**Figure 1. Taxonomic barplots and clustered heatmap of SEED subsys4 expression**

Samples were clustered based on their relative subsys4 expression, after removing rare subsys4 categories with mean read count <1 across all samples (far right heatmap and dendrogram). Complete linkage clustering was then used to draw the dendrogram based on euclidean distances between the zero-replaced, CLR-transformed subsys4 counts (the relative expression). The four major groups are identified on the plot as health-associated group 1 (H1) and group 2 (H2), and BV-associated group 1 (BV1) and group 2 (BV2). In alignment with the dendrogram-ordered samples, the taxonomic distribution by genus of each sample as a fraction of total reads are shown for V6 16S rDNA (leftmost barplot) and for the mapped mRNA reads (middle barplot). Reads belonging to the *Lactobacillus* genus are plotted by species-level assignment where possible. Nugent status and presence of corresponding metabolite samples (used for later analysis) are indicated. 16S sequence data are not available for the samples from the previous dataset, indicated with an “S” designation in the sample ID {Macklaim et al 2013}.