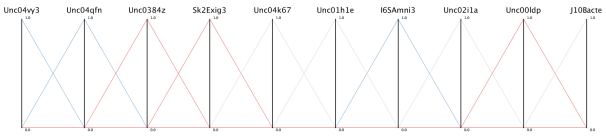
Quality Metrics to Guide Visual Analysis of High Dimensional Genomics Data

Supplemental Material

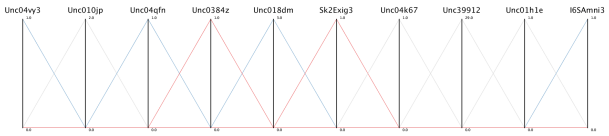
1 Examples for Preterm Infant Data

This section includes additional examples of visualization guided by the suggested quality metrics, of the Stewart et al. [3] dataset of samples from the gut microbiome of preterm infants, which includes 516 biological entities across 867 samples, complementing the examples in the paper. In these examples, the samples are classified by *Birth Mode*, with *Cesarean Birth* (red) and *Vaginal Birth* (blue) as sample groups.

Figure 1 displays the entities that are lowest ranked with the abundance and prevalence quality metrics. These show a typical pattern in genomics data, where a large number of entities are only detected in a small number of samples at very low counts, leading to low abundance and prevalence. These entities would commonly not be of great interest for analysis on their own, but for instance the combination of low prevalence and high abundance (entities detected in high counts in one or a small number of samples) can be of interest as it may indicate an outlier of interest to investigate further.



(a) The ten biological entities that was ranked lowest with the abundance quality metric.



(b) The ten biological entities that was ranked lowest with the prevalence quality metric.

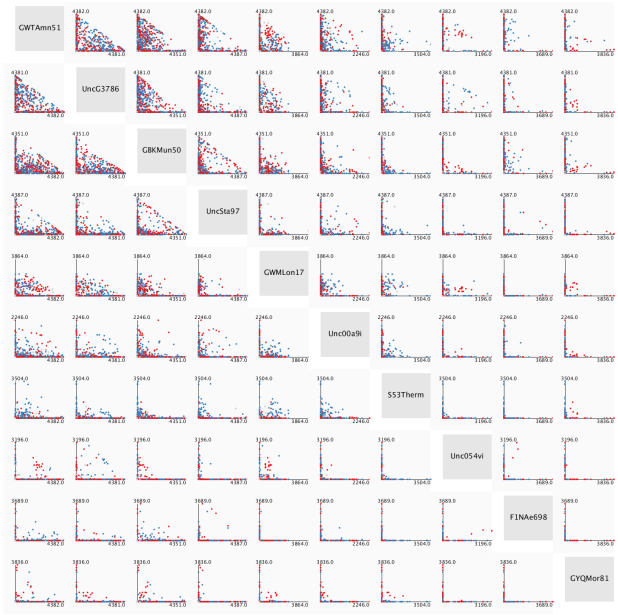
Figure 1: The entities that are lowest ranked with the abundance and prevalence quality metrics.

Figure 2 shows the biological entities with the highest values for the abundance and prevalence quality metrics, using Scatter Plot Matrices (SPloM). The data displayed here are the same as in figure 2 in the paper. The last three entities (bottom and left in both plots) are different for abundance and prevalence, and it is visible from these that for the prevalence SPloM (figure 2b), the entities are detected in a higher number of samples (samples with non-zero values) but the counts are lower (maximum count between 711 and 1790), compared to the abundance plot (figure 2a) where the entities are detected in fewer samples but at higher counts (maximum count between 3196 and 3836).

Figure 3 shows the biological entities that are lowest ranked with the sample group difference metrics. For both abundance and prevalence the least difference is found in entities which have been detected in only a single or a very small number of samples (similar to in figure 1), as this means it has near zero abundance and near zero prevalence for both sample groups. These entities would commonly not be of great interest for analysis, and could thus usually be removed from more detailed visual investigation of the data.

Figure 4 displays the ten highest ranked biological entities based on prevalence difference between sample groups (same entities as in figure 3 in the paper). Examples are visible of entities that are almost only prevalent in samples in the blue group (*Vaginal Birth*) (x-axis in first plot, both axes in second plot and y-axes in last two plots, for example), or that is mainly prevalent in samples in the red group (*Cesarean Birth*) (for example the x-axis in the rightmost plot). This can, for instance, help identifying biological entities that are possibly only prevalent under certain circumstances, as defined by the sample group.

