Identification of condition-specific microbial populations from human metagenomic data

Luxembourg Centre for Systems Biomedicine

Cedric C. Laczny, Patrick May, Nikos Vlassis, Paul Wilmes

Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg

Introduction

While mixed microbial communities (MCs) usually co-exist in a well-balanced state with the human host, several diseases have recently been associated with changes in the compositions of MCs, based on comprehensive culture-independent characterization via next-generation sequencing (metagenomics)[1,2].

However, the identification of individual populations in MCs is hampered by the limited availability of reference genomes, among others due to "unculturability" of most isolates under standard laboratory conditions. Hence, efficient means that do not rely on *a priori* knowledge are required for the deconvolution of individual MC compositions as well as for the comparison of collections of communities under different conditions, e.g., case vs. control.

This has led us to develop a reference-independent and alignment-free approach which enables the visualization of data-inherent taxonomic structure of short metagenomic fragments (~1,000nt) without prior training on subsets of longer fragments, e.g., above 5,000nt. Our approach[3] provides increased sensitivity and specificity over existing approaches, scales well with the usually big data amounts in metagenomic data (run-time of less than 1h for around 100,000 genomic fragments), and produces more intuitive two-dimensional visualizations (maps) compared to state-of-the-art methods (e.g., Emergent Self-Organizing Map (ESOM)-based approaches). The improved discriminatory performance is supported by independent parallel work[4].

