A beginner's guide to integrating multi-omics data from microbial communities

Anna Heintz-Buschart and **Johan A. Westerhuis** (Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands)

Microbial communities are immensely important and occur nearly everywhere, but their inner workings are still being discovered. The early years of microbiome research have been dominated by cataloguing the sheer diversity of microbes in these communities. Now, more and more studies try to understand connections between the microbes, between the way communities are built and how they function, and between their activity and the effects on their surroundings, including host organisms like humans. Omics measurements, or meta-omics as they are called when multiple organisms are measured at the same time, are a cornerstone in this endeavour. Here, we will discuss why their integration is important, how it can be achieved, what pitfalls may be avoided and which approaches are taken by integrative studies.

Omics analyses of microbiomes

Microbiomes are diverse communities of microorganisms that are responsible for essential environmental and host-associated processes. Many of their important functions are biochemical, involving the primary or secondary metabolism. These functions can be studied by omics analyses of the informative molecules of the central dogma of molecular biology, i.e., DNA, RNA, proteins and their metabolites. In this beginner's guide, we use the general term 'analyte' for either of these molecules, while we stick to DNA, RNA, protein and metabolite for specific cases. Technological advances in DNA and RNA sequencing and mass spectrometry of proteins and metabolites have driven considerable progress in cataloguing and understanding the molecular make-up and functioning of microbial communities in the last decades.

Omics technologies aim to measure as many analytes as possible within a system, e.g., all genes in a genome or all transcripts in a transcriptome (see Figure 1). The prefix 'meta-' indicates that the system under study comprises multiple species: in the case of a microbiome, this can consist of hundreds or thousands of different microbial taxa. Finally, multi-omics means the system-wide analysis of multiple analyte pools, e.g., the metagenome and metaproteome. The term is sometimes extended to multiple connected systems, e.g., the microbial metagenome and the host metabolome. Omics technologies generate large data volumes, whose processing and decoding have high computational demands. Nevertheless, these analyses are often less laborious than traditional, culture-based microbiological methods – in some cases, they are also the only option to learn about microbiomes when members of the microbiome are not culturable in isolation. Moreover, the phenotypes of isolates may not reflect their activity in a microbiome and in association with a host, because of interactions such as cross-feeding, inter-species signalling, chemical inhibition of competitors and immune responses.

Why integrate?

Each 'meta-omics level' is a proxy for the functions of the microbiome system. Each level provides information on only a part of this system. Integration of multiple omics levels can give more insight in the functioning of the whole system e.g., to answer questions on the production of a metabolite that may be beneficial or detrimental to a host (e.g., a short chain fatty acid in a gut bacterium). Measuring the metabolome may be a faithful proxy for the metabolite level, but the metabolome is a community measure. It does not enable us to distinguish between mechanisms, such as producers disappear or become inactive; producers invest into alternative metabolic pathways; and other community members metabolize our product of interest. Therefore, integration of, e.g., metagenomics, metatranscriptomics and metabolomics, including lipidomics, provides mechanistic hypotheses for a better understanding of the community state or dynamics. This can lead to mechanistic hypotheses,

Summary

- Microbial communities or microbiomes are made up of many different, interacting microbial species.
- To understand how microbiomes function, they are commonly studied by directly measuring 'meta-omes', e.g., the metagenome (the genomes of all present microbes), metatranscriptome (all transcripts of all microbes), metaproteome (all proteins), or metabolome (the metabolites of all microbes).
- Because every meta-omics data set provides only parts of the picture, integration of multiple omics data is key to mechanistic insights into microbial communities.
- Multi-omics study design and analysis makes provisions for biological and technical differences in the meta-omics data sets.
- Depending on the system under study and the research question, integration can make use of:
  1. common sequence information;
  2. data fusion methods to identify common patterns or interactions;
  3. prior knowledge of genetic or metabolic functions.

generalizable observations or indications of functionally important community members.

