MC-SPACE: Microbial communities from spatially associated counts engine

Anonymous Authors[1]

Abstract

Elucidating the biogeography of the gut microbiome is critical for understanding how the trillions of microbes that live in our intestines form complex communities and interact to maintain human health, or when disrupted, contribute to disease. A new technology, metagenomic plot sampling by sequencing (MaPS-seq), provides unprecedented micron-scale spatial data of entire microbiomes. However, the data is noisy and high-dimensional, making direct interpretation difficult. Here we present MC-SPACE, a Bayesian model that infers mixtures of spatially coherent microbial community subtypes and alterations in their prevalence due to perturbations, from MaPS-seq data. We apply MC-SPACE to a fecal microbiota transplantation (FMT) mouse study and find distinct microbial communities from donor mice that are spatially coherent and engraft into recipient mice, causing significant spatial restructuring of gut microbiomes. Our results highlight the ability of MC-SPACE to infer spatial microbiome structure from high-throughput data, and yield insights into the spatial dynamics of microbial colonization of the gut, which has potential to improve treatment for human diseases responsive to FMT.

1. Introduction

The microbiome, or collection of commensal microorganisms that live on and within us, is extremely complex and plays key roles in many prevalent human diseases, such as infectious (Van Nood et al., 2013), autoimmune, and other diseases. Thus, many efforts are underway to manipulate the microbiome for therapeutic purposes. For therapies to be effective, they must interact with a complex pre-existing microbial ecology. Spatial associations between microbes are thought to be important in these ecologies, because they influence many key factors such as maintenance of biodiversity (Reichenbach et al., 2007), microbe interactions with each other (Cordero & Datta, 2016) and their host, as well as the stability and plasticity of the microbiome (Bucci et al., 2016; Lee et al., 2022; Olsson et al., 2022).

However, the biogeography of the microbiome is relatively unexplored, in part due to technological limitations. Current imaging technologies require extensive experimental optimization (Amann & Fuchs, 2008; Mark Welch et al., 2017), are limited to profiling small numbers of targeted microbes often at low taxonomic resolution (Valm et al., 2012), and are challenging to scale to complex and diverse natural microbiomes. To address these issues, MaPS-seq (Sheth et al., 2019) was recently developed. The core idea behind MaPS-seq is to "freeze" microbes in place, barcode spatially proximate microbes in particles (typically 10 to 30 μm in diameter) and then disaggregate the material and interrogate it with high-throughput sequencing. MaPS-seq data has particular noise characteristics that make analysis challenging, including uneven particle amplification, variable read depth, and mixing effects, likely due to unencapsulated DNA contaminating particles.

To address these challenges, we developed MC-SPACE, a Bayesian model. Our contributions include: (1) automatic discovery of parsimonious spatially co-occurring groups of microbes, which we term *community subtypes*; (2) a noise model specifically tailored to MaPS-seq data, and (3) inference of changes in community subtype abundances due to experimental perturbations. Below, we first present experiments with semi-synthetic data to benchmark MC-SPACE against standard methods. We then apply MC-SPACE to a new mouse FMT dataset to demonstrate our method's ability to uncover microbial spatial dynamics.

Prior work. Current methods for analyzing MaPS-seq data mainly focus on detecting pairwise associations by binarizing operational taxanomic units (OTUs) in each particle and comparing to a null model of co-occurrence (Sheth et al., 2019; Urtecho et al., 2022). These methods fail to capture multiple associations in a community. A Gaussian mixture model was developed for MaPS-seq data in (Pasarkar et al., 2021) that recovers clusters of OTUs. However, none of these methods model the actual measurement noise in MaPS-seq data, and cannot capture changes due to perturbations,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

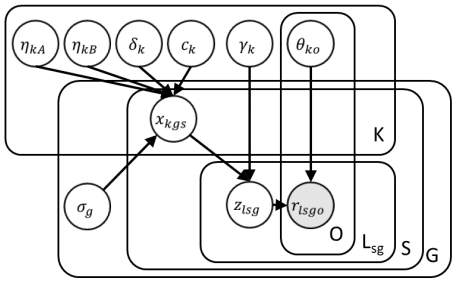

Figure 1. **The MC-SPACE model depicted using plate notation.**

which is key to elucidating microbial spatial dynamics.

