Towards the biogeography of butyrate-producing bacteria

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

Author email addresses:

Joel E. Brame: joel.brame@flinders.edu.au

Craig Liddicoat: craig.liddicoat@flinders.edu.au

Catherine A. Abbott: cathy.abbott@flinders.edu.au

Robert A. Edwards: robert.edwards@flinders.edu.au

Jake M. Robinson: jake.robinson@flinders.edu.au

Nicolas E. Gauthier: nicolas.gauthier@ufl.edu

Martin F. Breed: martin.breed@flinders.edu.au

Corresponding author: Joel Brame, Flinders University, College of Science and Engineering, Sturt Rd, Bedford Park, SA 5042, Telephone: +61 (0)8 8201 2113, Email: joel.brame@flinders.edu.au

Statement of Authorship

JB, CL, RE, and MB conceived and designed the study. JB performed data analysis, with input from CA, RE, NG, CL, and MB. JB and JR designed and developed the figures. JB wrote the manuscript, with input from all authors.

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/3ae2f1327a11f91793d8 .

Words in abstract: 147

Words in main text: 4847

Number of references: 47

Number of figures, tables, and text boxes: 5

Keywords: butanoate, ecosystem services, gut microbiome, metagenomics, microbial ecology, soil microbiota

ABSTRACT

Butyrate-producing bacteria are found in many ecosystems and organisms, including humans, and are vital to ecosystem functionality and human health. These bacteria ferment organic matter, producing the short-chain fatty acid butyrate. Despite their importance, their global biogeography is poorly studied. Here we developed new normalised butyrate production capacity (BPC) indices derived from global metagenomic (*n*=16,176) and Australia-wide soil 16S rRNA (*n*=1,331) data to geographically detail their environmental associations. We show that the highest BPC scores were in anoxic and fermentative environments. Among land types, higher BPC scores were in soils from seasonally productive sandy zones, temperate rural residential areas, and pastures and croplands. Geographical and climatic variables, and the presence of humans, were the primary drivers of BPC score variation. These new biogeographical insights could be integrated into microbially-focussed health and environmental policies to better harness the contributions of butyrate-producing bacteria to ecosystem functionality and human health.

INTRODUCTION

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 and non-human animals, gut-associated butyrate provides energy to colonic epithelial cells, maintains gut barrier stability, has anti-inflammatory and immunomodulatory effects, and strengthens the integrity of the blood-brain barrier (Rivière *et al.* 2016; Bedford & Gong 2018; Brame *et al.* 2021; Knox *et al.* 2022). As such, butyrate production by bacteria has become a vital part of host-microbe symbioses.

Butyrate-producing bacteria have been found in a wide variety of environments. Butyrate producer abundances in the human gut are well-studied, partly due to butyrate's important functional roles in humans (Stilling *et al.* 2016; Nilsen *et al.* 2020). Many non-human animals also retain these gut-associated bacteria, where they perform similar breakdown services for dietary fibre intake (Levine *et al.* 2013; Mallott & Amato 2022). Among studies from built settings, Ferraris et al. (2012) characterised premature human neonates' gut butyrate producer assemblages originating from the hospital environment, particularly species within the Clostridiaceae family. Anaerobic digesters and wastewater treatment plants regularly utilise substrates with high abundances of the butyrate producer-rich phylum Firmicutes (Yang *et al.* 2014). Among outdoor environment studies, Liu et al (2011) examined butyrate within rice paddy soils, although their focus was on butyrate oxidation pathways rather than butyrate synthesis. Indeed, butyrate producers have been found across

diverse environmental samples; however, to our knowledge, there has not been a global systematic examination of their distribution in outdoor environments.

Soils represent a key reservoir for microbial diversity in the environment (Fierer & Jackson 2006) and provide a focus for us to examine conditions that may support butyrate producers. Many native soil-associated butyrate-producing taxa are not typically associated with the human gut (Vital *et al.* 2014); instead, they may be involved in soil-plant-organic matter cycling processes and be ingested and expelled from the digestive tract of soil fauna (e.g., earthworms; Thakuria *et al.* 2010). At the same time, non-human animal gut-associated bacteria are regularly dispersed into the outdoor environment, where they can settle into soils (Blum *et al.* 2019). Such transmission is enabled by the ability of butyrate producers to either temporarily tolerate micro-aerobic niches (Miquel *et al.* 2013) or, more commonly, form a protective endospore to survive air exposure and then regerminate once inside a suitable anaerobic location (e.g., the gut after ingestion; Browne *et al.* 2017). Thus, soils are potentially an assemblage of both soil-associated and gut-associated butyrate-producing bacteria. Likewise, bacteria in the superficial layer of soils can be dispersed into the air (Polymenakou 2012) and transferred to hosts (Liddicoat *et al.* 2020). Human ingestion of soil-borne plants such as root vegetables, or even soil itself, can also directly transmit bacteria from soil to the host gut (Lang *et al.* 2014). As such, more detailed insight into the environmental drivers of soil butyrate producer abundances and distributions would substantially improve our epidemiological and ecological knowledge of these bacteria, including how they may be transmitted between species and the optimal conditions for their persistence.

Here, we combined global metagenomic datasets with continent-wide Australian 16S rRNA amplicon data to provide insight into the global biogeography of butyrate-producing bacteria. We developed two new normalised indices to estimate and compare the butyrate production capacity (BPC) of representative samples from both metagenomic (BPC_{meta}) and 16S rRNA amplicon (BPC_{16S}) data. Specifically, we aimed to: (a) determine the butyrate production capacities across a range of host and environmental sources using our novel BPC_{meta} scores; (b) provide a more detailed comparison of BPC_{meta} scores across subcategories within each broader source group; and (c) conduct an in-depth analysis of Australian soil samples using BPC_{16S} scores to describe environmental drivers of soil butyrate production capacity. By developing novel targeted formulae and collating multiple types of data and analyses, we were able to synthesise a novel and in-depth view into the global biogeography of butyrate-producing bacteria and the interlinkages between their hosts and environments.

