Temporal Dynamics of the Human Microbiome

Toby Kenney (tkenney@mathstat.dal.ca)
Dalhousie University https://orcid.org/0000-0002-5421-4325
Junqiu Gao
Dalhousie University
Hong Gu
Dalhousie University

Research

Keywords: Mean Reversion; Time Series; Sampling Frequency; Ornstein-Uhlenbek Process; Fisher Information

Posted Date: February 7th, 2020

DOI: https://doi.org/10.21203/rs.2.22873/v1

RESEARCH

Temporal Dynamics of the Human Microbiome

Toby Kenney*, Junqiu Gao and Hong Gu

*Correspondence:
tkenney@mathstat.dal.ca
Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada
Full list of author information is available at the end of the article

Abstract

Background
There has been a lot of research about the role of the microbiome in various processes. The research has focused almost exclusively on the structure of the microbiome at a single time-point. There have been several studies that measure the microbiome from a particular environment over time. However, even in these studies, little has been done to study the temporal dynamics of the microbiome. In this paper, we begin to rectify this situation. We analyse a widely studied microbial data set that contains a time series of the microbiome from four body sites on two healthy individuals. We choose a data set based on healthy individuals because we are interested in the baseline temporal dynamics of the microbiome.

Results
For this analysis, we focus on the temporal dynamics of individual genera, ignoring the interactions. We use simple stochastic differential equation models to assess the following three questions. (1) Does the microbiome exhibit temporal continuity? (2) Does the microbiome have a stable state? (3) To better understand the temporal dynamics, how frequently should data be sampled in future studies?

We find that a simple Ornstein-Uhlenbeck model which incorporates both temporal continuity and reversion to a stable state fits the data for all genera better than a Brownian motion model that contains only temporal continuity. The Ornstein-Uhlenbeck model also fits the data better than modelling separate time points as independent. Under the Ornstein-Uhlenbeck model, we calculate the variance of the estimated mean reversion rate (the speed with which each genus returns to its stable state). Based on this calculation, we are able to determine the optimal sample schemes for studying temporal dynamics.

Conclusions
There is evidence of temporal continuity for at least some genera; there is clear evidence of a stable state; and the optimal sampling frequency for studying temporal dynamics is in the range of one sample every 0.8–3.2 days.

Keywords: Mean Reversion; Time Series; Sampling Frequency; Ornstein-Uhlenbek Process; Fisher Information

Background

A significant number of microscopic organisms live in and around the human body. Research has shown that human microbiome plays a significant role in human health, for example [16] [11] [2] [14].

Technological development in DNA sequencing has permitted a more systematic study of the microbiome [5]. There has been substantial work studying the instantaneous structure of the microbiome, but the temporal dynamics of the microbiome are largely unstudied. The studies that exist suggest that the microbiome is generally stable. However, these suggestions have not been subjected to rigorous statistical analysis.

Several recent studies have suggested the temporal dynamics of the microbiome may have clinical relevance to IBD. For example, [17] find that under some β-diversity measures, there was more variation between multiple gut microbiome samples from individuals with Crohn's disease than healthy controls, and there was less variation in samples from individuals with ulcerative colitis. This suggests that the dynamics of the microbiome may be affected by these diseases. However, that study was based on samples taken at 3-month intervals, so the actual dynamics were not observed. Other studies such as [9] have also sampled IBD patients and healthy controls at 3-month intervals. [22] attempt to apply this to improve classification of IBD patients and healthy controls from microbiome data. By comparing longitudinal samples, they are able to improve classification accuracy. They argue that this may be caused by IBD patients having more variable microbiomes. However, it could also be explained by the fact that sampling more replicates will in general improve prediction.

Other studies have looked at the temporal dynamics of the microbiome when an individual's diet changes. [23] and [6] both took daily samples of individuals in a controlled feeding experiment. They both found that the microbiome reacts quickly to sudden changes in diet, and reverts to baseline when the controlled diet regime ends. This suggests that the temporal dynamics of the microbiome should be measured on a scale of days, rather than months.

In this paper, we look at the moving picture data set [4]. This is a time series from two healthy individuals, with approximately daily sampling. We are interested in healthy individuals because we want to better understand baseline temporal dynamics of the microbiome. This will help to interpret future work on how the dynamics change under certain conditions.

Since the microbiome is often considered as an ecological system, it is natural to model its temporal dynamics as a stochastic process. The observed stability suggests that a mean-reverting process may be appropriate. In this paper, we compare a mean-reverting process model with two alternative models: one alternative model is Brownian motion, which can be characterised as random drift; the second alternative model is an independent model, where observations at different time points are independent. By comparing with these models, we hope to confirm the widely held beliefs that the microbiome does show some temporal continuity, and that the system is subject to mean reversion, meaning that the system returns to its stable state whenever the composition randomly fluctuates away from that state. After confirming these aspects of the dynamics, we will also obtain an estimate of the time-scale under which the dynamics operate. For this paper, we focus on the dynamics of each individual genus and ignore interactions between different genera.

Another major issue in studying the temporal dynamics of the microbiome is sampling frequency. Sampling too frequently may result in not covering enough time

to observe the patterns, while large gaps between samples can lead to consecutive samples being uncorrelated. It is widely acknowledged that "An important question is how often to sample ..." [1]. Current knowledge on this topic largely consists of guesswork, based on what has been observed in studies conducted at different timescales. However, stochastic differential equation models, in addition to offering insightful explanations of the dynamics, also allow us to apply the powerful statistical theory of Fisher information. This theory provides the asymptotic variance of parameter estimates, based on the sampling scheme and the true parameter values. It is then a straightforward optimisation problem to determine which sampling scheme will generate the most accurate estimates of the temporal dynamics. We apply this approach based on the estimated dynamics from the moving picture data to determine the optimal sampling frequency for future studies.

Results

Summary of Data set Studied

We perform this analysis on the moving picture data set [4]. This data set follows two healthy individuals over 6-month and 15-month periods respectively. Four body sites were observed: gut, tongue, right palm and left palm. Samples are not collected at completely regular time intervals. Many samples are taken at daily intervals, but many intervals of multiple days are also present. Samples were sequenced using PCR on the V2 region of the 16S rRNA gene [5]. We aggregated the data at genus level and restricted our attention to abundant genera with total counts in a given environment greater than 10,000 for Person 1 or 20,000 for Person 2. We computed the proportion data by dividing every count for each abundant genus by the total count of all genera in that sample. We also analysed log-transformed data, where we took the logarithm of each proportion. Where the observed count was zero, we calculated the log-proportion as if the observed count were 0.3. Table 1 and Table 2 show the number of observations and abundant genera for each individual and each body site respectively.

Table 1 The number of observations for each individual and body site

	Gut	Tongue	Right Palm	Left Palm
Person 1	131	135	134	134
Person 2	336	373	359	365

Table 2 The number of abundant genera for each individual and body site

	Gut	Tongue	Right Palm	Left Palm
Person 1	17	11	12	45
Person 2	32	18	59	29

Testing Temporal Dependence of Microbial Dynamics

We perform a likelihood ratio test with null hypothesis an i.i.d. normal model, where there is no temporal dependence between observations, and alternative hypothesis an Ornstein Uhlenbek (OU) process which includes temporal dependence. The number of genera rejecting the null hypothesis using the proportion data for each person and body site is shown in Table 3. We see that many of the abundant

genera show strong evidence of dependence between different time points, particularly in more enclosed body sites, such as the gut. More exposed body sites show less evidence of temporal dependence. Since the i.i.d. normal model is a limiting case of the OU process when the rate of mean reversion tends to ∞, evidence of temporal dependence will be weaker in cases where the mean reversion is faster. It makes intuitive sense that exposed body sites could have faster mean reversion, because exposure to external influences is one of the driving factors that influence the microbiome towards its stable state, so body sites which are more exposed to external influences could be expected to exhibit faster mean reversion.

Table 3 The number of abundant genera with p-values less than 0.05, for the i.i.d. normal hypothesis against OU process on proportion data, for each individual and body site. Numbers after "/" are total number of abundant genera in the data set.