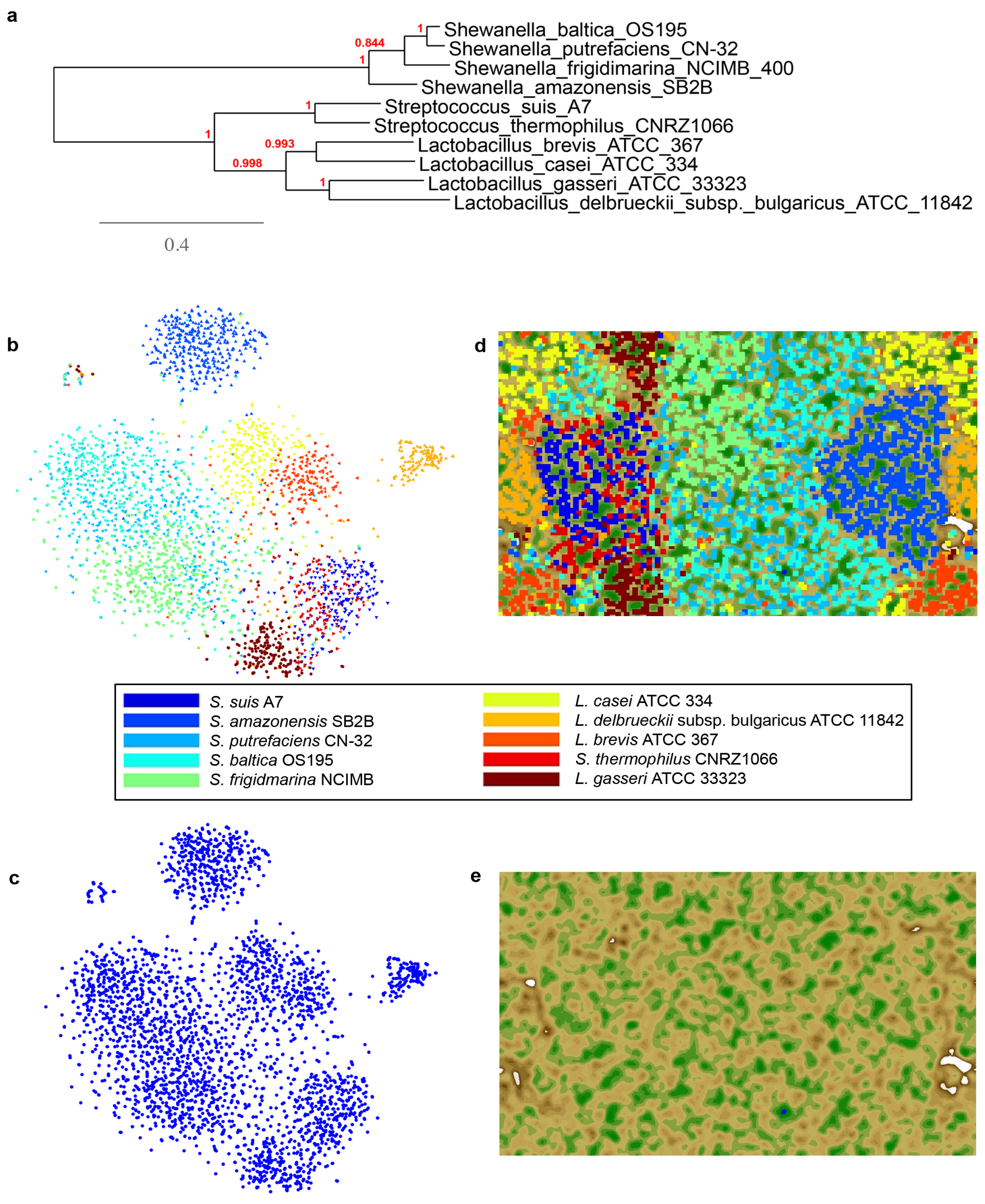
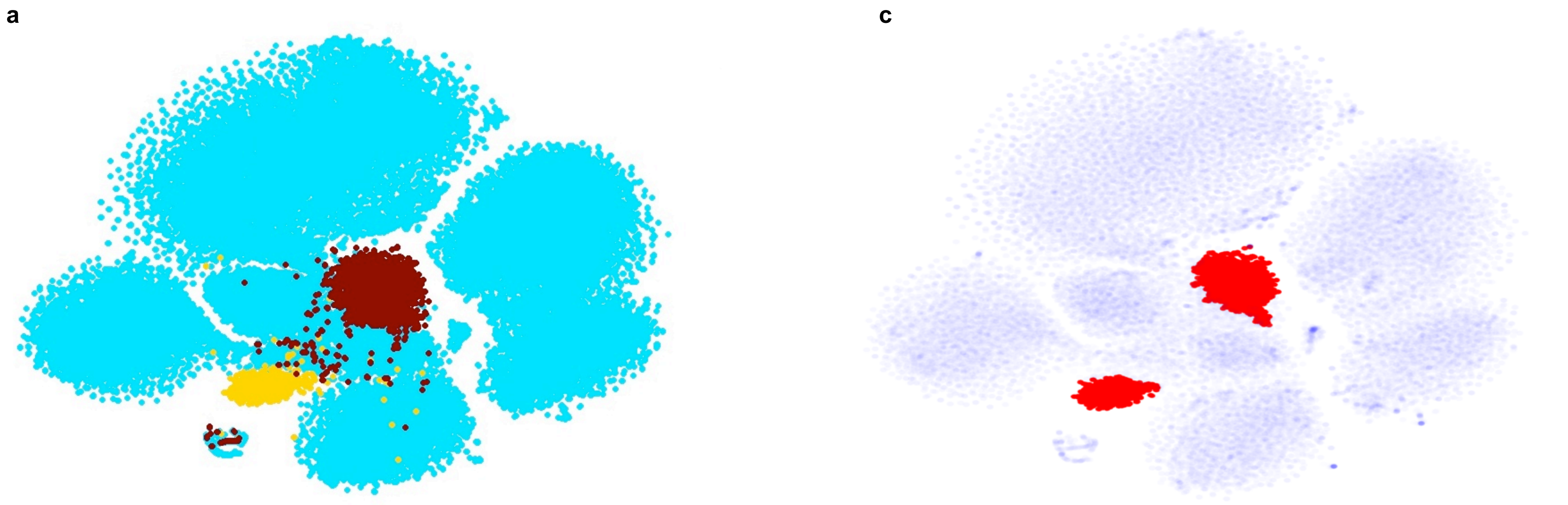


