Quantifying β-diversity over Phylogenetic Trees and Networks

Donovan Parks and Robert Beiko

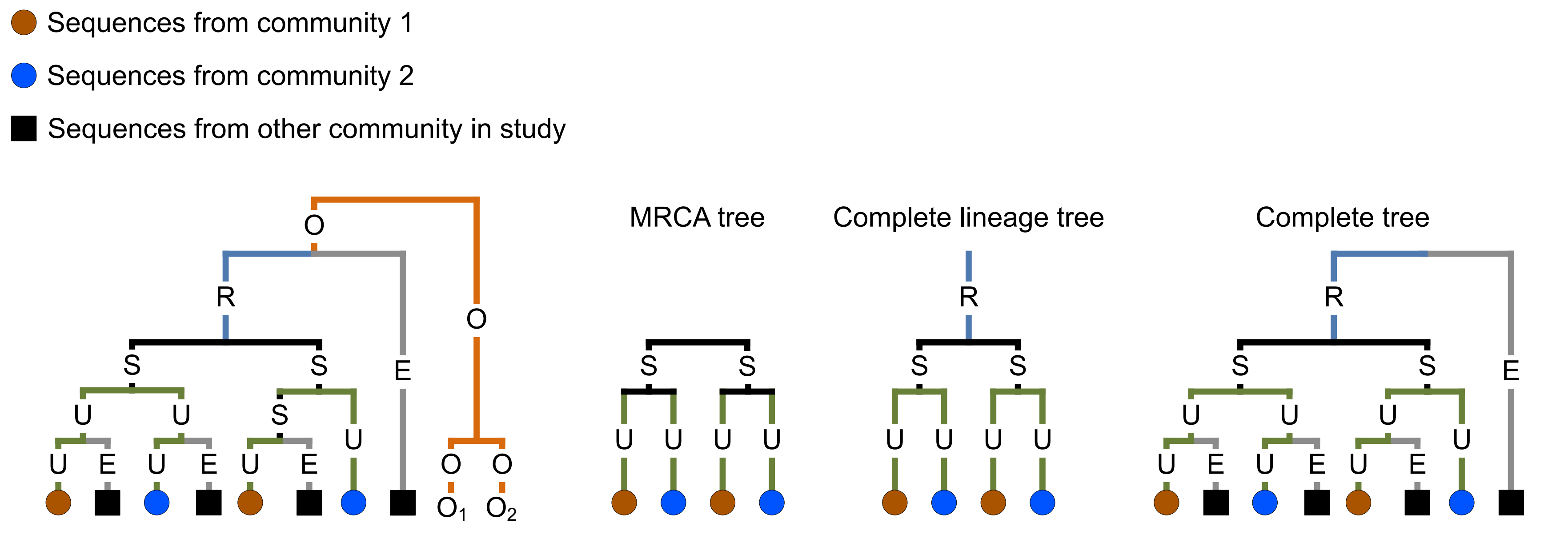
introduction

Phylogenetic β-diversity measures use gene trees to compare the taxonomic or metabolic diversity of communities. Such analyses yield quantitative diversity estimates that can be related to environmental and geographic factors. We have developed three new phylogenetic β-diversity measures. One of these is a modification of the widely used UniFrac statistic (Lozupone and Knight, 2005), while the other two use complementary definitions for determining which branches are shared by or external to a pair of communities.

measuring β-diversity over trees

Branches within a phylogenetic tree can be classified as *unique*, *shared*, *external*, *root*, or *outgroup* with respect to a pair of communities. Using this branch classification scheme we propose three subtrees over which the β-diversity between communities can be measured: the most recent common ancestor (MRCA) subtree, the complete lineage subtree, and the complete tree. Measuring β-diversity over these trees provides complementary information on how sequences are distributed across a phylogeny. Both qualitative and quantitative measures are proposed for each of these subtrees:

- *Qualitative measures*: suggest whether ecological factors prohibit taxa or gene families from occupying certain communities by considering the distribution of distinct sequences.
- *Quantitative measures*: suggest whether ecological differences between communities influence the abundance of taxonomic groups or gene families by considering the relative abundance of each unique sequence.