	Gut	Tongue	Right Palm	Left Palm
Person 1	15/17	8/11	8/12	22/45
Person 2	25/32	14/18	39/59	22/29

The likelihood ratios for each genus, along with the null distribution for each data set, are shown in Figure 1 in Appendix C. Many of the likelihood ratios are much larger than the critical values, indicating very significant evidence of temporal dependence for at least some genera.

Table 4 gives the same results for log-transformed data. The evidence of serial dependence is stronger in most cases after the log-transformation. We see that there is evidence of serial correlation for the vast majority of genera. This is expected, and indicates that changes in the community are at least partially due to the temporal dynamics of the system (rather than errors in the sequencing process, which would not be expected to show temporal dependence). Figure 2 in Appendix C shows the likelihood ratios compared to the null distributions. As for the proportion data, for some genera the evidence of temporal dependence is very strong.

Table 4 The proportion of abundant genera which reject the null hypothesis of i.i.d. log-normal distribution against an OU process on the log-proportion data, for each individual and body site. Numbers after "/" are total number of abundant genera.

	Gut	Tongue	Right Palm	Left Palm
Person 1	17/17	10/11	6/12	29/45
Person 2	30/32	18/18	55/59	25/29

Testing for Mean Reversion

We test for mean reversion using a likelihood ratio test between a null hypothesis of Brownian motion without drift, which has no mean reversion and an alternative hypothesis of an OU process, where the rate of mean reversion is given by the parameter η. The likelihood ratio statistics for each abundant genus in each body site are shown in Figure 3 in Supplementary Appendix C along with the null distribution and critical values. In addition to the statistical benefits of comparing nested models, setting the drift parameter to 0 in a Brownian motion is natural because the proportions of different genera are constrained to lie between 0 and 1, so a model with drift is not sustainable. We find that all the log-likelihood ratio

Table 5 Largest p-values for any genus in each data set for a likelihood ratio test between Brownian motion without drift, and an OU process.

	Person	Right Palm	Left Palm	Gut	Tongue
Proportion data	1	< 0.0002	< 0.0002	< 0.0002	< 0.0002
	2	< 0.0002	< 0.0002	< 0.0002	0.0006
Log Proportion data	1	< 0.0002	< 0.0002	0.0064	< 0.0002
	2	< 0.0002	< 0.0002	< 0.0002	0.0018

These p-values were calculated using a simulation of 5,000 values from the null distribution (Brownian motion without drift).

tests for the real genus data reject the null hypothesis. The largest p-value of any genus in each body site for each person are shown in Table 5. We see that there is very strong evidence rejecting Brownian motion in all data sets. This indicates that all abundant genera are subject to some mean reversion.

Figure 4 in Appendix C shows the same results for a log-transformed OU model. Again all abundant genera are subject to some mean reversion. Mean reversion is expected, since we know there are many mechanisms that keep the microbial system in a stable state. We see that the likelihood ratio statistics are larger on average for the palms than for the gut and the tongue, while the likelihood ratio statistics for the comparison with the i.i.d. log-normal distribution are larger for the gut and the tongue. This suggests stronger mean reversion in the palm microbiomes, and weaker mean-reversion in the gut and tongue microbiomes, which can be explained by the fact that the gut and tongue are enclosed systems with fewer external influences driving the microbiome back to the stable state.

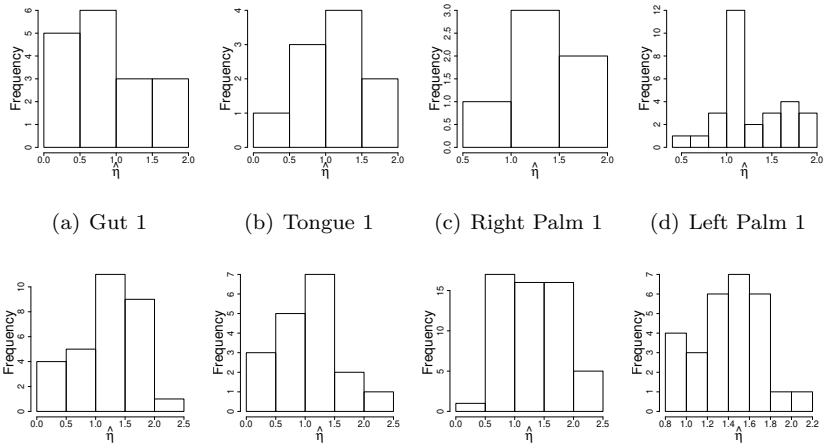
Variance of Estimated Mean Reversion Rates and Optimal Sampling Protocols

Next we look at the variance of our estimated value of η. Under an OU process, based on the theory of Fisher information, the asymptotic covariance of the parameter estimates in the OU model is given by the inverse of the Fisher information matrix given later in Proposition 1. The variance of $\hat{\eta}$ depends on η, but not on μ or σ. Histograms of the estimated values of η are given in Figure 1 for log proportion data and in Figure 2 for proportion data.

We see that many values of $\hat{\eta}$ are close to 1, with some larger values for the palms. For the moving picture data set, the standard deviations of $\hat{\eta}$ for different values of η are given in Table 6. In the Methods section, we are able to use the theory of Fisher information to determine the optimal sampling scheme for estimating the mean reversion rate η from a fixed number of observations. We compare the estimated standard deviations using the actual sampling scheme with the standard deviations that could be achieved by the optimal sampling scheme with a similar number of samples.

From Table 6, we see that $\hat{\eta}$ is a reasonable estimate for $\eta = 0.4$ and $\eta = 1$, with a coefficient of variation of 20–25% for all body sites for Person 1, and a coefficient of variation of about 14% for Person 2. The sampling scheme used achieves an accuracy close to the optimal sampling for genera where the true rate of mean reversion is in this range. For the genera with faster mean reversion, $\hat{\eta}$ is less accurate, and the accuracy could be improved by sampling more frequently.

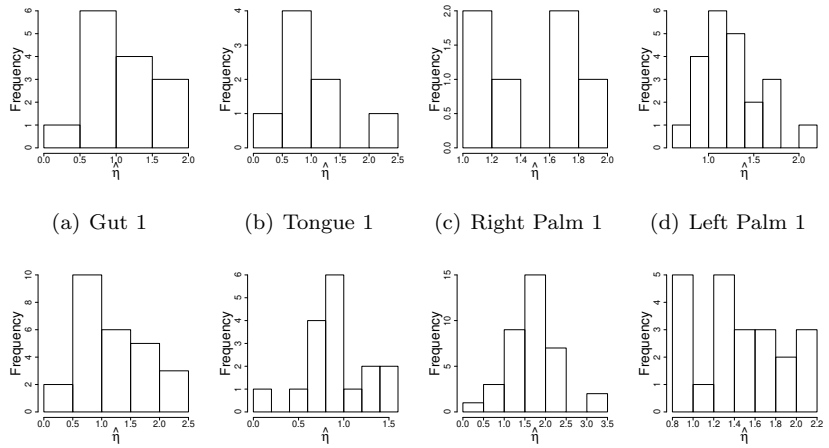
From Theorem 3 (see Methods section) we see that the optimal sampling scheme to study the temporal dynamics of a genus with mean reversion rate η is to sample regularly with time step $\frac{1.59362426}{2\eta}$. Thus, for some of the less quickly reverting



(a) Gut 1 (b) Tongue 1 (c) Right Palm 1 (d) Left Palm 1

(e) Gut 2 (f) Tongue 2 (g) Right Palm 2 (h) Left Palm 2

Figure 1 Distribution of $\hat{\eta}$ over genera for Person 1 (top) and Person 2 (bottom) for log-transformed data.



(a) Gut 1 (b) Tongue 1 (c) Right Palm 1 (d) Left Palm 1

(e) Gut 2 (f) Tongue 2 (g) Right Palm 2 (h) Left Palm 2

Figure 2 Distribution of $\hat{\eta}$ over genera for Person 1 (top) and Person 2 (bottom) for proportions. One outlier was removed from plot (a).