2. Methods

MC-SPACE is a sparse Bayesian mixture model (Figure 1) for discovering latent groups of spatially co-localized microbes, or community subtypes, as well as effects due to perturbations, from MaPS-seq data. The model considers pre-perturbation (A) and post-perturbation groups (C), as well as optional comparator groups (B). We model a common set of community subtypes for all groups. Assume we have O OTUs, S biological replicates, L_g particles per group, and a maximum of K community types. The generative process is then as follows:

1. Sample components $\theta_{ko} \sim \mathcal{LN}(0, 1)$ for each community subtype k and OTU o
2. Sample latent mixture weights $x_{kgs} \sim N(\eta_{kg}, \sigma_g^2)$ for each group g and subject s
3. Sample sparsity indicators $\gamma_k \sim \text{Bern}(\pi_\gamma)$ and compute community mixture weights $\beta_{kgs} = \frac{\gamma_k \exp(x_{kgs})}{\sum_j \gamma_j \exp(x_{jgs})}$
4. For each observed particle l in group g and subject s:
   (a) Sample a community type, $z_{lsg} \sim \text{Cat}(\boldsymbol{\beta}_{gs})$
   (b) Sample reads $r_{lsgo} \sim \text{Mult}(R_{lsg}, (1-\pi_g)\boldsymbol{\theta}_{z_{lsq}} + \pi_g \boldsymbol{B}_g)$ for all R_{lsg} reads in the particle

Here $\mathcal{LN}$ denotes the logistic-normal distribution. Group means η_{kA} and η_{kB} are sampled from a standard Normal prior. We explicitly model the effects of perturbations by linking the pre- and post-perturbation group means: $\eta_{kC} = \eta_{kA} + \delta_k c_k$, where $\delta_k \sim N(0, \rho_\delta^2)$ specify perturbation magnitudes and indicators $c_k \sim \text{Bern}(\pi_c)$ specify whether the perturbation effect is present for community k.

Community Mixing. To account for the mixing effect in MaPS-seq data, we added contaminating communities with parameters $B_{go} = \frac{\sum_s \sum_{l \in L_g} r_{lsgo}}{\sum_s \sum_{l \in L_g} \sum_j r_{lsgj}}$. This assumes the mixing effect corresponds to contamination from the bulk read distribution. Reads are then sampled from a mixture distribution, with weights π_g learned during inference.

3. Inference

Inference for latent mixture variables $\boldsymbol{\theta}$ and community proportions $\boldsymbol{x}$ was performed using amortized variational inference (Gershman & Goodman, 2014; Kingma & Welling, 2014), where parameters of approximating distributions are functions of the data. Specifically, we constructed inference networks that take a normalized representation $\tilde{r}_{lsg}$ of the data as input, and output the parameters for Gaussian approximating distributions. Normalized reads for each particle l are first passed into a fully-connected MLP encoder network. The outputs of the encoder are then averaged over all particles for all subjects, which is then passed through linear layers that output the mean and variance parameters of the approximating distribution. KL terms were computed analytically for all variables except for x_{kgs}, which we approximated using a Stochastic Gradient Variational Bayes estimator.

To balance the sparsity inducing-prior for γ with the data likelihood, we introduced an adjustable parameter ξ that multiples the corresponding KL term. This is similar to the scale factor used in the β-VAE model to learn more interpretable disentangled latent factors (Higgins et al., 2017). This scale parameter was chosen to give a stable clustering as we describe next.

Stability metric. To determine the setting for ξ, we took a stable clustering approach (Von Luxburg et al., 2010) and constructed a metric inspired by (Duan et al., 2019), originally developed for unsupervised disentanglement ranking for VAEs. Specifically, we trained our model on a range of ξ values, with 5 different initial seeds each. We then computed a similarity matrix F^{ij} for each pair of seeds i, j for a given setting of ξ. Each entry of R^{ij} is given by the Spearman correlation $F_{ab}^{ij} = \text{SpCorr}(z_{ia}(l), z_{jb}(l))$, where $z_{ia}(l)$ is the posterior probability of particle l being assigned to community a for model i. We then computed a stability score S_{ij} as, $S_{ij} = \frac{1}{d_a + d_b}\left[\sum_b \frac{r_a^2 \cdot \gamma_b}{\sum_a F_{ab}^{ij}} + \sum_a \frac{r_b^2 \cdot \gamma_a}{\sum_b F_{ab}^{ij}}\right]$, where $f_a = \max_a F_{ab}^{ij}$ and $d_a = \sum_a \gamma_a$. This metric will be larger for pairs of models that cluster particles into similar clusters, and smaller for models where the particle assignment is less stable across different seeds.

