

1 **Species-level verification of *Phascolarctobacterium* association to colorectal cancer**

2

3 Short Title:

4 ***Phascolarctobacterium* and colorectal cancer**

5

6 Cecilie Bucher-Johannessen^{1,2,3}, Thulasika Senthakumaran⁴, Ekaterina Avershina^{2,3}, Einar Birkeland⁵,
7 Geir Hoff^{6,7}, Vahid Bemanian⁸, Hege Tunsjø⁴, Trine B Rounge^{1,2,3}

8

9 **Affiliations:**

10 ¹ Department of Research, Cancer Registry of Norway - Norwegian Institute of Public Health, Norway

11 ² Department of Tumor Biology, Oslo University Hospital, Norway

12 ³ Center for Bioinformatics, Department of Pharmacy, University of Oslo

13 ⁴ Department of Life Sciences and Health, Oslo Metropolitan University, Oslo, Norway

14 ⁵ Center for Bioinformatics, Department of Informatics, University of Oslo

15 ⁶ Section for Colorectal Cancer Screening, Cancer Registry of Norway - Norwegian Institute of Public
16 Health, Norway

17 ⁷ Telemark Hospital, Norway

18 ⁸ Department of Pathology, Akershus University Hospital, Lørenskog, Norway

19

20 **Abbreviations:** ASV = Amplicon sequence variant

21 BCSN = Bowel Cancer Screening in Norway

22 CAZy = Carbohydrate Active Enzymes

23 CRC = Colorectal cancer

24 FIT = Fecal immunochemical test

25 KEGG = Kyoto Encyclopedia of Genes and Genomes

26 MAG = Metagenome Assembled Genomes

27 NORCCAP = Norwegian Colorectal Cancer Prevention

28 PERMANOVA = Permutational multivariate analysis of variance

29 SCFA = short chain fatty-acids

30 TCA = Tricarboxylic Acid

31 vOTU = virus Operational Taxonomic Unit

32

33 **Correspondence:** Trine B Rounge, University of Oslo, Oslo, Norway, t.b.rounge@farmasi.uio.no

34 Abstract

35 **Background and aims:** We have previously demonstrated an association between increased
36 abundance of *Phascolarctobacterium* and colorectal cancer (CRC) and adenomas in two
37 independent Norwegian cohorts. Here we seek to verify our previous findings using new cohorts
38 and methods. In addition, we characterize lifestyle and sex-specificity, the functional potential of
39 the *Phascolarctobacterium* species and their interaction with other microbial species.

40
41 **Methods:** We analyze *Phascolarctobacterium* with 16S rRNA sequencing, shotgun metagenome
42 sequencing and species-specific qPCR, using 2350 samples from three Norwegian cohorts -
43 CRCAhus, NORCCAP and CRCbiome - and a large publicly available dataset,
44 Curatedmetagenomedata. Using metagenome assembled genomes from the CRCbiome study we
45 explore genomic characteristics and functional potential of the *Phascolarctobacterium*
46 pangenome.

47
48 **Results:**
49 Three species of *Phascolarctobacterium* associated with adenoma/CRC were consistently
50 detected by qPCR and sequencing. Positive associations with adenomas/CRC were verified for
51 *P. succinatutens* and negative associations were shown for *P. faecium* and adenoma in
52 Curatedmetagenomedata. Men show higher prevalence of *P. succinatutens* across cohorts. Co-
53 occurrence among *Phascolarctobacterium* species was low (<6%). Each of the three species
54 show distinct microbial composition and form distinct correlation networks with other bacterial
55 taxa, although *Dialister invisus* was negatively correlated to all investigated
56 *Phascolarctobacterium* species. Pangenome analyses showed *P. succinatutens* to be enriched for
57 genes related to porphyrin metabolism and degradation of complex carbohydrates, whereas
58 glycoside hydrolase enzyme 3 was specific to *P. faecium*.

59
60 **Conclusion:**
61 We have verified that *P. succinatutens* is increased in adenoma and CRC and this species should
62 therefore be recognised among the most important CRC-associated bacteria.

63 **Keywords:** *Phascolarctobacterium*, colorectal cancer, microbiome

64

65 Introduction

66 Many studies have revealed associations between the microbiome and several intestinal diseases.
67 Among others, imbalance in microbial composition and enrichment of specific intestinal bacteria
68 have been associated to adenomas formation and their subsequent progression to CRC via the
69 adenoma-carcinoma pathway¹. The time span for the progression can vary from 5-10 years
70 depending on the specific pathway of tumorigenesis². However, less than 10% of the adenomas
71 are estimated to progress to cancer^{3,4}.

72

73 We have previously shown an increased abundance of Amplicon Sequence Variants (ASV)
74 belonging to the genus *Phascolarctobacterium* in CRC and adenoma cases when compared to
75 healthy controls in stool samples and tissue samples in two independent Norwegian cohorts^{5,6}.

76

77 Three species of *Phascolarctobacterium*; *Phascolarctobacterium succinatutens*,
78 *Phascolarctobacterium faecium*, and *Phascolarctobacterium wakonense* have been described
79 previously but they remain largely uncharacterized. While *P. wakonense* has been isolated from
80 common marmoset feces⁷, *P. succinatutens* and *P. faecium* are abundant in the human
81 gastrointestinal (GI) tract^{8,9}. *P. succinatutens* is estimated to be present in around 20% of human
82 fecal samples while prevalence of *P. faecium* varies between 40-90%, being strongly influenced
83 by host age⁸. The genus is Gram-negative, obligate anaerobic bacteria belonging to
84 *Negativicutes* class in the phylum *Firmicutes*. Both *P. faecium* and *P. succinatutens* use succinate
85 as energy source and can convert succinate into propionate^{9,10}. However, they lack the fumarate
86 reductase gene, an enzyme essential for the conversion of fumarate into succinate¹¹, thus they
87 only rely on the presence of succinate from the environment. Succinate, a tricarboxylic acid
88 (TCA) cycle intermediate in humans, is not abundant in the human diet but it is produced in the
89 GI tract by the host and bacteria such as those belonging to *Paraprevotella*⁹ and *Bacteroides*¹⁰.

90

91 Studies have reported an association between *Phascolarctobacterium* and adenoma/CRC.
92 Yachida et al.¹² observed an enrichment of *P. succinatutens* in early CRC stages, accompanied

93 by elevated succinate levels. Also, Zackular et al.¹³ and Peters et al.¹⁴ have reported higher
94 abundance of *Phascolarctobacterium* in fecal samples from adenoma/CRC cases. In contrast, a
95 small study by Sarhadi et al.¹⁵ found a reduced abundance of *Phascolarctobacterium* in fecal
96 samples from CRC compared to controls. While these studies showed an association between
97 adenoma/CRC and *Phascolarctobacterium* along with several other bacteria, none of them
98 conducted in-depth analyses on the species level.

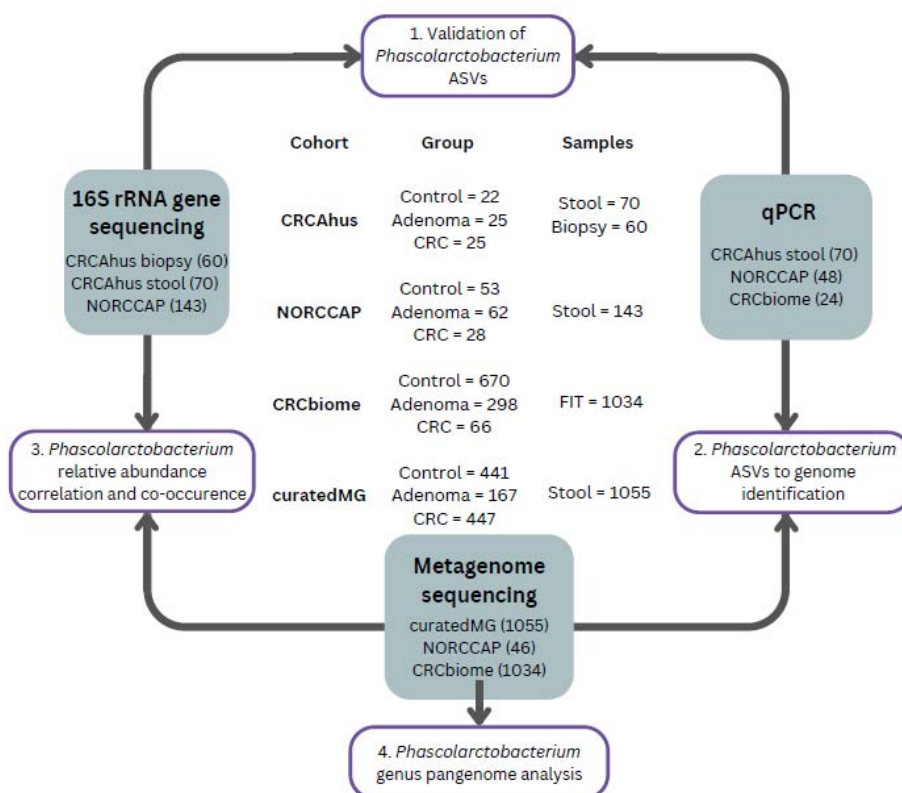
99
100 We aimed to verify the association between *Phascolarctobacterium* and adenoma/CRC at the
101 species level using independent cohorts and techniques, and to compare the genomic makeup of
102 *Phascolarctobacterium* across species.

103

104 Material and Methods

105 Study population and sample collection

106 Data from the CRC study from Akershus University Hospital (CRCAhus hereafter), Norwegian
107 Colorectal Cancer Prevention (NORCCAP) trial and CRCbiome study, and a publicly available
108 dataset, CuratedMetagenomicData, were included in this study (Figure 1).



109

110 Figure 1: Cohorts, processing workflow and data analyses included in this study.

111 ASV = Amplicon Sequence Variant

112 FIT = Fecal Immunochemical Test

113

114 The CRCAhus study (for details, see Senthakumaran et al.⁶) includes seventy-two participants
 115 (age 30-87) who underwent colonoscopy at Akershus University Hospital between 2014 and
 116 2017. Individuals included in the study were either referred for colonoscopy following the
 117 detection of polyps by CT or undergoing investigation for CRC due to unexplained bleeding or
 118 altered stool patterns for more than four weeks. Based on colonoscopy findings, the participants
 119 were classified into three categories: patients with cancer, patients with adenomatous polyps
 120 (diameter ≥ 10 mm), and healthy controls (no pathological findings). Either 2 or 4 biopsy
 121 samples from different locations in the colon were collected during colonoscopy for controls and
 122 cases, respectively. Each participant collected a stool sample in the RNALater RNA stabilizing
 123 buffer (Qiagen, Hilden, Germany) before colonoscopy or one week after colonoscopy. In total,

124 the study population included 72 participants. Of these, 70 (CRC = 23, adenoma = 25, controls =
125 22) provided stool samples, and 60 biopsy samples were included (one from each participant;
126 CRC = 23, adenoma = 20, controls = 17). Detailed information on the participants and sample
127 collection was described elsewhere ¹⁶.