● Sequences from community 1
● Sequences from community 2
■ Sequences from other community in study
MRCA tree
Complete lineage tree
Complete tree
O
R
S
U
E
O_1
O_2

U = unique, S = shared, E = external, R = root, O = outgroup

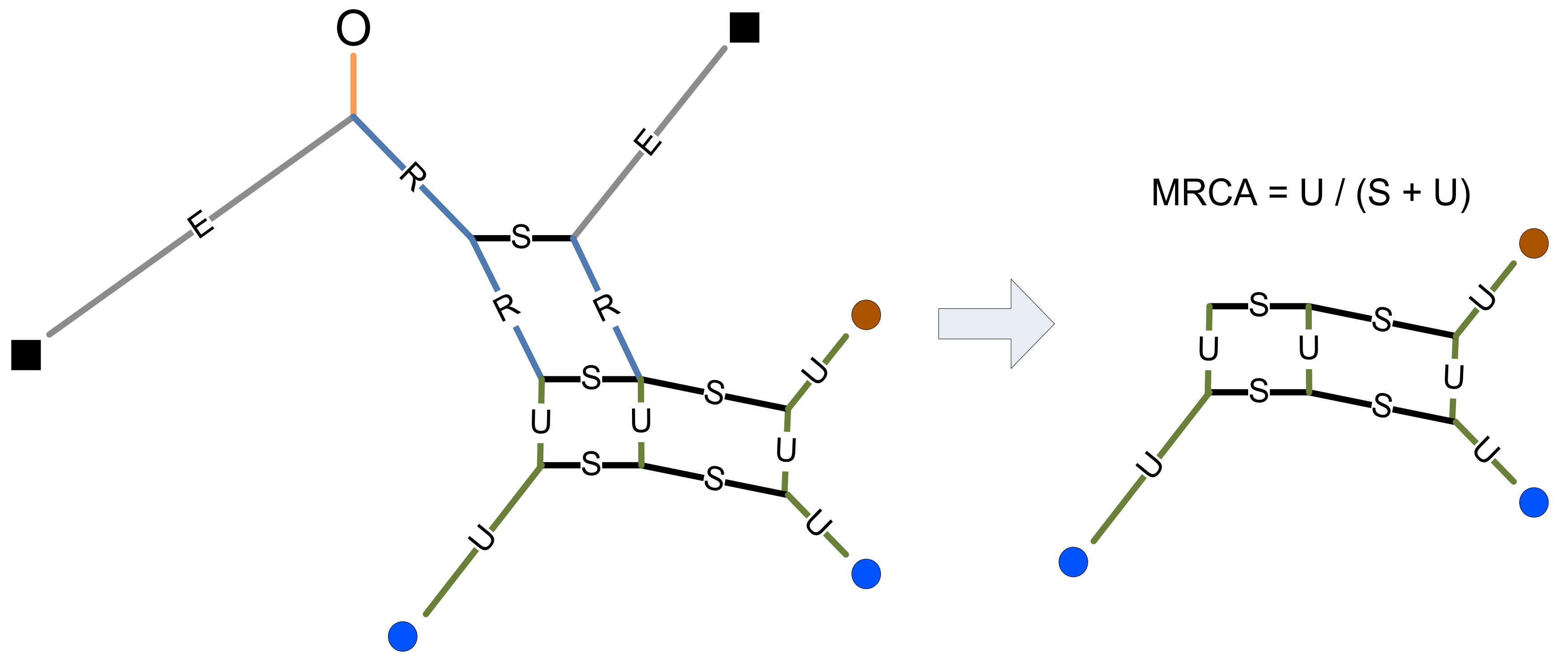
Measure	Qualitative	Quantitative
MRCA	U / (S+U)	$\frac{2\sum_{n=1}^{N}\left\|p_i^n - p_j^n\right\|W_n}{\sum_{n=1}^{N}\min\left(p_i^n + p_j^n, 2 - p_i^n - p_j^n\right)W_n + \sum_{n=1}^{N}\left\|p_i^n - p_j^n\right\|W_n}$
Complete lineage	U / (S+U+R)	$\frac{2\sum_{n=1}^{N}\left\|p_i^n - p_j^n\right\|W_n}{\sum_{n=1}^{N}\left(p_i^n + p_j^n\right)W_n + \sum_{n=1}^{N}\left\|p_i^n - p_j^n\right\|W_n}$
Complete tree	U / (S+U+R+E)	$\frac{\sum_{n=1}^{N}\left\|p_i^n - p_j^n\right\|W_n}{\sum_{n=1}^{N}\left(\max_c\left(p_c^n\right) - \min_c\left(p_c^n\right)\right)W_n}$

