Bioc3301-CAN WE DEFINE A SIGNIFICANT CORE MICROBIOME?

Intro

16S phenotyping

Variation of phenotype with environment

While humans are the most noticeable specie of the planet, most of the Earth is populated with very different lifeforms. Each Soil sample can contain up to hundred of thousands different bacterial and acheal species within. Because lifeforms depend on metabolites of the environment to survive, it can be expected that varying condition can modify the microbiome of soil, the exact variety of species. High concentration of one metabolite often incites microbes that can use it to proliferate, at the detriment of less adapted groups of prokaryotes.

Different prokaryotes can belong to different classification, often defined by the degree of similarity between them. General broad classification is often based on very defined traits and visible observation, leading to the 3 domains: bacteria, archea and eukaryotes. But Identifying the small differences between species cannot simply be done through phenotype, as closely related bacteria could look identical. Instead the study of the genetic variation of conserved gene gives a better idea of when species split and how similar they are. Many genes could be studied for this purpose, but the preference lies in short ribosomal RNA, as all domains contain it in one form or another. 16S rRNA forms part of the 30S subunit of prokaryotic ribosomes, and contains 9 conserved and 9 hypervariable regions. LITTLE WORD ABOUT RIBOSoMES AND WHY WE ALL HAVE THEM These hypervariable regions are so variable that they are often unique to a family or a genus, providing a way to identify bacteria from sequencing. 250bp reads of the V4 region of 16sRNA can provide enough uniqueness to identify the genus of a specie with high reliability. This can lead to the process of classification which is called taxonomy, and is composed of 9 levels in which lifeforms are slowly split in, from general life to individual species.

This metagenomics analysis of bacteria is predicated on fast and precise sequencing, often on short genes. This means NGS techniques are very suited, with Miseq being favoured, Illumina sequencing is nearly 100% reliable on short reads and can read up to 100BILLIONS in a day, proving it to be a very serious analysis tool.

Overall, microbes are known to vary and their abundance or presence can change dependent on the conditions. When metabolites, pH, dampness and other factors favour a specie, it will thrive and develop. With different conditions therefore comes different microbiomes, which allows this paper to ponder over the presence of some species across all sample and varying conditions. In essence, the aim would be to identify microbes that show an important presence across a wide range of conditions, showing a strong ability to adapt to its environment. According to the earth wide microbiome project, a microbiome can be defined as a functional community of microbes present over a wide range conditions. <https://www.nature.com/articles/nrgastro.2017.97>. Microbiome analysis across different conditions could help develop a general model of the functional groups in the soil . The benefits are also two folds, as understanding what specie is thriving can help understand the soil, and the exact conditions it is in. Abundance based observation of the most relevant microbiome could also help reveal the correlation between certain conditions, such as pH, and certain phylum of microbes.

SPECIES VARY.BUT NOT ALWYS (USE PRECISE EXEMPLES)-CAN WE EXPECT A CORE MICROBIOME-WHY COULD IT BE IMPORTANT

Methods and materials

Soil collection

Sample for metagenomics analysis were collected from Gordon Park, London. Sterilised tools were used to core the surface of the earth and collect half a Falcon tube’s worth of material. At the location of each 29 sample, precise geographical coordinates and observations were taken, including depth of the sample, moisture, vegetation, footfall, temperature and general aspect of the dirt. pH, phosphorus, nitrogen and potassium were also measured and recorded for each sample in the lab using the HANNA QuickSoil test.

Subsequent isolation of the gDNA from each sample was done using QIAGEN Power Soil isolation kit (Qiagen, Inc, Valencia, CA) in accordance to the manual, and verified using a 0.5% agarose gel with 3 µl of ethidum bromide

Subsequent PCR to amplify the 16SRNA region was performed in a MiniAmp Thermal Cycler. The targets for the primer were the conserved regions around the 16S SSU V4 region, to which universal primers 515f and 806r (FIG1) bind. 75 µl triplicates of the mixture (Fig1 C) were run for 25 cycles (Fig 1D). The resulting samples were run on 1.5% agarose gel with 3 µl.

