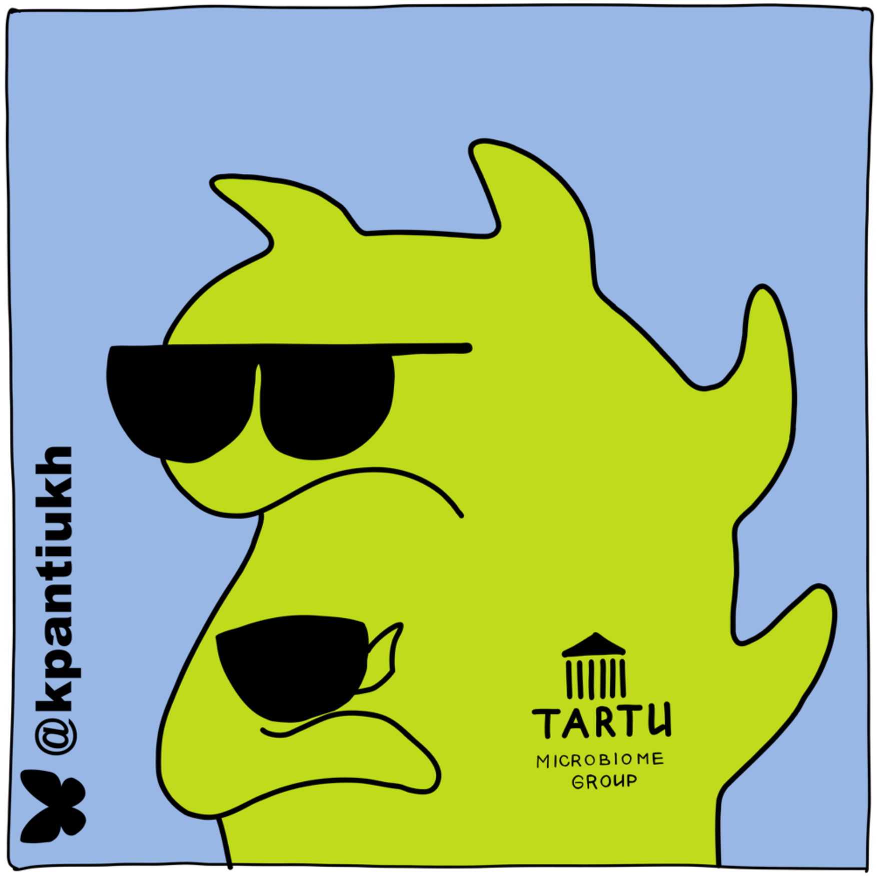


Metagenomics

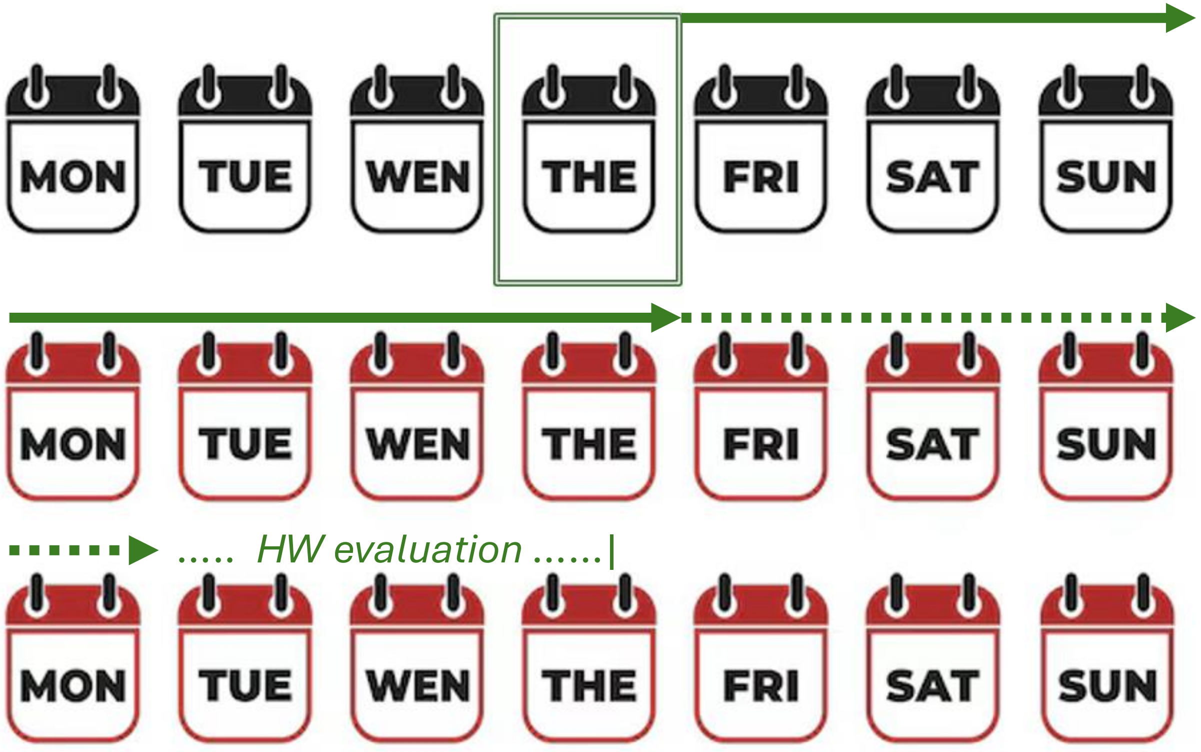
Lecture 3

Describing microbiome communities. Relative abundance and its limitations. Alpha- and beta-diversity metrics. Case–control study design

