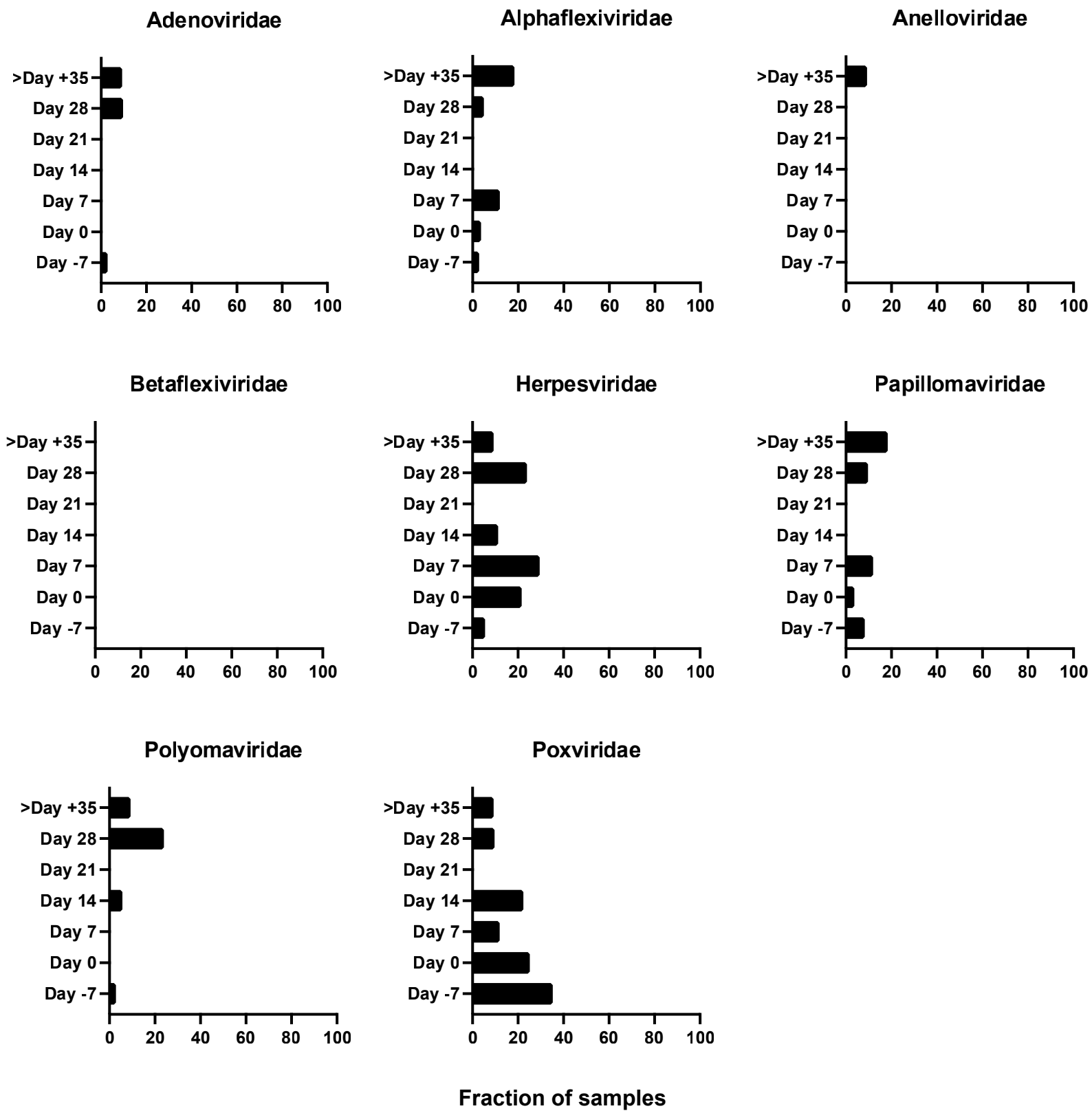
nature cancer

Article

https://doi.org/10.1038/s43018-023-00669-x

Bacteria and bacteriophage consortia are associated with protective intestinal metabolites in patients receiving stem cell transplantation

In the format provided by the authors and unedited



Virome pipeline utilizing metagenomic sequencing from virus-like particles purified from stool samples of allo-SCT patients was able to detect eukaryotic viral reads

Barplots showing proportion of patients in which the indicated eukaryotic viruses were detected at time-points relative to allo-SCT (day 0). Fraction of positive samples is shown.

