Microbiota composition is moderately associated with greenspace composition in a UK cohort of Twins.

Authors:

- Ruth C. E. Bowyer[1] (ruth.c.bowyer@kcl.ac.uk)
- Caoimhe Twohig-Bennett [2] (caoimhe.bennett@eruk.org.uk)
- Emma Coombes[2] (emma.coombes@uea.ac.uk)
- Philippa M. Wells[1] (philippa.wells@kcl.ac.uk)
- Tim Spector[1] (tim.spector@kcl.ac.uk)
- Andy P. Jones[2] (A.P.Jones@uea.ac.uk)[±]
- Claire J. Steves[1, 3 ±] * (claire.j.steves@kcl.ac.uk)

[±]**AP Jones and CJ Steves contributed equally**

*Corresponding author

Affiliations:

1. Department of Twin Research and Genetic Epidemiology, King's College London, 3-4th Floor South Wing Block D, St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK

2.Norwich Medical School, University of East Anglia, Norwich, Norfolk, NR4 7TJ

3. Department of Ageing and Health, St Thomas' Hospital, 9th floor, North Wing, Westminster Bridge Road, London SE1 7EH

Abstract

Exposure to natural environments, known as greenspace, appears to positively influence health, yet the mechanisms are unclear. Given that gut microbiota are associated with inflammatory disorders more prevalent in urban areas and individuals with lower greenspace exposure, microbiota may act as a mediator between greenspace and health. Using 2443 participants of the TwinsUK cohort, microbiota differences were compared in relation to rural/urban living and with quantiles of area-level greenspace at three different neighbourhood distances: 800m, 3000m and 5000m. Using microbiota data captured from faecal samples using 16S rRNA marker gene sequencing, small compositional differences in association with 3000m greenspace ($p=0.003$) in models adjusted for confounders of microbiota variance (sequencing depth, antibiotics use, body mass index, frailty, age, diet, region and socioeconomic variables) were observed. Differences in abundances of genus were observed for all measures of greenspace in adjusted models; a key pathogenic genus was increased in abundance in association with urbanicity (*Escherichia/Shigella*, logFC = 0.73742, padj < 0.001). Further, utilising the twin structure, within-pair differences in microbiota composition were compared and associations with 800m greenspace observed (factor level significance in association with greatest difference, $\beta= 0.08$, $p= 0.0162$) as were differences in *Escherichia/Shigella*. The microbiota signature of those with a greater exposure to greenspace, but not necessarily explicitly rural individuals, was distinct from other individuals, suggesting microbiota as a potential mediator for greenspace and health.

Keywords: Greenspace, microbiota, rural-urban classification, twin differences

Introduction

The human gut microbiome is the collective genome and surrounding environment of the assemblage of microorganisms inhabiting the gastrointestinal tracts (Goodrich et al., 2016; Marchesi and Ravel, 2015). It interacts with the immune system, is essential for nutrient processing, and offers protection from pathogens (Geva-Zatorsky et al., 2017; Litvak and Bäumler, 2019; Magnúsdóttir et al., 2015). The microbiome is therefore increasingly considered a target for clinical intervention. Whilst the extent of its influence on health is still being debated, there is evidence that it can be considered a biomarker and potential mediator of environmental exposures and health (Cresci and Bawden, 2015). Indeed, because of the relatively low influence of genetics on its composition, environmental factors are key to understanding the forces shaping its composition (Rothschild et al., 2018). Whilst influences such as diet, medication use and disease status have been more widely studied (for example: Ordiz *et al.*, 2015; Shreiner, Kao and Young, 2015; Jackson *et al.*, 2016; Belzer *et al.*, 2017; Weersma, Zhernakova and Fu, 2020), the influence of environmental factors more distal to the host on microbiota composition, such as where an individual lives, are less well understood (Schmidt et al., 2018). This is despite both the emerging importance of the microbiome and increasing evidence that exposures from the environment influence health.

High levels of exposure to natural environments may, for certain conditions, have similar therapeutic benefits equivalent in magnitude to some drug treatments (Twohig-Bennett and Jones, 2018) and consequently, urban greenspaces have been proposed as a means to mitigate the negative health impacts of urbanisation. Whilst frequently proposed because of the association of inflammatory and chronic immunological disease and urbanisation, it is unclear whether the mode of action for the observed benefit is mediated by the gut microbiota (Rook, 2013; Rook et al., 2013). Previous studies have suggested differences in the microbiota by urbanicity (Obregon-Tito et al., 2015), but have not assessed the relative importance of exposure to different types of environment.

Associations between any health trait and greenspace are likely confounded by the complex interdependency of social and behavioural factors that contribute to an individual's greenspace exposure (Robinson and Jorgensen, 2020; Stamper et al., 2016). This study makes use of the TwinsUK cohort, the UK's largest registry of twins, and aims to investigate how microbiota assemblages differ by greenspace or urbanicity.