Integration can also improve the detection of analytes in one data set by borrowing information from another: the measurements of the second level can be adjusted (e.g., choice of instrumentation or sampling depth); or the data mining can be adapted (e.g., by setting the search space in metaproteomics). Integration can serve to validate the results of models that are based on observations of one level. Or it can be part of an exploratory study of a system that is not yet well described and where information is lacking at all omics levels (as was the case in the recent description of the Asgard archaea). If the state of the microbiome is of interest for classification purposes, e.g., as a diagnostic tool, the integration of several omics levels can improve the sensitivity and/or specificity and suggest biomarker panels made up of different kinds of analytes. Several approaches for integration of multi-omics data exist. They can be roughly divided into three groups (Figure 2): (a) the flow of genetic information through omics levels; (b) data fusion models that identify common patterns or interactions; and (c) prior knowledge on functional units, e.g., metabolic pathways. Combinations of the three approaches can be employed in the same study. The development of strategies to combine and complement these approaches is an active field of research.

What to consider when planning a multi-omics study

Which omics levels contain the most important information for the question at hand? How much knowledge is there on the system at that level, and

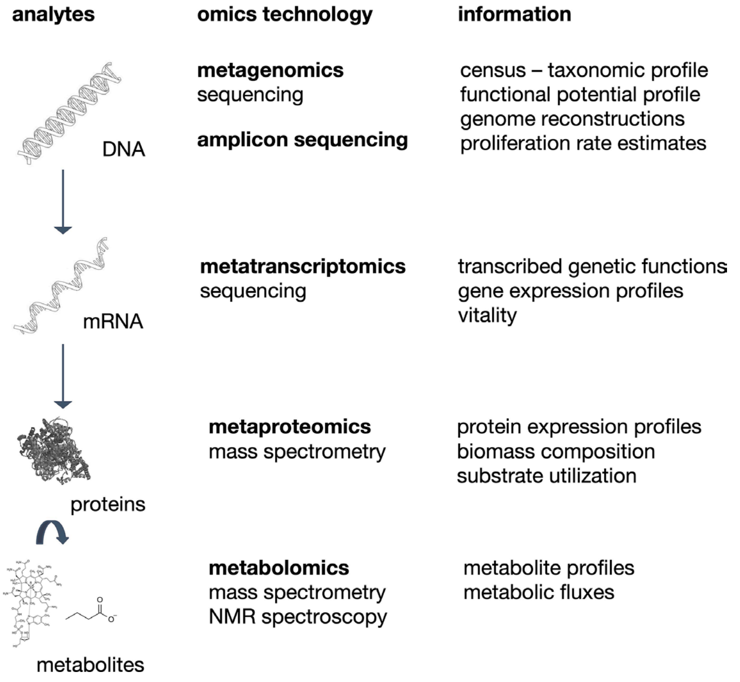


Figure 1. Meta-omics analytes, technologies, and associated information.

