Community Ecology: Final Project

Spring 2014

Yoav Ram

# Part 1: Community data analysis

I chose to analyze data from the Human Microbiome Project (<http://www.hmpdacc.org/>). I downloaded the Mother file of the Phase I (May 1) 16S variable regions 1-3 OTU (operational taxonomic unit) counts (<http://downloads.hmpdacc.org/data/HMMCP/finalData/hmp1.v13.hq.otu.counts.bz2>). This occurrences dataset contains 2,799 samples and 27,655 OTUs. The OTUs are described in the corresponding lookup file (<http://downloads.hmpdacc.org/data/HMMCP/finalData/hmp1.v13.hq.otu.lookup.bz2>). The samples metadata was retrieved (<http://www.hmpdacc.org/doc/ppAll_V13_map.txt>) and merged with the occurrences data. After filtering out unclassified OTUs, samples with zero counts, and OTUs with zero counts, 2,774 samples and 8,289 OTUs remained in the dataset.

The samples represent 173 individuals (85 males and 88 females), who visited between 1 and 3 times and were sampled in five major body sites (see Fig. 1), divided to up to 18 sub sites.

The analysis was performed using *R* with *vegan*. The code (analysis.R), this document (final\_project.docx), and an *IPython notebook* that was used to clean and format the dataset (munge.ipynb) are available online as a *git* repository at <https://github.com/yoavram/CommunityEcologyProject>.