Methods

Participants

Participants were members of TwinsUK, the UK's largest registry of mono and dizygotic twins, run out of the Department of Twin Research & Genetic Epidemiology, King's College London. Incepted to study the heritability of osteoarthritis and osteoporosis, the registry has grown to be one of the most clinically detailed twin cohorts worldwide (Verdi et al., 2019). Data from twins is collected approximately every four years either upon visit to the Clinical Research Facility, or via questionnaire collected on a more frequent basis over the telephone, sent in the post, or filled in and brought to visits. For the purposes of this analysis, all measures were matched based on date unless otherwise stated. 3218 samples from 2707 individuals from within the TwinsUK cohort with appropriately collected and processed 16S rRNA gene sequence data were considered for this analysis. 3010 samples were successfully geocoded; those where a postcode match could not be found were likely due to an administrative error, recording of a now disused postcode, or because the participant lived abroad. Participants were spread across the UK, with the majority in England (96.5%), of which 43.8% were living in South East of England or Greater London at time of microbiota sample. For primary analysis, the earliest sample provided by individuals was used where applicable with repeat samples provided by the same individual on the same day removed (sample chosen at random for removal), resulting in 2443 samples/individuals considered in this analysis.

Microbiome samples

Profiles of gut microbiota composition for each twin were available as a subset generated from faecal samples as previously described (Goodrich et al., 2016). Samples were collected and sequenced on a rolling basis between November 2011 and 2015. Participants stored samples in sealed ice packs and either posted them to the research department or provided them during clinical visit. Samples were stored at −80 °C and shipped frozen to Cornell University for DNA extraction and amplification of the V4 variable region of the 16S rRNA gene amplified in duplicate by PCR using 515F and 806R primers, following isolation of genomic DNA using the 'MoBio PowerSoil htp DNA isolation kit' from an aliquot of ~100mg of each sample. As described by Goodrich and colleagues 2014, PCR reactions comprised 2.5 U Easy-A high-fidelity enzyme, 1 × buffer, 10-100 ng DNA template, and 0.05 μM of each primer. Initial denaturation was carried out at 94°C for 3 min followed by 25 cycles of denaturation at 94°C for 45 seconds, followed by annealing at 50°C for 60 seconds, extension at 72°C for 90 seconds, and a final extension at 72°C for 10 minutes. A magnetic bead system was used to combine and purify the PCR reactions and washout contaminants. The QuantiT PicoGreen dsDNA Assay Kit was used to quantify the PCR amplicons, aliquots of which were combined for a final concentration of approximately 15 ng/μ. This 16S rRNA gene is a unit of the bacterial genome that exhibits extreme sequence conservation whilst containing 9 hypervariable regions that can be targeted for broad differentiation between microbial lineages (Tringe and Hugenholtz, 2008). For the limitations and advantages to this approach see Weinstock (2012) and Pollock et al. (2018).

The resulting sequences were analysed as amplicon sequence variants (ASVs) following the DADA2 pipeline (Callahan et al., 2016). Briefly, DNA sequences were demultiplexed, and separate forward and reverse read files were generated for each sample using QIIME (Caporaso et al., 2010). Quality of sequences was assessed, with ends trimmed to remove poor quality reads, error estimated within-sample for forward and reverse reads, and then the ASV algorithm applied. Forward and reverse ASVs were joined, and the total dataset merged. Chimeras were removed. Taxonomic assignment was via SILVA 1.3.2. Samples with less than 10,000 sequences or with only one viable read direction

were removed. ASVs can be considered as a 'taxonomic unit' of the microbiome, akin to a species of microorganism as best estimated using the 16S rRNA gene sequencing method, with taxonomic description at the lowest resolution possible for that sequence (i.e. some ASVs will be mapped to a species, others will only be assigned to a taxonomic order or family).

Residential location

To maximise the likelihood that the participant was living at the address when the sample was provided, residential postcodes (zipcodes) of study participants were linked to microbiota samples by the nearest date of a provided postcode prior to the sample date rather than the date outright. Matched postcodes were then geocoded based on their centroids (centre points) in R using package 'PostcodesioR' (Walczak, 2019), a wrapper that allows R interface with Postcodes.io and maps postcodes to Ordnance Survey postcode centroids.

Key predictor variables

The environmental exposure variables considered in this study were measures of urbanicity and greenspace, computed as follows.

To measure urbanicity, the Rural-Urban 2-fold classification (RUC) was linked to each individual's postcode as a measure of the urbanicity of their local area. The RUC was defined at the 2011 UK Census 'output-area' level (For data sources, please refer to *data availability)*. The measure was included to compare whether simply urbanicity, rather than greenspace, might shape microbiota composition.

For the measurement of greenspace in the residential vicinity of each study participant, buffer zones around each postcode were calculated using the Ordnance Survey Open Roads dataset that indicates the footpath and roads network across the UK. Using the QGIS 3.4.2 package, the area covering the

800m, 3000m and 5000m road and path networks from postcode centroids was converted to a polygon (closed geographical unit) representing the area surrounding an individual's home based on travel distance. Buffer distances were chosen to represent the effect of the immediate environment surrounding the home as well as the wider environment an individual might move through on a daily basis (Hillsdon et al., 2015).