4. Results

4.1. FMT dataset

We applied MC-SPACE to a mouse FMT dataset described in (Urtecho et al., 2022). Prior studies had revealed consistent distinct microbial compositions across mice from dif-

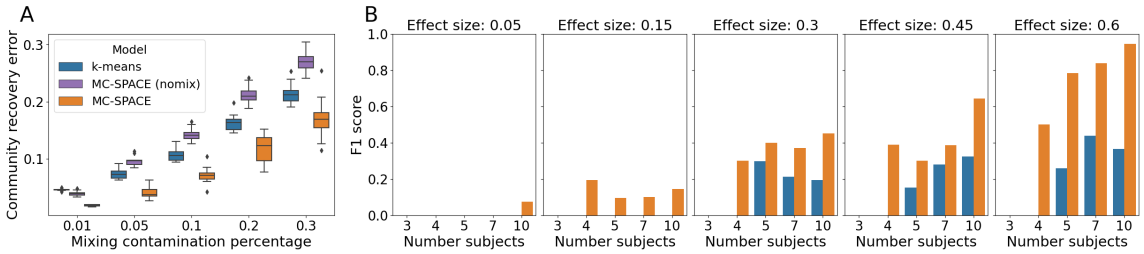

Figure 2. **Results on semi-synthetic data.** (A) Community reconstruction error for varying amounts of contamination noise. MC-SPACE (nomix) is the model without contamination communities included. Typical contamination percentages for real MaPS-seq data are around 0.05. (B) Comparison between k-means and MC-SPACE in detecting significant perturbations to community abundances. Effect sizes correspond to percent change in community distribution due to perturbation. For comparison, effect sizes detected on our real FMT data were 0.14 and 0.45, for 4 subjects. All results are from 10 datasets simulated with $K = 6$ underlying communities.

ferent vendors, with Jackson labs (Jax) mice having lower ecological diversity and Envigo mice showing higher diversity. Envigo mice flora were able to robustly invade mice from other vendors, and were resistant to invasion, while Jax mice were among the most susceptible to invasion. MaPS-seq was therefore used to investigate the spatial colonization dynamics of Jax mice receiving Envigo flora FMTs. The dataset consists of MaPS-seq data for 3 mice groups: Jax (pre-perturbation group), Envigo (comparator group), and Env2Jax (post-perturbation group, Envigo flora → Jax FMT), each with 4 biological replicates. After quality filtering, each sample consisted of a median of 677 particles (IQR=226.5) and 160 OTUs. Particle read depth had a median of 1798 reads (IQR=1867). Note the highly skewed and variable number of reads per particle, highlighting the noisy nature of MaPS-seq data.

4.2. Benchmarking with semi-synthetic data

To assess MC-SPACE's ability to recover underlying community structure and perturbation effects, we compared MC-SPACE to a *k*-means clustering method on semi-synthetic data. Semi-synthetic data was simulated from the communities MC-SPACE learned on the FMT dataset (Figure 3) using a bootstrapping-type procedure. Briefly, we randomly sampled $K = 6$ communities and their proportions $\boldsymbol{\beta}$ with replacement from inferred communities and randomly permuted OTU labels to avoid duplicated communities and preserve distributional properties. For perturbation experiments, we perturbed a randomly selected community with a fixed effect size, where the effect size is the change in abundance of a community due to perturbation. Subjects were then sampled with variances set to values inferred on real data. We then sampled 700 particles for each subject (corresponding to approximately the number of particles per subject in the real dataset), with read depths sampled from a negative binomial distribution fitted to the read depth distribution of the original dataset after filtering. Reads for each particle were then sampled from a mixture distribution with varying contamination mixture weights.

We first compared each model's ability to recover underlying communities on semi-synthetic data with varying levels of mixing contamination. As the true number of communities in real datasets is usually unknown, we ran each model with 20 communities as a conservative overestimate of the true number. To use *k*-means, we first applied a centered log-ratio (CLR) transformation to the reads normalized to relative abundance. The communities for *k*-means were then obtained by applying the softmax function, which is the inverse of the CLR transformation, to the cluster centers and converting them back to relative abundances. The reconstruction error was then calculated as $E = \frac{1}{K}\sum_k H(\boldsymbol{\theta}_k, \tilde{\boldsymbol{\theta}}_{c(k)})$, where $H(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{\sqrt{2}}\sqrt{\sum_i(\sqrt{a_i} - \sqrt{b_i})^2}$ is the Hellinger distance, $\boldsymbol{\theta}_k$ are ground truth communities, $\tilde{\boldsymbol{\theta}}_k$ are model inferred communities, and the map $c(k) = \underset{i}{\text{argmin}}\, H(\tilde{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_k)$. We found that without contamination communities MC-SPACE (nomix) was overly sensitive to contamination noise and obtained worse fits at high levels of contamination. In contrast, the full model, which explicitly addresses mixing contamination, consistently outperformed *k*-means (Figure 2A).