Supplementary Figure 1

Supplementary Notes

MOFA and MEFISTO

MOFA

MOFA receives multi-omics data from the same set of samples as input, and generates a model that infers a set of "Factors" that best explain patterns of covariation across samples. To this end, MOFA decomposes the input data matrix of each omics modality into a matrix of Factor weights for each feature (e.g., the abundance of a specific bacterial ASV or the concentration of a metabolite) and a matrix of Factor values for each sample, that is shared across all omics modalities. "[MOFA] can be seen as a versatile and statistically rigorous generalization of principal component analysis (PCA) to multi-omics data".[1] MOFA Factors can be interpreted as sets of covariations of features within and across omics modalities that explain major sources of variance in the multi-omics dataset. These covariations are reflected by the magnitude and direction of the Factor weights for the different features. Larger weights indicate a higher correlation with a Factor, while the positive or negative sign indicates the directionality of the correlation (a positive weight represents a positive association and *vice versa*). Indirectly, this means that weights of different features of the same Factor pointing in the same direction suggest positive correlation between the features, while the opposite suggests negative correlation. We show Factor weights unscaled, enabling comparison of Factor weights of the same omics modality across Factors.

Although the Factors estimated via MOFA are derived from all available omics modalities (16S, ITS1, virome and metabolomics, in this study) in combination, subsequent analysis allows to infer the proportion of variance explained by the model in each omics modality separately, both in total and by Factor.

Of note, large differences in the number of features between the different input modalities can be problematic. We nevertheless decided to include the metabolite data to obtain a comprehensive model representation of the multi-omics dataset that reflects covariation between all modalities. The four-modality model is capable of explaining a substantial fraction of metabolomics data variance, and comparison of the four-modality model with a model excluding metabolite data showed no differences in explained variance for the other modalities.

Given the longitudinal nature of our dataset, our model provides Factor values for individual patients at specific time-points. For clarification, a sample with a high Factor value has on average a higher abundance of positive (high-) weight features and a lower abundance of negative (high-) weight features of that Factor.

MEFISTO

MEFISTO[2] is an extension to the MOFA framework that allows for explicit modeling of spatial or temporal dependencies between samples, if these dependencies are provided as a covariate during model fitting. MEFISTO incorporates a continuous covariate, either space or time, and provides a scale value for each Factor that reflects the degree to which that Factor is dependent on the covariate. It is then possible to assess how smoothly a given Factor varies with the covariate, i.e., the degree of variation captured by that Factor that is attributable to the covariate.

Of note, the Factors identified by our MEFISTO model were almost identical to those obtained with our MOFA model in regard to Factor values, Feature weight/covariance structure, variance explained across omics modalities and thus comparable to our MOFA-identified Factors.

Bacteriome and Fungome Data Processing for MOFA/MEFISTO

Amplicon sequence variants (ASV) were generated in R (4.1.0) from demultiplexed reads using a dada2 (1.22)-based workflow[3]. Standard parameters were used, except gap penalty, band size and OMEGA_A values, which were set to -1, 32 and 1e-30 respectively. Taxonomy of ASVs was predicted by the IDTAXA algorithm of the DECIPHER[4] package after matching to the SILVA[5] 138.1 release 16S reference database (bacteria). The February 2020 release of the Unite database[6] (fungal ITS) was used to predict fungal taxonomy. Relative abundances of assigned ASVs were calculated with the phyloseq[7] package version 1.36. Taxonomy for unclassified bacterial and fungal taxa were further refined by alignment to the Smartgene IDNS3 reference database.

References

1. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, (2018).
2. Velten, B. *et al.* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 (2022).
3. Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J. & Holmes, S. P. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. *F1000Research* **5**, 1492 (2016).
4. Murali, A., Bhargava, A. & Wright, E. S. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* **6**, 140 (2018).
5. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).

6. Nilsson, R. H. *et al.* The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* **47**, D259–D264 (2019).
7. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).