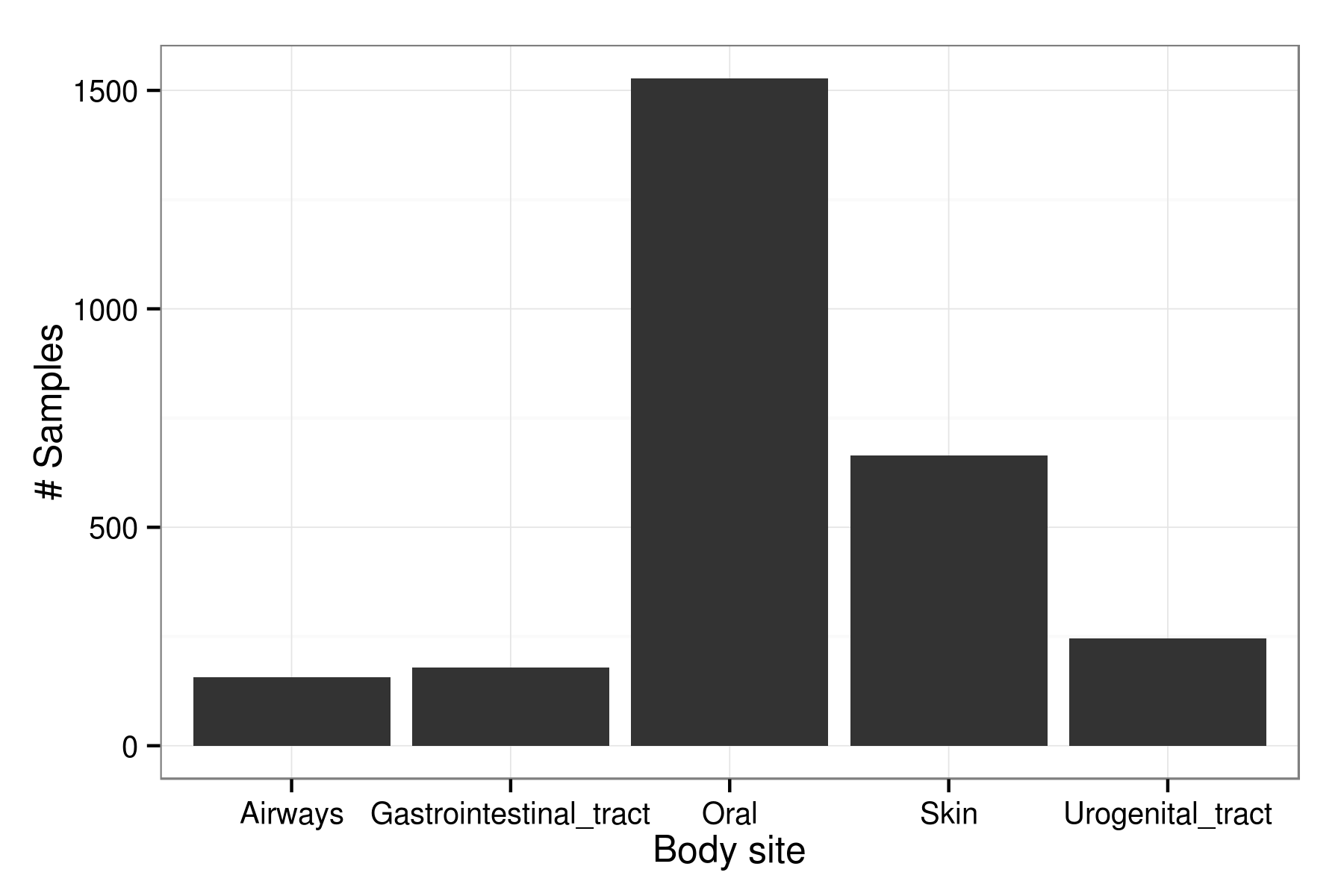


Figure 1 - # of samples per body site, summed over all sampled individuals.

## Species richness

The species richness analysis is summarized in Figs. 2-3. In general, it seems that the GI tract might have the richest community, followed by the oral, skin, airways, and finally, the UG tract. Interestingly, the two richest body sites also have the widest distribution of richness by sample.

