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

In order to further elucidate regional, climatic and genetic effects on OUT distributions in vole populations 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 as well as lineages (i.e. mismatches between mtDNA and nucDNA types). Further partition of OUT data into mtDNA-nucDNA combinations leads to overparamtrization of the model with few samples per mtDNA-nucDNA combination which most likely causes identifiability issues. Thus, this model was not implemented within our framework. 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/lineage. 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 *ij*-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)

To identify patterns and visually summarize the results of Bayesian parameter inference from hierarchical Dirichlet-multinomial empirical model we used state-of-the-art dimensionality reduction techniques such as *t*-distributed stochastic neighbor embedding (*t-*SNE) (van der Maaten & Hinton 2008) and potential of heat diffusion for affinity-based transition embedding (PHATE) (Moon *et al.* 2019). These techniques have been successfully applied in various areas of biology where data exhibit high dimensionality that requires projections on lower dimensions for visualization and identification of possible intrinsic patterns. Principle component analysis (PCA) has been widely used for this purpose however its sensitivity to noisy data and inability to optimize for 2D visualization drastically limit its implementation. Our main objective was twofold: (i) verify whether OTUs show any reginal specific distribution, i.e. whether vole OTU profiles differ between North, Contact zone, South, mtDNA type as well as lineage and (ii) how climate and genetic distance contribute to OTU distributions within each region. For the first part we performed OTU profile draws from posterior predictive distribution which is equivalent to simulations of unobservable OTU samples for a specific region. Using R implementations of *t-*SNE and PHATE under default parameters we were able to show that OTU distributions have clear regional specificity according to mtDNA (**Fig. 2**, first row) and lineage (**Fig. 3**, first row) models. Additionally, we showed that inferred probability parameter vectors () of multinomial likelihood function uniquely define OTU distributions (**Figs 2** and **3**, second row). Finally, we applied *t-*SNE and PHATE to the vectors of regression coefficient parameters for climate () and genetic distance (). The visualizations show that climate has an effect on reginal OTU distributions however it is not clearly separable within contact zone-South mtDNA (CS) and South mtDNA region (S) in mtDNA model (**Fig. 2**, third row). Similar tendencies were observed within the contact zone for three lineages (C1, C2 and C3) in lineage model (**Fig. 3**, third row). For climate, *t-*SNE and PHATE failed to clearly separate CS and CN regions in mtDNA model (**Fig. 2**, forth row), and North (N) as well as contact zone regions (C1, C2 and C3) in lineage model. Together these results indicate that although such parameters as climate and genetic distance may affect OTU distributions, their contribution within certain regions is not evident.

All code is freely available on GitHub (http: …. ).

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.

Moon KR, van Dijk D, Wang Z*, et al.* (2019) Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37**, 1482-1492.

van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605.

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.