We next compared MC-SPACE to *k*-means in detecting significant perturbations. We varied the number of subjects and effect sizes around values inferred from the real FMT dataset. Contamination mixture weights were set to values inferred on real data ($\pi_A = 0.0005$, $\pi_B = 0.002$, $\pi_C = 0.04$). To focus our analysis on each model's ability to detect perturbations, we ran each model with the true number of communities $K = 6$. For *k*-means, the inferred community distribution β_{kgs} is given as the proportion of particles assigned to community k for each group g and subject s. Significant changes between pre- and post-perturbation groups were then detected using a Wilcoxon rank-sum test followed by Benjamini-Hochberg (BH) correction. Changes with $p < 0.05$ were taken to be significant. For MC-SPACE, posterior probabilities $p(c_k|D)$ were converted to binary

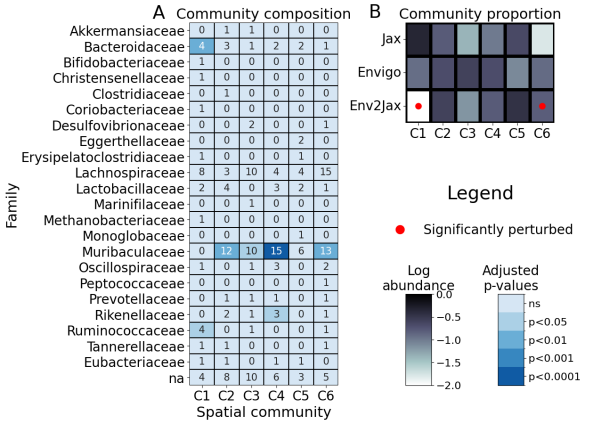

Figure 3. **MC-SPACE results for FMT study.** (A) Family level membership of learned communities and enrichment. Values correspond to number of OTUs present, with darker colors representing significant enrichment. (B) Proportions of learned communities in each group and significantly perturbed communities.

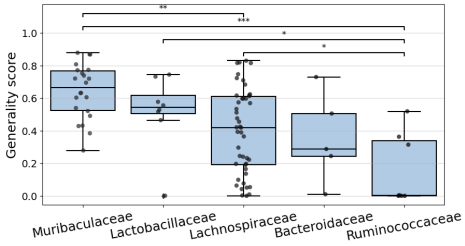

Figure 4. **Distribution of generality scores of OTUs from top 5 families.** Statistical significance was assessed with a Wilcoxon rank sum test followed by Benjamini Hotchberg correction (* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$).

outcomes for perturbation effects by applying a threshold at probability 0.5. Performance was then evaluated using the F1-score. As shown in (Figure 2B), MC-SPACE outperformed k-means in detecting perturbations across all effect sizes and number of subjects.

4.3. MC-SPACE learns perturbation effects on spatial communities in an FMT study

We analyzed the mouse FMT dataset with MC-SPACE to gain biological insights into the effects of FMT on the mammalian gut microbiome. The most stable clustering obtained by MC-SPACE consisted of 6 community subtypes (Figure 3). We first assessed the composition and enrichment of each community at the Family level. The presence of an OTU o in a community k was determined by using a threshold of $\theta_{ko} > 0.005$, equal to the threshold used for initial data filtering (Urtecho et al., 2022). To perform an enrichment analysis, we used the hypergeometric test followed by a BH correction. We found many of the communities present in all mice were significantly enriched in the Muribaculaceae Family. Community C1 was enriched in Bacteroidaceae and Ruminococcaceae (Figure 3A).

Of the 6 learned communities, MC-SPACE detected 2 as significantly perturbed (Figure 3B): C1 was significantly suppressed, and C6 was significantly enhanced post-FMT. Interestingly, C1 was the most abundant in Jax mice, making up 45% of the community distribution, and contained no taxa from the Muribaculaceae Family. In contrast, C1 is essentially absent in Env2Jax mice, and the remaining spatial communities post-FMT are significantly enriched in the Muribaculaceae Family and also contain some Lachnospiraceae Family OTUs (Figure 3A).

We next sought to further understand taxa responses to perturbations by examining their generalist behavior. Generalist behavior may allow taxa to more robustly engraft and colonize multiple community types, as they can utilize multiple resources and adapt to changes. In contrast, we expect more specialized taxa to be part of fewer communities and more susceptible to changes from perturbations. To quantify generalist behavior, we computed generality scores (Gerber et al., 2007), or normalized entropies, for each OTU o in the 5 most abundant families (Figure 4): $\text{GenS}_o = -\frac{1}{\log K}\sum_k h_{ko} \log h_{ko}$, where $h_{ko} = \frac{\theta_{ko}}{\sum_j \theta_{jo}}$. This score gives a measure of how much an OTU participates in multiple communities. Of the most abundant families, Muribaculaceae were the most generalized (Figure 4). This aligns with existing research suggesting that many Muribaculaceae are generalists and are able to utilize a diverse set of mucus-derived sugars (Pereira et al., 2020). In contrast, Ruminococcaceae were the least generalized, and primarily belonged to community C1. Many Ruminococcaceae were displaced post-FMT, which may be due to competition with invading taxa or changes in local nutrients post-FMT and their inability to adapt to new resources.