MATERIALS AND METHODS

Method overview

To compare the bacterial butyrate production capacity of samples across a broad spectrum of host and environmental sources, we chose to analyse samples of two types of bacterial sequencing data: shotgun metagenomic sequences and 16S rRNA amplicons. Each data type required interrogation of separate databases and the usage of separate statistical tools for analysis and visualisation. We developed novel formulae for each data type to estimate the butyrate production capacity (BPC) for each sample (detailed below). Each formula included a normalisation step to allow BPC scores between samples to be directly compared (**Table 1**).

Global shotgun metagenomic sample analysis

Gene selection and metagenome database interrogation

To characterise the global distribution of butyrate-producing bacteria, we analysed shotgun metagenomic datasets. To begin, 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 that coded 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 excluded (all gene decisions explained in **Supplementary Table 1**). The included genes *atoA* and *buk* participate in each of the two terminal pathways. The Integrated Microbial Genomes & Microbiomes (IMG/M) 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 yielded low numbers of samples 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 normalised 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 normalise estimates of total genomes in the samples. However, 115 samples did not include *pheS* count data and were subsequently removed from our analysis. Outliers with a pheS count <100 (*n*=5,479) and >50,000 (*n*=19) were also removed from our 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 our analysis (*n*=804). The remaining 16,176 samples were retained and analysed for our project. The samples originated from a broad range of sources, including soils and sediments, marine samples, human and non-human animal faecal samples, and wastewater samples.

BPC scores for shotgun 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, non-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 further analysis: soil samples were grouped by anthrome classification; aquatic samples were grouped by source subcategory; human samples were grouped by body compartment; non-human animal samples were grouped by vertebrate/invertebrate and by phylum; plants were grouped by compartment; agro-industrial samples were grouped by source site. To reduce bias within the anthrome "Class" categories, two studies whose samples accounted for >50% of the total class samples were removed from the analysis.

To validate our BPC_{meta} formula, we adapted it to estimate ethanol production (i.e., to investigate whether our BPC_{meta} formula was actually identifying butyrate production rather than just anaerobicity). This ethanol production pathway 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. A comparison of

the resulting EPC scores of the soil metagenomic samples with their BPC_{meta} scores showed a negligible correlation (**Supplementary Figure 1**).

Statistical tests were done in R (version 4.0.2; R Core Team 2021). The Shapiro-Wilk test was used to determine the normality of distribution. In each case where data did not fit a normal distribution either the non-parametric Kruskal-Wallis test or Wilcoxon rank-sum test was used to test for 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 3.3.5; Wickham 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.

Australian 16S rRNA amplicon sample analysis

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). We downloaded the zero-radius operational taxonomic units (zOTU) abundances and metadata for each resulting sample (*n*=3,023). We used the phyloseq package (McMurdie & Holmes 2013) for managing and cleaning the 16S rRNA data. We removed all “chloroplast” and “mitochondria” data. We removed low abundance zOTUs that did not occur in at least two samples and had total counts of <20, as these 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 resulting

sample count was *n*=2,795. To normalise the data and reduce bias in comparing samples, we transformed the zOTU abundances into relative abundances.

Because 16S rRNA data can have relatively poor resolution at the species level (Jovel *et al.* 2016), we focussed our BPC_{16S} derivation on family-level data. Using the Genome Taxonomy Database (GTDB) website interface and the set of putative butyrate-producing species (*n*=118) from Vital et al. (2014), we collated the families that included members in this list of butyrate-producers (*n*=54, **Supplementary Table 5**). This family list was matched with the Australian Microbiome Initiative taxonomy listings for each downloaded sample. Of the 54 taxonomic families with butyrate-producing bacteria analysed, 31 families had no representative (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 =

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

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

$$\cdots+$$

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

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. # butyrate-producing species (and total binomial species) in each family are from the GTDB.

Modelling environmental associations of BPC_{16S} scores

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

To associate environmental data with BPC scores, we first removed any non-continental Australia samples (*n*=342). Covariate data were then collated from a variety of sources and reflected a range of soil-forming factors (i.e., SCORPAN variables: S=soil; C=climate; O=organisms; R=relief; P=parent material; A=age; N=spatial location; McBratney *et al.* 2003) (see **Supplementary Table 6** for a list and description of all environmental covariate data). We identified 49 predictor variables (43 continuous, 6 categorical) as being relevant to our study, for which data sets were downloaded from the following sources: Australian Microbiome Initiative (e.g., organic carbon, clay content %, conductivity; Bissett *et al.* 2016), Atlas of Living Australia (e.g., annual temperature range, aridity index annual mean; Williams *et al.* ; Belbin 2011), Soil and Landscape Grid of Australia (e.g., Prescott index, topographic wetness index; O'Brien 2021), and Geoscience Australia (silica index; Cudahy *et al.* 2012). 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,

while 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 Variable Inflation Factor scores >15), and scatterplots generated often showed a curvilinear relationship with BPC_{16S} scores (scatterplots shown in **Supplementary Figure 2**). Therefore, we used two approaches (detailed below) that were less influenced by collinearity for subsequent analyses: principal components analysis into *k*-means clustering and decision tree modelling via Random Forest (Breiman 2001). Incomplete cases (n=1,122) were removed, leaving 1,331 samples for our analyses (**Supplementary Table 7**).

We scaled and analysed the 43 continuous predictor variables using principal component analysis (PCA) to reduce the dimensionality of the variables. PC1 and PC2 explained 27.4% and 14.0% of variance, respectively, and were selected for data visualisations (**Supplementary Figure 3**). We then used *k*-means clustering 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 both four and five clusters. We found that the additional fifth cluster more distinctly separated 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 examined using Kruskal-Wallis tests. We gave each cluster a generalised description and plotted the sample geospatial coordinates into maps using ggmap (Kahle & Wickham 2013) and Google maps to visualise their geographical distributions.

We then used 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 ($r > 0.8$ or VIF > 12) were removed (n=9). Tuning the hyperparameters of the model improve its performance (R^2) by 0.001. The R package spatialRF (Benito 2021) was used to minimise spatial autocorrelation of the residuals while fitting the spatial regression model. The resulting Random Forest decision tree model explained 35.3% of the variance in our BPC_{16S} dataset. 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**).