20 µl of the PCR DNA was run in the QIAquick PCR Purification Kit ((Qiagen, Inc, Valencia, CA), according to protocol, then followed

The concentration of DNA was measured using Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher/Invitrogen cat. no. P11496;) following the manufacturer’s instructions

The triplicates were then combined and the A260/A280 ratio was verified.

Then, 1 µl of 10nM of each sample were pooled and sent to sequencing at the illumina machine, along with the used primers.

Further metagenomics analysis of the sequencing was instigated using Qiime and Biopython based software on the Cirrus Hspc. A maximum of 64 cores were used for any one task, through a simple BASH command line and Vim. All the scripts and the workflow can be found on my Github

GITHUB\_\_\_\_

# The model of London was made in Gimp by combining soil pH data from UK Soil Observatory, collected by indivuals and professionals between 2015 and 2018, a map of central London, taken from <https://maproom.netcentral-london-map/> and the data provided by this experiment and analysis. All the areas have at least 1 pH reading, but were often generalised based on data collected in one specific part of the park. Other soil conditions were not taken into account to create the map.

FORGET ABOUT STATITICAL TEST-REDUNDANT WITH PCOA

Soil Results

Each soil sample was tested for numerous conditions that could impact the core microbiome of the earth. Turbidity and colour tests were conducted on samples to reveal approximate pH, potassium, phosphate and nitrogen content. Results indicate a varying range (table 1) of conditions within the soil. Correlation between the results, location and other metrics (depth of the sample) were shown to be inexistent through repeated ANOVA tests, as could be expected from random metrics. From the varying results, it seems that the range of conditions represented in this experiment cover a large area, but because of the general lack of identical conditions across sample, the reliability remains uncertain

Shannon entropy to use v4? (in intro?) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412817/>

Proteobacteria make up 40% of soil <http://aem.asm.org/content/72/3/1719.full?site=ApplEnvironMicrobiol&utm_source=TrendMDApplEnvironMicrobiol&utm_medium=TrendMDApplEnvironMicrobiol&utm_campaign=trendmdalljournals_0>

OUR RAREFACTION CURVE IS SHIT: <https://link.springer.com/article/10.1007%2Fs00374-017-1205-1>

Evaluating ilumina as a method: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315398/

Rarefaction plot and core diversity

Initial Alpha diversity calculation through hSPC lead to the creation of an individual-based rarefaction plot, summarizing all the different OTU counts in each sample. The curve represents how many unique species were measured in the sample versus the number of sequenced amplicons. The lack of plateau even after 500 reads in all samples indicates that the sequencing did not reach the limit, and new OTU reads were still revealing new microbes. But because the definition of “core” includes functional, high abundance is primordial. In that case, these phylum would have been identified early in the sequencing.

The first step was to identify which species were present in the sample. Visualisation of the data in bar charts reveal that the basic composition of most samples is similar. Proteobacteria, verrumicrobia and bacteroidites show high abundance (over 10%) in every sample, with low variation (Proteobacteria ranging from 20.7% to 39%). Statistical comparison of the OTU composition reveals that these abundance differences and presence of rare OTUs are significant (P=0.172). From the OTU summary, 11 phylum come out as core as in present in 100% of the samples.

But to assess of the differences between samples, a Jackknife PCoA plotted the beta score of all the samples. This showed a strong similarity of PD beta score across species, especially for samples 30, 24, 34,18,15,17, with their variance nearly overlapping. Some sample are far away from the rest and from each other (sample 35, 36, 12 and 25), even when considering variance, indicating a strong disparity of OUT composition. But PD score is not based on abundance, therefore tends to over value the presence of sparse species that do not impact the soil as much as the important core, making the difference observed in the plot majorly due to a pure difference of over all OTUs, not necessarily of core microbes.

Thereofore the core microbes identified before cannot be considered to all be significant. As revealed by the PCoA plot, non-abundance based analysis indicates disparity within samples because of rare OTUs. But when looking at abundance, 5 main phylum appear in all samples. Crossing abundance data with the core microbes reduces the impact of rare microbes, creating a more significant list of functional microbes to be counted as core. These abundant microbes correspond to in average 85% of all phylum present in the sample, indicating that they most likely are the ones having the most impact of the soil.