128
129 The NORCCAP trial (for details, see Holme et al.^{17,18} and Bretthauer et al.¹⁹), took place in 1999-
130 2001 and recruited participants (age 50-65) from the Norwegian counties of Oslo and Telemark.
131 Participants collected fecal samples at home in 20 ml vials and immediately stored them in their
132 freezers for up to 7 days, until transportation to the screening center during their sigmoidoscopy
133 screening appointments, and further storage at -20°C. Twenty-eight participants were diagnosed
134 with CRC at screening or diagnosed up to 17 years after screening (identified through cancer
135 registry linkage in 2015). Sixty-three participants had high risk adenomas classified at time of
136 sigmoidoscopy screening. Finally, 53 participants were included as healthy controls (no adenoma
137 or cancer diagnosis at screening or cancer during follow up). Participants with high-risk
138 adenomas were defined as having one or more adenomas ≥ 10 mm, with high-grade dysplasia or
139 villous components regardless of size; or having three or more adenomas regardless of their size,
140 dysplasia, and villosity.

141
142 The CRCbiome study (for details see Kværner et al.²⁰) recruited participants from the Bowel
143 Cancer Screening in Norway (BCSN) trial²¹ between 2017 and 2021. Participants in the BCSN
144 trial were invited for once only sigmoidoscopy or biennial fecal immunochemical test (FIT).
145 CRCbiome recruited participants (age 50 - 74) from the FIT arm, inviting those with a positive
146 FIT test (>15 μg hemoglobin/g feces) who were referred for colonoscopy. Based on diagnoses
147 retrieved from the BCSN database, participants were divided into three groups including 66 CRC
148 cases, 298 advanced adenomas (including advanced adenomas, and advanced serrated lesions)
149 and 670 controls (including no findings and those with non-advanced adenomas <3 mm). The
150 CRCbiome study aims to explore the influence of diet and lifestyle on the microbiome.
151 Participants completed two questionnaires prior to colonoscopy: A Food Frequency
152 Questionnaire (FFQ), encompassing 256 food items across 23 questions about consumption
153 frequency, portion sizes, and BMI; and a Lifestyle and Demographic Questionnaire (LDQ) with

154 10 items^{20,22}. From this a healthy lifestyle index (HLI) was developed as described in Kværner et
155 al 2023²³.

156
157 For external validation, we utilized the publicly available R package [dataset]
158 CuratedMetagenomicData²⁴ (accessed 22.03.2022, curatedMG hereafter), a comprehensive
159 dataset of 22,588 samples obtained from 93 independent datasets. Samples are collected from
160 various body sites, and raw data processed to generate relative abundance tables using
161 MetaPhlan3. We filtered the data to only include samples from stool and conditions including
162 CRC, adenomas, and healthy controls. This resulted in a subset of 1055 samples from seven
163 different studies where 447 were from CRC, 147 from adenomas, and 441 from healthy controls.
164 The largest study included in this dataset was from Yachida et al.¹² with 576 samples.

165 Ethical considerations

166 The CRCAhus study, BCSN trial, the CRCbiome study and the NORCCAP trial have been
167 approved by the Regional Committee for Medical and Health-related Research Ethics in
168 Southeast Norway (REK ref: 2012/1944, 2011/1272, 63148, and 22337 respectively). The
169 CRCAhus study also received approval from the data protection manager at Akershus University
170 Hospital. The BCSN trial is registered at clinicaltrials.gov (Clinical Trial (NCT) no.: 01538550).
171 All cohorts followed and performed experiments in compliance with the Declaration of Helsinki
172 Principles. For all cohorts, participants gave written informed consent prior to inclusion. All lab
173 procedures were conducted in accordance with relevant guidelines and regulations.

175 DNA extraction and sequencing

176 DNA from biopsies and fecal samples from the CRCAhus were extracted using AllPrep
177 DNA/RNA Mini Kit (Qiagen, Hilden, Germany) and PSP Spin Stool DNA Kit (Stratec
178 Molecular GmbH, Berlin, Germany), respectively as described by¹⁶. Amplicon sequencing of 16S
179 rRNA V4 region was performed on the Illumina MiSeq platform (Illumina Inc., San Diego, CA,
180 USA) using the MiSeq reagent kit v2 as previously described⁶. PCR amplification of 16S rRNA
181 V4 region was performed using 16S forward primer (16Sf V4: GTGCCAGCMGCCGCGGTAA)
182 and 16S reverse primers (16Sr V4: GGACTACHVGGGTWTCTAAT)²⁵.

183
184 DNA extraction of the NORCCAP samples was performed using the QIASymphony automated
185 extraction system and a QIASymphony DSP Virus/Pathogen Midi Kit (Qiagen, Hilden,
186 Germany), with an off-board lysis protocol that included modifications. The process involved
187 bead beating of the samples, followed by a mixture with a lysis buffer, and subsequent incubation
188 for lysis. Amplification of 143 samples was carried out using a TruSeq (TS)-tailed 1-step
189 amplification protocol²⁶. For 16S rRNA sequencing, the V3-V4 region was targeted using the
190 primers S-D-Bact-0341-b-S-17 (5'CCTACGGGNGGCWGCAG'3) and SD-Bact-0785-a-A-21
191 (5'GACTACHVGGGTATCTAATCC'3)²⁷. Sequencing was performed using the Illumina MiSeq
192 instrument generating paired-end reads of 2x300 bp. A subset of the samples, 46, was also
193 metagenome sequenced using the Riptide protocol (Twist Bioscience, CA, USA) and sequenced
194 on an Illumina NovaSeq platform, generating paired-end reads of 2x130 bp. We have previously
195 shown the feasibility of using these long-term stored samples for microbiome analyses²⁸.

196
197 For CRCbiome, DNA extraction followed a similar protocol as NORCCAP, but with the
198 inclusion of an extra washing step during lysis. The sequencing libraries for 1034 CRCbiome
199 samples were prepared in line with the Nextera DNA Flex Library Prep Reference Guide with
200 the modification of reducing the reaction volumes to a quarter of the recommended amounts.
201 Sequencing was performed on the Illumina Novaseq system generating 2x151 bp paired-end
202 reads (Illumina, Inc., CA, USA).

203
204 **Bioinformatics and taxonomic profiling**
205 16S rRNA sequencing data from CRCaHus and NORCCAP were processed using Quantitative
206 Insights Into Microbial Ecology (QIIME2²⁹, version 2021.2.0 and 2020.2.0, respectively) with
207 the DADA2-plugin as described previously^{5,6}, resulting in ASV. For comparative analysis with
208 metagenomic data, ASV counts were transformed to relative abundances using the
209 transform_sample_counts function from the Phyloseq package³⁰ (v1.26.1) where each ASV
210 count was divided by the total count of ASVs in the sample.

211

212 NORCCAP metagenomic reads were processed using Trimmomatic³¹ (v0.66.0) for quality
213 trimming, discarding sequences below a quality threshold of 30 across four bases and those
214 shorter than 30 base pairs. Bowtie2³² (v2.4.2) and Samtools³³ (v1.12) was used for removal of
215 human-derived sequences. Taxonomic profiling was conducted using MetaPhlan3³⁴ (v3.0.4)
216 with default settings.

217
218 For the CRCbiome samples sequencing reads were processed using two different approaches.
219 First, raw reads were trimmed using Trimmomatic (v0.36) and reads mapping to the human
220 genome (hg38) and PhiX were removed using Bowtie2 (v2.3.5.1). Read-based taxonomy was
221 determined at the species level and quantified as relative abundance determined by MetaPhlan3
222 using the mpa_v30_Chocophlan_201901 (v3.0.7) database. Second, metagenome assembled
223 genomes (MAGs) were created using the framework Metagenome-ATLAS³⁵ (v2.4.3). Low-
224 quality reads were filtered and human and phiX sequences removed using BBTools³⁶. Reads
225 were then assembled via MetaSpades³⁷ (v3.13) and grouped into genomes with DASTool³⁸ (v1.1),
226 utilizing MetaBat³⁹ (v2.2) and MaxBin⁴⁰ (v2.14) for genomic bin identification. Genome
227 dereplication was conducted using dRep⁴¹ (v2.2) based on 95% identity over 60% genome
228 overlap. Genomes with completeness >90% and contamination < 10%, determined using
229 CheckM⁴² were kept. GTDB-Tk (v1.3) assigned a taxonomy against the GTDB database⁴³ (v95).
230 Metagenome-assembled genome abundance was estimated by median read depth across 1000-bp
231 bins of each genome and scaled by reads per million. The taxonomic classification approach was
232 employed for those analyses where consistency and comparability across datasets was necessary.
233 The MAGs were used for analyses including only CRCbiome samples which encompassed diet
234 analyses, genomic characterization and functional potential of the individual
235 *Phascolarctobacterium* species.

236
237 Species-specific quantification of *Phascolarctobacterium* by qPCR
238 To verify our previous findings between *Phascolarctobacterium* ASVs and adenoma/CRC^{5,6}, we
239 developed species-specific qPCR assays. BLAST (Basic Local Alignment Search Tool) search
240 identified the ASVs as *Phascolarctobacterium succinatutens* and *Phascolarctobacterium* sp. 377.
241 As *P. faecium* is also prevalent in the human GI tract we decided to include this species as well,

242 and genomes from *P. succinatutens* (YIT 12067), *P. faecium* (JCN 30894) and *P. sp 377*
243 (AB739694.1) were used for qPCR assay development. IDT PrimerQuest Tool (Integrated DNA
244 Technologies, Leuven, Belgium) was used for primer and probe design. The primers and the
245 probes were synthesized by TiB Molbiol (Berlin, Germany) and are listed in Table 1. The
246 analytical specificity of the *Phascolarctobacterium* qPCR assays were tested using 50 different
247 bacterial strains, obtained mostly from Culture Collection University of Gothenberg (CCUG) and
248 clinical isolates from Akershus University Hospital (Table S1). Limit of detection (LOD) was
249 determined using 10-fold serial dilution of DNA from pure bacterial suspensions. qPCR assays
250 were performed using Brilliant III Ultra-fast QPCR master mix (Integrated DNA Technologies,
251 USA) with 2 µl DNA in 20 µl reaction volume. Amplification was conducted on QuantStudio5
252 Real-Time PCR systems (Thermo Fisher Scientific, Waltham, MA, USA). Cycling conditions for
253 the *Phascolarctobacterium* assays were as follows: an initial denaturation of 95 °C for 5 min,
254 followed by 40 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 30 s.

255
256 In the CRCAhus cohort, all 70 fecal samples were subjected to species-specific qPCR analysis.
257 For the NORCCAP cohort, 38 samples with reads mapping to the genus level of
258 *Phascolarctobacterium* in both the 16S rRNA and metagenome datasets were subjected to
259 species-specific qPCR, along with 10 samples with no reads mapping to *Phascolarctobacterium*.
260 Additionally, 24 samples from the CRCbiome cohort were selected to verify detection of
261 *Phascolarctobacterium* and identify their genomes. The CRCbiome cohort included 12 samples
262 with reads from *P. succinatutens*, four samples from *P. faecium*, four samples from *P. sp 377*,
263 and four samples without *Phascolarctobacterium* reads. Total bacterial DNA load in each sample
264 was estimated using the universal 16S rRNA as target. The primer and probe sequences and the
265 cycling conditions for the universal 16S rRNA gene amplification has been described
266 elsewhere⁴⁴. qPCR data was analyzed with the ΔC_t method ($\Delta C_t = C_{tTarget} - C_{tTotal\ DNA}$) using the
267 16S rRNA gene as a reference. Relative abundance was calculated by $2^{-\Delta C_t}$.

