Integrating Sample Similarities into Latent Class Analysis: A Tree-Structured Shrinkage Approach

Mengbing Li[1], Daniel Park[3], Maliha Aziz[3], Cindy M Liu[3], Lance Price[3], Zhenke Wu[1,2*]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[2]Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI 48109, USA

[3]Environmental and Occupational Health, Milken Institute School of Public Health,

The George Washington University, Washington, DC 20052

**email:* zhenkewu@umich.edu

SUMMARY: This paper is concerned with using multivariate binary observations to estimate the proportions of unobserved classes with scientific meanings. We focus on the setting where additional information about sample similarities is available and represented by a rooted binary weighted tree. Leaves in the given tree represent groups of observations with shorter distances between them indicating higher similarity. We propose a novel data integrative extension to classical latent class models (LCMs) with tree-structured shrinkage that enables 1) borrowing of information across leaf nodes, 2) data-driven groupings of observations with distinct vectors of class proportions, and 3) individual-level probabilistic class assignment given the observed multivariate binary measurements. We derive and implement a scalable posterior inference algorithm in a variational Bayes framework. Extensive simulations show more accurate estimation of class proportions than alternatives based on suboptimal use of the additional sample similarity information. We demonstrate the method by using mobile genetic elements to estimate the proportions of unobserved zoonotic *E. coli* isolates mapped over a phylogenetic tree which summarizes core-genome similarities. Model limitations and extensions are also discussed.

KEY WORDS: Gaussian Diffusion; Latent Class Models; Phylogenetic Tree; Zoonotic Infectious Diseases; Spike-and-Slab Prior; Variational Bayes.

1. Introduction

1.1 *Motivating Application*

The fields of infectious disease epidemiology and microbial ecology need better tools for tracing the transmission of microbes between humans and other vertebrate animals (i.e., zoonotic transmissions). The COVID-19 pandemic provides a monumental example of the devastating potential of zoonotic transmissions, but each year less staggering outbreaks and sporadic transmissions cumulatively exact a heavy burden on society resulting in numerous infections, hospitalizations, and deaths in the United States (CDC, 2019). The transmission patterns of frank zoonotic pathogens such as *Salmonella* and SARS-CoV2 (the virus that causes COVID-19) can be traced by disease cases, which often present within a predictable period after a successful exposure. In contrast, the epidemiology of colonizing opportunistic pathogens (COPs), such as *Escherichia coli* (*E. coli*), *Staphylococcus aureus* and *Enterococcus spp.*, can be much more cryptic due to their ability to asymptomatically colonize the human body for indefinite periods prior to initiating an infection, transmitting to another person, or being shed without a negative outcome (e.g., Price et al., 2017). Some COPs can colonize many different vertebrate hosts and cross-species transmissions can go unrecognized. Estimating the probability of clinical isolates of zoonotic origin in the population and for each isolate would provide important insights into the natural history of infections and inform more effective intervention strategies, such as eliminating high-risk clones from livestock via vaccination.

Core-genome phylogenetic analysis, the current gold standard for infectious disease investigations, has limitations for analyzing the epidemiology of COP infections (e.g., Besser et al., 2019). The isolates that comprise an infectious disease outbreak are typically derived from a recent common ancestor and can be identified as being part of an outbreak based on a small number of single nucleotide polymorphisms (SNPs) in the core genome (i.e., the

regions of the genome shared by all isolates). In contrast, the strains that comprise the annual sporadic disease burden for a COP species can be extremely diverse and would require an unachievable sampling effort of humans and animals to resolve their origins based on core-genome phylogenetics. Whereas the core genome includes the genes essential to life of an organism irrespective of host or setting, the accessory genome is comprised of genes and mobile genetic elements (MGEs) that enable niche adaptation (e.g., Lindsay and Holden, 2004). Host-associated accessory elements that provide selective advantages in particular hosts may be lost and gained as COPs transmit among hosts. Interrogating an isolate's repertoire of host-associated accessory elements could reveal its host origins and differentiate recent zoonotic spillover events, which may represent actionable transmission phenomena, versus historic host switch events, which are no longer actionable.

Recent research on two COP species, *E. coli* and *S. aureus*, has demonstrated the utility of complementing core-genome phylogenetic trees with host-associated MGEs to resolve host origins (e.g., Liu et al., 2018; Sieber et al., 2018). However, in both cases only a single host-associated MGE was used and analysis was largely limited to visual inspection of how each element fell on the scaffold of the evolutionary tree with leaf nodes representing distinct core-genome multi-locus sequence types (STs, Maiden et al., 1998). For this approach to reach its full potential, a statistical model must be developed that integrates phylogenetic information with the presence and absence of multiple host-associated MGEs to estimate the probability with which the isolates were derived from a particular host in each ST-specific population and for each individual isolate.

1.2 *Integrating Sample Similarities into Latent Class Analysis*

Latent class models (e.g., Lazarsfeld, 1950; Goodman, 1974) is designed to uncover unobserved classes of observations with distinct scientific meanings using multiple categorical measures. LCMs are examples of latent variable models that assume the observed dependence

among multivariate responses is induced by variation among unobserved or "latent" variables. Any multivariate discrete data distribution can be approximated arbitrarily closely by an LCM with a sufficiently large number of classes (Dunson and Xing, 2009, Corollary 1). The most commonly used LCMs assume the class membership indicators for the observations are drawn from a population with the same vector of class proportions.

In this paper, we will focus on latent class analysis of multivariate binary responses with additional sample similarity information represented by a rooted binary weighted tree. We assume known entities at the leaves. Each leaf may contain multiple independent observations or samples, each associated with the multivariate binary responses which are then combined to form the rows of a binary data matrix $\mathbf{Y}$. In the motivating application, the latent class represents the unobserved host origin (human or non-human) to be inferred by the presence or absence of multiple MGEs. The additional sample similarity information is represented by a maximum likelihood phylogenetic tree (e.g., Scornavacca et al., 2020). The leaf nodes represent distinct contemporary core-genome *E. coli* STs.

To integrate the tree-encoded sample similarity information into a latent class analysis, ad hoc groupings of the leaf nodes may be adopted. From the finest to the coarsest leaf grouping, one may 1) analyze data from distinct lineages one at a time, 2) manually form groups of at least one leaf node and fit separate LCMs, or 3) fit all the data by a single LCM. However, all these methods pose significant statistical challenges. First, separate latent class analysis may have low accuracy in estimating latent class proportions and other model parameters for rare lineages. Second, observations of similar lineages may have similar propensities in host jump resulting in similar host origin class proportions. Modeling these similarities could lead to gain in statistical efficiency. Finally, approaches based on coarse ad hoc groupings may obscure the study of the variation in the latent class proportions across different parts of the tree.

Fully probabilistic approaches that integrate tree-structured information into latent variable models appeared primarily in topic modeling literature (e.g., Airoldi and Bischof, 2016; Ghahramani et al., 2010). However, these authors did not consider tree-encoded sample similarities. Thomas et al. (2019) studied tree-structured shrinkage in the context of matched logistic regression. In this paper, we focus on integrating the tree-encoded sample similarity information into latent class analysis. We assume the tree information is given and not computed from the multivariate binary measurements.

1.3 *Primary Contributions*

In this paper, we propose a tree-integrative LCM framework to 1) discover groups of observations where the multivariate binary measurements have distinct vectors of latent class proportions; 2) accurately estimate the latent class probabilities for each discovered leaf group and assign probabilities of an individual sample belonging to a few latent classes; 3) leverage the relationship among the observations to boost the accuracy of the estimation of latent class proportions. Without pre-specifying the group, the automatic data-driven approach enjoy robustness by avoiding potential mis-specification of the grouping structure. On the other hand, the discovered data-driven groups dramatically reduce the dimension of leaf nodes into fewer homogeneous subgroups of leaf nodes hence improving interpretation. In addition, the proposed approach shows better accuracy in estimating the latent class proportions in terms of root mean squared errors, indicating the advantage of the shrinkage. On posterior computation, we derive a scalable posterior inference algorithm based on variational inference (VI). The VI algorithm also overcomes previously reported issues regarding spike-and-slab priors (George and McCulloch, 1997).

The rest of the paper is organized as follows. Section 2.2 defines tree-related terminologies and formulates LCMs. Section 3 proposes the prior for tree-structured shrinkage in LCMs. Section 4 derives a variational Bayes algorithm for inference. Section 5 compares the perfor-

mances of the proposed and alternative approaches via simulations. Section 6 illustrates the approach by analyzing an *E. coli* data set. The paper concludes with a brief discussion.

2. Model

We first introduce necessary terminologies and notations to describe a rooted binary weighted tree. LCMs are then formulated for data on the leaf nodes of the tree.

