Title: *Biogeography of human-health associated butyrate-producing bacteria*

Authors: *Joel E. Brame[a], Craig Liddicoat[a,b], Catherine A. Abbott[a], Robert A. Edwards[a], Jake M. Robinson[a], Nicolas E. Gauthier[c], Martin F. Breed[a]

a. College of Science and Engineering, Flinders University, Bedford Park, SA 5042, Australia
b. School of Public Health, The University of Adelaide, SA 5005, Australia
c. Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

* Corresponding author: Joel E. Brame

Abstract

Butyrate-producing bacteria are found in many ecosystems and organisms, including humans. They ferment organic matter, producing the by-product butyrate, a short-chain fatty acid with important roles in human health. Several human diseases have been associated with a decreased abundance of butyrate-producing bacteria in the gut. Outdoor environments can potentially replenish the abundance of these commensal bacteria in humans. However, the environmental sources and exposure pathways remain poorly understood. Here we developed new normalized Butyrate Production Capacity (BPC) indices derived from global metagenomic (n=16,176) and Australian soil 16S rRNA (n=1,285) data to geographically detail the environments that associate with bacterial butyrate production potential. We show that the highest BPC scores were in anoxic and fermentative environments, including plant rhizospheres and the gut of vertebrates. Among land types, higher BPC scores were in soils from temperate urban hinterlands and bogs. Climatic and geographical variables were the primary drivers of BPC score variation across land types. We show that the potential for ambient human exposure to health-promoting butyrate-producing bacteria should be highest in residential woodlands, dense urban environments with moderate rainfall, and particular pastures and croplands. This new biogeographic understanding of how and where humans are exposed to these important health-promoting microbes should be integrated into health and environmental policies to improve public health outcomes.

Keywords:

butanoate, gut microbiome, ecosystem services, soil microbiota, urban green space, microbiome exposure

1. INTRODUCTION

[1]Butyrate-producing bacteria are associated with host organisms and free-living in the environment. They have critical roles in breaking down organic products including fibres(Baxter et al., 2019) and cellulose(Goldfarb et al., 2011). Given suitable organic substrates and anaerobic conditions, these bacteria can produce butyrate, a short-chain fatty acid, as a metabolic by-product of fermentation. In soils, butyrate is associated with the suppression of soil-borne plant pathogens(Poret-Peterson et al., 2019). In humans, butyrate and the presence of butyrate-producing bacteria have direct implications for many health outcomes(Liu et al., 2018; Valles-Colomer et al., 2019). For example, during childhood, a delay in the assemblage of butyrate producers can contribute to atopic illnesses(Roduit et al., 2019). During adulthood, a reduced abundance of butyrate-producing bacteria in the human gut is associated with several immune-related diseases including inflammatory bowel disease and multiple sclerosis(Miyake et al., 2015; Parada Venegas et al., 2019). Thus, improved human health outcomes associate with an adequate supply of butyrate producers. However, poor diet, lifestyle, and antibiotics can cause the loss of butyrate producers(Sonnenburg and Sonnenburg, 2019). Beyond probiotic(Chen et al., 2020) and prebiotic(Cantu-Jungles et al., 2019) supplementation, strategies for replenishing the abundance of gut butyrate-producing bacteria that support human health remain poorly developed.

The human living environment influences the abundance of butyrate-producers in the human gut(Nurminen et al., 2018). Outdoor environmental microbiomes contribute to the indoor environmental microbiomes(Alfven et al., 2007; Parajuli et al., 2018). Therefore, living in rural, agricultural, and more biodiverse areas can promote ambient exposure to diverse

[1] BPC = Butyrate production capacity

microbiota, which then modify the internal gut microbiomes and promote immunotolerance(Liddicoat et al., 2020; Ottman et al., 2019; Rothschild et al., 2018).

However, in urban environments, residents' exposure to beneficial environmental bacteria with which humans have co-evolved has waned(Rook et al., 2013). Indeed, urbanisation is associated with a rise in immune-related chronic diseases(Flies et al., 2019). Thus, regular exposure to biodiverse outdoor environments, especially during childhood, could provide a strategy to increase exposure to health-promoting bacteria, including butyrate producers(Liddicoat et al., 2020; Mohammadkhah et al., 2018; Selway et al., 2020). Detailed insights into the biogeography of butyrate-producing bacteria will further the understanding of the exposure pathways and links between human and environmental health.

Here, we utilized global metagenomic datasets and a focussed regional analysis of continent-wide Australian 16S rRNA amplicon data to provide insight into the global biogeographical distribution of butyrate-producing bacteria. We developed new indices to estimate the butyrate production capacity (BPC) of representative samples from both metagenomic (BPC_{meta}) and 16S rRNA amplicon (BPC_{16S}) data.

2. MATERIALS AND METHODS

2.1. Gene selection and metagenome database interrogation

The butanoate (butyrate) synthesis pathways were reviewed using Seed viewer subsystems (https://pubseed.theseed.org/) and the KEGG pathway (https://www.genome.jp/kegg/pathway.html) to determine the genes coding for enzymes that are part of the butyrate production pathway. Based on these pathways, the following two genes were chosen for further analysis: *buk* (butyrate kinase) and *atoA* (acetate-CoA:acetoacetyl-CoA transferase subunit beta). *ACADS/bcd* (butyryl-CoA dehydrogenase)

and *ptb* (phosphate butyryltransferase) were analysed but subsequently excluded (all gene decisions explained in **Supplementary Table 1***)*. The genes *atoA* and *buk* participate in each of the two terminal pathways. The IMG/M genome database(Chen et al., 2021) was then searched for each butyrate-related gene to obtain their mean counts among genomes with at least one copy of either gene (**Supplementary Table 2**).

We next searched global metagenomic databases for *atoA* and *buk* to find metagenomes that suggest the potential presence of butyrate-producing bacteria. Initial gene and translated gene searches of metagenomics data at searchsra.org using bowtie2 and diamond, respectively, yielded low numbers of samples returned and/or high E-values. The largest datasets came from searching IMG/M using EC numbers for each butyrate-production enzyme (butyrate kinase = EC 2.7.2.7, acetate-CoA:acetoacetyl-CoA transferase subunit beta = EC 2.8.3.8) as well as three enzymes with single-copy genes (phenylalanine—rRNA ligase = EC 6.1.1.20; guanylate kinase = EC 2.7.4.8; alanine—tRNA ligase = EC 6.1.1.7). Sample datasets with the genes *atoA* (*n*=19,993) and *buk* (*n*=16,263) were downloaded as our starting point for metagenomics analysis. We found 14,407 datasets with both genes and datasets with one gene but not the other (*atoA n*=6,330 and *buk n*=1,856), which created an initial dataset of 22,593 metagenomic samples (**Supplementary Table 3**).

Counts for each butyrate-production gene were normalized by dividing by counts of the single-copy gene *pheS*, which codes for the protein phenylalanine—tRNA ligase alpha subunit and was used as a proxy for total genome count. Counts for two other single-copy genes (*GUK1*: guanylate kinase and *alaS*: alanine—tRNA ligase) were also inspected, but they were not used because *GUK1* searches showed low counts, and *alaS* showed slightly different but proportional counts to *pheS*, which validated the usage of *pheS* to normalize estimates of total genomes in the samples. However, 115 samples did not include *pheS* count data and were removed from our analysis. To minimize skewed data, outliers with a pheS

count <100 (*n*=5,479) and >50,000 (*n*=19) were removed from analysis. In addition, samples where the ratio (*buk*+*atoA)*/*pheS* was >30%, implying an inflated dominance of the two genes of interest, were removed from the analysis (*n*=804). The remaining 16,176 samples were analysed for this project.