268

269 Table 1: List of primers and TaqMan probes used in this study

Target	Primer and probe Sequence 5' - 3'	Amplicon size	Reference
--------	-----------------------------------	---------------	-----------

<i>P. succinatutens</i>	16S rRNA	Fwd: GGGACAACATCCCGAAAGG Rev: GCCATCTTTCACAGCATCCT Probe: ACCGAATGTGACAGCAATCTCGCA	73	This study
<i>P. faecium</i>	16S rRNA	Fwd: CCATCCTTTAGCGATAGCTTACT Rev: ACATTCCGAAAGGAGTGCTAATA Probe: AGGCCATCTTTCATCCTGCCA	98	This study
<i>Phascolarctobacterium</i> sp. 377	16S rRNA	Fwd: GTAGGCAACCTGCCCTTAG Rev: CCATCCTTTAGCGATAGCTTACAT Probe: ATGTGACGCTCCTATCGCATGAGG	127	This study
Total bacterial DNA	16S rRNA	Fwd: AATAAATCATAAACTCCTACGGGAGGCAGCAGT Rev: AATAAATCATAACCTAGCTATTACCGCGGCTGCT Probe: CGGCTAACTMCGTGCCAG	204	Brukner et al ⁴⁴

270
271 Genome analyses
272 *16S rRNA gene*
273 To compare ASVs across studies, we made a phylogenetic tree based on the V4 region from
274 NORCCAP and reference genomes for the three *Phascolarctobacterium* species identified in
275 CRCAhus (*P. succinatutens*, *P. faecium*, and *P. sp.377*). Initially, we created a BLAST database
276 of the 16S V4 region of the CRCAhus ASVs using the makeblastdb (v2.13.0) command from
277 BLAST+ NCBI toolbox with default settings⁴⁵. Blastn was then employed to extract the
278 corresponding V4 region from the NORCCAP and reference sequences (*P. succinatutens* (GCA
279 017851075.1), *P. sp 377* (AB739694.1) and *P. faecium* (AP025563.1)). The V4 sequences from
280 CRCAhus, NORCCAP and reference genomes were then aligned by Multiple Alignment using
281 Fast Fourier Transform (MAFFT, v7)⁴⁶. A maximum likelihood phylogenetic tree was
282 constructed with IQ-TREE⁴⁷ (v2.2) using F18+F substitution model and bootstrapping set to
283 1000. The resulting tree was visualized using Interactive Tree of Life (iTOL, v6)⁴⁸. ASVs with

284 over 97% similarity to a reference sequence were collapsed into one ASV for all subsequent
285 analyses.

286

287 *Metagenome data*

288 All CRCbiome genomes belonging to the *Phascolarctobacterium* genus were annotated using
289 Dram⁴⁹ (v1.4) with default settings using the databases KOfam⁵⁰ (accessed 31.10.2022),
290 dbCAN⁵¹ (accessed 08.09.2022) and Uniref90⁵² (accessed 14.11.2022). Identified protein-coding
291 gene sequences were then used as input for a pangenome analysis using Roary⁵³ (v3.13), based
292 on identification of gene clusters with 70% identity cutoff for protein similarity. Gene clusters
293 within the species-specific core were defined as those found in 95% of the genomes from one
294 species and in 0% of the other two. Genus-level core was defined as those genes present in $\geq 95\%$
295 of genomes regardless of species. All genomes were aligned using MAFFT⁵⁴ (v7.520) and a tree
296 was constructed using IQ-TREE (v2.2) with GTR+F+R7 substitution model and visualized using
297 ITOL (v6). Pairwise average nucleotide identity (ANI) between the genomes was calculated
298 based on tetranucleotide frequencies using the Python package pyani (v0.2.12).

299

300 *Statistics*

301 Associations between *Phascolarctobacterium* species abundance and participant characteristics
302 were evaluated in separate linear models for each species and variable. Abundance was coded as
303 the dependent variable and participant characteristics as independent variables, adjusting for sex,
304 age and screening center (Telemark or Oslo for NORCCAP, and Moss or Bærum for
305 CRCbiome), or study of origin (total 7 studies for curatedMG). Here, relative abundances were
306 log transformed, with 0 replaced by a pseudo count, defined as half the lowest observed relative
307 abundance of the feature. The participant characteristics evaluated included clinical group (CRC,
308 adenoma, or controls), lifestyle and dietary factors. Diet variables included were energy intake
309 (kcal/day), macronutrients (in energy percentage (E%)), and alcohol and fiber (in g/day) as
310 described in⁵⁵. Lifestyle and demographic variables included were national background,
311 education, occupation, marital status, body mass index, physical activity level, use of antibiotics
312 and antacids in the past three months, smoking and snus habits and the healthy lifestyle index
313 (further details in Kværner et al.²³ and Istvan et al.⁵⁵). The relationship between

314 *Phascolarctobacterium* species relative abundance and FIT values. was assessed using an ordinal
315 logistic regression model adjusted for sex and age, implemented with the function polr from the
316 R package MASS⁵⁶ (v7.3-60). Here, FIT values (in µg hemoglobin/g feces) were categorized into
317 four groups based on their level of hemoglobin (group 1 = 15-20, group 2 = 20-35, group 3 = 35-
318 70, group 4 = >70). Group differences in prevalence of *Phascolarctobacterium* species were
319 evaluated using a chi-squared test.

320
321 To assess the correlation between the relative abundance of *Phascolarctobacterium* species as
322 estimated using NGS and qPCR, we performed Spearman correlation analysis. Pairwise co-
323 occurrence of *Phascolarctobacterium* species was quantified as a percentage, calculated by
324 dividing the number of sample pairs featuring two species by the total sample count within the
325 dataset and multiplying by 100. To evaluate whether the dominant *Phascolarctobacterium*
326 species were associated with distinct microbial communities, a permutational multivariate
327 analysis of variance (PERMANOVA) test was conducted using the adonis2 function from the
328 vegan package⁵⁷ (v.2.5-7) based on Bray-Curtis distances of relative species abundance. Here,
329 participants were categorized according to *Phascolarctobacterium* presence: those with reads
330 exclusively mapping to *P. succinatutens*, *P. sp 377*, or *P. faecium*; those with reads mapping to
331 two or more species; and those with no *Phascolarctobacterium* reads. The PERMANOVA test
332 was adjusted for sex, age, and screening center. Cor_test from the package rstatix⁵⁸ (v.0.7.0) was
333 used to calculate Spearman's correlation between relative abundance of the three
334 *Phascolarctobacterium* species and all other species or virus OTUs (vOTUs)⁵⁵. Before species-
335 correlation analysis, a 5% prevalence filtration was performed. Correlation networks were
336 visualized using Cytoscape⁵⁹ (v3.9.0).

337
338 Utilizing the results from the pangenome analyses, a chi-squared or Fisher's exact test were used
339 to identify significant deviations in the prevalence of carbohydrate-active enzymes (CAZY) and
340 Kyoto Encyclopedia of Genes and Genomes (KEGG) genes across CRC, adenoma, and control
341 groups. Additionally, we compared the prevalence of CAZY and KEGG genes within each
342 *Phascolarctobacterium* species against the other two species combined. The KEGG genes that

343 had varying distribution across species were used for the pathway overrepresentation analysis
 344 with MicrobiomeProfiler⁶⁰ (v1.4.0).

345
 346 All statistical analyses were performed using the R software (v4.1.0), with main package being
 347 tidyverse⁶¹ (v.1.3.1). Nominal statistical significance was considered for $p < 0.05$. Adjustment for
 348 multiple testing was performed using the Benjamini-Hochberg false discovery rate (FDR)⁶², with
 349 $FDR < 0.05$ being considered statistically significant. Code available on
 350 https://github.com/Rounge-lab/Phascolarctobacterium_CRC.

351

352 Results

353 Participant characteristics

354 In total, data from 2350 participants from three Norwegian CRC-related cohorts and the
 355 international collection of datasets available as curatedMG were analyzed (Table 2). The
 356 distribution of men and women was similar across datasets, with the percentage of women
 357 ranging from 39 to 44%.

358

359 Table 2: Participant characteristics

	CRCAhus	NORCCAP	NORCCAP	CRCbiome	CuratedMG
	** n=72	16S	MG***	n=1034	n=1055
		n=143	n=46		
CRC, n (%)	25 (35)	28 (20)	7 (15)	66 (6)	447 (42)
Adenoma, n (%)	25 (35)	62 (43)	17 (37)	298 (29)	167 (16)
Control, n (%)	22 (30)	53 (37)	22 (48)	670 (65)	441 (42)
Male, n (%)	42 (58)	86 (60)	28 (61)	582 (56)	622 (59)
Female, n (%)	30 (42)	57 (40)	18 (39)	452 (44)	433 (41)
Age, median (range)	67.5 (30-87)	57 (51-65)	58,5 (53-65)	67 (55-77)	64 (21-88)
		<i>16S</i>		<i>Metagenome</i>	

<i>P. succinatutens</i> , n (%)*	13 (16)	37 (26)	11 (24)	156 (15)	315 (30)
<i>P. sp 377</i> , n (%)*	9 (13)	1 (0.7)	1 (2)	88 (8)	68 (6)
<i>P. faecium</i> , n (%)*	22 (31)	17 (12)	4 (9)	335 (32)	320 (30)

360 * from NGS relative abundance

361 **CRCAhus included 72 participants with both stool (70) and biopsy samples

362 ***NORCCAP samples subset with metagenome sequencing

363

364 Phylogenetic comparison of the *Phascolarctobacterium* ASVs and reference
365 genomes

366 We assessed the phylogenetic relationship between *Phascolarctobacterium* ASVs, including two
367 CRC-associated ASVs and *Phascolarctobacterium* reference genomes. The CRC-associated
368 ASVs from NORCCAP⁵ and CRCAhus⁶ studies clustered in proximity to 16S rRNA gene
369 sequences from *P. succinatutens* and *P. sp 377* genomes, respectively (Figure 2A).
370 *Phascolarctobacterium* ASVs from paired biopsy and fecal samples (CRCAhus) clustered
371 exclusively together. The CRCAhus ASVs clustered with *P. succinatutens* (6 ASVs), *P. sp 377*
372 (4 ASVs) and *P. faecium* (3 ASVs) reference genomes. In NORCCAP stool samples, 37, 1 and 2
373 ASVs clustered with *P. succinatutens*, *P. sp 377* and *P. faecium* references, respectively. These
374 results show that the ASVs represent three distinct species of *Phascolarctobacterium*, and that
375 the CRC-associated ASVs represent two independent *Phascolarctobacterium* species.

