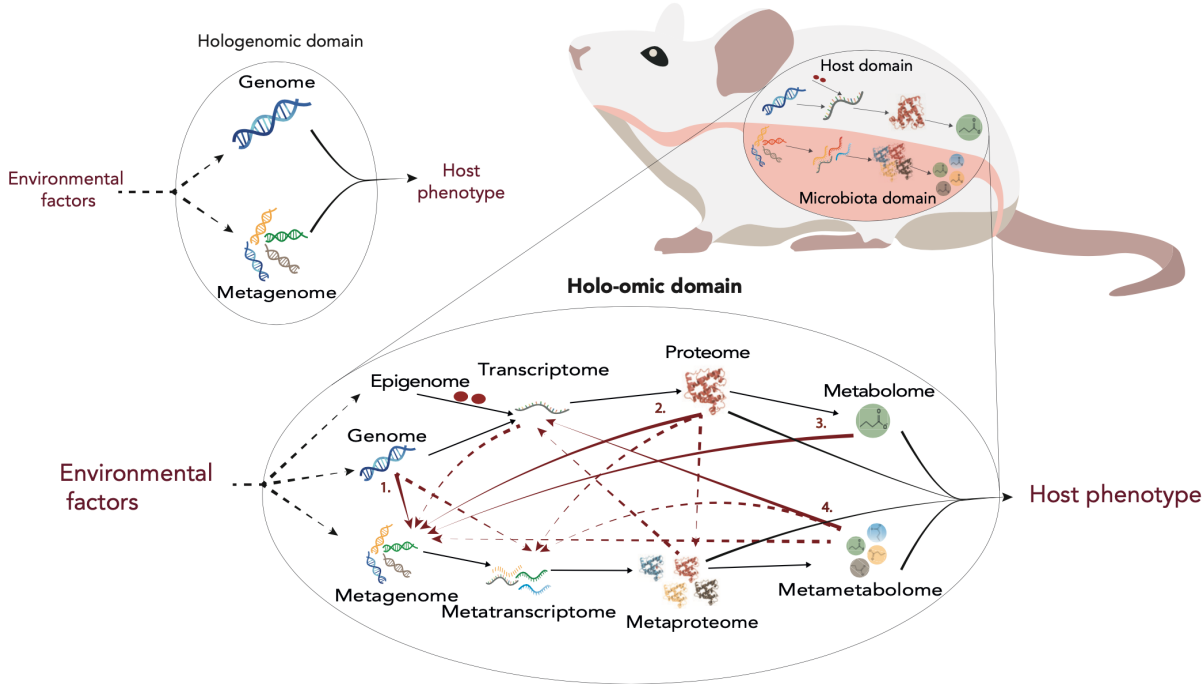
A Practical Guide to Holo-Omics

Contents

About this guidebook

The practical guide to holo-omics is under construction, and its contents are still preliminary. We expect to have an initial complete version by May 2023.



Holo-omics overview. Modified from Nyholm et al. 2020 [1]

The **practical guide to holo-omics** is a compilation of methodological procedures to generate, analyse and integrate holo-omic data, i.e., multi-omic data jointly generated from hosts and associated microbial communities [1, 2]. This guide extends the contents of the article **"A practical introduction to holo-omics"**, which aims at guiding researchers to the main critical steps and decision points to perform holo-omic studies. While the article focuses on discussing pros and cons of using multiple available options, the aim of this guide is to compile protocols and pipelines to be implemented by researchers. The **practical guide to holo-omics** is presented in two formats:

- Website (http://www.holo-omics.science/)
- PDF document (http://www.holo-omics.science/holo_omics_workbook.pdf)

This guide is presented as a final output of the H2020 project HoloFood. More information about this EU Innovation Action that ran between 2019 and 2023 can be found in the HoloFood Website and the CORDIS website.

Contents

- **Introduction**: general information about holo-omics, employed data types and study design considerations.
- **Laboratory procedures**: methods and procedures for generating raw omic data of hosts and microbial communities.
- **Bioinformatic procedures**: methods and procedures for processing raw omic data into quantitative datasets to be analysed through statistics.
- **Statistical procedures**: methods and procedures for analysing and integrating holo-omic data.