RESULTS

Global distribution of butyrate-producing bacteria

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**).

Butyrate production capacity of different environments

Human samples

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 (Kruskal-Wallis test: H=1136, 4 d.f., *P*<0.001; **Figure 2A**).

Non-human animal host-associated samples

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 non-human 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: Wilcoxon rank sum test: W=22,592, *P*<0.001). We then compared non-human animal samples by taxonomic phylum (between-group differences: Kruskal-Wallis test: H=331, 4 d.f., *P*<0.001; **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 versus the human gut showed the human gut had slightly higher BPC scores (3.19, *n*=800) than the primate gut (3.12, *n*=26).

Plant samples

Our dataset included 1,006 plant-associated samples. These were subcategorised 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 had the lowest median BPC_{meta} score (1.76, *n*=30; between-group differences: Kruskal-Wallis test: H=105, 3 d.f., *P*<0.001; **Figure 2C**).

Soil samples

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 5**). At 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: Kruskal-Wallis test: H=186, 5 d.f., *P*<0.001; **Figure 2D**).

Aquatic samples

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: Kruskal-Wallis test: H=530, 4 d.f., *P*<0.001; **Figure 2E**).

Agricultural and industrial samples

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: Kruskal-Wallis test: H=431, 6 d.f., *P*<0.001; **Figure 2F**). Activated sludge is a bacteria-rich product formed in aeration tanks with aerobic conditions.

Environmental characteristics associate with soil BPC scores

When clustering the Australian 16S rRNA soil samples on environmental metadata, each cluster matched with a distinct land type, which were given the following descriptive titles: "arid inland clay plains" (cluster 1), "seasonally productive sandy soils" (cluster 2), "sandy inland deserts" (cluster 3), "temperate urban hinterland" (cluster 4), and "wet, cold, acidic, vegetated montane" (cluster 5). Mapping the geographical locations of the sample sites showed consistency with the clustered land type descriptions (**Figure 3A**). Median BPC_{16S} scores varied significantly between the clusters (Kruskal-Wallis test: H=164, 4 d.f., *P*<0.001; **Figure 3B**). The highest median BPC_{16S} scores came from the seasonally productive sandy soils (0.93, *n*=385) and temperate urban hinterland cluster (0.86, *n*=445). The lowest median BPC_{16S} score came from the wet, cold, acidic, vegetated montane cluster (0.49, *n*=205). Principal components analysis of all continuous environmental predictor variables showed two distinct patterns of axial distribution, environmental wetness and greenness and soil fertility, and dimensions 1 and 2 explained 27.4% and 14.0% of the variation in the data, respectively (**Figure 3C**).

The cluster with the highest median BPC_{16S} score was the seasonally productive sandy soil, which has low clay content, low levels of cations, high precipitation seasonality, and

generally moderate topographic relief, Prescott index, and mean temperature. The cluster with the second-highest median BPC_{16S} score was temperate urban hinterland, which is generally moderate in elevation, annual rainfall, topographic relief, clay, and soil fertility and has high levels of zinc and manganese. The cluster with the lowest median BPC_{16S} score, wet, cold, acidic, vegetated montane, had high elevation and topographic relief, colder 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 inland desert, also had distinct characteristics (**Supplementary Table 8**). A separate analysis of categorical variables also showed the highest BPC_{16S} scores among specific land cover types ("built-up" and "annual crops and highly modified pastures") and land use types ("rural residential" and "plantations") (**Supplementary Table 9**).

The variable importance plot from Random Forest decision tree analysis (**Figure 4A)** showed that the eight top predictor variables were elevation, longitude, latitude, topographic relief, precipitation seasonality, iron, Prescott index, and population density, suggesting that geography was a strong driver of the abundance of butyrate-producing bacteria in soils, with climate and human presence also having notable effects. Low-to-moderate elevation, moderate Australian latitude (with several large cities), and moderate topographic relief were key drivers of the higher BPC_{16S} scores (**Figure 4B**). Partial dependence plots for each variable are shown in **Supplementary Figure 4**.

DISCUSSION

We report biogeographical patterns of butyrate-producing bacteria via our novel BPC formulae that use both metagenomic and 16S rRNA bacterial data. Our BPC_{16S} results

showed that geographical and climatic variables and human presence are key environmental drivers of soil butyrate production capacity. Soils in locations with moderate elevation and topographic relief, with greater seasonal variation in precipitation, had higher BPC scores. These geographic findings suggest that butyrate production is favoured in areas with a temperate climate with seasonal flushes of green growth. We also show that human and non-human animal presence substantially increased soil butyrate producer abundances, which probably reflects host-to-host transmission of gut spore-forming butyrate producers. Among the many patterns identified across samples, we highlight two of them below: anaerobic conditions and human presence.

Anaerobic conditions affect BPC scores

We show that BPC_{meta} scores were higher in anaerobic environments. Our analysis of plant and non-human animal data showed that compartments exposed to air, such as leaf surfaces and human skin, had the lowest median BPC_{meta} scores of their respective groups. Similarly, wastewater treatment plant samples showed that activated sludge, a product formed from aeration tanks, had among the lowest BPC_{meta} scores of its category. Samples from anaerobic digesters and the chordate gut, which each employ anaerobic processes, had the highest BPC_{meta} scores. These findings are consistent with previous work showing that butyrate-producing bacteria thrive in anaerobic environments (Conrad 2020) and support the validity of our BPC_{meta} formula.

Effect of human presence on BPC scores

We report an association between the presence of human and BPC scores. Soils from dense settlements showed the highest BPC_{meta} scores. In contrast, soils from ‘wildlands’ had low median BPC_{meta} scores. These results suggest that the presence of humans probably

contributes to butyrate production, possibly via inadvertent dispersal of gut-associated bacteria into the environment (Blum *et al.* 2019). Our 16S rRNA data from Australian soils also support this human-butyrate production connection. Australia's major cities and hinterlands are coastal, often with river-floodplain systems and areas of fertile soils that were attractive to European settlers. Our BPC_{16S} results suggest an association between human presence, local biodiversity, and high BPC scores. Still, the direction of influence and bacterial transfer remains a compelling question that may be of future research interest.