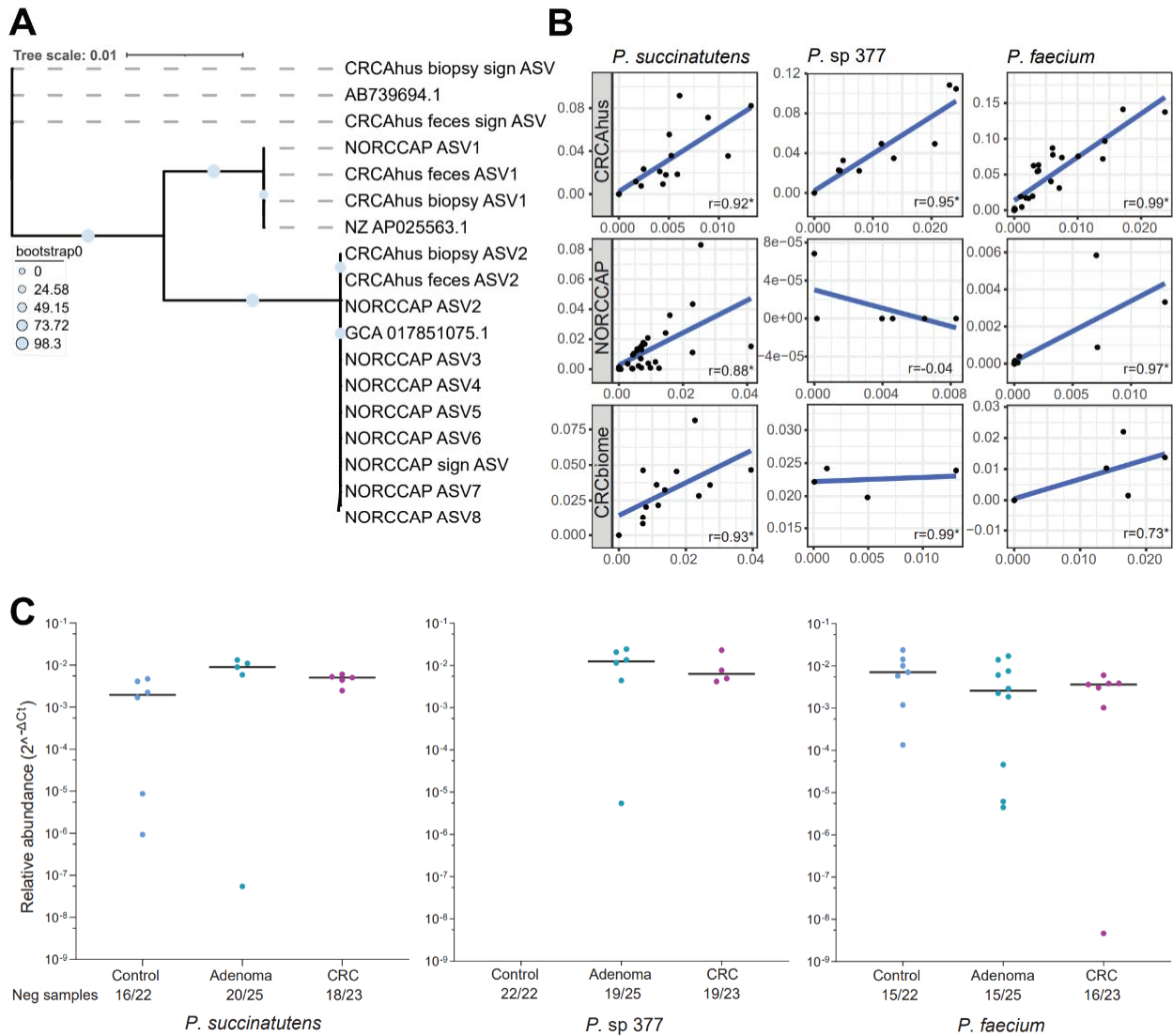
376

377 qPCR confirms phylogenetical distinct ASVs and CRC-association for
378 *Phascolarctobacterium spp.*

379 To validate the phylogenetic discordance between *Phascolarctobacterium* ASVs identified in
380 CRCAhus and NORCCAP, we established qPCR assays for *P. succinatutens*, *P. sp 377* and *P.*
381 *faecium*. Analytical specificity assessed for a panel of 50 bacterial species revealed all three
382 assays to exhibit 100 % specificity for each targeted *Phascolarctobacterium* species (Table S1).
383 LOD for *P. succinatutens* and *P. faecium* assays were 1fg/μl. qPCR assays detection rates for

384 samples with sequencing reads for *P. succinatutens*, *P. sp 377* and *P. faecium* were 96 %, 94 %
385 and 100 %, respectively. The qPCR additionally detected (presence) of 4, 3 and 11 of *P.*
386 *succinatutens*, *P. sp 377* and *P. faecium*, respectively, where 3, 1 and 9 samples (7 from
387 NORCCAP) had low abundance (Ct >32). With regards to metagenome data from NORCCAP
388 and CRCbiome, qPCR analysis also confirmed presence of the three species in these samples
389 (Table S2). qPCR only detected five additional samples with either *Phascolarctobacterium* that
390 did not have sequencing reads in the long-term stored NORCCAP samples. There was high
391 concordance (100 %) between *Phascolarctobacterium* relative abundance detection in
392 CRCbiome FIT samples and qPCR. Overall, this indicates high qPCR sensitivity across sample
393 types and storage conditions.

394
395 Our results showed a high concordance between relative abundance from 16S rRNA gene
396 sequencing, shotgun metagenome sequencing and qPCR. Spearman's correlation coefficients of
397 0.92, 0.95, and 0.97 for *P. succinatutens*, *P. sp 377* and *P. faecium*, respectively (all $p < 0.01$,
398 Figure 2B and Table S2) was observed in CRCAhus. In NORCCAP 16S *P. succinatutens* and *P.*
399 *faecium* showed a significant positive correlation (0.88, 0.73, $p < 0.05$), but *P. sp 377* did not (-
400 0.05, $p = 0.7$). In the NORCCAP MG and CRCbiome cohorts, *P. succinatutens*, *P. sp 377* and *P.*
401 *faecium* showed positive correlations (0.99, 0.97 and 0.73 for NORCCAP MG, all $p < 0.01$, and
402 0.93, 0.99 and 0.84 for CRCbiome, all $p < 0.01$). In accordance with 16S rRNA sequencing-based
403 detection in the CRCAhus study, qPCR results identified *P. sp 377* in 6/25 adenomas and 4/23
404 CRC cases but was absent from the control group (Figure 2C).



405
 406 Figure 2: A) Phylogenetic tree showing that ASVs from CRCAhus and NORCCAP cluster with reference
 407 genomes for *P. succinatutens* (GCA 017851075.1), *Phascolarctobacterium* sp. (AB739694.1), and
 408 *faecium* (AP025563.1). The CRC associated ASV from CRCAhus cluster in proximity to *P. sp 377* and
 409 the CRC associated ASV from NORCCAP cluster in proximity to *P. succinatutens*. B) Scatter plot
 410 illustrating the relationship between relative abundance from NGS data on y-axis and relative abundance
 411 from qPCR on x-axis. Each point represents one sample. Data is presented for *P. succinatutens*, *P. sp 377*
 412 and *P. faecium* per dataset (CRCAhus feces, NORCCAP and CRCbiome). NORCCAP and CRCbiome
 413 samples were selected based on relative abundance data. C) Relative abundance of
 414 *Phascolarctobacterium* spp in fecal samples from CRCAhus. While *P. succinatutens* and *P. faecium* were
 415 present in all three groups, the uncultured *P. sp 377* was not found in the control group. Each point

416 represents one sample. Number of negative samples with 0 abundance is indicated on the x-axis (neg
417 samples).

418 * = p-value < 0.001.

419 r = Spearman's correlation coefficient

420

421 CuratedMG metagenomes confirms association between CRC and *P. succinatutens*

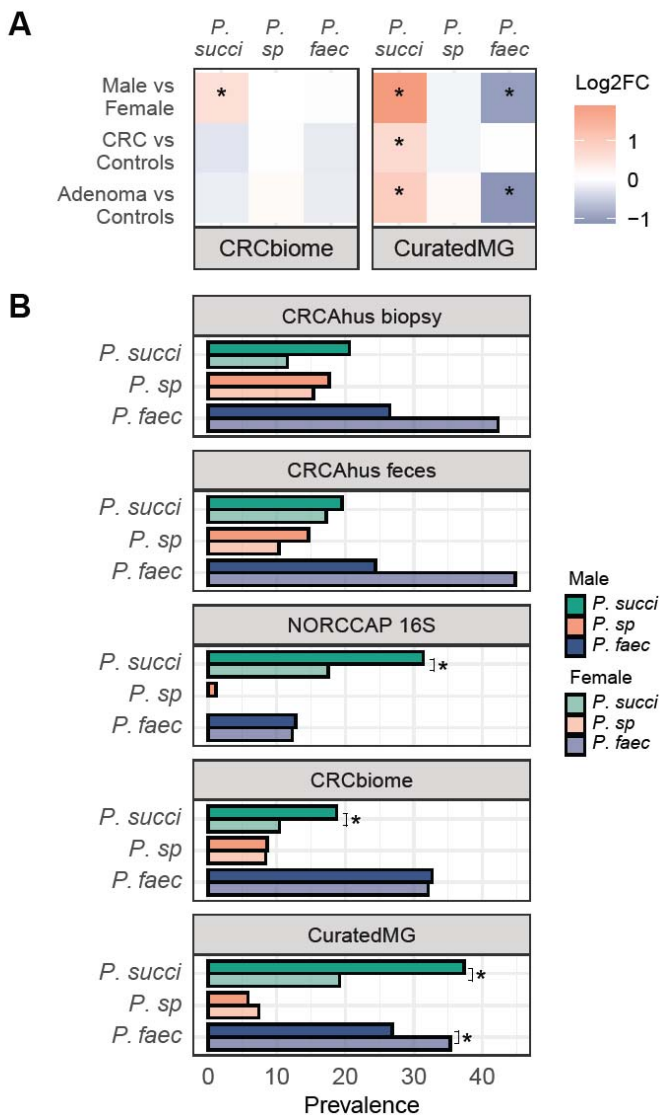
422 We further investigated the association between adenoma/CRC cases and abundance of the three
423 *Phascolarctobacterium* species in two large and independent CRC-related datasets, namely
424 CRCbiome and curatedMG. The results showed a positive association between *P. succinatutens*
425 and adenomas/CRC in curatedMG (Figure 3A, Table S3, all p<0.05). *P. sp 377* was not
426 associated with adenomas or CRC in either dataset. *P. faecium* was negatively associated with
427 adenomas in curatedMG (p=0.014).