Protocols, exercises and tutorials

This guide contains example data and bits of code (mostly shell and R) to reproduce data generation and analysis procedures. Code boxes look like the following:

```
shao4d_perm <- shao4d %>%
  tax_transform("identity", rank = "genus") %>%
  dist_calc("aitchison") %>%
  dist_permanova(
    variables = c("birth_mode", "sex", "number_reads"),
    n_perms = 99, # you should use more permutations in your real analyses!
    n_processes = 1
  )
#> Dropping samples with missings: 15
#> 2022-11-24 01:15:20 - Starting PERMANOVA with 99 perms with 1 processes
#> 2022-11-24 01:15:21 - Finished PERMANOVA

shao4d_perm %>% perm_get()
#> Permutation test for adonis under reduced model
#> Marginal effects of terms
#> Permutation: free
#> Number of permutations: 99
#>
#> vegan::adonis2(formula = formula, data = metadata, permutations = n_perms, by = by, parallel = paral
#>               Df SumOfSqs      R2       F Pr(>F)
#> birth_mode     1    10462 0.09055 29.3778   0.01 **
#> sex            1      402 0.00348  1.1296   0.31
#> number_reads   1     1117 0.00967  3.1364   0.01 **
#> Residual     287   102209 0.88462
#> Total        290   115540 1.00000
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Authors

Antton Alberdi Antton is an Associate Professor at the University of Copenhagen whose research is focused on understanding how animal-microbiota interactions shape basic and applied biological processes. Antton is the corresponding author of the **Practical Guide to Holo-omics**.

antton.alberdi@sund.ku.dk | www.alberdilab.dk

- Morten T Limborg, University of Copenhagen
- Iñaki Odriozola, University of Copenhagen
- Jacob A Rasmussen, University of Copenhagen

Protocol and script contributors

- Carlotta Pietroni, University of Copenhagen
- Raphael Eisenhofer, University of Copenhagen
- Jorge Langa, University of Copenhagen

Other relevant people

- Tom Gilbert (HoloFood project coordinator), University of Copenhagen
- Anna Fotaki (HoloFood project manager), University of Copenhagen

How to cite this work

Instructions to cite this work will be eventually added.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817729.

Part I

INTRODUCTION

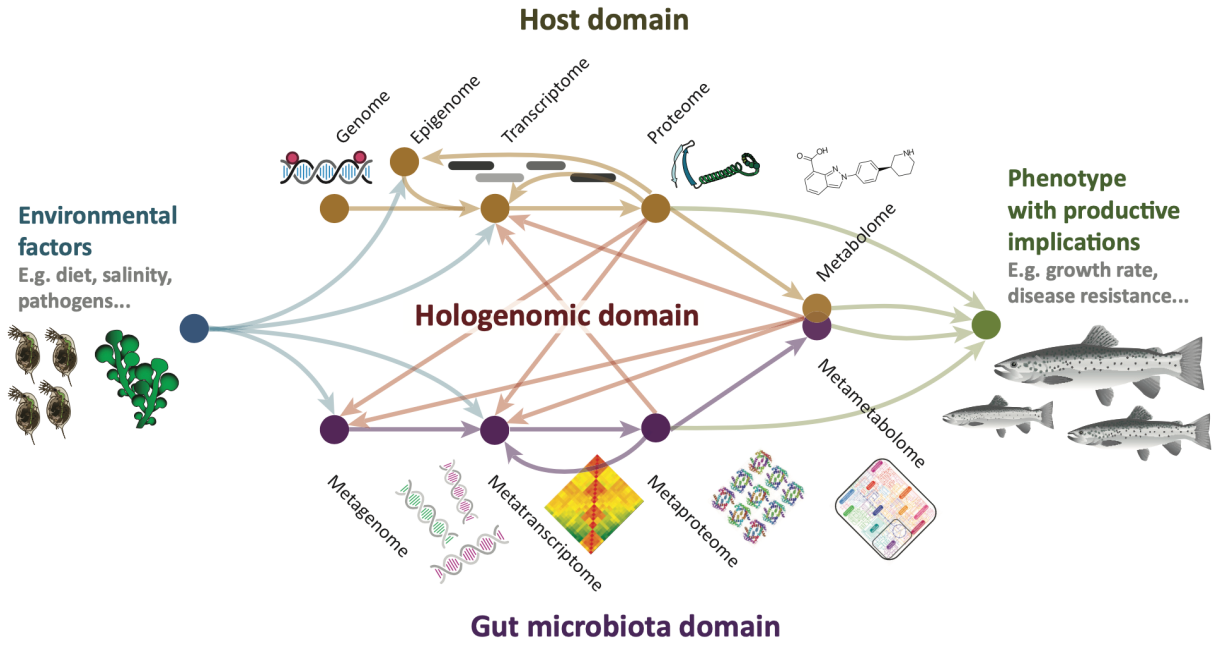
1 Introduction to holo-omics

Holo-omics refers to the methodological approach that jointly generates and analyses multi-omic data from hosts and associated microbial communities [1]. The holo-omic approach to host-microbiota interactions relies on three major assumptions:

1. Host-associated microorganisms interact not only with each other but also with their host [3].
2. These interactions affect, either positively or negatively, central biological processes of hosts and microorganisms [4].
3. The interplay can be traced using biomolecular tools.

1.1 Omic layers

Nucleic acid sequencing and mass spectrometry technologies that enable tracking the biomolecular pathways linking host and microbial genomic sequences with biomolecular phenotypes by generating (meta)transcriptomes, (meta) proteomes, and (meta)metabolomes. The same technologies also enable epigenomic and exposomic profiling, which can further contribute to disentangling the biochemical associations between host-microbiota-environment interactions and their effect on host phenotypes.



Overview of omic layers. Modified from Limborg et al. 2018 [2].

In this workbook we consider seven omic layers that require specific data generation and analysis strategies before integrating them into multi-omic statistical models:

- Nucleic acid sequencing-based
  - Host genomics - **HG**
  - Host transcriptomics - **HT**
  - Microbial metagenomics - **MG**
  - Microbial metatranscriptomics - **MT**
- Mass spectrometry-based
  - Host proteomics - **HP**
  - Microbial metaproteomics - **MP**
  - (Meta)metabolomics - **ME**

Acknowledging the distinct biological and structural characteristics of these seven omic layers is essential to design experiments and analytical pipelines for better solving the complex puzzle of host-microbiota interactions.

1.2 Host genomics (HG)

Contents to be added

1.3 Host transcriptomics (HT)

Contents to be added

1.4 Microbial metagenomics (MG)

Contents to be added

1.5 Microbial metatranscriptomics (MT)

Contents to be added

1.6 Host proteomics (HP)

Contents to be added

1.7 Microbial metaproteomics (MP)

Contents to be added

1.8 (Meta)metabolomics (ME)

Contents to be added

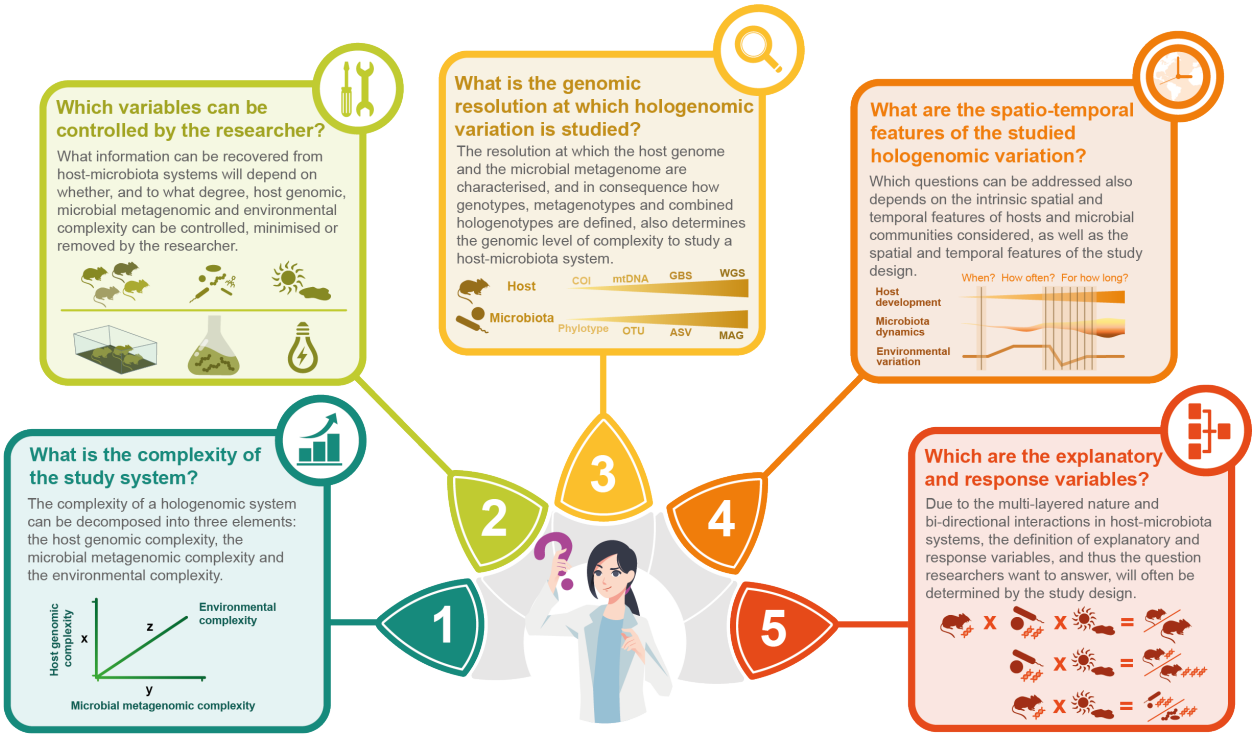
2 Study design considerations

The contents of this section have been extracted and modified from the article Disentangling host–microbiota complexity through hologenomics published in ***Nature Reviews Genetics*** in 2022 by the authors of the **Holo-omics Workbook**.

Holo-omic approaches can be used to understand how the combined features of hosts and microorganisms shape biological processes relevant for hosts (such as adaptation), for microorganisms (such as meta-community dynamics) or both [5].

Depending on the aims and features of the study system, holo-omics can be implemented using different study designs, model systems and techniques. This landscape of possibilities is shaped around five essential questions that need to be considered when designing and interpreting hologenomic studies, which relate to five core topics:

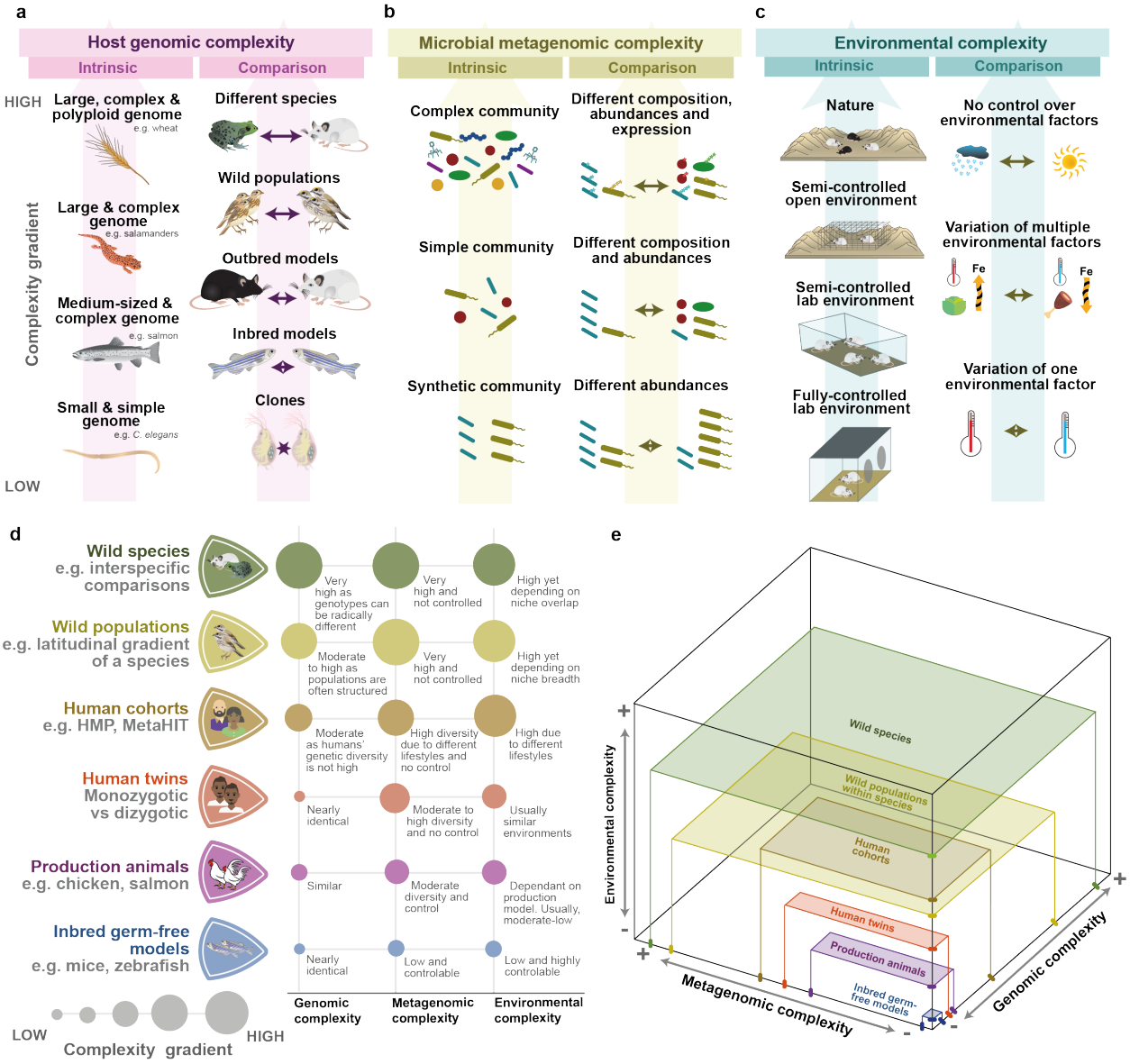
1. **Hologenomic complexity**
2. **Control of variables**
3. **Molecular resolution**
4. **Spatiotemporal factors**
5. **Explanatory and response variables**



2.1 Hologenomic complexity

The contents of this section have been extracted and modified from the article Disentangling host–microbiota complexity through hologenomics published in ***Nature Reviews Genetics*** in 2022 by the authors of the **Holo-omics Workbook**.

Hologenomic complexity can be broadly defined as the amount of information relevant to the study that the biological system under analysis contains and it can be decomposed into three major elements: host genomic, microbial metagenomic and environmental complexity [5]. Within each of these elements, two sources of complexity can be defined: the intrinsic complexity of the system under study, including host genome size and number of bacterial genomes, and the complexity introduced by the degree of difference between the organisms under comparison such as gene expression differences versus distinct genomes.



Decomposition of hologenomic complexity. (a-c) The design and interpretation of hologenomic studies depend on the host genomic (part a), microbial metagenomic (part b) and environmental (part c) complexity of the system under study. Within each axis of complexity, two types of gradients can be defined based on whether the features are intrinsic to the system or introduced by the researcher through the selection of groups under comparison. **(d)** Six examples of study systems with different levels of genomic, metagenomic

and environmental complexity. **(e)** Three-dimensional representation of the complexity of the examples. The area of the plain represents the combined host genomic and microbial metagenomic complexity of the system, while the height represents the environmental complexity. The combined three-dimensional volume represents the overall hologenomic complexity of the system. HMP: Human Microbiome Project.

2.2 Control of variables

The contents of this section have been extracted and modified from the article Disentangling host–microbiota complexity through hologenomics published in ***Nature Reviews Genetics*** in 2022 by the authors of the **Holo-omics Workbook**.

Controlling the complexity of hologenomic variables is essential for addressing specific research questions. Broadly speaking, the more detailed and mechanistic the question under study, the greater the required control. For instance, research on specific biomolecular processes using laboratory models will require a higher level of control than studying biogeographical patterns of host–microbiota interactions in wild organisms. The control of hologenomic variables can be achieved through a number of strategies.

2.2.1 Controlling host genomes

The control over host genomic complexity largely depends on the model organisms studied and the technical approaches employed. In laboratory organisms that can reproduce asexually, such as water fleas (Daphnia, Crustacea) and Lamiaceae plants, absolute control over host genotypes can be achieved by using clonal organisms [6]. When clones cannot be used, inbred laboratory animals can provide a high level of genomic homogeneity. The use of groups of genetically homogeneous hosts allows the effects of contrasting environmental conditions or specific microbial comunities to be compared. Clonal and inbred models also enable the effects of a specific host genetic factor to be studied in a controlled genomic background through the application of targeted techniques for modulating gene expression (such as RNA-mediated interference) or for genomic engineering (such as CRISPR–Cas9). Working with humans and wild organisms does not enable such a degree of control over the genotypes studied unless in vitro models, such as organ-on-a-chip co-cultures of animal tissues and microbial communities, are generated62. When this level of control is not possible, coarse control over host genotypes can be achieved through contrasting animals from different populations or from closely related species63, while greater control can be achieved through comparing individuals across different degrees of kinship, such as monozygotic versus dizygotic twins38, and family members to other individuals64.

2.2.2 Controlling microbial metagenomes

Control over microbial metagenomic complexity is usually achieved through modulating microbial communities. Some strategies, such as modification of dietary regimes or the administration of microbiota-targeted additives or prebiotics, aim to modify microbial ecosystems by changing nutrient availability. However, unless compounds that match unique enzymatic capabilities of specific microorganisms are used, it is difficult to accurately modulate the microbiota owing to the complexity of ecological relationships among microorganisms. Alternative approaches to modify microbial communities include inoculation of target bacteria (such as probiotics) and faecal microbiota transplantation. The efficacy and accuracy of these methods is also variable; there is no guarantee that inoculated bacteria will establish or modulate the microbiota, while transplantation does not enable accurate control over the microbial community introduced or the secondary elements that are transplanted along with bacteria. These issues complicate interpretation of results; for example, bacteriophages transferred alongside bacteria may severely impact the gut microbiota composition. A higher level of control could potentially be achieved through transplanting synthetic microbial communities. While this approach has been successfully implemented in diverse in vitro setups the complexity of microbial communities still hinders its efficient use as a routine scientific procedure in live animals.

2.2.3 Controlling the environment

In most laboratory studies, environmental complexity is reduced so that no, or very few, environmental parameters (usually only experimental treatments) vary among groups and subjects. Climate chambers and aquaria enable experiments by providing absolute control of abiotic conditions, such as light/dark cycles,

humidity and temperature variations. Outdoor common garden experiments do not provide full control over environmental factors, but they ensure the effect on the systems being compared is identical. Some natural systems can also provide special conditions that enable environmental features to be controlled, such as cuckoo nestlings that are bred by other birds or salmon populations that breed in the same rivers in alternating years. Research on wild organisms usually incorporates more complex and dynamic environmental conditions. When controlling them is not possible, collection of relevant environmental metadata to be incorporated as covariates in the statistical analyses is useful. A century of ecological research has revealed the advantages of each of these approaches. On the one extreme, laboratory microcosms allow the most reductive control. On the other extreme, studies in the macrocosm of the real world provide perspective on emergent properties of natural ecosystems that cannot be anticipated solely based on microcosms.

