BIOC 3301 Project Report

A Tale of Two Parks: Biogeographic patterns of microbial diversity in urban park soils, a comparative study.

K. Anonymous Dickens

Abstract

The soil microbiome is one of the greatest sources of biological diversity. Recent advances in Next Generation Sequencing and bioinformatics have enabled us to start gaining insight into the structure of below-ground communities. Here, we explore the effects of urbanisation on microbial composition and diversity. We assess the bacterial and archaeal communities of Gordon Square in London, and compare them with those of Central Park in New York City described by a previous study. Despite the lower number of distinct phylotypes in Gordon Square, both parks were found to contain similarly structured communities. Highly predictable microbial diversity patterns based on physicochemical soil properties were found in Central Park, but not in Gordon Square, which was correlated with a lower variability of *Acidobacteria* abundance across samples. These results point at the impact that different management regimes can have on microbial communities exposed to anthropogenic pressures.

Introduction

Soil is one of the richest and most biodiverse environments on Earth. It contains organisms from all trees of life (Archaea, Bacteria and Eukarya)(1,2) which carry out essential roles in their ecosystems, such as nitrogen fixation or degradation of organic matter (3). Despite its importance, the role of microbial diversity in specific environmental dynamics remains largely unexplored, as well as the exact composition of the microbial communities. Recent developments in sequencing technology, such as Next Generation Sequencing (4), together with the development of powerful bioinformatics tools, have provided unprecedented insight into the complex structure of these below-ground communities. The collection of metagenomic and metatranscriptomic information into open databases, together with powerful statistical data analysis tools, enables us to place every new finding in a global context and to start finding significant patterns that push the boundaries of our understanding of microbial ecology.

The Earth Microbiome Project (5) is an international collaborative project that attempts to survey the global microbial diversity. It provides a series of standardised methods for the collection and analysis of metagenomic data to crowd-source a comprehensive and robust exploration of biogeographic patterns of microbial populations. Most of the studies tied to the Earth Microbiome Project have focused on natural environments, and there is little information on the microbial structure of urban soils, and how they compare to different habitats across Earth.

A 2014 study (6) aimed to address this issue by surveying the soil diversity of 596 samples across Central Park in New York, and then comparing 52 of them to 52 samples from a global dataset obtained from two previous studies (7,8). Very similar patterns of diversity were found between the Central Park samples and the global database, and concluded that the mosaic nature of the environments of the park and the diverse management regimes of each promoted this range of microbial diversity.

The aim of this report is to further explore the effect of anthropogenic pressures in urban soil diversity by comparing the microbial communities of two urban parks in two of the largest metropoles in the world: Central Park in New York and Gordon Square in London, and confirm whether the patterns found in Central Park are common to other urban environments, or a result of the particular management regime of the park.

Materials and Methods

a) Soil collection and measurements.

On the same day, 27 soil samples were collected from different locations across Gordon Square (Figure S1) at a depth of approximately 10cm. The locations were heterogeneous and included lawn, near trees, shrubs, or herbaceous areas, with varying degrees of soil moisture. All samples were selected from areas with minimal footfall. For each soil sample, the pH, nitrogen, potassium and phosphorus levels were measured using the Hanna's Quick Soil Test Kit HI3895.

b) Genomic DNA extraction and amplification of the 16S rRNA V4 region

Each soil sample was split into three 250mg biological triplicates to reduce the impact of potential experimental bias. The PowerSoil® DNA Isolation Kit by MoBio was used to extract the genomic DNA, and each resulting DNA isolate was split into triplicates before PCR amplification.

The V4 region of the 16S rRNA was amplified using the updated primers from the Earth Microbiome project (5) (fwd-barcoded: 515FB-806RB) whose complementary sequences are: FWD:GTGYCAGCMGCCGCGGTAA; REV:GGACTACNVGGGTWTCTAAT. These have been optimised to minimise biases against specific taxa and allow for accurate taxonomic assignment of sequences. A linker sequence, pad, barcode, and adapter were added in to later carry out Illumina Sequencing. PCR amplification was performed in a volume of 25µl using BioMix™ from Bioline following manufacturer's instructions: adding 5pmol of each primer and 1µl of undiluted template. After an initial denaturation (99°C, 1min), 25 cycles were carried out each including denaturation (94°C, 15 s), annealing (50°C, 45 s), and elongation (72°C, 30 s). Ending with a final elongation step (72°C, 5min).