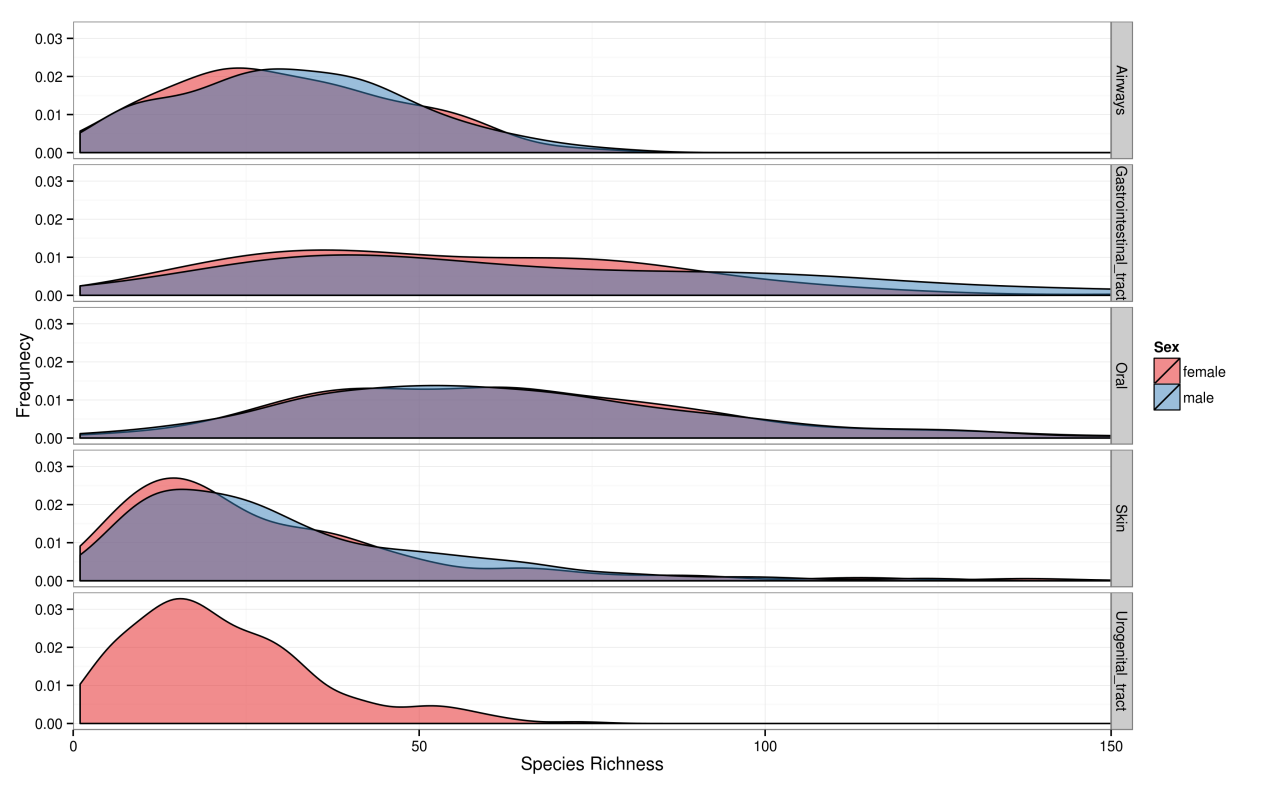


Figure 2 – Species richness distribution by sex (females in red, males in blue) and body sites (from top to bottom: airways, GI tract, oral, skin, UG tract). The richness distribution is similar between the sexes (confirmed by rarefaction which is not shown and consistent with similar number of male and female participants), but very different between body sites, where the GI tract and oral sites have a much wider distribution compared to the airways, skin and UG tract. See Fig. 3 for rarefaction curves.

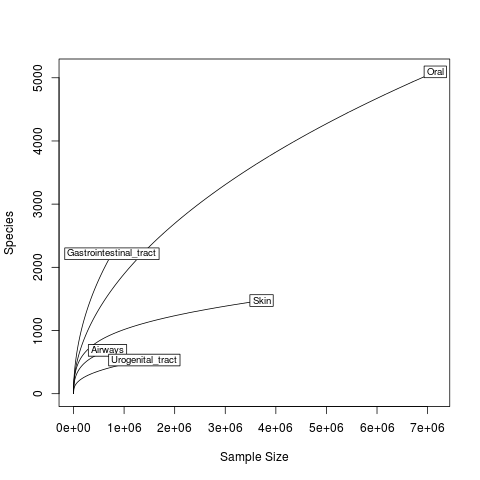


Figure 3 – Rarefaction curves for species richness by body site. This figure adds more information to Fig. 1, suggesting that the GI tract might show the highest richness if the sampling effort was increased.

## Species diversity

Fig. 4 shows the distribution of true diversity (Shannon index) by sex and body site. Similar to species richness, there seems to be little effect of sex on diversity, and a significant effect of body site, with greater diversity for the "richer" body sites (see Fig. 3) and lower diversity for the "poorer" body sites.