5. Discussion and future work

Overall, our results suggest MC-SPACE's noise model and explicit modeling of perturbation effects provide significant advantages in uncovering biologically relevant spatial dynamics from MaPS-seq data. There are multiple possibilities for future work, such as using MC-SPACE to inform dynamical systems models for determining interactions between microbes, or incorporating prior biological knowledge such as phylogenetic relationships into the model. A current limitation of the model is in determining which taxa are present/absent in a community in a threshold independent manner. This can be addressed by explicitly modeling OTU sparsity in each community subtype. The variational inference framework also allows for easy generalization of MC-SPACE to other types of studies such as time-series experiments involving perturbations and multiple subjects.

References

Amann, R. and Fuchs, B. M. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology*, 6(5):339–348, 2008.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney, M. L., Liu, Q., et al. Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome biology*, 17:1–17, 2016.

Cordero, O. X. and Datta, M. S. Microbial interactions and community assembly at microscales. *Current opinion in microbiology*, 31:227–234, 2016.

Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C. P., Lerchner, A., and Higgins, I. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*, 2019.

Gerber, G. K., Dowell, R. D., Jaakkola, T. S., and et al. Automated discovery of functional generality of human gene expression programs. *PLoS Comput Biol*, 3(8):e148, 2007.

Gershman, S. and Goodman, N. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Kingma, D. P. and Welling, M. Stochastic gradient vb and the variational auto-encoder. In *Second international conference on learning representations, ICLR*, volume 19, pp. 121, 2014.

Lee, J.-Y., Tsolis, R. M., and Bäumler, A. J. The microbiome and gut homeostasis. *Science*, 377(6601): eabp9960, 2022.

Mark Welch, J. L., Hasegawa, Y., McNulty, N. P., Gordon, J. I., and Borisy, G. G. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proceedings of the National Academy of Sciences*, 114(43):E9105–E9114, 2017.

Olsson, L. M., Boulund, F., Nilsson, S., Khan, M. T., Gummesson, A., Fagerberg, L., Engstrand, L., Perkins, R., Uhlén, M., Bergström, G., et al. Dynamics of the normal gut microbiota: A longitudinal one-year population study in sweden. *Cell Host & Microbe*, 30(5):726–739, 2022.

Pasarkar, A. P., Joseph, T. A., and Pe'er, I. Directional gaussian mixture models of the gut microbiome elucidate microbial spatial structure. *Msystems*, 6(6):e00817–21, 2021.

Pereira, F. C., Wasmund, K., Cobankovic, I., Jehmlich, N., Herbold, C. W., Lee, K. S., Sziranyi, B., Vesely, C., Decker, T., Stocker, R., et al. Rational design of a microbial consortium of mucosal sugar utilizers reduces clostridiodes difficile colonization. *Nature Communications*, 11(1):5104, 2020.

Reichenbach, T., Mobilia, M., and Frey, E. Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. *Nature*, 448(7157):1046–1049, 2007.

Sheth, R. U., Li, M., Jiang, W., Sims, P. A., Leong, K. W., and Wang, H. H. Spatial metagenomic characterization of microbial biogeography in the gut. *Nature biotechnology*, 37(8):877–883, 2019.

Urtecho, G., Moody, T. M., Huang, Y., Sheth, R. U., Richardson, M., Lekan, O., Velez-Cortes, F., Ricaurte, D., and Wang, H. H. Spatiotemporal dynamics during niche remodeling by super-colonizing microbiota in the mammalian gut. *bioRxiv*, pp. 2022–10, 2022.

Valm, A. M., Welch, J. L. M., and Borisy, G. G. Clasifish: principles of combinatorial labeling and spectral imaging. *Systematic and applied microbiology*, 35(8): 496–502, 2012.

Van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., de Vos, W. M., Visser, C. E., Kuijper, E. J., Bartelsman, J. F., Tijssen, J. G., et al. Duodenal infusion of donor feces for recurrent clostridium difficile. *New England Journal of Medicine*, 368(5):407–415, 2013.

Von Luxburg, U. et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3): 235–274, 2010.