2.2. BPC scores for metagenomic samples

To derive the Butyrate Production Capacity (BPC_{meta}) score for each sample with metagenomic data, the following formula was developed:

Sample BPC_{meta} score =

$$\log10\left(\sum_{i=1}^{n}\left[\left(\frac{CountGene1}{MeanGene1Copies}\right)/Count\ SCG + \left(\frac{CountGene2}{MeanGene2Copies}\right)/Count\ SCG\right] * 10{,}000\right)$$

where: SCG = single copy gene (*pheS*)

Gene1 = *buk*, Gene 2 = *atoA*

CountGene1, CountSCG are from global metagenomics sample datasets

MeanGeneXCopies = mean count of copies of gene X among all genomes found from searches of gene X within the IMG/M genome database.

Once BPC_{meta} scores were computed and added to the spreadsheet using Excel formulas, the samples were sorted into six categories: soil and terrestrial sediments, aquatic, human, animal, plant, and agro-industrial (**Supplementary Table 4**). An additional "Excluded" category was created for samples that did not fall within our research question, such as subsurface, contaminated, and experimentally altered samples. Samples were then grouped by subcategories for statistical testing and results of interest: soil samples grouped by anthrome classification; aquatic samples grouped by source subcategory; human samples grouped by body compartment; animal samples grouped by vertebrate/invertebrate and by

phylum; plants were grouped by compartment; agro-industrial samples grouped by source site. For anthrome "Class" categories, to reduce bias, individual studies (n=2) whose samples accounted for >50% of the total class samples were removed from the analysis.

To determine if our BPC_{meta} formula was only identifying anaerobicity rather than specifically butyrate production, we adapted the BPC_{meta} formula to represent ethanol production, a pathway that also requires anaerobic conditions. The butyrate synthesis genes were replaced with the terminal gene for alcohol dehydrogenase (*ADH*, EC 1.1.1.1) to derive an Ethanol Production Capacity (EPC) score. We then compared the EPC scores of the soil metagenomic samples in section 3.3 with their BPC_{meta} scores (**Supplementary Figure 1**).

Statistical tests were then performed in R (version 4.0.2Team, 2021). Shapiro-Wilk test was used to determine the normality of distribution. In each case, the data did not fit a normal distribution, and either the non-parametric Kruskal-Wallis test or Wilcoxon rank-sum test was then used to test the significance of between-group variation. Due to a high n in some subgroups, a post hoc Dunn test with Bonferroni correction was used to compare subgroup pairwise differences at α=0.05. ggplot2 (version 3.3.5Wickham, 2016) was used for data visualisation. Mapping of soil samples was performed from 2,850 sample metadata coordinates after excluding 360 samples with coordinates with less than two decimal points and 153 samples with no coordinates.

2.3. BPC scores for 16S rRNA amplicon samples

To assemble 16S rRNA gene abundance data in Australian soil samples, the Australian Microbiome Initiative(Bissett et al., 2016) database was queried for the following parameters: Amplicon = "27F519R", Kingdom = "bacteria", Environment = "is soil", Depth = "between 1 and 10" (cm). The zOTU abundances and metadata for each resulting sample (n=3,023) were downloaded. We used the phyloseq package(McMurdie and Holmes, 2013) for managing and cleaning the 16S rRNA data. We removed all "chloroplast" and

"mitochondria" data, which have non-bacterial origins. We removed low abundance zOTUs that did not occur in at least two samples and had total counts of <20, which may have arisen from processing errors. In addition, we kept only samples with total sequences between 30,000 and 500,000 to remove outliers and samples with low read depth. The final sample size was *n*=2,795.

16S rRNA data often have relatively poor resolution at the genus and species level, so we focussed our BPC_{16S} derivation on family-level data. Using the Genome Taxonomy Database (GTDB) website interface and a set of putative butyrate-producing species (*n*=118) from Vital et al.(Vital et al., 2014), we identified the families with members from our species list (*n*=54, **Supplementary Table 5**). This family list was then matched with the Australian Microbiome Initiative taxonomy listings for each downloaded sample. Of the 54 taxonomic families with butyrate-producing bacteria analyzed, 31 families had no representative zero-radius operational taxonomic units (zOTUs) in any sample. The proportion of butyrate-producers in each family was used to estimate the abundance of butyrate-producing taxa within each sample and a corresponding BPC_{16S} score, as follows:

Sample BPC_{16S} score =

$$\log10\left(\sum_{i=1}^{n=54}\left[(CountZOTUs\ from\ butyrate\ producing\ family\ 1)\left(\frac{\#\ butyrate\ producing\ species\ in\ family\ 1}{\#\ species\ in\ family\ 1}\right)+\right.\right.$$

$$(CountZOTUs\ from\ butyrate\ producing\ family\ 2)\left(\frac{\#\ butyrate\ producing\ species\ in\ family\ 2}{\#\ species\ in\ family\ 2}\right)+$$

$$\left.\left.\cdots+(CountZOTUs\ from\ butyrate\ producing\ family\ n)\left(\frac{\#\ butyrate\ producing\ species\ in\ family\ n}{\#\ species\ in\ family\ n}\right)\right]\right.$$

*10,000)

Where: family1= Acetonemaceae, family 2 = Acidaminococcaceae, … (see **Supplementary Table 5** for a list of all 54 families)

Count zOTUs in each butyrate-producing family are from Australian Microbiome Initiative datasets

169 # butyrate-producing species (and total binomial species) in each family are from the
170 GTDB.
171
172 16S rRNA amplicon datasets are often rarefied to normalize for sampling effort. However, in
173 soils the butyrate-producing bacteria are often rare taxa, which could more easily be missed
174 with rarefaction. Therefore, we chose to utilize unrarefied data. In addition, to reduce data
175 handling requirements, zOTU abundances, rather than relative abundances, were used for
176 analysis.

177 **2.4. Regional environmental correlation modelling for BPC_{16S} scores**

178 To provide further biogeographical context to butyrate-producing bacteria in soils, BPC_{16S}
179 scores were associated with geographically paired environmental metadata. We chose 16S
180 rRNA amplicon-based studies for this analysis because many soil studies in Australia have
181 utilized 16S data, and the Australian Microbiome Initiative facilitated access to a continental
182 coverage of data collected via a common sampling and bioinformatic protocol. By selecting a
183 manageable spatial scale (Australia only), we efficiently examined associations of a larger
184 pool of environmental characteristics with BPC_{16S} scores. We also used environmental
185 metadata from sources that focus solely on Australia (e.g., Atlas of Living Australia), which
186 differs from the metadata sources utilized in our global analyses (e.g., anthropogenic biomes).

187 For the analysis of potential environmental influences on the BPC scores, covariate data were
188 collated from a variety of sources and reflect a range of soil-forming factors (i.e. SCORPAN
189 variables(McBratney et al., 2003); S=soil; C=climate; O=organisms; R=relief; P=parent
190 material; A=age; N=spatial location) (see **Supplementary Table 6** for a list and description
191 of all environmental covariate data). We identified 49 predictor variables (43 numeric, 6
192 categorical) as being relevant to our study, for which data sets were downloaded from the
193 following sources: Australian Microbiome Initiative(Bissett et al., 2016) (e.g., organic

carbon, clay content %, conductivity), Atlas of Living Australia(Belbin, 2011; Williams et al.) (e.g., annual temperature range, aridity index annual mean), Soil and Landscape Grid of Australia(O'Brien, 2021) (e.g., Prescott index, topographic wetness index), and Geoscience Australia(Cudahy et al., 2012) (silica index). We used the best available resolution of source data as supplied to avoid introducing additional noise or bias into our analyses. For example, certain analytical test results were available from sample metadata corresponding to 16S rRNA amplicon data, and other environmental covariate data were extracted from gridded spatial environmental layers at points corresponding to the site locations.

Analysis of the predictor variables showed multiple instances of collinearity (e.g., r>0.80 or high Variable Inflation Factor scores >12), and scatterplots generated often showed a curvilinear relationship with BPC_{16S} scores (scatterplots shown in **Supplementary Figure 2**). Therefore, two method sequences less influenced by collinearity were chosen for subsequent analysis: principal components analysis into *k*-means clustering and decision tree modelling via Random Forest. Incomplete cases (n=1,510) were removed, leaving 1,285 samples in the analyses.