But average abundance is also limited, as the conditions of the soil impact its microbial composition. Therefore, the effect of pH on abundance needed to be verified, to identify which varied more or less with acidity. When observing abundance of phylum with pH, the major core phylum still show highest abundance but low change across pH, with correlation with pH close to 0. Less abundant phylum varies, with some very variable (Nitrospirae, Chlroflexi) and some with low change (Gemmatimonadetes). What comes out is that the average correlation is relatively low (average = 0.003), but the minimum and maximum variation within a phylum is high. Less abundant phylum showed an increased rate of variation with pH, with Nitrospirae having a 0.268 average correlation with pH, while proteobacteria had a correlation of 0.07 . But within each phylum, the correlation of each genus varies greatly. Noticeably, the range of values is greatest in large phylum where the overall average correlation is close to 0, such as Firmicutes and Proteobacteria.

To assess the effects of pH on microbes in a more precise way, as to observe the variation within phylum of core classes, their correlation with pH changes were graphed. Within a very abundant phylum (as acido), \_\_\_\_ shows low abundance, while \_\_\_and \_\_\_\_ show high (\_\_\_). Furthermore the correlation with pH score also shows variation within the phylum. Microbes such as \_\_\_\_ and \_\_\_\_\_ have a score of \_\_\_, while their overall phylum has a score of \_\_\_\_. This indicates that looking at core phylum creates or more precise overview of the microbe composition, but also the changes pH have on their abundance. Splitting it up allowed to observe that some classes vary much more than their phylum, such as acidobacteria-6, with 0.133 positive correlation with pH while its phyla only had 0.07.

From there, the pH of London soil was used to create a model of microbiomes in the city. Observing the model shows that the city has numerous different conditions, even within the same park. But soil closer to industry (south west), tend to have a lower pH, and therefore a microbiome composed majorly composed of alphaproteobacteria and spartobacteria. Community parks and areas tend to have a more neutral pH, possibility due the effects of vegeration, with a core microbiome composed of acidobacteria-6 and deltaproteobacteria, with higher than average presence of thermophilia. Soils near water tended to have a higher pH, and a microbiome composed of acidobacteria-6 and alphaproteobacteria, with peak levels of thaumarcheota.

To identify each microbe, the 16S V4 sequences in illumina were compared and paired using Qiime with a corresponding data base of V4 sequences. As the variable V4 sequences are unique to each genus (at least), this provided a map of each microbe in the samples. The abundance of each sequence was also linked to the abundance of the sample. Fig2. This analysis was based on the sequenced data, which isn’t 100% reliable. Using a PHRED cut of score of Q20, or 1/1000 error rate, the V4 sequences were not reliable enough to give 100% reliability for genus identification.

This showed all the phylum present in each sample. To test the significance of the results and if each samples OTU composition differed, statistical analysis was run to observe if the microbiome of each sample differed from each other in a meaningful fashion. The returned p (p=0.0367) value indicated a significant difference for every sample, indicating mostly that no samples had exactly similar species or abundance for each specie. Fig3

But to assess of the differences between samples, a Jackknife PCoA plotted the beta score of all the samples. The beta score of samples correspond to a diversity between samples. Its a mathematical calculated score, as a ratio of local and global diversity. The beta score was measured using a phylogeny diversity (PD-Whole tree) weighting, calculated as "the sum of the lengths of all those branches that are members of the corresponding minimum spanning path”, reducing the impact of rarer OTUs. The Jackknife method of making the plot corresponded to introducing 100 random permutations to the samples beta score calculations and OTU identification, to visualize what random variance could bring. Results indicate that the beta scores varied between sample but were most often found clustered together, nearly superimposing. Fig4. This showed a strong similarity of microbe composition across species, especially for samples 30, 24, 34,18,15,17. Some sample are far away from the rest and from each other (sample 35, 36, 12 and 25), even when taking into account variance, indicating a strong disparity of OTU. But PD score is not based on abundance, therefore tends to over value the presence of sparse species that do not impact the soil as much as the important core, making the difference observed in the plot majorly due to a pure difference of over all OTUs, not necessarily of core microbes. .