Kateryna Pantiukh
pantiukh@ut.ee

GitHub

Revision of homework deadline



..... *HW evaluation* ......|

Microbiome

A community of **microorganisms** that lives in a specific environment

- *Bacteria*
- *Archaea*
- *Viruses*
- *Microeucarites*

- Primar degrades
- Primary fermenters
- Secondary fermenters
- Sinks

Microbiome

A community of **microorganisms** that lives in a specific environment

- *Bacteria*
- *Archaea*
- *Viruses*
- *Microeucarites*

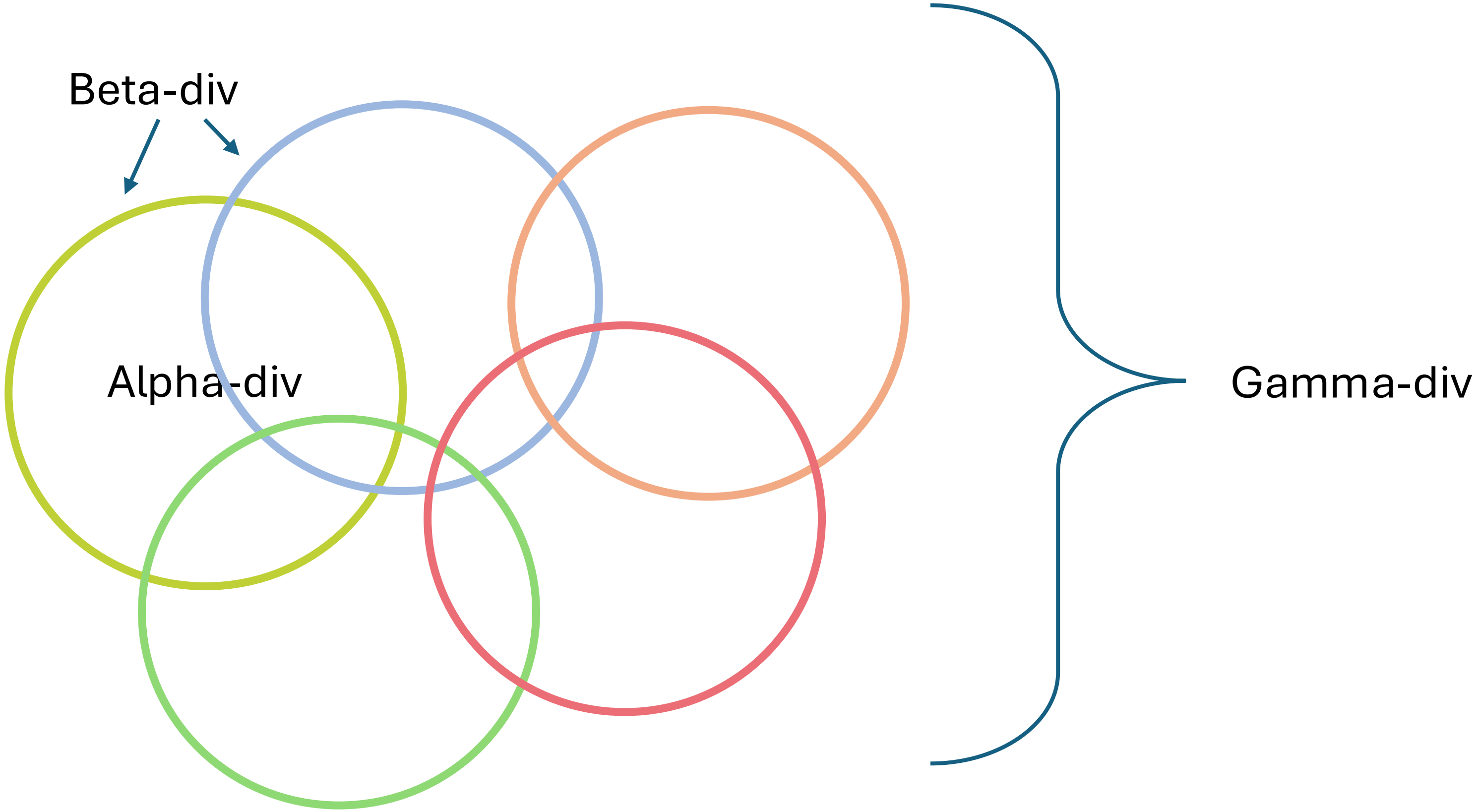
- Primar degrades
- Primary fermenters
- Secondary fermenters
- Sinks

Community may have different level of complexity

… may be evaluated with
Community **diversity indexes**

alpha-, beta-, gamma-

Microbiome



Microbiome

A community of **microorganisms** that lives in a specific environment

- *Bacteria*
- *Archaea*
- *Viruses*
- *Microeucarites*

- Primar degrades
- Primary fermenters
- Secondary fermenters
- Sinks

Community may have different level of complexity

... may be evaluated with Community **diversity indexes**

alpha-, beta-, gamma-

... the first and simplest measures that can reveal differences between communities

Microbiome

A community of **microorganisms** that lives in a specific environment

- *Bacteria*
- *Archaea*
- *Viruses*
- *Microeucarites*

- Primar degrades
- Primary fermenters
- Secondary fermenters
- Sinks

*Associations between **community complexity** and the feature of interest*

Community **diversity indexes**

Microbiome

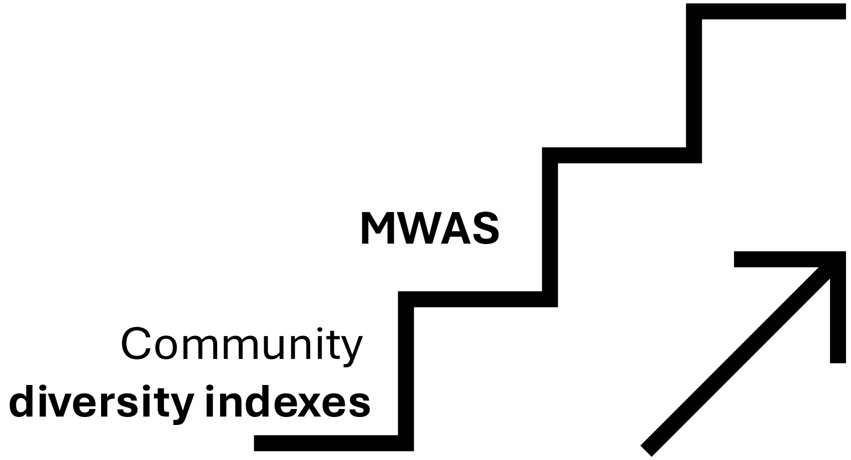
A community of **microorganisms** that lives in a specific environment

- *Bacteria*
- *Archaea*
- *Viruses*
- *Microeucarites*

- Primar degrades
- Primary fermenters
- Secondary fermenters
- Sinks

MWAS – <u>m</u>icrobiome <u>w</u>ide <u>a</u>ssociation <u>s</u>tudy

*Associations between **specific species** and the feature of interest*



MWAS

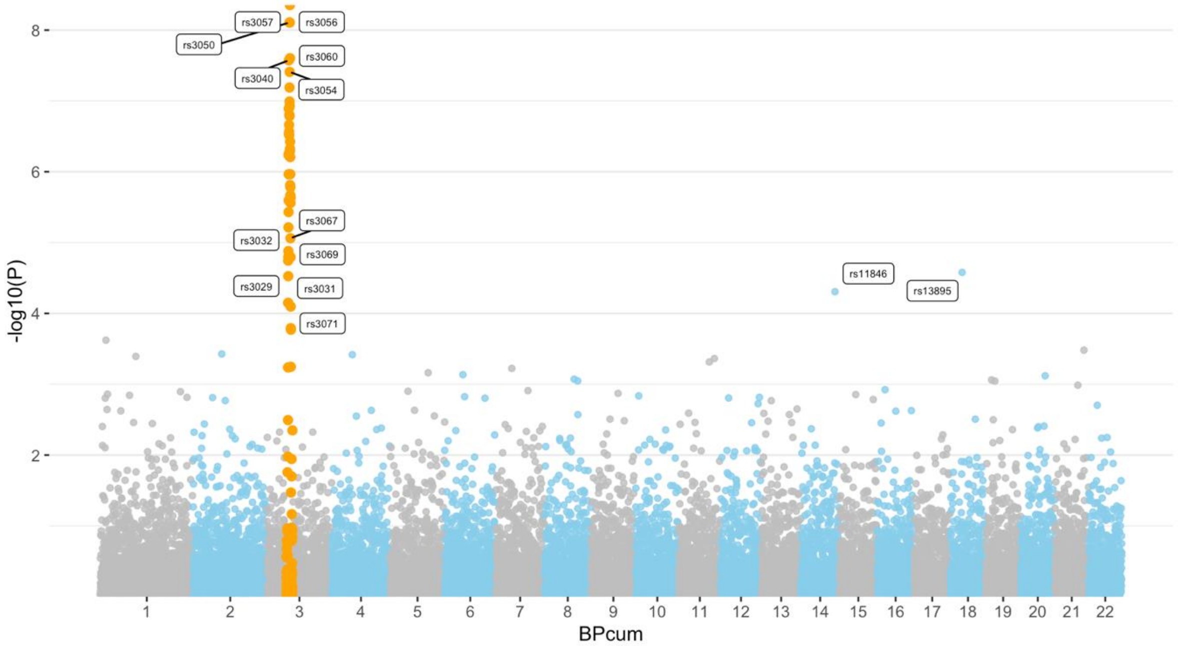
MWAS – microbiome wide association study