To further understand the relationships between environmental influences, we scaled and analysed the 43 continuous predictor variables using principal component (PC) analysis to reduce the dimensionality of the variables. BPC_{16S} scores were excluded from this analysis to avoid response variable influence. PC1 and PC2 demonstrated the highest explanation of variance (27.2% and 14.2% of variance explained, respectively) and were thus selected for truncation to maximize data visualisation (**Supplementary Figure 3**). *k*-means clustering was then performed on scaled original data to assign the samples into clusters. The optimal number of clusters was examined using the "elbow" method, silhouette method, and gap statistic method. While four was considered an optimal number of clusters, we examined results using both four-cluster and five-cluster analyses and found that the additional fifth

cluster more distinctly separated out the resulting land types. Thus, the five-cluster approach was selected for analysis. The resulting cluster data was collated, and BPC_{16S} scores were then matched and returned to the data set. Medians were calculated for each variable in each cluster, revealing environmental trends distinct to each cluster. Between-cluster significance was tested using the Kruskal-Wallis test. We gave each cluster a generalized description and plotted the sample geospatial coordinates into maps using ggmap(Kahle and Wickham, 2013) and Google maps to visualize their geographical distribution.

We then utilized Random Forest regression modelling(Breiman, 2001) to discern variable importance results and obtain partial dependence plots for each variable against BPC_{16S} scores. Only the 43 continuous variables were included in this analysis. The model fit was estimated using out-of-bag error from the bootstrap. To reduce multicollinearity, highly correlated predictor variables (r>0.80, n=9) were removed. Tuning the hyperparameters of the model did not improve its performance, so the original model was used. The R package spatialRF was used to minimize spatial autocorrelation of the residuals while fitting the spatial regression model. The resulting Random Forest decision tree model could explain 52.6% of the variance. The variable importance plot was created using random permutations for each predictor variable's values in out-of-bag data, then calculating the mean decrease in node impurity. Thirty model repetitions were used to create the plot of variable importance. Partial dependence plots were then generated and confirmed the non-linear relationship of most variables with BPC_{16S} scores (**Supplementary Figure 4**).

3. RESULTS

3.1. Global distribution of butyrate-producing bacteria

To characterize the global distribution of butyrate-producing bacteria, we analysed shotgun metagenomic datasets (*n*=16,176) from the IMG/M database(Chen et al., 2021). Samples originated from a broad range of sources, including soils and sediments, marine samples, human and animal faecal samples, and wastewater samples. Our novel BPC_{meta} index (see Methods) was established using the counts of two terminal genes in the butyrate production metabolic pathway, weighted by the mean count of each gene in bacterial genomes (full workflow shown in **Supplementary Figure 5**). The genes selected for analysis were *buk* (butyrate kinase) and *atoA (*acetate-CoA:acetoacetyl-CoA transferase subunit beta). Both enzymes catalyse the final steps in converting butyryl-CoA into butyrate, also referred to as butanoate (**Supplementary Table 1**). We grouped gene count data, BPC_{meta} scores, and sample metadata into six general source categories: soil, aquatic (including freshwater, brackish, and marine waters), non-human animal host-associated, human host-associated, plant-associated (e.g., root-associated, rhizosphere), and agro-industrial (e.g., anaerobic digesters, agricultural soil).

Metagenomes with genes for butyrate production were found on every continent, in every ocean, and in 89 countries (**Figure 1A**). Overall highest median BPC_{meta} scores were found in human host-associated (2.99, *n*=1 553) and non-human animal host-associated samples (2.91, *n*=771), with the lowest median BPC_{meta} scores in aquatic samples (1.93, *n*=6 017) (**Figure 1B**).

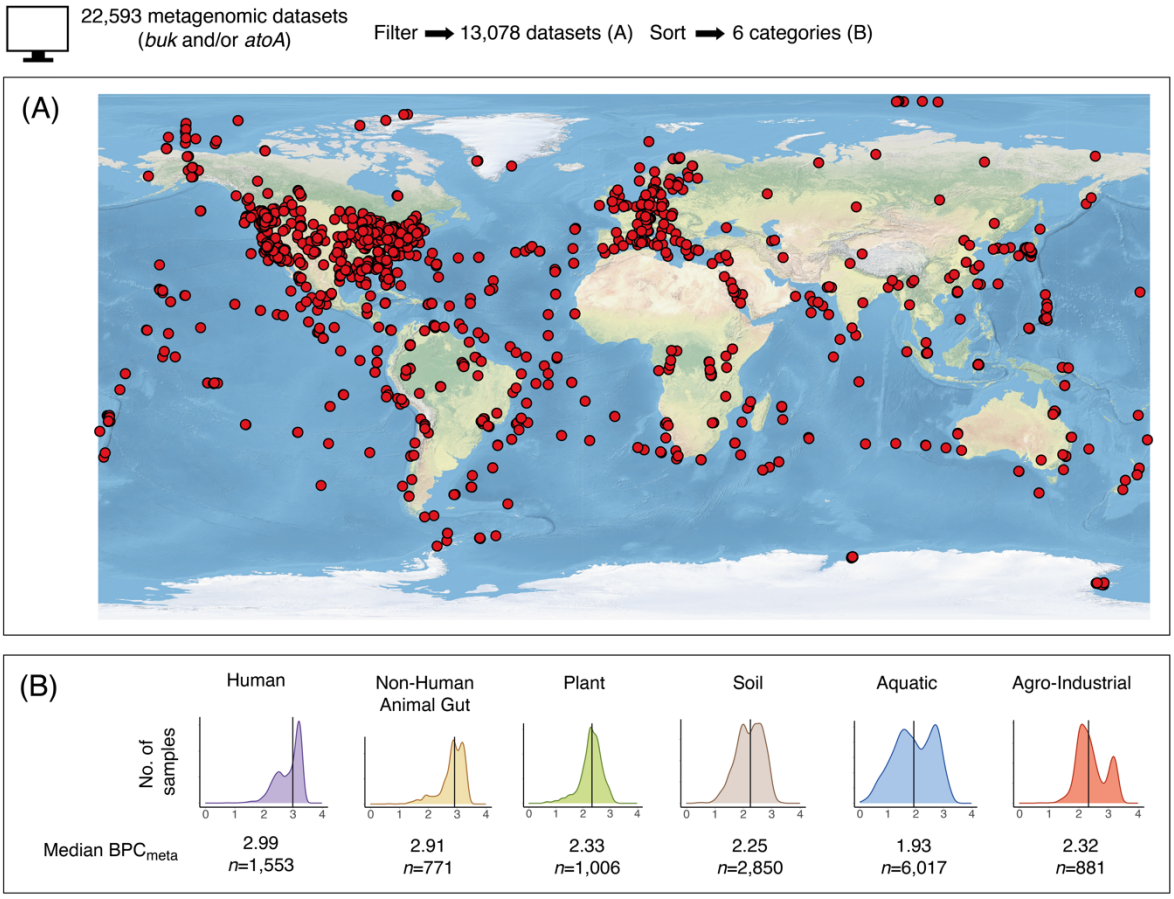

Fig.1: Butyrate-producing bacteria are found on every continent, in every ocean, and in 89 countries. (A) Map showing study locations of samples with *buk* and/or *atoA* genes. (B) Density plots showing frequency distributions of sample BPC_{meta} scores in the six highest-level groupings (x axis=BPC_{meta} score). BPC_{meta} score medians rather than means are presented due to non-normal BPC_{meta} score distributions. The range of sample BPC_{meta} scores was from 0.02 to 3.39. Bimodal peaks in five of the six categories may represent divergence between environments supportive and unsupportive of fermentative activity (discussed below). *n* is the number of samples.

3.2. Butyrate production capacity of different environments

To examine the global biogeographical distribution of butyrate producers more closely, we further subdivided each category into subcategories. Human samples were sorted into five body compartments: skin, nasal, oral, genital, and gut. The highest median BPC_{meta} score came from the gut (3.19, *n*=800), with faecal samples acting as a proxy for the anaerobic gut

environment. The lowest median BPC_{meta} score came from the skin (1.86, *n*=17), which is exposed to aerobic conditions. Between-group differences were statistically significant (H=1136, 4 d.f., *P*<0.001, Kruskal-Wallis test; **Figure 2A**).