428

429 Sex-specificity of *P. faecium* and *P. succinatutens*

430 We also observed an association between *Phascolarctobacterium* species and sex. Men exhibited
431 a higher abundance of *P. succinatutens* in CRCbiome and curatedMG datasets (Figure 3A, Table
432 S3, p<0.05), while women showed a higher abundance of *P. faecium* in curatedMG. Subsequent
433 presence/absence analysis confirmed a higher presence of *P. succinatutens* in men across
434 NORCCAP MG, CRCbiome, and curatedMG, and a greater prevalence of *P. faecium* in women
435 in curatedMG (Figure 3B, Table S4, all p<0.05).

436



437

438

439 Figure 3: A) Summary of multivariate linear models adjusting for sex, age and, for region (CRCbiome)

440 and for study (curatedMG). Colour indicates log2 fold change, with red indicating a higher abundance and

441 blue indicating a lower abundance compared to the reference group. Significantly increased abundance of

442 *P. succinatutens* was observed in adenoma/CRC compared to controls in curatedMG and a lower

443 abundance of *P. faecium* in adenomas versus controls. Significantly higher abundance of *P. succinatutens*

444 in males were observed in both CRCbiome and curatedMG . B) Percentage of samples containing each of

445 the three *Phascolarctobacterium* species, categorized by sex. Men displayed a higher prevalence of *P.*

446 *succinatutens* compared to women in the NORCCAP, CRCbiome and curatedMG datasets. *P. succi* = *P.*

447 *succinatutens*; *P. sp* = *P. sp* 377; *P. faec* = *P. faecium*

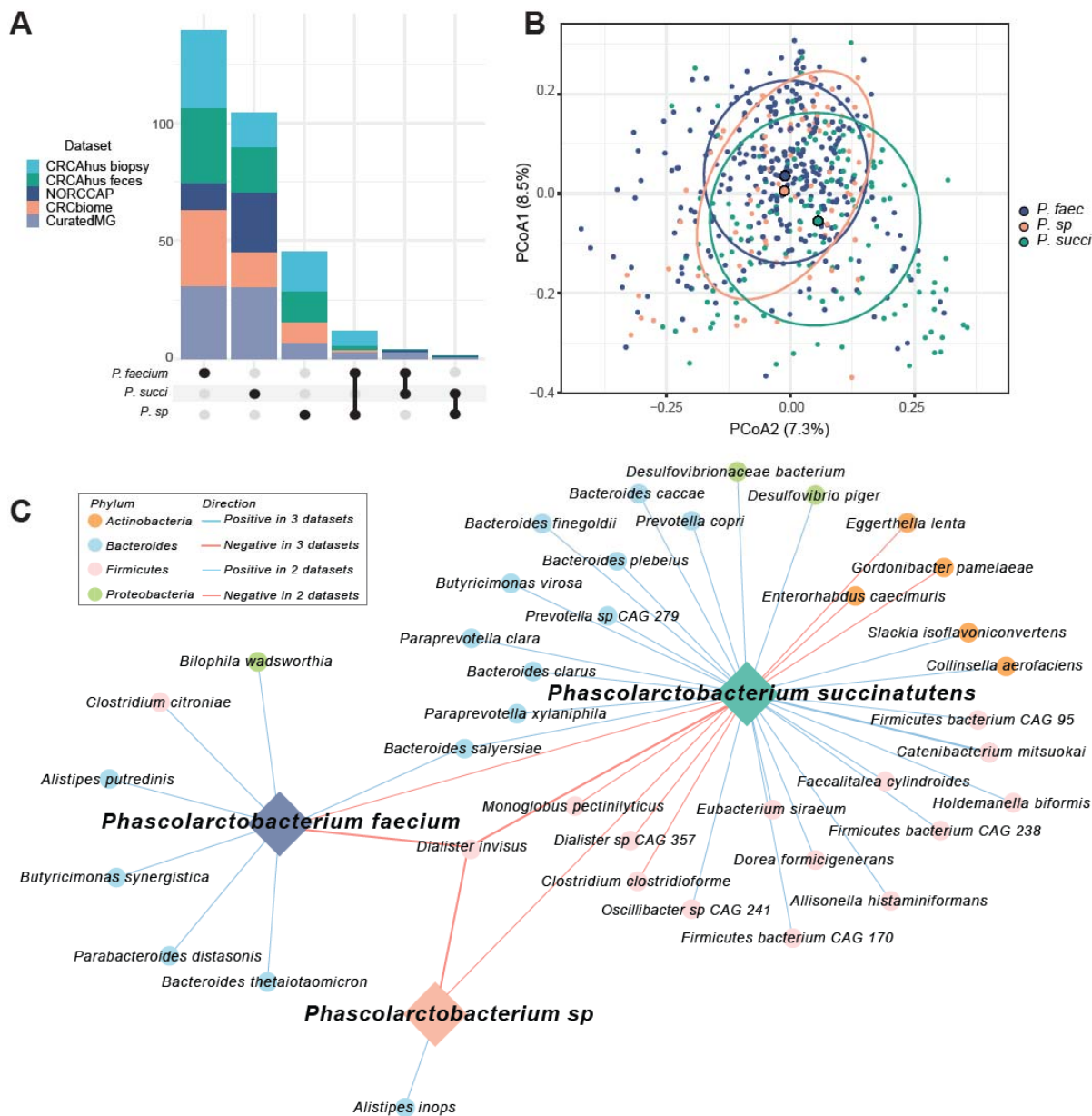
448 * = p-value < 0.05.

449
450 *Phascolarctobacterium* species are mutually exclusive and have distinct microbial
451 partners

452 We further investigated the characteristics of the microbiome, their prevalence in participants'
453 microbiomes, and their interactions with other microbes. We explored the extent of
454 *Phascolarctobacterium* species co-occurrence across samples and study populations. We found a
455 low rate of co-occurrence between the *Phascolarctobacterium* species in all datasets (Figure 4A
456 and Table S5). The highest pairwise co-occurrence was observed in CRC_Ahus biopsy samples
457 between *P. faecium* and *P. sp 377* (6%). For all other datasets, co-occurrence was less than 3%
458 and no samples had all three species across datasets. There was also a significant compositional
459 difference between samples with different dominating *Phascolarctobacterium* species in
460 CRCbiome (p=0.001, $R^2=0.02$, Figure 4B and Figure S1) and curatedMG (p=0.001 and
461 $R^2=0.02$).

462
463 For all datasets we identified 321 species with significant correlation to one or more
464 *Phascolarctobacterium* species where 248 showed a positive correlation. Forty-one species
465 showed consistent correlations across metagenome datasets. *Dialister invisus* exhibited negative
466 correlations with all three *Phascolarctobacterium* species (Figure 4C and Table S6) suggesting
467 that this species could also be mutually exclusive. On the other hand, *Bacteroides salyersiae* was
468 positively correlated to both *P. faecium* and *P. succinatutens*. There were also 5 other
469 *Bacteroides* species that showed a positive correlation to *Phascolarctobacterium* species. We
470 have recently characterized viral diversity in CRCbiome samples⁶³. Here we detected 12 vOTUs
471 with significant associations to one or more *Phascolarctobacterium* species (Figure S2). In
472 contrast to the predominantly positive associations observed between bacteria, 11 out of 12
473 significant associations for viruses were negative, and only one had a positive association with *P.*
474 *faecium*, but not with other *Phascolarctobacterium*.

475



476
477 Figure 4: A) Upset plot illustrating the co-occurrence of *Phascolarctobacterium* species in all five
478 datasets. No samples had all three species present across datasets. B) PCoA plot showing the microbial
479 composition for the CRCbiome samples, where the groups are defined based on presence of one
480 dominating *Phascolarctobacterium* species. PERMANOVA test showed a significant difference between
481 the three groups ($p = 0.001$) with an R^2 of 0.02. C) Correlation network plot of the 41 species with FDR-
482 significant, consistent correlations across at least two of the metagenome datasets (NORCCAP MG,
483 CRCbiome, and curatedMG). Edge colors represent phyla. Red line color indicates negative correlations

484 and blue indicates positive correlations. Line thickness indicates number of datasets the correlation was
485 observed in. *P. succi* = *P. succinatutens*; *P. sp* = *P. sp 377*; *P. faec* = *P. faecium*

486 Association of *Phascolarctobacterium* species abundance with education but not
487 with diet and fecal blood concentration

488 *P. faecium* and *P. succinatutens* both use succinate as a primary carbon source, therefore we
489 investigated whether the relative abundance of *Phascolarctobacterium* species was associated
490 with diet and other lifestyle factors. Here, we employed the CRCbiome dataset with dietary and
491 lifestyle information. After adjusting for sex, age and screening center, there was a significant
492 association to alcohol consumption and increased abundance of *P. sp 377* ($p=0.018$ and
493 $\text{padj}>0.05$; Table S7). High-school ($\text{padj}=0.04$) and university education ($p=0.005$ and
494 $\text{padj}>0.05$) was associated with lower abundance of *P. succinatutens*. University education
495 ($p=0.03$ and $\text{padj}>0.05$) and those not married or cohabitating ($p=0.03$ and $\text{padj}>0.05$) was
496 associated with higher and lower abundance of *P. faecium*, respectively. The concentration of
497 blood in stool was not associated with the abundance of either of the three
498 *Phascolarctobacterium* species (all $p>0.05$).