The $25m^2$ Land Cover Map of Great Britain 2015 (LCM, version 1.2) was downloaded from the Centre for Ecology & Hydrology via the 'Digimap' data portal. The LCM uses satellite data to classify landcover across the UK into 21 classes. Buffer polygons were overlaid with the LCM, and percentage of each landcover type falling within each was calculated, with all the 21 classes that were not considered as 'urban' or 'suburban' being considered as greenspace. For the purposes of this analysis, greenspace percentages were defined as within-population quantiles. For details on the data sources please refer to the data availability section.

Statistical analysis

All statistical analyses were undertaken in RStudio using R version 3.6.3. Missing covariate data (Body Mass Index BMI, frailty, diet and antibiotic use) were imputed using the 'missForest' R package, an iterative imputation method based on random forest, that can impute both continuous and categorical data (Stekhoven and Buhlmann, 2012). Covariates were compared for their differences by the key outcome variables (urbanicity and greenspace) using pairwise Wilcoxon rank sum tests or chi-squared tests depending on the variable. Means, standard deviations (for continuous variables) and percentage of the dominant class (factor variables) were calculated.

Compositional (beta diversity) analysis

A compositional framework was favoured for comparison of inter-individual community composition (sometimes referred to as 'Beta diversity') (Gloor et al., 2017). Hypothesising that the composition

and presence of rarer microbes were more likely to differ geographically than ubiquitous ones, and equally, that it was of interest to understand whether the composition of the conserved microbial function differed spatially, ASVs were subset to those 'abundant' (>20% of the population, a frequently used cut off in microbiota analysis) and 'rare' (ASVs in between 7.5-20% of the population). As a noise-reduction strategy, abundant ASV table were 'collapsed' to genus (i.e. transformed counts were grouped together where their taxonomic assignment suggested they belonged to the same genera).

For this aspect of the analysis, following methods of Tipton et al. (2018), a pseudocount was added to the ASV table (prior to subsetting) using a Bayesian-multiplicative replacement of count zeros via the 'zCompositions' package (Palarea-Albaladejo and Martín-Fernández, 2015). Samples were subset as described above (i.e. rare, abundant and genus), and relative abundance of the samples was calculated in 'phyloseq' (McMurdie et al., 2013), followed by a centred-log ratio (CLR) transformation. Finally, Aitchison's distance calculated on each table.

Our analysis was framed by the use of permutational analysis of variance (PERMANOVA) using the vegan package (Oksanen et al., 2015) to assess variation in microbiota composition in association with outcome variables (i.e. how different are microbial assemblages in individuals living in areas of high greenspace compared to low greenspace). The model works by comparing group with the null hypothesis that group centroids and dispersion (defined by a distance matrix) are equivalent for all groups (Anderson, 2001). Homogeneity of variance and pairwise comparison of factors were assessed using *beta.dispers* and *pairwise.adonis* respectively.

PERMANOVA describes how variation of (in this case) a distance matrix can be attributed to different factors and variables. When specifying the model, terms are added sequentially, and therefore where factors are correlated, the extent of variance that they explain diminishes if they are added after a corollary (Anderson, 2001). Therefore, a nested approach was taken, as follows - **Model 1:** where models were performed with the habitat variable adjusted for family structure (to account for twin relatedness) and library size (a technical variable representing the number of

sequences per sample). **Model 2:** models adjusted as in Model 1 with the addition of their geographic region (to account for differences simply as a facet of distance/location, with the regions grouped as follows: 1. East Midlands, 2. East of England, 3. London, 4. North England & Scotland, 5. South East, 6. South West, 7. West Midlands and 8. Yorkshire and the Humber). **Model 3:** A saturated model, as in Model 2 with the inclusion of potential mediatory and confounding biological variables, all of which have previously been shown to associate with microbiota composition in this cohort. They were: body mass index (BMI), a measure of dietary quality in the form of the healthy eating index (HEI) (Bowyer et al., 2018), the frailty index – a measure of health deficit derived using Rockwood's method (Searle et al., 2008), age at microbiota sample, antibiotic use in the previous month, highest educational attainment and the area-based Index of Multiple Deprivation (IMD). The final two were considered as key measures of socioeconomic status, with the latter being an important capture of area-level deprivation. To understand the relative variance explained by each variable comparatively, where statistically significant Model 3 was repeated to include all terms marginally.

As a result of the above methodology, three outcome measures were assessed in the sequential manner as laid out above based on Aitchinson's distance matrix calculated on abundant ASVs, rare ASVs and genera, with the four key variables of interest as: 1) RUC, 2) 800m greenspace quantiles, 3) 3000m greenspace quantiles and 4) 5000m greenspace quantiles, totalling 36 models. Therefore the multiple-testing alpha threshold was set as 0.004 for the PERMANOVAs, by assuming very conservatively independence of the 3 Aitchinson's distances and 4 outcome variables, but not adjusting to account for the sequential design (i.e. 0.05/(3x4)).