Figure 1) Simulated community of 10 organisms: (a) Phylogenetic tree based on rpoB gene. Taxonomic structures **(b,c)** based on our approach vs. **(d,e)** based on ESOMs

Strategy

In addition to the visualization of data-inherent taxonomic structure from individual MCs, the overlay of the resulting taxonomic structures enables the visual comparison of MC compositions of samples derived from different conditions[5]. To facilitate the identification of genomic fragments belonging to condition-specific microbial populations, we propose the concept of “condition-specific weights”. Condition-specific clusters can thus be intuitively identified without prior annotation or taxonomic identification.



b

N_x denotes the set of n nearest neighbors of point x

$$r_x = \frac{1}{n}\#\{y \in N_x : condition(y) = 0\}$$

$$w_x = \left[\frac{r_x - 0.5}{0.5}\right]^2$$

Figure 2) Illustration of the concept of condition-specific weights: (a) Two-dimensional embedding of simulated genomic fragments demonstrating two condition-specific microbial populations (points in yellow or brick-red colour, respectively) and shared microbial populations (points in light-blue colour). **(b)** Formulae used to compute the “condition-specific weights” (w_x). **(c)** Differential plot. Points in red have a high weight (w_x >= 0.9), thus suggesting condition-specificity. The remaining points are plotted transparently.

Funding by: Fonds National de la Recherche Luxembourg

ATTRACT A09/03
AFR PHD/4964712

Results

We have applied our approach to a metagenomic dataset obtained via metagenomic sequencing of fecal samples from ten healthy human individuals. We observe a pronounced cluster structure (as well as a collection of smaller clusters) specific to one individual (P03) uniquely following a vegetarian diet.