p_i^n: proportion of sequences from community i descendant from branch n

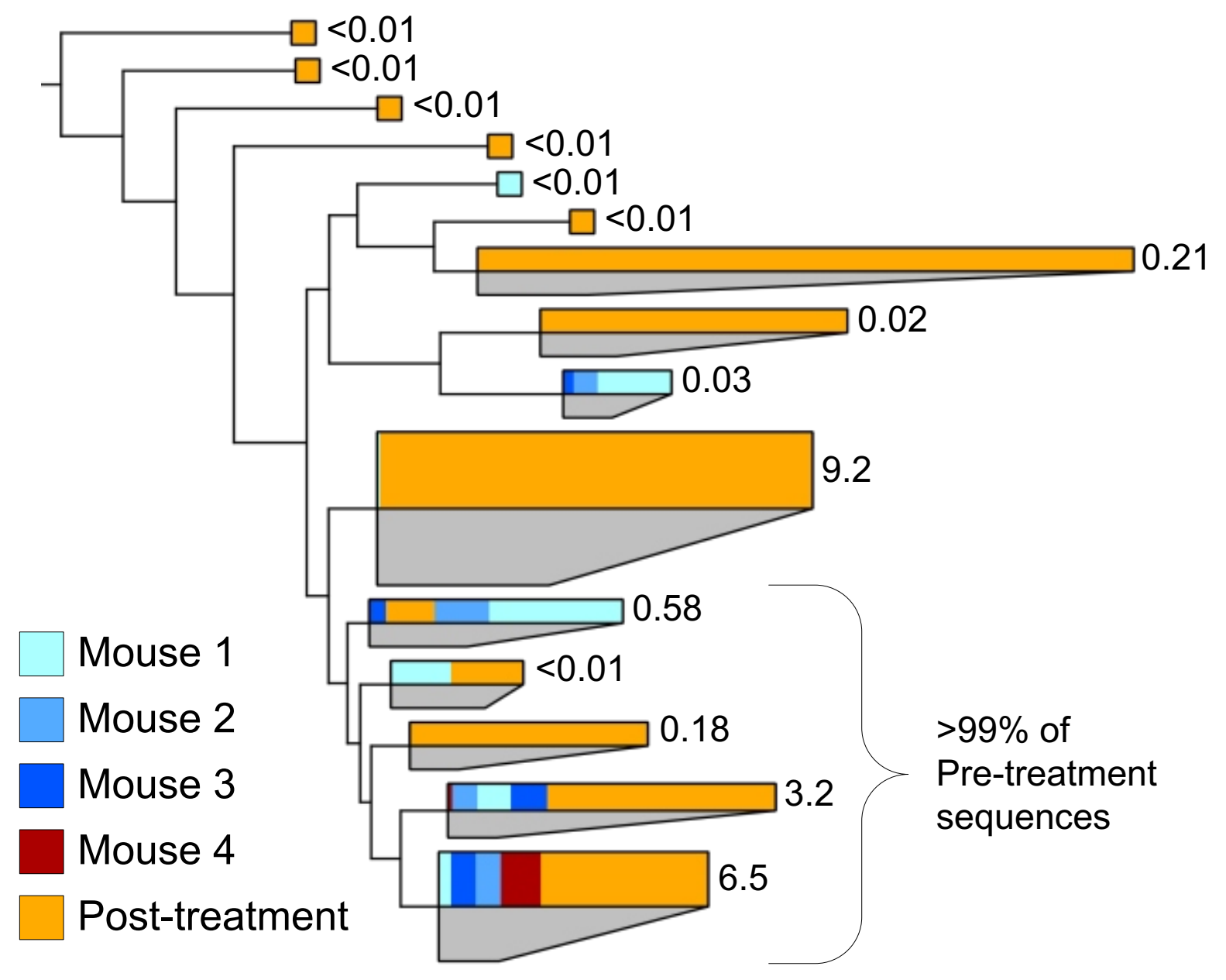
W_n: weight (length) of branch n

measuring β-diversity over networks

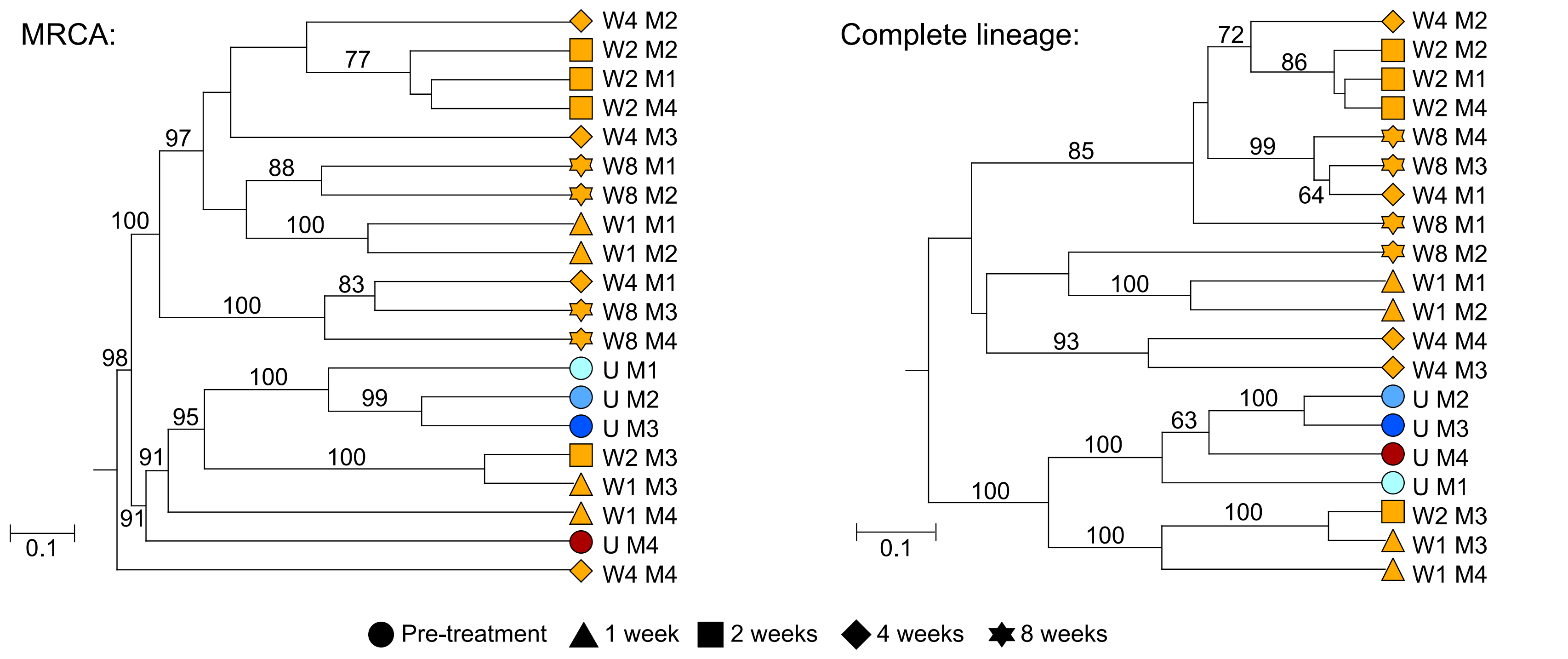
Split networks can be used to represent uncertainty or conflict in the evolutionary history of a set of sequences. Removing parallel edges within a split network induces a bipartition on the set of sequences. This is analogous to removing a branch from a phylogenetic tree. The definitions used to classify branches as *unique*, *shared*, *external*, *root*, or *outgroup* can be extended to the splits of a rooted split network. This permits the proposed qualitative measures to be applied to rooted split networks (example below). All the proposed quantitative measures can be calculated over rooted split networks, while the quantitative MRCA and complete tree measures can also be applied to unrooted split networks. **Measuring β-diversity over a split network provides a measure of community similarity that is averaged over phylogenetic uncertainty or conflict**.