Differences in ASV and genera abundances

Differences in ASV abundances were assessed using the DESeq2 pipeline (Ben J Callahan et al., 2016; Love et al., 2014) using both the abundant ASVs and genera tables described above, transformed using the variance stabilising method internal to DESeq2. Models were fitted both unadjusted and adjusted for sequencing depth, antibiotics in the month prior to sample collection, Body Mass Index, frailty, age, diet, index of multiple deprivation, highest educational attainment and

region of living. Results were false-discovery rate adjusted using the the Benjamini–Hochberg procedure.

Twin-pair differences

Twin pair differences were assessed in two ways. First, the within-in pair dissimilarity was extracted from the 'abundant' distance matrix (described above) and used as the outcome variable in linear regressions, adjusting for within-pair difference in antibiotic use, BMI, education attainment, frailty, diet, index of multiple deprivation and library size. Difference was calculated as the absolute value between each individual and their co-twin; all covariates are as described above. Second, variance stabilised genus abundances (derived within the DESeq2 pipeline) for those statistically significant in the population wide analysis were tested for difference between twins discordant for the relevant factor of interest using paired Wilcox rank sum tests. Discordance was defined as where one twin lived in a rural area versus one twin living in an urban area for RUC, and where one twin lived in a different quantile of greenspace.

Results

From the 3218 samples from 2707 individuals considered for analysis, 3010 were successfully geocoded and, after removal of repeats, 2433 samples were used in this analysis. Descriptive statistics (**Table 1**) suggest broadly the same sub-populations were observed across each quantile and by rural/urban differences, although there were some key differences: age was significantly lower in Q1 of 3000m greenspace and 5000m quantiles compared to all four other quantiles ($p<0.001$) and to a lesser extent in 800m greenspace quantiles (Q1:Q3 $p = 0.0022$, Q1:Q4 $p= 0.008$) and in urban dwellers compared with rural ($p=0.002$) and IMD differed by quantiles of 3000m and 5000m greenspace, primarily the most deprived compared to the others (Q1:Q2,Q3,Q4 $p<0.001$, Q2:Q3 $p=0.01$, Q3:Q4 $p=0.007$, Q1:Q2,Q3,Q4 $p<0.01$, Q2:Q3 $p=0.0003$, Q3:Q4 $p=0.00125$).

256 'Out-of-bag' imputation error (OOB error) was: 0.000169 (normalised root mean squared error,
257 NRMSE – continuous variables) and 0.0131 (Proportion of falsely classified- PFC – categorical
258 variables) suggesting the imputed dataset was suitable for analysis.

259 Percentage greenspace in buffers surrounding homes was highly correlated between the different
260 distances (**Figure S1**), although there were many more individuals with <5% estimated greenspace
261 within the 800m buffer (**Figure S2**).

262 ***Microbiota differences are observable in association with greenspace and to a lesser extent***
263 ***urbanicity***

264

265 Whilst there were statistically significant associations in unadjusted models for all greenspace buffers
266 and rare and abundant microbiota compositions, (**Table S3**) saturated models did not generally show
267 differences between any greenspace buffers nor urban-rural categories apart from a nominally
268 significant association with 800m and rare microbiota composition (p=0.02). However, when
269 considering genus differences, there were statistically significant differences observed according to
270 urban-rural status and across quartiles of all neighbourhood sizes (**Table 2**). Only 3000m greenspace
271 buffers showed statistically significant differences across quartiles in microbiota compositions in
272 saturated models, with post-hoc pairwise comparisons suggesting the association was primarily driven
273 by the fourth quantile representing the highest amount of greenspace. Comparison of the relative
274 variance explained by each included variable (marginal model) suggested that whilst overall variance
275 explained was low, greenspace composition explained nearly as much as diet (HEI) and more than the
276 included measures of socioeconomic status and health deficit (**Figure 1i).** Sensitivity analysis where
277 one twin was removed at random in each pair suggested diminished associations, with few models
278 passing the calculated multiple testing threshold of 0.004 (Sensitivity analysis S1).

279

280

281 **---- Figure 1 here -----**

Differences in abundance of ASVs and genus were observed by urbanicity and greenspace

Results of DESeq2 tests of differences of ASV (n=297) and genus (n=87) abundances suggested several associations – particularly in adjusted models (Figure1.ii-iii). The clearest signal was the suggested increase in abundance of *Escherichia/Shigella* with urbanicity, which was replicated across models of adjusted and unadjusted ASVs and genus (Genus, adjusted: logFC = 0.73742, padj < 0.001). Genera annotated as *Streptococcus* and *Prevotella_9* also increased in association with urbanicity, whereas *Ruminiclostridium* also decreased in unadjusted models, but these signals were not replicated in adjusted models. Only nominal significance was observed in unadjusted ASV models of greenspace, whereas ASVs were associated with adjusted models of 3000m and 5000m greenspace buffer, including one annotated at species level *Roseburia inulinivorans* with 3000m (logFC= 0.19951, padj=0.03957) and *Lachnoclostridium* with 5000m; the latter was replicated when unadjusted genus models, but not adjusted models (Figure1ii-iii. Table S4). Other genera were observed in association with all three greenspace buffers in adjusted models; *Haemophilus* decreased in association with 800m greenspace.; *Escherichia/Shigella* decreased in association with 3000m greenspace and *Ruminococcaceae_UCG_014* increased; *Anaerostipes* increased in association with 5000m greenspace (Figure1iii, Table S4)