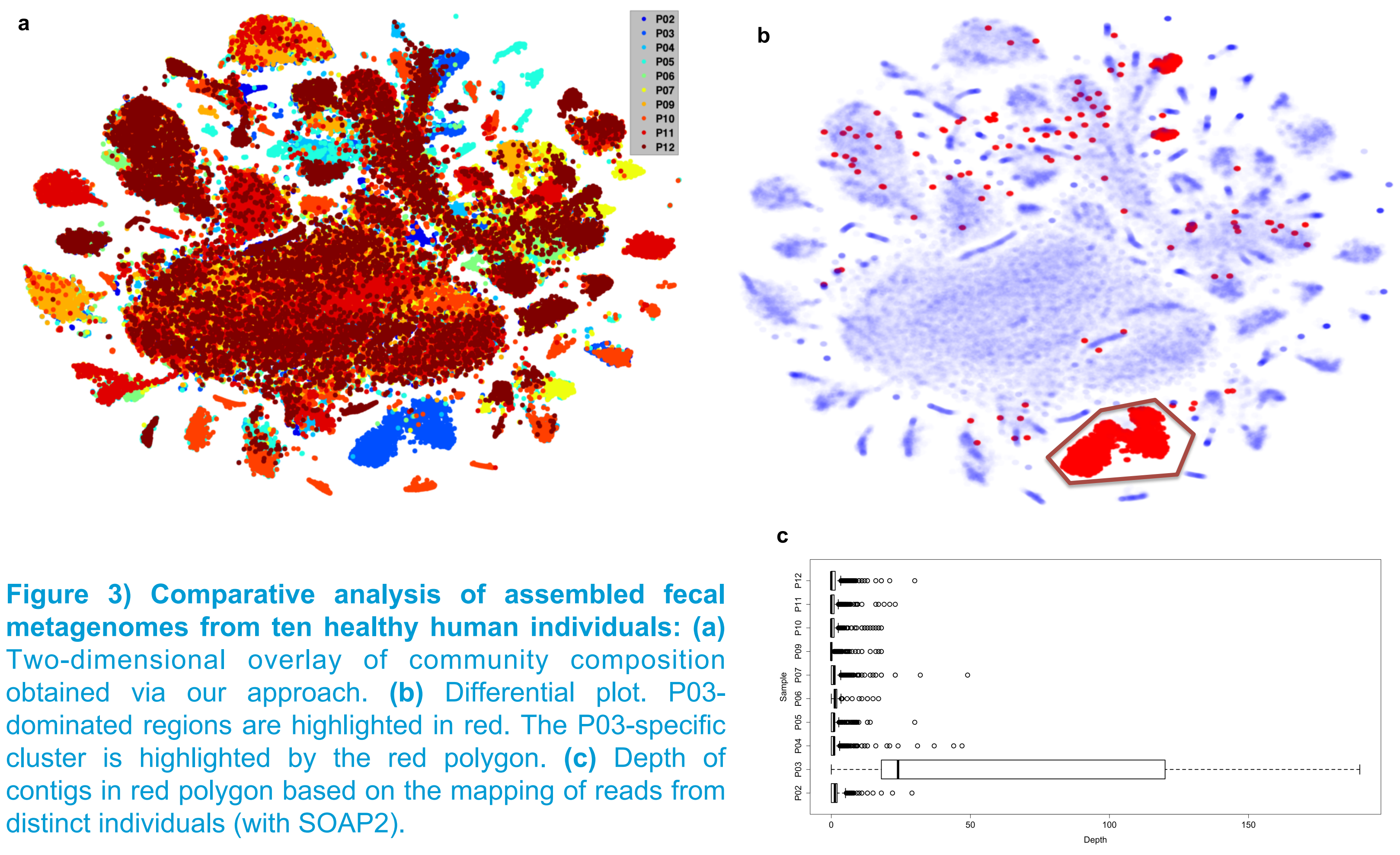


Figure 3) Comparative analysis of assembled fecal metagenomes from ten healthy human individuals: (a) Two-dimensional overlay of community composition obtained via our approach. **(b)** Differential plot. P03-dominated regions are highlighted in red. The P03-specific cluster is highlighted by the red polygon. **(c)** Depth of contigs in red polygon based on the mapping of reads from distinct individuals (with SOAP2).