2.1 *Rooted Binary Weighted Trees*

A rooted binary tree is a graph $\mathcal{T} = (\mathcal{V}, E)$ with node set $\mathcal{V}$ and edge set E where there is a root node u_0 and each node has at most two child nodes. Let $p = |\mathcal{V}|$ represent the total number of leaf and non-leaf nodes. Let $\mathcal{V}_L \subset \mathcal{V}$ be the set of leaf nodes, and $p_L = |\mathcal{V}_L| < p$. We typically use u to denote any node ($u \in \mathcal{V}$) and v to denote any leaf node ($v \in \mathcal{V}_L$). Each edge in a binary tree defines a *clade*: the group of leaf nodes below it. Splitting the tree at an edge creates a partition of the leaf nodes into two groups. For any node $u \in \mathcal{V}$, the following notations apply: $c(u)$ is the set of offspring of u; $pa(u)$ is the parent of u; $d(u)$ is the set of descendants of u including u, and; $a(u)$ is the set of ancestors of u including u. In Figure 3(a), if $u = 2$, then $c(u) = \{6, 7, 8\}$, $pa(u) = \{1\}$, $d(u) = \{2, 6, 7, 8\}$, and $a(u) = \{1, 2\}$. The phylogenetic tree in our motivating application is a nested hierarchy of the STs for $N = 3,126$ observations, where the $p_L = |\mathcal{V}_L| = 133$ leaves represent distinct STs and the $p - p_L = 132$ internal (non-leaf) nodes represent ancestral "species" or speciation events leading up to the leaf descendants.

Edge-weighted graphs appear as a model for numerous problems where nodes are linked with edges of different weights such as distance. In particular, the edges in $\mathcal{T}$ are attached with weights where $w : E \to \mathbb{R}^+$ is a weight function. Let $(\mathcal{T}, w)$ be a rooted binary weighted tree. A path in a graph is a sequence of edges which joins a sequence of distinct vertices. For a path P in the tree connecting two nodes, $w(P)$ is defined as the sum of all the edge

weights along the path, often referred to as the "length" of P. The distance between two vertices u and u', denoted by $dist_{\mathcal{T},w}(u,u')$ is the length of a shortest (with minimum length) (u,u')-path. $dist(\mathcal{T},w)$ is a distance: it is symmetric and satisfies the triangle inequality. In our motivating application, the distance between two nodes provides a measurement of the similarity or divergence between any two core-genome sequences of the input set. In this paper, we use h_u to represent the edge length between a node u and its parent node $pa(u)$. For the root node u_0, there are no parents, i.e. $pa(u_0)=\emptyset$; we set $h_{u_0}=1$.

2.2 *Latent Class Models for Data on the Leaf Nodes*

Although LCMs can deal with multiple categorical responses in general, in this paper, we focus on presenting the model and algorithm using multivariate binary responses and their application to the motivating data. Let $\mathbf{Y}_i^{(v)}=(Y_{i1}^{(v)},\ldots,Y_{iJ}^{(v)})^{\mathsf{T}}\in\{0,1\}^J$ be the vector of binary responses for observation $i\in[n_v]$ that is nested within leaf node $v\in\mathcal{V}_L$. Throughout this paper $[Q]:=\{1,\ldots,Q\}$ represents the set of positive integers smaller than or equal to Q. Let $N=\sum_{v\in\mathcal{V}_L}n_v$ represent the total sample size. Let $\mathbf{Y}=\{Y_{ij}^{(v)},i\in[n_v],v\in\mathcal{V}_L,j\in[J]\}$ represent the binary data matrix. The LCM is specified in two steps:

$$\textsf{class indicator}: \quad I_i^{(v)}\mid\boldsymbol{\pi}_v\sim\textsf{Categorical}_K\left\{\boldsymbol{\pi}_v\right\},\boldsymbol{\pi}_v\in\mathcal{S}_{K-1}, \tag{1}$$

$$\textsf{data}: \quad Y_{ij}^{(v)}\mid I_i^{(v)}\sim\textsf{Bern}\left\{\theta_{j,I_i^{(v)}}\right\},\ \text{independently for feature } j\in[J], \tag{2}$$

and independently for observation $i\in[n_v]$ and leaf node $v\in\mathcal{V}_L$. Let $\mathbf{I}=\{I_i^{(v)}:i\in[n_v];v\in\mathcal{V}_L\}$ represent the latent class indicators and $Z_{ik}^{(v)}=\mathbf{1}\{I_i^{(v)}=k\}$, $k\in[K]$; Here $\mathbf{1}\{A\}$ is an indicator function which equals 1 if statement A is true and 0 otherwise. We have assumed observations in different leaf nodes v have potentially different vectors of class proportions $\boldsymbol{\pi}_v=(\pi_{v1},\ldots,\pi_{vK})^{\mathsf{T}}\in\mathcal{S}_{K-1}$, $v\in\mathcal{V}_L$. Here $\mathcal{S}_{K-1}=\{\boldsymbol{r}\in[0,1]^K:\sum_{k=1}^K r_k=1\}$ is the probability simplex. $\theta_{jk}\in[0,1]$ is the positive response probability for feature $j\in[J]$ in class $k\in[K]$. In our motivating application, the MGEs adapt to the unobserved host

origin (i.e., latent class) which can be characterized by class-specific response probability profiles $\boldsymbol{\theta}_{\cdot k} = (\theta_{1k}, \ldots, \theta_{Jk})^{\mathsf{T}}$, $k \in [K]$. Because the latent class indicators $I_i^{(v)}$'s are assumed to be unobserved, the observed data likelihood for N independent observations is $\prod_{v \in \mathcal{V}_L} \prod_{i=1}^{n_v} \sum_{k=1}^{K} \pi_{vk} \mathbb{P}\left(\boldsymbol{Y}_i^{(v)} \mid I_i^{(v)} = k, \{\boldsymbol{\theta}_{\cdot k}\}\right)$.

Throughout this paper, we assume that we wish to classify individuals into the the same set of K classes so that probabilistic assignment has coherent interpretation based on $\boldsymbol{\theta}_{\cdot k}$ across the samples. However, we do not assume that observations are drawn from a homogeneous population with the same vector of class proportions. Figure 1 provides a schematic of the data generating mechanism given $\boldsymbol{\pi}_v$ for three leaf nodes.

[Figure 1 about here.]

3. Prior Distributions

We first specify a prior distribution for $\{\boldsymbol{\pi}_v : v \in \mathcal{V}_L\}$. Because the number of observations in a leaf node may vary, we propose a tree-structured prior to borrow information across nearby leaf nodes. The prior encourages collapsing certain parts of the tree so that observations within a collapsed leaf group share the same latent class proportions. In particular, we extend Thomas et al. (2019) to deal with rooted binary weighted trees in an LCM setting. The prior specification is completed by priors for the class-specific response probabilities.

Tree-structured prior for latent class proportions $\boldsymbol{\pi}_v$. We specify a spike-and-slab Gaussian diffusion process prior along a rooted binary weighted tree based on a logistic stick-breaking parameterization of the vector of latent class proportions.

We reparameterize $\boldsymbol{\pi}_v$ with a stick-breaking representation: $\pi_{vk} = V_{vk} \prod_{s<k}(1 - V_{vs})$, for $k \in [K]$, where $0 \leqslant V_{vk} \leqslant 1$, for $k \in [K-1]$ and $V_{vK} = 1$. As the stick-breaking analogy suggests, at Step 1, V_{v1} of the unit-length stick is broken resulting in a segment of length $\pi_{v1} = V_{v1}$ and leaving a stick segment of length $1 - V_{v1}$; V_{v2} of what remains is broken again

in Step 2 ($\pi_{v2} = V_{v2}(1 - V_{v1})$ and so on. At the final Step K, we fix $V_{vK} = 1$ to keep all of the remaining stick segment of length $\pi_{vK} = \prod_{s<K}(1 - V_{vk})$.

We further logit-transform $V_{vk}, k \in [K-1]$, to facilitate the specification of a Gaussian diffusion process prior without range constraints. In particular, let $\eta_{vk} = \sigma^{-1}(V_{vk})$, $k \in [K-1]$, $v \in \mathcal{V}_L$, where $\sigma(x) = 1/\{1 + \exp(-x)\}$ is the sigmoid function. The logistic stick-breaking parameterization is completed by

$$\pi_{vk} = \{\sigma(\eta_{vk})\}^{\mathbf{1}\{k<K\}} \prod_{s<k} \sigma(-\eta_{vs}), k \in [K], \tag{3}$$

which affords simple and accurate posterior inference via variational Bayes (see Section 4).