Modest twin pair differences in greenspace and urbanicity reflect observations of difference in microbiota composition

There were 975 twin pairs within the subset used for this analysis. Within-pair difference in microbiota composition was associated with within-pair difference in 800m greenspace at factor level (Twin pair regression, factor level significance in association with greatest difference, β= 0.08, p= 0.0162, but comparison with null model not statistically significant at p=0.055), but not RUC or other greenspace buffers. This suggests a small, statistically significant difference in microbiota comparison between twins where one lives in an area of low greenspace immediate to their home compared with the co-twin who lives in an area of high greenspace immediate to their home. The only other variable

that was statistically significant in these twin pair models apart from sequencing depth was within-pair difference in IMD.

Of the five genus that were significant in the population-wide analysis, *Escherichia/Shigella* abundances were statistically significantly different between twins discordant for rural-urban living (Wilcox rank, paired, V = 125, p-value = 0.02) (Figure 2).

---Figure 2 here---

Discussion

Subtle differences of microbiota in association with the environmental composition around an individual's home postcode were observed. The microbiota signature of those with a greater neighbourhood exposure to greenspacewas distinct from other individuals, and a difference in ASV and genus abundance was observed particularly between rural and urban dwelling individuals. Differences in microbiota composition and geographic factors have previously been cited. In a study of traditional populations in India, Dehingia and colleagues observed microbiota variation differed within geographic region (Dehingia et al., 2016). He *et al.* note regional differences explained the greatest variance in their study of microbiota divergence between districts of China (He et al., 2018). In this study's PERMANOVA analysis, findings of He and colleagues were not replicated, instead observing factors other than environmental influences to explain a greater variance of microbiota composition. This may in part be explained by differences in how covariates were measured, such as diet which was included as a single composite measure whereas He and colleagues included 19 different dietary 'items' which by themselves would individually capture less variance than a composite. Reflecting this, and as observed previously in this data (Bowyer et al., 2019), adonis-R^2 values are low which suggests that whilst statistically significant, variation explained by all factors is low. Indeed, because of the high inter-individual variation in the microbiota, measured factors in

general explain low proportions of microbiota variance and so should be considered relatively to one another (Falony et al., 2016; Rothschild et al., 2018).

Compositional differences in rarer species were associated with local greenspace availability. This could indicate an increase of exposure to rarer microbes in association with more greenspace (although compositional differences assessed in this manner do not reflect directional differences), but this was not reflected in subsequent twin-pair analysis, and importantly did not pass the multiple testing threshold in saturated models. This suggests that other confounders could explain a degree of variance associated with greenspace and microbiota. Area-level socioeconomic deprivation, included in saturated models via the Index of Multiple Deprivation (IMD) could in part explain this as IMD and greenspace are correlated, and has previously been cited as a confounder to studies of greenspace and health. Unpicking this should be the focus of future work. Of particular interest in the twin-pair models was that difference in IMD, along with 800m greenspace difference, was significantly associated with twin dissimilarity. Importantly, being a composite area-based measure, the IMD does not necessarily reflect the socioeconomic status of an individual living in an area (this is known as the 'ecological fallacy') but rather the area-level socio-economics at the population level (i.e. environmental deprivation). Considered together therefore, this strengthens the suggestion that the physical environment might exert pressures that shape the microbiota.

Little compositional difference by explicitly measured urbanicity (as opposed to high greenspace exposure) was observed, despite the existence of previous evidence to suggest the microbiota differs between rural and urban environments. For example, a comparison between Russian cities and villages observed differences in the associated microbiota, with the rural microbiome being more health associated (Tyakht et al., 2014). The difference between urban and rural living in the UK however may be small compared to the environment in that study. Nevertheless, the increase of *Escherichia/Shigella* in association with urbanicity both in population-wide and twin pair models does suggest subtle differences in microbiota between urban and rural individuals. *Escherichia* and *Shigella* are genetically closely related pathogens (and not differentiable by a 16s rRNA marker gene approach) and whilst these species are commonly observed as members of the commensal microbiota,

they are known pathogens and generally not associated with health. The increased colonisation in urban individuals could suggest that the resistance to pathogens conferred by a healthy microbiota is diminished in these areas.