Table 6 Estimated standard deviations of $\hat{\eta}$ from the moving picture data for various body sites and true mean reversion rates

Body Site	$\eta = 0.4$	$\eta = 1$	$\eta = 1.5$	$\eta = 2$
Person 1 Gut	0.093985	0.229590	0.404363	0.682644
Person 1 Tongue	0.093338	0.224076	0.392306	0.660776
Person 1 L. Palm	0.093542	0.225305	0.395091	0.665951
Person 1 R. Palm	0.093542	0.225305	0.395091	0.665951
Optimal Sampling 131 samples	0.086844	0.217093	0.325639	0.434185
Optimal Equal-Spaced 131 Samples	0.087189	0.217972	0.326959	0.435945
Person 2 Gut	0.058055	0.143934	0.256814	0.436208
Person 2 Tongue	0.056130	0.133687	0.234737	0.395993
Person 2 L. Palm	0.056856	0.137175	0.242194	0.409543
Person 2 R. Palm	0.056524	0.135688	0.239088	0.403997
Optimal Sampling 351 samples	0.052923	0.132307	0.198460	0.264614
Optimal Equal-Spaced 351 Samples	0.053137	0.132843	0.199265	0.265686

genera, it would be better to sample slightly less frequently, while for some of the more frequently reverting genera, we should aim to sample more frequently to understand the temporal dynamics. Some of the fastest mean-reverting genera have $\hat{\eta}$ more than 2. For this value of η, it would be best to sample 2.5 times per day. Obviously, this may be impractical for some environments. Figure 3 shows the effect of the sampling scheme on our estimates of η. We compare the actual sampling scheme, a sampling scheme optimised for $\eta = 1$, and the best results that can be obtained for each particular η.

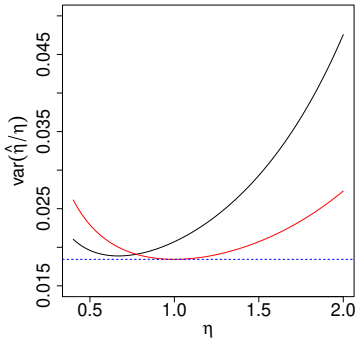


Figure 3 Variance of $\hat{\eta}$ under actual sampling for Person 2's gut (black), under a sampling scheme optimised for $\eta = 1$ (red), and under a sampling scheme optimised for each η separately (blue).

From Figure 3, we see that the actual sampling gives fairly good estimates for smaller values of η, but for larger values of η, the variance of $\hat{\eta}$ is about twice as large as it would be if the sampling were optimised for $\eta = 1$, and nearly three times as large as it could be with more frequent sampling. We see that optimising for $\eta = 1$ would be good at controlling the variance for most of the values of η estimated from the data. We conclude that daily sampling is fairly good, and should be used for future studies. Ideally, for the rates observed here, we should aim to sample slightly more frequently than once per day. Among regular sampling schemes, sampling approximately once every 18 hours would be ideal for this data set, but small variations in time between samples could increase the range of values over which our estimates of η are accurate, so there is some flexibility about the sampling scheme.

From Proposition (2) in the Methods section, we have that if the samples are evenly spaced with time difference Δ_t, then $\mathrm{Var}\,(\hat{\eta}) = (I^{-1})_{\eta\eta} = \frac{e^{2\eta\Delta_t}-1}{n{\Delta_t}^2}$ does not depend on σ or μ and is inversely proportional to n. Figure 4 shows the relation of $\mathrm{Var}\,(\hat{\eta})$ with Δ_t for fixed $\eta = 1$ and $n = 100$.

From Figure 4, we see that the accuracy of our estimated mean reversion rate is not harmed too much by sampling slightly more or less frequently, though as the sampling frequency deviates further from the optimal value, the effect of sampling frequency becomes more significant.

Rates of Mean Reversion for Different Genera, People and Body Sites

We now look at how the rates of mean reversion vary between different genera, body sites and individuals. Table 7 shows the estimates $\hat{\eta}$ for the genera which are

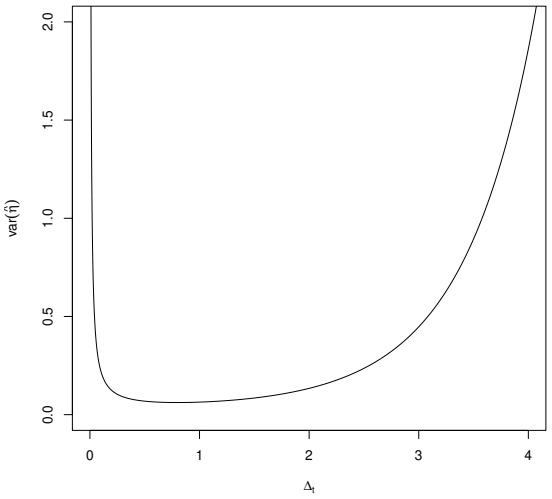


Figure 4 Relation between time difference Δ_t and $\mathrm{Var}(\hat{\eta})$ for $\eta = 1$, $n = 100$

abundant in at least five of the eight environments, based on the log-transformed data.

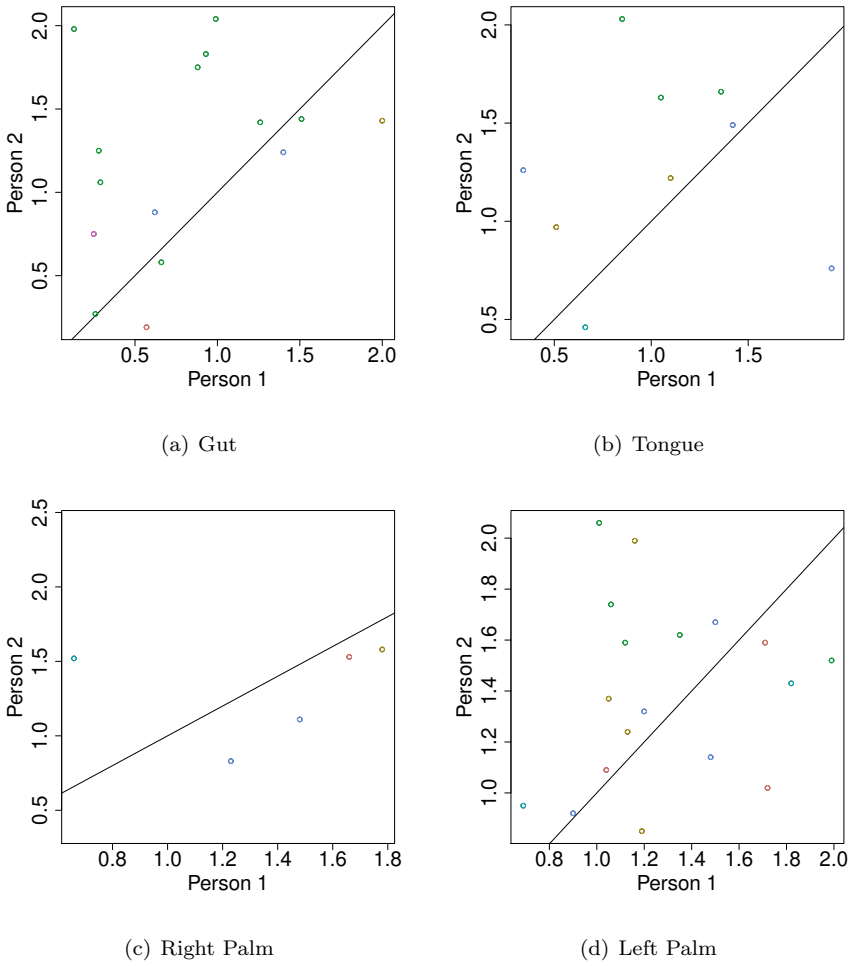
Table 7 Comparison of estimated η values (with standard errors) for log-transformed OU across body sites and individuals for common genera.