Intriguingly, seasonally-variable sandy soils had high BPC scores. This finding does not appear to follow the above patterns of higher BPC scores in more fertile soils and anaerobic conditions. However, seasonally-variable sandy soils showed the highest precipitation seasonality. Seasonal rainfall could lead to flushes of green growth, followed by a suitable climate to enable microbial breakdown. Sandy soils would indicate less protection of organic matter via association with clay minerals, which could enhance microbial breakdown of existing organic matter (Johnston *et al.* 2009). This suggests a high turnover system, and most cropping lands were found in this cluster.

Future research directions and limitations

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, but their updates are not uniform across databases. We used the Genome Taxonomy Database to classify our list of butyrate-producing bacteria, but it showed occasional discrepancies with the classification system applied to the

Australian soil 16S rRNA data. Thus, utilising multiple taxonomic classification systems likely means that some butyrate producers were not identified in our data. This could affect the reliability of our results and a future sensitivity analysis is warranted.

To maximise 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 within a species), among which some may be butyrate producers and others not. Analysis at the strain level could provide a higher resolution of data, which we recommend is a future research priority. Our data did not measure butyrate production directly; this would require different data types (e.g., metatranscriptomic data) and pipelines. Furthermore, 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 m; thus, our mountain cluster modelling may not fit other countries with higher mountains.

Finally, our data depends on the capacity of laboratory DNA extraction methods to open endospores. Sampling methods that expose the samples to air may inadvertently cause the sporulation of bacteria. Such methods may subsequently reduce the quantities of DNA extracted from spore-formers, several of which could be butyrate-producing bacteria (Browne *et al.* 2016). Consistency across sampling and DNA extraction methods in future studies could help improve butyrate-producing bacterial abundance data reliability.

Conclusions

Butyrate-producing bacteria provide critical ecosystem services for both hosts, including humans, and soils. Understanding the distribution and environmental drivers of these bacteria is essential to building new knowledge of their roles in ecosystem functionality and human health (Brame *et al.* 2021). We present new evidence that outdoor ecosystems influenced by human-associated processes are abundant reservoirs of butyrate producers. Because half of the world's population now lives in cities (Rydin *et al.* 2012), understanding the influence of the presence of humans on nearby green spaces is essential to biodiversity research and urban green space design (Delgado-Baquerizo *et al.* 2018; Kondo *et al.* 2018; Watkins *et al.* 2020). Public health policymakers also have a critical need for research to guide the development of urban green infrastructure that supports its residents (Robinson *et al.* 2021). Our study helps advance these research areas. Future assessment of butyrate-production capacity across fine spatial scales (i.e., below global and continental, as used here) will help provide greater detail for city infrastructure planning and further microbiome-based public health and ecology research.

Acknowledgements

We would like to acknowledge that this work is supported by funding from the Flinders Foundation, Australian Research Council [grants LP190100051] and the New Zealand Ministry of Business Innovation and Employment [grant UOWX2101].

References

1.
Baxter, N.T., Schmidt, A.W., Venkataraman, A., Kim, K.S., Waldron, C. & Schmidt, T.M. (2019). Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. *mBio*, 10, 13.

2.
Bedford, A. & Gong, J. (2018). Implications of butyrate and its derivatives for gut health and animal production. *Anim Nutr*, 4, 151-159.

3.
Belbin, L. (2011). The Atlas of Living Australia's spatial portal. In: *Proceedings of the Environmental Information Management Conference (EIM 2011)* (ed. Jones, M, B. & Gries, C.) Santa Barbara, pp. 39-43.

4.
Benito, M. (2021). spatialRF: Easy spatial regression with random forest. *R package version*, 1.

5.
Bissett, A., Fitzgerald, A., Meintjes, T., Mele, P.M., Reith, F., Dennis, P.G. *et al.* (2016). Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. *GigaScience*, 5, s13742-13016-10126-13745.

6.
Blum, W.E., Zechmeister-Boltenstern, S. & Keiblinger, K.M. (2019). Does soil contribute to the human gut microbiome? *Microorganisms*, 7, 287.

7.
Brame, J.E., Liddicoat, C., Abbott, C.A. & Breed, M.F. (2021). The potential of outdoor environments to supply beneficial butyrate-producing bacteria to humans. *Science of The Total Environment*, 777, 146063.

8.
Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

9.
Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D. *et al.* (2016). Culturing of 'unculturable'human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533, 543-546.

10.
Browne, H.P., Neville, B.A., Forster, S.C. & Lawley, T.D. (2017). Transmission of the gut microbiota: spreading of health. *Nat. Rev. Microbiol.*, 15, 531-543.

11.
Chen, I.-M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M. *et al.* (2021). The IMG/M data management and analysis system v. 6.0: new tools and advanced capabilities. *Nucleic acids research*, 49, D751-D763.

12.
Conrad, R. (2020). Methane production in soil environments—anaerobic biogeochemistry and microbial life between flooding and desiccation. *Microorganisms*, 8, 881.

13.
Cudahy, T., Caccetta, M. & Thomas, M. (2012). National ASTER Map of Australia. *Scale*, 1, 1000000.

14.
Delgado□Baquerizo, M., Reith, F., Dennis, P.G., Hamonts, K., Powell, J.R., Young, A. *et al.* (2018). Ecological drivers of soil microbial diversity and soil biological networks in the Southern Hemisphere. *Ecology*, 99, 583-596.

15.

Ellis, E.C., Gauthier, N., Goldewijk, K.K., Bird, R.B., Boivin, N., Díaz, S. *et al.* (2021). People have shaped most of terrestrial nature for at least 12,000 years. *Proceedings of the National Academy of Sciences*, 118.

16.

Ferraris, L., Butel, M.J., Campeotto, F., Vodovar, M., Rozé, J.C. & Aires, J. (2012). Clostridia in premature neonates' gut: incidence, antibiotic susceptibility, and perinatal determinants influencing colonization. *PLoS One*, 7, e30594.

