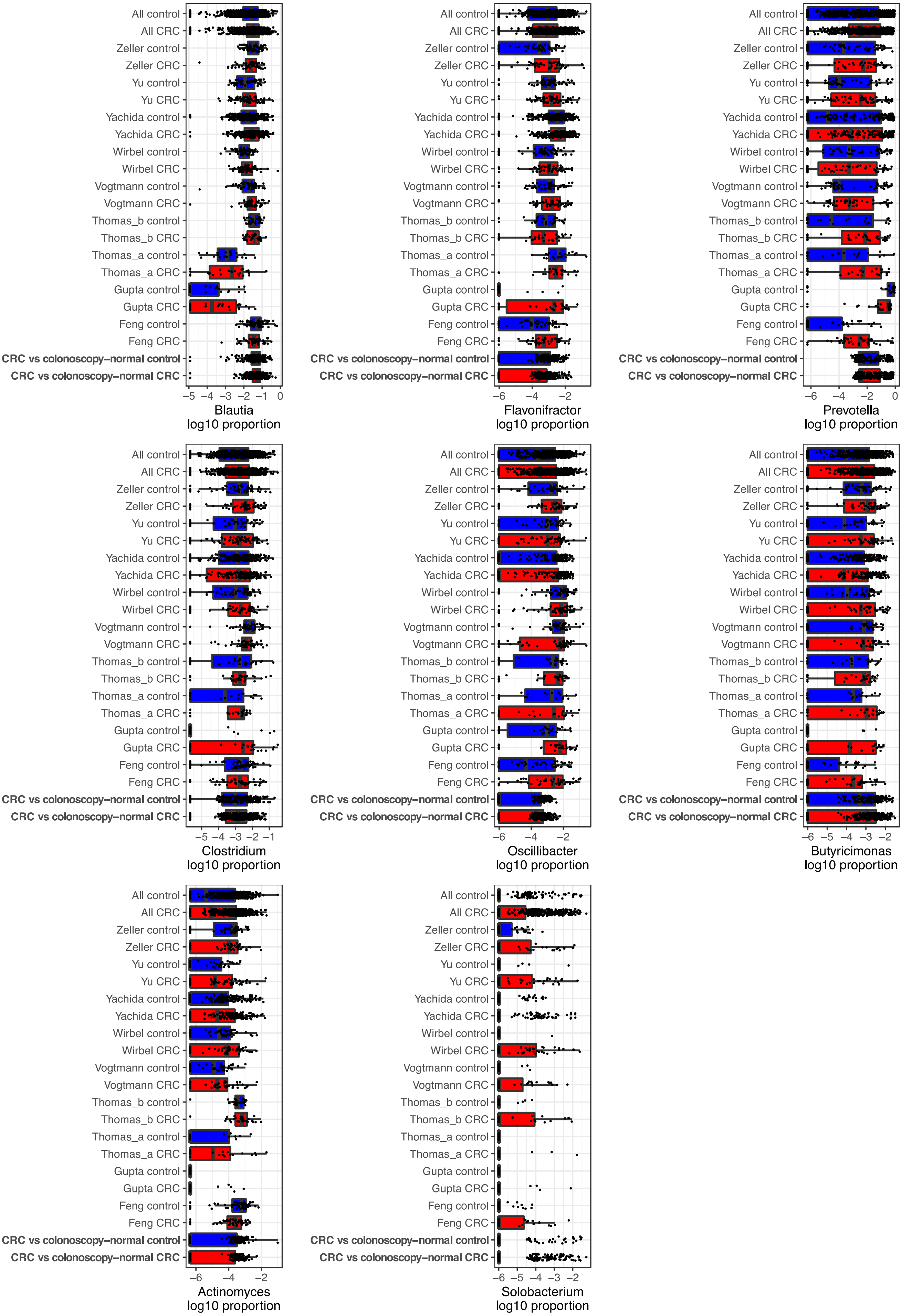
Supplementary Figure 5E. Distributions of relative abundance of genera of greatest importance. The boxplots labelled ‘All’ are a summary of all of the studies.