The strengths of this study lie in the investigation of an understudied influencer of microbiota variance using samples taken from a large and well characterised population cohort integrated with environmental measures computed within a geographical information system, and the use of twin structure to do so. There are however several limitations. Sensitivity analysis re-running PERMANOVAs revealed that whilst nominal statistical significance remains and R^2 values are similar, the results no longer pass our multiple testing threshold of 0.004. This may be due to the reduction in numbers, but we cannot exclude that there might be an effect of twin relatedness. Where the outcome of interest is only moderately heritable as in the case of the microbiota (Rothschild et al., 2018) the twin design has limited power and thus analysis is paired with a population-wide approach. Observations were weak and certain geographic variables of interest that may correlated with greenspace and therefore act as confounders were not included. For example, air pollution could promote negative health consequences by influencing the microbiota; exposure to nitro-polycyclic aromatic hydrocarbons through diesel fuel emissions have been suggested to promote nitroreduction by the bacteria that results in a microbiota dependent increase in carcinogenesis (Claus et al., 2016). Whilst the complex interdependency between socioeconomics, health and environmental exposure is partially mitigated via the use of twin pair and adjustment this may not fully account for these potential confounders, and inferences are limited by use of static data to capture environmental exposure. Dynamic data reflecting the behaviour of an individual, and direct measurement of their use and time spent in greenspace would address some of the uncertainties raised here. A further limitation was the inclusive definition of "greenspace" as being any area of non-urban environment, which therefore does not consider factors such as the accessibility of land or the types of habitat present. A final limitation is the use of 16S rRNA data, which whilst appropriate for initial observational studies such as this, gives less insight than whole genome sequencing data which would better be able to

confirm the differences in microbial-induced immunoregulation between individuals via functional annotation, and additionally allow capture of non-bacterial microorganisms.

Conclusion

This study has observed difference in the microbiota of twin pairs with different environmental exposures, suggesting elements of the host's habitat may contribute to the complex compositions of microbial assemblages. Considered together, results are suggestive that there are geographic patterns in the microbiota observable in this dataset which do not seem to be accounted for by diet, BMI and frailty. Modest microbiota variance may be attributable to greenspace within a close to moderate distance to an individual's home and urbanicity. Findings highlight the potential importance of considering non-lifestyle factors that influence the composition of the microbiota and add to the literature exploring the microbiome-environment-health axis.

Figure legends

Figure 1.i. Comparison of adonis-R^2 values from saturated Genera PERMANOVA (adonis) marginal model, demonstrating the comparative variance of each variable in a saturated model. Stars indicate statistical significance at 2000 permutations: ***= p<0.0001, ** = p < 0.001, *= p< 0.004. ii. Volcano plots of differences in abundance of ASVs adjusted for sample sequence depth, antibiotics in the month prior to sample collection, Body Mass Index, frailty, age, diet, index of multiple deprivation, highest educational attainment and region. Log fold change (logFC) represent change in relation to A. Urban living, B. Higher greenspace percentage in 800m, C. Higher greenspace percentage in 3000m and D. Higher greenspace percentage in 5000m. Grey point = non-significant (n.s), blue point = nominal significance (p<0.05) and red point = fdr-adjusted significance (q<0.05) **iii.** Differences in abundance of genus adjusted as in ii.

Figure 2. *Escherichia/Shigella* variance stabilised abundances in twins discordant for rural-urban living. Dashed lines link pairs.

Supplementary figure legends

Figure S1. Correlation of percentage greenspace within 800m, 3000m and 5000m of an individual's residence.

Figure S2. Estimate greenspace percentage within 800m, 3000m and 5000m of an individual's residence.

428

Declarations

Ethics approval and consent to participate

Favourable ethical opinion was granted by the formerly known St. Thomas' Hospital Research Ethics Committee (REC). Following restructure and merging of REC, subsequent amendments were approved by the NRES Committee London—Westminster (TwinsUK, REC ref: EC04/015, 1 November 2011); use of microbiota samples was granted NRES Committee London—Westminster (The Flora Twin Study, REC ref: 12/LO/0227, 1 November 2011).

Availability of data and material

The environmental exposures used in this paper were calculated thanks to several open-source geographic datasets, as follows:

- The Rural-Urban classification:
    - England/Wales 2011 (Office for National Statistics): https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification
    - Scotland 2016 (Scottish Government): https://www.gov.scot/publications/scottish-government-urban-rural-classification-2016/
- OS open roads 2017 (Ordnance Survey): https://www.ordnancesurvey.co.uk/business-government/products/open-map-roads
- 25m^2 Land Cover Map of Great Britain 2015 (UK Centre for Ecology & Hydrology): https://www.ceh.ac.uk/services/land-cover-map-2015

- The Index of Multiple Deprivation
    - English (2019): https://www.gov.uk/government/statistics/english-indices-of-deprivation
    - Scottish (2016): https://www2.gov.scot/Topics/Statistics/SIMD
    - Welsh (2019): https://statswales.gov.wales/Catalogue/Community-Safety-and-Social70 Inclusion/Welsh-Index-of-Multiple-Deprivation/WIMD-2019

The European Bioinformatics Institute (EBI) accession numbers for the microbial DNA sequences reported in this paper is ERP015317.

The processed ASV data, along with phenotypic data that can be used to recreate this analysis is available following reasonable request to the TwinsUK data access committee. Information on data access and how to apply is available at http://www.twinsuk.ac.uk/data-access/submission-procedure-2/. Please contact the corresponding author for further detail.

Funding

TwinsUK receives funding from the Wellcome Trust (WT081878MA), the National Institute for Health Research (NIHR) Clinical Research Facility at Guy's & St Thomas' NHS Foundation Trust and NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. This work was also supported by the Chronic Disease Research Foundation.