GWAS – genome wide association study

*Associations between **specific genome variations** (SNP or indel) and the feature of interest*

Genotyping

Manhattan plot

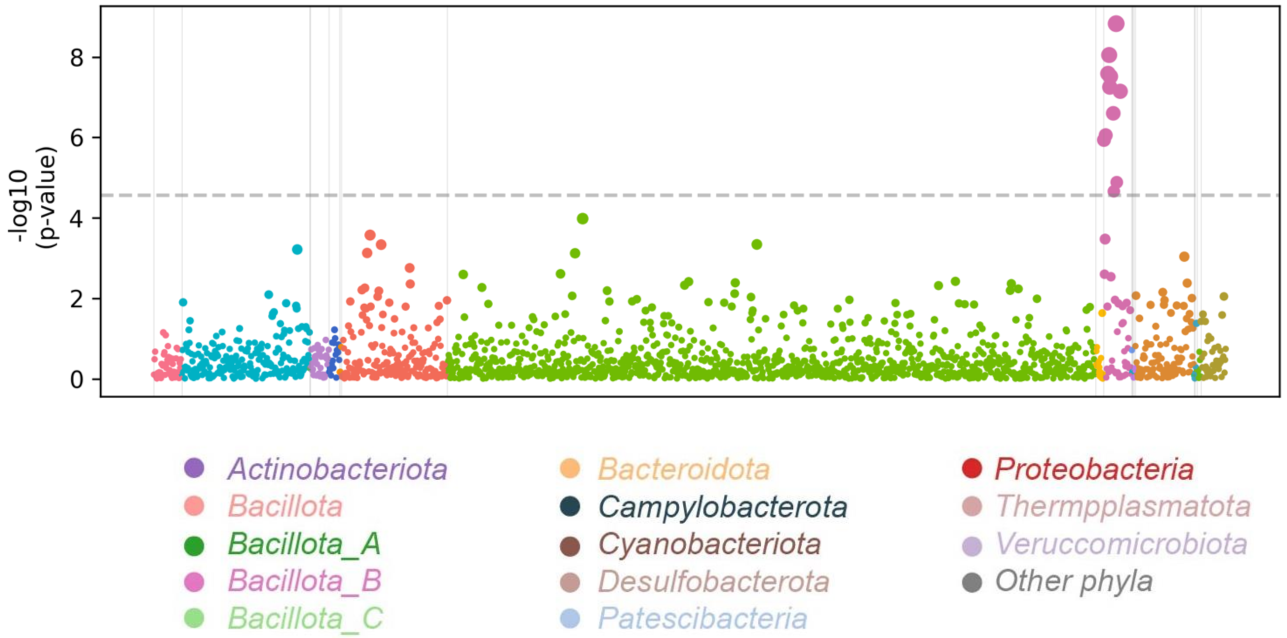


MWAS

MWAS – microbiome wide association study

Bacteria sp. instead of SNP

Manhattan plot



MWAS

MWAS – microbiome wide association study

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{j=2}^{p} \beta_j C_{ij} + \epsilon_i$$

Where:

- Y_i is the phenotype (e.g., disease status, quantitative trait) for sample i
- X_i is the abundance (or presence/absence) of a microbial feature in sample i
- C_{ij} are covariates (age, sex, sequencing batch, etc.)
- β_0 is the intercept
- β_1 is the effect size of the microbial feature
- ϵ_i is the residual error

MWAS

MWAS – microbiome wide association study

{ *Age, sex, BMI, Stool type* 💩 }

Heart Desease status = Intersept coeficient + ∑(effect of species × abundance of species) + ∑(effect of covariates × covariates) + error

MWAS

MWAS – microbiome wide association study

Heart Desease status = Intersept coeficient + ∑(effect of species × abundance of species) + ∑(effect of covariates × covariates) + error

Heart Desease status = Intersept coeficient + ∑(effect of species × abundance of species) + ∑(effect of covariates × covariates) + error

Heart Desease status = Intersept coeficient + ∑(effect of species × abundance of species) + ∑(effect of covariates × covariates) + error

→ !! Correction for multiple testing

MWAS

MWAS – microbiome wide association study

!! Correction for multiple testing

Bonferoni correction

Significance level / number of tests

0.05 / 3 = 0.0166

Corrected significance level = 0,0166

MWAS

MWAS – microbiome wide association study

Age, sex, BMI, Stool type 💩

Heart Desease status = Intersept coeficient + ∑(effect of species × abundance of species) + ∑(effect of covariates × covariates) + error

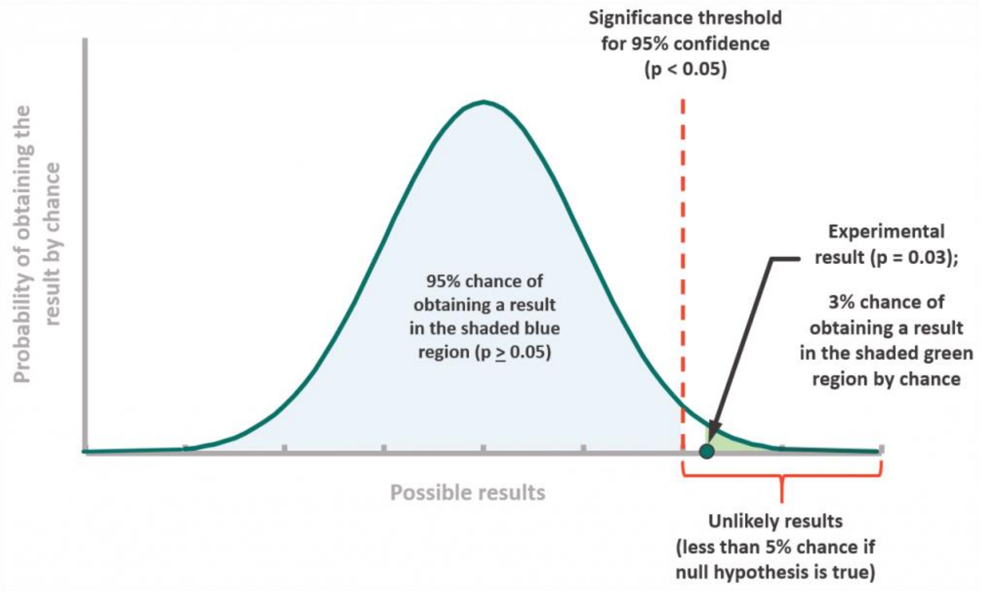
Main outcome: p-value and beta coeficient

p-value

MWAS – microbiome wide association study

null hypothesis: the true effect is zero

The p-value is the probability of observing a result **at least as extreme as the one you got**, assuming that null hypothesis is true.