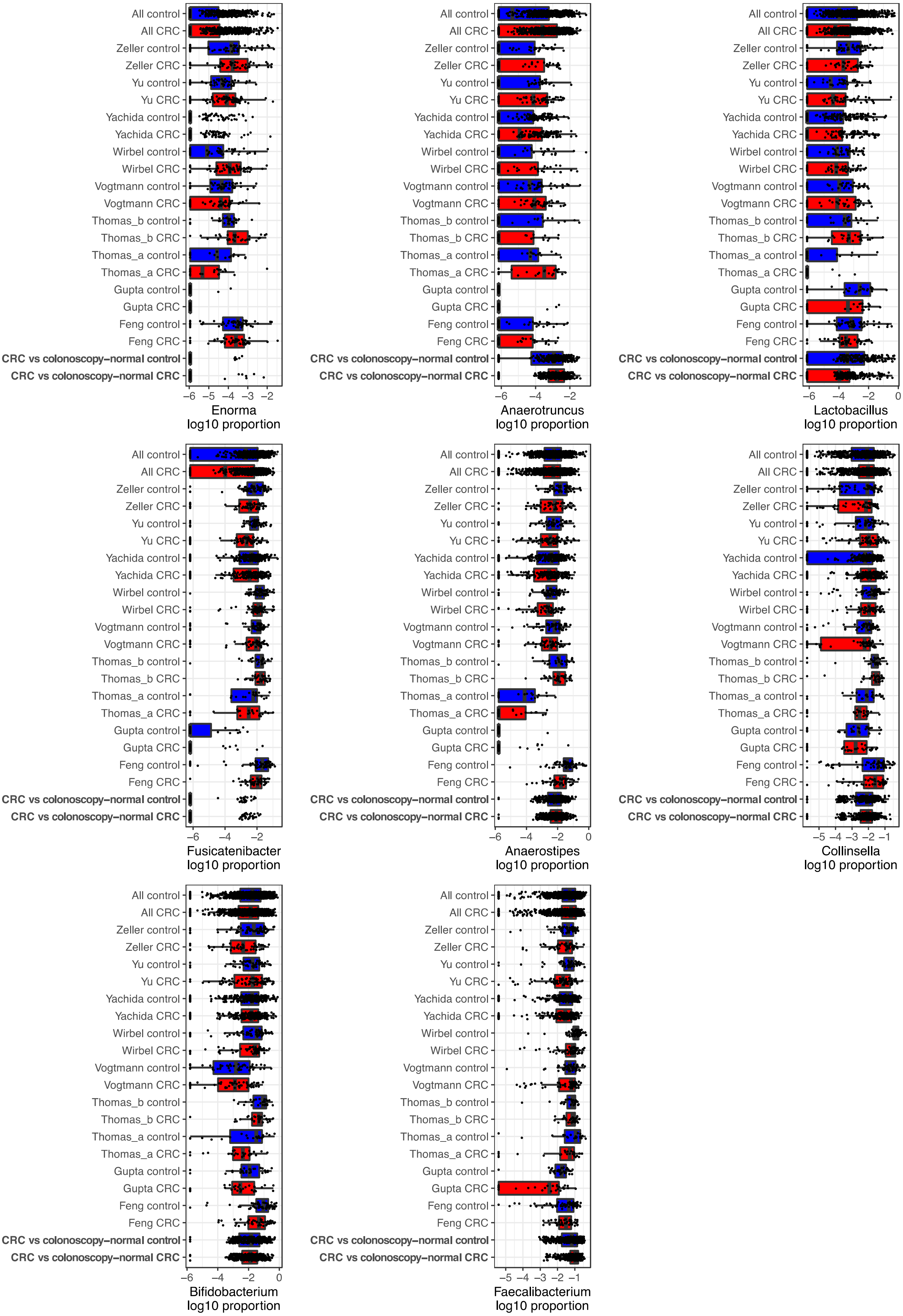


	Thomas_a	Thomas_b	Yachida	Zeller	Wirbel	Yu	CRC vs colonoscopy–normal	Vogtmann	Feng	Gupta
Parvimonas	58	1	1	3	1	1	6	26	2	22
Peptostreptococcus	69	2	2	2	9	2	3	12	8	27
Alistipes	15	23	9	6	25	21	5	16	12	20
Fusobacterium	37	11	4	1	11	4	8	48	7	34
Streptococcus	38	26	11	5	31	37	2	11	5	19
Escherichia	1	5	35	27	39	40	18	1	13	7
Lachnoclostridium	24	51	23	7	4	5	29	9	16	29
Odoribacter	6	6	7	22	21	48	4	43	37	3
Parabacteroides	12	35	13	16	26	17	20	31	24	5
Blautia	14	33	3	24	23	13	26	2	34	30
Flavonifractor	18	32	8	8	27	36	44	19	10	2
Prevotella	16	14	33	43	47	18	23	8	1	1
Clostridium	3	29	29	18	28	41	31	5	32	16
Oscillibacter	25	16	45	19	40	43	34	4	11	10
Butyricimonas	11	43	38	29	17	32	40	38	3	12
Actinomyces	32	4	39	38	16	22	53	20	17	36
