Bi270a Final Project

Drew Honson

5 December 2023

Introduction

Geography has a major influence on the composition of mammalian microbiomes.[1–3] One study suggests that latitude is the strongest predictor of the ratio of Firmicutes to Bacteroidetes in the human gut.[1] A study of individuals in China's Guangdong Province suggested that districts and even certain neighborhoods have characteristic microbiomes.[4] The district an individual lived in predicted their gut microbial content better than a wide variety of other markers including diet, smoking, age, and weight.[4]

In mice, a survey of caecal communities from eight sampling locations in Western Europe found that geographic location predicted microbial content better than heredity as measured by mitochondrial DNA.[3] Interestingly, however, several geographically distant locations had highly similar caecal communities as measured by principle component analysis: Divonne les Bains, Louan, and Espelette in France (Figure 1a-b). This raises the question of whether a variable beyond geography may explain the similarity of these regions.

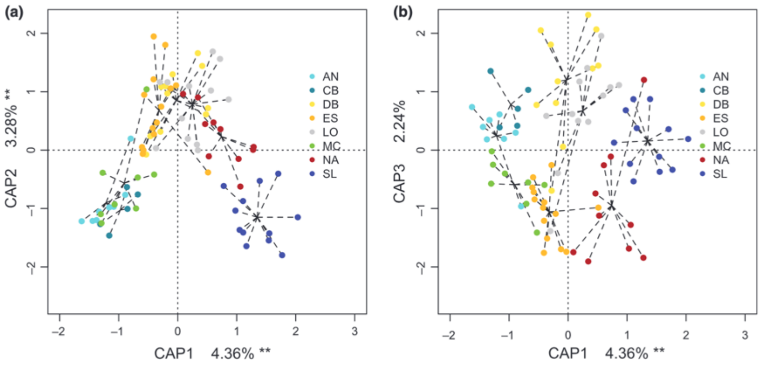

Figure 1. Figures 4a and 4b reproduced from Linnenbrink et al 2013. The visualizations are of a three-component PCA of mucosal microbiota. a) PCA visualization plotting components 1 and 2. b) PCA visualization comparing components 1 and 3.

Western Europe, and France in particular, has an incredibly high density of passenger rail.[5,6] Intuitively, it seems possible that in addition to humans these rail networks may transport rodent stowaways. If so, the practical distance between two locations may be smaller than raw geographic distance would suggest. This project aimed to determine how commuter rail distances compared to geographic distance in predicting the similarity of mouse caecal content in two locations.

Methods

The following analyses used three sources of data: metagenomic and location data from Linnenbrink et al 2013, geographic distances as calculated by the Haversine formula, and public transportation distance drawn from Google Maps' API.[3,7] All analyses were performed using Python 3.

Metagenomic data were retrieved from MGnify study MGYS00000516 using a custom script in notebook 01) pulldata.ipynb. 201 individual samples were pulled and annotated for their geographic location. I examined the distribution of read counts and the complexity of the taxonomic content and found that all samples and study locations were highly similar (notebook 02) preprocessing.ipynb). None had such dramatically divergent coverage that I thought it would impact downstream analyses. I also looked at the variation in abundance for individual taxa. One Family, Clostridiaceae, had dramatically higher variance in abundance than other clades. Given that Clostridiaceae contains a number of pathogens, I decided to look more closely at its prevalence in individual mice. Five mice out of the 201 represented the vast majority of Clostridiaceae signal. Because these mice were only in two locations and the abnormal abundance of this Family might suggest they were ill, I excluded them from subsequent analyses. I also removed taxa present in fewer than 10% of samples.

To create a metric for passenger rail distance between two places, I started with the geographic distance and then applied penalties based on the number of legs, number of stops, and the distance that cannot be traversed by rail (i.e. the legs for which Google can only find automobile routes). The details of these calculations are found in 03) transittimes.ipynb.

I performed PCA using the LAPACK Single Value Decomposition method as distributed in sci-kit learn. I first treated all taxa at any level as dimenions for PCA, then I repeated the analysis after collapsing taxa down to the Phylum level.

I next calculated Bray-Curtis distance at both the full taxon and Phylum levels and correlated it with geographic and transit-adjusted distances. Finally, I performed PCA using Bray-Curtis distance on the Phylum level.