For a leaf node $v \in \mathcal{V}_L$, let

$$\eta_{vk} = \sum_{u \in a(v)} \xi_{uk}, k \in [K-1], \tag{4}$$

Here η_{vk} is defined for leaf nodes only and ξ_{uk} is defined for all the nodes. Suppose v and v' are leaf nodes and siblings in the tree such that $pa(v) = pa(v')$, setting $\xi_{vk} = \xi_{v'k} = 0$ for $k \in [K-1]$, implies $\eta_{vk} = \eta_{v'k}$, for $k \in [K-1]$, hence $\boldsymbol{\pi}_v = \boldsymbol{\pi}_{v'}$. More generally, a sufficient condition for M leaf nodes η_{vk}, $v \in \{v_1, \ldots, v_M\}$ to fuse is to set $\xi_{uk} = 0$ for any u that is an ancestor of any of $\{v_1, \ldots, v_M\}$ but not common ancestors for all v_m. That is, to achieve grouping of observations that share the same vector of latent class proportions, in our model, it is equivalent to parameter fusing. In the following, we specify a prior on the ξ_{uk} that *a priori* encourages sparsity, so that closely related observations are likely grouped to have the same vector of class proportions. The fewer distinct ancestors two nodes have, the more likely the parameters η_{vk} are fused, because the prior would encourage fewer auxiliary variables ξ_{uk} to be set to zero. In particular, we specify

$$\xi_{uk} = s_u \alpha_{uk}, \forall\ u \in \mathcal{V} \tag{5}$$

$$\alpha_{uk} \sim N(0, \tau_{1k\ell_u} h_u), \text{ independently for } k \in [K-1], \forall\ u \in \mathcal{V}, \tag{6}$$

$$s_{u_0} = 1, \text{ and } s_u \sim \mathsf{Bernoulli}(\rho_{l_u}), \text{ independently for } u \in \mathcal{V} \setminus u_0, \tag{7}$$

$$\rho_\ell \sim \mathsf{Beta}(a_\ell, b_\ell), \text{ independently for } l \in [L], \tag{8}$$

where $N(m, s)$ represents a Gaussian density function with mean m and variance s. $\tau_{1k\ell}$ is the unit-length variance and controls the degree of diffusion along the tree which may differ by dimension k and node level l_u where $\ell_u \in [L]$ represents the "level" or "hyperparameter set indicator" for node u. For example, in simulations and data analysis, we will assume that the root node for the diffusion process has a prior unit-length variance distinct from other non-root nodes. For the root node u_0 with $s_{u_0} = 1$, α_{u_0k} determines the starting point of the diffusion of η_{uk}. Finally, smaller values of ρ_{ℓ_u} encourage parameter fusion of sibling nodes.

REMARK 1: Equations (4)-(8) define a Gaussian diffusion process initiated at α_{u_0k}:

$$\eta_{uk} \mid \{\xi_{u'k}, u \in a(u)\}, s_u, \tau_{1k\ell_u}, h_u \sim N\left(\sum_{u' \in a(u)} \xi_{u'k}, s_u \tau_{1k\ell_u} h_u\right), \tag{9}$$

independently for $k \in [K-1]$, for any non-root node $u \neq u_0$; also see the seminal formulation by Felsenstein (1985). To aid the understanding of this Gaussian diffusion prior, it is helpful to consider a special case of $s_u = 1$ and $\ell_u = 1$, $\forall u \in \mathcal{V}$. For two leaf nodes $v, v' \in \mathcal{V}_L$, the prior correlation between η_{vk} and $\eta_{v'k}$ is

$$\mathrm{Corr}(\eta_{vk}, \eta_{v'k}) = \frac{\sum_{u \in a(v) \cap a(v')} h_u}{\{dist_{\mathcal{T},w}(u_0, v) dist_{\mathcal{T},w}(u_0, v')\}^{1/2}}, \tag{10}$$

When v and v' have the same number of ancesters ($|a(v)| = |a(v')|$) and all edges have identical weight $h_u = c, \forall u$, the prior correlation is the fraction of common ancestors. Note that $\boldsymbol{\eta}_v$ fully determines $\boldsymbol{\pi}_v$ in (3) and induces correlations among $\{\boldsymbol{\pi}_v, v \in \mathcal{V}_L\}$.

Priors for class-specific response probabilities. Let $\gamma_{jk} = \log\{\theta_{jk}/(1-\theta_{jk})\}$. We specify

$$\gamma_{jk} \sim N(0, \tau_{2jk}), \text{ independently for feature } j \in [J] \text{ and class } k \in [K]. \tag{11}$$

Joint distribution. The joint distribution of data and unknown quantities is

$$\pi(\mathbf{Y}, \mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho}; \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{a}, \boldsymbol{b}) = \prod_{v\in\mathcal{V}_L}\prod_{i=1}^{n_v}\prod_{k=1}^{K}\left[\{\sigma(\eta_{vk})\}^{\mathbf{1}\{k<K\}}\prod_{s<k}\{1-\sigma(\eta_{vs})\}\prod_{j=1}^{J}\sigma\left(X_{ij}^{(v)}\gamma_{jk}\right)\right]^{Z_{ik}^{(v)}}$$
$$\times\prod_{u\in\mathcal{V}}\prod_{k=1}^{K-1}\left(\frac{1}{\sqrt{2\pi\tau_{1kl_u}h_u}}\exp\left\{-\frac{1}{2\tau_{1kl_u}h_u}\alpha_{uk}^2\right\}\right)\times\prod_{j=1}^{J}\prod_{k=1}^{K}\left(\frac{1}{\sqrt{2\pi\tau_{2jk}}}\exp\left\{-\frac{1}{2\tau_{2jk}}\gamma_{jk}^2\right\}\right)$$
$$\times\prod_{u\in\mathcal{V}}\rho_{l_u}^{s_u}(1-\rho_{l_u})^{1-s_u}\cdot\prod_{l=1}^{L}\frac{1}{\mathsf{Beta}(a_l,b_l)}\rho_l^{a_l-1}(1-\rho_l)^{b_l-1}, \tag{12}$$

where $X_{ij}^{(v)} = 2Y_{ij}^{(v)} - 1 \in \{-1, 1\}$, and sparsity indicators $\boldsymbol{s} = \{s_u : u \in \mathcal{V}\}$, logit-transformed positive response probabilities $\boldsymbol{\gamma} = \{\gamma_{jk}, j \in [J]; k \in [K]\}$, logistic stick-breaking parameters $\boldsymbol{\alpha} = \{\alpha_{uk} : u \in \mathcal{V}, k \in [K-1]\}$ and $\boldsymbol{\varrho} = (\rho_1, \ldots, \rho_L)^\mathsf{T}$, $\boldsymbol{a} = (a_1, \ldots, a_L)^\mathsf{T}$, and $\boldsymbol{b} = (b_1, \ldots, b_L)^\mathsf{T}$. By setting $s_u = 0$ for non-root nodes in Equation (5), a single vector of class proportions $\boldsymbol{\pi} = \boldsymbol{\pi}_{u_0}$ results: it omits the tree $\mathcal{T}$ that encodes sample similarities and simplifies to the classical LCM. Throughout the paper, we use function $\pi(A; B)$ to denote a probability density or mass function of quantities in A with parameters B; when it is unambiguous, we simply use $\pi(A)$. Figure 2 shows a directed acyclic graph (DAG) that represents the model likelihood and prior specifications.

[Figure 2 about here.]

4. Posterior Inference Algorithms

Calculating a posterior distribution often involves intractable high-dimensional integration over the unknowns in the model. Traditional sequential sampling approaches such as Markov chain Monte Carlo (MCMC) remains a widely used tool to generate approximate samples from the posterior distribution based on which inference is drawn. They can be powerful in evaluating multidimensional integrals. However, they do not guarantee closed-form posterior distribution and monotonic improvement in the approximation. Variational inference (VI) excels exactly along these dimensions and has been used in the context of classical latent

class analysis (e.g., Grimmer, 2011). The idea of VI is to formulate the problem of calculating the posterior distribution into an optimization one that is amenable to iterative solvers. The objective function of VI is typically a lower bound of the marginal likelihood with unknowns integrated out. VI requires a user-specified family of distributions that can be expressed in tractable forms while being flexible enough to approximate the true posterior; the approximating distributions and their parameters are referred to as "variational distributions" and "variational parameters", respectively. VI algorithms finds the best variational distribution that optimizes the objective function. VI has been widely applied in Gaussian Carbonetto et al. (2012); Titsias and Lázaro-Gredilla (2011) and binary likelihoods (e.g., Jaakkola and Jordan, 2000; Thomas et al., 2019). Also see Blei et al. (2017) for a detailed review.

We use VI algorithm to conduct posterior inference. A generic VI updating algorithm applied to our model would involve taking expectations of the joint distribution (12) with respect to sigmoid-transform Gaussian random variables which do not have closed-form expression due to non-conjugacy. In this paper, we use a technique introduced by Jaakkola and Jordan (2000) which bounds the sigmoid function from below by a Gaussian kernel with a tuning parameter, hence restores the calculation of closed-form expectations. Specifically, we will use the inequality

$$\sigma(x) \geqslant \sigma(\psi) \exp\{(x-\psi)/2 - g(\psi)(x^2-\psi^2)\} := h(x,\psi), \tag{13}$$

with $g(\psi) = \frac{1}{2\psi}[\sigma(\psi) - \frac{1}{2}]$ where ψ is a tuning parameter.

Based on the logistic stick-breaking parameterization of $\boldsymbol{\pi}_v$ in (3), we bound π_{vk} using the Jaakkola-Jordan technique in (13). For class $k \in [K]$,

$$\pi_{vk} = \{\sigma(\eta_{vk})\}^{\mathbf{1}\{k<K\}} \prod_{s<k} \sigma(-\eta_{vs}) \geqslant \{h(\eta_{vk}; \phi_k^{(v)})\}^{\mathbf{1}\{k<K\}} \prod_{s<k} h(-\eta_{vs}; \phi_s^{(v)}).$$

When constructing lower bounds for $\{\pi_{vk}\}$, we suggest logisitic stick-breaking parameterization over classical softmax parameterization of $\boldsymbol{\pi}_v$, because currently the latter would need

additional intermediate lower bounds (Titsias, 2016) before applying the Jaakkola-Jordan technique, resulting in generally looser lower bounds and poor estimation $\boldsymbol{\pi}_v$.