Solobacterium	62	19	26	26	3	23	57	25	33	61
Enorma	20	3	60	35	2	34	71	35	45	74
Anaerotruncus	5	58	28	48	52	29	16	32	59	55
Lactobacillus	36	24	50	40	51	42	1	49	25	24
Fusicatenibacter	34	27	30	30	30	3	67	18	4	33
Anaerostipes	35	38	17	4	22	28	24	13	9	25
Collinsella	2	30	5	34	19	7	22	21	30	23
Bifidobacterium	4	18	19	11	32	24	15	33	21	15
Faecalibacterium	8	34	15	31	5	8	7	3	28	4

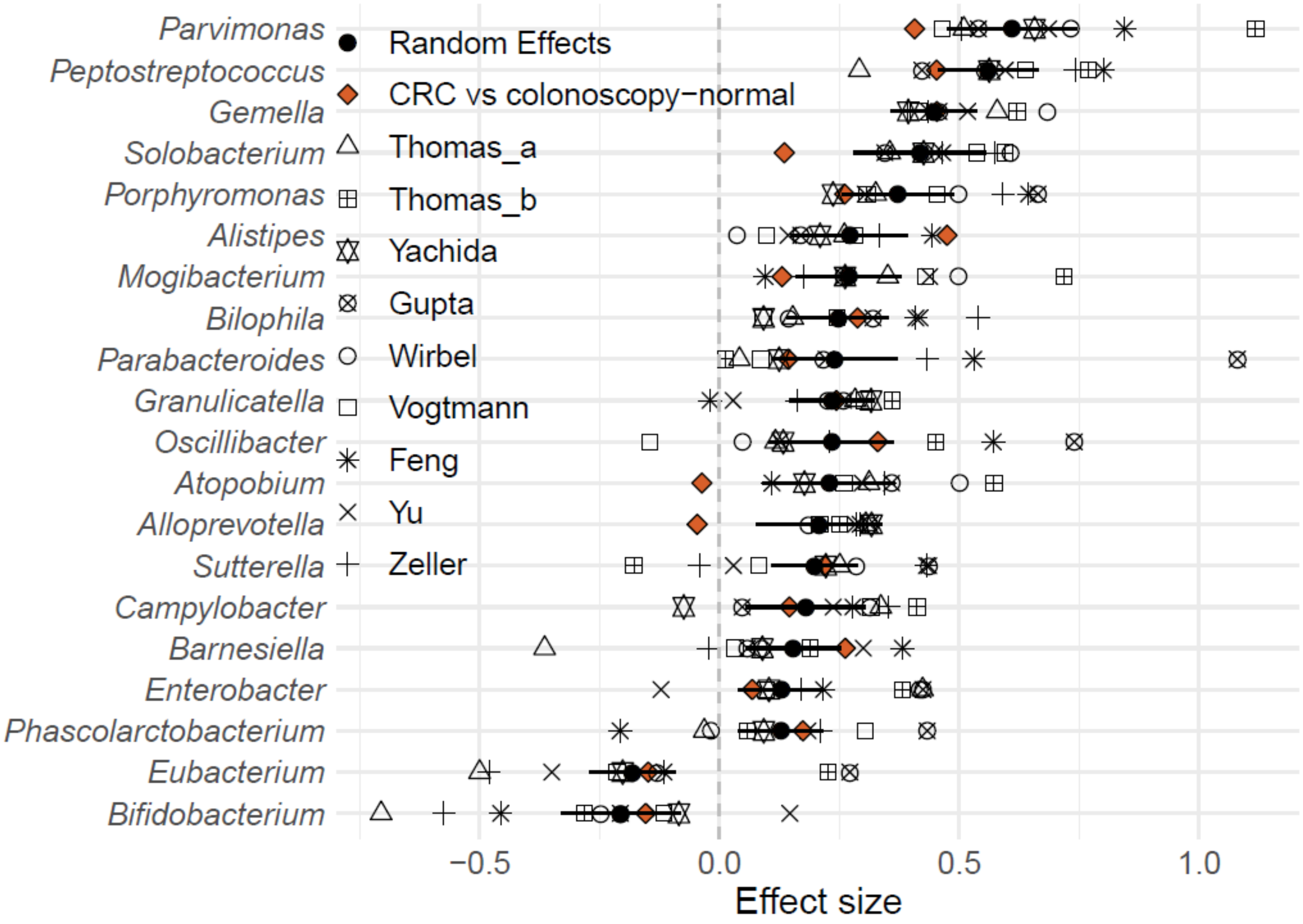
Supplementary Figure 5F. Comparison of RF feature ranking across datasets. The bacteria "CRC vs colonoscopy-normal" total RF model compared to external faecal metagenomic datasets. Importance of each genus for cross-validation performance using gini values. Only genera in the top five in at least one dataset are displayed.