17.

Fierer, N. & Jackson, R.B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences*, 103, 626-631.

18.

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

19.

Goldfarb, K.C., Karaoz, U., Hanson, C.A., Santee, C.A., Bradford, M.A., Treseder, K.K. *et al.* (2011). Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. *Frontiers in microbiology*, 2, 94.

20.

Johnston, A.E., Poulton, P.R. & Coleman, K. (2009). Soil organic matter: its importance in sustainable agriculture and carbon dioxide fluxes. *Advances in agronomy*, 101, 1-57.

21.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T. *et al.* (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology*, 7, 459.

22.

Kahle, D.J. & Wickham, H. (2013). ggmap: spatial visualization with ggplot2. *R J.*, 5, 144.

23.

Knox, E.G., Aburto, M.R., Tessier, C., Nagpal, J., Clarke, G., O'Driscoll, C.M. *et al.* (2022). Microbial-Derived Metabolites Induce Actin Cytoskeletal Rearrangement and Protect Blood-Brain Barrier Function. *iScience*, 105648.

24.

Kondo, M.C., Fluehr, J.M., McKeon, T. & Branas, C.C. (2018). Urban green space and its impact on human health. *International journal of environmental research and public health*, 15, 445.

25.

Lang, J.M., Eisen, J.A. & Zivkovic, A.M. (2014). The microbes we eat: abundance and taxonomy of microbes consumed in a day's worth of meals for three diet types. *PeerJ*, 2, e659.

26.

Levine, U.Y., Looft, T., Allen, H.K. & Stanton, T.B. (2013). Butyrate-producing bacteria, including mucin degraders, from the swine intestinal tract. *Applied and Environmental Microbiology*, 79, 3879-3881.

27.

Liddicoat, C., Sydnor, H., Cando-Dumancela, C., Dresken, R., Liu, J., Gellie, N.J.C. *et al.* (2020). Naturally-diverse airborne environmental microbial exposures modulate the gut microbiome and may provide anxiolytic benefits in mice. *Science of the Total Environment*, 701, 134684.

28.

Liu, P., Qiu, Q. & Lu, Y. (2011). Syntrophomonadaceae-affiliated species as active butyrate-utilizing syntrophs in paddy field soil. *Applied and environmental microbiology*, 77, 3884-3887.

29.

Mallott, E.K. & Amato, K.R. (2022). Butyrate Production Pathway Abundances Are Similar in Human and Nonhuman Primate Gut Microbiomes. *Molecular biology and evolution*, 39, msab279.

30.

McBratney, A.B., Santos, M.M. & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117, 3-52.

31.

McMurdie, P.J. & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8, e61217.

32.

Miquel, S., Martin, R., Rossi, O., Bermúdez-Humarán, L., Chatel, J., Sokol, H. *et al.* (2013). Faecalibacterium prausnitzii and human intestinal health. *Curr. Opin. Microbiol.*, 16, 255-261.

33.

Nilsen, M., Madelen Saunders, C., Leena Angell, I., Arntzen, M.Ø., Lødrup Carlsen, K.C., Carlsen, K.-H. *et al.* (2020). Butyrate Levels in the Transition from an Infant-to an Adult-Like Gut Microbiota Correlate with Bacterial Networks Associated with Eubacterium Rectale and Ruminococcus Gnavus. *Genes*, 11, 1245.

34.

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

35.

Polymenakou, P.N. (2012). Atmosphere: a source of pathogenic or beneficial microbes? *Atmosphere*, 3, 87-102.

36.

Poret-Peterson, A.T., Albu, S., McClean, A.E. & Kluepfel, D.A. (2019). Shifts in soil bacterial communities as a function of carbon source used during anaerobic soil disinfestation. *Frontiers in Environmental Science*, 6, 160.

37.

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

38.

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

39.

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

40.

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

41.

Stilling, R.M., van de Wouw, M., Clarke, G., Stanton, C., Dinan, T.G. & Cryan, J.F. (2016). The neuropharmacology of butyrate: the bread and butter of the microbiota-gut-brain axis? *Neurochemistry international*, 99, 110-132.

42.

Thakuria, D., Schmidt, O., Finan, D., Egan, D. & Doohan, F.M. (2010). Gut wall bacteria of earthworms: a natural selection process. *The ISME journal*, 4, 357-366.

43.

Vital, M., Howe, A.C. & Tiedje, J.M. (2014). Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta) genomic Data. *mBio*, 5, 11.

44.

Watkins, H., Robinson, J.M., Breed, M.F., Parker, B. & Weinstein, P. (2020). Microbiome-Inspired Green Infrastructure: A Toolkit for Multidisciplinary Landscape Design. *Trends in Biotechnology*.

45.

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

46.

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.

47.

Yang, Y., Yu, K., Xia, Y., Lau, F.T., Tang, D.T., Fung, W.C. *et al.* (2014). Metagenomic analysis of sludge from full-scale anaerobic digesters operated in municipal wastewater treatment plants. *Appl. Microbiol. Biotechnol.*, 98, 5709-5718.

692 **Table 1**. Overview of our analyses and data types used.