499 Pangenome variability among *Phascolarctobacterium* species

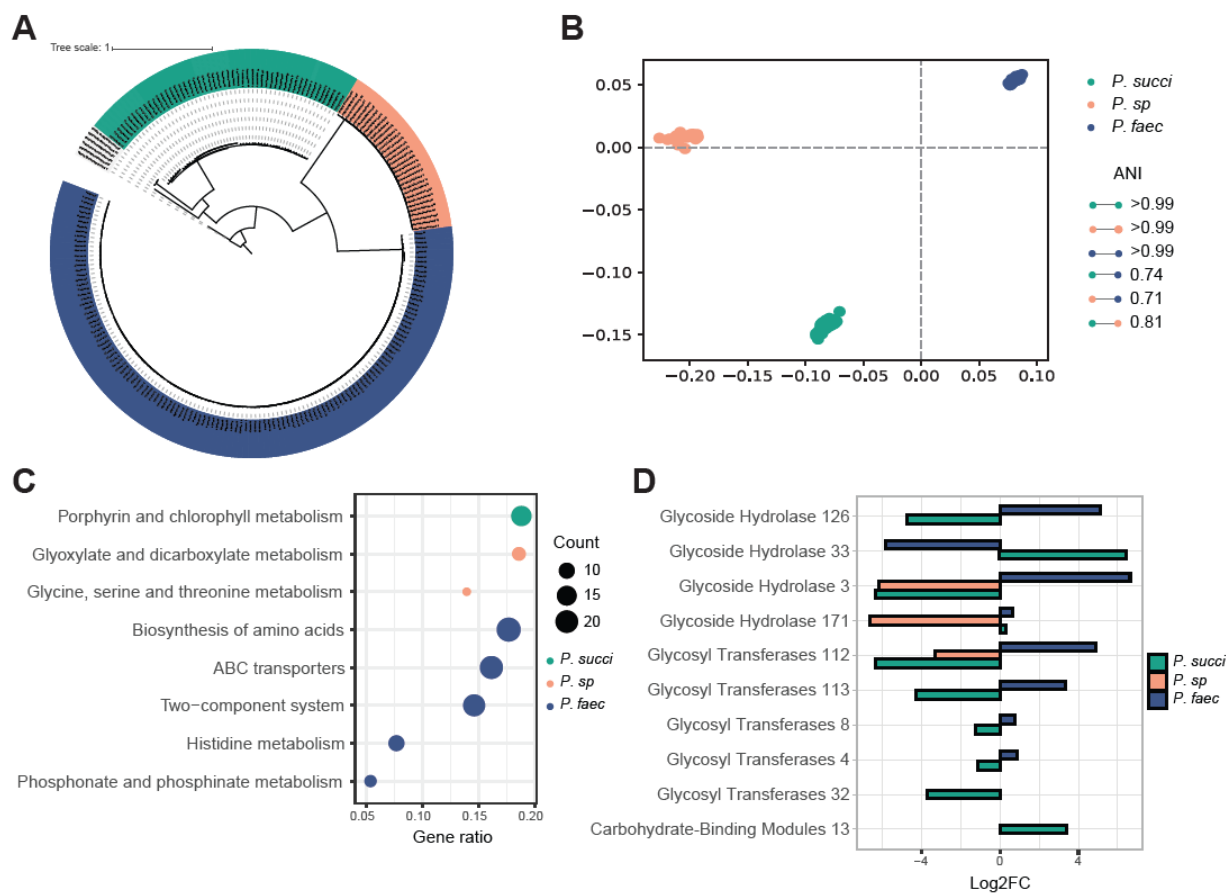
500 Based on metagenome sequencing data from CRCbiome, 221 high quality genomes of the
501 *Phascolarctobacterium* genus were identified. Fifty-two genomes were annotated as *P.*
502 *succinatutens*, 131 as *P. faecium*, and 32 as *P. sp 377* (Figure 5A). Mean within-species ANI was
503 99.9%, 99.9% and 99.8% for *P. faecium*, *P. sp 377* and *P. succinatutens* respectively, and mean
504 between-species ANI of 73.9% (Figure 5B).

505
506 Pangenome analysis for all *Phascolarctobacterium* genomes identified 25 847 gene clusters, with
507 1423 of them being ubiquitous ($\geq 95\%$) within a species, and not found in the others (species-
508 specific cores). On average, each genome contained 2065 gene clusters. Specifically, the average
509 for *P. succinatutens* was 2071, *P. sp 377* 1752, and *P. faecium* 2153 gene clusters. Only 197
510 gene clusters were identified in $\geq 95\%$ of *Phascolarctobacterium* genomes (genus-level core).
511 17127 gene clusters were annotated with UniRef, 1804 with KEGG pathways, and 65 with CAZy

512 annotations. All species-specific cores had multidrug resistance genes, metalloβ-lactamases, 2-
513 thiouracil desulfurase enabling H₂S production and contained various virulence factors. For
514 example, *P. succinatutens* genomes contained amylovoran and holin-like protein genes (Table
515 S8); *P. sp 377* - holin-like protein genes (Table S9); and *P. faecium* - heme-binding protein,
516 exfoliative toxin, hemolysis and immunity protein genes (Table S10).

517
518 There was an over-representation of genes within the porphyrin and chlorophyll metabolism
519 KEGG pathway in *P. succinatutens*. Glyoxylate and diglyxolate metabolism and glycine, serine
520 and threonine metabolism KEGG pathways were over-represented in *P. sp 377*. *P. faecium*
521 genomes were enriched in histidine metabolism, ABC transporters, two component system,
522 phosphonate and phosphinate metabolism, and biosynthesis of amino acids KEGG pathway
523 genes (all $p_{adj} < 0.05$, Figure 5C, Table S11).

524
525 With regard to carbohydrate-active enzymes, two CAZy belonging to glycoside hydrolases
526 family and one belonging to carbohydrate binding molecules family were significantly more
527 prevalent in *P. succinatutens* compared to the two other species. Three CAZy belonging to
528 glycoside hydrolases and four belonging to glycosyl transferases family were more prevalent in
529 *P. faecium* compared to the two others (all $p_{adj} < 0.05$, Figure 5D, Table S12). Glycoside
530 hydrolase 171 was present in all *P. succinatutens* and *P. faecium*, but completely missing in *P.*
531 *sp*. Glycoside hydrolase 33 was exclusively found in *P. succinatutens* (88% of genomes) and
532 glycoside hydrolase 3 exclusively in *P. faecium* (98% of genomes).



533
 534
 535 Figure 5: *Phascolarctobacterium* species genome comparison. A) Core-genome maximum likelihood tree
 536 representing all *Phascolarctobacterium* genomes in CRCbiome used for pangenome analyses, 52
 537 genomes from *P. succinatutens*, 32 from *P. sp 377*, and 131 from *P. faecium*. B) Multi-dimensional
 538 scaling of *Phascolarctobacterium* genomes based on their pairwise ANI distances. The average within
 539 species ANI and between species ANI are presented in the legend. C) Enrichment analysis of pathways
 540 with a significant over-representation of KEGG genes from either *P. succinatutens*, *P. sp 377*, or *P.*
 541 *faecium*. KEGG genes included in the analyses were those that were significantly different between
 542 whichever species against the two others combined, as determined by a chi-square test ($p_{adj} < 0.05$). Size
 543 of the dot point represents the number of KEGG genes within the relevant pathway. D) Log2FC of the
 544 significantly different CAZy enzymes between one species versus the two others combined as determined
 545 by a chi-square test ($p_{adj} < 0.05$). Only samples from CRCbiome are included in these analyses. ANI =
 546 Average Nucleotide Identity; Log2FC = Log2 fold-change; *P. succi* = *P. succinatutens*; *P. sp* = *P. sp 377*;
 547 *P. faec* = *P. faecium*.

548

549 Discussion

550 Based on our findings in two independent Norwegian cohorts, we replicate an association
551 between increased abundance of *P. succinatutens* and adenoma/CRC in the large international
552 curatedMG dataset. Three species were identified within the *Phascolarctobacterium* genus to be
553 nearly mutually exclusive, forming distinct microbial communities, potentially defining a CRC-
554 relevant microbial state. *P. succinatutens* was more common in men, in line with their increased
555 CRC risk. Together, this puts *P. succinatutens* on the list of highly relevant and reproducibly
556 CRC-associated bacteria.

557

558 In this study, we describe three distinct species within the genus *Phascolarctobacterium*. These
559 were *P. succinatutens*, *P. faecium* and one uncultured species referred to as *P. sp 377*, all with a
560 between species ANI of <95% and a limited core genome. Using qPCR we linked > 200 high
561 quality genomes from the species mentioned encompassing four datasets to our previously
562 identified CRC-associated 16S rRNA gene ASVs. qPCR assay was more sensitive than NGS in
563 detecting low abundant *Phascolarctobacterium*.

564

565 We observed a mutually exclusive relationship between *Phascolarctobacterium* species across
566 datasets and regardless of methods. The three different species of *Phascolarctobacterium* formed
567 species-specific bacterial and viral networks, in addition to different overall community structure.
568 *P. faecium* composition was more similar to those without any *Phascolarctobacterium*, whereas
569 *P. succinatutens* was markedly distinct. These distinct community structures could indicate
570 competition for resources or niche adaptation. Interestingly, all *Phascolarctobacterium* species
571 were negatively correlated with *Dialister* and tended to have positive correlations with
572 *Bacteroides*, suggesting that these community structures extend beyond the
573 *Phascolarctobacterium* genus.

574

575 Bacteria in the large intestine ferment complex carbohydrates and fibers and produce short-chain
576 fatty acids (SCFA), primarily acetate, butyrate and propionate. SCFAs, and especially butyrate,

577 have been proposed as potential biomarkers for CRC as they play a role in strengthening of the
578 gut barrier and modulation of immune responses⁶⁴. Succinate is an SCFA precursor and serves
579 as a substrate for several bacteria, including *Phascolarctobacterium* and *Dialister*⁶⁵. This
580 common reliance on succinate makes them potential competitors and might explain the observed
581 negative correlations. Positive feedback loop between succinate-producing *Bacteroides*
582 *thetaiotaomicron* and both *Dialister hominis*⁶⁶ and *P. faecium*¹⁰ has been demonstrated.

583
584 The three *Phascolarctobacterium* species shared only a small conserved genus level core genome
585 of about 0.76% of their genes, supporting distinct niche adaptation. For example, we observed
586 significant variations in metabolic capacity. Interestingly, glycoside hydrolase family 33 was
587 found only in *P. succinatutens*. Glycoside hydrolase family 33 comprises sialidases that break
588 down sialic acid from diet (mainly red meat) and potentially from the mucus layer in the
589 intestine⁶⁷ causing inflammation^{68,69}. In contrast, Glycoside hydrolase family 3 was found
590 exclusively, and in almost 100% of *P. faecium* genomes and is involved in a range of
591 mechanisms including bacterial pathogen defense, cell-wall remodeling, energy metabolism, and
592 cellulosic biomass degradation⁷⁰. Carbon starvation protein, a membrane protein, was found to be
593 unique to *P. sp 377*. Carbon starvation is exhibited by bacteria when they experience depletion of
594 carbon sources for their metabolic process⁷¹ and may provide *P. sp 377* a selective advantage in
595 nutrient limited conditions.

596
597 Bacterial virulence factors are employed in bacterial warfare, and are often detrimental to host
598 health⁷²⁻⁷⁴. We found different virulence factors for the three species. Holin-like protein was
599 present in only *P. succinatutens* and *P. sp 377*. Holin-like proteins control cell wall lysis by
600 producing pores in the cell membrane and can be involved in biofilm formation⁷⁵ contributing to
601 chronic inflammation in the colon, a known risk factor for CRC^{76,77}. Another gene involved in
602 biofilm formation, TabA, was specific to *P. succinatutens*. We also found an overrepresentation
603 of porphyrin and chlorophyll metabolism in *P. succinatutens*. Succinate is the main precursor and
604 porphyrin is an intermediate of heme production, which is closely linked to the TCA cycle.
605 Succinyl-CoA is the intermediate compound of succinate in the TCA cycle and is released upon
606 production of an ATP molecule⁷⁸. In our previous work we showed a lower abundance of several

607 pathways related to heme biosynthesis in high risk adenomas compared to healthy controls⁵.
608 Haem-binding uptake protein (Tiki superfamily) and hemolysin III protein were identified as
609 distinct to *P. faecium*. Tiki proteins may function as Wnt proteases, counteracting the Wnt
610 signaling pathway⁷⁹, a pathway which is commonly deregulated in CRC⁸⁰. Hemolysin III
611 exhibits hemolytic activity and contributes to the destruction of erythrocytes by pore formation⁸¹.
612 Together, our findings from the pangenome analyses contribute to a deeper understanding of the
613 functional diversity of *Phascolarctobacterium* species in the CRC microbiome.