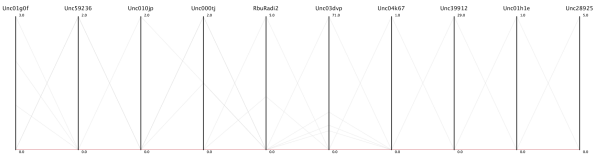
Figure 5 displays the same entity subsets as in figure 4 in the paper. These are the biological entities that have been highest ranked on abundance difference between the two sample groups, using a cluster separation metric in figure 5a and difference in average abundance in figure 5b. In figure 5a there are a number of visible examples of where the groups are clearly separated and the blue samples (*Cesarean Birth*) have considerably higher abundance (for the entities represented by both axes in the first plot, by the y-axes in the second, third and fourth plot, and



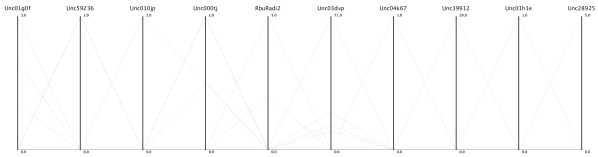
(a) The ten biological entities that was highest ranked with the abundance quality metric.

(b) The ten biological entities that was highest ranked with the prevalence quality metric.

Figure 2: The biological entities with highest values for the abundance and prevalence quality metrics, represented using SPloM



(a) The ten biological entities that was ranked lowest with the abundance based sample group difference metric, using the Davies-Bouldin index as separation metric.



(b) The ten biological entities that was ranked lowest with the prevalence based sample group difference metric.

Figure 3: The biological entities that are lowest ranked with the sample group difference metrics.

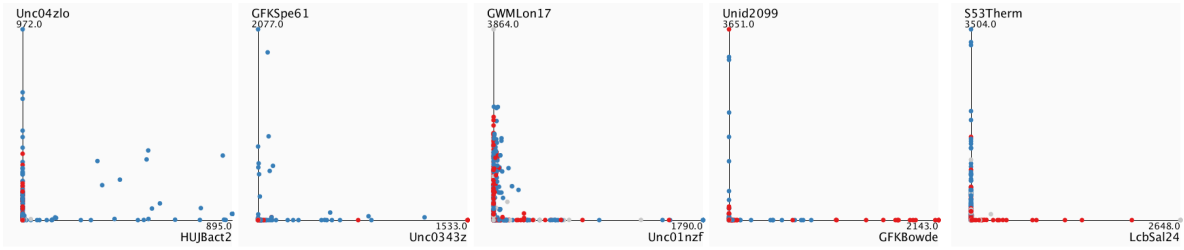
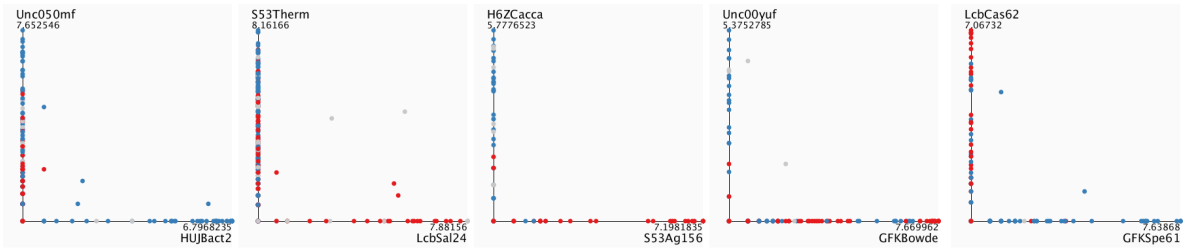


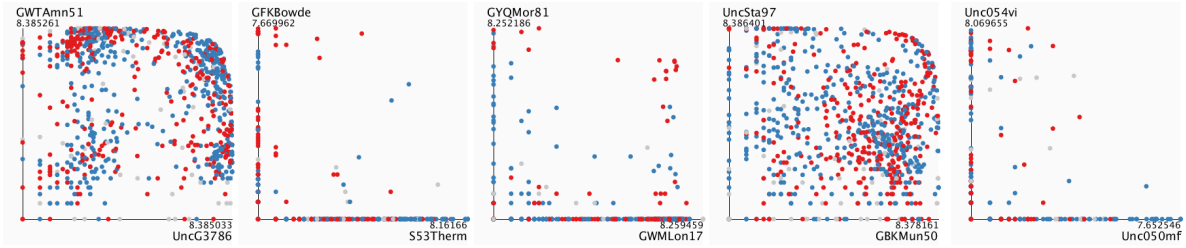
Figure 4: The highest ranked entities for prevalence difference between sample groups.

the x-axis in the fifth plot), and where the abundance is generally higher for red samples (*Vaginal Birth*) (x-axes in second, third and fourth plot, and y-axis in fifth plot). The sample groups are considerably less separated when the mean difference is used as a metric (figure 5b), indicating that a cluster separation based metric is generally better for identifying in which biological entities there are an abundance difference between sample groups.