Sequenced data type	**Shotgun metagenomic reads**	**16S rRNA amplicon reads**
Formula (see methods for details)	BPC_{meta}	BPC_{16S}
Sample locations	Global (terrestrial and aquatic)	Australia
Sample sources	All available	Soil, 1-10 cm depth
Database(s) interrogated	Gene data (for *buk* and *atoA* genes) and sample metadata: Integrated Microbial Genomes & Microbiomes	Amplicon data and sample metadata: Australian Microbiome Initiative Environmental data: Atlas of Living Australia, Soil and Landscape Grid of Australia, Socioeconomic Data and Applications Center, and Geoscience Australia Taxonomic data: Genome Taxonomy Database
Downloaded sample count	22,593	3,023
Normalisation step	Single copy gene *pheS* data from IMG/M	Relative abundance
Sample BPC score range	0.02 – 3.39	0.003 – 12.59
Statistics and visualisation	After grouping samples into categories, between-group variance and mapping	Principal components analysis, *k*-means clustering, mapping, decision tree modelling

Figure 1

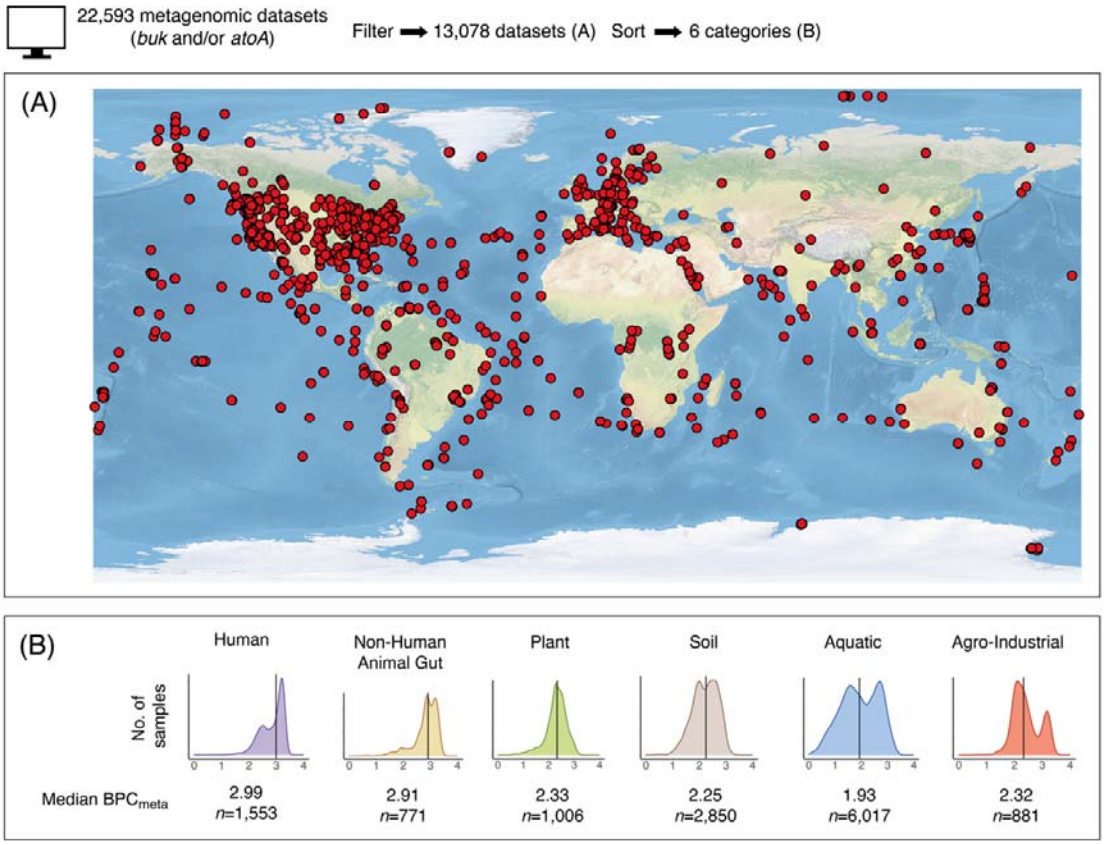

22,593 metagenomic datasets
(buk and/or atoA)
Filter ➡ 13,078 datasets (A)
Sort ➡ 6 categories (B)
(A)
(B)
Human
Non-Human Animal Gut
Plant
Soil
Aquatic
Agro-Industrial
No. of samples
0 1 2 3 4
Median BPCmeta
2.99
n=1,553
2.91
n=771
2.33
n=1,006
2.25
n=2,850
1.93
n=6,017
2.32
n=881