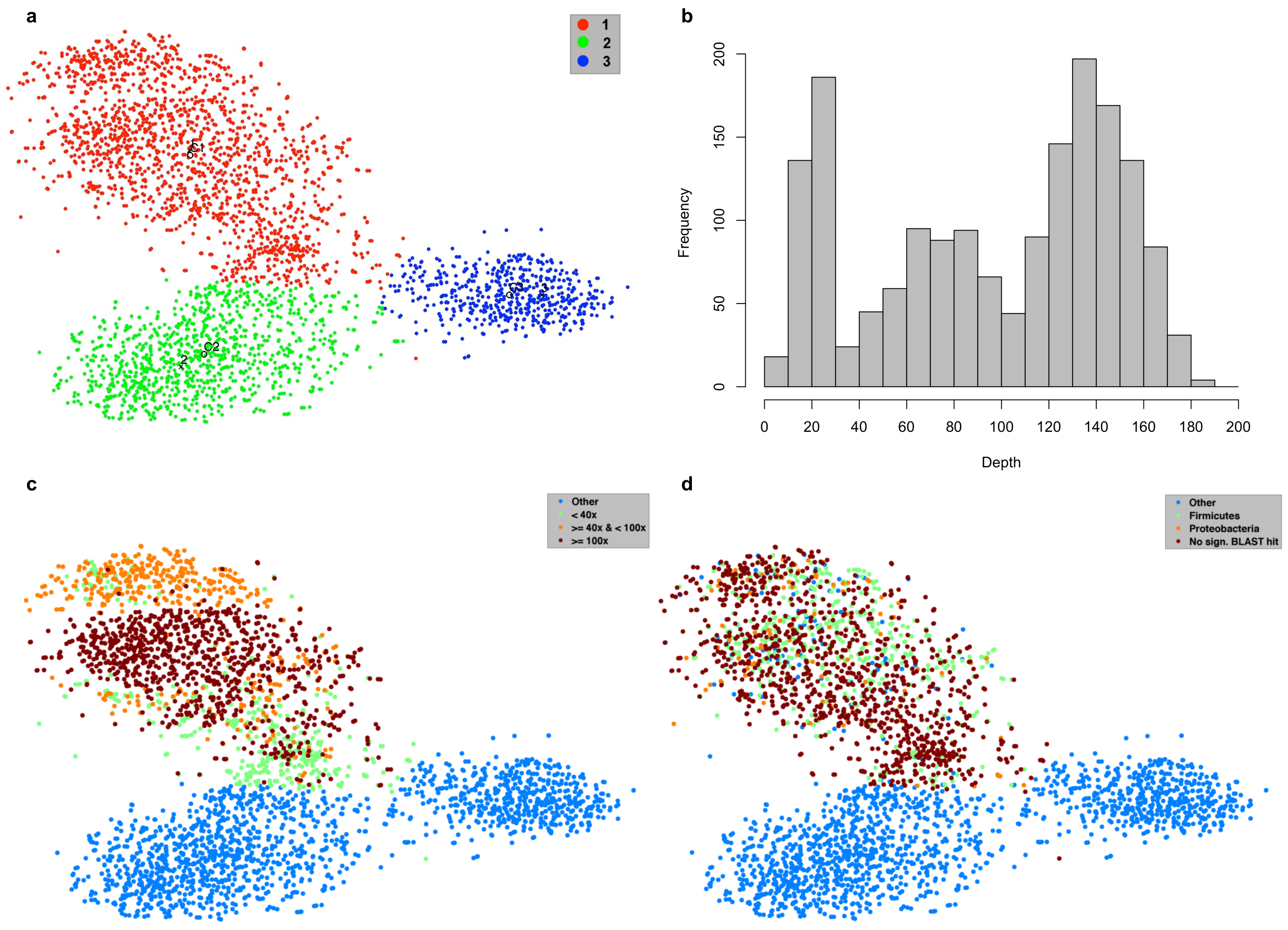


Figure 4) Examination of the P03-specific contigs: (a) Map of the P03-specific contigs as highlighted by the red polygon in Fig. 3b. Human-augmented clustering identified three subclusters (subcluster 1: red, subcluster 2: green, subcluster 3: blue). **(b)** Histogram of the depth of contigs in subcluster 1 (P03 reads aligned with SOAP2). **(c)** Points labeled according to their depth. Low (< 40x), medium (>= 40x & < 100x), and high depth(>= 100x) regimes were defined. Integration of the depth indicates three subpopulations in subcluster 1. **(d)** Integration of taxonomic information. P03-specific contigs were aligned against all existing reference genomes via NCBI's BLAST webservice and the Least-Common-Ancestor (LCA) was defined by the LCA-option in MEGAN.

Summary & Outlook

Besides the deconvolution of single microbial community compositions, our approach allows for the reference-independent identification of condition-specific microbial populations from human metagenomic data. It is robust with regards to parameter values, scales well with current metagenomic dataset sizes, and can be applied even to small cohorts. Computation of the weights (w_x) and weight-based selection in the high-dimensional space followed by reduction to two dimensions can potentially improve the identification of condition-specific genomic fragments. However, high dimensionality may require learning a specific distance function[6], which is the topic of future work.

References

1. Qin, J. *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. **Nature** 490, 55–60 (2012).
2. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. **PNAS** 109, 594–9 (2012).
3. Laczny, C. C. *et al*. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. In revision.
4. Gisbrecht, A. *et al*. Nonlinear dimensionality reduction for cluster identification in metagenomic samples. **17th Int Conf Inf Vis** (2013).
5. Patent filed at the Luxembourg Patent Office and pending.
6. Weinberger, K. Q. & Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. **J Mach Learn Res** 10, 207–244 (2009).

www.uni.lu/lcsb • University of Luxembourg • LCSB • 7, avenue des Hauts Fourneaux • L-4362 Esch-sur-Alzette • T. + 352 / 46 66 44-6973