We focus on variational distributions that can be factorized as

$$q(\mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho}) = q(\boldsymbol{\gamma}) \underbrace{\prod_{u \in \mathcal{V}} q(s_u, \boldsymbol{\alpha}_u)}_{q(\boldsymbol{s}, \boldsymbol{\alpha})} \cdot \underbrace{\prod_{v \in \mathcal{V}_L} \prod_{i=1}^{n_v} q(\boldsymbol{Z}_i^{(v)})}_{q(\mathbf{Z})} \cdot \underbrace{\prod_{l=1}^{L} q(\rho_l)}_{q(\boldsymbol{\varrho})},$$

where $q(\boldsymbol{Z}_i^{(v)})$ is a multinomial distribution with variational parameters $\boldsymbol{r}_i^{(v)} = \left(r_{i1}^{(v)}, \ldots, r_{iK}^{(v)}\right)$, and $r_{ik}^{(v)}$ represents the approximate posterior probability of observation i in leaf node v belonging to class k and $\sum_{k=1}^{K} r_{ik}^{(v)} = 1$. We bound the marginal likelihood by an evidence lower bound (ELBO) $\mathcal{E}^*(q)$:

$$\begin{aligned} \log \pi(\mathbf{Y}) &\geqslant \int q(\mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho}) \log \left[\frac{\pi(\mathbf{Y}, \mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho})}{q(\mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho})} \right] \mathrm{d}\mathbf{Z}\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\alpha} \\ &\geqslant \int q(\mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho}) \log \left[\frac{h^*(\boldsymbol{\psi}, \boldsymbol{\gamma}, \mathbf{Z}) h^{**}(\boldsymbol{\phi}, \boldsymbol{s}, \boldsymbol{\alpha}, \mathbf{Z}) \pi(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho})}{q(\mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho})} \right] \mathrm{d}\mathbf{Z}\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\alpha} := \mathcal{E}^*(q), \quad (14) \end{aligned}$$

where

$$h^*(\boldsymbol{\psi}, \boldsymbol{\gamma}, \mathbf{Z}) = \prod_{v \in \mathcal{V}_L} \prod_{i=1}^{n_v} \prod_{k=1}^{K} \left\{ \prod_{j=1}^{J} h\left(X_{ij}^{(v)} \gamma_{jk}, \psi_{jk} \right) \right\}^{Z_{ik}^{(v)}},$$

$$h^{**}(\boldsymbol{\phi}, \boldsymbol{s}, \boldsymbol{\alpha}, \mathbf{Z}) = \prod_{v \in \mathcal{V}_L} \prod_{i=1}^{n_v} \prod_{k=1}^{K} \left\{ \{h(\eta_{vk}; \phi_k^{(v)})\}^{\mathbf{1}\{k<K\}} \prod_{m<k} h(-\eta_{vm}; \phi_m^{(v)}) \right\}^{Z_{ik}^{(v)}}.$$

Our algorithm iterates to find the optimal variational distribution q that maximizes the $\mathcal{E}^*(q)$. In addition, we adopt an approximate Empirical Bayes approach by optimizing the VI objective function over the hyperparameters $\{\tau_{1k\ell}, \tau_{2jk}, \ell \in [L], j \in [J], k \in [K]\}$. Relative to specifying weakly informative but often non-conjugate hyperprior for the variance parameters, optimizing hyperparameter is more practically convenient. Because updating the hyperparameters means changing the prior, we update the hyperparameters every d complete VI iterations. Pseudocode in Algorithm 1 outlines the steps of the VI procedure; see the exact updating formula in Appendix A1.

Algorithm 1: Pseudocode for Variational Algorithm to Integrate Sample Similarities into Latent Class Models

Data:

(a) Multivariate binary data $\mathbf{Y}$
(b) A binary weighted rooted tree $(\mathcal{T}, w)$ with leaf nodes $\mathcal{V}_L \subset \mathcal{V}$, edge lengths $\boldsymbol{h} = (h_u)_{u\in\mathcal{V}}$,
(c) The leaf id for each observation (n_v observations in leaf node $v \in \mathcal{V}_L$)

Fixed Hyperparameters:

(a') The number of classes $K \geqslant 2$; levels $\ell_u \in [L]$ for all nodes $u \in \mathcal{V}$
(b') Hyperparameters for the prior probability of $s_u = 1$: (a_l, b_l), $l \in [L]$

Initialize:

(a'') $t \longleftarrow 0$; Initialize $q_t(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ `// (see Step 0 in Appendix A1)`
(b'') Set an initial ELBO $\mathcal{E}_0^* \longleftarrow 0$

$t \longleftarrow 1$; $\mathcal{E}_1^* \longleftarrow \mathcal{E}_0^* + 2\epsilon$
while $|\mathcal{E}_t^* - \mathcal{E}_{t-1}^*| > \epsilon$ **do**
  $q_t(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \longleftarrow q_{t-1}(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$
  $\boldsymbol{\phi}^{(t)} \longleftarrow \boldsymbol{\phi}^{(t-1)}$; $\boldsymbol{\psi}^{(t)} \longleftarrow \boldsymbol{\psi}^{(t-1)}$
  $\boldsymbol{\tau}_1^{(t)} \longleftarrow \boldsymbol{\tau}_1^{(t-1)}$; $\boldsymbol{\tau}_2^{(t)} \longleftarrow \boldsymbol{\tau}_2^{(t-1)}$
  for $v \in \mathcal{V}_L$ **do**
    for $i \in [n_v]$ **do**
      for $k \in [K]$ **do**
        $r_{ik}^{(v),t} \longleftarrow \operatorname{argmax}_{r_{ik}^{(v)}} \mathcal{E}_t^*(q)$ `// (See Step 1a in Appendix A1)`
  $q_t(\boldsymbol{\gamma}) \longleftarrow \operatorname{argmax}_{q_t(\boldsymbol{\gamma})} \mathcal{E}_t^*(q)$ `// (see Step 1b in Appendix A1)`
  for $u \in \mathcal{V}$ **do**
    $q_t(s_u, \boldsymbol{\alpha}_u) \longleftarrow \operatorname{argmax}_{q_t(s_u,\boldsymbol{\alpha}_u)} \mathcal{E}_t^*(q)$ `// (see Step 1b in Appendix A1)`
  for $l \in [L]$ **do**
    $q_t(\rho_l) \longleftarrow \operatorname{argmax}_{q_t(\rho_l)} \mathcal{E}_t^*(q)$ `// (see Step 1c in Appendix A1)`
  for $k \in [K]$ **do**
    `// update local variational parameters for tigher lower bounds`
    for $v \in \mathcal{V}_L$ **do**
      $\phi_k^{(v),(t)} \longleftarrow \operatorname{argmax}_{\phi_k^{(v)}} \mathcal{E}_t^*(q)$
    for $j \in [J]$ **do**
      $\psi_{jk}^{(t)} \longleftarrow \operatorname{argmax}_{\psi_{jk}} \mathcal{E}_t^*(q)$ `// (see Step 2 in Appendix A1)`
  if $t \ mod \ d = 0$ **then**
    for $k \in [K]$ **do**
      for $l \in [L]$ **do**
        $\tau_{1kl}^{(t)} \longleftarrow \operatorname{argmax}_{\tau_{1kl}} \mathcal{E}_t^*(q)$
      for $j \in [J]$ **do**
        $\tau_{2jk}^{(t)} \longleftarrow \operatorname{argmax}_{\tau_{2jk}} \mathcal{E}_t^*(q)$ `// (see Step 3 in Appendix A1)`
  $\mathcal{E}_t^* \longleftarrow ELBO(q_t)$ `// (see Step 4 in Appendix A1)`
  $t \longleftarrow t + 1$

Return: $q_{t-1}(\boldsymbol{\gamma})$, $q_{t-1}(\boldsymbol{s}, \boldsymbol{\alpha})$, $\{q_{t-1}(\boldsymbol{Z}_i^{(v)})\}$, $q_{t-1}(\boldsymbol{\varrho})$, $\{\mathcal{E}_1^*, \ldots, \mathcal{E}_{t-1}^*\}$

4.1 *Posterior Summaries*

Two sets of point and interval estimates for $\{\boldsymbol{\pi}_v : v \in \mathcal{V}_L\}$ result from our VI algorithm: 1) data-driven grouped estimates ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$), and 2) leaf-specific estimates ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$). For 1), we select the median posterior model by choosing the nodes with approximate posterior inclusion probabilities greater than 0.5: $\mathcal{U} = \{u : E_{q_t}[s_u] > 0.5\}$ (see Step 1b, Appendix A1). The approximate posterior mean and 95% credible intervals (CrIs) can be calculated as follows: draw random samples of α_{uk} from a Gaussian distribution with mean $E_{q_t}[\alpha_{uk} \mid s_u = 1]$ and variance $V_{q_t}[\alpha_{uk} \mid s_u = 1]$, for $k \in [K-1]$; compute the corresponding latent class proportions based on (3) with $s_u = \mathbf{1}\{u \in \mathcal{U}\}$; based on the derived random samples on the scale of $\boldsymbol{\pi}_v$, compute the means and 95% CrIs marginally for π_{vk}, $k \in [K]$. As a comparison, for 2), we define leaf-specific estimates $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$ by the mean of (3) where η_{uk} follows a Gaussian distribution with mean $\sum_{u \in a(v)} E_{q_t}[s_u \alpha_{uk}]$ and variance $\sum_{u \in a(v)} V_{q_t}[s_u \alpha_{uk}]$, for $k \in [K]$; We also used Monte Carlo simulation to calculate the posterior means and 95% CrIs. In general, leaf-specific estimates of the class proportions differ across the leaf nodes. In contrast, the data-driven grouped estimates $\{\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}\}$ induces dimension reduction. Similarly, we obtain the posterior means and quantiles for the class-specific response probabilities: $\theta_{jk} = \sigma(\gamma_{jk})$ by simulating γ_{jk} from Gaussian distributions with mean $E_{q_t}[\gamma_{jk}]$ and variance $V_{q_t}[\gamma_{jk}]$.