Acknowledgements

The authors of this paper wish to express our appreciation to all study participants of the TwinsUK cohort for donating their samples and time. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre

based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

We thank Dr. Julia K Goodrich, Dr. Ruth E Ley and the Cornell technical team for generating the microbial data..

References

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Belzer, C., Chia, L.W., Aalvink, S., Chamlagain, B., Piironen, V., Knol, J., de Vos, W.M., 2017. Microbial Metabolic Networks at the Mucus Layer Lead to Diet-Independent Butyrate and Vitamin B12 Production by Intestinal Symbionts. MBio 8, e00770-17. https://doi.org/10.1128/mBio.00770-17

Bowyer, R.C.E., Jackson, M.A., Le Roy, C.I., Lochlainn, M.N., Spector, T.D., Dowd, J.B., Steves, C.J., 2019. Socioeconomic status and the gut microbiome: A twinsuk cohort study. Microorganisms 7. https://doi.org/10.3390/microorganisms7010017

Bowyer, R.C.E., Jackson, M.A., Pallister, T., Skinner, J., Spector, T.D., Welch, A.A., Steves, C.J., 2018. Use of dietary indices to control for diet in human gut microbiota studies. Microbiome 6. https://doi.org/10.1186/s40168-018-0455-y

Callahan, Benjamin J, McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–3. https://doi.org/10.1038/nmeth.3869

Callahan, Ben J, Sankaran, K., Fukuyama, J.A., McMurdie, P.J., Holmes, S.P., 2016. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. F1000Research 5, 1492. https://doi.org/10.12688/f1000research.8986.2

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336. https://doi.org/10.1038/nmeth.f.303

Claus, S.P., Guillou, H., Ellero-Simatos, S., 2016. The gut microbiota: A major player in the toxicity of environmental pollutants? npj Biofilms Microbiomes. https://doi.org/10.1038/npjbiofilms.2016.3

Cresci, G.A., Bawden, E., 2015. Gut Microbiome: What We Do and Don't Know . Nutr. Clin. Pract. 30, 734–746. https://doi.org/10.1177/0884533615609899

508 Dehingia, M., Thangjam devi, K., Talukdar, N.C., Talukdar, R., Reddy, N., Mande, S.S., Deka, M., Khan, M.R., 2016. Gut
509 bacterial diversity of the tribes of India and comparison with the worldwide data. Sci. Rep. 5, 18563.
510 https://doi.org/10.1038/srep18563

511 Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer,
512 M., Vandeputte, D., Tito, R.Y., Chaffron, S., Rymenans, L., Verspecht, C., De Sutter, L., Lima-Mendez, G., D'hoe,
513 K., Jonckheere, K., Homola, D., Garcia, R., Tigchelaar, E.F., Eeckhaudt, L., Fu, J., Henckaerts, L., Zhernakova, A.,
514 Wijmenga, C., Raes, J., 2016. Population-level analysis of gut microbiome variation. Science (80-.). 352, 560–564.
515 https://doi.org/10.1126/science.aad3503

516 Geva-Zatorsky, N., Sefik, E., Kua, L., Pasman, L., Tan, T.G., Ortiz-Lopez, A., Yanortsang, T.B., Yang, L., Jupp, R., Mathis,
517 D., Benoist, C., Kasper, D.L., 2017. Mining the Human Gut Microbiota for Immunomodulatory Organisms. Cell 168,
518 928-943.e11. https://doi.org/10.1016/J.CELL.2017.01.022

519 Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome Datasets Are Compositional: And
520 This Is Not Optional. Front. Microbiol. 8, 2224. https://doi.org/10.3389/fmicb.2017.02224

521 Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G.,
522 Ley, R.E., 2016. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 19, 731–743.
523 https://doi.org/10.1016/j.chom.2016.04.017

524 He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., Chen, M.-X., Chen, Z.-H., Ji, G.-Y., Zheng, Z.-D.-X.,
525 Mujagond, P., Chen, X.-J., Rong, Z.-H., Chen, P., Lyu, L.-Y., Wang, X., Wu, C.-B., Yu, N., Xu, Y.-J., Yin, J., Raes,
526 J., Knight, R., Ma, W.-J., Zhou, H.-W., 2018. Regional variation limits applications of healthy gut microbiome
527 reference ranges and disease models. Nat. Med. 24, 1532–1535. https://doi.org/10.1038/s41591-018-0164-x

528 Hillsdon, M., Coombes, E., Griew, P., Jones, A., 2015. An assessment of the relevance of the home neighbourhood for
529 understanding environmental influences on physical activity: how far from home do people roam? Int. J. Behav. Nutr.
530 Phys. Act. 12, 100. https://doi.org/10.1186/s12966-015-0260-y

531 Jackson, M.A., Goodrich, J.K., Maxan, M.-E., Freedberg, D.E., Abrams, J.A., Poole, A.C., Sutter, J.L., Welter, D., Ley,
532 R.E., Bell, J.T., Spector, T.D., Steves, C.J., 2016. Proton pump inhibitors alter the composition of the gut microbiota.
533 Gut 65, 749–56. https://doi.org/10.1136/gutjnl-2015-310861