Results

I was unable to replicate the results of the original paper in that I saw no correlation between microbial content and geographic location. This correlation could have been visible in two of my analyses. The first considers the coordinates of the sampling site in isolation **(Figure 2a-d)**. In other words, the ultimate visualization of the data will treat all sites as completely independent; their physical proximity to one another is irrelevant. If geography is a primary determinant of microbial content, unbiased PCA of either the full taxonomic order or the Phylum should show clusters that reflect the sampling site. This was not the case; neither two nor three component analysis revealed any structure that

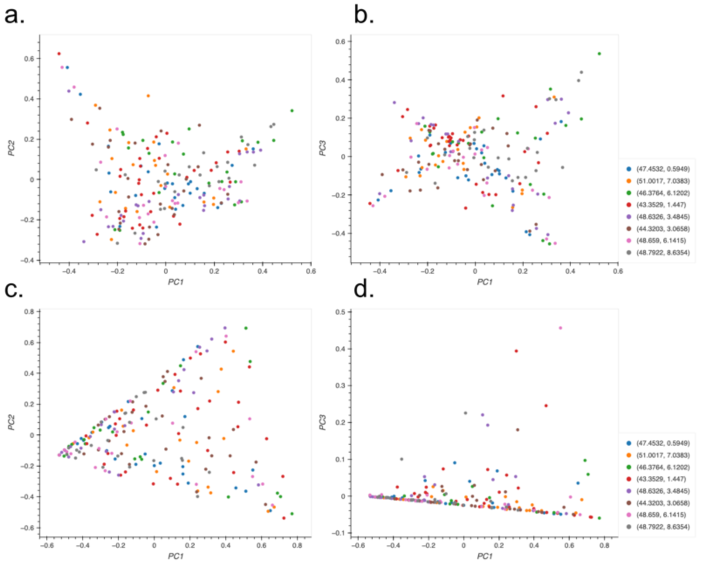

Figure 2. Three component PCAs performed with sci-kit learn for full taxonomic orders (a, b) or Phylum only (c, d). Components 1 and 2 are plotted in (a) and (c) and components 1 and 3 are plotted in (b) and (d).

could have indicated sampling site. Indeed, all geographic locations seem intermingled in every area of the plot.

The second analysis considers the relative distances between sites **(Figure 3a-b)**. These distances have three values: beta diversity as measured by Bray-Curtis dissimilarity, geographic distance, and transit distance. If my hypothesis is correct, the Bray-Curtis distance should have a positive correlation with both geographic and transit distance, but the transit distance should have a higher correlation. In fact, I observed no correlation between microbiome distance and either physical distance measure. The Bray-Curtis values had high variability and were approximately identical when comparing mice sampled 1000 km away from each other as when comparing mice collected at the same site.

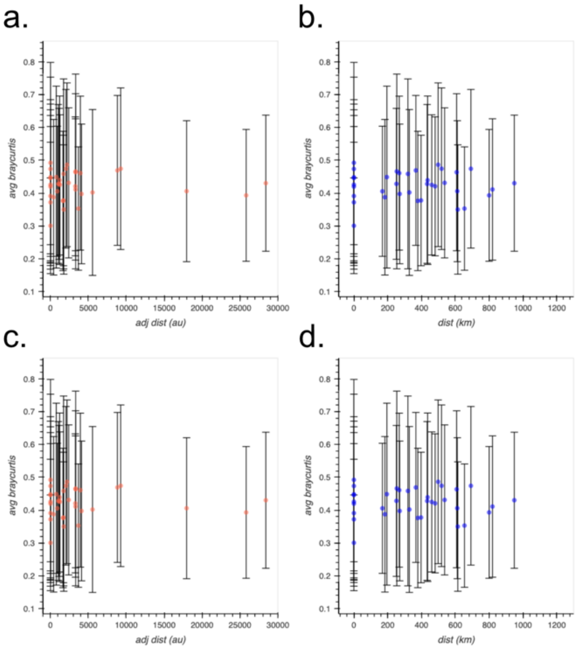

Figure 3. Bray-Curtis distances for the full taxonomic order (a, b) and Phylum only (c, d) plotted against transit-adjusted distance (a, c) and geographic distance (b, d).

Neither of these results were encouraging, but the original figure in the paper that drew my attention was a PCA performed on Bray-Curtis distances. Strangely, although the paper argues that geography is the most significant factor in determining intestinal microbial content, this PCA was performed on mucosal samples. I decided to use the caecal data with which I was working to attempt to replicate this result **(Figure 4)**. Once again, I was unable to replicate the published result and saw no clustering indicative of geography.