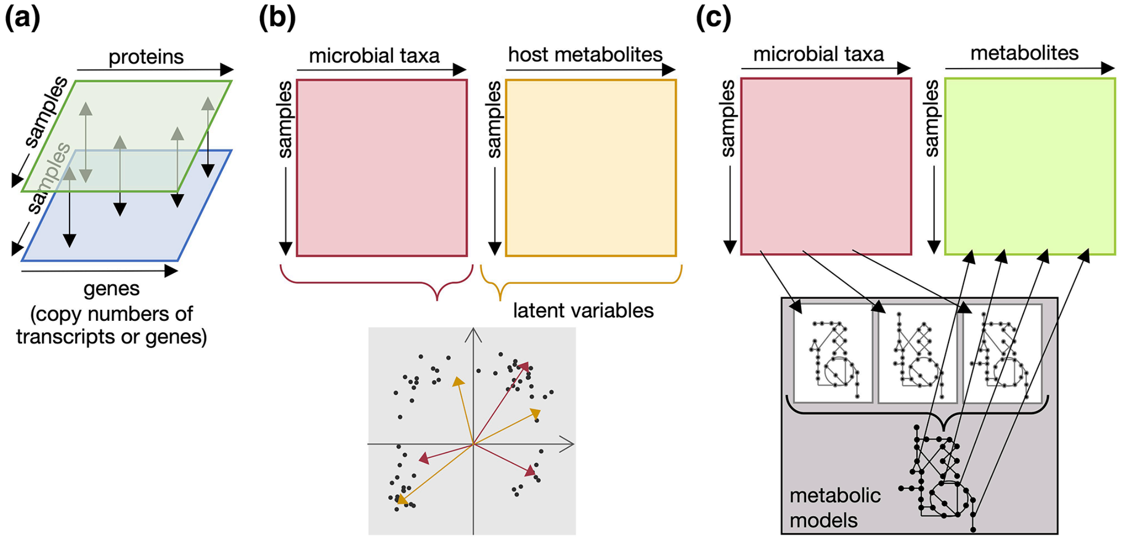


Figure 2. Schematic representation of integration strategies. Each meta-omics data set is visualized as a matrix of samples × analytes. In all examples, the samples that are measured at the two meta-omics levels are representative of the same systems. (a) Due to sequence identity, the analytes (e.g., proteins and genes) can be directly related across the meta-omics levels; (b) the relationship between the analytes in the two data sets is unknown and data fusion methods are employed to reveal common patterns within the samples (black dots) together with the most important analytes (arrows); (c) for well-described systems, links between the analytes in the two data sets are established based on existing knowledge, e.g., genome scale metabolic models. Metabolic models of each species (white boxes) are combined into common compartment models of the whole community and their host (grey box), of which the metabolites are obtained. Approaches to extending metabolic models to community level are still in their infancy.

can this be taken advantage of? For metagenomics or metatranscriptomics data, curated collections of genomes from isolates or metagenomes exist, including gene annotations and gene function predictions. The tens or hundreds of thousands of genomes in these collections far outnumber the species recorded in phenotype databases. Most of the millions of recognized gene families or orthologous groups are also not well described. Therefore, databases that link gene identities or orthologous groups to metabolic reactions, pathways, kinetics or models can only provide a means to integrate a limited proportion of sequence-based information with metabolomes.

How reliable are the different omics data sets? Each omics data set is affected by its technology's limitations. For example, metaproteomics can only reach a shallower sampling depth than sequencing-based analyses and often does not reach the same phylogenetic resolution. The time scales in which the analytes can be measured accurately also vary: DNA can be very stable, but it can also represent dead and dormant microorganisms. RNA can be very unstable and does not always yield reliable results when long experimental handling times are required. Proteins and metabolites have specific life times, with some being very labile and others outlasting their producers in the environment.

Should more omics or more samples be measured? There are usually many more analytes than samples, which aggravates problematic characteristics of omics data sets. The uncertainty in the identification of the analytes may also affect the observed abundance, including not reporting an analyte. Most omics data sets are not truly quantitative, e.g., the number of reads linked to one taxon does not specify a cell count, and the peak intensities of different metabolites cannot be compared, as each metabolite has its own response factor which depends on the type of metabolite. Frequently, the technological sampling or measurement depth is not adequate to capture all analytes and very high and very low levels cannot be measured with the desired accuracy.

How related are the omics data sets? Preferably, the different omics are measured from the same samples or subjects, to reduce the effects of individual variation on the integration. There are also technical differences in the data sets: the total counts in metagenomics and -transcriptomics data are capped by the sequencing technology (and, therefore, compositional), while for metabolites the abundance of one metabolite does not *per se* affect the detectable amount of all others (but there can be specific effects on detectability, as in ion suppression). Metatranscriptomics functional data has a high proportion of zeroes, which mainly occur due to