mouse ileum microbiota



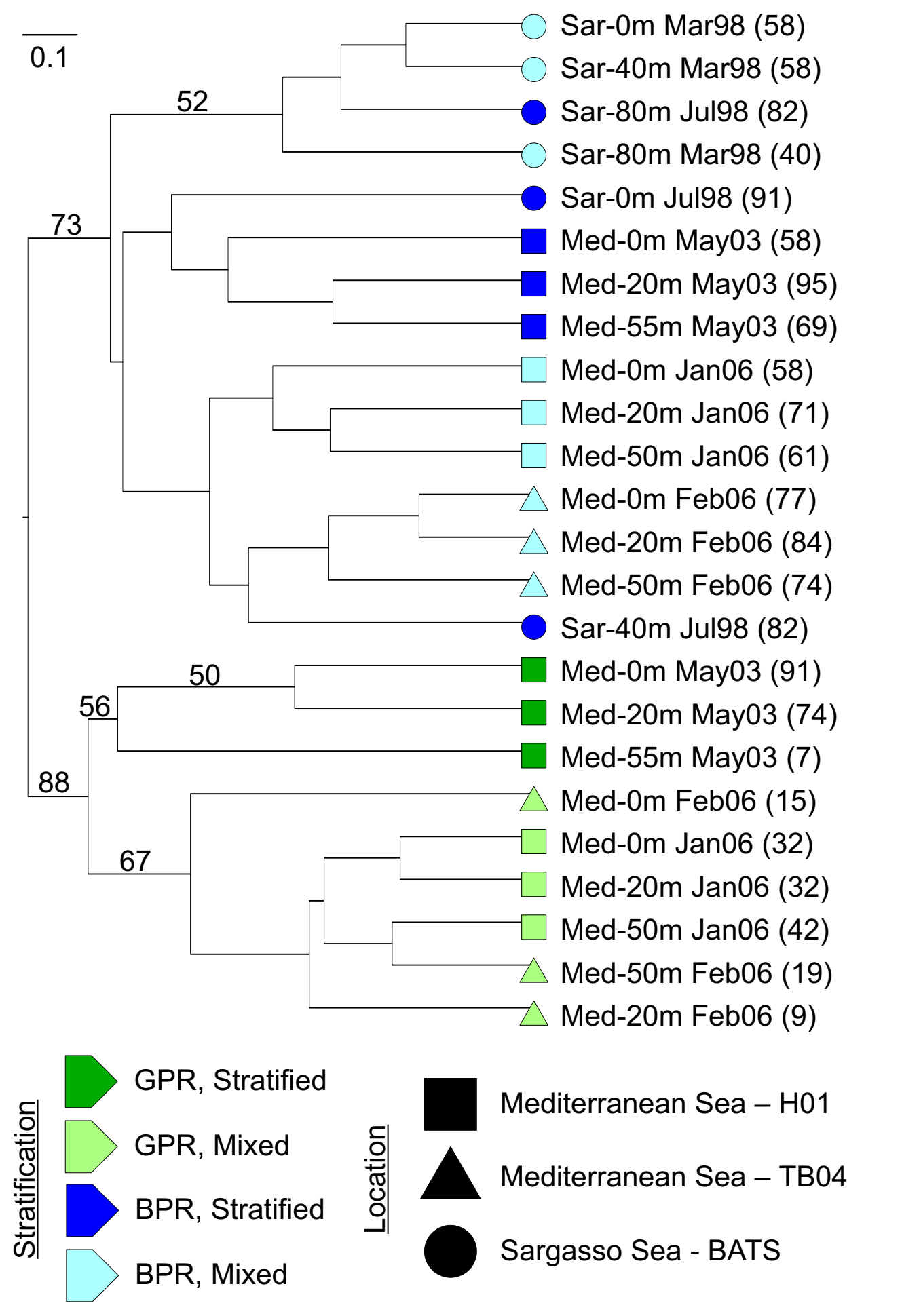
Antibiotic treatment can result in significant changes in intestinal microbiota that may lead to increased susceptibility to infections caused by antibiotic-resistant bacteria. Here we apply the proposed weighted MRCA and complete lineage measures to a set of ileum microbiota sampled by Ubeda *et al.* (2010) from four mice before and after antibiotic treatment. A 16S rDNA phylogeny of the bacterial strains from the 20 ileum samples shows a substantial shift in phylogenetic diversity after treatment (right). Strains from the four pre-treatment samples are mainly confined to lineages comprising Erysipelotrichaceae and Lactobacillaceae species with 96% of sequences from mouse 4's microbiota being restricted to Lactobacillaceae. Applying the quantitative MRCA and complete lineage measures to this phylogeny reveals contrasting patterns of sample similarity (see UPGMA cluster trees below). The complete lineage measure clusters all pre-treatment samples together. This clustering reflects the pre-treatment samples being restricted to Erysipelotrichaceae and Lactobacillaceae species, whereas post-treatment samples primarily consist of Clostridiales Incertae Sedis XIV, Lachnospiraceae, Enterococcaceae, and Streptococcaceae species. In contrast, the MRCA measure is specifically designed to identify communities that are similar across the phylogenetic diversity spanned by a pair of samples and is therefore sensitive to the unique nature of the ileum microbiota within mouse 4.