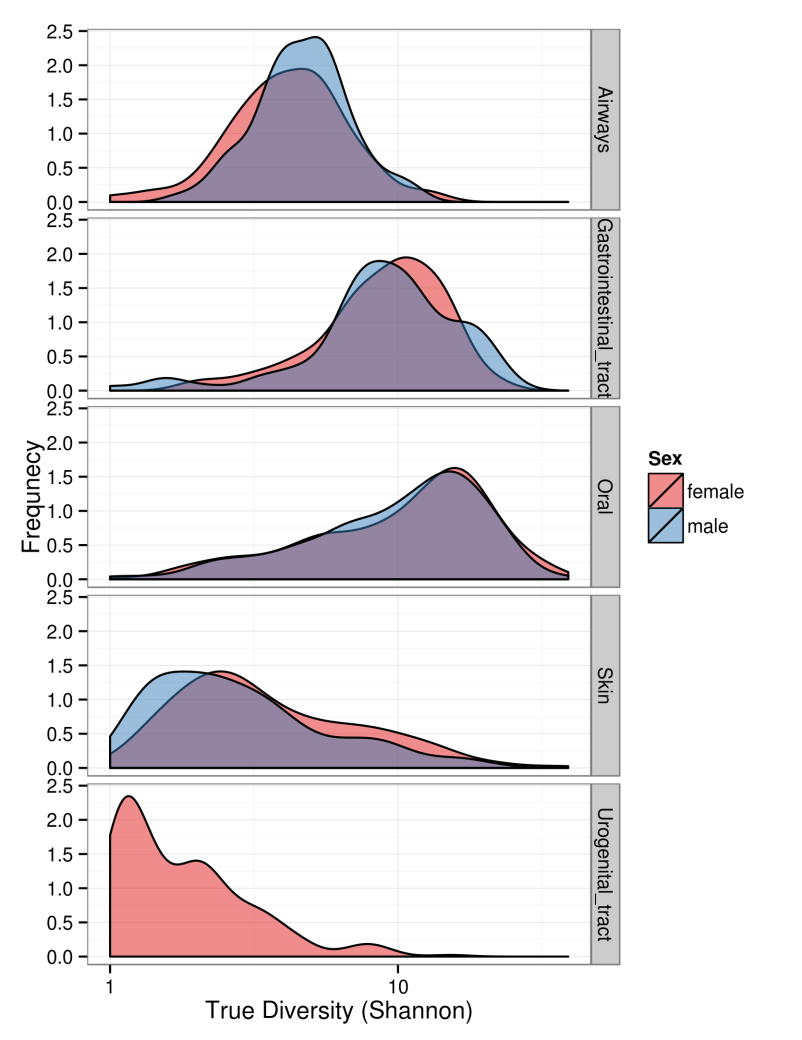


Figure 4 – species "true" diversity by body site and sex. The distribution of true diversity (Shannon index) by sex (red for female, blue for male) and body site (from top to bottom: airways, GI tract, oral, skin, and UG tract).

The diversity profile (Fig. 5) is consistent with this observation, and the trend of diversity as a function of the Hill number is similar in all body sites, rapidly declining as the effect of rare species decreases.

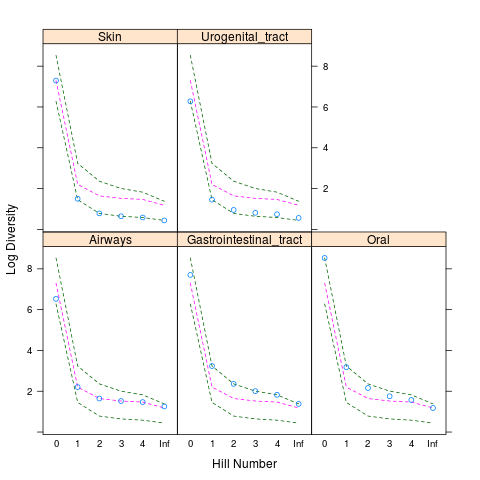


Figure 5 – Diversity profile by body site. The Log of the diversity as a function of Hill numbers for the different body sites.

## Beta diversity

First, I analyzed the beta diversity of the different body sites. From the beta diversity triplot in Fig. 6 it seems that the beta diversity is mainly influenced by species that occur in one body site but not in the other one (b and c), rather than by species that occur in both body sites (a).

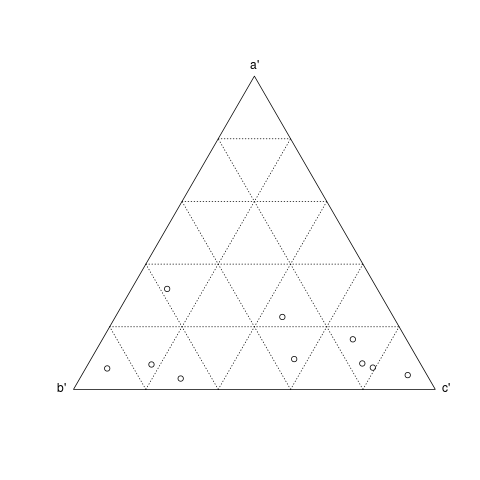


Figure 6 – Beta diversity triplot by body site.

Next, I analyzed three different beta diversity indices: Jaccard , Simpson , and Lennon (see Fig. 7). These indices were chosen because they present relatively different features. Jaccard's index (top) shows that the oral body sites share more species than the other body sites, on average, but there is a lot of variance. Simpson's index (middle) is similar, but inverse, and puts less weight on species that occur only in one person but in in another person. Lennon's index (bottom) checks the symmetry in segregation of species between two persons; the results are all around 0.5 but it seems that the oral samples have an increased variance, meaning that some persons had more unique species.