All control
All CRC
Zeller control
Zeller CRC
Yu control
Yu CRC
Yachida control
Yachida CRC
Wirbel control
Wirbel CRC
Vogtmann control
Vogtmann CRC
Thomas_b control
Thomas_b CRC
Thomas_a control
Thomas_a CRC
Gupta control
Gupta CRC
Feng control
Feng CRC
CRC vs colonoscopy−normal control
CRC vs colonoscopy−normal CRC
−6 −4 −2
Parvimonas
log10 proportion
−6 −5 −4 −3 −2
Peptostreptococcus
log10 proportion
−6 −4 −2 0
Alistipes
log10 proportion
−6 −4 −2
Fusobacterium
log10 proportion
−6 −4 −2 0
Streptococcus
log10 proportion
−6 −4 −2 0
Escherichia
log10 proportion
−6 −4 −2
Lachnoclostridium
log10 proportion
−6 −5 −4 −3 −2 −1
Odoribacter
log10 proportion
−6 −4 −2
Parabacteroides
log10 proportion


All control
All CRC
Zeller control
Zeller CRC
Yu control
Yu CRC
Yachida control
Yachida CRC
Wirbel control
Wirbel CRC
Vogtmann control
Vogtmann CRC
Thomas_b control
Thomas_b CRC
Thomas_a control
Thomas_a CRC
Gupta control
Gupta CRC
Feng control
Feng CRC
CRC vs colonoscopy−normal control
CRC vs colonoscopy−normal CRC
−5
−4
−3
−2
−1
0
Blautia
log10 proportion
−6
−4
−2
Flavonifractor
log10 proportion
−6
−4
−2
0
Prevotella
log10 proportion
−5
−4
−3
−2
−1
Clostridium
log10 proportion
−6
−4
−2
Oscillibacter
log10 proportion
−6
−5
−4
−3
−2
Butyricimonas
log10 proportion
−6
−4
−2
Actinomyces
log10 proportion
−6
−5
−4
−3
−2
Solobacterium
log10 proportion



Supplementary Figure 5G. Distributions of relative abundance of genera of greatest importance. The boxplots labelled ‘All’ are a summary of all of the studies.



Supplementary Figure 5H. Genera prioritised by gFOBT amplicon and shotgun metagenomic-based regression models.