The PCR triplicates were pooled together before amplicon purification with QIAquick (Qiagen), and then split again for DNA quantification to ensure a similar concentration in all triplicates. The SpectraMax Quant AccuClear Nano dsDNA Assay Kit was used for fluorometric quantification of PCR product. Finally, all triplicates derived from the same original soil sample were pooled together at equimolar proportions (1nM).

c) Sequencing

Prior to sequencing, library quality was assessed with the Agilent D1000 ScreenTape System. Sequencing was carried out at the UCL Institute of Neurology using the Illumina MiSeq System at a concentration of 10pM.

d) Computational analysis of sequencing results

Data analysis was carried out using the using the Quantitative Insights into Microbial Ecology (QIIME) 1.9.1 pipeline (9). The raw forward and reverse read data was paired, and then demultiplexed and quality filtered to include only reads of Phred quality score >Q19. Operational Taxonomic Unit (OTU) picking was carried out at 97% similarity with closed reference and open reference using the reference databases Greengenes 13.8 (10) and Silva 128 (11). The UCLUST algorithm (12) was used to cluster sequences into OTUs, and select representative sequences, which were aligned using PyNAST (13). The FASTTREE algorithm (14) was used to generate a maximum-likelyhood phylogenetic tree.

Diversity metrics were calculated after applying a rarefaction depth of 39504 minimum sequences per sample to ensure even sampling (Table 1, Figure S2). Alpha diversity was computed using several metrics (Chao 1 estimator, observed OTUs, PD whole tree, Shannon index and Simpson index). Beta diversity was determined by computing weighted and unweighted UniFrac distances (calculated through relative branch lengths in the phylogenetic tree) (15) , and were then interpreted through principal coordinates analysis (PCoA) on the basis of several metadata features, including soil characteristics such as moisture, Nitrogen, Phosphorus, Potassium, and pH.

Finally, t-tests were performed to determine the correlation between variables, such as the between incidence of novel OTUs and proportional abundance of a phylotype, or between the soil pH and the abundance of *Acidobacteria*. We applied a significance threshold of 0.05 with Bonferroni correction (0.05 / 27 = 0.0018).

e) Full data and results

New York Central Park data was obtained from (6).

Gordon Square data and scripts used are available in Github: github.com/Kharyb/3301-Metagenomics

Results

Sequencing generated a total of 8.6 million read pairs (2x 150bp) with high quality (~91% of bases had Phred qualities >Q30). After quality filtering 3.5 million sequences remained. Taxonomic profiling with 97% sequence similarity to the Greengenes or Silva database was performed with closed and open reference. The results are summarized in table 1.

	Greengenes 13.8		Silva 128	
	Closed ref.	**Open ref.**	**Closed ref.**	**Open ref.**
Sequences conserved	2.9 million	3.3 million	3 million	3.3 million
Max sequences/sample	0.19 million	0.22 million	0.20 million	0.22 million
Min sequences/sample	34514	39504	34859	39477
Mean	107505.741	124137.9	111590.6	124129.444

Table 1. OTU picking results. The value in red was chosen as the rarefaction depth in further downstream analysis.

Closed reference OTU picking against Greengenes successfully mapped 83% of the sequences to a total of 11941 OTUs. Silva performed slightly better, assigning taxonomy to 86% of the sequences to 16176 OTUs. This highlights the fact that Silva 128 is a more complete and recently updated database, however, for the purpose of maintaining a methodologically homogeneous comparison with the Central Park study, Greengenes was used in the analyses described in this paper.

A total of 3 archaeal phyla and 49 bacterial phyla were successfully detected. However, these only matched 17.8% of the OTUs in the samples (at a 97% or greater sequence identity) (Figure 1A). This result is similar to that obtained in Central Park (Figure 1B).