5. Simulation

5.1 *Design and Performance Metrics*

We conduct a simulation study to evaluate the performance of the proposed tree-integrative LCM. We compare our model to a few alternatives with ad hoc grouping of observations in terms of accuracy in estimating the latent class proportions $\{\boldsymbol{\pi}_v, v \in V_L\}$ and the ability to correctly estimate leaf groups. Data were generated under two scenarios with a small and large number of leaf nodes. Appendix A2 details the true parameter settings of the

simulations. Figure 3(a) and 3(b) visualize the trees used in the simulation with 11 and 133 leaves, respectively.

We simulated 100 datasets for different sample sizes ($N = 500, 2000$), where N is the number of observations each of which belongs to a leaf node in the rooted tree. For each N, we set $n_v \approx N/p_L$ for $\nu \in \mathcal{V}_L$ (with rounding where needed) to investigate balanced leaf nodes and set n_v to be approximately $\frac{1}{5}N/p_L$ or $\frac{4}{5}N/p_L$ with equal chance for mimicking unbalanced observations across leaf nodes. For observations in a leaf node v, we simulate the observations $\boldsymbol{Y}_i$ according to an LCM with class proportions $\boldsymbol{\pi}_v$ and class-specific response probabilities $\theta_{jk}, j \in [J]$, $k \in [K]$, where J is the number of binary measures for each observation and K is the number of classes. We simulated data for different dimensions $J = 20, 80$ and different class numbers $K = 3, 6$.

For each simulated dataset, we fit the proposed model using the VI algorithm (Section 4). Based on the proposed model, we compute the data-driven grouped estimates ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$) and leaf-specific estimates $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$ that do not perform median posterior spike-and-slab model selection as defined in Section 4.1. Our primary interest is in $\{\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}\}$; $\{\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}\}$ are for comparisons. In addition, we also tested a few approaches based on ad hoc leaf node groupings: 1) True grouping analysis (fit separate LCMs to obtain estimates in each of the true groups); 2) Single group LCM analysis (omit the tree information); 3) Ad hoc grouping 1 (manual grouping coarser than the true grouping); 4) Ad hoc grouping 2: analysis based on classical LCM for each leaf node. All analyses assume $\{\theta_{jk}\}$ that does not vary by leaf nodes.

We used three model performance metrics. First, we computed the root mean squared errors (RMSE) for an estimate $\widehat{\boldsymbol{\pi}}_v$ where $\mathsf{RMSE}(\widehat{\boldsymbol{\pi}}_\nu) = \sqrt{(Kp_L)^{-1}\sum_{k=1}^{K}\sum_{v\in V_L}\{\widehat{\pi}_{vk} - \pi_{vk}\}^2}$. Second, we compared the true grouping of the latent class proporions used to generate the data with the grouping estimated by the proposed model given the data via adjusted Rand

Index (ARI, Hubert and Arabie, 1985). ARI is a chance-corrected index that takes value between -1 and 1 with values closer to 1 indicating better agreement.

5.2 *Simulation results*

Figure 3 shows a representative comparison among the RMSEs for different models and over increasing sample sizes for the small and large trees under $J = 20$ and uneven sample sizes over the leaf nodes. For sample sizes $N = 500$ and $N = 2,000$, the proposed methods with data-driven grouping ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$) produces similar or better RMSE than analyses based on ad hoc leaf groupings. The proposed approach achieves similar RMSE as $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$ without the posterior median node selection, indicating little accuracy is lost in exchange for dimension reduction when using data-driven grouped estimates. The RMSE for $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ is smaller compared with a refined leaf-node-level grouping because of the smaller sample sizes in the leaf nodes to accurately estimate their $\boldsymbol{\pi}_v$. Under a larger J or K or even sample sizes in the leaf nodes, we observed similar relative advantage in terms of RMSE for $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ of the proposed method.

On the other hand, based on a single LCM or other approaches based on ad hoc groupings of leaf nodes, the individual-specific variational posterior of class probabilities $\{\boldsymbol{r}_i^{(v)}\}$ can be averaged within in each leaf nodes to produce a local estimates of the $\boldsymbol{\pi}_v$. However, the ad hoc post-processing cannot fully address the issue of assessment of posterior uncertainty nor data-driven grouping of leaf nodes. The proposed method overcomes both issues in terms of estimating leaf-specific class proportions in a coherent modeling framework.

We compared the ARI that assesses how well the proposed methods discovered the leaf groups relative to the truth. The accuracy of group discovery increases with sample sizes with other settings fixed. Although the groups discovered are not perfect, the RMSE are comparable to the estimates based on true leaf node groupings.

[Figure 3 about here.]

6. *E. Coli* Data Application

6.1 *Background and Data*

E. coli infections cause millions of urinary tract infections (UTIs) in the US each year (e.g., Johnson and Russo, 2002). Many studies have shown that extraintestinal pathogenic *E. coli* (ExPEC) strains routinely colonize food animals and contaminate the food supply chain serving as a likely link between food-animal *E. coli* and human UTIs (e.g., Johnson et al., 2005). The scientific team adopted a novel strategy of augmenting fine-scale core-genome phylogenetics with interrogation of accessory host-adaptive MGEs (see Section 1.1). The scientific goal is to accurately estimate the population proportions of *E. coli* isolates with human and non-human host-origins across genetically diverse but related *E.coli* sequence types (STs).

We restrict our analysis to $N = 3,126$ *E.coli* isolates in a well-defined collection from humans and retail meat obtained over a 12-month period in Flagstaff, Arizona, US. Each isolate belongs to one of $p_L = 133$ different STs (leaf nodes in the phylogenetic tree) that are identified via a multilocus sequence typing scheme based on short-read DNA sequencing. A total of $J = 17$ MGEs were curated and associated with functional annotations. Each ST was represented by at least four isolates ($n_v \geqslant 4$, $\forall v \in \mathcal{V}_L$). We constructed rooted, maximum-likelihood phylogenies using core-genome SNP data for the 133 STs. Figure 4 shows the estimated phylogenetic tree for the STs, each of which is overlaid in the same row with the empirical frequencies of the MGEs and of the observed sources (human clinical vs foodborne samples). The MGEs vary in their observed frequencies across lineages, indicating potential between-lineage differences in the relative proportions of human and non-human host-origins of the *E.coli* samples. We apply the proposed tree-integrative LCM to 1) estimate the population proportions of *unobserved* human and non-human host-origins for all *E.coli* STs with data-driven groupings of the STs for dimension reduction; and 2) to produce isolate-

level probabilistic host-origin assignment. A subset of preliminary data is analyzed in this paper for illustrating the proposed method. Inclusion of additional samples and/or MGEs may change findings. The final results and the detailed workflow of MGE discovery will be reported elsewhere.

[Figure 4 about here.]

6.2 *Data Results*

We apply the proposed approach to 1) estimating the host-origin probabilities for each leaf node and 2) probabilistic assignment of host-origin for each individual sample. We assume the origin for each isolate is in an unobserved class of human vs food animals ($K = 2$). We first estimate the class proportions for each leaf node based on the data-driven grouped estimates $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ which are shown in Figure 5(a). The estimated class-specific response profiles $(\widehat{\theta}_{jk}, j = 1, \ldots, K)$ exhibit host-specific enrichment of MGEs (Figure 5(b)). For example, MGE 3, 11-17 are estimated with probability of between 0.15 and 0.71 being present in class 1, with log odds ratios (class 1 vs class 2: $OR(\widehat{\theta}_{j1}, \widehat{\theta}_{j2})$) greater than one. The functional annotations of these MGEs reveal that class 1 is likely associated with food-animal hosts. In contrast, MGE 4-10 are estimated to be present in class 2 with probability between 0.35 to 0.82 with log odds ratios greater than one relative to the corresponding estimated response probabilities in class 1. The results suggest the MGEs are highly associated with different types of host-origins.

[Figure 5 about here.]

The analysis discovered 26 groups of STs (leaf nodes), for each of which the tree-integrative LCM estimates distinct vectors of the latent class proportions ($\boldsymbol{\pi}_v^{\mathsf{dgrp}}$). For ST Group 15 and 16, the class proportions are about evenly split between the two classes. For many other ST groups, the class proportions are almost entirely dominated by human or non-human

host-origin. For example, ST Groups 1 to 4 showed high probabilities of non-human host-origin of *E.coli* based on their high class 1 proportion estimates (the estimated class 1 has higher response probabilities for non-human-host-specific MGEs). The results suggest multiple nearby lineages underwent rare recent cross-species transmissions.