Supplementary Methods

Routine NHSBCSP processing of gFOBT

A gFOBT (Hema Screen, Immunostics, Inc) is posted to adults aged 60-74 every two years. gFOBT preparation is performed by participants at home. Participants apply two subsamples from a stool to two of the six squares of the gFOBT. They repeat this with squares 3-4 using a second stool and squares 5-6 using a third stool. Participants store gFOBT at room temperature at home until preparation is complete. Participants then post the gFOBT back to the NHSBCSP Hub, whereupon a strip of card is removed from the reverse of the gFOBT and developer solution (Hema Screen, Immunostics, Inc) (containing hydrogen peroxide and ethanol) is applied. If haemoglobin is present, blue discolouration occurs. If five or six squares turn blue, the result is deemed 'blood-positive' and colonoscopy is offered. If one to four squares turn blue, the result is deemed 'unclear' and up to two further gFOBT are dispatched. If no colour change occurs, the result is deemed 'blood-negative' and screening is complete. 2% of participants are offered colonoscopy.[1] CRC is detected at 10% of colonoscopies, adenoma at 40% and 50% reveal a normal bowel or non-neoplastic condition.[1]

References

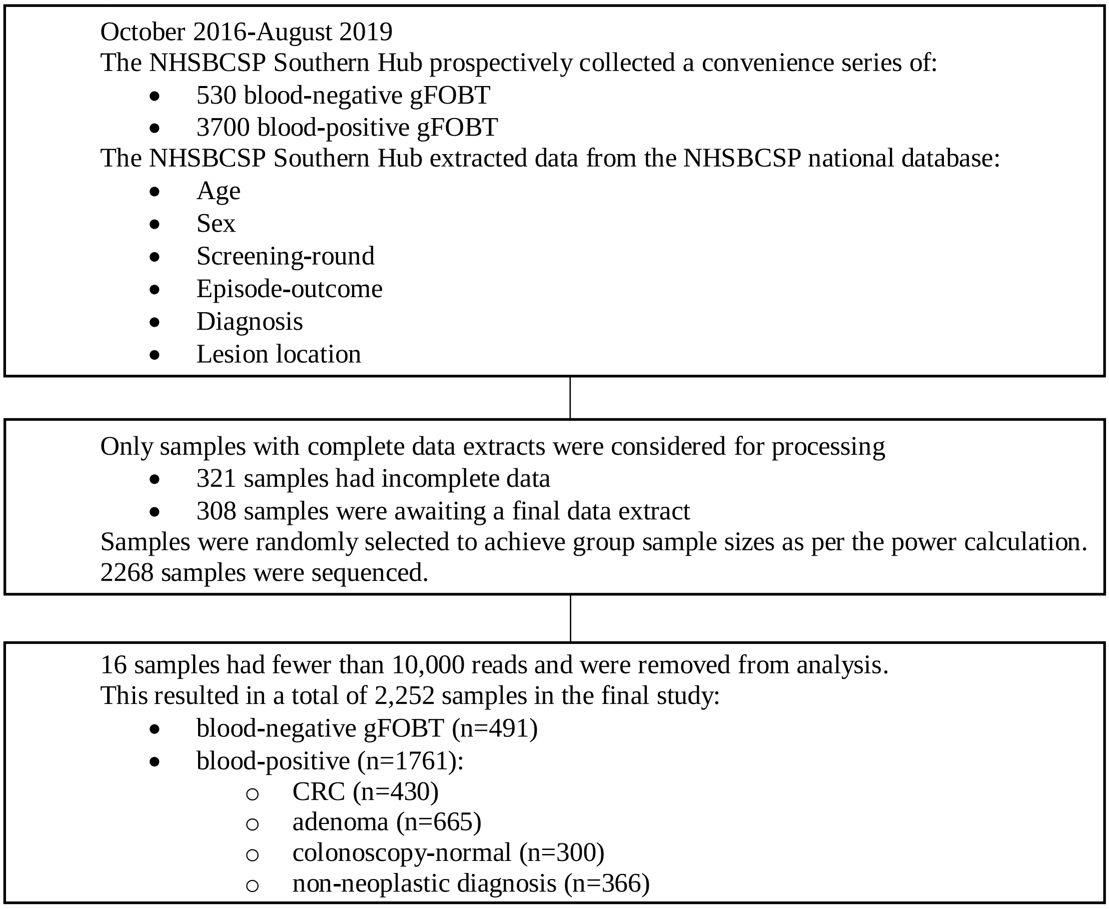
1. Bowel cancer screening: the facts (FOB test kit). https://www.gov.uk/government/publications/bowel-cancer-screening-benefits-and-risks (accessed 24.9.19.