Using the taxa summary, a list of phylum present in 100% of samples was easily made. WHICH ONES\_\_\_This could be taken as the core microbiome of the samples. But the Jackknife PCoA indicated that abundance was not considered in beta score calculation, which is why some species are not clustered even if they possess the same core. Therefore, simply relying on OTU presence would be misleading in creating a representative core microbiome.

To verify the assess the different abundance in core microbes, the metrics were crossed with the core phylum described before. Microbes such as Proteobacteria, Verrucomicrobia and Acidobacteria dominate the abundance charts, while others (chloroflexi, proteoacrhea) show low abundance in the samples. When obersving abundance of phylum with pH, the major core phylum still show highest abundance but low change across pH, with correlation with pH close to 0. Less abundant phylum showed an increased rate of variation with pH, with Nitrospirae having a 0.268 average correlation with pH, while proteobacteria had a correlation of 0.07 . But within each phylum, the correlation of each genus varies greatly. Noticeably, the range of values is greatest in large phylum where the overall average correlation is close to 0, such as Firmicutes and Proteobacteria. Average range of correlation within a phylum is 0.78, and the average difference to the average is 0.456, indicating strong variation within phylums.

To assess the effects of pH on microbes in a more precise way, as to observe the variation within classes, the abundance of core classes and their correlation with pH changes were graphed. Within a very abundant phylum (as acido), \_\_\_\_ shows low abundance, counteracted by \_\_\_\_. Furthermore the correlation with pH score also shows variation within the phylum. Microbes such as \_\_\_\_ and \_\_\_\_\_ have a score of \_\_\_, while their overall phylum has a score of \_\_\_\_. Splitting it up allowed to observe that some classes vary much more than their phylum, such as acidobacteria-6, with 0.133 positive correlation with pH while its phyla only had 0.07.

Abundance of core phylum with pH was therefore measured and plotted. Some pH metrics, such as 5.7, only has 1 representative sample and were deemed statistically unreliable.

Using the correlation between core microbiome composition and pH, a model of London’s core microbiome in relation with pH was set up. General soil pH was taken from the UK soil website and generalised to the area. Each soil is most likely

While proteobacteria dominate the overall phylum abundance, at pH 9, the archaea Crenarchaeota is the most abundant microbe, and unique proteoarcheas are less represented

To emphasise and enumerate the exact effect of pH on phylum, a statistical Pearson test was also used to reveal which species correlated the most with pH changes. Smaller phylum showed an increase rate of variation with pH, with Nitrospirae having a 0.268 average correlation with pH. But within each phylum, the correlation of each genus varies greatly. Noticeably, the range of values is greatest in large phylum where the overall average correlation is close to 0, such as Firmicutes and Proteobacteria. Therefore, the same analysis was run with the core classes of microbes. This allowed to get a better understanding of exact changes in microbe compoisition with pH. Splitting it up allowed to observe that some classes vary much more than their phylum, such as acidobacteria-6, with 0.133 positive correlation with pH while its phyla only had 0.07.

This observational data can therefore be combined with a map of London’s soil to create a model of the core phylum present in the soil in relation to the pH. The relatively strong correlation of several phylum (Nitrospirae, Chloroflexi) with pH indicates that there is chance these results will be representative of the actual state of the soil. As the samples came from London originally, several general conditions should be roughly identical (pollution, light levels, temperature), increasing the reliability of the data for the model, but confirming its inadequacy on a larger scale.

Discussion

Identification of a core microbiome based on abundance

What to discuss-composition of microbiome. While data is limited (rarefaction), its good enough for phylum analysis.

16S taxonomy identification is far from a perfect method. Combining with illumina sequencing produces numerous quality issues. But the rarefaction curve, while it does show that the sequencing didn’t really reach rare OTUs, confirms that the data can be used for a functional microbiome, as abundance (opposite of rarity) is an important part.