Non-human animal host-associated samples included in our analysis (*n*=771) were either direct or proxy (e.g., faecal) measures of animal gut microbiota (*n*=448) or were non-gut but host-associated samples (e.g., attine ant fungus gardens, gutless marine worms, *n*=323). We first compared animal groupings by vertebrates (median BPC_{meta} score=3.11, *n*=389) and invertebrates (median BPC_{meta} score=2.76, *n*=382) (between-group differences statistically significant, W=22,592, *P*<0.001, Wilcoxon rank sum test). We then compared non-human animal samples by taxonomic phylum (between-group differences statistically significant, H=331, 4 d.f., *P*<0.001, Kruskal-Wallis test; **Figure 2B**), where Chordata had the highest median BPC_{meta} score (3.11, *n*=389) and Porifera (sponges), which lack a gut, had the lowest BPC_{meta} scores (1.87, *n*=34). A further comparison of the median BPC_{meta} scores of the primate gut (3.12) versus the human gut (3.19) corroborates the findings of a recent related study that showed a higher abundance of butyrate production pathway genes in humans versus most non-human primates(Mallott and Amato, 2022).

Our dataset included 1,006 plant-associated samples. These were subcategorized into four groups by plant compartment: leaf surface, plant litter, rhizosphere, and root. Root samples had the highest median BPC_{meta} score (2.50, *n*=123). Leaf surface samples, which are exposed to aerobic conditions, had the lowest median BPC_{meta} score (1.76, *n*=30). Between-group differences were statistically significant (H=105, 3 d.f., *P*<0.001, Kruskal-Wallis test; **Figure 2C**).

Soil samples (*n*=2,850) were sorted using the anthropogenic biome (anthrome) categories(Ellis et al., 2021; Gauthier et al., 2021), representing varying densities of human population and land use (anthrome classes and world map shown in **Supplementary Figure**

6). Here, we used the "Level" category of anthromes. The highest median BPC_{meta} scores came from both "Dense settlements" (includes classes "urban" and "mixed settlements"; median BPC_{meta} score=2.38, *n*=297) and "Cultured" (includes "woodlands" classes and the "inhabited drylands" class; median BPC_{meta} score=2.36, *n*=1076). The lowest median BPC_{meta} score (1.94, *n*=167) came from the anthrome level "Wildlands", which has the lowest human influence (between-group differences statistically significant, H=186, 5 d.f., *P*<0.001, Kruskal-Wallis test; **Figure 2D**)

Aquatic samples (*n*=6,017) were sub-grouped into five categories: marine, freshwater, brackish water and estuary, springs, and inland saltwater. The highest median BPC_{meta} score (2.52, *n*=911) was found in inland saltwater samples, and marine samples had the lowest median BPC_{meta} score (1.67, *n*=2047) (between-group differences statistically significant, H=530, 4 d.f., *P*<0.001, Kruskal-Wallis test; **Figure 2E**).

Agricultural and industrial ("agro-industrial") samples (*n*=881) were from a wide variety of sources and materials. We grouped them into seven source types, which include two sample types from wastewater treatment plants (i.e., activated sludge from aeration tanks and anaerobic digesters). The highest median BPC_{meta} scores (3.16, *n*=120) were from anaerobic digester samples. The lowest median BPC_{meta} scores were from the agricultural soils (2.16, *n*=486) and activated sludge (2.23, *n*=50) (between-group differences statistically significant, H=431, 6 d.f., *P*<0.001, Kruskal-Wallis test; **Figure 2F**). Activated sludge is a bacteria-rich product formed in aeration tanks with aerobic conditions.

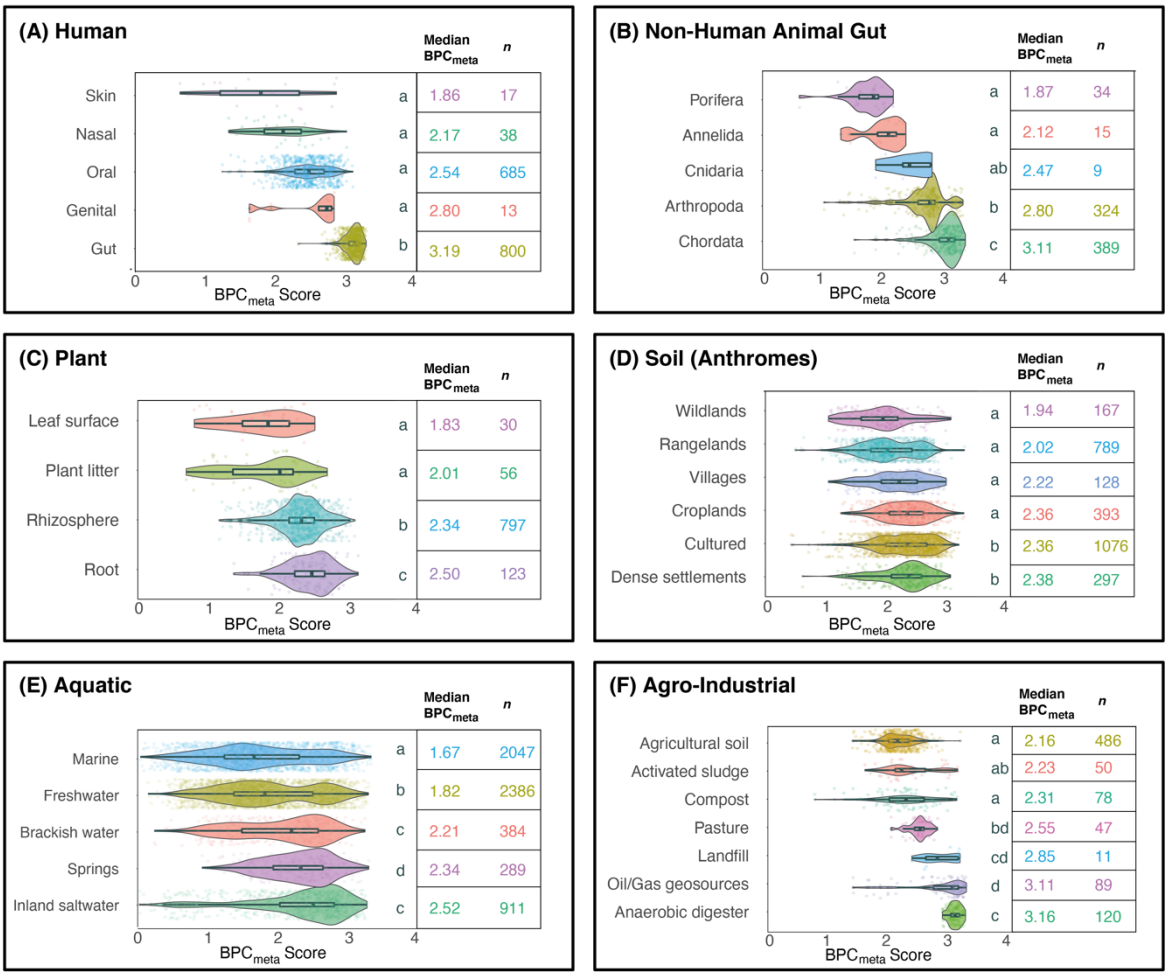

Fig.2: BPC_{meta} scores vary between host communities and environmental sources. BPC_{meta} score density plots by group subcategories. (A) BPC_{meta} scores of humans, sorted by body compartment. (B) BPC_{meta} scores of non-human animal-associated microbial communities, sorted by class. Note that Porifera do not possess a gut. (C) BPC_{meta} scores of plant-associated samples, grouped into compartments. (D) BPC_{meta} scores of soil samples, grouped into anthropogenic biomes (anthromes) levels that represent human influence on land use. The level "Cultured" includes woodlands and inhabited drylands. (E) BPC_{meta} scores of aquatic ecosystem samples, grouped into source site categories. (F) BPC_{meta} scores of agricultural and industrial samples, grouped by source site. Activated sludge and anaerobic digesters are common components of wastewater treatment plants. In each of (A)-(F), Kruskal-Wallis tests show that between-group differences were significant at $P<0.001$. Medians sharing a letter are not significantly different by the adjusted Dunn test at the 5% level of significance. Boxes show the interquartile range.