Figure 6 shows examples where a subset of entities have been selected based on a summarised correlation quality metric, with a prevalence threshold of 50 %. In figure 6a the entities are ordered using the correlation based ordering described in Johansson and Johansson [1], while the entities in figure 6b are not ordered by the correlation metric. Through the subset selection and ordering, the coexistence of *GWMLon17* (a *Bifidobacterium*), *HUJBact2* (an



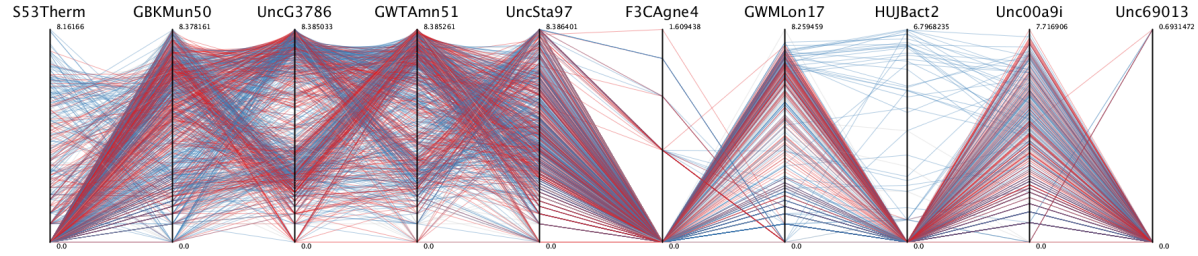
(a) Sample group difference identified using the Davies-Bouldin index.



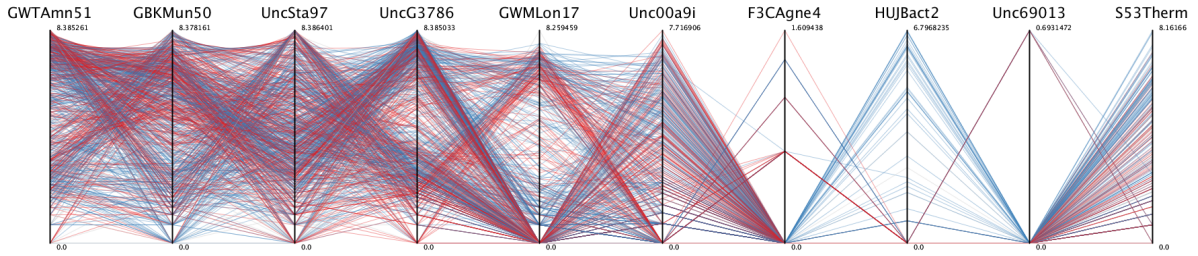
(b) Sample group difference identified based on average abundance.

Figure 5: The highest ranked biological entities for abundance based sample group separation, using two different approaches for measuring the difference, visualized using Scatter Plots with logarithmic scaling.

Actinomyces) and *Unc00a9i* (a *Veillonella*) in some blue samples is visible in figure 6a (the blue lines at the top of the 2nd to 4th axis from the right). This is not as clearly visible in figure 6b where the metric based ordering is not applied (where the entities are represented by the 6th, 5th and 3rd axes from the right), demonstrating the benefit of utilising metrics both for ordering and subset selection.



(a) Axes are ordered based on the pairwise Q_{Sim}.



(b) Axes are not ordered based on pairwise Q_{Sim} correlation.

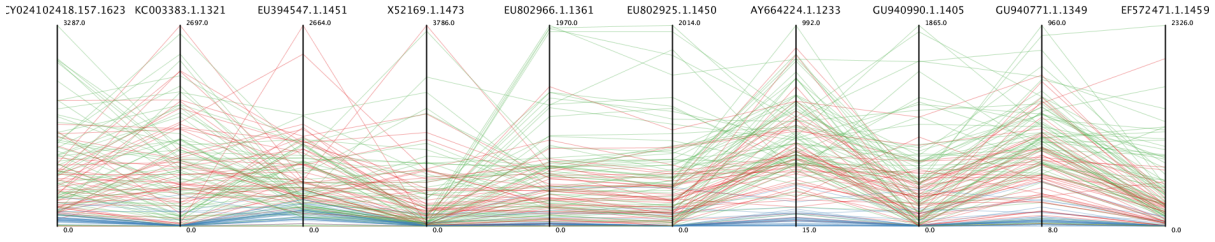
Figure 6: Selection based on summarised Pearson correlation.