Average abundance of each phyla introduces 5 main microbes, which share 85% of the abundance of all the samples. But average abundance across all sample ignores what leads to high abundance-being suited to the environment. Therefore splitting the abundance in relation of the pH of the sample revealed relatively little difference. The 5 main phyla are still, at every pH, the most abundant ones, and their correlation with pH is low (except for Bacteroideties). A pattern appears, where phyla with high abundance over have low correlation with pH, while lesser abundant phyla tend to vary more. These phyla are also the largest phyla englobing the most species in general. Therefore, the pattern possibly due to the high amount of different classes within them rather than preferential conditions. Due to taxonomy being mostly based on similarity of genes, it leads to large classes of prokaryotes not sharing any metabolic similarity (like preference for one pH). While the 5 phyla identified are the most functional microbes present in 100% of sample, the high variance of abundance and correlation with pH make it so they offer nearly no predictive power to the model.

Bringing the same analysis to the level of core classes reveals greater variation. Because classes are a smaller level of classification, indicating more similarities between microbes, the effects of pH appear clearer. Classes such as RB\_25 show a correlation score of 0.42 while its phylum has 0.07, showing that it is much more sensitive to pH as its phylum suggest. Furthermore, because of the splitting up of the phyla, some classes in less abundant phyla are just as abundant as those in abundant core phyla. This is because the most abundant core phyla often have several abundant core classes, which add up to the high abundance, while some less abundant phyla are only composed of one. This indicates that the abundance of core classes is more representative of the effects of pH and not of larger groupings. Similarly, the correlation score with pH of the individual classes are generally above those of their respective phyla, especially for RB\_25 show a correlation score of 0.42 while its phylum has 0.07. But the pattern described before is still present. This indicated that a similar situation is most likely also present at the class level. Because a Q20 PHRED score only allows to be certain about classes, this is the limit, but it is likely that it is repeated at the order, family or even genus level. This pattern of average abundance and correlation with pH being due to genetic differences within groups that do not show similar reaction to acidity should continue until the species level. The lower down taxonomic grouping, the more similar the members are, until the level where there is just one-species. The only place where correlation of abundance with pH will be 100% due to the ability for the microbe to grow in pH is at that level. Before that, the slight differences between members of a genus will cause averaging. Further studies should be made to understand at which taxonomic level will the correlation with pH will be truly significant.

This is the main limitation of the model of London soils created. But even as a gross generalisation it can be useful

Importantly, the definition of microbiome is “functional”, therefore the chosen microbes should be both present in every sample and highly abundant. With average abundance across samples, 5 main phylum come out. But this abundance is not constant across conditions. Because ions (what they do in bacteria), there is a change of abundance. But phylums are too general of a classification, and they microbes in it are only linked by 16S RNA-not by their similar reaction to conditions. They are generally quite separated. Therefore the change of phylum with pH is hardly reliable, and doesn’t indicate much. Therefore, go down to class level. While its still limited, the correlation with pH appears much clearer. And we see that abundance of classes and phylums are not equal, with now nitrospirae or \_\_\_ having the same abundance as some proteobacteria.

From this data, the model of London made can be used to create a approximation of the microbiome. This can be used to \_\_\_\_. But there are still limitation. Non representative conditions, especially in London as the conditions are never the same (even in one park), and class not being precise enough

Initial results from the experiment indicated that the depth of the sequencing was not enough to find all the OTUs present in the sample. The lack of plateau in the rarefaction curve indicate that rarer geni were missed. But phylum are very large classes of microbes, encompassing many classes and geni, were most likely not missed. Even if some were, they must have had very low abundance, which is integral to the definitions of core microbiome, as the functionality is inherently linked to its impact on the soil, often a result of its abundance.