534 Litvak, Y., Bäumler, A.J., 2019. The founder hypothesis: A basis for microbiota resistance, diversity in taxa carriage, and
535 colonization resistance against pathogens. PLOS Pathog. 15, e1007563. https://doi.org/10.1371/journal.ppat.1007563

536 Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with

DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Magnúsdóttir, S., Ravcheev, D., de Crécy-Lagard, V., Thiele, I., 2015. Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. Front. Genet. 6, 148. https://doi.org/10.3389/fgene.2015.00148

Marchesi, J.R., Ravel, J., 2015. The vocabulary of microbiome research: a proposal. Microbiome 3, 31. https://doi.org/10.1186/s40168-015-0094-5

McMurdie, P.J., Holmes, S., Kindt, R., Legendre, P., O'Hara, R., 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One 8, e61217. https://doi.org/10.1371/journal.pone.0061217

Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A.T., Lewis, C.M., 2015. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. 6, 6505. https://doi.org/10.1038/ncomms7505

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2015. vegan: Community Ecology Package. R package version 2.0-10. URL: https://CRAN.R-project.org/package=vegan

Ordiz, M.I., May, T.D., Mihindukulasuriya, K., Martin, J., Crowley, J., Tarr, P.I., Ryan, K., Mortimer, E., Gopalsamy, G., Maleta, K., Mitreva, M., Young, G., Manary, M.J., 2015. The effect of dietary resistant starch type 2 on the microbiota and markers of gut inflammation in rural Malawi children. Microbiome 3, 37. https://doi.org/10.1186/s40168-015-0102-9

Palarea-Albaladejo, J., Martín-Fernández, J.A., 2015. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. 143, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

Robinson, J.M., Jorgensen, A., 2020. Rekindling old friendships in new landscapes: The environment–microbiome–health axis in the realms of landscape research. People Nat. 2, 339–349. https://doi.org/10.1002/pan3.10082

Rook, G.A., 2013. Regulation of the immune system by biodiversity from the natural environment: an ecosystem service essential to health. Proc. Natl. Acad. Sci. U. S. A. 110, 18360–18367. https://doi.org/10.1073/pnas.1313731110

Rook, G.A.W., Lowry, C.A., Raison, C.L., 2013. Microbial "Old Friends", immunoregulation and stress resilience.

Evolution (N. Y). 46–64. https://doi.org/10.1093/emph/eot004

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., Shilo, S., Lador, D., Vila, A.V., Zmora, N., Pevsner-Fischer, M., Israeli, D., Kosower, N., Malka, G., Wolf, B.C., Avnit-Sagi, T., Lotan-Pompan, M., Weinberger, A., Halpern, Z., Carmi, S., Fu, J., Wijmenga, C., Zhernakova, A., Elinav, E., Segal, E., 2018. Environment dominates over host genetics in shaping human gut microbiota. Nature 555, 210–215. https://doi.org/10.1038/nature25973

Schmidt, T.S.B., Raes, J., Bork, P., 2018. The Human Gut Microbiome: From Association to Modulation. Cell 172, 1198–1215. https://doi.org/10.1016/j.cell.2018.02.044

Searle, S.D., Mitnitski, A., Gahbauer, E.A., Gill, T.M., Rockwood, K., 2008. A standard procedure for creating a frailty index. BMC Geriatr. 8, 1–10. https://doi.org/10.1186/1471-2318-8-24

Shreiner, A.B., Kao, J.Y., Young, V.B., 2015. The gut microbiome in health and in disease. Curr. Opin. Gastroenterol. 31, 69–75. https://doi.org/10.1097/MOG.0000000000000139

Stamper, C.E., Hoisington, A.J., Gomez, O.M., Halweg-Edwards, A.L., Smith, D.G., Bates, K.L., Kinney, K.A., Postolache, T.T., Brenner, L.A., Rook, G.A.W., Lowry, C.A., 2016. The Microbiome of the Built Environment and Human Behavior: Implications for Emotional Health and Well-Being in Postmodern Western Societies. Int. Rev. Neurobiol. 131, 289–323. https://doi.org/10.1016/BS.IRN.2016.07.006

Stekhoven, D.J., Buhlmann, P., 2012. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tringe, S.G., Hugenholtz, P., 2008. A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol. 11, 442–446. https://doi.org/10.1016/J.MIB.2008.09.011

Twohig-Bennett, C., Jones, A., 2018. The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. Environ. Res. 166, 628–637. https://doi.org/10.1016/j.envres.2018.06.030

Tyakht, A. V, Alexeev, D.G., Popenko, A.S., Kostryukova, E.S., Govorun, V.M., 2014. Rural and urban microbiota. Gut Microbes 5, 351–356. https://doi.org/10.4161/gmic.28685

Eryk Walczak (2021). PostcodesioR: API Wrapper Around 'Postcodes.io'. R package version 0.3.0. URL: https://CRAN.R-project.org/package=PostcodesioR.

Weersma, R.K., Zhernakova, A., Fu, J., 2020. Interaction between drugs and the gut microbiome. Gut 69, 1510 LP – 1519.

592 https://doi.org/10.1136/gutjnl-2019-320204

593

594