unrooted proteorhodopsin phylogeny

Proteorhodopsin genes provide aquatic bacteria with a light-driven proton pump suggesting that photosynthesis may play a significant role in the metabolism of aquatic ecosystems. Sabehi *et al.* (2007) extracted proteorhodopsin sequences from 15 environmental samples. Six samples were taken at the BATS station in the Sargasso Sea in March when deep water mixing occurs, and three in July when the water is highly stratified. Samples were taken at depths of 0, 40, and 80 m. Analogous samples were taken from the Mediterranean Sea in January (mixed, H01), February (mixed, TB04) and May (stratified, H01) at depths of 0, 20, and 50-55 m.

Proteorhodopsins are preferentially 'blue-absorbing' (BPR) or 'green-absorbing' (GPR). To investigate the distribution of these spectral genotypes, we applied the quantitative MRCA measure to a maximum likelihood phylogeny. Since there is no established rooting for proteorhodopsin phylogenies, we leave the tree unrooted and take advantage of the ability to apply the quantitative MRCA measure to unrooted trees. The resulting UPGMA clustering reveals clear structuring by genotype and secondary structuring by geography during periods of deep water mixing.



further information

A manuscript is in preparation along with software implementing the proposed measures. For more information, please contact Donovan (parks@cs.dal.ca) or Rob (beiko@cs.dal.ca).

references

Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl *Environ Microbiol* 71: 8228-8235.

Sabehi G, Kirkip BC, Rozenberg M, Stambler N, Polz MF, Béjà O. (2007). Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. *ISME J* 1: 48-55.

Ubeda C, Taur Y, Jenq RR, Equinda MJ, Son T, Samstein M, et al. (2010). Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J Clin Invest* 120: 4332-4341.

GenomeAtlantic

DALHOUSIE UNIVERSITY
Inspiring Minds

Les Fiducies Killam Trusts