Looking at the microbe composition of each sample, the similarity seems striking. While statistical differences were shown, the JackKnife PCoA plot indicated that most sample sat within a similar variance of each other. This is a result of the PD whole tree method used to calculate the beta scores. Because abundance is not considered, only similarity on a phylogeny tree, the numerous shared OTUs make the samples look similar. In the same vein, some samples (35, 36, 12 and 25) sit far from others because of the presence of some rarer microbes. But the use of abundance metrics shows the limit of this method. Out of the 11 Phylum present across all 35 samples, the relative abundance varies greatly, due to the different conditions. Core phylums such as Proteobacteria, Verrucomicrobia and Acidobacteria dominate the abundance charts and could be considered the most functional core microbe in the soil. Other phylums, such as chloroflexi, show low abundance in all the samples, indicating that while it might impact the PD score as much as other core microbes, it shouldn’t have the same significance to the soil. The average abundance for a core microbe in the samples is about 2.4%, with 5 outliers sitting above 10%, and 7 sitting under 1%. This exacerbates the idea that not all core microbes are equal and have the same significance on the soil. The average abundant microbe has about 10x the numbers of less abundance phylums, indicating their importance on the soil. Nearly 85% of the soil comes from 5 phylum, and this can be considered to be the ones that have the most impact

ACTUAL NUMBERS

Creation of a model of London for microbiome

With an idea of which microbes are the most important to the soil, based on abundance, a possible model of the London soil could be made to give an idea of soil composition. But the abundance of these phylum vary a lot across composition, and the average microbiome would not be representative. Oberving the very abundance phylum, their correlation with pH is often low (average 0.10), meaning their general abundance doesn’t tend to change. But looking in each phylum at the maximum and minimum correlation, it can be understood that a large amount of variation is present. Dropping down to the class level allows to identify which microbes really vary with pH, making the results more significant in the different soils of London. Within each phylum, core classes vary greatly, both in abundance and correlation with pH. Acidobacteria RB\_25, with only a 0.8% abundance, has a 0.43 correlation with pH. Both differ greatly from its phylum (19% abundance and 0.07 correlation), indicating the variance of classes. These results also show that some dominant phylum are made up of numerous abundant classes, while others, such as Nitrospirae, are mostly represented by one. This level the field, and classes such as Thaumarcheota suddenly appear with a similar abundance to several acidobactrias, showing that they could be having a similar impact.

Furthermore, the variation with pH of the classes is also interesting. Most show an increased variation with pH then their relative phylum. Most of the classes show a preference for one extreme of the pH scale, with acidophiles (Thermophilae, Spartobacteria or acidimicrobia), basophiles (acidobacteria-6, alphaproteobacteria or thaumarcheota), or even neutrophils. This indicates that it can be possible to make a significant guess on the composition of a soil microbiome based on its pH. Low pH soils should present high amounts of acidophiles, especially the more naturally abundant ones, but also good amounts of the bacterias that don’t vary with pH, such as TK10 or Pedospharae. Therefore, crossing a map of London soil pH with this should be able to create a significant model of the microbes in the soil.

While this model can be useful, it is also inherently flawed. The use of classes to observe the core microbes presents the same issue as the use of phylum, as the taxa categories are usually to large to encompass microbes that all react similarly. As the categorisation is based on ribosome RNA similarity, there are many unpredictable variation within each classes. Therefore the limited power of illumina, allowing only 100% precision down to class, makes the model limited in terms of genuine predictive power for anything more precise than class, which is too large to really understand the impact on the soil. Furthermore, this model only looked at variation of microbes with pH, but several other conditions would modify the abundance of certain phylum. To grow, microbes desperately need metabolites, and while some can produce them or harvest them from air, most depend on the soil. Therefore, the variation recorded here only serve as an possible outline, but could be refined with variation with phosphate, nitrates or phosphorous in the soil. Furthermore, most samples were taken from one park, with already a lot of varying conditions. Because London has a lot of unique conditions (light, temperature, pollution), these core microbes and their abundance is most likely unique to this area. In terms of precise prediction, the different conditions of different parks and small scale variations of the soil would indicate that these results are only representative of one soil.

Therefore this model of London is a general approximation, which gives a gross outline of the possible microbe composition.