2 Examples for the Tara Oceans Dataset

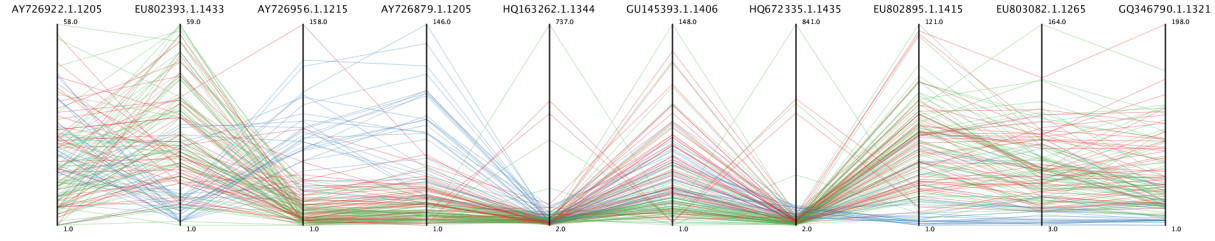
This section includes further examples of visualization guided by some of the suggested quality metrics, using the public Tara Oceans dataset. The dataset, which is described in detail in Sungawa et al. [4] and Pesant et al. [2] includes data from 139 samples, with a total of 35650 biological species (i.e. biological entities) detected. In the

examples here, samples are coloured by *Layer of Origin* (as also mentioned in [4, 2]), with three categories: *Surface* (green) which is the top layer, *Deep Chlorophyll Maximum* (red) which is the middle layer, and *Mesopelagic* (blue) which is the deepest level.

Figures 7 and 8 display the ten highest ranked biological entities using the abundance and prevalence metrics, using PC and SPloM respectively. From the abundance based selection (figures 7a and 8a) it is visible that the overall most abundant entities are mainly abundant in high counts in the green and red layers, while only at lower counts or not at all in the blue samples. The most prevalent entities (figures 7b and 8b) have on the other hand been detected in all samples (visible from that the lowest value on the PC axes are above 0), but are generally detected at low counts (between 58 and 148 as maximum values on the axes, compared to 960 to 3786 for the high abundance entities).



(a) The ten biological entities that were highest ranked by the abundance metric.



(b) The ten biological entities that were highest ranked by the prevalence metric.

Figure 7: The highest ranked biological entities, based on abundance and prevalence, displayed using PC.

One aim of analysing a dataset like the Tara Oceans dataset may be to understand the difference in the microbiome of different ocean layers. The suggested metrics for sample group differences can support this by suggesting entities for further investigation where the differences are relatively big. Figure 9a displays a subsets of ten biological entities with a clear separation of sample groups, based on the abundance separation metric using the Davies-Bouldin index. The differences in abundance between the blue, red and green samples are clearly visible, indicating a number of biological entities that are abundant in higher counts in the deepest layer (blue) but in low counts in the surface layer (green), as well as two (third and fifth axis) that are abundant in higher counts in the surface and middle layers. Figure 9b provides a comparison what the result would have been if using the difference of average values of sample groups instead of the cluster separation metric. While differences are visible also when using the average value approach, the green and red samples are considerably more mixed, thus confirming that the cluster separation approach in figure 9a provides a better metric for sample group separation.

Figure 10 displays the ten highest ranked biological entities based on prevalence difference between the sample groups, visualized using PC and Scatter Plots. As prevalence is only based on if an entity is detected or not, independent of its count, the result is different to the abundance difference. The ten entities display a similar prevalence pattern with high prevalence in the deepest layer (blue), some prevalence in the middle layer (red) and no or nearly no prevalence at all in the surface layer (green). A potential conclusion to draw from this is that there are some species that generally only exist in the deeper layer of the ocean, and that the selected set of entities represent examples of such species.

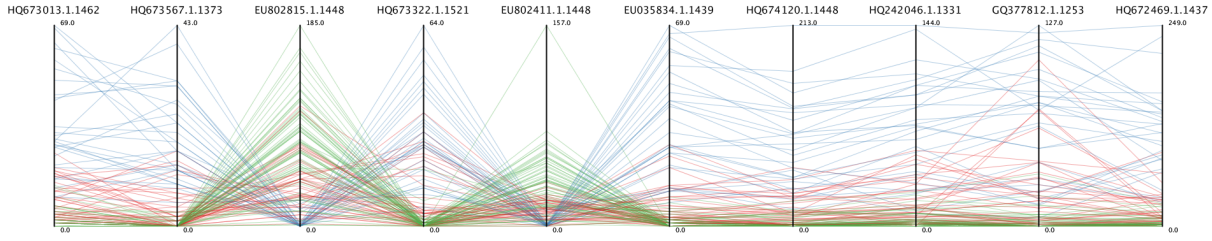
References

[1] Sara Johansson and Jimmy Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000, 2009.