Genus	Person	Right	Left	Gut	Tongue
Actinomyces	1	1.66(0.47)	1.04(0.24)		
	2	1.53(0.25)	1.09(0.15)		1.18(0.16)
Porphyromonas	1	1.16(0.27)	1.05(0.24)		0.51(0.11)
	2		1.37(0.21)	1.74(0.33)	0.97(0.13)
Prevotella	1	1.78(0.53)	1.16(0.27)	2.00(0.68)	1.10(0.25)
	2	1.58(0.26)	1.99(0.40)	1.43(0.24)	1.22(0.17)
Neisseria	1		1.50(0.40)		0.34(0.08)
	2	1.92(0.38)	1.67(0.29)		1.26(0.18)
Fusobacterium	1	0.66(0.14)	0.69(0.15)		0.66(0.14)
	2	1.52(0.25)	0.95(0.13)		0.46(0.06)
Veillonella	1		1.12(0.26)		1.05(0.24)
	2	2.23(0.52)	1.59(0.26)		1.63(0.27)
Haemophilus	1		1.20(0.28)		1.42(0.36)
	2	2.29(0.55)	1.32(0.20)		1.49(0.23)

We see that there is substantial variation in the estimated mean reversion rates across different body sites, even for the same genus. This is also shown in Figure 5 and Figure 6, which include all genera abundant in both environments. Full estimates for η for the log-transformed OU process for all abundant genera for both people at all body sites are given in Appendix C Table 1.

From Figures 5 and 6, we see that there is some correlation between the estimated mean reversion rates for a given genus in different environments, but it is fairly weak.

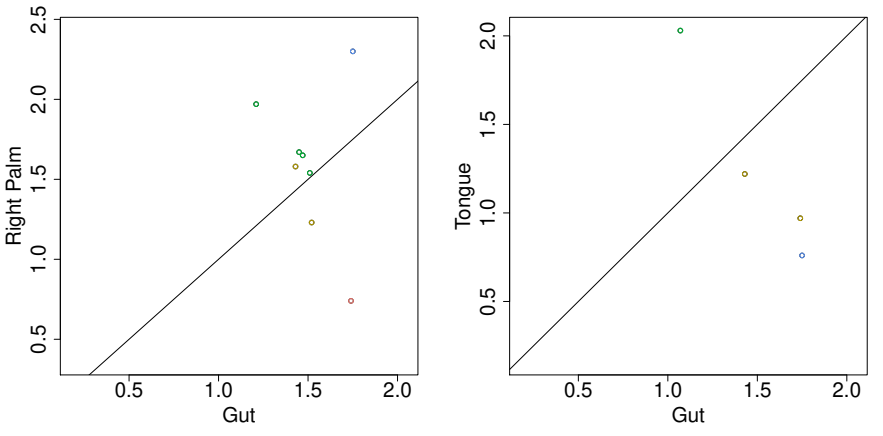
Figure 7 shows the estimated mean reversion for the most abundant genera in each body site for Person 2 under a log-OU model. The genera are grouped by taxonomy, so genera in a given taxonomic grouping (phylum, class, order or family) are adjacent. From the plot, we see that estimated mean reversion rates are lower on average for the gut and the tongue, and higher for the palms, but there is



(a) Gut (b) Tongue

(c) Right Palm (d) Left Palm

Figure 5 Comparison of estimated η values for log-transformed data for common genera between two individuals for the same body site. The lines $y = x$ are shown on each plot. Points are coloured by phylum.



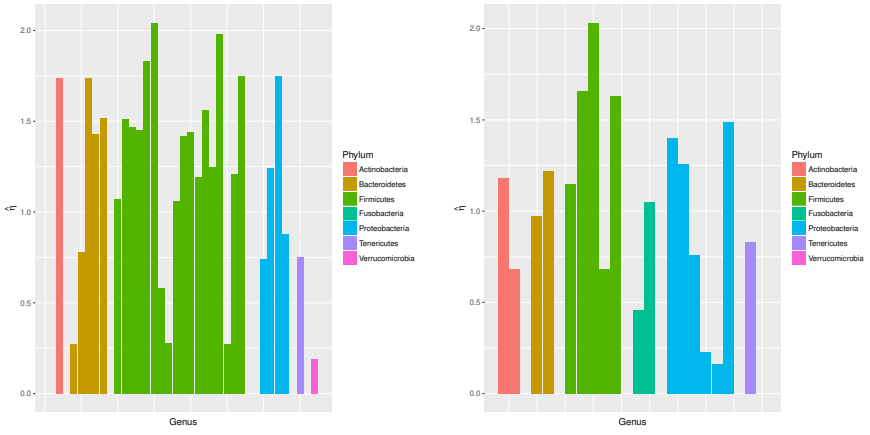
(a) Gut — Right Palm

(b) Gut — Tongue

(c) Tongue — Left Palm

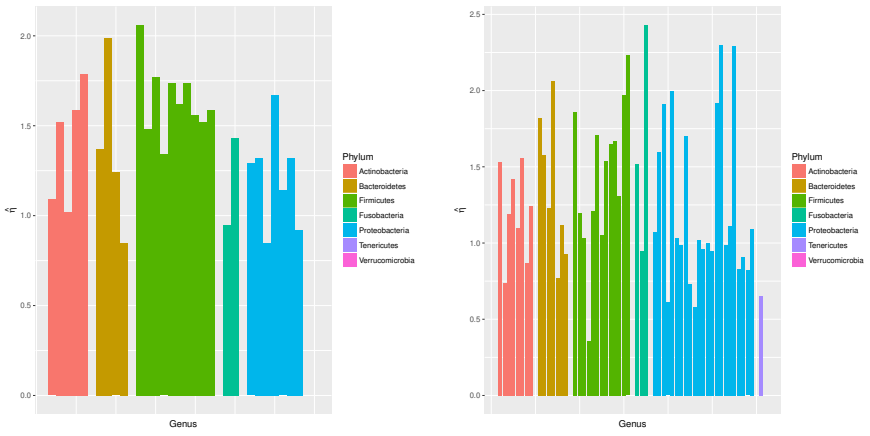
(d) Right Palm — Left Palm

Figure 6 Comparison of estimated η values for common genera between two body sites for Person 2. The lines $y = x$ are shown on each plot. Points are coloured by phylum.



(a) Gut

(b) Tongue



(c) Left Palm

(d) Right Palm

Figure 7 Comparison of estimated η values at log-scale for common genera at each body site for Person 2. The genera are arranged by taxonomy.

a lot of variation between different genera in each environment. Figure 7 shows little similarity in the temporal dynamics of phylogenetically close genera. Most Firmicutes tend to have slightly faster mean reversion than other phyla, though a few show low mean reversion rates in the gut. Mean reversion rates are generally higher on the palms, which makes intuitive sense, since the palms are more exposed systems, so would be expected to be subject to external influences, which would reset any imbalance's that might arise in the system. A similar figure showing the estimated mean reversion rates for various genera for Person 1 is Appendix C, Figure 5. Again, there is no clear taxonomic pattern in the estimated mean reversion rates.

Figures covering the distribution of $\hat{\sigma}^2$ are in Appendix C. Figure 6 in Appendix C shows the distribution of $\hat{\sigma}^2$ arranged by taxonomy. There are several genera for which $\hat{\sigma}^2$ is extremely large. Upon inspection, these are conditionally rare genera, which are often absent from the samples, and occasionally occur in large blooms. It seems that $\hat{\sigma}^2$ is larger for the palms. This makes sense, since the palms are exposed to more external influences which can affect the microbial community. Interestingly $\hat{\sigma}^2$ is lower for the tongue than for the gut, indicating smaller random fluctuations. Given that the tongue is more exposed than the gut, this is slightly surprising. It can be partially explained because the most abundant genera in the tongue are more abundant, and more abundant genera are expected to be more stable. However, even if we compare genera with the same stable level in the gut and the tongue, the estimated value of $\hat{\sigma}$ is lower for the tongue.

We compare the estimated values of η on the proportion data, and on log-transformed data. We see that there is some correlation but not too much, so the log-transformation does make a noticeable difference to our estimated rate of mean reversion.