614
615 We replicate our previous findings of an association between increased abundance of *P.*
616 *succinatutens* and adenomas/CRC. Several studies have reported similar associations at the genus
617 level^{13, Peters, 2016 #1244,15}, with few having looked at species level. Both our previous work
618 including 17 years of follow-up⁵, and Yachida et al.¹² found increased abundance of *P.*
619 *succinatutens* in the early stages of CRC. We observed lower abundance of *P. faecium* in
620 adenomas and also low levels of co-occurrences between *P. succinatutens* and *P. faecium*. This
621 could indicate that the gut community might shift from a low-risk *P. faecium* community to a
622 high-risk *P. succinatutens* community in early cancerogenesis.

623
624 Noteworthy, we found a higher prevalence of *P. succinatutens* in men than in women across
625 cohorts independent of the colonoscopy outcome. Men have an elevated risk for CRC⁸², often
626 attributed to lifestyle and dietary factors^{83,84}. We did, however, not find an association between
627 *Phascolarctobacterium* abundance and host diet and lifestyle, nor with presence of blood in
628 stool. On the contrary, the observed association with education could be a proxy for
629 socioeconomic status where low socioeconomic status have been linked to increased risk of
630 CRC^{85,86}.

631
632 Here we report consistent findings of *Phascolarctobacterium* across cohorts with different
633 methods, which emphasizes the reliability of our results and strengthens the validity of the study.
634 However, this study has some limitations. All participants in the CRCbiome study are FIT
635 positive and therefore have blood in their stool something which has been suggested to alter the
636 microbiome composition⁸⁷ and could also be a sign of colonic inflammation. It may also

637 introduce selection bias in the cohort. This may provide a reason for why we did not observe an
638 association between *Phascolarctobacterium* abundance and adenoma/CRC in the CRCbiome
639 cohort.

640
641 External factors like smoking, diet and gut flora may influence different stages along the
642 adenoma-carcinoma sequence of events leading to bowel cancer. The interplay between
643 *Phascolarctobacterium* species revealed in this study adds further to this complexity revealing
644 possible CRC-associated microbial networks and genomic characteristics.

645 Conclusion

646 Our study reveals that three *Phascolarctobacterium* species form distinct microbial communities
647 in the gut, each possessing different virulence factors and metabolic capabilities. We found that
648 microbiome composition varies significantly according to which *Phascolarctobacterium* species
649 is dominating. The verification of the *P. succinatutens* association with adenomas and CRC, and
650 the observation of increased abundance of *P. faecium* in controls, suggests that the gut
651 community might shift from a low-risk *P. faecium* community to a high-risk *P. succinatutens*
652 community in early cancerogenesis.

653
654
655 **Funding:** This work was funded by the South-Eastern Norway Regional Health Authority (project
656 number 2020056 and 2022067), Oslo Metropolitan University (project number 202401) and Akershus
657 University Hospital. The CRCbiome study was funded by grants from the Norwegian Cancer Society
658 (project number 190179 and 198048). Sequencing of the NORCCAP samples was funded by the Cancer
659 Registry of Norway funds.

660
661 **Disclosures:** Authors have no conflict of interest to disclose.

662
663 **Author Contributions:**

664 TBR and HT designed the research.

665 CBJ, EB, EA, TBR, TS, HT and VB conducted the research.

666 CBJ, EB, EA and TS analyzed data or performed statistical analysis.

667 CBJ and TS drafted the paper.

668 All authors read and approved the final manuscript.

669

670 **Data availability:**

671 Data from the CRCbiome project have been deposited in the database Federated EGA under accession
672 code EGAS50000000170 (<https://ega-archive.org/studies/EGAS50000000170>) and the

673 Curatedmetagenomedata is available here: <https://waldronlab.io/curatedMetagenomicData/index.html>.

674 Due to the sensitive nature of the data derived from human subjects, including personal health
675 information, analyses and sharing of data from cohorts in this project must comply with the General Data
676 Protection Regulation (GDPR). Data processors must have approval from the Regional Committee for
677 Medical Research in Norway (REC), legal basis according to GDPR Article 6 and 9 and the need for a
678 Data Protection Impact Assessment (DPIA) according to GDPR article 35 must be considered. Requests
679 for data access can be directed to corresponding author Trine B Rounge. The custom R scripts used in this
680 study are available at: https://github.com/Rounge-lab/Phascolarctobacterium_CRC.

681

682 **Acknowledgments**

683 We would like to acknowledge Jan-Inge Nordby for his work on preparing both NORCCAP and
684 CRCbiome samples, and for performing the DNA extractions. Elina Vinberg has also contributed with
685 sample handling and project coordination in both NORCCAP and CRCbiome projects. Library
686 preparation and sequencing of NORCCAP and CRCbiome samples were performed at the FIMM
687 Technology Centre supported by HiLIFE and Biocenter Finland. Therefore, we would like to thank Tiina
688 Hannunen, Harri A. Kangas, and Pekka J. Ellonen for their service and good cooperation. We would also
689 like to thank the members of our research groups Maja Jacobsen, Ane Sørli Kværner, Paula Berstad and
690 Paula Istvan. Thank you for the great working environment and for fruitful discussions. We thank the
691 Department of Multidisciplinary Laboratory Science and Medical Biochemistry at Akershus University
692 Hospital for providing laboratory facilities. We are grateful to Tone M. Tanneæs, Aina E.F. Moen, Gro
693 Gunderson, Eva Smedsrud and John Christopher Noone for their contribution in sample extraction and
694 sequencing.

695 **References**

696

- 697 1. Liang, S., *et al.* Gut microbiome associated with APC gene mutation in patients with intestinal
698 adenomatous polyps. *Int J Biol Sci* **16**, 135-146 (2020).
- 699 2. Vacante, M., Ciuni, R., Basile, F. & Biondi, A. Gut Microbiota and Colorectal Cancer
700 Development: A Closer Look to the Adenoma-Carcinoma Sequence. *Biomedicines* **8**(2020).
- 701 3. Muto, T., Bussey, H.J. & Morson, B.C. The evolution of cancer of the colon and rectum. *Cancer*
702 **36**, 2251-2270 (1975).
- 703 4. Brenner, H., *et al.* Risk of progression of advanced adenomas to colorectal cancer by age and sex:
704 estimates based on 840,149 screening colonoscopies. *Gut* **56**, 1585-1589 (2007).
- 705 5. Bucher-Johannessen, C., *et al.* Long-term follow-up of colorectal cancer screening attendees
706 identifies differences in *Phascolarctobacterium* spp. using 16S rRNA and metagenome
707 sequencing. *Frontiers in Oncology* **13**(2023).
- 708 6. Senthakumaran, T., *et al.* Microbial dynamics with CRC progression: a study of the mucosal
709 microbiota at multiple sites in cancers, adenomatous polyps, and healthy controls. *Eur J Clin*
710 *Microbiol Infect Dis* **42**, 305-322 (2023).
- 711 7. Shigeno, Y., Kitahara, M., Shime, M. & Benno, Y. *Phascolarctobacterium wakonense* sp. nov.,
712 isolated from common marmoset (*Callithrix jacchus*) faeces. *Int J Syst Evol Microbiol* **69**, 1941-
713 1946 (2019).
- 714 8. Wu, F., *et al.* *Phascolarctobacterium faecium* abundant colonization in human gastrointestinal
715 tract. *Exp Ther Med* **14**, 3122-3126 (2017).
- 716 9. Watanabe, Y., Nagai, F. & Morotomi, M. Characterization of *Phascolarctobacterium*
717 *succinatutens* sp. nov., an asaccharolytic, succinate-utilizing bacterium isolated from human
718 feces. *Appl Environ Microbiol* **78**, 511-518 (2012).
- 719 10. Ikeyama, N., *et al.* Microbial interaction between the succinate-utilizing bacterium
720 *Phascolarctobacterium faecium* and the gut commensal *Bacteroides thetaiotaomicron*.
721 *Microbiologyopen* **9**, e1111 (2020).
- 722 11. Ogata, Y., *et al.* Complete Genome Sequence of *Phascolarctobacterium faecium* JCM 30894, a
723 Succinate-Utilizing Bacterium Isolated from Human Feces. *Microbiol Resour Announc* **8**(2019).
- 724 12. Yachida, S., *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific
725 phenotypes of the gut microbiota in colorectal cancer. *Nat Med* **25**, 968-976 (2019).
- 726 13. Zackular, J.P., Rogers, M.A., Ruffin, M.T.t. & Schloss, P.D. The human gut microbiome as a
727 screening tool for colorectal cancer. *Cancer Prev Res (Phila)* **7**, 1112-1121 (2014).
- 728 14. Peters, B.A., *et al.* The gut microbiota in conventional and serrated precursors of colorectal
729 cancer. *Microbiome* **4**, 69 (2016).

- 730 15. Sarhadi, V., *et al.* Gut Microbiota and Host Gene Mutations in Colorectal Cancer Patients and
731 Controls of Iranian and Finnish Origin. *Anticancer research* **40**, 1325-1334 (2020).
- 732 16. Tunsjo, H.S., *et al.* Detection of *Fusobacterium nucleatum* in stool and colonic tissues from
733 Norwegian colorectal cancer patients. *Eur J Clin Microbiol Infect Dis* **38**, 1367-1376 (2019).
- 734 17. Holme, O., *et al.* Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and
735 mortality: a randomized clinical trial. *JAMA* **312**, 606-615 (2014).
- 736 18. Bretthauer, M., *et al.* NORCCAP (Norwegian colorectal cancer prevention): a randomised trial to
737 assess the safety and efficacy of carbon dioxide versus air insufflation in colonoscopy. *Gut* **50**,
738 604-607 (2002).
- 739 19. Bretthauer, M., *et al.* Design, organization and management of a controlled population screening
740 study for detection of colorectal neoplasia: attendance rates in the NORCCAP study (Norwegian
741 Colorectal Cancer Prevention). *Scand J Gastroenterol* **37**, 568-573 (2002).
- 742 20. Kvaerner, A.S., *et al.* The CRCbiome study: a large prospective cohort study examining the role
743 of lifestyle and the gut microbiome in colorectal cancer screening participants. *BMC cancer*
744 (2020).
- 745 21. Randel, K.R., *et al.* Colorectal cancer screening with repeated fecal immunochemical test versus
746 sigmoidoscopy: baseline results from a randomized trial. *Gastroenterology* (2020).
- 747 22. Kvaerner, A.S., *et al.* Associations of red and processed meat intake with screen-detected
748 colorectal lesions. *Br J Nutr*, 1-36 (2022).
- 749 23. Kvaerner, A.S., *et al.* Associations of the 2018 World Cancer Research Fund/American Institute
750 of Cancer Research (WCRF/AICR) cancer prevention recommendations with stages of colorectal
751 carcinogenesis. *Cancer Med* **12**, 14806-14819 (2023).
- 752 24. Pasolli, E., *et al.* Accessible, curated metagenomic data through ExperimentHub. in *Nat Methods*,
753 Vol. 14 1023-1024 (2017).
- 754 25. Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K. & Schloss, P.D. Development of a
755 dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the
756 MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**, 5112-5120 (2013).
- 757 26. Raju, S.C., *et al.* Reproducibility and repeatability of six high-throughput 16S rDNA sequencing
758 protocols for microbiota profiling. *J Microbiol Methods* **147**, 76-86 (2018).
- 759 27. Klindworth, A., *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical
760 and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**, e1 (2013).
- 761 28. Rounge, T.B., *et al.* Evaluating gut microbiota profiles from archived fecal samples. *BMC*
762 *Gastroenterol* **18**, 171 (2018).