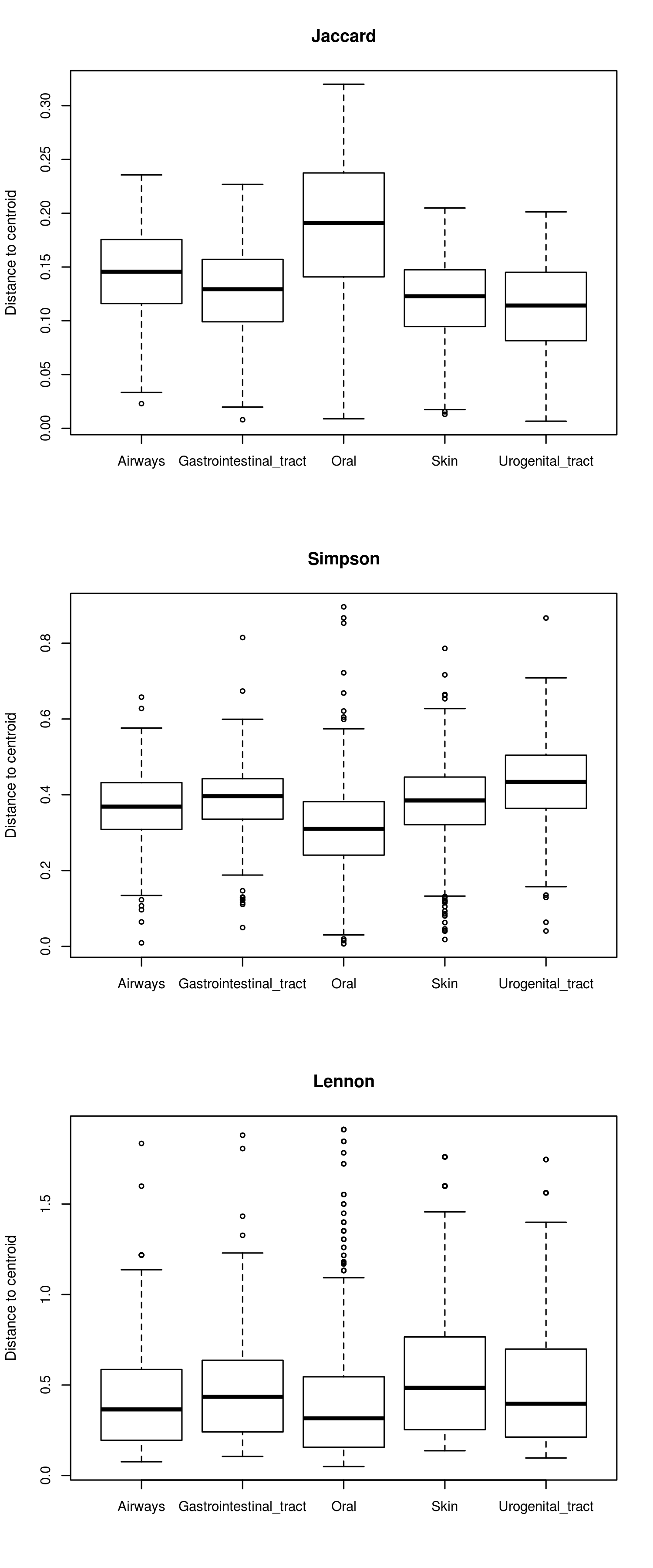


Figure 7 – Beta diversity by body site.

## Ordination

I performed nMDS on the data to find a species and sites ordination. The stress level was ~0.2.

### Sites ordination

The sites ordination (Fig. 8) clearly shows that samples are clustered by body sites, but not sex, with the only overlap between skin and airways. Consistent with our observation from the beta diversity analysis, the variance in the oral samples is larger (demonstrated by the relatively many green markers far from the centroid of the green cluster); however, this can be a result of an increased sampling effort in the oral sites (Fig. 1).