Figure 2

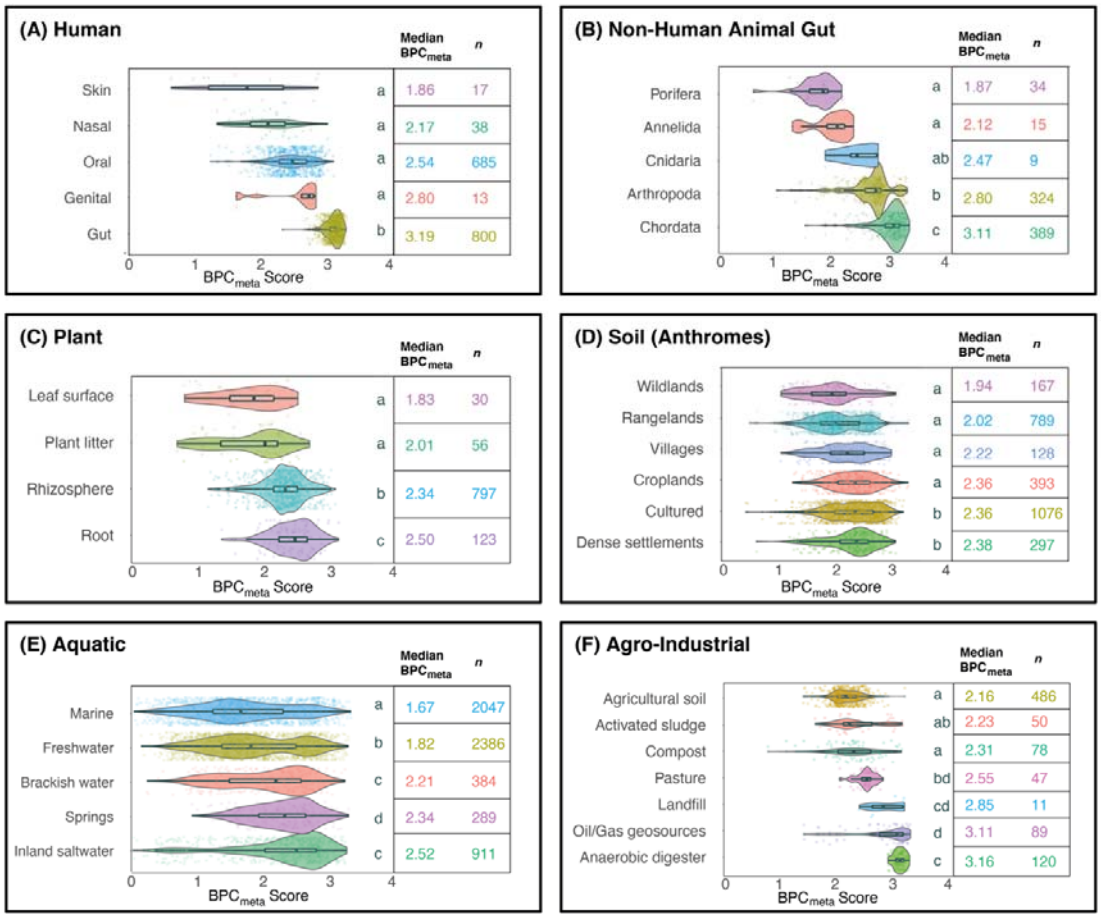

(A) Human
Median BPCmeta
n
Skin a 1.86 17
Nasal a 2.17 38
Oral a 2.54 685
Genital a 2.80 13
Gut b 3.19 800
0 1 2 3 4
BPCmeta Score
(B) Non-Human Animal Gut
Median BPCmeta
n
Porifera a 1.87 34
Annelida a 2.12 15
Cnidaria ab 2.47 9
Arthropoda b 2.80 324
Chordata c 3.11 389
0 1 2 3 4
BPCmeta Score
(C) Plant
Median BPCmeta
n
Leaf surface a 1.83 30
Plant litter a 2.01 56
Rhizosphere b 2.34 797
Root c 2.50 123
0 1 2 3 4
BPCmeta Score
(D) Soil (Anthromes)
Median BPCmeta
n
Wildlands a 1.94 167
Rangelands a 2.02 789
Villages a 2.22 128
Croplands a 2.36 393
Cultured b 2.36 1076
Dense settlements b 2.38 297
0 1 2 3 4
BPCmeta Score
(E) Aquatic
Median BPCmeta
n
Marine a 1.67 2047
Freshwater b 1.82 2386
Brackish water c 2.21 384
Springs d 2.34 289
Inland saltwater c 2.52 911
0 1 2 3 4
BPCmeta Score
(F) Agro-Industrial
Median BPCmeta
n
Agricultural soil a 2.16 486
Activated sludge ab 2.23 50
Compost a 2.31 78
Pasture bd 2.55 47
Landfill cd 2.85 11
Oil/Gas geosources d 3.11 89
Anaerobic digester c 3.16 120
0 1 2 3 4
BPCmeta Score

Figure 3

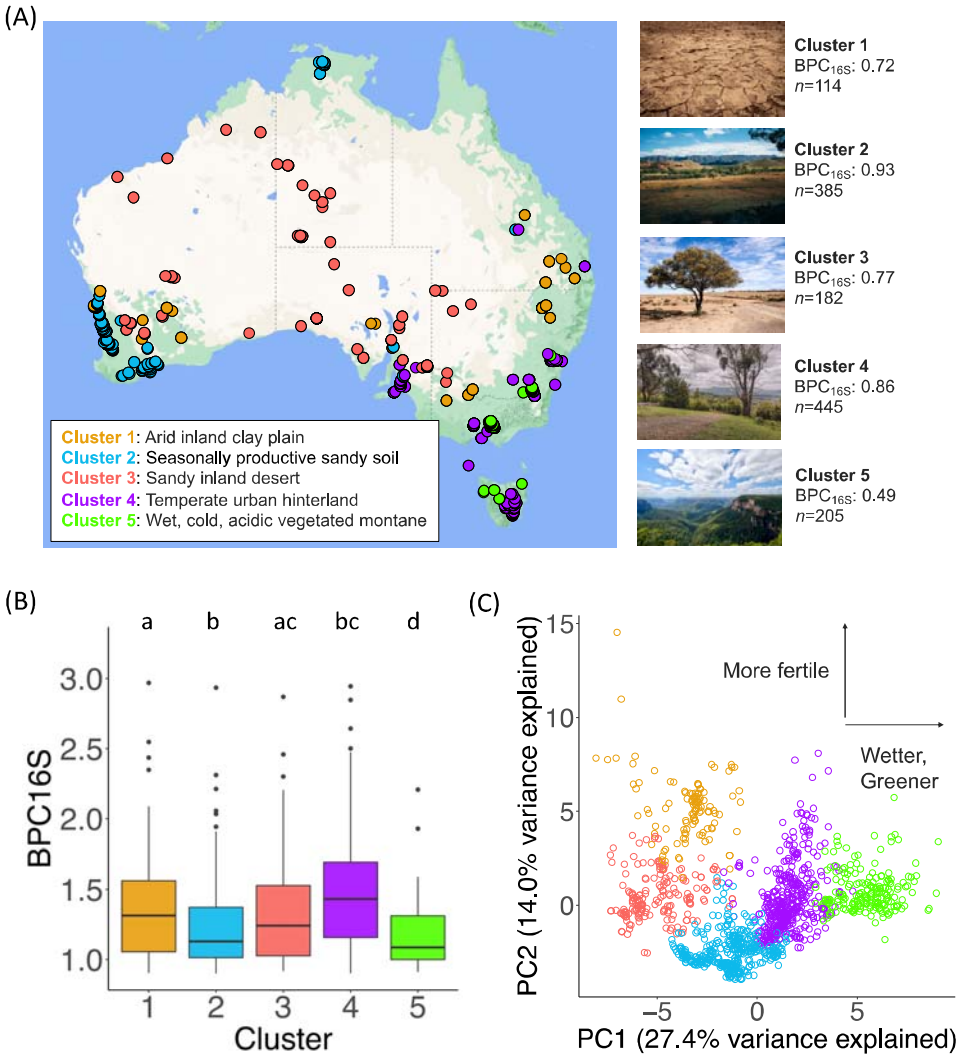
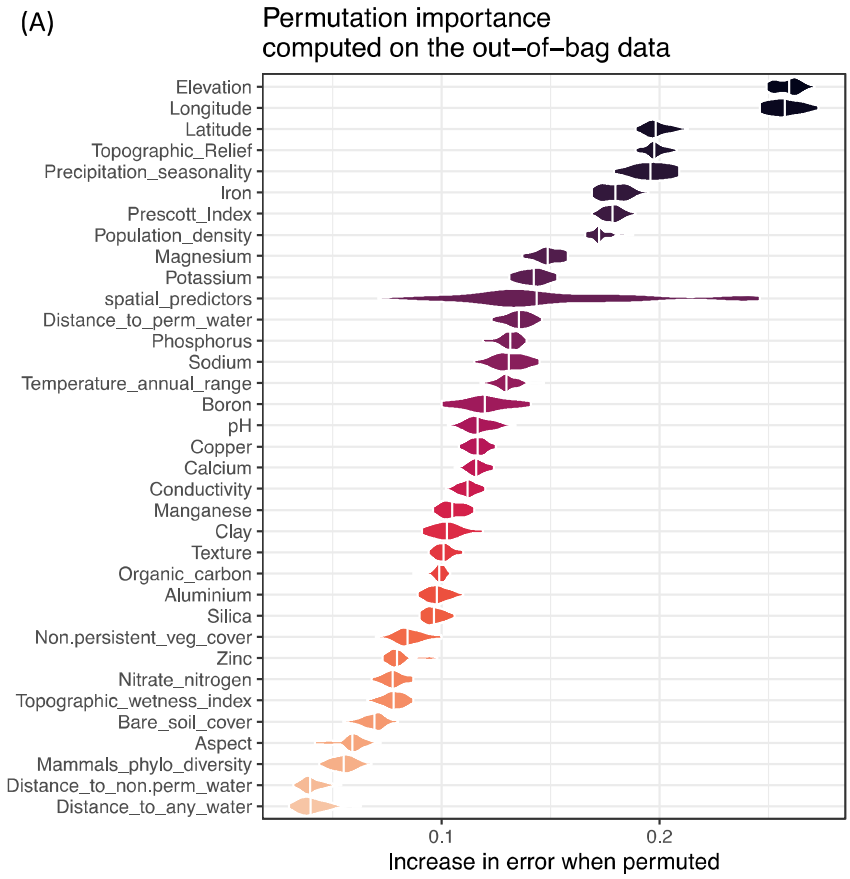