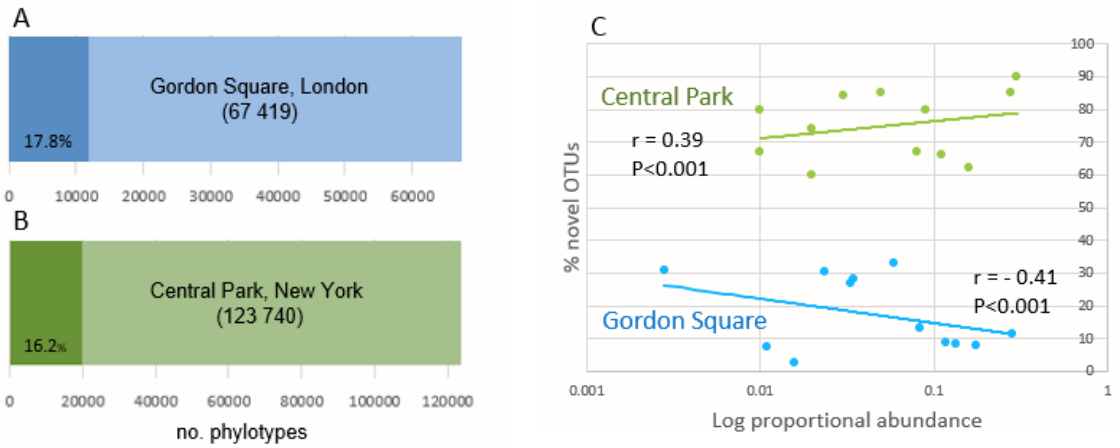

Figure 1. Urban soils harbour high levels of undescribed biodiversity. A) Only 17.8% of the bacterial and archaeal species found in Gordon Square and **B)** 16.2% of those found in New York matched the Greengenes database (Adapted from (6)). **C)** In each park, opposite correlations were found between abundance of a phyla and the amount of novel OTUs found belonging to that phyla: A negative correlation in Gordon Square (r = -0.41, P<0.001), and a positive correlation in Central Park (r = 0.39, P<0.001).

Notably, there was a strong correlation between the abundance of a phylum and the amount of novel OTUs found to correspond to it (Figure 2C). This correlation was found to be statistically significant (P<0.001) for both Central Park and Gordon Square samples, however Central Park data indicates

a positive correlation (r = 0.39, P<0.001) and a much higher average percentage of novel OTUs, while Gordon Square data indicates a negative correlation (r = -0.41, P<0.001) and a lower average percentage of novel OTUs for each phylum.

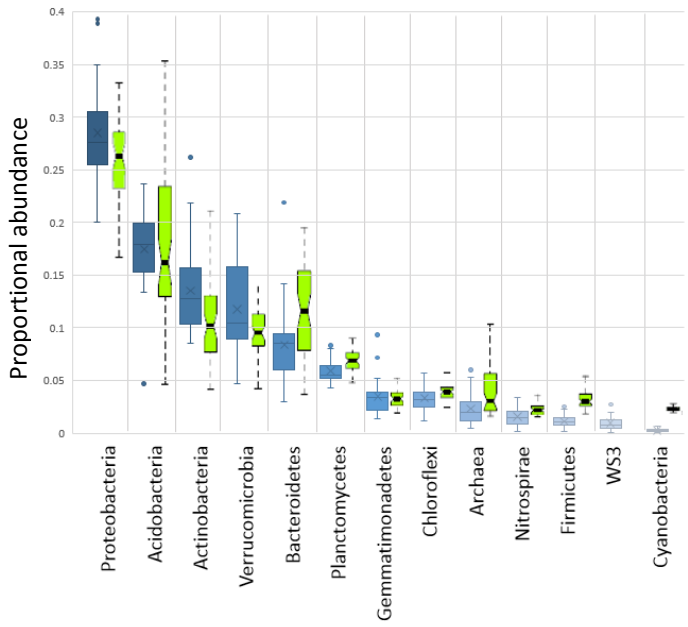

Figure 2. Relative abundances of the dominant bacterial and archaeal phyla. Soil diversity and abundance in Gordon Square (blue gradient) is similar to that of Central Park (green). WS3 was not reported in Central Park but was the 12th most abundant bacterial species in Gordon Square.

Gordon Square is home to large numbers of bacteria and archaea, which was evidenced in the high alpha diversity results obtained, however Central Park was found to have almost 2 times more species: 67,419 different bacterial and archaeal species were found in Gordon Square, and 123,740 were found in central park (Figure 1 A and B).

Of the 50 bacterial phyla detected in Gordon Square, the most abundant were: *Proteobacteria, Acidobacteria, Actinobacteria, Verrucomicrobia,* and *Bacteroidetes.* These were also the most abundant Phyla in Central Park. In both geographies the proportional abundance and variability across samples was found to be similar (Figure 2), with the exceptions of *Bacteroidetes*, whose mean abundance in Gordon Square was lower, *Acidobacteria*, whose abundance although similar had a much lower variability, and Archaea whose average abundance was lower than that found in Central Park (Figure 2).