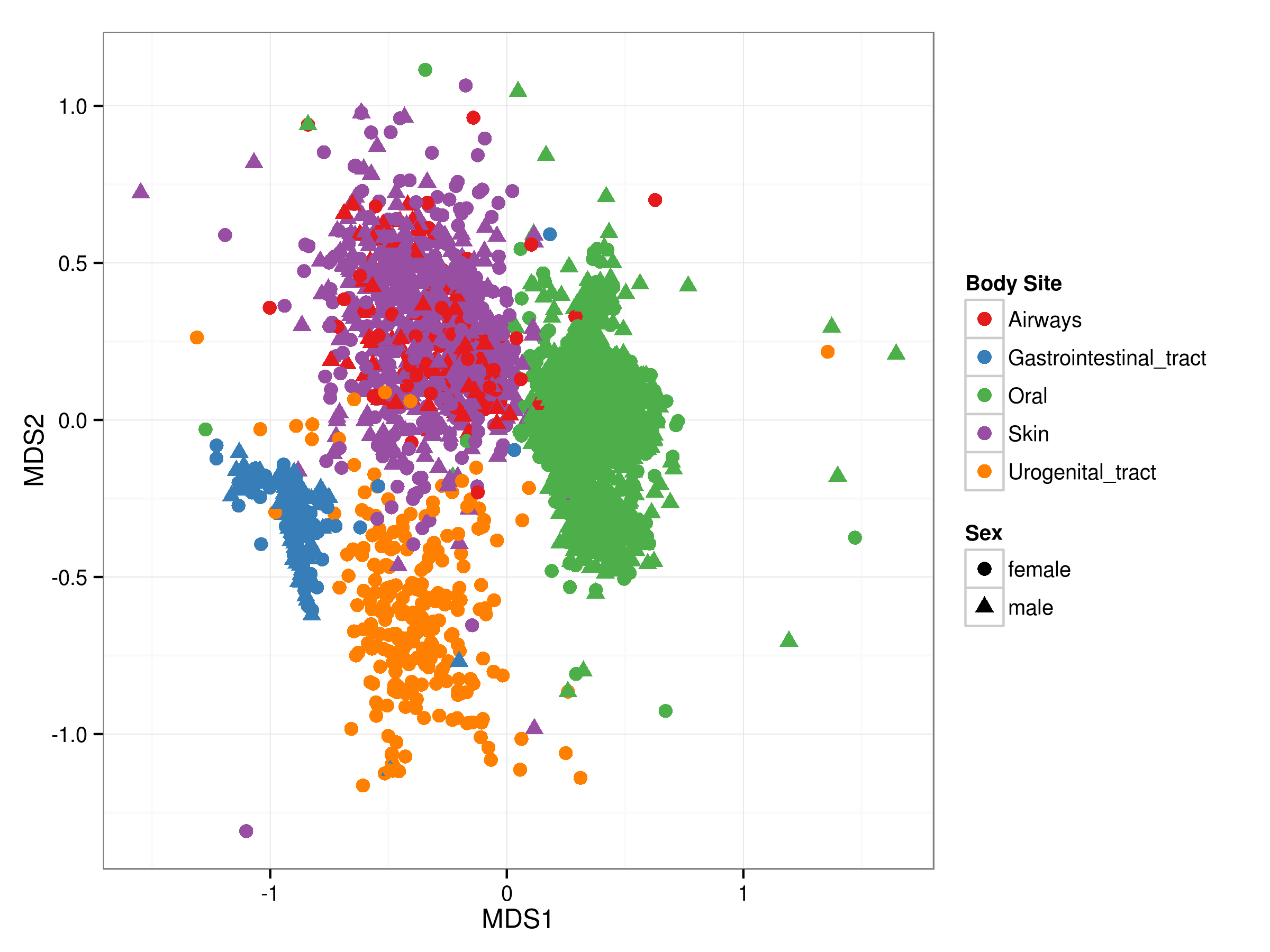


Figure 8 – Site ordination (nMDS). Sites are designated by body site (marker color) and sex (marker shape).

### Species ordination

The species ordination (Fig. 9) demonstrates that there are 3-4 distinct species cluster. The left and right clusters are composed of frequent species (species that appear in many samples); the species in the left cluster are mostly bacteroidetes, while the species in the right cluster are from all the kingdoms. There is another cluster in the middle of the ordination plot (or maybe two cluster, one in the middle and one at the bottom); this cluster is composed of rare species of all kingdoms.

It seems that the species ordination has caught something here – especially interesting is the left, homogeneous cluster – but the separation is either lacking, either due to high dimensionality of the underlying data, the complex and somewhat arbitrary concept of OUT, and the possibly messy composition of bacterial communities.