3.3. Butyrate production capacity of soils

Butyrate production involves fermentation and is characteristically an anaerobic activity. In land types where soils undergo lengthy periods of desiccation (e.g., desert soils) or have regular disruption from human activity (e.g., agricultural soils), butyrate production would be expected to decrease. By contrast, in soils from land types with longer annual periods of waterlogging or inundation, with sufficient organic matter and where human disruption of soils is minimal, we would expect more anaerobic conditions and therefore also butyrate production. Landscapes which combine cooler and wetter climate patterns, appreciable levels of primary production, flat terrain, and typically acidic soil pH have been associated with the conditions that support organic matter degradation via fermentation(Pemberton, 2005). Thus, land type and soil oxygenation appear to be drivers for butyrate production activity and selection for fermentative bacterial taxa.

We compared the median BPC_{meta} scores across soil metagenome projects that used the GOLD Ecosystem Classification path(Ivanova, 2010; Mukherjee et al., 2021) (**Figure 3**). We show that bogs had the highest median BPC_{meta} score (2.75). Bogs and their associated peat have submerged layers of decaying plant matter. Anoxic conditions and depletion of inorganic oxidants delay the full degradation of organic matter(Conrad, 2020), sometimes for thousands of years. Thus, the microbial content of bogs includes fermenters and methanogenic archaea.

Orchard soil and deserts had the lowest median BPC_{meta} scores (1.48 and 1.77, respectively). Hot, dry temperatures and regular soil turnover should not favour anaerobic butyrate production. However, desert soil crusts (comprising a resilient biofilm and associated microbiota(Cania et al., 2020), discussed below) and the propensity of Bacillota (formerly Firmicutes) to form endospores(Browne et al., 2016) may maintain dormant butyrate production potential until wetter conditions arrive.

With our focus on human exposure to commensal butyrate-producing bacteria, accessibility of the land type to human visitation is central to assessing the feasibility of exposure. Bogs have a high median BPC_{meta} score, but their anaerobic microbial activity occurs primarily beneath the waterlogged surface; therefore, direct human exposure to their butyrate producers would be challenging. Our results suggest that agricultural land and forest soils may be more reasonable for human exposure due to moderate butyrate production capacity and higher human accessibility.

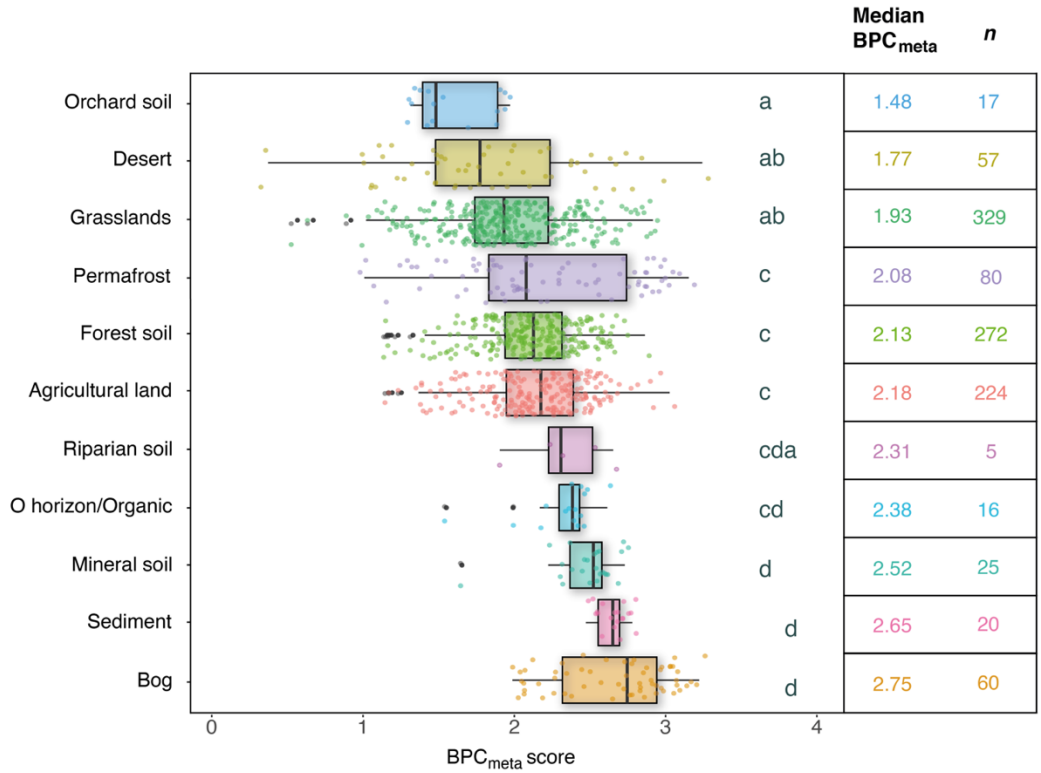

Fig.3: Soil ecosystem data show that land types with persistent anaerobic conditions have high BPC_{meta} scores. Median BPC_{meta} score boxplots of soil ecosystem categories. Between-group differences were statistically significant, H=233, 10 d.f., $P<0.001$, Kruskal-Wallis test. Medians sharing a letter are not significantly different by the adjusted Dunn test at the 5% level of significance. Boxes show the interquartile range.

3.4. Environmental characteristics associate with BPC scores

To examine regional-scale variations in environmental influences on BPC scores, we analysed 1,285 surface soil samples from across Australia for associations between environmental metadata and BPC scores derived from bacterial 16S rRNA amplicon data (BPC_{16S}). 16S rRNA amplicons are commonly used to quantify bacterial taxonomic abundances, and here we examined an extensive dataset collected using consistent protocols across the continent of Australia(Bissett et al., 2016). For associated environmental metadata, covariate data with 49 variables (**Supplementary Table 7**) were downloaded and analysed. All continuous predictor variables (n=43) were analysed using principal components analysis to discern relationships between the variables. The environmental origins of samples were visualized by plotting coordinates of the top two principal components and grouping into five land type clusters via k-means clustering. This process revealed a mapping of samples to distinct land types that were distributed across Australia (**Figure 4A**) and corresponded with differences in the predictor variables across the clusters (**Figure 4B**). The cluster plot (**Figure 4C**) showed clear separation of land type clusters with dominant influences of moisture and fertility, consistent with the map display and later modelling of key drivers of BPC_{16S} scores.