Discussion

In this paper, we have studied the temporal dynamics of the most abundant genera in the microbiomes at various body sites. We found evidence of temporal dependence and mean reversion in the moving picture data set. For enclosed body sites, the abundant genera show stronger evidence of dependence than for external sites. Furthermore, all of the abundant genera show evidence of mean reversion. This provides statistical support for previous intuitive observations about the temporal stability of the microbiome.

Our model also estimates the time-scale of mean reversion for each genus. Under the OU model, the expected abundance decays exponentially towards the mean, and never reaches it. A common way to describe the time-scale in this context is the *half-life*, which is the time until the expected abundance is half way towards the mean. For the OU process, this time is $\lambda = \frac{\log(2)}{\eta}$. For the real data example, our estimated values of η were between 0.4 and 2, which corresponds to a half-life of 0.35 days to 1.7 days.

These results are consistent with the results of diet change studies, e.g. [23, 6]. Those studies observed that the microbiome can respond to changes in diet within one day. In the OU framework, a change in diet could be modelled as a change in equilibrium state. Under the OU model, the system would then adjust to the new

equilibrium state with the same temporal dynamics. The time taken to revert half way towards the new system would be between 0.35 and 1.7 days, which is in line with the results from those studies.

The OU model also has some relation to the long-term variance of the system. The IBD studies [17, 9] found differences between IBD patients and healthy controls in the long-term variance of the system. Under the OU model, the long-term variance is $\frac{\sigma^2}{2\eta}$. A difference in long-term variance between two populations could be explained in several ways:

- An increase in σ^2. This would correspond to more rapid fluctuation in the microbial communities. It is unclear what biological processes could cause such a change.
- A decrease in η. This would correspond to a weakening of the mean-reversion mechanisms involved. For example, the host immune system might respond in an abnormal way to changes in the microbiome, reducing its stabilising effect.
- Temporal variation in the stable state. The OU model assumes that the stable state is fixed. However, there is strong evidence that this state is influenced by many external factors such as diet, lifestyle, antibiotics, etc. If these external factors vary more, or are more influential in IBD patients, then we would expect the asymptotic variance to increase.
- Variation in sampling bias. The sampling procedure is known to introduce large biases into microbiome data. It is conceivable that some of the many factors influencing this such as consistency of the stool, or blood in the stool, could lead to higher sampling variance in IBD patients than controls.
- An artefact of the methodology. The studies [17, 9] were based on beta-diversity measures, which could be sensitive to changes in the stable state. Since the stable state is different for IBD patients, comparisons of the variability of the microbiome for healthy controls and IBD patients depend heavily on the choice of measurement.

Further work is needed, with more frequent sampling, to determine which of these cases actually explains the observed results.

We can also examine how the half-life varies between different genera and environments. There is significant variation, but is is only weakly associated with the most obvious differences. The half-life is generally shorter for exposed environments such as the palms. This suggests that the external environment may act as a stabilising force, quickly driving the system back to the equilibrium state. The more enclosed states are not driven back so quickly, allowing the state to drift away from the equilibrium state for longer. On the other hand, the palms exhibit larger random fluctuations, so the long-term variance is actually larger for the palms than for the gut or tongue.

The rates of mean reversion for a fixed genus in multiple environments show weak correlation. This seems reasonable, since part of the mean reversion process is expected to depend on the characteristics of the particular genus, while the remainder is expected to be driven by the environment. We expect the strength of the correlation to vary according to the similarity between the environments. This seems to be the case, but the number of genera that are abundant in both environments is too small to make reliable conclusions about this.

Comparing the rates of mean reversion with the taxonomy of the genera, we do not see very strong relationships, even within a single environment. Mean reversion does appear to be slightly faster on average for Firmicutes than for Proteobacteria, but the variation within each phylum is much larger than the between-phylum variation. This suggests that more of the factors controlling microbial dynamics are related to particular genera, rather than higher level taxonomy. It would be interesting to study the dynamics at even higher resolution, but doing so would require adapting the methodology to better take account of the noise caused by sequencing.

We also derived the Fisher information matrix for the OU process and applied the theory of Fisher information to measure the accuracy of our estimated mean reversion velocity. For the moving picture data, we calculated the variance of our estimates, and showed that the estimates were reasonable. We also used Fisher information to determine the most efficient sampling schemes for future studies. If we insist on a minimum time difference between samples, in order to make our estimates more robust to model misspecification, then the optimal sampling scheme is to sample equally-spaced time points with difference $d_i \approx \frac{0.80}{\eta}$. Even in the case where the data perfectly follow an OU process, this sampling scheme is very efficient, causing a less than 1% increase in the variance of the estimator, compared to the theoretically best sampling scheme under the model. Given that we know the OU model is not perfect because of sequencing and other issues, we recommend this sampling scheme. We performed simulations to confirm that the asymptotic theory used here applies fairly accurately to our finite sample cases.

The optimal sampling scheme varies with the true rate of mean reversion. Thus, the optimal time difference is different for different body sites, or if we are interested in a particular subset of genera or OTUs. The optimal sampling frequency for enclosed body sites is slightly smaller than for exposed body sites (meaning we should have shorter intervals between samples for studying exposed body sites). The efficiency of the estimation remains reasonable if the sampling rate differs slightly from the optimal rate, so there is some flexibility in sampling. For the real data, sampling rates of approximately one sample per day should be adequate for studying the dynamics of these communities.

One limitation of the data set studied in this paper is that the sampling times are only available to the nearest day. We do not know what time of day the samples were taken. We have assumed that they were collected at the same time of day. However, if this assumption is not true, then the time of collection could affect the estimates under the OU model. It is unlikely that the difference will be large, but future studies would be able to estimate temporal dynamics more accurately if the time of sample collection and processing were available at higher resolution.

Future work

The OU process used in this paper is a very simple model with a linear velocity parameter mean reversion. It does not capture many of the aspects of the real data. In this section, we discuss some of the possible improvements to the model. The Fisher information theory from this paper can be extended to these improved models. We expect the estimated optimal sampling times not to be overly sensitive to the exact model specification, so that the same sampling scheme can be used to obtain good parameter estimates for multiple models.

One improvement to the model would be to allow multiple stable states. This could be achieved either by using a non-linear mean-reversion term or by a hidden Markov model where the equilibrium value varies following some process. While both of these models would describe a system with multiple stable states, they would have slightly different dynamics, and very different biological interpretations. Under the nonlinear equation, the microbiome drifts between stable states under its own dynamics. This change in state could then cause phenotypic changes in the host or environment. For example, dysbiosis might cause illness in the host. Under the hidden Markov model, the dynamics of the equilibrium state model external forces affecting the system. For example, an immune flare-up in response to some allergen might result in different dynamics of the microbiome. The distinction between these two models is of extreme clinical importance. Under the nonlinear model, monitoring the microbiome might provide early prediction of dysbiosis, and there is the potential for microbiome-based remedies. Under the hidden Markov model, changes to the microbiome occur after the external system change, so monitoring the microbiome offers less advance warning. Furthermore, if the microbiome changes are symptomatic rather than causal, microbiome-based interventions are unlikely to persist, or to remedy other symptoms. We hope to be able to distinguish between the two types of dynamics by comparing the fit of the respective two models.

It is widely believed that the temporal dynamics of the microbiome are driven by interactions between different OTUs, rather than each OTU acting independently. Therefore, it would be appropriate to develop a model which incorporates interactions between OTUs. There are already several differential equation models that have been used to model multiple systems. There is a natural multivariate version of the OU process, where the stable state is replaced by a vector, the mean reversion velocity η is replaced by a matrix, and the random fluctuations are vector-valued. Alternatively, a number of multivariate differential equations used in ecology to model the growth of multiple populations, such as the Generalised Lotka-Volterra model or the Holling type-II model, could be equipped with a stochastic term to incorporate the random effects. Whatever model is used, parameter estimation would be challenging because of the high dimensionality. To deal with the high-dimensionality, some conditions to ensure sparsity of the estimated interaction parameters would be appropriate.