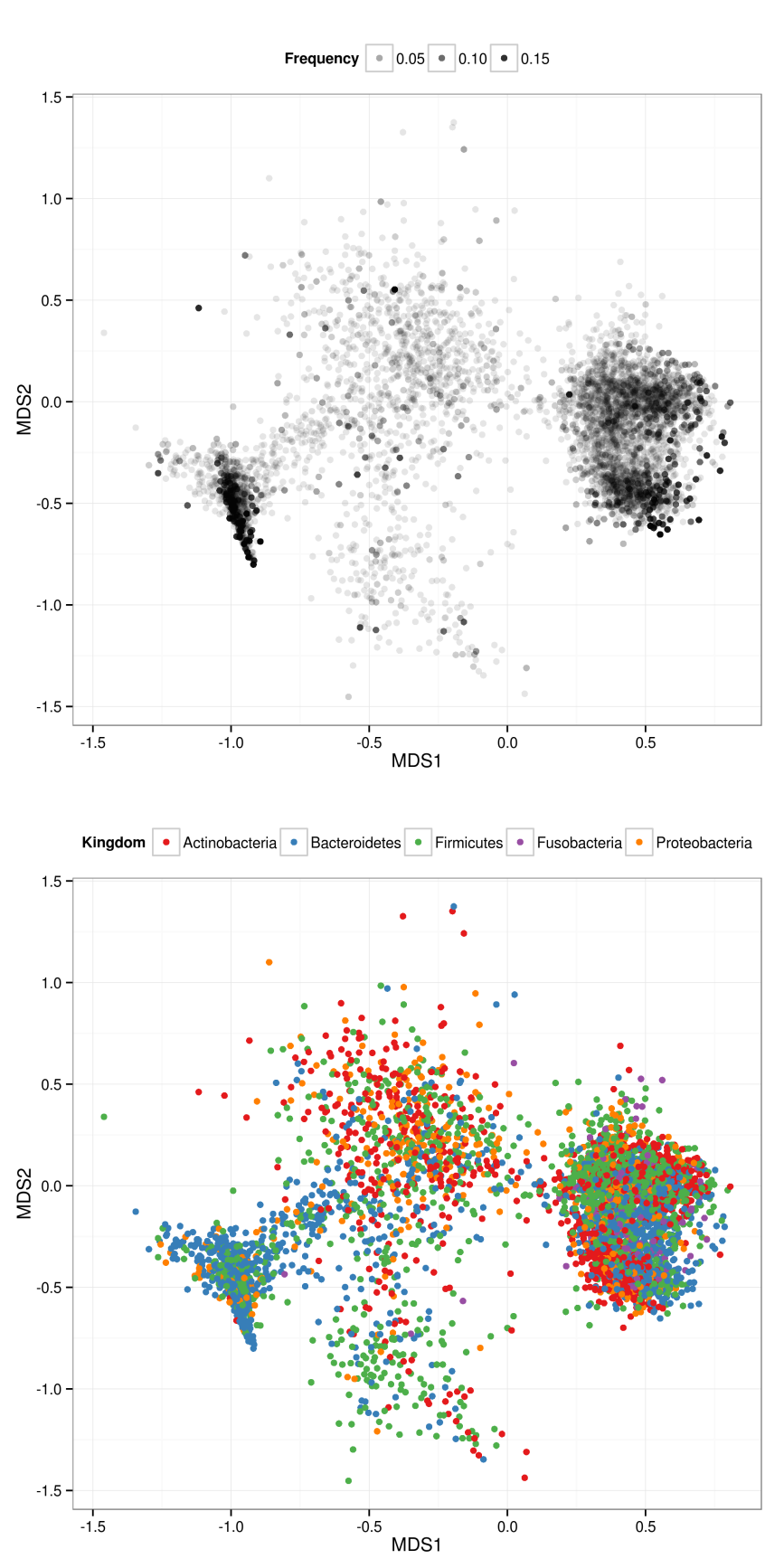


Figure 9 – Species ordination. The top panel shows the species by their frequency in the samples (the darker the marker the more frequent the species); the bottom panel shows the species by their Kingdom (denoted by different colors; only the five most frequent kingdoms are shown).