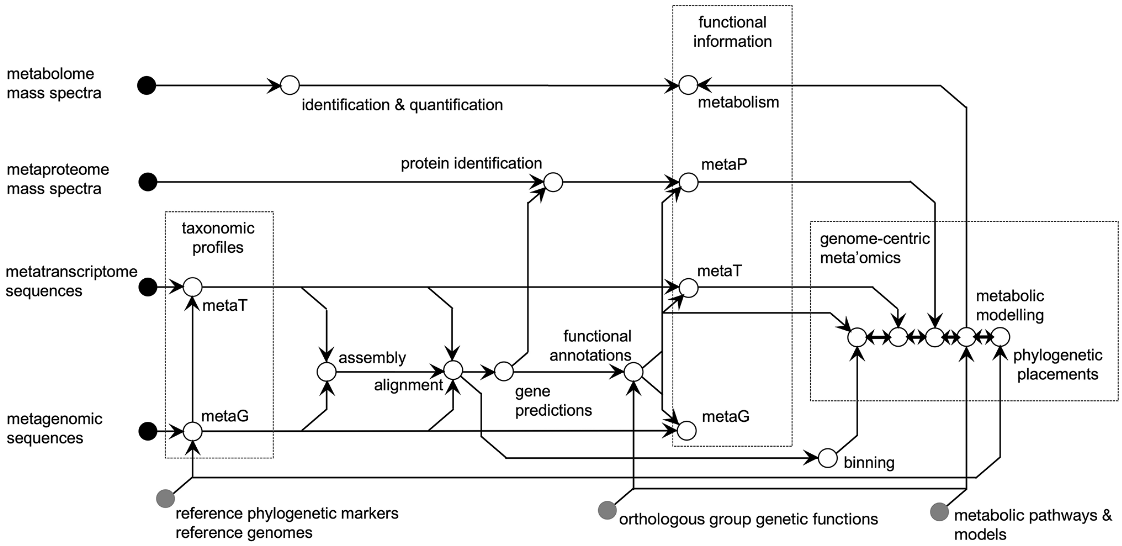


Figure 3. Example of information flow in a study integrating multiple meta-omics levels. Black circles indicate raw data, grey circles represent knowledge represented in databases, white circles are major data processing steps; boxes connect data sets that are connected by common information.

under-sampling, while the high (strain-level) resolution of metagenomics yields taxa which are absent in most samples. It's important to take these differences into account when designing studies, when choosing omics levels and in the multi-omics integration.

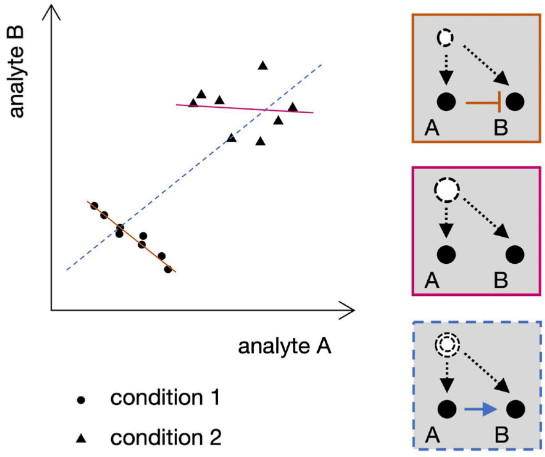


Figure 4. The interpretation of correlations between analytes depends fully on the relationship between the samples. Here, analytes A and B are limited by a third (white circle in the boxes, e.g. a precursor or nutrient source) which has one fixed level under condition 1 and another under condition 2. The correlation between analytes A and B over all samples of both conditions simultaneously is positive (blue box) and could suggest positive feedback from A to B, while focusing on specific conditions, negative correlation by inhibition (orange) or no correlation (red) should be concluded.

Applications

After a few pioneering multi-omics studies of human-associated and soil microbiomes, multi-omics investigations of microbiomes have become more frequent in the last 5 years – contributing to human, animal, plant, environmental, and biotechnological research.

Sequence- and genome-centric integration

As mentioned earlier, metagenomics, metatranscriptomics and metaproteomics lend themselves to integration due to the sequence identity: metatranscriptomics reads can be mapped onto assembled metagenomes or can be co-assembled with metagenomic reads (Figure 3). This increases the detectability of highly expressed genes in low-abundant taxa. Proteins must be identified based on protein or peptide databases, and it has been demonstrated that this process is aided by the use of metagenomic information from the same sample. Integration has been used to determine the correlation between transcript and protein abundance, the level of variation of the different omics levels and hence the potential to find mechanistically important players in either data set.

