**Hierarchical Dirichlet-multinomial Empirical Bayesian model of OTU distribution**

In order to further elucidate XXX we implemented a simple alternative modeling strategy using Hierarchical Dirichlet-Multinomial Bayesian framework. The Multinomial likelihood for modeling OTU counts has been proposed in several studies previously (Holmes *et al.* 2012; Wadsworth *et al.* 2017). The Dirichlet-Multinomial model provides a flexible framework for inclusion of additional parameters that may affect OTU distributions via log-linear regression. First, we explicitly model host-associated OTU distributions depending on host-specific geographic locality (North, contact zone and South), mtDNA and/or DNA types as well as lineages. Additionally, in the proposed model we incorporated such parameters as genetic distance () and climate () into a regression hyperprior. In order to reconstruct hyperprior distributions for and we implemented an empirical Bayes approach. Specifically, we assumed Gaussian hyperpriors for both and with parameters and respectively. Then, the values of these parameters were derived by maximum-likelihood estimators, and , applied to normalized empirical distributions of and for within each geographic region. Below we provide a formal description of our model with its graphical representation (**Fig. 1**).

Observed data (observed OTU’s count):

Where *i* and *j* correspond to *i*-th sample and *j*-th OTU

= vector *i* of observed j OTU counts

= a probability vector that defines expected number of each OTU abundance in *i*-th vector

= expected *j*-th OTU abundance in *r*-th region

Each parameter includes effects of climate (*c*), genetic distance (*dis*) and a random effect () distributed according to a Gaussian distribution with the corresponding hyperparameters.

represent regression coefficients distributed according to Gaussian distributions with high variance.

The hyperparameters and were estimated from the normalized genetic distance and climatic data, respectively.



**Figure 1** A graphical representation of the hierarchical Dirichlet-multinomial empirical Bayesian model. The solid line circles denote stochastic nodes whereas squares denote constant nodes. Dashed elements represent deterministic nodes defined by applying a function to another variable. The graphical model was constructed following convention of (Höhna et al. 2014; Wadsworth et al. 2017)

Höhna S, Heath TA, Boussau B*, et al.* (2014) Probabilistic Graphical Model Representation in Phylogenetics. *Systematic Biology* **63**, 753-771.

Holmes I, Harris K, Quince C (2012) Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One* **7**, e30126.

Wadsworth WD, Argiento R, Guindani M*, et al.* (2017) An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* **18**, 94.