2.3 Molecular resolution

The contents of this section have been extracted and modified from the article Disentangling host–microbiota complexity through hologenomics published in ***Nature Reviews Genetics*** in 2022 by the authors of the **Holo-omics Workbook**.

The complexity of a study system is not only determined by its inherent properties and study design, but also the techniques and procedures employed to analyse it. Researchers can decide how much a system is simplified by altering the resolution of the hologenomic features under study; in essence, zooming in or zooming out.

2.3.1 Resolution of host genotypes

In host-microbiota studies, host genotypes can be defined at different levels, including species, breeds, populations, strains, sex or individuals. Genotypes can be defined as categorical variables, without analysing the differences between them, or can be studied in more detail through considering their actual genetic content and establishing correlations among them. When using an evolutionary perspective, phylogenetic relationships between genotypes are established based on phylogenomic markers, which usually vary above population and species level, but not among individuals. This implies that genomic variability among the individuals included within each genotype is overlooked. Studying the effect of interindividual genomic variability on host-microbiota systems, such as identifying candidate host genomic variants associated with microbial features, requires a higher level of resolution. This is achieved through defining genotypes at the individual level, and using techniques based on whole genome resequencing that enable the complexity of host genomes to be screened at a much finer level, so that differences between the individuals contrasted are not only defined based on their kinship, but also the functional properties of their genomic variants. Currently, this approach requires high quality reference genomes from which high density SNP profiles of individuals can be generated, for example through SNPchip or resequencing studies. The genomic resolution could be further refined by incorporating structural variants, methylation patterns, or even, we hypothesise, chromosome 3D folding structure as revealed through techniques such as Hi-C. In doing so, researchers can identify associations between SNPs or gene variants and specific microbiota traits, such as the relative abundance of certain taxa or the enrichment of a given function, and thus identify mechanisms by which a host exerts control over composition and function of its associated microbiota

Resolution of microbial metagenotypes

The structure and resolution at which microbial metagenotypes are defined also affects the complexity of the metagenome under analysis. Metagenotypes can be defined as arrays of microbial taxa, microbial genes or a combination of both. The most common approach to define them is to rely on short marker sequences targeted for metabarcoding purposes, such as the 16S rRNA or the internal transcribed spacer (ITS). However, these procedures often do not enable reliable taxonomic assignment at genus or species level, do not capture strain level community dynamics, and are prone to generate biased functional inferences, as bacteria with identical marker genes (particularly those associated with wild taxa) might carry very different catalogues of genes. Thus, while useful for estimating microbial diversity and obtaining preliminary insights into functionality, targeted sequencing approaches do not provide conclusive evidence about the metabolic capabilities of the microbiota, particularly when working with non-human systems.

By contrast, if appropriate strategies and adequate sequencing depths are employed, shotgun metagenomics enables bacterial genome sequences to be recovered, from which genes can be predicted and annotated to create a gene catalogue that can define a metagenotype. However, these genes are not randomly distributed, but enclosed within genomes of specific bacteria or other microorganisms, with a particular combination of genes that shape their expression and the specific biological features (such as oxygen affinity, reproduction time, metabolic capacity) that determine their ecology. Hence, a more refined characterisation of microbial metagenotypes can be achieved through binning algorithms that enable bacterial genome reconstruction from metagenomic mixtures, yielding metagenome-assembled genomes (MAGs). Nevertheless, unless short-read sequencing is combined with long-read approaches, it is challenging to capture multi-copy genes such as the 16S rRNA marker gene 103, which is often employed in metabarcoding studies and therefore represents a useful link to a large number of existing studies. Machine learning-based solutions to link 16S rRNA marker gene sequences with MAGs are, however, being developed 104. Finally, regardless of the approach used to define the microbial metagenotype, the complexity of microbial communities will often require dimensionality reduction to increase statistical power 105,106. This can be achieved by defining co-abundance clusters, ecological guilds or more complex strategies that also consider temporal features of microbiota variation, such as compositional tensor factorisation.

Resolution of envirotypes

Characterisation of environmental factors that affect the host-microbiota system under study enable the definition of envirotypes, a term drawn from crop sciences that is useful for accounting for the environmental factors in the hologenomic context. Any different physical place, or place sampled at different time points, will be exposed to a different environment, as conditions will seldom be identical between two spatial and temporal points. Hence, the resolution at which the composite of environmental factors is considered will define whether these two environments will be considered different envirotypes or not. For example, if only considering water temperature, killer whales sampled in the Arctic and the Antarctic seas experience the same envirotype. However, if the biotic composition is also considered in the definition of the environment, the Arctic and the Antarctic will need to be split into two distinct envirotypes, as some killer whales will have access to penguins while others will not. The same principle applies to laboratory setups or mesocosm experiments: a temperature shift of 2-3 ^{o}C might not be considered relevant under some experimental setups, while it can define different envirotypes under other study designs. Finally, failure to recognise environmental factors that affect host-microbiota interactions, and thus define relevant envirotypes, can lead to increased noise and decreased capacity to achieve statistical significance.

2.4 Spatiotemporal factors

The contents of this section have been extracted and modified from the article Disentangling host–microbiota complexity through hologenomics published in ***Nature Reviews Genetics*** in 2022 by the authors of the **Holo-omics Workbook**.

Spatial factors

Spatial resolution. Microbial communities associated with animal and plant hosts vary not only across coarse body parts, but also at the micro-scale, such as between the lumen and the intestinal crypts. Thus, the resolution at which a body site is defined will also determine how a hologenomic system is characterised. For example, the animal gastrointestinal tract can be considered a single sampling unit, 4-5 units or hundreds of micro-units, depending on the sampling and data processing strategies employed. Naturally, each level of resolution will allow different questions to be addressed and will require the use of different technologies and analytical approaches.

Temporal factors

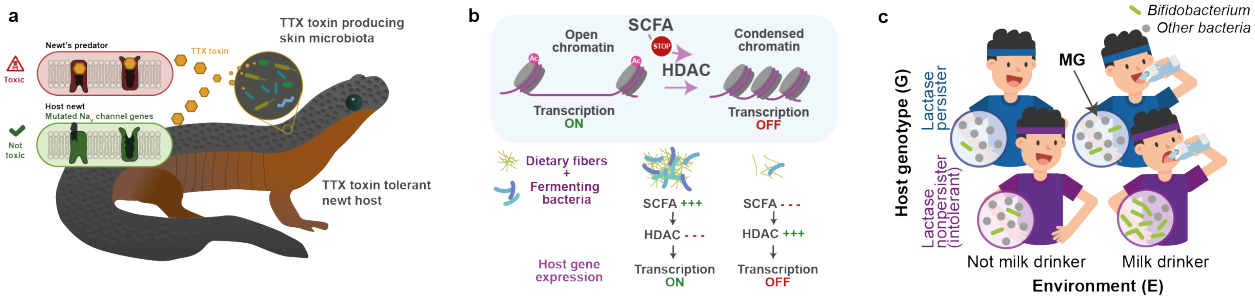
Temporal features to be considered include when, how often, and for how long host-microbiota systems are to be analysed. When a host is first exposed to microbes with regard to temporal benchmarks (number of days or years) must be considered, as should the order in which it is exposed to them. Priority effects relate to how

the order of species arrivals in an ecosystem shape the potential for subsequently arriving taxa to establish themselves. Although originally discussed at the macroorganismal level in the context of plant communities, the phenomenon is also relevant for building host-associated microorganism communities, for example as documented in the human gut. In addition, microbial communities are known to vary daily, seasonally and relative to life-stage patterns. Hence, the extent and frequency of sampling determine which of these dynamics will be observed or, conversely, missed. Finally, it is important to consider that the consequences of changes at one time period or life stage may appear only later in time, thus detection of such effects obviously requires that the subsequent period is also studied. For example, interventional animal experiments show that when the immune system develops early in life, there is a window of opportunity where the gut microbiota composition shapes the risk of developing diseases in the future.

2.5 Explanatory and response variables

The contents of this section have been extracted and modified from the article Disentangling host–microbiota complexity through hologenomics published in ***Nature Reviews Genetics*** in 2022 by the authors of the **Holo-omics Workbook**.

Host genomic and microbial metagenomic data generated under hologenomic setups can take on different roles when generating statistical models. While the environment is most often considered as an explanatory variable (though one can also study how the hologenome affects the environment), the host genome and the microbial metagenome are sometimes viewed as explanatory and sometimes as response variables, depending on the aim of the research. In many cases, directionality is set by the researcher rather than the biological system itself, as host-microbiota systems contain many bi-directional interactions and circular processes, which complicate the establishment of causal relationships. Here, we define three basic models in which the three main variables (genome, metagenome and environment) are assigned different roles to address different types of fundamental questions.



P = G x MG x E
Phenotype: toxin-tolerant toxic newt
Genotype: toxin-tolerant NaV channel genes
Metagenotype: toxin-producing skin bacteria
Environment: predators

G = MG x E
Genotype: immune responsive genome
Metagenotype: fermenting bacteria
Environment: dietary fibers

MG = G x E
Metagenotype: *Bifidobacterium*-enriched
Genotype: lactase persister / nonpersister
Environment: milk consumption

Examples of biological processes addressed by the different models of host-microbiota interactions. **a)** How does the hologenome shape animal phenotypes? Only the combination of specific host genomic (G) and microbial metagenomic (MG) features, probably developed due to a selective force exerted by the presence of predators (E) enables rough-skinned newts to have skin toxicity, an ecologically relevant phenotypic trait (P). **b)** How do the microbial metagenome and environment shape host genomic features? SCFA-producing bacteria along with a fibre-rich diet enhance chromatin accessibility and thus activate immune gene expression. **c)** How do the host genome and the environment shape microbial genomic features? Only the combination of a lactase nonpersister genotype combined with the milk-drinking envirotype generates a microbial metagenotype characterised by enrichment of Bifidobacterium.

Phenotype as a product of genotype, metagenotype and envirotype

This is the main model used when hologenomics is conducted to ascertain how genome-metagenome-environment interactions affect the biological properties of a host, such as disease susceptibility, performance or fitness. It is an especially common and relevant model for health, agricultural, and ecological and evolutionary