Modified version of the QIAamp DNA Mini Kit protocol

From each developed gFOBT, three alternate squares of faecally-loaded card were dissected and processed as a combined sample. 800µl of Buffer ASL was added. Samples were incubated at 23°C on a Thermomixer Comfort (Eppendorf UK) at 850rpm for one hour. Samples were centrifuged. Supernatant was transferred to pathogen lysis tubes (S) (Qiagen, Germany). Samples were agitated (Vibrax VXR, IKA, UK) at a motor setting of 1800-2200 for ten minutes. Samples were incubated at 95°C on the Thermomixer at 850rpm for 15 minutes. Samples were centrifuged at 18625g for one minute. Supernatant was transferred to a tube containing 173µl of 10M ammonium acetate. Samples were vortexed and placed on ice for five minutes. Samples were centrifuged at 18625g for five minutes. Supernatant was transferred to a tube containing 725µl of propan-2-ol, vortexed and placed on ice for 30 minutes. Samples were centrifuged at 18625g for ten minutes, supernatant was discarded and 1ml of 70% ethanol was added. Samples were centrifuged at 18625g for five minutes, supernatant discarded and 500µl 70% ethanol was added. Samples were centrifuged at 18625g for three minutes, supernatant discarded and samples left for ten minutes to evaporate residual ethanol. 200µl tris-EDTA was added. After ten minutes, samples were vortexed and added to tubes containing 200µl of Buffer AL (QIAamp DNA Mini Kit). 15µl of Proteinase K (QIAamp DNA Mini Kit) was added, samples were vortexed and incubated at 70°C on the Thermomixer at 650rpm for ten minutes. The QIAamp DNA Mini Kit protocol was then followed. To elute DNA, 100µl of UV-irradiated molecular biology grade water was added to samples for five minutes before centrifuging at 18625g for one minute.

Additional sample information

Flow diagram of samples



October 2016-August 2019
The NHSBCSP Southern Hub prospectively collected a convenience series of:

- 530 blood-negative gFOBT
- 3700 blood-positive gFOBT

The NHSBCSP Southern Hub extracted data from the NHSBCSP national database:

- Age
- Sex
- Screening-round
- Episode-outcome
- Diagnosis
- Lesion location

Only samples with complete data extracts were considered for processing

- 321 samples had incomplete data
- 308 samples were awaiting a final data extract

Samples were randomly selected to achieve group sample sizes as per the power calculation.
2268 samples were sequenced.

16 samples had fewer than 10,000 reads and were removed from analysis.
This resulted in a total of 2,252 samples in the final study:

- blood-negative gFOBT (n=491)
- blood-positive (n=1761):
    - CRC (n=430)
    - adenoma (n=665)
    - colonoscopy-normal (n=300)
    - non-neoplastic diagnosis (n=366)

Table 1. Time between faecal collection and DNA extraction by group.

	Time until DNA extraction (days)		
Group	**Minimum**	**Maximum**	**Median**
Blood-negative	46	558	119
CRC	57	670	389
Adenoma	57	686	399
Colonoscopy-normal	55	564	362
Non-neoplastic condition	61	706	530

Table 2. Table of non-neoplastic sample diagnoses. Of the non-neoplastic samples, lesion data was available for 333 of the 366 samples. Many samples had more than one diagnosis recorded; the commonest diagnosis was ‘diverticulosis’.

Diagnosis	Number
Diverticulosis	203
Non-dysplastic polyp	96
Haemorrhoids	90
Inflammatory bowel disease	41
Angiodysplasia	17
Radiation proctitis	6
Diverticulitis	4
Benign submucosal lesion	2
Stricture	2
Melanosis	1
Mucosal Prolapse	1