An important concept in integrated meta-omics is genome-centric analyses: genomes are reconstructed from metagenomics data and the other omes are mapped to them. The genome, therefore, gives context to the observations (e.g., other genetic functions in the same genome, abundances in different samples). Based

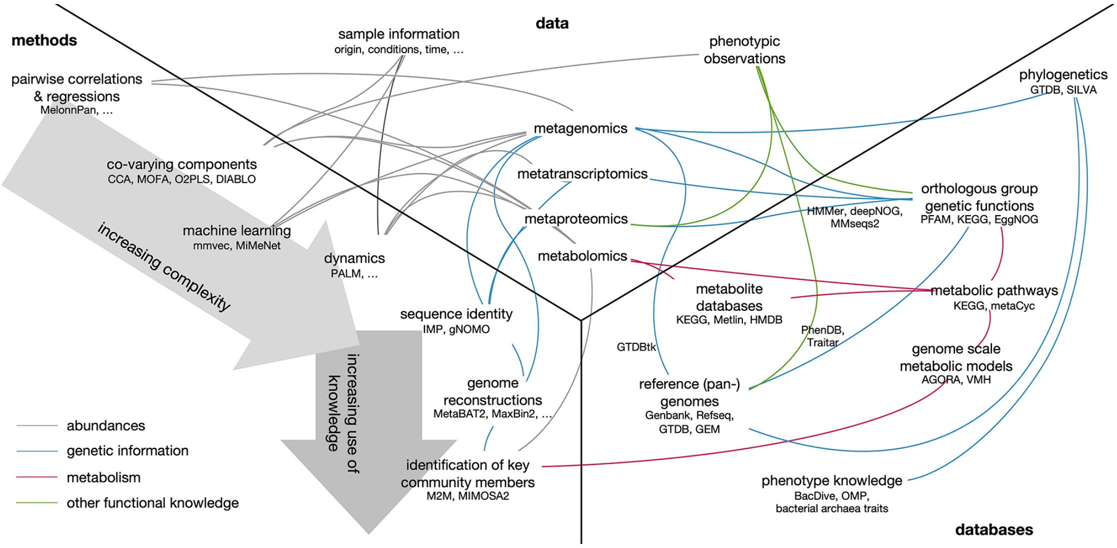


Figure 5. Meta-omics data with examples of integration methods. The methods make use of numerical, sequence or metabolism information and to different extents of existing databases. Abbreviations/method names: AGORA – assembly of gut organisms through reconstruction and analysis, BacDive – bacterial diversity metadatabase, CCA – canonical correlation analysis, deepNOG – deep network architecture to assign EggNOG5 orthologous groups, DIABLO – data integration analysis for biomarker discovery using latent variable approaches for omics studies, EggNOG – database of orthology relationships, functional annotation and gene evolutionary histories, gNOMO – multi-meta-omics pipeline for non-model organisms, GTDB – genome taxonomy database, GTDBtk – GTDB toolkit, GEM – genome scale metabolic models, HMDB – human metabolome database, HMMer – methods using profile hidden Markov models, IMP – integrated meta-omic pipeline, KEGG – Kyoto Encyclopedia of Genes and Genomes, MaxBin2 – automatic tool for binning metagenomics sequences, MetaBAT2 – metagenome binning with abundance and tetra-nucleotide frequencies, metaCyc – highly curated, non-redundant reference database of small-molecule metabolism, Metlin – metabolite and chemical entity database, MiMeNet – Microbiome-metabolome network, MIMOSA2 – model-based integration of metabolite observations and species abundances, mmvec – microbe-metabolite vectors, MOFA – multi-omics factor analysis, M2M – Metage2Metabo (a software system for the characterization of metabolic complementarity starting from annotated individual genome), OMP – ontology of microbial phenotypes, O2PLS – two-way orthogonal PLS (partial least squares), PALM – pipeline for the analysis of longitudinal multi-omics data, PhenDB – prediction of bacterial phenotypes, SILVA – automatic software pipeline for sequence retrieval, quality assignment and alignment of ribosomal RNA genes, Traitar – microbial trait analyzer, VMH – virtual metabolic human, MMseqs2 – many-against-many sequence searching.

on phylogenetic relationships, genomes can be linked to prior phenotypic or biochemical knowledge. Examples of applications include functional analyses of human gut, soil, ground- and wastewater microbiomes. An advantage of these methods is that they are applicable to both described and completely unknown organisms. Genome-centric approaches to metabolomics integration based on reference strain metabolite profiles and co-cultures are currently being developed.