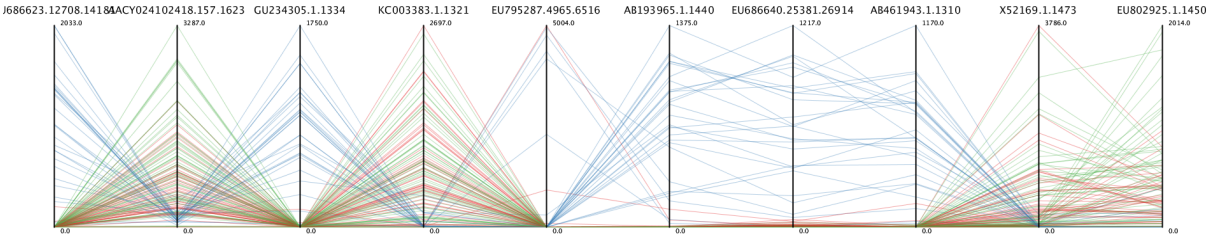
(a) The ten biological entities that were highest ranked by the abundance metric.

(b) The ten biological entities that were highest ranked by the prevalence metric.

Figure 8: The highest ranked biological entities, based on abundance and prevalence, displayed using SPloM.



(a) Sample group difference identified using cluster separation with the Davies-Bouldin index



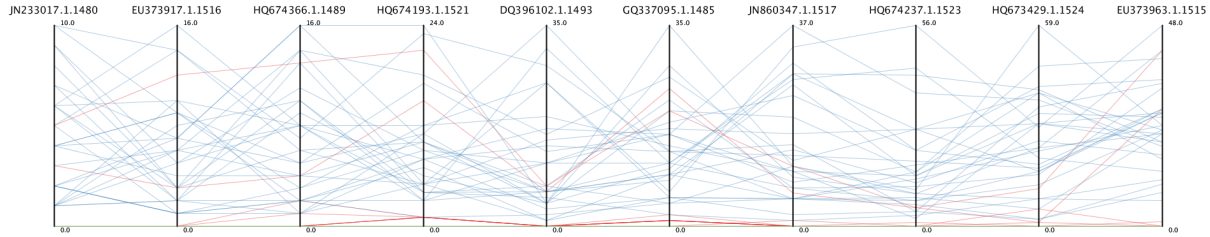
(b) Sample group difference identified based on average abundance.

Figure 9: Biological entities with highest abundance difference between sample groups, using different metrics.

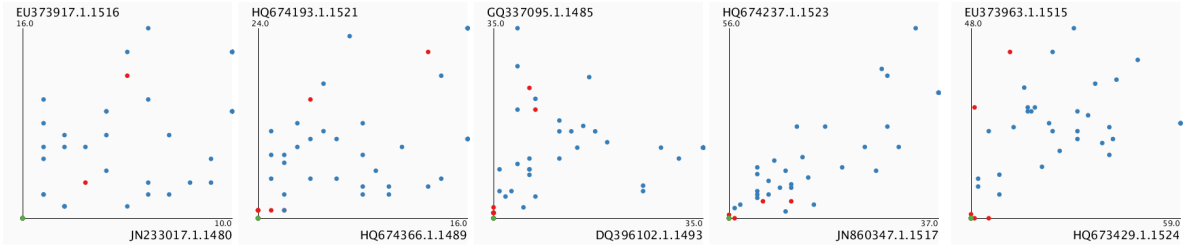
[2] Stéphane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Sabrina Speich, Romain Troublé, et al. Open science resources for the discovery and analysis of tara oceans data. *Scientific data*, 2(1):1–16, 2015.

[3] Christopher J Stewart, Nicholas D Embleton, Elizabeth Clements, Pamela N Luna, Daniel P Smith, Tatiana Y Fofanova, Andrew Nelson, Gillian Taylor, Caroline H Orr, Joseph F Petrosino, et al. Cesarean or vaginal birth does not impact the longitudinal development of the gut microbiome in a cohort of exclusively preterm infants. *Frontiers in microbiology*, 8:1008, 2017.

[4] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237), 2015.



(a) The ten highest ranked entities visualized with PC.



(b) The ten highest ranked entities visualized with Scatter Plots.

Figure 10: The highest ranked biological entities for prevalence difference between sample groups.