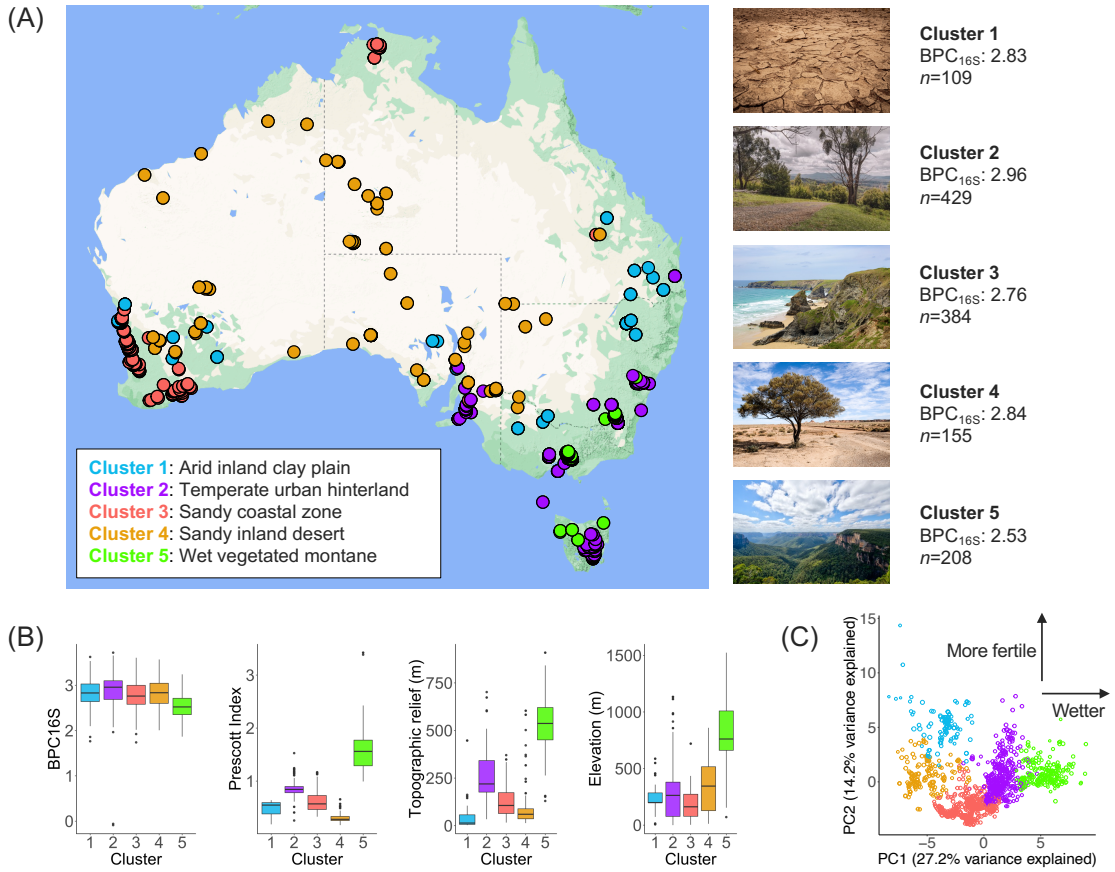

Fig.4: Clustering of environmental data shows five distinct land types. Analysis of environmental variables associated with Australian soil samples and their BPC_{16S} scores. (A) Map of Australian soil samples clustered on 43 continuous environmental variables, five cluster distribution, mapped using R package ggmap and Google maps. Photographs were downloaded from Unsplash.com under CC0 license. (B) Boxplots for BPC_{16S} scores and the top 3 variables from (D) across each of the 5 clusters. Between-cluster BPC_{16S} score differences were statistically significant (H=170, 4 d.f., P<0.001, Kruskal-Wallis test). Boxes show the interquartile range. (C) First two principal components coloured by k-means clusters. The x-axis can be broadly interpreted as environmental wetness and associated variables (e.g., vegetation cover). The y-axis can be broadly interpreted as soil fertility and the presence of cations.

BPC_{16S} scores varied significantly between the clusters (H=170, 4 d.f., *P*<0.001, Kruskal-Wallis test). The highest median BPC_{16S} scores came from the temperate urban hinterland

cluster (2.96, *n*=429) and the sandy desert cluster (2.84, *n*=155). The lowest median BPC_{16S} score came from the wet vegetated montane cluster (2.53, *n*=208).

The cluster with the highest median BPC_{16S} score was the temperate urban hinterland soil and is generally moderate in elevation, annual rainfall, topographic relief, clay, and soil fertility, and has high levels of zinc and manganese (**Supplementary Table 8**). A separate analysis of categorical variables also showed the highest BPC_{16S} scores among urban land cover types ("built-up") and land use types ("rural residential"), further reflecting the potential association with human population density (**Supplementary Table 9**). The cluster with the second-highest median BPC_{16S} score was, intriguingly, from sandy inland deserts, with moderate elevation, dry climate, low soil nutrient content, low soil organic carbon content, minimal vegetation cover, and higher annual mean temperature. The cluster with the lowest median BPC_{16S} score, wet vegetated montane, had high elevation and topographic relief, cold mean annual temperature, high annual rainfall and aridity index, consistent rainfall levels throughout the year, high soil organic carbon content, and high soil iron and aluminium content. The two additional clusters, arid inland clay plains and sandy coastal zones, also had distinct characteristics (**Supplementary Table 8**). Because our 16S rRNA data came only from Australia, our modelling may not be generalisable to global conditions that exceed the ranges of our environmental covariate data. For example, the height of mountains in Australia does not exceed 2,745 metres; thus, our mountain cluster modelling may not fit other countries with higher mountains.

Random Forest decision tree analysis(Breiman, 2001) was then used to better understand how the continuous predictor variables influenced BPC_{16S} scores. The resulting variable importance plot (**Figure 5)** showed that the five top predictor variables were Prescott Index, topographic relief, elevation, longitude, and latitude, suggesting climate and geography are

the strongest overall environmental drivers of the abundance of butyrate-producing bacteria in soils. Partial dependence plots for each variable are shown in **Supplementary Figure 4**.

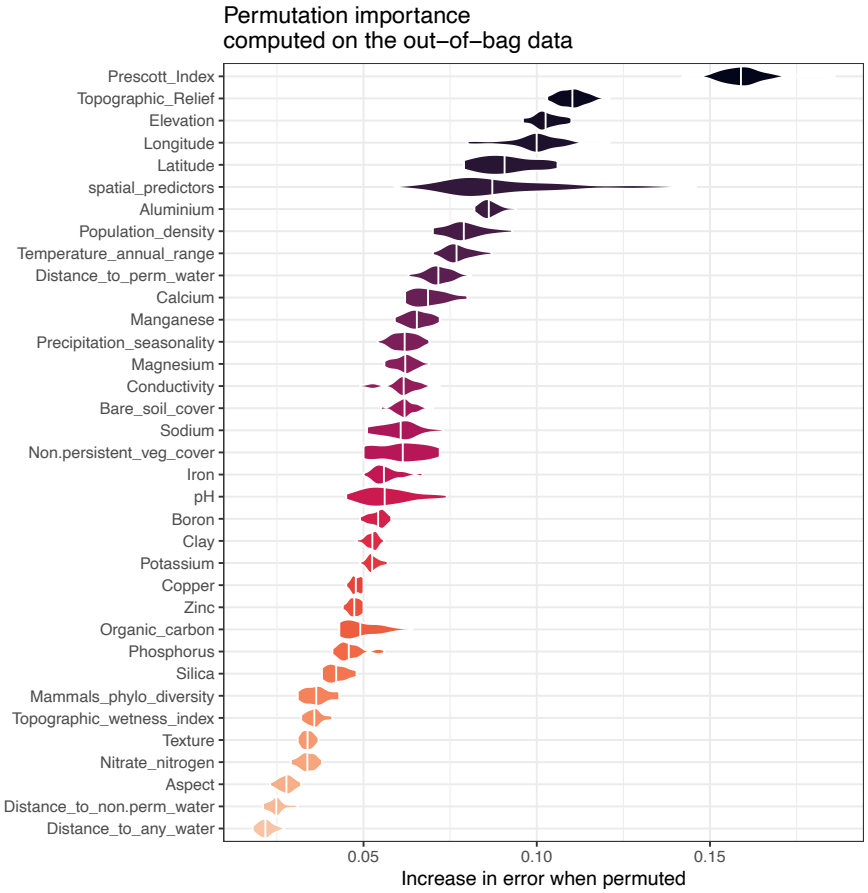

Fig. 5: Variable importance results from Random Forest decision tree modelling. Random Forest variable importance results from 43 continuous environmental predictor variables. The model was fitted using out-of-bag errors from the bootstrap. The variable importance was determined using random permutations of predictor variables and the mean decrease in node impurity.