Data fusion

Probably the most common combination of omes are metagenomics and metabolomics: it has been studied in most microbiomes, from various human niches, over model and non-model terrestrial and marine animals, to plant rhizospheres, to soil and to biotechnological mixed communities. However, this integration is challenging: there is, of course, no sequence-identity to rely on. Due to the absence of biologically meaningful links, applicable methods are called data 'fusion methods', as opposed to integration methods. In the simplest case, fusion is attempted by pair-wise correlations of, e.g., all microbial taxa with all metabolites, where correlations that are above a certain threshold are represented in a 'correlation network graph'. However, in this example, it is likely that many real processes are not observed except for cases where metabolites can only be produced and metabolized by single taxa or by sets of highly correlating taxa. More advanced methods estimate multivariate correlations between data sets by calculating linear combinations (components) of the analytes in one data set that co-vary highly with linear combinations of the other data set(s) (see Figure 2b). These covarying components are said to describe the 'common' or 'joint' information. Most of these methods assume analyte levels to be symmetrically distributed, but new methods are being developed that take the zero-inflated structure of microbiome data into account. Note that for interpreting correlations, it is of extreme importance to consider over which samples the correlation is calculated as they can change due to changing experimental conditions (Figure 4).

All data fusion methods have in common that biological interpretation is done afterwards, which is a problematic situation: the omics way of working that revolutionized biological research suffers from the curse of dimensionality, as more and more analytes are measured in an untargeted approach for a small set of samples. To model such data in a meaningful manner, thousands of samples would be needed to find the relevant analytes between the noise. If studies have a limited number of samples, it is of essential importance that the biological function of each feature is known and used to link analytes within and between datasets.

Integration with prior knowledge

Knowledge-based analyses of metabolic networks, which represent both sequence-based omics and metabolomics, have been applied in soil, wastewater treatment and human microbiomes. Here, the omics levels are summarized as functions of the whole community. Because they are additionally linked by known metabolic pathways, these approaches are successful in providing insights into metabolic pathways that respond to changing conditions (e.g., drought stress, temperature, or nutrition).

The microbial metagenome has become an omics level that is included in multi-omics studies of complex organisms such as plants, e.g., *Brassica rapa*, mice, cows and humans – especially in the context of metabolic and inflammatory diseases. Systematic references for mechanistic links of host and microbiome are not yet well developed, especially outside of human physiology. Hence, the associations between microbiome taxa or functions and host gene expression, epigenetics, metabolism and phenotype must necessarily be established by data fusion methods. Another trend in host-focused studies is to attempt classification of individuals (e.g., as a diagnostic tool for colorectal carcinomas) based on multi-omics biomarker panels, where supervised data fusion approaches are applied.

Outlook

Integration strategies adapt to and are facilitated by technological advances: for instance, recent research in a biogas reactor community has demonstrated several ways of how quantitative measurements at multiple meta-omics levels provide better functional explanations of community phenotype. High-quality multi-species metabolic models, methods for metagenomics-based construction of metabolic models and the integration of multi-omics measurements into such models are important research fields. Measurements and integration of the spatial structure of microbial communities will play a bigger role in the future. Large, openly accessible multi-omics data sets, databases with genetic and metabolite information and data standards (Figure 5) are developed, maintained and grown thanks to individual and community efforts. A key challenge for meta-omics integration will be the development of methods that combine the approaches described earlier to make sound use of data and meaningful knowledge – and to use the gained information to develop new research questions. ■

Further reading

Readers who are interested in knowing more are referred to recent reviews which cover multi-omics integration from different points of view:

- Sequence-based omics: Heintz-Buschart A, Wilmes P. Human Gut Microbiome: Function Matters. *Trends Microbiol*. 2018 26 (7): 563–574. doi: 10.1016/j.tim.2017.11.002.
- Metabolomics: Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK, Dorrestein PC. Mass spectrometry-based metabolomics in microbiome investigations. *Nat Rev Microbiol*. 2021 doi: 10.1038/s41579-021-00621-9.
- Multi-omics networks and beyond: Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, Jiang Y. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Front Genet*. 2019 10: 995. doi: 10.3389/fgene.2019.00995.
- Metabolic modelling: Colarusso A, Goodchild-Michelman I, Rayle M, Zomorrodi AR. Computational modeling of metabolism in microbial communities on a genome-scale. *Curr Opin Syst Biol*. 2021 26: 46–57. doi: 10.1016/j.coisb.2021.04.001.