We also compare the results based on two ad hoc groupings of the leaf nodes, (a) single group latent class analysis (omitting the tree information); (b) latent class analysis based on an ad hoc grouping of leaf nodes into three groups (three clades shown in different colors in Appendix Figure S1a) as selected by the scientific team. Appendix Figure S1 shows the results based on a three-clade LCM. The response probability profiles are estimated to be similar to the ones obtained from the proposed tree-integrative LCM approach. The single LCM produced estimated class 1 proportion of 0.624; The LCM based on the ad hoc three-clade grouping produced coarser estimates (estimated class 1 proportion: 0.980 (95% CrI : $0.976, 0.984$)) relative to the results obtained from the proposed approach. In particular, data-driven estimates identified two local leaf nodes (ST1141 and ST10) with a total sample size of 102 that have estimated class 1 proportion (0.828 $(0.750, 0.890)$), highlighting the inability of misspecified ad hoc leaf groups to uncover local differences in the latent class proportions. To compare different models, we performed 10-fold cross-validation to compute the mean predictive log-likelihood as a criterion. Of note, because of small sample sizes in some leaf nodes, a naive cross-validation may by chance result in a training set without any observation in some leaf nodes. We therefore randomly keep one observation per leaf node and perform a 10-fold cross-validation using the remaining observations as testing data. Across models with different K, two-class the tree-integrative LCM is best supported by the data.

On an individual isolate level, the proposed model can assign a quantitative value for the probability that an isolate was derived from a particular host. For example, by incorporating additional observed sample source information, we can visualize the approximate posterior

probability of the true host-origin agreeing with the *observed* sample source category of the same *E.coli* isolate $S_i^{(v)}$ (e.g., 1 for foodborne samples; 2 for human clinical samples (UTI or blood)): $r^{(v)}_{i,S_i^{(v)}}$; we refer to it as "posterior concordance probability (PCP)" for sample i in leaf node v. The histogram of PCPs for all isolates is shown in Figure 5(c). Low estimated PCPs, e.g., below a user-specified threshold of 0.5, indicate likely recent host-jump. In future studies, the selected isolates may subject to further examination to estimate timing of host transmissions based on *in vitro* stability data of each MGE.

7. Discussion

The genetic diversity of COPs that infect people and colonize the more than 9 billion food animals in the US challenge our ability to differentiate zoonotic infections from those endemic to humans. Researchers have augmented core-genome phylogenetics with the interrogation of host-associated MGEs. However, based on the presence or absence of multiple MGEs, methods that further integrate phylogenetic tree information for host-origin proportion estimation remains underdeveloped. In this paper, we proposed a tree-integrative LCM for analyzing multivariate binary data. We formulated the scientific question in terms of inferring latent class proportions that may vary in different parts of a tree. We propose a Gaussian diffusion prior for logistic stick-breaking parameterized latent class proportions and designed a scalable approximate posterior inference algorithm. Our *E. coli* data analysis revealed MGEs are disproportionately associated with specific host origins. Combined with external sample source information, the model can help identify isolates that underwent recent host jump, paving the way for further isolate-level host-origin validation.

Our study had some limitations. First, the MGE data we analyzed may represent a fraction of the host-associated accessory elements. By design, additional accessory elements identified in future studies can be readily integrated and evaluated in the proposed framework. Second, host-associated accessory elements are lost and gained over time as *E. coli* strains transition

across hosts. For infections that were zoonotic in nature, we did not observe how much time had lapsed between the cross-species host jump and the actual infection. Our model partly accounted for these uncertainties by the imperfect positive response probabilities. However, the timings may drive the presence or absence of multiple MGEs, resulting in potential statistical dependence given the true class of host-origin. Deviations from local independence assumption may impact model-based inference (e.g., Pepe and Janes, 2006; Albert and Dodd, 2004).

Further extensions in the tree-integrative LCM framework may improve model applicability. First, when a subset of observations are not mapped in the tree at random, the algorithm can add additional unobserved leaf node indicators to be inferred along with other parameters. Second, the tree integrated into LCM in this paper is estimated from core-genome sequences with statistical uncertainty. An additional layer of the prior in the tree space may be added into the model to fully account for upstream uncertainty in tree estimation. Finally, *E. coli* isolates may vary in additional factors such as the hosts' clinical characteristics. Regression extensions may refine the understanding of variation in latent class proportions and positive response probabilities that are driven by covariates. We leave these topics for future work.

Acknowledgment

The research is supported in part by a Precision Health Investigator Award from University of Michigan, Ann Arbor (ML, ZW); an award from Wellcome Trust (LBP, MA and CML; award number 201866); and National Institutes of Health (NIH) grants R01AR073208(ZW), P30CA04659(ZW), and 1R01AI130066-01A1(LBP).

Data Availability Statement

An `R` package "`lotR`" is freely available at `https://github.com/zhenkewu/lotR`. The data that support the findings in this paper are available from the corresponding author upon reasonable request.

References

Airoldi, E. M. and Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* **111,** 1381–1403.

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60,** 427–435.

Besser, J. M., Carleton, H. A., Trees, E., Stroika, S. G., Hise, K., Wise, M., and Gerner-Smidt, P. (2019). Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. *Foodborne pathogens and disease* **16,** 504–512.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association* **112,** 859–877.

Carbonetto, P., Stephens, M., et al. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis* **7,** 73–108.

CDC (2019). Antibiotic resistance threats in the United States. Atlanta, GA: US Department of Health and Human Services, CDC; 2019.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 1042–1051.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist* **125,** 1–15.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica* pages 339–373.

Ghahramani, Z., Jordan, M. I., and Adams, R. P. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in neural information processing systems*, pages 19–27.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Grimmer, J. (2011). An introduction to bayesian inference via variational approximations. *Political Analysis* **19,** 32–47.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* **2,** 193–218.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10,** 25–37.

Johnson, J. R., Delavari, P., O'Bryan, T. T., Smith, K. E., and Tatini, S. (2005). Contamination of retail foods, particularly turkey, from community markets (minnesota, 1999–2000) with antimicrobial-resistant and extraintestinal pathogenic escherichia coli. *Foodbourne Pathogens & Disease* **2,** 38–49.

Johnson, J. R. and Russo, T. A. (2002). Extraintestinal pathogenic escherichia coli:"the other bad e coli". *Journal of Laboratory and Clinical Medicine* **139,** 155–162.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In Stouffer, S., editor, *The American Soldier: Studies in Social Psychology in World War II*, volume IV, pages 362–412. Princeton University Press, Princeton, NJ.

Lindsay, J. A. and Holden, M. T. (2004). Staphylococcus aureus: superbug, super genome? *Trends in microbiology* **12,** 378–385.

Liu, C. M., Stegger, M., Aziz, M., Johnson, T. J., Waits, K., Nordstrom, L., Gauld, L., Weaver, B., Rolland, D., Statham, S., et al. (2018). Escherichia coli st131-h22 as a foodborne uropathogen. *MBio* **9,**.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. (1998). Multilocus sequence typing: a portable

approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95,** 3140–3145.

Pepe, M. S. and Janes, H. (2006). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8,** 474–484.

Price, L. B., Hungate, B. A., Koch, B. J., Davis, G. S., and Liu, C. M. (2017). Colonizing opportunistic pathogens (cops): the beasts in all of us. *PLoS pathogens* **13,** e1006369.

Scornavacca, C., Delsuc, F., and Galtier, N. (2020). *Phylogenetics in the Genomic Era*. No commercial publisher — Authors open access book.

Sieber, R. N., Skov, R. L., Nielsen, J., Schulz, J., Price, L. B., Aarestrup, F. M., Larsen, A. R., Stegger, M., and Larsen, J. (2018). Drivers and dynamics of methicillin-resistant livestock-associated staphylococcus aureus cc398 in pigs and humans in denmark. *MBio* **9,**.

Thomas, E. G., Trippa, L., Parmigiani, G., and Dominici, F. (2019). Estimating the effects of fine particulate matter on 432 cardiovascular diseases using multi-outcome regression with tree-structured shrinkage. *Journal of the American Statistical Association* pages 1–11.

Titsias, M. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in neural information processing systems* **24,** 2339–2347.

Titsias, M. K. (2016). One-vs-each approximation to softmax for scalable estimation of probabilities. In *Advances in Neural Information Processing Systems*, pages 4161–4169.

Supporting Information

Web Appendices and Figures referenced in Sections 4, 5 and 6.