- 763 29. Bolyen, E., *et al.* Reproducible, interactive, scalable and extensible microbiome data science
764 using QIIME 2. *Nat Biotechnol* **37**, 852-857 (2019).
- 765 30. McMurdie, P.J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and
766 graphics of microbiome census data. *PloS one* **8**, e61217 (2013).
- 767 31. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
768 data. *Bioinformatics* **30**, 2114-2120 (2014).
- 769 32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment
770 of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
- 771 33. Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079
772 (2009).
- 773 34. Beghini, F., *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse
774 microbial communities with bioBakery 3. *Elife* **10**(2021).
- 775 35. Kieser, S., Brown, J., Zdobnov, E.M., Trajkovski, M. & McCue, L.A. ATLAS: a Snakemake
776 workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC*
777 *bioinformatics* **21**, 1-8 (2020).
- 778 36. Bushnell, B. BMap: a fast, accurate, splice-aware aligner. (Lawrence Berkeley National
779 Lab.(LBNL), Berkeley, CA (United States), 2014).
- 780 37. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. metaSPAdes: a new versatile
781 metagenomic assembler. *Genome research* **27**, 824-834 (2017).
- 782 38. Sieber, C.M., *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and
783 scoring strategy. *Nature microbiology* **3**, 836-843 (2018).
- 784 39. Kang, D.D., *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
785 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
- 786 40. Wu, Y.-W., Simmons, B.A. & Singer, S.W. MaxBin 2.0: an automated binning algorithm to
787 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607 (2016).
- 788 41. Olm, M.R., Brown, C.T., Brooks, B. & Banfield, J.F. dRep: a tool for fast and accurate genomic
789 comparisons that enables improved genome recovery from metagenomes through de-replication.
790 *The ISME journal* **11**, 2864-2868 (2017).
- 791 42. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. CheckM: assessing
792 the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome*
793 *Res* **25**, 1043-1055 (2015).
- 794 43. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P. & Parks, D.H. GTDB-Tk: a toolkit to classify
795 genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927 (2019).

- 796 44. Brukner, I., Longtin, Y., Oughton, M., Forgetta, V. & Dascal, A. Assay for estimating total
797 bacterial load: relative qPCR normalisation of bacterial load with associated clinical implications.
798 *Diagn Microbiol Infect Dis* **83**, 1-6 (2015).
- 799 45. Information, N.C.f.B. & Camacho, C. *BLAST (r) command line applications user manual*,
800 (National Center for Biotechnology Information (US), 2008).
- 801 46. Madeira, F., *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic
802 Acids Res* **50**, W276-W279 (2022).
- 803 47. Trifinopoulos, J., Nguyen, L.T., von Haeseler, A. & Minh, B.Q. W-IQ-TREE: a fast online
804 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* **44**, W232-235 (2016).
- 805 48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree
806 display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021).
- 807 49. Shaffer, M., *et al.* DRAM for distilling microbial metabolism to automate the curation of
808 microbiome function. *Nucleic Acids Res* **48**, 8883-8900 (2020).
- 809 50. Aramaki, T., *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and
810 adaptive score threshold. *Bioinformatics* **36**, 2251-2252 (2020).
- 811 51. Zhang, H., *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.
812 *Nucleic Acids Res* **46**, W95-W101 (2018).
- 813 52. Wang, Y., *et al.* A crowdsourcing open platform for literature curation in UniProt. *PLoS biology*
814 **19**, e3001464 (2021).
- 815 53. Page, A.J., *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**,
816 3691-3693 (2015).
- 817 54. Katoh, K., Rozewicki, J. & Yamada, K.D. MAFFT online service: multiple sequence alignment,
818 interactive sequence choice and visualization. *Brief Bioinform* **20**, 1160-1166 (2019).
- 819 55. Istvan, P., *et al.* Exploring the gut DNA virome in fecal immunochemical test stool samples
820 reveals associations with lifestyle in a large population-based study. *Nature Communications*
821 **15**(2024).
- 822 56. Ripley, B., *et al.* Package ‘mass’. *Cran r* **538**, 113-120 (2013).
- 823 57. Oksanen, J., *et al.* Vegan: community ecology package. Ordination methods, diversity analysis
824 and other functions for community and vegetation ecologists. 2.3-1 (2015).
- 825 58. Kassambara, A. rstatix: Pipe-friendly framework for basic statistical tests. *R package version 0.6.*
826 *0* (2020).
- 827 59. Shannon, P., *et al.* Cytoscape: a software environment for integrated models of biomolecular
828 interaction networks. *Genome research* **13**, 2498-2504 (2003).

- 829 60. Chen, M. & Yu, G. MicrobiomeProfiler: An R/shiny package for microbiome functional
830 enrichment analysis. (2023).
- 831 61. Wickham, H., *et al.* Welcome to the Tidyverse. *Journal of open source software* **4**, 1686 (2019).
- 832 62. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful
833 Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*
834 **57**, 289-300 (1995).
- 835 63. Istvan, P., *et al.* Exploring the gut virome in fecal immunochemical test stool samples reveals
836 novel associations with lifestyle in a large population-based study. *medRxiv*, 2023.2008.
837 2024.23294548 (2023).
- 838 64. Hou, H., *et al.* Gut microbiota-derived short-chain fatty acids and colorectal cancer: Ready for
839 clinical translation? *Cancer Lett* **526**, 225-235 (2022).
- 840 65. Wei, Y.H., Ma, X., Zhao, J.C., Wang, X.Q. & Gao, C.Q. Succinate metabolism and its regulation
841 of host-microbe interactions. *Gut Microbes* **15**, 2190300 (2023).
- 842 66. Sakamoto, M., *et al.* *Alistipes communis* sp. nov., *Alistipes dispar* sp. nov. and *Alistipes*
843 *onderdonkii* subsp. *vulgaris* subsp. nov., isolated from human faeces, and creation of *Alistipes*
844 *onderdonkii* subsp. *onderdonkii* subsp. nov. *Int J Syst Evol Microbiol* **70**, 473-480 (2020).
- 845 67. Lipnicanova, S., Chmelova, D., Ondrejovic, M., Frecer, V. & Miertus, S. Diversity of sialidases
846 found in the human body - A review. *Int J Biol Macromol* **148**, 857-868 (2020).
- 847 68. Bell, A., Severi, E., Owen, C.D., Latousakis, D. & Juge, N. Biochemical and structural basis of
848 sialic acid utilization by gut microbes. *The Journal of biological chemistry* **299**, 102989 (2023).
- 849 69. Juge, N., Tailford, L. & Owen, C.D. Sialidases from gut bacteria: a mini-review. *Biochem Soc*
850 *Trans* **44**, 166-175 (2016).
- 851 70. Geoff Fincher, B.M., Harry Brumer. Glycoside Hydrolase Family 3 in *CAZypedia*, Vol. 2024
852 (2023).
- 853 71. Rasmussen, J.J., *et al.* *Campylobacter jejuni* carbon starvation protein A (CstA) is involved in
854 peptide utilization, motility and agglutination, and has a role in stimulation of dendritic cells. *J*
855 *Med Microbiol* **62**, 1135-1143 (2013).
- 856 72. Cheng, W.T., Kantilal, H.K. & Davamani, F. The Mechanism of *Bacteroides fragilis* Toxin
857 Contributes to Colon Cancer Formation. *Malays J Med Sci* **27**, 9-21 (2020).
- 858 73. Pleguezuelos-Manzano, C., *et al.* Mutational signature in colorectal cancer caused by genotoxic
859 pks(+) *E. coli*. *Nature* **580**, 269-273 (2020).
- 860 74. Casterline, B.W., Hecht, A.L., Choi, V.M. & Bubeck Wardenburg, J. The *Bacteroides fragilis*
861 pathogenicity island links virulence and strain competition. *Gut Microbes* **8**, 374-383 (2017).

- 862 75. Saier Jr, M.H. & Reddy, B.L. Holins in bacteria, eukaryotes, and archaea: multifunctional
863 xenologues with potential biotechnological and biomedical applications. *Journal of bacteriology*
864 **197**, 7-17 (2015).
- 865 76. Dejea, C.M., *et al.* Microbiota organization is a distinct feature of proximal colorectal cancers.
866 *Proc Natl Acad Sci U S A* **111**, 18321-18326 (2014).
- 867 77. Raskov, H., Kragh, K.N., Bjarnsholt, T., Alamili, M. & Gogenur, I. Bacterial biofilm formation
868 inside colonic crypts may accelerate colorectal carcinogenesis. *Clin Transl Med* **7**, 30 (2018).
- 869 78. Stojanovski, B.M., *et al.* 5-Aminolevulinic synthase catalysis: The catcher in heme biosynthesis.
870 *Mol Genet Metab* **128**, 178-189 (2019).
- 871 79. Sanchez-Pulido, L. & Ponting, C.P. Tiki, at the head of a new superfamily of enzymes.
872 *Bioinformatics* **29**, 2371-2374 (2013).
- 873 80. Bienz, M. & Clevers, H. Linking colorectal cancer to Wnt signaling. *Cell* **103**, 311-320 (2000).
- 874 81. Ramarao, N. & Sanchis, V. The pore-forming haemolysins of bacillus cereus: a review. *Toxins*
875 (*Basel*) **5**, 1119-1139 (2013).
- 876 82. Siegel, R.L., Wagle, N.S., Cercek, A., Smith, R.A. & Jemal, A. Colorectal cancer statistics, 2023.
877 *CA Cancer J Clin* **73**, 233-254 (2023).
- 878 83. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality,
879 survival, and risk factors. *Prz Gastroenterol* **14**, 89-103 (2019).
- 880 84. Gunter, M.J., *et al.* Meeting report from the joint IARC-NCI international cancer seminar series: a
881 focus on colorectal cancer. *Ann Oncol* **30**, 510-519 (2019).
- 882 85. Doubeni, C.A., *et al.* Socioeconomic status and the risk of colorectal cancer: an analysis of more
883 than a half million adults in the National Institutes of Health-AARP Diet and Health Study.
884 *Cancer* **118**, 3636-3644 (2012).
- 885 86. Aarts, M.J., Lemmens, V.E., Louwman, M.W., Kunst, A.E. & Coebergh, J.W. Socioeconomic
886 status and changing inequalities in colorectal cancer? A review of the associations with risk,
887 treatment and outcome. *European journal of cancer (Oxford, England : 1990)* **46**, 2681-2695
888 (2010).
- 889 87. Chenard, T., Malick, M., Dube, J. & Masse, E. The influence of blood on the human gut
890 microbiome. *BMC Microbiol* **20**, 44 (2020).

891