4. DISCUSSION

Butyrate-producing bacteria have critical roles in human health. Their assemblage in the human gut receives extensive attention, but the original environmental sources of these bacteria remain poorly understood. Here, we characterized the presence of butyrate-producing bacteria in outdoor environments worldwide. We compared outdoor samples with

437 human and non-human animal gut samples to provide insights into the potential for these
438 bacteria to transfer to people spending time outdoors. Our novel BPC score formulae from
439 both 16S rRNA and metagenomic bacterial data were validated through several lines of
440 evidence. We identified specific geographical trends that advance the characterization of
441 butyrate-producing bacteria in outdoor locations. Of the many patterns revealed in our results,
442 we highlight two of them here: anaerobic conditions and human presence.

443 First, BPC_{meta} scores were, as expected, higher in anaerobic environments, which agrees with
444 previous work showing that butyrate-producing bacteria thrive in anaerobic
445 environments(Conrad, 2020; Riviere et al., 2016). Our analysis of plant and animal data
446 shows that compartments exposed to air, such as leaf surfaces and human skin, had the lowest
447 median BPC_{meta} scores of their respective group. Similarly, wastewater treatment plant
448 samples showed that activated sludge, a product formed from aeration tanks, had among the
449 lowest BPC_{meta} scores of its category. Yet, samples from anaerobic digesters and chordate gut
450 samples, which each employ anaerobic processes, had the highest BPC_{meta} scores. These
451 findings are consistent with the literature and, including evidence of the specificity of the
452 BPC_{meta} formula toward butyrate production (**Supplementary Figure 1**), support the validity
453 of our BPC_{meta} formula.

454 Second, our results show an association between human presence and BPC scores. Soils from
455 the anthrome level 'dense settlements' showed the highest BPC_{meta} scores within its group. In
456 contrast, soils from the anthrome level 'wildlands' had a low median BPC score. These
457 results suggest that the presence of humans may in fact contribute to the BPC score, possibly
458 from gut-associated bacteria being inadvertently dispersed into the environment. Our 16S
459 rRNA data from Australian soils also support this connection between human presence and
460 BPC scores. Australia's major cities and hinterlands are coastal, often with river-floodplain
461 systems and areas of fertile soils that were attractive to the European settlers. Thus, evidence

suggests an association between human presence, soil fertility, and high BPC scores, but the direction of influence raises a compelling question: Are the higher soil BPC scores in urban areas due to the presence of humans (and perhaps also their chordate pets), whose digestive products are introduced into the environment, or are humans drawn to live in areas with naturally high soil fertility and high BPC scores due to their high capacity for primary production? These questions may be of future research interest.

Intriguingly, inland sandy deserts had relatively high BPC scores. This finding does not follow the pattern of higher BPC scores in more fertile soils, temperate areas, and anaerobic conditions. Our data were unable to identify a specific reason for this finding; however, it should be noted that desert microbiota often form biological soil crusts (biocrusts), which are densely packed microbial structures that are desiccation-resistant and include photosynthesizers such as cyanobacteria(Garcia-Pichel et al., 2001). Therefore, it may be possible that butyrate production potential is conserved in these bacteria and biocrusts but remains dormant until more favourable environmental conditions prevail following rainfall. Upon wetting, soil biocrust microbial activity rapidly accelerates, and growing biomass can create anoxic microniches that could favour fermentative processes such as butyrate production(Angel et al., 2011).

During the development of our methods, several limitations of our study became apparent. Analyses of shotgun metagenomic sequences and bacterial 16S rRNA amplicons rely on reference databases that are continually being developed but are incomplete. Missed or incomplete sequence identification could affect the reliability of our formulae. Likewise, taxonomy databases are regularly updated due to new information, but their updates are not uniform across databases. We used the GTDB database to classify our list of butyrate-producing bacteria, but it showed occasional discrepancies with the classification system on which the downloaded Australian soils 16S rRNA data are based. Thus, the utilisation of

multiple taxonomic classification systems likely means that some butyrate producers were not identified in our data. This could affect the reliability of our BPC formulae.

To maximize the precision of our butyrate producer database, we chose to use species-level classifications via GTDB representative species. This may have inadvertently created inconsistent data from species with multiple strains (sometimes hundreds of strains are present), among which some may be butyrate producers and others not. In addition, some strains may display pathogenicity. Thus, analysis at the strain level could provide a higher resolution of data, which could be a future research opportunity. Finally, our data is dependent on the capacity of laboratory DNA extraction methods to open endospores. Because butyrate-producers tend to thrive in anaerobic environments, they often form endospores when exposed to air, protecting them until they can germinate in a new anaerobic environment. Thus, sampling methods that expose the samples to air may inadvertently cause sporulation of bacteria. Such methods may subsequently reduce the quantities of DNA extracted from spore-formers, a number of which may be butyrate-producing bacteria(Browne et al., 2016). Consistency across sampling and DNA extraction methods among future studies could help improve butyrate-producing bacterial abundance data reliability.

Time spent in natural and biodiverse settings is known to offer human health benefits(Kondo et al., 2018; Lai et al., 2019). The transfer of health-beneficial microbes to people spending time in green spaces could be a key mechanism of such health benefits. Urban green space designers rely on evidence of these health benefits to identify particular green space attributes that could be utilized in urban design, such as the abundance of health-beneficial butyrate-producing bacteria in soils and plants. Because half of the world's population now lives in cities(Rydin et al., 2012), policy makers and urban green space designers have a critical need for research to guide the development of green infrastructure(Robinson et al., 2021) that

supports the health of its residents. Our study helps advance such research. We applied our methods on a broad biogeographical scale. Future assessment of butyrate-production capacity across finer metropolitan-level scales will provide greater precision for city infrastructure planning and further microbiome-based public health research.

Acknowledgements

We are pleased to acknowledge that this work is supported in part through a grant from Flinders Foundation.

Data availability

The datasets generated during and/or analysed in the current study are available in Supplementary information, and all datasets and custom R code are available on figshare at https://figshare.com/s/18ebb617daee5935a870.

CRediT author statement

Joel Brame: Conceptualization, Methodology, Formal analysis, Project Administration, Writing – Original draft preparation, Visualization, Funding acquisition. **Craig Liddicoat:** Conceptualization, Software, Methodology, Writing – Reviewing and editing. **Catherine Abbott:** Visualization, Writing – Reviewing and editing. **Robert Edwards:** Conceptualization, Methodology, Resources, Writing – Reviewing and editing. **Jake Robinson:** Visualization, Writing – Reviewing and editing. **Nicolas Gauthier:** Formal analysis, Methodology, Writing – Reviewing and editing. **Martin Breed:** Conceptualization, Methodology, Project Administration, Visualization, Writing – Reviewing and editing, Funding acquisition.

References

Alfven T, Braun-Fahrlander C, Brunekreef B, von Mutius E, Riedler J, Scheynius A. Allergic diseases and atopic sensitisation in children related to farming and anthroposophic lifestyle - the PARSIFAL study. Allergy 2007; 62: 455-455.

Angel R, Matthies D, Conrad R. Activation of methanogenesis in arid biological soil crusts despite the presence of oxygen. PloS one 2011; 6: e20453.

Baxter NT, Schmidt AW, Venkataraman A, Kim KS, Waldron C, Schmidt TM. Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. mBio 2019; 10: 13.

Belbin L. The Atlas of Living Australia's spatial portal. In: Jones M, B. & Gries, C., editor. Proceedings of the Environmental Information Management Conference (EIM 2011), Santa Barbara, 2011, pp. 39-43.

Bissett A, Fitzgerald A, Meintjes T, Mele PM, Reith F, Dennis PG, et al. Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. Gigascience 2016; 5: s13742-016-0126-5.

Brame JE, Liddicoat C, Abbott CA, Breed MF. The potential of outdoor environments to supply beneficial butyrate-producing bacteria to humans. Science of The Total Environment 2021; 777: 146063.

Breiman L. Random forests. Machine learning 2001; 45: 5-32.

Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable'human microbiota reveals novel taxa and extensive sporulation. Nature 2016; 533: 543-546.

Cania B, Vestergaard G, Kublik S, Köhne JM, Fischer T, Albert A, et al. Biological soil crusts from different soil substrates harbor distinct bacterial groups with the potential to produce exopolysaccharides and lipopolysaccharides. Microbial ecology 2020; 79: 326-341.