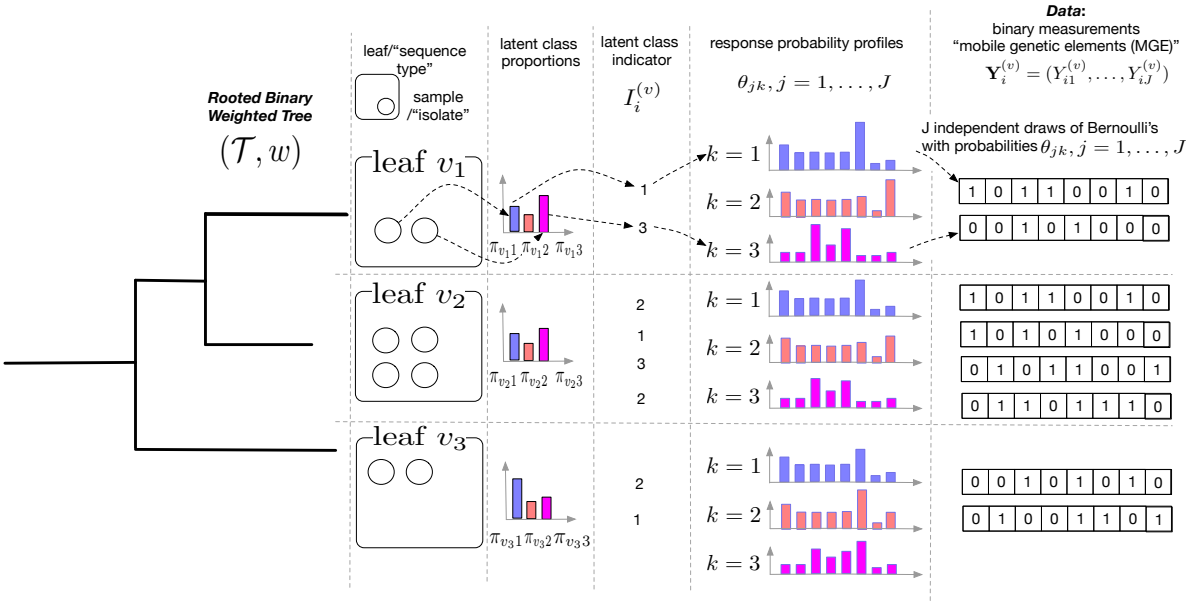


Figure 1: Schematic representation of a hypothetical rooted binary weighted tree with three leaves and data generated based on the proposed model with $K = 3$ latent classes.

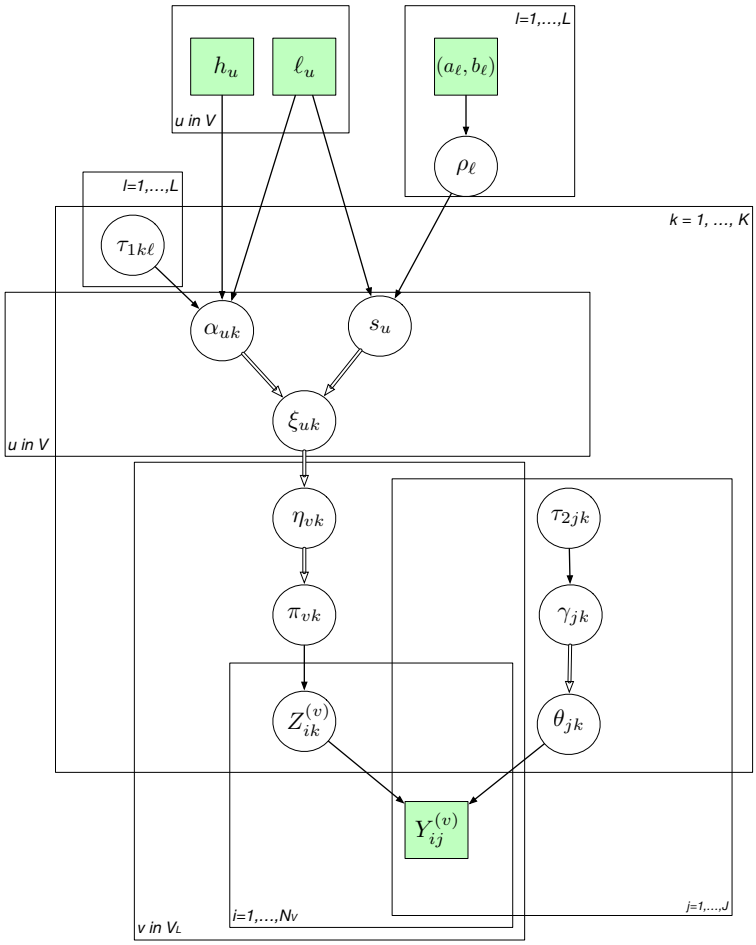
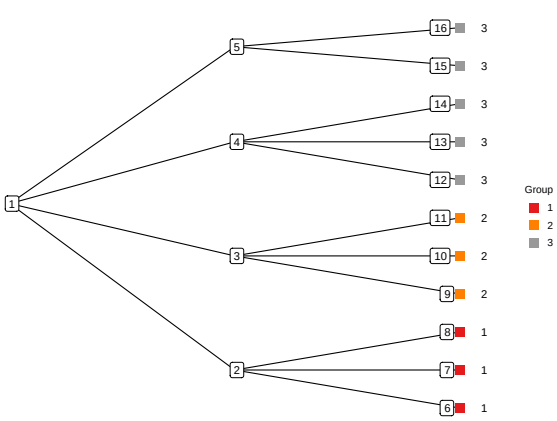
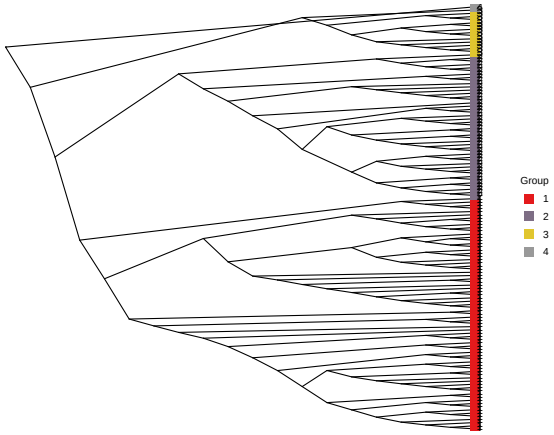


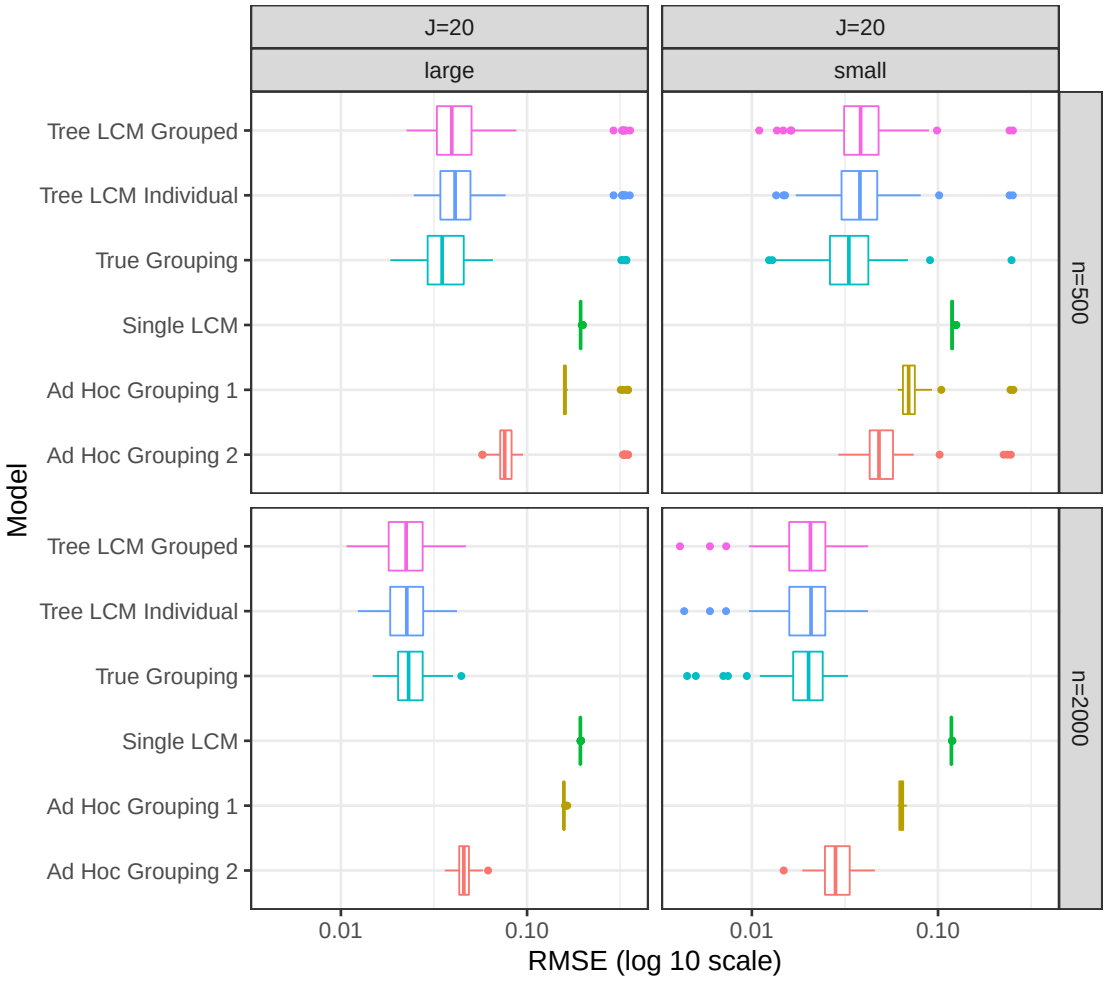
Figure 2: The directed acyclic graph (DAG) representing the structure of the model likelihood and priors. The quantities in squares are either data or hyperparameters; the unknown quantities are shown in the circles. The arrows connecting variables indicate that the parent parameterizes the distribution of the child node (solid lines) or completely determines the value of the child node (double-stroke arrows). The rectangular "plates" where the variables are enclosed indicate that a similar graphical structure is repeated over the index; The index in a plate indicate nodes, hyperparameter levels, leaf nodes, subjects, classes and features.