p-value

MWAS – microbiome wide association study

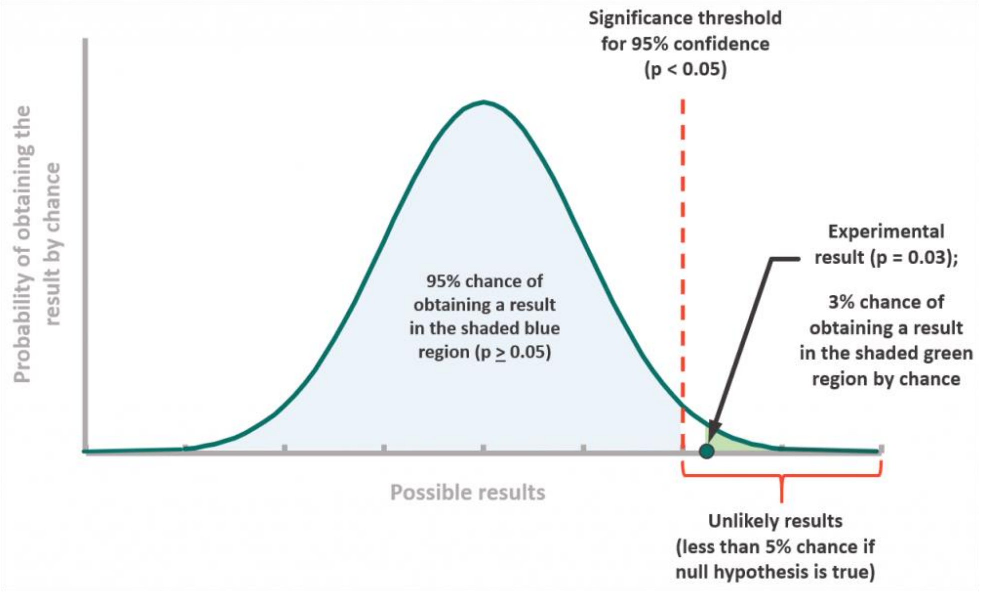
What it tells you:

- A **small p-value** means your result would be unlikely if the effect were truly zero.
- A **large p-value** means your data are quite compatible with no effect.

What it does *not* tell you:

- It is **not** a measure of effect size or importance!!!

* - Effect size - beta

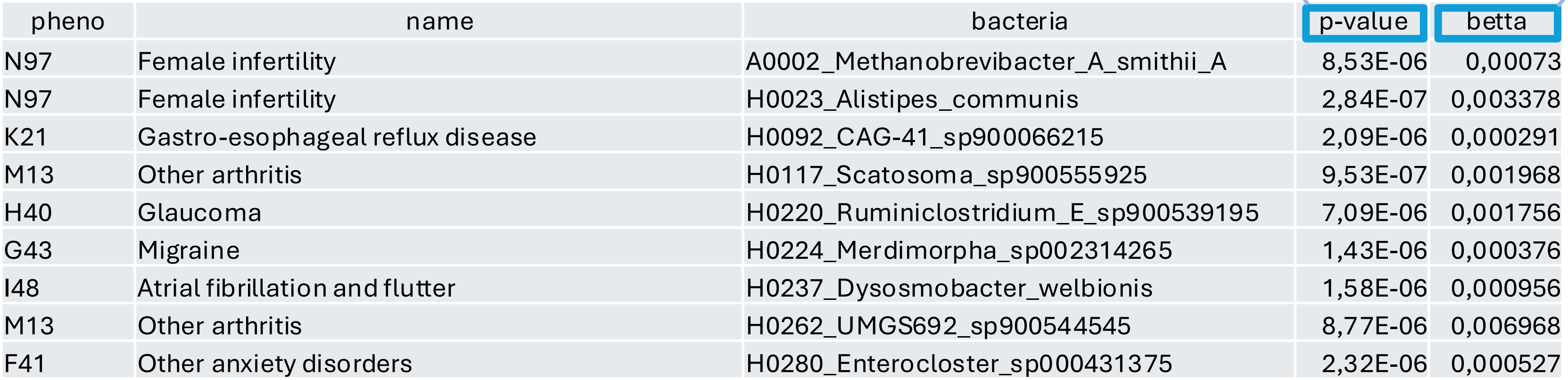


MWAS

MWAS – microbiome wide association study

If p-value < level of significance
We consider the association
Statistically significant

- Positive betta – positive correlation
- Negative betta – negative correlation

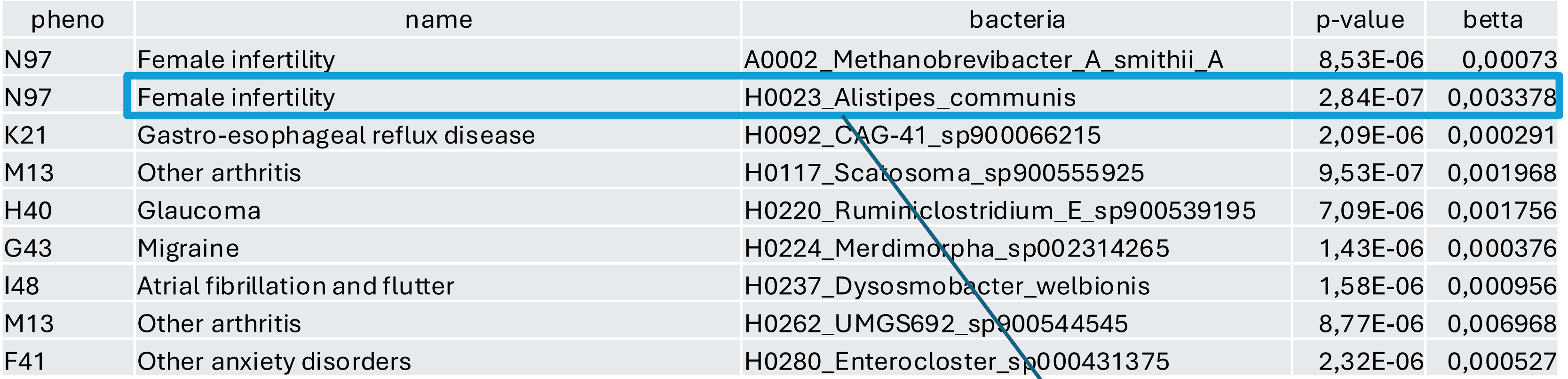


pheno	name	bacteria	p-value	betta
N97	Female infertility	A0002_Methanobrevibacter_A_smithii_A	8,53E-06	0,00073
N97	Female infertility	H0023_Alistipes_communis	2,84E-07	0,003378
K21	Gastro-esophageal reflux disease	H0092_CAG-41_sp900066215	2,09E-06	0,000291
M13	Other arthritis	H0117_Scatosoma_sp900555925	9,53E-07	0,001968
H40	Glaucoma	H0220_Ruminiclostridium_E_sp900539195	7,09E-06	0,001756
G43	Migraine	H0224_Merdimorpha_sp002314265	1,43E-06	0,000376
I48	Atrial fibrillation and flutter	H0237_Dysosmobacter_welbionis	1,58E-06	0,000956
M13	Other arthritis	H0262_UMGS692_sp900544545	8,77E-06	0,006968
F41	Other anxiety disorders	H0280_Enterocloster_sp000431375	2,32E-06	0,000527