We should also incorporate measurement error in sampling to derive more accurate estimates. It is well-known that microbiome data are subject to significant errors and biases in DNA extraction and sequencing. These errors could have an impact on the estimated parameters of our model. By developing a model which incorporates as much as we know about the error in the sequencing procedure, we should be able to obtain more accurate parameter estimates. These errors are more significant at lower taxonomic levels, so by modelling the error structure, we will be able to apply our model to lower taxonomic levels, which could reveal more interesting biological patterns.

Another aspect we would like to include in the model in future is the non-homogeneity of the sample. Given that most microbes are several micrometres in diameter, it is very plausible that entirely separate microbial communities could coexist, separated by mere millimetres. In such a case, a faecal sample that has

travelled the entire length of the gut, or even a sample collected from different areas on opposite sides of the tongue, would be expected to be a mixture of these different communities. However, the exact proportion of each community included in the sample will vary randomly between samples. This means that even if each of the microbial communities perfectly follows the stochastic differential equation, the overall sample will exhibit more complicated dynamics. Creating a model to include this effect will involve solving major statistical difficulties, but is a long-term goal for better modelling the temporal dynamics of microbial communities.

Conclusions

There is clear evidence of temporal dependence among many of the abundant genera at all body sites. There is very strong evidence for mean reversion for all genera at all body sites. This provides support for previous observations about the temporal stability of the microbiome, but in a more statistically rigorous framework. Our model also estimates the time-scale of mean reversion. We estimate the time for each genus to revert half way towards its mean. This time varies between about 0.35 days for some of the most stable genera to about 1.7 days for the less stable genera. There is a large variation in all environments, but on average, mean reversion is slightly faster in exposed environments. The rate of mean reversion for a single genus varies between environments, but shows some weak correlation across different environments. The rate of mean reversion is not strongly associated with the taxonomy, though there are some general trends (e.g. mean reversion is on average slightly faster for Firmicutes than for Proteobacteria).

Using the Fisher information matrix, it is possible to estimate optimal sampling strategies for studying temporal dynamics of the microbiome. Based on this calculation, daily sampling is close to optimal for most genera. Estimates for fast-reverting genera would be improved by more frequent sampling. More accurate sample collection time data would also improve the accuracy of the estimated mean reversion.

Methods

Review of Ornstein-Uhlenbeck Process

The Ornstein-Uhlenbeck process is a very simple stochastic differential equation in a single variable subject to mean reversion, meaning that, while fluctuating randomly, the variable's values trend towards a stable mean value. This process has been used extensively to model situations where mean reversion is expected in a wide range of areas including physics [13], finance [20] and biology [19] [18]. It combines a linear mean-reversion term with a Brownian motion noise term. We therefore begin by reviewing Brownian motion.

Brownian Motion

Brownian motion is a simple model of the behaviour of a system undergoing random fluctuations. It is the limiting process of a random walk as the step size and time between steps both converge to zero in a certain way. A thorough introduction of Brownian motion can be found in [10].

A stochastic process $W_t, t \geqslant 0$ with state space $\mathbb{R}$ is a *Standard Brownian Motion* (also called a *Wiener process*) if for any $0 \leqslant s \leqslant t$, $W_t - W_s$ is normally distributed with mean 0 and variance t, and $W_t - W_s$ is independent of $\{W_r | 0 \leqslant r \leqslant s\}$.

If $W_t, t \geq 0$ be a standard Brownian motion. A stochastic process $\{X_t | t \geqslant 0\}$ given by

$$X_t = x_0 + \mu_{\mathrm{BM}} t + \sigma W_t, t \geq 0$$

is called a *Brownian motion* with drift parameter $\mu_{\mathrm{BM}} \in \mathbb{R}$, variance parameter $\sigma^2 > 0$, and starting point $x_0 \in \mathbb{R}$.

If $X_t, t \geqslant 0$ follows Brownian motion with drift μ_{BM} and variance σ^2 then for any $s, t \geqslant 0$, $X_{s+t} - X_s \sim N(\mu_{\mathrm{BM}} t, \sigma^2 t)$, and $X_{s+t} - X_s$ is independent of $\{X_r | 0 \leqslant r \leqslant s\}$.

Ornstein-Uhlenbeck Process

The OU process X_t is defined by the following linear stochastic differential equation (SDE)

$$dX_t = \eta(\mu - X_t)dt + \sigma dW_t$$

Where $\eta > 0$ is the velocity of the reversion process and μ is the stable state and W_t is a Wiener process. We see that when $X_t > \mu$, the average derivative of X_t is negative, meaning that on average X_t will decrease; when $X_t < \mu$, the average derivative is positive. Thus in all cases X_t will on average tend towards μ, but the Brownian motion term adds random fluctuation to its trajectory. The rate at which X_t trends towards the stable state grows larger as X_t moves further away from the stable state. η represents the average rate at which X_t reverts to the stable state, while σ represents the magnitude of the random fluctuations. A more complete introduction to the OU process can be found in any textbook on stochastic differential equations, for example [15].

There is a well-known explicit solution available to the OU process:

$$X_{s+t} | X_s \sim N\left(\mu + e^{-\eta t}(X_s - \mu), \sigma^2 \frac{1 - e^{-2\eta t}}{2\eta}\right)$$

Testing Temporal Dependence and Mean Reversion of Microbial Dynamics

We use likelihood ratio tests to test for temporal dependence and for mean reversion. For testing temporal dependence, the hypotheses are:

H_0**:** X_t follow i.i.d. Normal distributions.

H_1**:** X_t follow an OU mean reverting process.

For testing for mean reversion, the hypotheses are:

H_0**:** X_t follow Brownian motion without drift ($\mu_{\mathrm{BM}} = 0$).

H_1**:** X_t follow an OU mean reverting process.

We perform separate tests for each person, body site and genus. Since we expect the majority of genera to reject the null hypotheses, we do not worry about multiple test correction, which reduces false positives in cases where the majority of tests cannot reject the null hypotheses. The i.i.d. normal model is a limiting case of the OU process model as $\eta \to \infty$, $\sigma \to \infty$ with $\frac{\sigma^2}{\eta}$ fixed. The difference between the i.i.d. normal model and the OU process is just the serial correlation, so if the log-likelihood ratio test rejects the i.i.d. normal case, it suggests that there is serial dependence.

Brownian motion without drift is a special case of an OU process where the mean-reversion parameter is 0. Thus, testing against Brownian motion is a natural way to test for mean reversion.

The likelihood ratio statistics for these tests are not guaranteed to follow the usual χ^2 distribution. In the case of the i.i.d. normal hypothesis, this is because it is a limiting case, rather than an internal parameter value. In the Brownian motion case, the mean parameter μ of the OU process vanishes, so it is a non-identifiable special case. There is some theory on the asymptotic behaviour of these statistics in the case where the samples are equally spaced [21], but it does not apply in our case where the samples are not evenly spaced. Instead, we find the critical values for our hypothesis tests by simulation.

We simulate 5000 data sets using the same time points as the original data, under the null hypothesis. (We use a different simulation for each person and body site, as the time-points are different.) The likelihood ratio statistic is scale-invariant and translation-invariant for both tests (see Supplemental Appendix A.3 for a proof) so the null distribution is the same for any values for the parameters of the null distribution. Therefore, we only need to perform one simulation for each person and body-site. For the normal distribution we use $\sigma = 1$ and $\mu = 0$ for the simulation. For Brownian motion, we simulate with $x_0 = 0$, drift $\mu_{\mathrm{BM}} = 0$ and variance $\sigma = 1$.

For the Brownian motion simulation, the η estimated for the OU mean-reverting process should be close to zero. To reduce the effect of rounding errors, we use a Taylor expansion approximation to evaluate the parameter estimates and the log-likelihood. Details of this approximation are in Supplemental Appendix A.2.

Variance of Estimated Mean Reversion Rates and Optimal Sampling Protocols

For estimating the variance of parameter estimates, we will use the statistical theory of Fisher Information. A detailed review of Fisher information can be found in [12] or [8]. For a model with parameter vector $\theta = (\theta_1, \ldots, \theta_k)^T$, the Fisher information matrix at a vector $\theta = \theta_0$ is defined by $I = (I_{ij})_{i=1,\ldots,k, j=1,\ldots,k}$ where

$$I_{ij} = -\mathbb{E}_{X|\theta}\left(\frac{\partial^2 l(\theta; X)}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\theta_0} \right)$$

where $l(\theta; X)$ is the log-likelihood of data X at parameter value θ.