(a) Scenario 1 with a smaller tree



(b) Scenario 2 with a larger tree



(c) The RMSE comparisons across multiple models and scenarios

Figure 3: Simulation studies show the proposed model produces grouped estimates $\widehat{\boldsymbol{\pi}}_v^{\text{dgrp}}$ with similar or smaller RMSEs compared to alternatives.

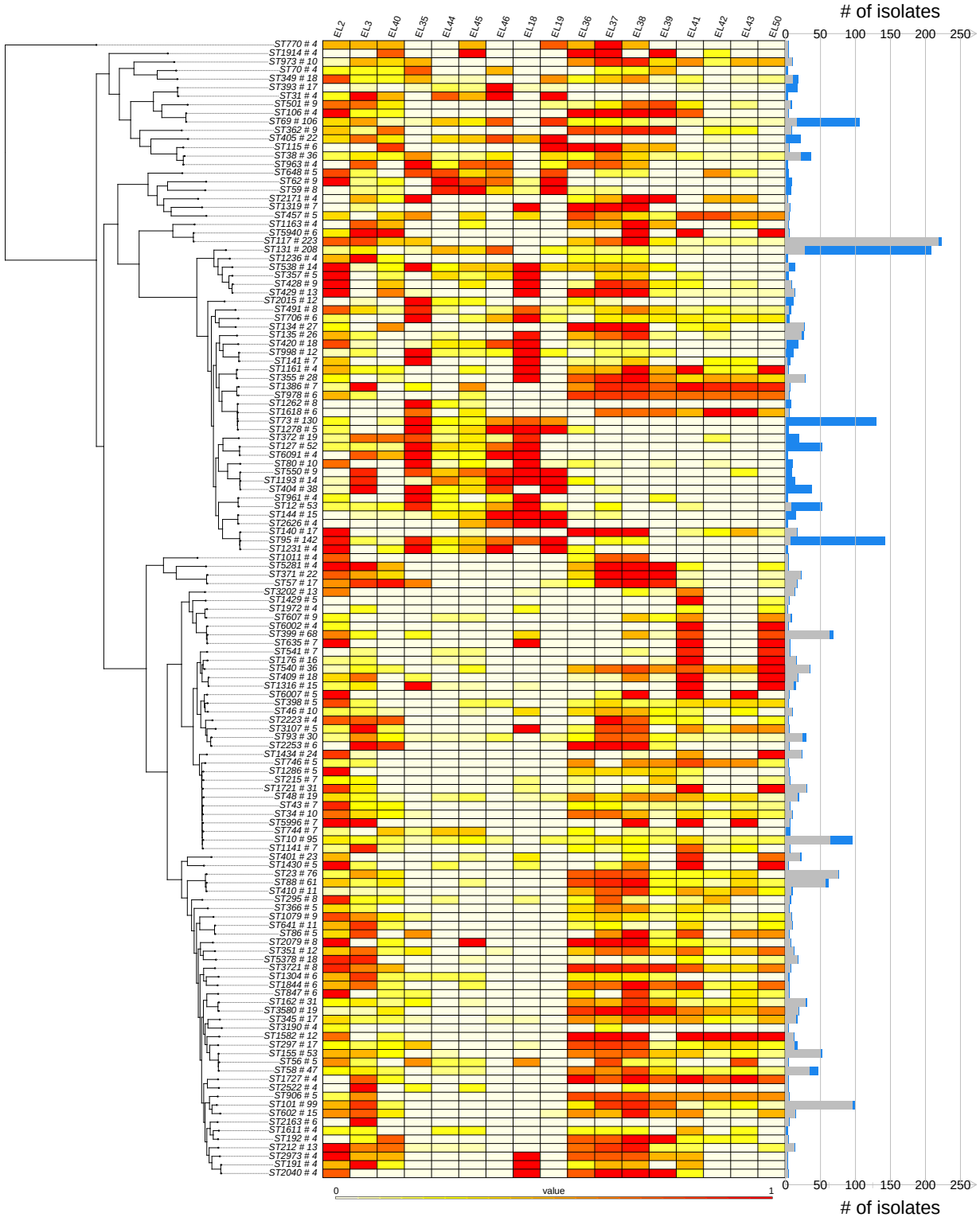
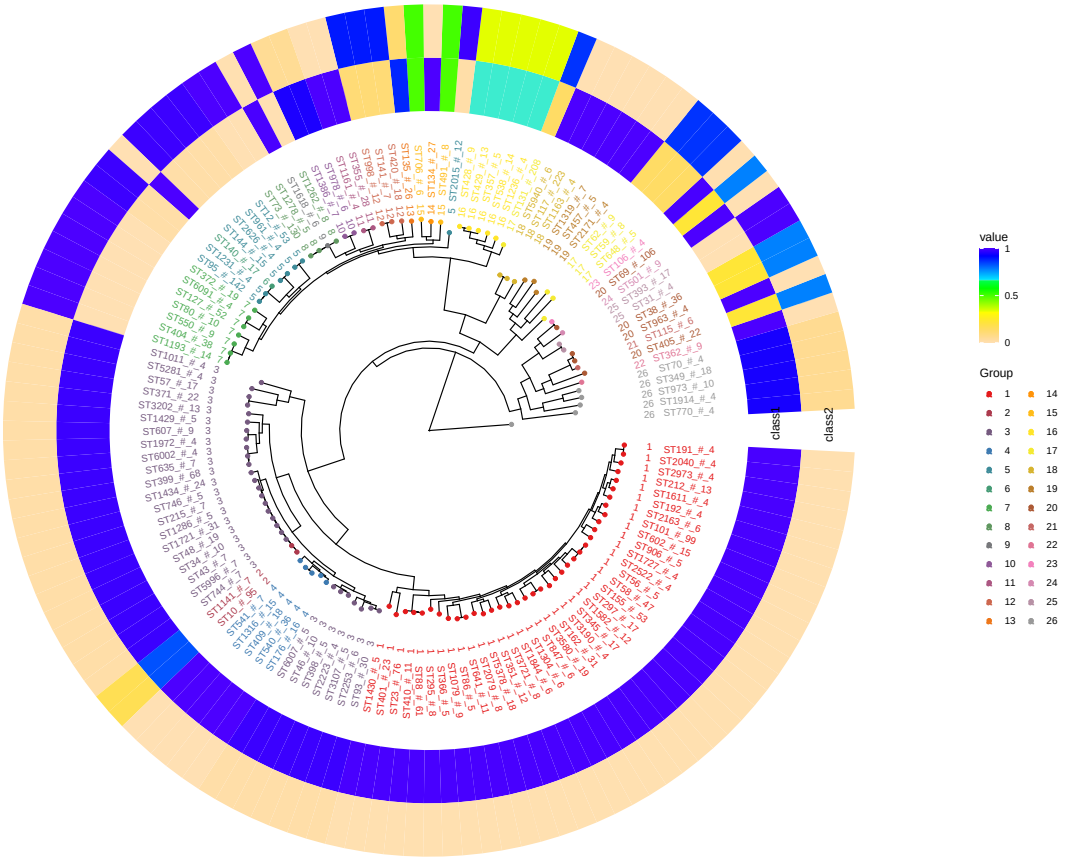
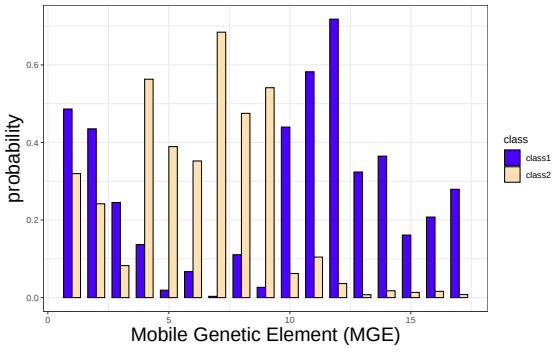


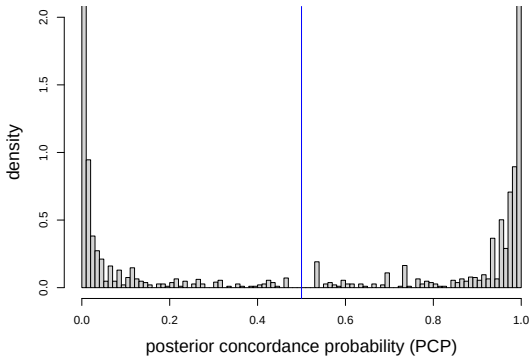
Figure 4: The empirical frequencies for $J = 17$ MGEs within each ST mapped in the core-genome phylogenetic tree. The bars on the right indicate the total number isolates of each ST; the gray and blue bars represent the number of isolates obtained from apparent non-human and human sources, respectively. The core-genome phylogenetic tree on the left margin maps $N = 3,126$ *E. coli* isolates into $p_L = 133$ STs (leaf nodes).



(a) Estimated groups and class proportions



(b) The estimated class-specific response probabilities



(c) Histogram of host-source posterior concordance probability (PCP)

Figure 5: a) Data results with estimated leaf groups and latent class proportions by group. ST names (ST_#_isolates) are aligned to the tips of the circular tree , which are colored by discovered leaf groups. The circular heatmap shows the estimated latent class proportions ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$, $v \in \mathcal{V}_L$); b) and c): see the captions of the subfigures.