MWAS

MWAS – microbiome wide association study

level of significance = 5 * 10-6



pheno	name	bacteria	p-value	betta
N97	Female infertility	A0002_Methanobrevibacter_A_smithii_A	8,53E-06	0,00073
N97	Female infertility	H0023_Alistipes_communis	2,84E-07	0,003378
K21	Gastro-esophageal reflux disease	H0092_CAG-41_sp900066215	2,09E-06	0,000291
M13	Other arthritis	H0117_Scatosoma_sp900555925	9,53E-07	0,001968
H40	Glaucoma	H0220_Ruminiclostridium_E_sp900539195	7,09E-06	0,001756
G43	Migraine	H0224_Merdimorpha_sp002314265	1,43E-06	0,000376
I48	Atrial fibrillation and flutter	H0237_Dysosmobacter_welbionis	1,58E-06	0,000956
M13	Other arthritis	H0262_UMGS692_sp900544545	8,77E-06	0,006968
F41	Other anxiety disorders	H0280_Enterocloster_sp000431375	2,32E-06	0,000527

Alistipes putredinis is positively associated with female infertility status, indicating higher relative abundance in diagnosed individuals compared with controls

MWAS

MWAS – microbiome wide association study

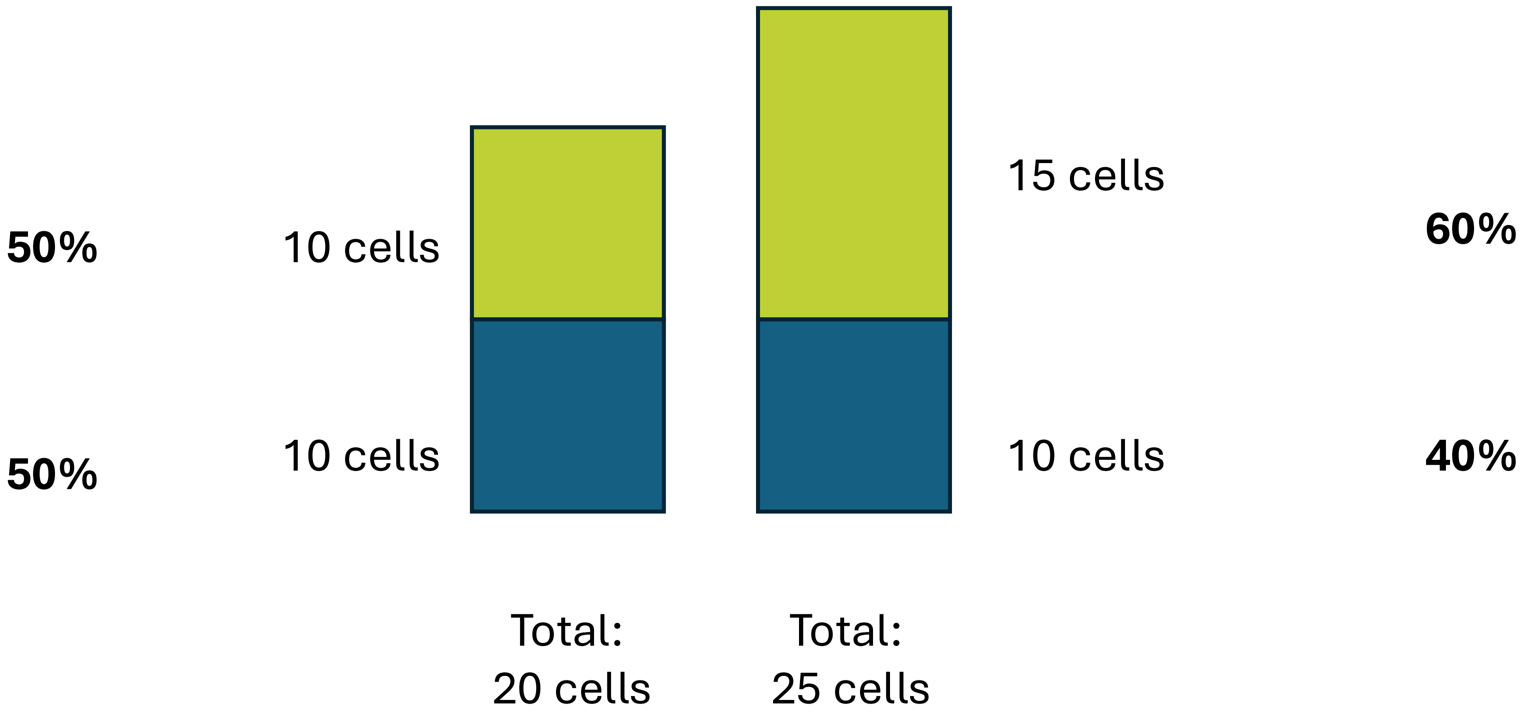
level of significance = 5 * 10-6

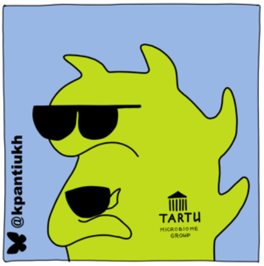
pheno	name	bacteria	p-value	betta
N97	Female infertility	A0002_Methanobrevibacter_A_smithii_A	8,53E-06	0,00073
N97	Female infertility	H0023_Alistipes_communis	2,84E-07	0,003378
K21	Gastro-esophageal reflux disease	H0092_CAG-41_sp900066215	2,09E-06	0,000291
M13	Other arthritis	H0117_Scatosoma_sp900555925	9,53E-07	0,001968
H40	Glaucoma	H0220_Ruminiclostridium_E_sp900539195	7,09E-06	0,001756
G43	Migraine	H0224_Merdimorpha_sp002314265	1,43E-06	0,000376
I48	Atrial fibrillation and flutter	H0237_Dysosmobacter_welbionis	1,58E-06	0,000956
M13	Other arthritis	H0262_UMGS692_sp900544545	8,77E-06	0,006968
F41	Other anxiety disorders	H0280_Enterocloster_sp000431375	2,32E-06	0,000527

Alistipes putredinis is positively associated with female infertility status, indicating higher relative abundance in diagnosed individuals compared with controls, however the estimated effect size is small