Despite the high degree of variability, no significant clustering was found by PCoA of the UniFrac distances for any of the soil characteristics measured. This contrasts with the high predictability of the microbial communities found in Central Park, for which the main indicator was pH. In particular, a strong correlation (r = -0.50, P<0.001) was found between the abundance of *Acidobacteria* and the soil pH in Central Park (Figure 3A), however no such correlation was found in Gordon Square (Figure 3B).

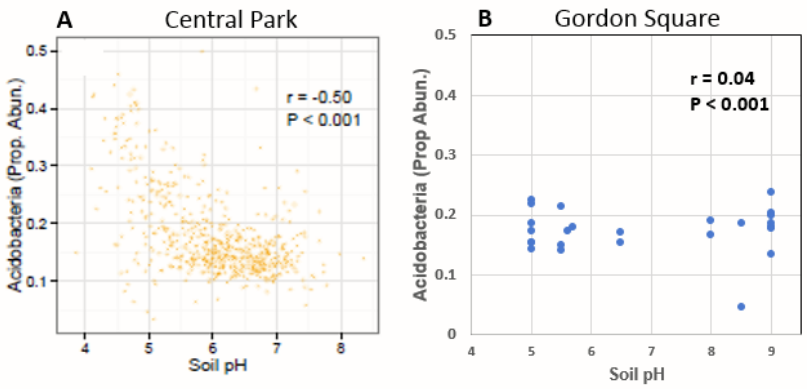

Figure 3. Correlation between soil pH and relative abundance of *Acidobacteria* in both parks. A) (Figure on the left obtained from supplementary materials (6)) The strong negative correlation between pH and abundance of *Acidobacteria* found in Central Park (r = -0.50, P<0.001) was not found in **B)** Gordon Square (r = 0.04, P<0.001).

Discussion

Despite the efforts of the Earth Microbiome Project, soil diversity of bacterial and archaeal species is largely unexplored (16). The underrepresentation of sequenced phylotypes in the used databases is evidence of this, however sequencing of the V4 region only allows phylum level resolution, which suggests that the number of different species reported is likely to be inaccurate. A main limitation of this study is the restriction to the Greengenes database. Silva is a more recently updated and more complete database and consequently fewer novel OTUs would have been obtained with Silva (mapped 22.9% of the OTUs in Gordon Square as opposed to 17.8% with Greengenes).

Analysis of the Gordon Square taxa suggested that more abundant phylotypes are better represented in the database, yielding a lower number of novel OTUs, while the rare taxa that are found at lower abundances are more likely to be represented by a novel OTU than a mapped one. Analysis of the same taxa abundances and relative novel OTUs from Central Park found that the opposite was true, and there was a positive correlation between both factors. The variation in the number of novel OTUs found could to be due to the difference in clustering algorithms used. UCLUST was used on the Gordon Square data, while USEARCH was used on Central Park data. However, expected variations in the number of novel OTUs picked with each, cannot justify the extent to which the correlations differ, and further analysis of this relationship is needed.

Both parks had highly variable abundancies of bacterial phyla as shown by the high alpha diversities, but same 11 phyla were found to be the most abundant in both parks (Figure 2). Interestingly, *Acidobacteria* was the most variable phylotype in Central Park and its abundance was highly correlated to variations in soil pH. In Central Park, pH was found to be the main determinant of bacterial community structure, a correlation that has been observed consistently in the literature (8,17). The absence of this correlation in the Gordon Square study can be explained by the low sensitivity of the pH test used, and also due to the fact that the average pH was higher in Gordon Square, than in Central Park (Figure 3). This may also explain the lower variability and lower average abundance of *Acidobacteria* found in Gordon Square given that a higher proportional abundance was found at lower pHs (Figure 3A), particularly between pHs 4 and 5, values that were not detected in any Gordon Square sample. The pH range in Gordon Square could also explain the lower abundance and diversity of Archaea: Ramirez et al (6) found that most of the phylotypes found in global samples but absent from Central Park were the ones tied to extreme pH values, which could explain the absence of these in Gordon Square.