research 19,125–127. One clear example of a phenotype shaped by host genomic, microbial metagenomic and environmental factors was recently reported for rough-skinned newts. The study showed that bacteria on the skin of the newts produce a deadly neurotoxin from which the newt is protected by mutations in five host genes that encode the NaV channels normally targeted by the toxin. Thus, this 'toxic newt' phenotype is the result of both host and microbial genes, which likely evolved under the pressure exerted by an environmental factor, namely the presence of predators.

Genotype expression influenced by metagenotype and envirotype

When studying how core host genomic features, which contribute to shaping phenotypes, are affected by the microbiota, host genomic features become the response variable. Unlike the microbial metagenome, the genome sequence of the host organism is not variable, but microorganisms can induce chromatin remodelling and DNA methylation, and thus modulate the bioactivity of molecular receptors and host gene expression. A well-studied pathway that links the microbiota with host gene expression involves modulation of the activity of host histone deacetylases (HDAC) by short chain fatty acids (SCFA) produced by intestinal microorganisms. HDACs remove histone lysine acetyl groups, which leads to chromatin condensation and transcriptional silencing of genes. Increased SCFA concentrations inhibit histone deacetylases, thereby enhancing chromatin accessibility and activating gene expression. A metagenotype with a higher capacity to produce SCFAs combined with an envirotype characterised as a fibre-rich diet (required to produce SCFAs), therefore contributes to boost immune response through activating host immune gene expression.

Metagenotype as a product of genotype and envirotype

This model assumes the inverse causal directionality between the host genome and microbial metagenome to that described above. Candidate host genes related to microbiota features can be identified through GWAS in which the metagenotype (or derived metrics such as diversity or abundance of specific microbial taxa, genes or metabolic functions) are treated as a phenotypic trait. For instance, the increased abundance of lactose degrader Bifidobacteria in humans has been shown to be associated with lactase nonpersister genotype and consumption of milk (envirotype). Once candidate genes are known, targeted analyses in which natural or human-controlled genomic variability (such as the number of copies of the amylase-encoding gene in humans) can be contrasted under controlled environmental conditions to ascertain the effect on metagenotypes (such as the abundance of Ruminococcaceae bacteria in the gut microbiota).

Part II

LABORATORY PROCEDURES

3 About labwork

General considerations

Although the generation of each omic data layer requires dedicated protocols to be implemented, there are multiple general considerations that apply to all laboratory processes. In the following we list three of the most relevant ones.

External contamination The risk of external contamination is a relevant issue that must be actively tackled when generating multi-omic data. External contamination refers to any molecule of interest that unintendedly is added to the sample, and analysed with the target molecules. As shotgun sequencing entails analysing all available nucleic acids in a sample, human saliva, skin microbiome, or microbes present in water and reagents are some of the sources of external contamination. Incorporating DNA and RNA from these sources can distort the biological signal, which can lead researchers into incorrect conclusions. The following measures contribute to minimise external contamination:

- Always wear gloves and work in sterile environments, such as clean laminar flow cabinets.
- Use filtered pipette tips.
- Separate pre-PCR and post-PCR laboratories.
- Process and sequence blank controls.

Internal cross-contamination Another common type of contamination is that happening among samples. During the many pipetting actions laboratory protocols entail, is not uncommon to transfer small amounts of samples to adjacent tubes or wells. This can obviously distort the sample, and lead researchers into incorrect conclusions. The following measures contribute to minimise internal cross-contamination:

- Process all batches in an identical way for errors to be detectable.
- Avoid pipetting from the top of the tube to minimise sprays.
- Process and sequence blank controls.

Batch effects The last global consideration is to be aware of batch effects and try to minimise or account for their impact in downstream analyses. Batch effects are almost unavoidable in holo-omic data generation, because samples are usually processed in batches. Each batch can suffer different types of technical biases, including the aforementioned contamination issues, but also other problems derived from the use of different reagent stocks, different researchers executing identical protocols in slightly different ways, or storing samples for variable time periods. The critical measure to minimise the impact of batch effects and account for them in downstream analysis is to randomise samples. Randomising means randomly assigning samples from different contrasting groups to the different batches, minimising correlation between batches and experimental groups. If this is done, laboratory batches can be included as covariates in statistical analyses, which enable accounting and controlling for batch effects in the final results.

Procedures for generating multi-omic data

This chapter contains sections dedicated to each of the omic layers included in the workbook.

- **Nucleic-acid sequencing-based approaches**
  - DNA/RNA extraction for **HG**, **HT**, **MG** and **MT**
  - Sequencing library preparation for **HG** and **MG**
  - Sequencing library preparation for **HT**
  - Sequencing library preparation for **MT**
- **Mass spectrometry-based approaches**
  - Protein extraction for **HP** and **MP**
  - Metabolite extraction for **ME**

4 DNA/RNA extraction

Hundreds or (probably) thousands of different protocols and variations exist for extracting and purifying nucleic acids. Protocols can be classified based on methodological (e.g., chemical vs. physical DNA/RNA isolation, column-based vs bead-based, commercial vs. open-access).

Sample preprocessing

Bead-beating Contents to be added

Freeze-heat shock Contents to be added

Tissue digestion Contents to be added

Chemical isolation

Based on chemical separation of nucleic acids from the rest of molecules.

Physicochemical isolation

Based on physical separation of nucleic acids from the rest of molecules.

Column-based Contents to be added

Bead-based Contents to be added

Available protocols

Contents to be added

5 Protein/metabolite extraction

Laboratory protocols for protein/metabolite extraction

6 Sequencing library preparation

Sequencing of DNA and RNA molecules require sequencing libraries to be prepared, which entails modifying original nucleic acid molecules to allow sequencing platforms to identify the target molecules and perform the sequencing.

Sequencing strategies and platforms

Library features are specific to each sequencing platform, which requires selecting in advance the sequencing strategy to be employed. Pure nucleic acid sequencing-based strategies can be broadly divided in two groups. Short-read sequencing (SRS) platforms provide large amounts of data yet with short sequencing reads (typically 150 nucleotides). In contrast, long-read sequencing (LRS) platforms yield much longer sequences (thousands or even million of nucleotides), yet with a lower throughput, and typically lower sequence quality. The SRS market is dominated by two main companies with proprietary platforms, namely Illumina and BGI, although PacBio recently released their own SRS platform called ONSO. The LRS market is also dominated by two different companies with proprietary technologies, which are Oxford Nanopore (ONT) and Pacific Biosciences (Pacbio).

Sequencing enterprises, as well as auxiliary biotechnological companies, provide library preparation kits that can be more or less customised for different purposes.

Technology	Platforms	Sequencing type	Company
Sequencing by synthesis (SBS)	MiSeq, NovaSeq	Short-read sequencing	Illumina
Combinatorial probe-anchor synthesis (cPAS)	DNBSeq	Short-read sequencing	BGI
Sequencing by binding (SBB) technology	Onso	Short-read sequencing	PacBio
Single Molecule Real-Time sequencing (SMRT)	Sequel, Revio	Long-read sequencing	PacBio
Nanopore sequencing	MinION, GridION, PromethION	Long-read sequencing	Oxford Nanopore

Some of the most widely used sequencing technologies and platforms.

PCR-based vs. PCR-free library preparation

Sequencing library preparation procedures can be split into two main groups depending on whether they PCR-amplify or not the DNA templates. Unlike in the case of targeted amplicon sequencing, in which the objective is to amplify a specific target region, the aim of including a PCR step in shotgun-based library preparation is to increase the molarity of the library and/or to attach indices (see below) to the adaptors.

Learn more about PCR-based and PCR-free library preparation in this article by Jones et al. [7].

Indices and multiplexing

Usually, library preparation also entails tagging molecules with unique sample identifiers known as indices, which enable pooling molecules derived from multiple samples in a single sequencing run. This can be achieved in PCR-free protocols by using adaptors containing unique indices per sample, or by using indexed amplification primers in PCR-based library preparation protocols.

Learn more about indices and multiplexing in this article by Kircher et al. [8].

Unique molecular identifiers (UMIs)

Unique molecular identifiers (UMIs) are a type of molecular barcoding that provides error correction and increased accuracy during sequencing by uniquely tag each molecule (rather than each pool of molecules derived from a sample) in a sample library. UMIs are used for a wide range of sequencing applications, many around PCR duplicates in DNA and cDNA. UMI deduplication is also useful for RNA-seq gene expression analysis and other quantitative sequencing methods.

Learn more about unique molecular identifiers in this article by Kivioja et al. [9].

Contents of this section were created by Antton Alberdi.

6.1 Host genomics and microbial metagenomics

The library preparation strategies for generating host genomic (HG) and microbial metagenomic (MG) data are generally the same.

DNA fragmentation Short-read sequencing libraries requires DNA to be sheared to the desired fragment-length (usually 400-500 nucleotides), which can be achieved using either chemical (e.g., restriction enzymes) or physical (e.g. ultrasonication) procedures. Some long-read sequencing libraries intend to keep the largest DNA molecules possible, although some others recommend fragmenting to optimal mid-length molecules (e.g., around 10,000 nucleotides for Pacbio HiFi). After fragmentation, many library preparation protocols require repairing molecule ends by converting 5'-protruding and/or 3'-protruding ends to 5'-phosphorylated, blunt-end (see below) molecules.

Adaptor ligation In shotgun libraries adaptors are merged to DNA template molecules through chemical ligation (e.g., using a ligase enzyme). The ligation process is slightly different depending on whether the DNA template has blunt- or sticky-ends. In blunt ends, both strands are of equal length – i.e. they end at the same base position, leaving no unpaired bases on either strand, while in sticky ends, one strand is longer than the other. Some protocols deliberately create sticky-ends from blunt-end fragmented DNA molecules by adding a single adenine base to form an overhang by an A-tailing reaction. This A overhang allows adapters containing a single thymine overhanging base to pair with the DNA fragments.

An example of a blunt-end molecule:

```
5'-GATCTGACTGATGCGTATGCTAGT-3'
3'-CTAGACTGACTACGCATACGATCA-5'
```

An example of a sticky-end molecule:

```
5'-GATCTGACTGATGCGTATGCTAGT-3'
3'-CTAGACTGACTACGCATACGATC-5'
```

List of available protocols

Type	Name	Author/owner	Protocol/Article
SRS	Blunt-End Single-Tube (BEST) library prep protocol	Open access	Article
SRS	Santa Cruz Reaction (SCR) single-stranded library prep protocol	Open access	Article
SRS		Illumina	Protocol
LRS	SMRTbell prep kit 3.0 for PacBio HiFi Sequencing	Pacbio	Protocol

6.2 Host transcriptomics

Library preparation for host transcriptomics (HT) requires some extra steps to the already described procedures for host genomics and microbial metagenomics. This is due to two main reasons. First, because RNA molecules cannot be directly built into most sequencing libraries, and require instead to generate complementary DNA (cDNA) before library preparation. Second, because gene transcripts tend to be overwhelmingly dominated by rRNA and mtDNA genes, which are often not of interest for the researcher.

Sample quality assessment Before starting any library preparation protocol assessing the quality of RNA samples is strongly recommended. While traditionally assessed through agarose gel electrophoresis, nowadays RNA quality assessment is performed on electropherogram profiles, which are produced by nucleic acid fragment analysis instruments (e.g. Bioanalyzer, Fragment Analyzer). Traditionally, a simple model evaluating the 28S to 18S rRNA ratio was used as a criterion for RNA quality. However, the most common metric currently employed for assessing the preservation quality of RNA is the RNA integrity number (RIN), which accounts for more RNA features for assessing sample quality [10]. RIN values range from 10 (intact RNA) to 1 (totally degraded RNA). For example, the poly(A) enrichment procedures explained below require high quality RNA (RIN $>$ 8), because RNA degradaation to breaks within the transcript body and due to the selection of the poly(A) tail, the 3' ends are enriched while the more 5' sequences would not be captured, leading to a strong 3' bias for degraded RNA inputs.

DNA removal Depending on the RNA extraction method employed, it is not rare trace amounts of genomic DNA (gDNA) to be co-purified with RNA. Contaminating gDNA can interfere with reverse transcription and may lead to false positives, higher background, or lower detection in sensitive applications such as RT-qPCR. The traditional method of gDNA removal is the addition of DNase I to RNA extracts. DNase I must be removed prior to cDNA synthesis since any residual enzyme would degrade single-stranded DNA. Unfortunately, RNA loss or damage can occur during DNase I inactivation treatment. As an alternative to DNase I, double-strand–specific DNases are available to eliminate contaminating gDNA without affecting RNA or single-stranded DNAs.

Stranded vs. non-stranded transcriptomics RNA-Seq libraries can be stranded or non-stranded (unstranded), a decision that affects data analysis and interpretation. Stranded RNA-Seq (also referred to as strand-specific or directional RNA-Seq) enables researchers to determine the orientation of the transcript, whereas this information is lost in non-stranded, or standard, RNA-Seq. Non-stranded RNA-Seq is often sufficient for measuring gene expression in organisms with well-annotated genomes, as with a reference transcriptome, it is possible to infer orientation for most of the sequencing reads. As there are fewer steps

than stranded library preparation, the benefits of this approach are lower cost, simpler execution, and greater recovery of material, which renders non-stranded RNA-Seq the preferred option for holo-omic analyses. In contrast, stranded RNA-Seq is useful if the aims include annotating genomes, identifying antisense transcripts or discovering novel transcripts.

cDNA conversion Most RNA-Seq experiments are carried out on instruments that sequence DNA molecules, rather than RNA. This implies that RNA conversion to cDNA is a required step before library preparation. The synthesis of cDNA from an RNA template is carried out via reverse transcription using reverse transcriptases. In nature, these enzymes convert the viral RNA genome into a complementary DNA (cDNA) molecule, which can integrate into the host's genome, among other processes.

Reverse transcription, similar to PCR, requires the use of primers. Two main types of primers:

- **Random primers**: this type of primers are oligonucleotides with random base sequences. They are often six nucleotides long and are usually referred to as random hexamers. While random primers help improve cDNA synthesis for detection, they are not suitable for full-length reverse transcription of long RNA. Increasing the concentration of random hexamers in reverse transcription reactions improves cDNA yield but results in shorter cDNA fragments due to increased binding at multiple sites on the same template
- **oligo(dT) primers**: this type of primers consist of a stretch of 12–18 deoxythymidines that anneal to poly(A) tails of eukaryotic mRNAs (see the section below for further details).

Reverse transcription reactions for cDNA library construction and sequencing involve two main steps: first-strand cDNA synthesis and second-strand cDNA synthesis.

- **First-strand cDNA synthesis**: this initial step generates a cDNA:RNA hybrid through the below-described three-step process.
  - **Primer annealing**: in this step primers are attached to the RNA template, which usually happens before reverse transcriptase and necessary components (e.g., buffer, dNTPs, RNase inhibitor) are added.
  - **DNA polymerisation**: in this step the complementary DNA is polymerised by the reverse transcriptase enzyme. With oligo(dT) primers (Tm ~35–50°C), the reaction is often incubated directly at the optimal temperature of the reverse transcriptase (37–50°C), while random hexamers typically have lower Tm (~10–15°C) due to their shorter length. Using a thermostable reverse transcriptase allows, a higher reaction temperature (e.g., 50°C), to help denature RNA with high GC content or secondary structures without impacting enzyme activity. With such enzymes, high-temperature incubation can result in an increase in cDNA yield, length, and representation.
  - **Enzyme deactivation**: in this final step temperature is increased to 70–85°C, depending upon the thermostability of the enzyme, to deactivate the reverse transcriptase.
- **Second-strand cDNA synthesis**: in this second step the first-strand cDNA is used as a template to generate double-stranded cDNA representing the RNA targets. Synthesis of double-stranded cDNA often employs a different DNA polymerase to produce the complementary strand of the first cDNA strand.

rRNA depletion through poly-A enrichment Ribosomal RNA (rRNA) helps translate the information in messenger RNA (mRNA) into protein. It is the predominant form of RNA found in most cells, which can make over 80% of cellular RNA despite never being translated into proteins itself. In consequence, most reads derived from RNA belong to rRNAs, unless depletion strategies are implemented.

Excessively abundant rRNA sequences can be depleted using multiple strategies, which are covered in the Microbial metatranscriptomics. The most broadly employed enrichment strategy when dealing with eukaryotic organisms is rRNA depletion through poly-A enrichment. This strategy relies on the fact that mature coding mRNAs of eukaryotic organisms contain polyA tails, long chains (tens to hundreds) of adenine nucleotides that are added to primary RNA transcripts to increase the stability of the molecule. However, not all

transcripts contain poly(A) tails. microRNAs, small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs), some long non-coding RNAs (lncRNAs), and even protein-coding mRNAs such as histone mRNAs do not contail poly(A) tails, thus will be removed together with rRNA during poly(A) selection. If interested in quantifying expression of such transcripts the use of alternative methods is recommended.

The most broadly employed strategies to deplete rRNA through poly-A enrichment rely either on hybridisation with Oligo(dT)-attached magnetic beads or oligo(dT) priming during cDNA conversion step. In the former strategy, poly(A)-containing RNA molecules hybridise with Oligo(dT) stretches attached to magnetic beads. Following hybridisation, the supernatant consisting of non-polyadenylated molecules is removed. The beads are washed prior to elution of the poly(A)-selected RNA in water or buffer.

List of available protocols

Type	Name	Author/owner	Protocol/Article
oligo(dT) hybridisation oligo(dT) priming	Dynabeads Oligo (dT)25-61005	Thermo Fisher	Protocol

Contents of this section were created by Antton Alberdi.

6.3 Microbial metatranscriptomics

Sequencing library preparation for microbial metatranscriptomics faces the same challenges as host genomics, but the fact that prokaryotic mRNA have no poly-A tails makes it impossible to apply oligo(dT)-based rRNA depletion strategies. There are three other alternatives through which prokaryotic rRNA can be depleted. These three strategies require designing oligos, probes or guides whose sequences complement the DNA sequences that should be removed. Most commercial kits contain probes designed to remove rRNA sequences of the most commonly employed animal hosts (Human/Mouse/Rat), as well as bacteria, but custom probes targeting any genes could be employed. The first two methods shown below are implemented before library-preparation, thus independent reactions must be ran for each sample. The last strategy is implemented after library preparation, which enables multiple indexed libraries to be pooled, and thus performing a single reaction per pool.

Capture-based rRNA depletion

This method relies on capture rRNA with complimentary oligos that are coupled to paramagnetic beads. Unwanted transcripts get bound to beads, which can then be retained using a magnet, while the non-hybridasing transcripts remain in the elution.

RNAse-based rRNA depletion

A more recent technological upgrade to capture-based rRNA depletion is to, instead of using paramagnetic beads, degrade RNA:DNA hybrids using RNase H [11].

CRISPR/Cas9-based rRNA depletion

The newest method of all three relies on the DNA claveage capacity of the Cas9 enzyme [12]. In this method, custom-designed guides are used for the Cas9 enzyme to cleveage unwanted sequences. This strategy is applied once libraries are prepared, and before the final PCR amplification is conducted. When the targetted molecules are cleaveged, they lack one of the two adaptors, and therefore they are not amplified, resulting in a considerable depletion compared to the rest of the library.

List of available protocols

Name	Strategy	Author/owner	Protocol/Article
Custom capture-based depletion	Capture-based	Open Source	Article [13]
Legacy Ribo-Zero	Capture-based	Illumina	
Custom RNAse-based depletion	RNAse-based	Open Source	Article [11]
Ribo-Zero Plus	RNAse-based	Illumina	
NEBNext® rRNA Depletion Kit	RNAse-based	NEB	
DASH	Cas9-based	Open Source	Article [14]

Contents of this section were created by Antton Alberdi.

Part III

BIOINFORMATIC PROCEDURES

7 About bioinformatics

Bioinformatic processing of raw sequencing and mass spectrometry data is the computational step that precedes statistical analyses and integration of multi-omic data. Through bioinformatic processing raw data are converted into meaningful bits of information, usually drastically decreasing the size of the data sets that are used for downstream analyses.

Raw sequencing and mass spectrometry-based data files used in holo-omic analyses are typically in the realm of gigabytes (Gb) or even terabytes (Tb). Many of the performed operations require large amounts of memory (some more than 1Tb), which makes it impossible to process data in personal computers. Instead, most bioinformatics tasks are performed in computational clusters with access to large amounts of memory and many CPUs and GPUs, which enable parallelising computational tasks thus speeding up data processing time.

However, for the sake of simplicity and practicality, the example datasets included in this Workbook have been considerably downscaled to enable reproducing the exercises in personal computers.

All bioinformatic analyses included in the **Holo-omics workbook** are conducted in a Unix command line Shell environment (BASH/SH). You can find the details to set-up your SHELL environment in the section Prepare your Shell environment.

7.1 Prepare your shell environment

If the comment chunks of the code (text after #) is creating you problems, use the following code to disable interactive comments and avoid issues when copy-pasting code:

```
setopt interactivecomments
```

Required software

Bioinformatic pipelines for processing omic data require the use of dozens of software. All the required software are listed in the conda environment installation file available here.

Install conda / miniconda

Conda is an open-source package management system and environment management system that quickly installs, runs, and updates packages and their dependencies. If **conda** is not installed in your system, the first step is to install **miniconda** (a free minimal installer for conda, enough to run the bioinformatic analyses explained in the workbook). Miniconda installers for Linux, Mac and Windows operating systems can be found in the following website: https://docs.conda.io/en/latest/miniconda.html

Once conda or miniconda is installed in your system, you should be able to create and manage your conda environments. You can test whether conda has been succesfully installed using the following code:

```
conda -V
#> conda 22.11.1 #or whatever version you have installed
```

Install mamba (optional)

An optional step is to install **mamba**, which is a reimplementation of the conda package manager in C++, which speeds up many of the processes. Mamba can be installed through the command line using the `conda install` option.

```
conda install mamba -n base -c conda-forge
```

Create a conda environment

All the bioinformatic analyses explained in this workbook will be run within an environment containing all the necessary software. The file that specifies which software to install in the environment is available here, and can be retrieved using wget (as shown in the code below), or downloading from the Internet browser. If using the latter option, don't forget to provide the absolute path to the 'holo-omics-env.yaml' file in the `mamba create` command.