The main use of Fisher information is the result that under common conditions, maximum likelihood estimates of parameters are asymptotically normal with variance given by the inverse of the Fisher information matrix. More explanation and a proof can be found in many textbooks on statistical theory, for example [3] [7].

Fisher Information Derivation for OU Mean Reverting Process

To apply this to OTU temporal dynamics, we first calculate the Fisher information matrix for an OU process with parameters η, μ and σ.

Proposition 1 *For an OU process with parameter vector $\theta = (\mu, \eta, \sigma)^T$, sampled at time points $t_0 = 0, t_1, \ldots, t_n$, the Fisher information matrix is given by*

$$I = \begin{bmatrix} [I(\theta)]_{\mu,\mu} & [I(\theta)]_{\mu,\eta} & [I(\theta)]_{\mu,\sigma} \\ [I(\theta)]_{\mu,\eta} & [I(\theta)]_{\eta,\eta} & [I(\theta)]_{\eta,\sigma} \\ [I(\theta)]_{\mu,\sigma} & [I(\theta)]_{\eta,\sigma} & [I(\theta)]_{\sigma,\sigma} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{1-e^{-\eta d_i}}{1+e^{-\eta d_i}} & 0 & 0 \\ 0 & \sum_{i=1}^{n}\left(\frac{d_i^2 e^{-2\eta d_i}(1+e^{-2\eta d_i})}{(1-e^{-2\eta d_i})^2} - \frac{2}{\eta}\frac{d_i e^{-2\eta d_i}}{(1-e^{-2\eta d_i})} + \frac{1}{2\eta^2}\right) & -\frac{n}{\sigma\eta} + \frac{2}{\sigma}\sum_{i=1}^{n}\frac{d_i e^{-2\eta d_i}}{(1-e^{-2\eta d_i})} \\ 0 & -\frac{n}{\sigma\eta} + \frac{2}{\sigma}\sum_{i=1}^{n}\frac{d_i e^{-2\eta d_i}}{(1-e^{-2\eta d_i})} & \frac{2n}{\sigma^2} \end{bmatrix}$$

where $d_i = t_i - t_{i-1}$.

The proof of this proposition is in Supplemental Appendix B.1. In the case of equally spaced samples, we are able to simplify the Fisher information matrix. The following proposition is obtained by setting all $d_i = \Delta_t$, performing straightforward simplifications and inverting.

Proposition 2 *For an OU process with parameter vector* $\theta = (\mu, \eta, \sigma)^T$*, sampled at time points* $t_i = i\Delta_t$ *for some constant time spacing* Δ_t *and* $i = 0, \ldots, n$*, the inverse of the Fisher information matrix is given by*

$$I^{-1} = \frac{1}{n}\begin{bmatrix} \frac{\sigma^2}{2\eta}\frac{1+e^{-\eta\Delta_t}}{1-e^{-\eta\Delta_t}} & 0 & 0 \\ 0 & \frac{e^{2\eta\Delta_t}-1}{{\Delta_t}^2} & \sigma\left(\frac{e^{2\eta\Delta_t}-1}{2\eta{\Delta_t}^2} - \frac{1}{\Delta_t}\right) \\ 0 & \sigma\left(\frac{e^{2\eta\Delta_t}-1}{2\eta{\Delta_t}^2} - \frac{1}{\Delta_t}\right) & \sigma^2\left(\frac{(1+e^{-2\eta\Delta_t})}{2(1-e^{-2\eta\Delta_t})} - \frac{1}{\eta\Delta_t} + \frac{(e^{2\eta\Delta_t}-1)}{4\eta^2{\Delta_t}^2}\right) \end{bmatrix}$$

Determining Optimal Sampling

The parameter that best describes the temporal dynamics is η, the rate of mean reversion. Therefore, if our objective is to understand the temporal dynamics of the microbiome, accurate estimation of η is important. We will therefore focus on the sampling scheme that minimises $Var(\hat{\eta}) = [I(\theta)]^{-1}_{\eta,\eta}$. Theorem 3 gives the solution to this minimisation problem. The proof is in Appendix B.2.

Theorem 3 *(1) The optimal sampling scheme to minimise* $\mathrm{Var}(\hat{\eta})$ *under an OU process is to sample the observations with time difference infinitesimal with probability* p *and equal to* $d_i = \frac{s^\dagger}{2\eta}$ *with probability* $1-p$*, where the value* $s^\dagger$ *is the solution to*

$$2\left(\frac{1}{s^\dagger} - \frac{1}{e^{s^\dagger}-1}\right) - 1 = 2s^\dagger\left(\frac{1}{s^\dagger} - \frac{1}{e^{s^\dagger}-1}\right)^2\left(\frac{1}{s^\dagger} - \frac{1}{e^{s^\dagger}-1} - 1\right)$$

and

$$p = \frac{1}{2} - \frac{{s^\dagger}^2(e^{s^\dagger}-1)}{4(e^{s^\dagger}-1-s^\dagger)^2}$$

Numerically these values can be solved as $s^\dagger = 1.956493$ *and* $p = 0.1572033$*. For this optimal sampling scheme,* $\mathrm{Var}(\hat{\eta}) = \frac{6.12679\eta^2}{n}$.

(2) Let s_l be the solution to

$$s_l^2(e^{s_l}-1)-c_0(e^{s_l}-1)^2=2(b_0(e^{s_l}-1)-s_l)^2$$

with $c_0=\sup\frac{x^2}{e^x-1}=0.6476102$ and $b_0=1-\sqrt{1-c_0}=0.4063757$. Numerically, this is $s_l=0.5844618$. If samples from an OU process must be collected with time difference $d_i\geqslant\frac{s_l}{2\eta}$, then the optimal sampling scheme is to sample evenly-spaced observations with $d_i=\frac{s_0}{2\eta}$, where s_0 is the solution to

$$(2-s_0)(e^{s_0}-1)=s_0$$

Numerically, $s_0=1.59362426$, and for this sampling scheme $\mathrm{Var}\left(\frac{\hat{\eta}}{\eta}\right)=\frac{6.176555\eta^2}{n}$.

(3) If σ is known for an OU process, then the optimal sampling scheme to minimise $\mathrm{Var}(\hat{\eta})$ is to sample evenly-spaced observations with time difference $d_i=\frac{s_k}{2\eta}$ where s_k is the solution to

$$4(e^{s_k}-1)^2+s_k^2e^{s_k}(3+e^{s_k})=6s_ke^{s_k}(e^{s_k}-1)+2s_k(e^{s_k}-1)$$

Numerically, we get $s_k=5.109858$. In this case $\mathrm{Var}\left(\frac{\hat{\eta}}{\eta}\right)=\frac{1.964279\eta^2}{n}$.

The optimality of sampling with infinitesimal time differences is slightly counter-intuitive. Under an OU model, as time difference tends to zero, the Brownian motion term dominates, and therefore, samples with infinitesimal time differences are most efficient for estimating σ. Therefore, the optimal sampling scheme from Theorem 3(1) uses some samples which are very informative about σ, and others which give information about both σ and η. We can see from part (3) of the theorem that the remaining samples are still much more frequent than would be needed if we already knew σ. This makes sense, because in the OU process, the long-term variance is given by $\frac{\sigma^2}{2\eta}$, so if σ is known then we can estimate η, even using samples at very large time difference.

When using samples with small time differences, σ is estimated from the ratios $\frac{(X_{t_i}-X_{t_{i-1}})^2}{t_i-t_{i-1}}$. Because this involves differences between close quantities, it is very sensitive to any model misspecification, such as measurement error in t or X. Therefore, in practice, it may be sensible to limit the frequency of sampling as in part (2) of Theorem 3. With less frequent sampling, the OU model is likely to be a better fit for the data, so the sampling scheme from part (2) is likely to be more useful in practice. Even in the perfect case with no model misspecification, the loss from using equally spaced samples as in part (2) over the scheme given by part (1) is relatively small.