Overall, Gordon Square is a homogeneous park with a limited variety of ecosystems and physicochemical soil properties, while Central Park is a mosaic of cover types and management regimes, with heterogeneous biogeographic patterns. Next steps should focus on determining whether the reduced diversity found in Gordon Square is caused by the limited diversity in ecosystems, or due to anthropogenic factors such as pollution. Further studies in larger parks more

akin to Central Park should be carried out, such as Hyde Park, which will allow to discern the effects of geography and urbanisation from the physicochemical soil features. Furthermore, moving beyond structural descriptions, metatranscriptomics (18) studies will allow to understand the functions of the organisms found in the soil, and the nature of their interactions between each other and with humans, especially in pathogenic processes. A better understanding of these communities can lead to the development of management regimes for promoting soil fertility (19), and even have a public health impact in matters like allergies (20).

References

1. Wall DH, Bardgett RD, Kelly E. Biodiversity in the dark. Nat Geosci. 2010 May 1;3(5):297–8.
2. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci U S A. 2012 Dec 26;109(52):21390–5.
3. Wall DH, Bardgett RD, Behan-Pelletier V HJ, Jones H, Ritz K, Six J, Strong DR van der P, WH. Soil ecology and ecosystem services. USA: Oxford University Press, editor. 2012.
4. Schöler A, Jacquiod S, Vestergaard G, Schulz S, Schloter M. Analysis of soil microbial communities based on amplicon sequencing of marker genes. Biol Fertil Soils. 2017 Jul 4;53(5):485–9.
5. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017 Nov 1;
6. Ramirez KS, Leff JW, Barberán A, Bates ST, Betley J, Crowther TW, et al. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. Proceedings Biol Sci. 2014 Nov 22;281(1795):20141988.
7. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, et al. Functional metagenomic profiling of nine biomes. Nature. 2008 Apr 12;452(7187):629–32.
8. Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol. 2009 Aug 1;75(15):5111–20.
9. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May 11;7(5):335–6.
10. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses

of bacteria and archaea. ISME J. 2012 Mar;6(3):610–8.

11. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan;41(Database issue):D590-6.

12. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19):2460–1.

13. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010 Jan 15;26(2):266–7.

14. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. PLoS One. 2010 Mar 10;5(3):e9490.

15. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011 Feb;5(2):169–72.

16. Pace NR. A molecular view of microbial diversity and the biosphere. Science. 1997 May 2;276(5313):734–40.

17. Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, et al. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J. 2010 Oct 6;4(10):1340–51.

18. Carvalhais LC, Dennis PG, Tyson GW, Schenk PM. Application of metatranscriptomics to soil environments. J Microbiol Methods. 2012 Nov;91(2):246–51.

19. Chaparro JM, Sheflin AM, Manter DK, Vivanco JM. Manipulating the soil microbiome to increase soil health and plant fertility. Biol Fertil Soils. 2012 Jul 13;48(5):489–99.

20. Lynch S V, Boushey HA. The microbiome and development of allergic disease. Curr Opin Allergy Clin Immunol. 2016 Apr;16(2):165–71.

Appendix

Figure S1. Sampling locations. A) Central Park, New York. (green markers indicate samples used in the original paper for comparison with global data). Generated with CartoDB. Obtained from (6) **B)** Gordon Square, London. Generated with Google Maps.

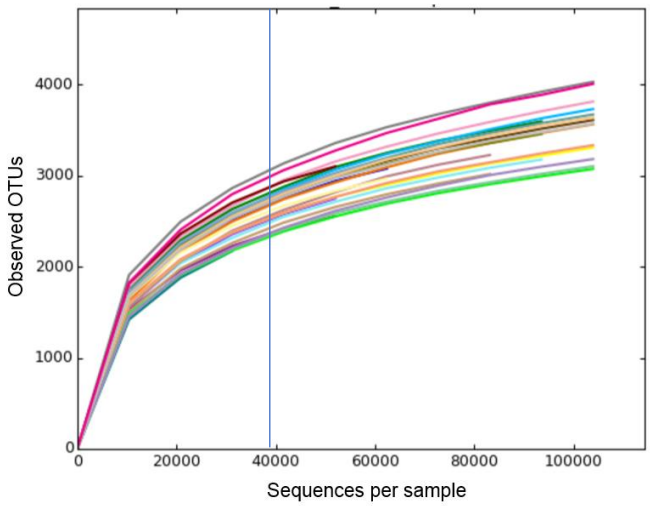

Figure S2. Alpha rarefaction curves. Different colours represent different samples. No cut-off was applied in order to show the variability in sampling depths. Most samples reach a plateau indicating that most of the microbial diversity has been sampled. Blue vertical line indicates rarefaction cut-off point applied for diversity analyses (39504 random sequences/sample).