```
wget https://raw.githubusercontent.com/holo-omics/holo-omics.github.io/main/bin/holo-omics-env.yaml #do
conda update conda #to ensure everything is updated
conda deactivate #deactivate any conda environment before creating a new one
conda env create -f holo-omics-env.yml
rm holo-omics-env.yaml #remove installer file
```

As the environment contains dozens of softwares, the process of creating it will take a while. It is recommended to have a good Internet connection to speed-up software download. Once the installation is over, you can double-check whether the environment has been successfully created using the following script:

```
conda activate holo-omics
#> (holo-omics) anttonalberdi@Anttons-MBP ~ %
```

The (holo-omics) specifies the environment you are at. To get out of the environment use:

```
conda env list
#> base                  *  /Users/anttonalberdi/miniconda3
#> holo-omics               /Users/anttonalberdi/miniconda3/envs/holo-omics
```

Activate the holo-omics conda environment

Whenever running the holo-omic analyses explained in this workbook, it will be necessary to activate the holo-omics environment through the following command:

```
conda activate holo-omics
#> (holo-omics) anttonalberdi@Anttons-MBP ~ %
```

Install software in conda environment

Content to be added.

```
#Example code goes here
```

7.2 Example data for bioinformatics

Contents to be added here.

```
#Example code goes here
```

7.3 Using snakemake for workflow management

Contents to be added here.

```
#Example code goes here
```

8 Sequencing data preprocessing

The first step of the bioinformatic pipeline is to pre-process the raw sequencing data to prepare them for downstream analyses.

Preprocess the reads using fastp

Raw sequencing data require an initial preprocessing to get rid off low-quality nucleotides and reads, as well as any remains of sequencing adaptors that can mess around in the downstream analyses. An efficient way to do so is to use the software **fastp**, which can perform all above-mentioned operations in a single go and directly on compressed files.

fastP documentation can be found here.

```
fastp \
    --in1 {input.r1i} --in2 {input.r2i} \
    --out1 {output.r1o} --out2 {output.r2o} \
    --trim_poly_g \
    --trim_poly_x \
    --low_complexity_filter \
    --n_base_limit 5 \
    --qualified_quality_phred 20 \
    --length_required 60 \
    --thread {threads} \
    --html {output.fastp_html} \
    --json {output.fastp_json} \
    --adapter_sequence {params.adapter1} \
    --adapter_sequence_r2 {params.adapter2}
```

Splitting host and non-host data

Depending on the sample type employed for data generation, sequencing data might contain only host reads, only microbial reads, or a mixture of both. For example, blood sampled from an animal is expected to only contain host DNA/RNA reads (unless an infection is ongoing), while DNA extracted from a microbial culture is only expected to contain microbial DNA/RNA reads (unless human contamination has happened). In contrast, intestinal content samples, faecal samples, leave samples or root samples can contain both host and microbial nucleic acids.

```
bowtie2-build \
    --large-index \
    --threads {threads} \
        {output.rn_catted_ref} {output.rn_catted_ref}
```

Index host genome

```
# Map reads to catted reference using Bowtie2
bowtie2 \
    --time \
    --threads {threads} \
    -x {input.catted_ref} \
    -1 {input.r1i} \
    -2 {input.r2i} \
| samtools view -b -@ {threads} - | samtools sort -@ {threads} -o {output.all_bam} - &&

# Extract non-host reads (note we're not compressing for nonpareil)
samtools view -b -f12 -@ {threads} {output.all_bam} \
| samtools fastq -@ {threads} -1 {output.non_host_r1} -2 {output.non_host_r2} - &&

# Send host reads to BAM
samtools view -b -F12 -@ {threads} {output.all_bam} \
| samtools sort -@ {threads} -o {output.host_bam} -
```

Map samples to host genomes

9 Host genomics (HG) data processing

Contents to be added.

9.1 Host reference genome

Contents to be added.

9.2 Host genome resequencing

Contents to be added.

10 Microbial metagenomics (MG) data processing

Microbial metagenomic data processing can be conducted following different strategies. Decision on which approach to use should be based on the aims of the study, available reference data, amount of generated data, and many other criteria. In this workbook we consider three main approaches that require different bioinformatic pipelines to be implemented.

- **Reference-based approach:** it relies on a reference database of microbial genomes to which sequencing reads can be mapped to obtain estimations of relative proportion of reads belonging to each of the genomes available in the reference database. It is the simplest and computationally less expensive approach, yet it completely relies on a complete and representative reference database.
- **Assembly-based approach:** it is based on assembling sequencing reads into longer DNA sequences known as contigs, which can then be used to predict genes and perform functional analyses. The main

limitation of this approach is that the entire metagenome (set of contigs) in each sample is considered as a single unit, thus overlooking which bacterial genome each detected gene belongs to.
- **Genome-resolved approach:** it is the most advanced of the three approaches, and the strategy that provides the largest amount of information, as the aim of this approach is to directly reconstruct all the genomes in a metagenome. This is achieved by binning contigs into Metagenome-Assembled Genomes (MAGs), which can then be taxonomically and functionally annotated to perform sound community-level analyses.

10.1 Reference-based

Assembly-based.

10.2 Assembly-based

Assembly-based approaches can be divided in two main strategies:

- Individual assembly-based
- Coassembly-based

10.2.1 Individual assembly-based

Two of the most popular metagenome assemblers are **Megahit** and **MetaSpades**. Metaspades is considered superior in terms of assembly quality, yet memory requirements are much larger than those of Megahit. Thus, one of the most relevant criteria to choose the assembler to be employed is the balance between amount of data and available memory. Another minor, yet relevant difference between both assemblers is that Megahit allows removing contings below a certain size, while MetaSpades needs to be piped with another software (e.g. bbmap) to get rid off barely informative yet often abundant short contigs.

```
megahit \
    -t {threads} \
    --verbose \
    --min-contig-len 1500 \
    -1 {input.r1} -2 {input.r2} \
    -o {params.workdir}
    2> {log}
```

10.2.1.1 Individual assembly using Megahit

```
metaspades.py \
    -t {threads} \
    -k 21,33,55,77,99 \
    -1 {input.r1} -2 {input.r2} \
    -o {params.workdir}
    2> {log}

# Remove contigs shorter than 1,500 bp using bbmap
reformat.sh \
    in={params.workdir}/scaffolds.fasta \
    out={output.assembly} \
    minlength=1500
```

10.2.1.2 Individual assembly using MetaSpades

10.2.1.3 Get assembly statistics using Quast

The metagenome assemblies can have very different properties depending on the amount of data used for the assembly, the complexity of the microbial community, and other biological and technical aspects. It is therefore convenient to obtain some general statistics of the assemblies to decide whether they look meaningful to continue with downstream analyses. This can be easily done using the software **Quast**.

```
quast \
    -o {output.report} \
    --threads {threads} \
    {input.assembly}
```

TO BE CONTINUED FROM HERE: https://github.com/earthhologenome/EHI_bioinformatics/blob/main/0_Code/2_Individual_Assembly_Binning.snakefile

10.2.2 Coassembly-based

10.3 Genome-resolved

Contents to be added.

11 Host transcriptomics (HT) data processing

Contents to be added.

12 Microbial metatranscriptomics (HT) data processing

Contents to be added.

13 Host proteomics (HP) data processing

Contents to be added.

14 Microbial metaproteomics (MP) data processing

Contents to be added.

Part IV

STATISTICAL PROCEDURES

15 About statistics

Statistics is probably the most challenging step of holo-omic studies, due to two main factors: the extreme complexity of the data, often containing thousands of features, and the limited sample size, often in the realm of the dozens of sampling units. This combination renders many holo-omic datasets rather statistics unfriendly.

A step-by-step approach

In this workbook we strongly encourage researchers to proceed step-by-step when dealing with holo-omics data and biological questions.

Initial quantitative exploration of omic layers The analysis of any multi-omic data should begin with independent analysis of each omic layer to learn about its structure and variability before jumping to multi-omic data integration.

- **Data transformations:** multivariate datasets consist of different data types (e.g., presence-absence of taxa, counts of genes, community-level metabolic capacity index of a function, concentrations of metabolites across samples) that may require specific transformation before applying statistical techniques.
- **Unsupervised exploration of omic layers:** include exploratory techniques, such as cluster analysis and ordination-based visualisation methods, which reveal the structure and main patterns of the omic datasets without prior information about experimental design. These procedures might reveal that the observations are structured into meaningful groups or that variables can be reduced to fewer dimensions.
- **Supervised analysis of omic layers:** this type of analyses incorporate information of experimental design and aim at testing and estimating the effects of the experimental factors (e.g., dietary treatment, drug administration) or variables of interest (e.g., age of the experimental subjects, geographic location of studied populations) on different omic layers.

Multi-omic data integration When it comes to multi-omic data integration, the approaches can be broadly categorised into two types: multi-staged analysis and meta-dimensional or simultaneous analysis.

- **Multi-staged integration:** leverages the central dogma of molecular biology to assume that the variation in omic datasets is hierarchical, such that variation in DNA leads to variation in RNA and so on to determine the phenotype
- **Meta-dimensional integration:** considers the possibility that the phenotype is the product of the combination of variation across all omic layers, with the presence of complex inter-omic interactions.

All statistical analyses included in the **Holo-omics workbook** are conducted in R environment. You can find the details to set-up your R environment in the section Prepare your R environment.

15.1 Prepare your R environment

All statistical analyses included in the **Holo-omics workbook** are conducted in R environment [15]. R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS, and in order to use it, R or RStudio must be installed in your local computer or remote server.

Required packages

In order to reproduce the analyses shown in the workbook, a rather long list of R packages must be installed. Packages are the fundamental units of reproducible R code, which include reusable R functions, the documentation that describes how to use them, and sample data.

- ape
- DESeq2
- distillR
- ggplot2
- tidyverse
- vegan
- (. . .)

Package installation

Packages are installed programatically using three main ways: through CRAN, Bioconductor or Github.

Install package from CRAN CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Packages stored in CRAN can be installed using the following code:

```
install.packages("package_name")
#e.g.
install.packages("vegan")
```

Install package from Bioconductor Bioconductor is a free, open source and open development software project for the analysis and comprehension of genomic data generated by wet lab experiments in molecular biology. Packages included in Bioconductor can be installed using the following code:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("package_name")
#e.g.
BiocManager::install("DESeq2")
```

Install package from Github GitHub is a code hosting platform for version control and collaboration. Packages stored in R can be installed using the following code after installing the package devtools:

```
library(devtools)
install_github("github_repository_name_of_the_package")
#e.g.
install_github("anttonalberdi/distillR")
```

15.2 Example data for statistics

Contents to be added here.

16 Single omic analyses

Here is a review of existing methods.

17 Data transformations

Here is a review of existing methods.

18 Unsupervised exploration

Unsupervised methods include exploratory techniques, such as cluster analysis and ordination-based visualisation methods, which reveal the structure and main patterns of the omic datasets without prior information about experimental design. These procedures might reveal that the observations are structured into meaningful groups or that variables can be reduced to fewer dimensions. Researchers can then opt for using the outputs of these analyses, rather than the original multidimensional datasets, for the multi-omic data integration. Most clustering and ordination techniques are computed from association matrices, thus it is essential to do an appropriate pre-transformation of the data and choice of association coefficient, since this influences the final outcome of the analyses

- Cluster analysis
- Dimension reduction and ordination

18.1 Cluster analysis

Clustering procedures group features or observations into homogeneous sets by minimising within-group and maximising among-group distances

18.1.1 Hierarchical clustering

Hierarchical clustering produces a stratified organisation of features or observations where relatively similar objects are grouped together. The clustering can be performed using different criteria to measure the distance between clusters, which will affect the final outcome of the analysis (e.g., single linkage, complete linkage, average linkage and Ward's minimum variance).

```
#Example code goes here
```

A useful exploratory analysis to reveal general patterns in an omic layer can be obtained by simultaneous application of hierarchical clustering to the rows and columns of the data matrix, and visualising the results in a heatmap.

```
#Example code goes here
```

18.1.2 Disjoint clustering

Disjoint clustering techniques aim at separating the objects into individual, usually mutually exclusive, and in most cases, unconnected clusters. K-means clustering is one of the most typical algorithms where objects are assigned to k clusters using an iterative procedure that minimises the within-clusters sums of squares. Other available clustering methods include twinspan, self-organising maps, dbscan and Dirichlet multinomial mixtures (DMM). DMM were specifically developed to analyse MG data but can be equally useful for other sequencing-based omic datasets.

```
#Example code goes here
```

18.2 Dimension reduction and ordination

Ordination is a method complementary to data clustering, which enables displaying differences among samples graphically through reducing the dimensions of the original data set, so that similar objects are near and dissimilar objects are farther from each other.

18.2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is one of the most widely applied methods for ordination. PCA generates new synthetic variables (principal components) that are linear combinations of the original variables and capture as much variance of the original data as possible. The principal components are orthogonal to each other and correspond to the successive dimensions of maximum variance of the scatter of points. The distance preserved among objects is euclidean and the relationships among variables are linear, thus PCA should generally be applied after appropriate transformations.

```
#Example code goes here
```

18.2.2 Principal Coordinate Analysis (PCoA)

Content to be added here.

```
#Example code goes here
```

18.2.3 Non-metric Multidimensional Scaling (NMDS)

Content to be added here.

```
#Example code goes here
```

18.2.4 t-Distributed Stochastic Neighbour Embedding (t-SNE)

Content to be added here. Requires many data points.

```
#Example code goes here
```

18.2.5 Uniform manifold approximation and projection (UMAP)

Content to be added here.

```
#Example code goes here
```

18.2.6 Potential of heat diffusion for affinity-based transition embedding (PHATE)

Content to be added here.

```
#Example code goes here
```

19 Supervised analysis

The supervised analyses of omic layers, in contrast to unsupervised ones, incorporate information of experimental design, and can be divided into two types of problems: regression and classification. A regression problem is when the output of the model is a numeric variable or a matrix, such as the phenotypic characteristics of the host or the omic data sets themselves. These methods aim at testing and estimating the effects of the experimental factors (e.g., dietary treatment, drug administration) or variables of interest (e.g., age of the experimental subjects, geographic location of studied populations) on different omic layers, or associating the omic layers with host phenotypic features. A classification problem is when the output of the model is categorical. In the context of multi omic studies, classification methods aim at classifying observations into their experimental groups (e.g. health status, dietary treatment) based on their features on different omic layers.

19.1 Regression methods

Independently testing the effects of the experimental factors of interest on different omic layers can be very informative to get an overall picture of how the host and the microbiome are responding to the environment and/or the experimental treatment.

19.1.1 PERMANOVA

Content to be added here.

```
#Example code goes here
```

19.1.2 ANOSIM

Content to be added here.

```
#Example code goes here
```

19.1.3 Redundancy analysis (RDA)

Content to be added here.

```
#Example code goes here
```

19.1.4 Canonical Correspondence Analysis (CCA)

Content to be added here.

```
#Example code goes here
```

19.1.5 Generalised linear modelling (GLM)

Content to be added here.

```
#Example code goes here
```

19.1.6 Generalised linear mixed modelling (GLMM)

Content to be added here.

```
#Example code goes here
```

19.2 Classification methods

It is in classification problems where ML algorithms have proven most useful.

19.2.1 Random Forests (RF)

Content to be added here.

```
#Example code goes here
```

19.2.2 Support Vector Machines (SVM)

Content to be added here.

```
#Example code goes here
```

20 Multi-omic integration

Here is a review of existing methods.

21 Multi-staged integration

Here is a review of existing methods.

22 Meta-dimensional integration

Here is a review of existing methods.

Part V

RESOURCES

23 Useful links

Genomics

- **Data Wrangling and Processing for Genomics (website):**

Shell command line usage

- **Introduction to the Command Line for Genomics (website):** general overview of basic command line usage.

R usage (General usage and programming)

- **Intro to R and RStudio for Genomics (website):**
- **Efficient R programming (website):** best practices for programming in R.

R usage (Graphics and visualisation)

- **Fundamentals of Data Visualization (website):** guide to making visualisations that accurately reflect the data, tell a story, and look professional.
- **R Graphics Cookbook (website):** a practical guide that provides more than 150 recipes to generate high-quality graphs using ggplot2.

Statistics

- **An Introduction to Statistical Learning (book):** freely available book about general statistical learning covering regression and classification problems through linear modelling and machine learning.
- **High dimensional statistics with R (website):** virtual lesson specialised in dealing with high dimensional data.

24 References

1. Nyholm L, Koziol A, Marcos S, Botnen AB, Aizpurua O, Gopalakrishnan S, et al. Holo-Omics: Integrated Host-Microbiota multi-omics for basic and applied biological research. iScience. 2020;23:101414.
2. Limborg MT, Alberdi A, Kodama M, Roggenbuck M, Kristiansen K, Gilbert MTP. Applied hologenomics: Feasibility and potential in aquaculture. Trends Biotechnol. 2018;36:252–64.
3. Fischer CN, Trautman EP, Crawford JM, Stabb EV, Handelsman J, Broderick NA. Metabolite exchange between microbiome members produces compounds that influence drosophila behavior. Elife. 2017;6.
4. Wu H-J, Wu E. The role of gut microbiota in immune homeostasis and autoimmunity. Gut Microbes. 2012;3:4–14.
5. Alberdi A, Andersen SB, Limborg MT, Dunn RR, Gilbert MTP. Disentangling host-microbiota complexity through hologenomics. Nat Rev Genet. 2022;23:281–97.
6. Mushegian AA, Arbore R, Walser J-C, Ebert D. Environmental sources of bacteria and genetic variation in behavior influence Host-Associated microbiota. Appl Environ Microbiol. 2019;85.
7. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci U S A. 2015;112:14024–9.
8. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the illumina platform. Nucleic Acids Res. 2012;40:e3.
9. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2011;9:72–4.
10. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: An RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7:3.
11. Huang Y, Sheth RU, Kaufman A, Wang HH. Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. Nucleic Acids Res. 2020;48:e20.
12. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of abundant sequences by hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016;17:41.
13. Kraus AJ, Brink BG, Siegel TN. Efficient and specific oligo-based depletion of rRNA. Sci Rep. 2019;9:12281.
14. Prezza G, Heckel T, Dietrich S, Homberger C, Westermann AJ, Vogel J. Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. RNA. 2020;26:1069–78.
15. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2008.