Cantu-Jungles TM, Rasmussen HE, Hamaker BR. Potential of Prebiotic Butyrogenic Fibers in Parkinson's Disease. Frontiers in Neurology 2019; 10: 8.

Chen D, Jin D, Huang S, Wu J, Xu M, Liu T, et al. Clostridium butyricum, a butyrate-producing probiotic, inhibits intestinal tumor development through modulating Wnt signaling and gut microbiota. Cancer letters 2020; 469: 456-467.

Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The IMG/M data management and analysis system v. 6.0: new tools and advanced capabilities. Nucleic acids research 2021; 49: D751-D763.

Conrad R. Methane production in soil environments—anaerobic biogeochemistry and microbial life between flooding and desiccation. Microorganisms 2020; 8: 881.

Cudahy T, Caccetta M, Thomas M. National ASTER Map of Australia. Scale 2012; 1: 1000000.

Ellis EC, Gauthier N, Goldewijk KK, Bird RB, Boivin N, Díaz S, et al. People have shaped most of terrestrial nature for at least 12,000 years. Proceedings of the National Academy of Sciences 2021; 118.

Flies EJ, Mavoa S, Zosky GR, Mantzioris E, Williams C, Eri R, et al. Urban-associated diseases: Candidate diseases, environmental risk factors, and a path forward. Environment international 2019; 133: 105187.

Frei R, Heye K, Roduit C. Environmental influences on childhood allergies and asthma—The Farm effect. Pediatric Allergy and Immunology 2022; 33: e13807.

Fu X, Ou Z, Zhang M, Meng Y, Li Y, Wen J, et al. Indoor bacterial, fungal and viral species and functional genes in urban and rural schools in Shanxi Province, China–association

with asthma, rhinitis and rhinoconjunctivitis in high school students. Microbiome 2021; 9: 1-16.

Garcia-Pichel F, López-Cortés A, Nubel U. Phylogenetic and morphological diversity of cyanobacteria in soil desert crusts from the Colorado Plateau. Applied and environmental microbiology 2001; 67: 1902-1910.

Gauthier N, Ellis E, Goldewijk KK. Anthromes 12K DGG (V1) Full Dataset. Harvard Dataverse 2021; 10.

Goldfarb KC, Karaoz U, Hanson CA, Santee CA, Bradford MA, Treseder KK, et al. Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. Frontiers in microbiology 2011; 2: 94.

Ivanova N. A call for standardized classification of metagenome projects. 2010.

Kahle DJ, Wickham H. ggmap: spatial visualization with ggplot2. R J. 2013; 5: 144.

Kondo MC, Fluehr JM, McKeon T, Branas CC. Urban green space and its impact on human health. International journal of environmental research and public health 2018; 15: 445.

Lai H, Flies EJ, Weinstein P, Woodward A. The impact of green space and biodiversity on health. Frontiers in Ecology and the Environment 2019; 17: 383-390.

Liddicoat C, Sydnor H, Cando-Dumancela C, Dresken R, Liu JJ, Gellie NJC, et al. Naturally-diverse airborne environmental microbial exposures modulate the gut microbiome and may provide anxiolytic benefits in mice. Science of the Total Environment 2020; 701: 11.

Liu H, Wang J, He T, Becker S, Zhang G, Li D, et al. Butyrate: a double-edged sword for health? Advances in Nutrition 2018; 9: 21-29.

Mallott EK, Amato KR. Butyrate Production Pathway Abundances Are Similar in Human and Nonhuman Primate Gut Microbiomes. Molecular biology and evolution 2022; 39: msab279.

McBratney AB, Santos MM, Minasny B. On digital soil mapping. Geoderma 2003; 117: 3-52.

McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one 2013; 8: e61217.

Miyake S, Kim S, Suda W, Oshima K, Nakamura M, Matsuoka T, et al. Dysbiosis in the gut microbiota of patients with multiple sclerosis, with a striking depletion of species belonging to clostridia XIVa and IV clusters. PLOS One 2015; 10: e0137429.

Mohammadkhah AI, Simpson EB, Patterson SG, Ferguson JF. Development of the gut microbiome in children, and lifetime implications for obesity and cardiometabolic disease. Children 2018; 5: 160.

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, et al. Genomes OnLine Database (GOLD) v. 8: overview and updates. Nucleic acids research 2021; 49: D723-D733.

Nurminen N, Lin J, Grönroos M, Puhakka R, Kramna L, Vari HK, et al. Nature-derived microbiota exposure as a novel immunomodulatory approach. Future Microbiology 2018; 13: 737-744.

O'Brien L. slga: Data Access Tools for the Soil and Landscape Grid of Australia, 2021, pp. R package.

Ottman N, Ruokolainen L, Suomalainen A, Sinkko H, Karisola P, Lehtimäki J, et al. Soil exposure modifies the gut microbiota and supports immune tolerance in a mouse model. Journal of allergy and clinical immunology 2019; 143: 1198-1206. e12.

Parada Venegas D, De la Fuente MK, Landskron G, González MJ, Quera R, Dijkstra G, et al. Short chain fatty acids (SCFAs)-mediated gut epithelial and immune regulation and

its relevance for inflammatory bowel diseases. Frontiers in Immunology 2019; 10: 277.

Parajuli A, Grönroos M, Siter N, Puhakka R, Vari HK, Roslund MI, et al. Urbanization reduces transfer of diverse environmental microbiota indoors. Frontiers in Microbiology 2018; 9: 84.

Pemberton M. Australian peatlands: a brief consideration of their origin, distribution, natural values and threats. Journal of the Royal Society of Western Australia 2005; 88: 81.

Poret-Peterson AT, Albu S, McClean AE, Kluepfel DA. Shifts in soil bacterial communities as a function of carbon source used during anaerobic soil disinfestation. Frontiers in Environmental Science 2019; 6: 160.

Riviere A, Selak M, Lantin D, Leroy F, De Vuyst L. Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut. Frontiers in Microbiology 2016; 7: 21.

Rivière A, Selak M, Lantin D, Leroy F, De Vuyst L. Bifidobacteria and butyrate-producing colon bacteria: importance and strategies for their stimulation in the human gut. Frontiers in Microbiology 2016; 7: 979.

Robinson JM, Watkins H, Man I, Liddicoat C, Cameron R, Parker B, et al. Microbiome-Inspired Green Infrastructure: a bioscience roadmap for urban ecosystem health. arq: Architectural Research Quarterly 2021; 25: 292-303.

Roduit C, Frei R, Ferstl R, Loeliger S, Westermann P, Rhyner C, et al. High levels of butyrate and propionate in early life are associated with protection against atopy. Allergy 2019; 74: 799-809.

Rook GA, Lowry CA, Raison CL. Microbial 'Old Friends', immunoregulation and stress resilience. Evolution, Medicine, and Public Health 2013; 2013: 46-64.

Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature 2018; 555: 210-215.

Rydin Y, Bleahu A, Davies M, Dávila JD, Friel S, De Grandis G, et al. Shaping cities for health: complexity and the planning of urban environments in the 21st century. The lancet 2012; 379: 2079-2108.

Selway CA, Mills JG, Weinstein P, Skelly C, Yadav S, Lowe A, et al. Transfer of environmental microbes to the skin and respiratory tract of humans after urban green space exposure. Environment International 2020; 145: 106084.

Sonnenburg ED, Sonnenburg JL. The ancestral and industrialized gut microbiota and implications for human health. Nature Reviews Microbiology 2019; 17: 383-390.

Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021.

Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nature Microbiology 2019; 4: 623-632.

Vital M, Howe AC, Tiedje JM. Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta) genomic Data. mBio 2014; 5: 11.

Wickham H. Ggplot2: Elegant graphics for data analysis (2nd ed). Springer-Verlag. Springer-Verlag, New York, 2016.

Williams K, Stein J, Storey R, Ferrier S, Austin M, Smyth A, et al. 0.01 degree stack of climate layers for continental analysis of biodiversity pattern, version 1.0. v2. CSIRO. Data Collection.