Simulation

The asymptotic normality of MLE theorem states that for a large enough sample size, the asymptotic behaviour of MLEs can be described by the Fisher information matrix. However, it does not specify what sample size is needed for this asymptotic approximation to be reasonable. We therefore conduct a simulation study to confirm that the asymptotic approximation can be used for realistic sample sizes.

Simulation Design

In order to test the estimated covariance matrix, we simulate data sets under an OU model, with $\mu = 0$, $\eta \in \{0.5, 1, 2\}$ and $\sigma \in \{1, 2, 4\}$, with time step $\Delta_t = 1$, with different sample sizes n. These values cover a range of scenarios similar to the values estimated from the real data. For each sample, we compute the MLEs $\hat{\eta}$, $\hat{\sigma}$ and $\hat{\mu}$. We compare the covariance matrices estimated from 100,000 simulations with the asymptotic covariance matrices predicted using the Fisher information matrix, using the following matrix dissimilarity measure

$$d^2(A, B) = \frac{\|A - B\|^2}{\|A\|^2}$$

where $\|A\|$ is the Euclidean norm of A, i.e. $\|A\|^2 = \sum_{i,j} {A_{ij}}^2$. We measure the dissimilarity $d^2(I^{-1}, \hat{V})$, where I is the Fisher information matrix, and $\hat{V}$ is the empirically estimated covariance matrix.

Simulation Results

The simulation results are shown in Table 8. We see that as expected, the observed covariance matrix gets closer to the Fisher information matrix as sample size n increases. (Note that the errors in the table are relative errors). For $\eta = 0.5$ and $\eta = 1$, the approximation becomes reasonably accurate, (with an error of about 10% in the covariance matrix), somewhere between sample size 100 and 500. Thus, for the moving picture data set, we expect the inverse of the Fisher information matrix to provide a fairly accurate estimate of the variance of the parameter estimates.

Table 8 Distance between inverse Fisher information matrix and sample covariance matrix for various sample sizes.

		n					
η	σ	5000	1000	500	100	50	10
0.5	1	0.054	0.053	0.052	0.058	0.125	0.996
	2	0.066	0.065	0.064	0.055	0.052	0.998
	4	0.066	0.065	0.064	0.055	0.040	0.995
1	1	0.008	0.008	0.011	0.139	0.330	0.968
	2	0.057	0.051	0.046	0.106	0.266	1.000
	4	0.122	0.109	0.099	0.056	0.121	0.853
2	1	0.001	0.033	0.147	0.078	0.021	12.281
	2	0.002	0.028	0.123	0.073	0.047	12.168
	4	0.005	0.020	0.102	0.045	0.098	7.586

For small sample sizes (mostly $n \leqslant 100$) there were a few simulations where $\hat{\eta} = \infty$, i.e. there is no estimated correlation between consecutive samples. These simulations were removed from the variance calculation.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Availability of data and material
The data analysed in this paper are from (Caporaso *et al.*, 2011), and are study 550 in *Qiita*.

Competing interests
The authors declare that they have no competing interests.

Funding
The first and third authors gratefully acknowledge funding from NSERC, grants RGPIN/4945-2014 and RGPIN-2017-05108, respectively. NSERC had no role in the design or interpretation of the study or the writing of the paper.

Authors' contributions
TK and HG conceived and designed the study. JG fitted the models and performed the likelihood ratio tests. JG calculated the Fisher Information. TK simplified the Fisher information calculation to the form presented in the Appendix. TK proved Theorem 3 about the optimal sampling scheme. TK interpreted the results of the analysis. TK designed the simulation. TK and JG ran the simulation. TK and HG wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements
We are grateful to our colleagues Lam Ho and Edward Susko for helpful suggestions on the project.

Additional Files
Additional File 1 — Supplementary Appendices
Appendix A provides detailed calculations of the log-likelihood calculations. Appendix B provides the detailed calculation of the Fisher information of the OU process, and proves the main theoretical result about the optimal sampling scheme. Appendix C provides additional results from the data analysis.

References

1. Celeste Allaband, Daniel McDonald, Yoshiki Vázquez-Baeza, Jeremiah J. Minich, Anupriya Tripathi, David A. Brenner, Rohit Loomba, Larry Smarr, William J. Sandborn, Bernd Schnabl, Pieter Dorrestein, Amir Zarrinpar, and Rob Knight. Microbiome 101: Studying, analyzing, and interpreting gut microbiome data for clinicians. *Clin. Gastroenterol. Hepatol.*, 17:218–230, 2019.
2. Fredrik Bäckhed, Hao Ding, Ting Wang, Lora V Hooper, Gou Young Koh, Andras Nagy, Clay F Semenkovich, and Jeffrey I Gordon. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences*, 101(44):15718–15723, 2004.
3. Peter J Bickel and Kjell A Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, 2015.
4. J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome biology*, 12(5):R50, 2011.
5. Elizabeth K Costello, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.
6. Lawrence A. David, Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, Sudha B. Biddinger, Rachel J. Dutton, , and Peter J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature.*, 505:559–563, 2014.
7. Morris H DeGroot and Mark J Schervish. *Probability and statistics*. Pearson Education, 2012.
8. B Roy Frieden. *Science from Fisher information: a unification*. Cambridge University Press, 2004.
9. Jonas Halfvarson, Colin J. Brislawn, Regina Lamendella, Yoshiki Vázquez-Baeza, William A. Walters, Lisa M. Bramer, Mauro D'Amato, Ferdinando Bonfiglio, Daniel McDonald, Antonio Gonzalez, Erin E. McClure, Mitchell F. Dunklebarger, Rob Knight, and Janet K. Jansson. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, 2:17004, 2017.
10. Ioannis Karatzas and Steven E Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.
11. Jeremy E Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D Fricker, Jesse Stombaugh, Rob Knight, Largus T Angenent, and Ruth E Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585, 2011.
12. Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.
13. Don S Lemons and Paul Langevin. *An introduction to stochastic processes in physics*. JHU Press, 2002.
14. Thomas T MacDonald and Sven Pettersson. Bacterial regulation of intestinal immune responses. *Inflammatory bowel diseases*, 6(2):116–122, 2000.
15. Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.
16. Chana Palmer, Elisabeth M Bik, Daniel B DiGiulio, David A Relman, and Patrick O Brown. Development of the human infant intestinal microbiota. *PLoS biology*, 5(7):e177, 2007.
17. Victoria Pascal, Marta Pozuelo, Natalia Borruel, Francesc Casellas, David Campos, Alba Santiago, Xavier Martinez, Encarna Varela, Guillaume Sarrabayrouse, Kathleen Machiels, Severine Vermeire, Harry Sokol, Francisco Guarner, and Chaysavanh Manichanh. A microbial signature for crohn's disease. *Gut*, 66:813–822, 2017.
18. Luigi M Ricciardi and Laura Sacerdote. The ornstein-uhlenbeck process as a model for neuronal activity. *Biological cybernetics*, 35(1):1–9, 1979.

19. Rori V Rohlfs, Patrick Harrigan, and Rasmus Nielsen. Modeling gene expression evolution with an extended ornstein–uhlenbeck process accounting for within-species variation. *Molecular biology and evolution*, 31(1):201–211, 2013.
20. Leung Tim Siu-tang and Li Xin. *Optimal mean reversion trading: Mathematical analysis and practical applications*, volume 1. World Scientific, 2015.
21. A. Szimayer and R. Maller. Testing for mean reversion in processes of ornstein-uhlenbeck type. *Statistical Inference for Stochastic Processes*, 7:95–113, 2004.
22. Yoshiki Vázquez-Baeza, Antonio Gonzalez, Zhenjiang Zech Xu, Alex Washburne, Hans H Herfarth, R Balfour Sartor, and Rob Knight. Guiding longitudinal sampling in ibd cohorts. *Gut*, 67:1743–1745, 2018.
23. Gary D. Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A. Keilbaugh, Meenakshi Bewtra, Dan Knights, William A. Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D. Bushman, and James D. Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334:105–108, 2011.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.pdf