## Assembly rules

I studied the correlation in the species composition between the different sites. Fig. 10 shows a correlation map (

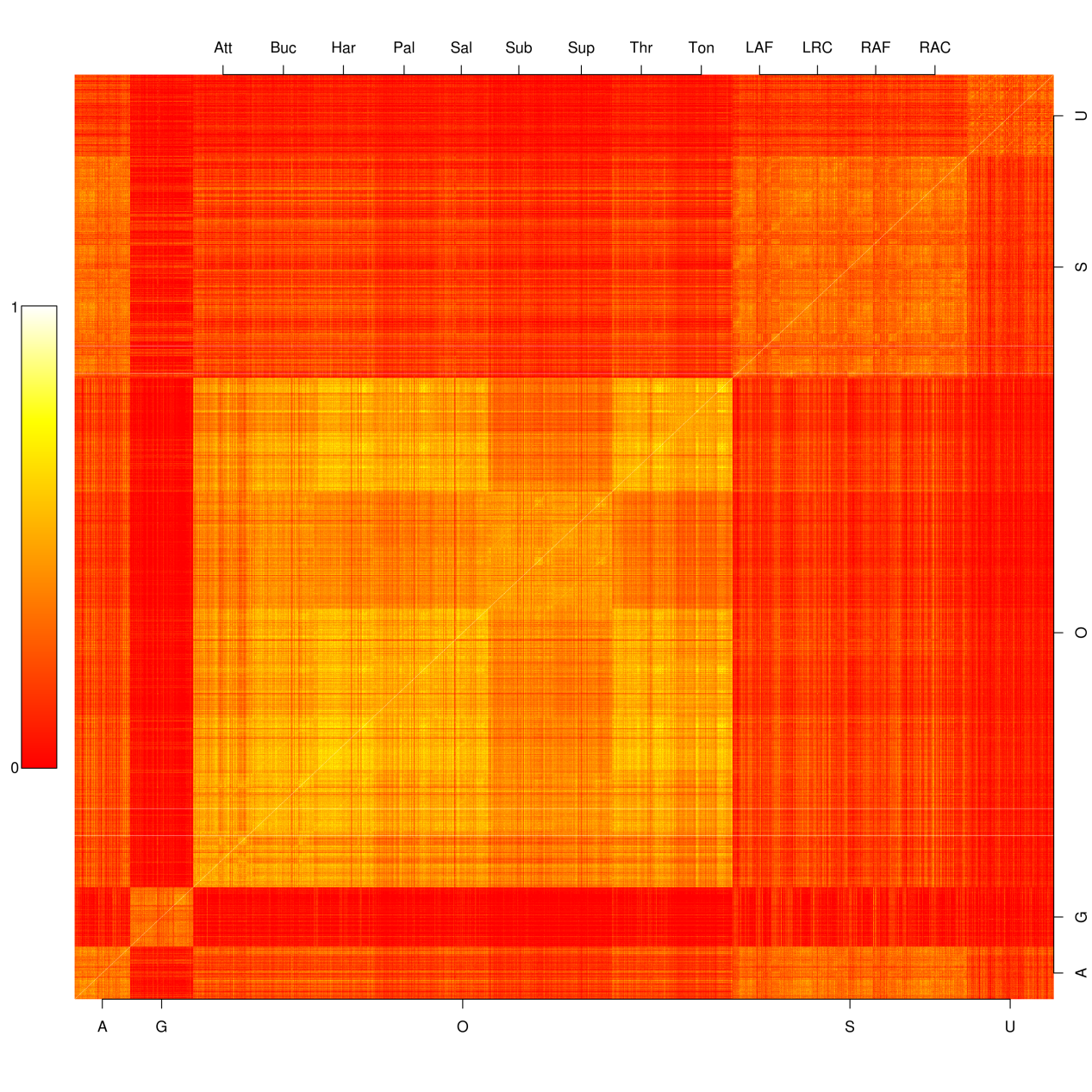


Figure 10 – Correlation between samples. Yellow for high correlation, red for low correlation. The samples are ordered by body site (bottom and left axes: A – airways; G – GI tract; O – oral; S – skin; U – UG tract) and sub-sites for the oral and skin sites (top axis, oral: Att - Attached Keratinized gingiva; Buc - Buccal mucosa; Har - Hard palate; Pal - Palatine Tonsils; Sal – Saliva; Sub - Subgingival plaque; Sup - Supragingival plaque; Thr – Throat; Ton - Tongue dorsum; skin: LAF - Left Antecubital fossa; LRC - Left Retroauricular crease; RAF - Right Antecubital fossa; RAC - Right Retroauricular crease.

# Part 2: Open questions