Abundance metrics

relative abundance PROBLEM





MWAS

MWAS – microbiome wide association study

INPUT:
relative abundance

	Tom	Mary
Species 1	0.1	3.7
Species 2	0.0	0.2
Species 3	2.3	0.0

INPUT:
presence-absence

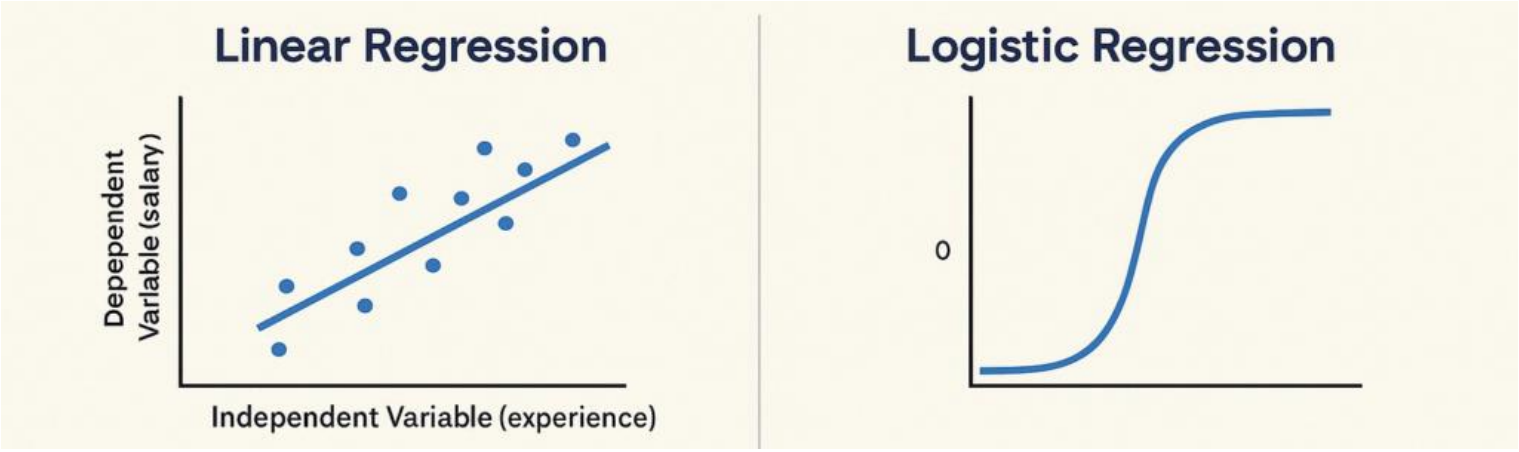
	Tom	Mary
Species 1	1	1
Species 2	0	1
Species 3	1	0

MWAS

MWAS – microbiome wide association study

INPUT:
relative abundance

INPUT:
presence-absence



Effect size - odds ratios

MWAS – microbiome wide association study

Example:

- Predictor: Alistipes putredinis present vs absent
- Outcome: Female infertility
- Logistic regression gives OR = 2.0

Interpretation:

Individuals with Alistipes putredinis present have twice the odds of being diagnosed with female infertility compared to individuals without this bacterium, holding other variables constant.

If OR = 0.5, the interpretation flips:

Individuals with Alistipes have half the odds of infertility compared to those without it.

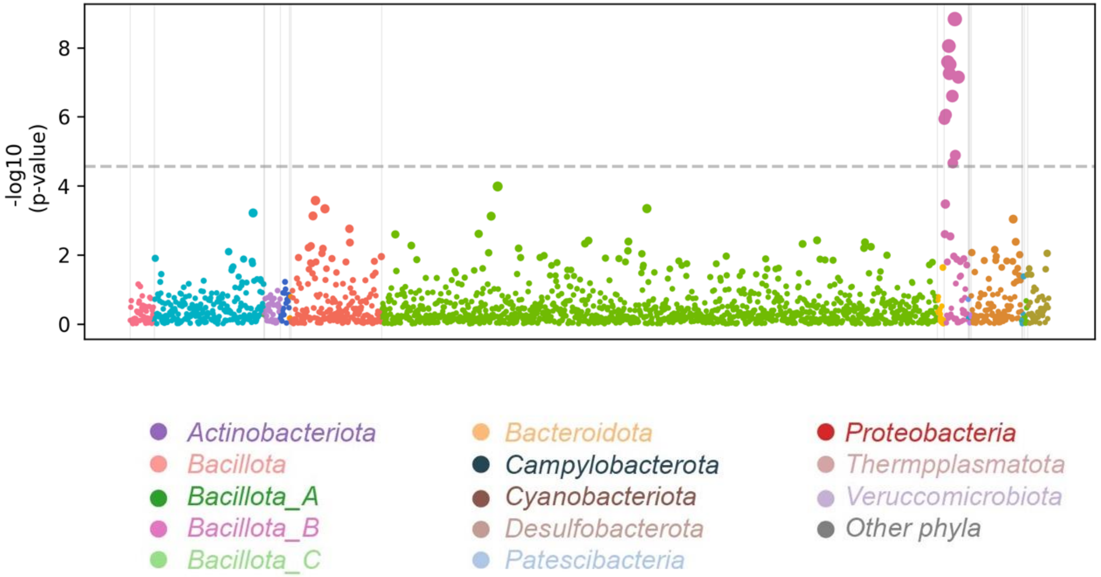
MWAS

Visualisation example

Linear regression

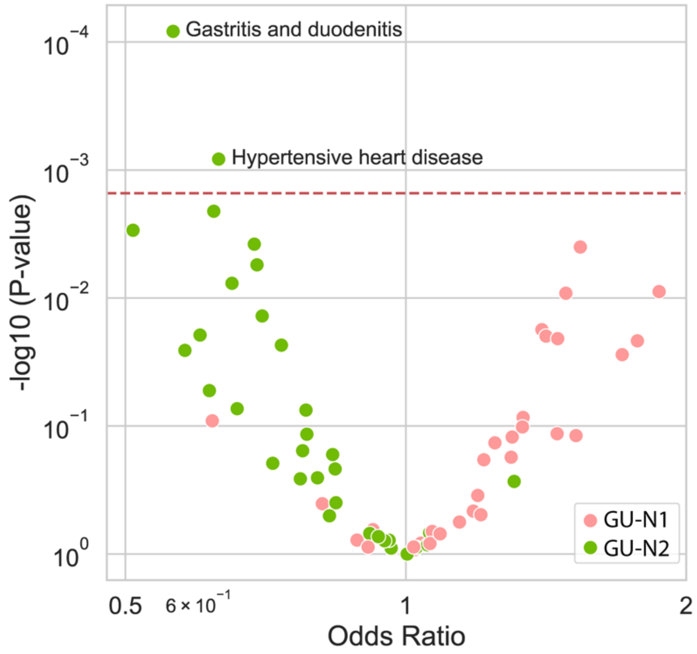
(relative abundance input)

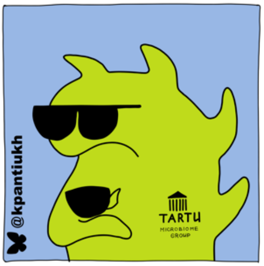
Manhattan plot



Logistic regression

(presence-absence input)