(A)
Cluster 1: Arid inland clay plain
Cluster 2: Seasonally productive sandy soil
Cluster 3: Sandy inland desert
Cluster 4: Temperate urban hinterland
Cluster 5: Wet, cold, acidic vegetated montane
Cluster 1
BPC16S: 0.72
n=114
Cluster 2
BPC16S: 0.93
n=385
Cluster 3
BPC16S: 0.77
n=182
Cluster 4
BPC16S: 0.86
n=445
Cluster 5
BPC16S: 0.49
n=205
(B)
a
b
ac
bc
d
BPC16S
3.0
2.5
2.0
1.5
1.0
1
2
3
4
5
Cluster
(C)
More fertile
Wetter, Greener
PC2 (14.0% variance explained)
15
10
5
0
−5
0
5
PC1 (27.4% variance explained)

Figure 4


(A)
Permutation importance
computed on the out−of−bag data
Elevation
Longitude
Latitude
Topographic_Relief
Precipitation_seasonality
Iron
Prescott_Index
Population_density
Magnesium
Potassium
spatial_predictors
Distance_to_perm_water
Phosphorus
Sodium
Temperature_annual_range
Boron
pH
Copper
Calcium
Conductivity
Manganese
Clay
Texture
Organic_carbon
Aluminium
Silica
Non.persistent_veg_cover
Zinc
Nitrate_nitrogen
Topographic_wetness_index
Bare_soil_cover
Aspect
Mammals_phylo_diversity
Distance_to_non.perm_water
Distance_to_any_water
0.1
0.2
Increase in error when permuted

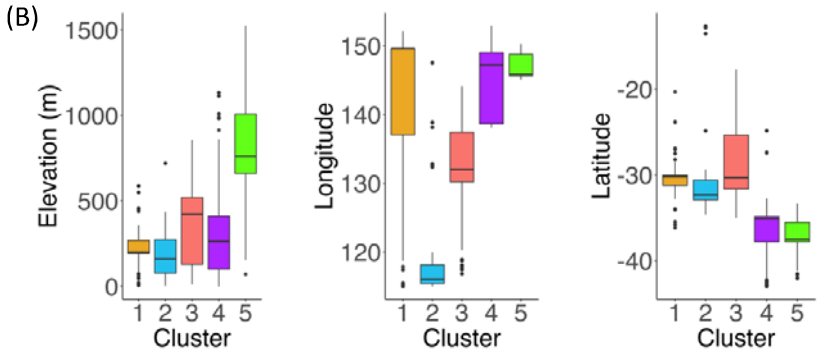

(B)
Elevation (m)
1500
1000
500
0
1 2 3 4 5
Cluster
Longitude
150
140
130
120
1 2 3 4 5
Cluster
Latitude
-20
-30
-40
1 2 3 4 5
Cluster

Figure Legends

Figure 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}). 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.

Figure 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% significance level. Boxes show the interquartile range.

Figure 3: Clustering of environmental data shows five distinct land types. (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 of median BPC_{16S} scores across each of the five clusters. Medians sharing a letter are not significantly different by the adjusted Dunn test at the 5% significance level. Boxes show the interquartile range. (C) The first two principal components coloured by k-means clusters. The x-axis can be broadly interpreted as environmental wetness and greenness and associated variables (e.g., vegetation cover). The y-axis can be broadly interpreted as soil fertility and the presence of cations.

Figure 4: Variable importance results from Random Forest decision tree modelling. (A) 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. (B) Boxplots for the top three variables from (A) across each of the five clusters. In each of elevation, longitude, and latitude, Kruskal-Wallis tests show that between-group differences were significant at $P<0.001$. Boxes show the interquartile range.