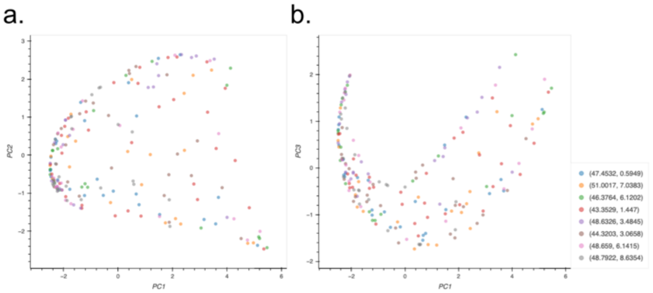

Figure 4. Three component PCA of Bray-Curtis distances. a) Component 1 vs component 2. b) Component 1 vs component 3.

Conclusions and Discussion

Based on my analysis, I was unable to satisfactorily evaluate the impact of human transit networks on wild mouse microbiomes. The primary reason for this failure was that I was unable to replicate the results of the original work. My analysis was different from the original paper but I think my approach should have been able to arrive at similar conclusions. The first major difference was that the original analysis kept all taxa but normalized the read depth per sample to 1000 reads using a uniform removal of reads. In contrast, I removed taxa that were present in fewer than 10% of samples, which reduced the overall number of taxa by over 90% (293 to 22 taxonomic orders). I repeated my analysis using all 293 taxa, but it did not materially impact my results. I wonder whether downsampling the data so aggressively for more complex samples led to biases due to the random dropout of low abundance taxa.

I also used a different PCA as I could not find a Python distribution that exactly matched their capscale method. The primary difference between capscale and LAPACK Singular Value Decomposition is that capscale analyzes multiple response variables. Given that the only variable I wanted to include in the analysis were caecal content, it is unclear to me what benefit capscale would have for this analysis unless an additional variable, such as mitochondrial DNA markers, was included. For this reason, I think that my PCA should have been adequate or superior to capscale for specifically testing whether geographic clusters emerge from dimensionality reduction of metagenomic data.

Overall, I think this exercise demonstrates the importance of having openly available code from initial processing of the data through visualization. It is highly possible that I've made errors in my analysis that could be remedied with more information on how the figures from the paper were originally generated. I would also be interested to see additional analyses of this dataset, as well as wild mouse or rat microbiome studies that span larger geographic areas. Norway rats might be particularly interesting, as they have populated the globe through shipping and colonial exploration leading to large amounts of genetic drift and huge habitat diversity.[8] Studies like this as well as expanded surveys of human microbiota would help us better understand how adaptation and happenstance contribute to the composition of our gut flora.

References

1. Suzuki, T. A. & Worobey, M. Geographical variation of human gut microbial composition. *Biol. Lett.* **10**, 20131037 (2014).

2. Gaulke, C. A. & Sharpton, T. J. The influence of ethnicity and geography on human gut microbiome composition. *Nat. Med.* **24**, 1495–1496 (2018).

3. Linnenbrink, M. *et al.* The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol. Ecol.* **22**, 1904–1916 (2013).

4. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).

5. Kuester, F. Is the German Rail Freight System broken? A portrait of DB Cargo | Combined Transport. *Combined Transport* https://combined-transport.eu/german-railway-system (2017).

6. Boudet, A. & Steinmann, L. The State once again supports rail freight | The echoes. *Les Echos* https://www.lesechos.fr/industrie-services/tourisme-transport/letat-une-enieme-fois-au-chevet-du-fret-ferroviaire-1226685 (2020).

7. Google. *Google Maps API v3.55*. (2023).

8. Puckett, E. E. *et al.* Global population divergence and admixture of the brown rat (Rattus norvegicus). *Proc. R. Soc. B: Biol. Sci.* **283**, 20161762 (2016).

Supplementary Material

All code used in retrieving and analyzing the data is available at https://github.com/dhonson-lncrna/DH_SymbiosisProject. To replicate the analyses in this report, run the notebooks in the following order: 01, 02, 03, 04. To perform the analyses without filtering taxa present in fewer than 10% of samples, run notebooks 02a and 04a. This will only work if notebooks 01 and 03 have previously been run.