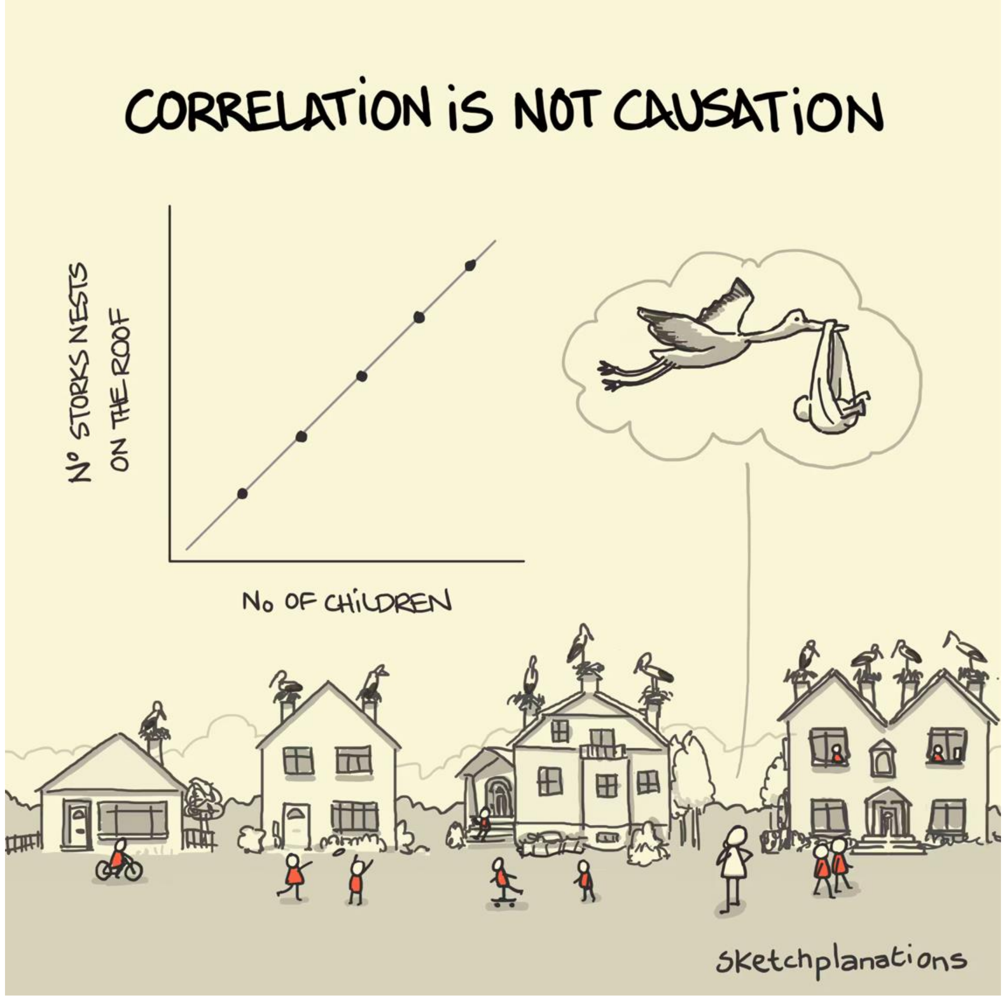




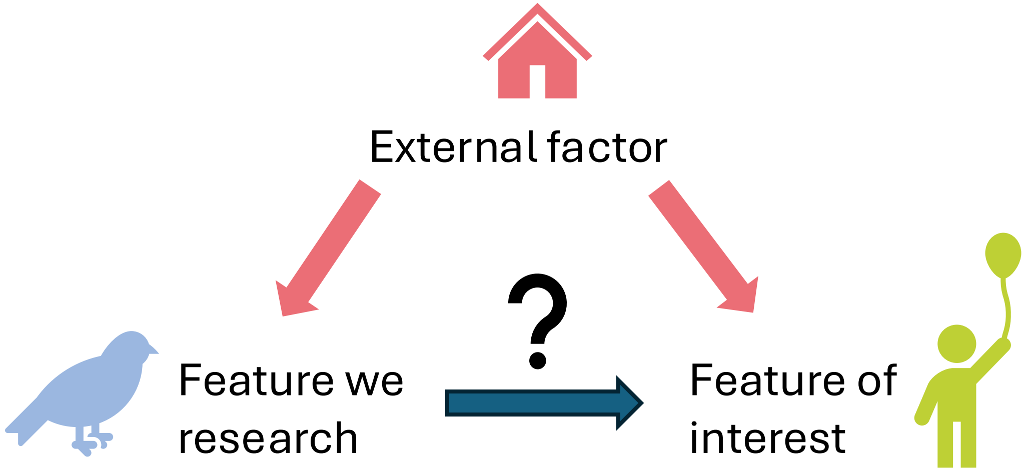
Causation issue

Correlation != Causation

Causation issue



Causation issue

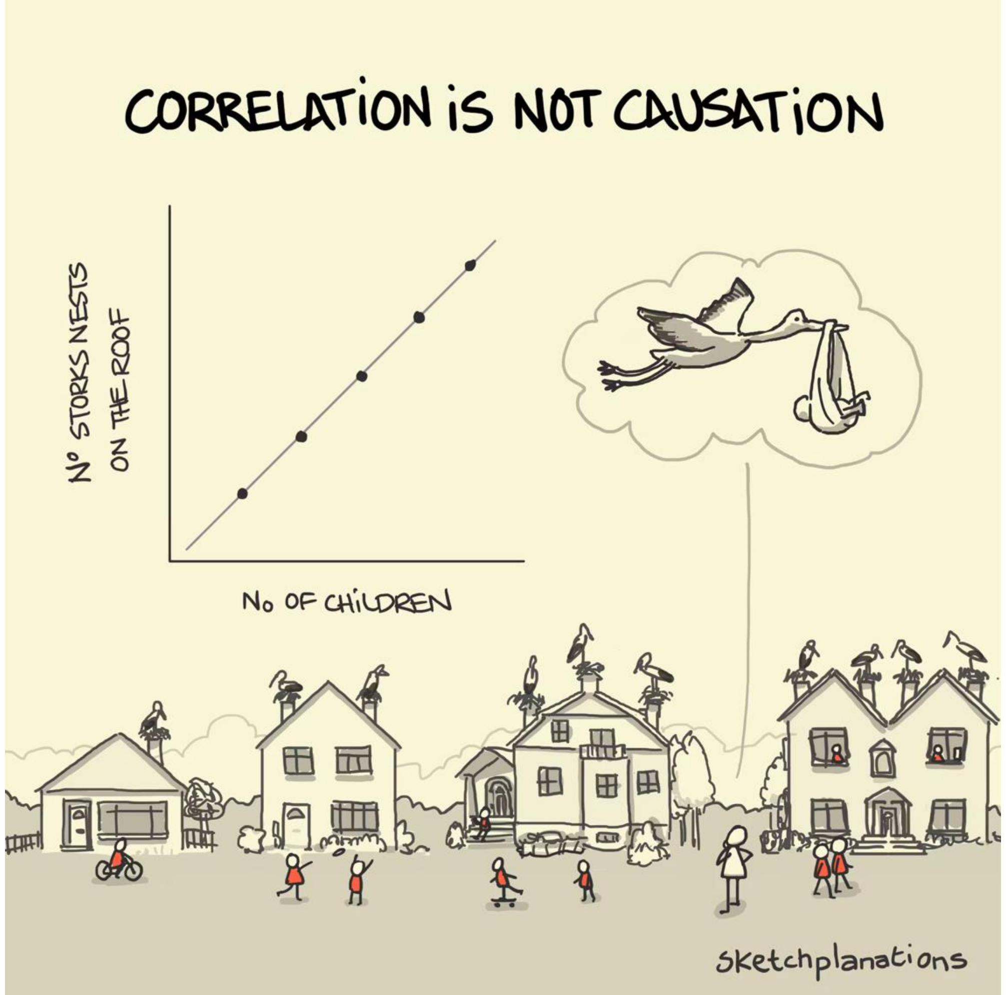


External factor

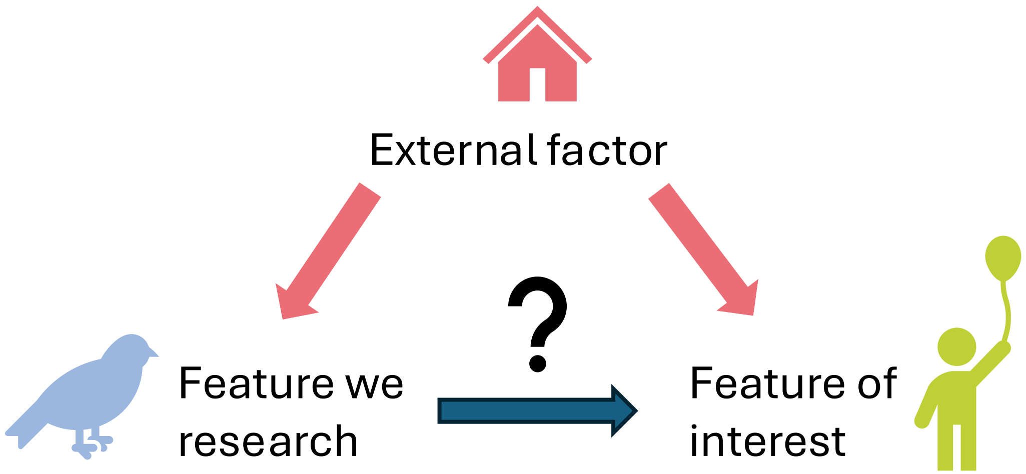
Feature we research

?

Feature of interest



Causation issue



External factor

Feature we research

?

Feature of interest

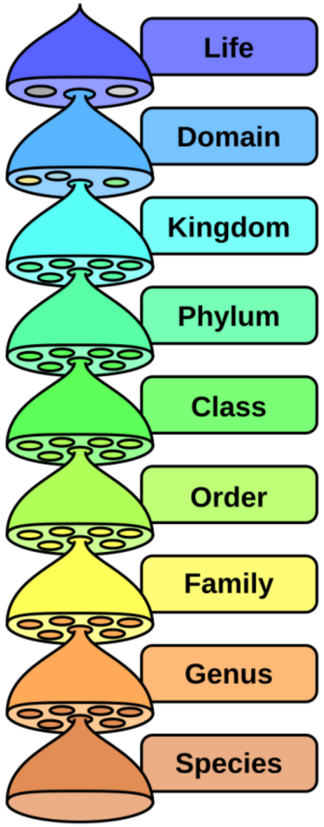
WHAT WE CAN DO ABOUT IT?

1. Even whe causation questionable, correlation may be use as a predictor

Usefull for early diagnosis

2. We can design additional experiments to check causeation

Different level abundance tables



	Tom	Mary
Genus 1	0.1	3.7
Genus 2	0.0	0.2
Genus 3	2.3	0.0

	Tom	Mary
Species 1	0.1	3.7
Species 2	0.0	0.2
Species 3	2.3	0.0

Low dimensionality → fewer multiple-testing problems.

Species & strain level:
can link associations to particular functions or pathogenic potential.

BUT: group size metters!

How to decide what taxonomic level to use?

1.Cohort size and statistical power

- species/strain have many rare or zero-count taxa.
- Smaller cohorts may not provide enough observations per taxon to detect associations reliably.
- Broader levels (phylum, class, family) aggregate taxa, increasing counts and power but may hide specific effects.

2.Expected prevalence of taxa

- Rare taxa are often absent in many samples.
- Presence/absence models can work for very rare taxa, but effect size estimates become unstable.
- Focus on taxa that occur in a meaningful fraction of samples (e.g., >10–20% prevalence).

Tip: Check prevalence at different levels before deciding; sometimes grouping into higher levels improves coverage.

3.Biological interpretability

- Broad levels show general trends (e.g., Bacteroidetes increase) but often lack actionable insights.
- Genus or species level allows linking findings to metabolic pathways, pathogenicity, or prior literature.
- Strain-level associations are most informative for functional or mechanistic hypotheses but require high-resolution sequencing.

How to decide what taxonomic level to use?

4.Multiple testing burden

- Species/strain levels increase the number of tests, requiring stricter p-value correction and reducing statistical power.