Specific methodological questions are addressed in the current literature:

(Continued)

Further reading (*Continued*)

- Multi-omics power calculation: Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, Tegnér J, Westerhuis JA, Conesa A. Harmonization of quality metrics and power calculation in multi-omic studies. *Nat Commun*. 2020 11 (1): 3092. doi: 10.1038/s41467-020-16937-8.
- Quantitative multi-omics: Delogu F, Kunath BJ, Evans PN, Arntzen MØ, Hvidsten TR, Pope PB. Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nat Commun*. 2020 11 (1): 4708. doi: 10.1038/s41467-020-18543-0.

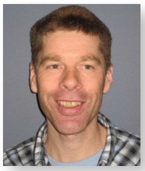
Data analysis packages may be found here:

- Multiple omics analytics: http://mixomics.org/
- General data fusion methods: https://cran.r-project.org/web/packages/multiblock/
- R.JIVE: O'Connell MJ, Lock EF. R.JIVE for exploration of multi-source molecular data. *Bioinformatics*. 2016 32 (18): 2877-9. doi: 10.1093/bioinformatics/btw324. https://cran.r-project.org/web/packages/r.jive/

Example applications of multi-omics analyses can be found here:

- A pioneering study in permafrost soils: Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, Shah MB, VerBerkmoes NC, Lee LH, Mavrommatis K, Jansson JK. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*. 2015 521 (7551): 208–212. doi: 10.1038/nature14238.
- Time-resolved multi-omics in wastewater treatment: Herold M, Martínez Arbas S, Narayanasamy S, Sheik AR, Kleine-Borgmann LAK, Lebrun LA, Kunath BJ, Roume H, Bessarab I, Williams RBH, Gillece JD, Schupp JM, Keim PS, Jäger C, Hoopmann MR, Moritz RL, Ye Y, Li S, Tang H, Heintz-Buschart A, May P, Muller EEL, Laczny CC, Wilmes P. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat Commun*. 2020 11 (1): 5281. doi: 10.1038/s41467-020-19006-2.
- Human gut microbiome multi-omics in colorectal cancer: Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, Hosoda F, Rokutan H, Matsumoto M, Takamaru H, Yamada M, Matsuda T, Iwasaki M, Yamaji T, Yachida T, Soga T, Kurokawa K, Toyoda A, Ogura Y, Hayashi T, Hatakeyama M, Nakagama H, Saito Y, Fukuda S, Shibata T, Yamada T. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med*. 2019 25 (6): 968–976. doi: 10.1038/s41591-019-0458-7.
- Hertel J, Heinken A, Martinelli F, Thiele I. Integration of constraint-based modeling with fecal metabolomics reveals large deleterious effects of Fusobacterium spp. on community butyrate production. *Gut Microbes*. 2021 13: 1. doi: 10.1080/19490976.2021.1915673.

Anna Heintz-Buschart is an assistant professor for Microbial Metagenomics at the Swammerdam Institute for Life Sciences. She earned a PhD in the wet-lab on molecular microbiology and she still believes that our world belongs to the microbes. Because of that, she develops bioinformatics methods to analyse large-scale meta-omics data and integrate multiple omics levels, to facilitate biological interpretations and predictions. She has not quite decided what her favourite microbe-related system is: the human microbiome, biotechnology, soil ecology or biodiversity research. Email: a.u.s.heintzbuschart@uva.nl. twitter handle: @_a_h_b_ Email: a.u.s.heintzbuschart@uva.nl



Johan A. Westerhuis is an assistant professor for Biosystems Data Analysis at the Swammerdam Institute for Life Sciences. He obtained his PhD at the University Centre for Pharmacy at the University of Groningen on 'Multivariate statistical modeling of the pharmaceutical process of wet granulation and tableting' using multiblock (path-) models. After a postdoc at McMaster University, Hamilton, ON, on batch process monitoring and multiblock methods, he joined the Biosystems Data Analysis group at the Universiteit van Amsterdam. He teaches statistics and biochemical data analysis at Bachelor's and Master's levels and supervises PhD students and postdocs in metabolomics and microbiome data analysis. Email: j.a.westerhuis@uva.nl Email: j.a.westerhuis@uva.nl