- Consider pre-filtering low-abundance taxa or focusing on taxa with prior evidence to reduce false negatives.

5.Sequencing depth and MAGs aviability

- Low-depth sequencing may not resolve species or strains reliably.
- High-resolution analysis is only meaningful if the data can support it; otherwise, stick to genus/family.

6.Hierarchical approach

- Start broad to detect global shifts, then zoom in to finer levels for taxa showing signals.
- This balances power, interpretability, and control of multiple testing.

How to decide what taxonomic level to use?

It’s about finding the right balance and testing many different options before discovering what works

- Small cohort, rare taxa, low sequencing depth → use higher taxonomic levels (family/genus).
- Large cohort, common taxa, high-resolution data → species or strain level can be explored.
- Always consider prevalence, expected group size, and interpretability.

Preparing an abundance table

1. Filter low-abundance and rare taxa

    - Remove taxa that are present in very few samples (<1% prevalence)
    - Remove taxa with extremely low relative abundance (not popular)

    This reduces sparsity, improves statistical power, and decreases the multiple-testing burden.

2. Handle zeros & Compositional transformation

    - Zero counts are common in microbiome data.
    - Perform CLR transformations

Bacterial species

main characteristics

Ecological niche / lifestyle

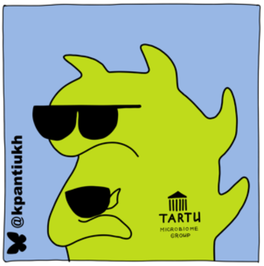
Functional traits / metabolism

Genomic

Morphology and structural features

Interaction with other microbes

Clinical or industrial relevance



Bacterial species

main characteristics

Ecological niche / lifestyle

Functional traits / metabolism

Genomic

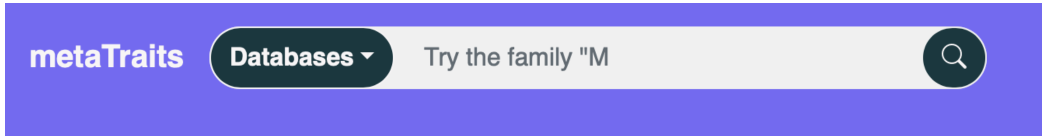
Morphology and structural features

Interaction with other microbes

Clinical or industrial relevance

- **Habitat:** gut, oral cavity, soil, water, skin, etc.
- **Host association:** commensal, symbiont, opportunistic pathogen, obligate pathogen.
- **Temperature preference:** psychrophile, mesophile, thermophile.
- **pH tolerance** and other environmental tolerances.

https://metatraits.embl.de



Bacterial species

main characteristics

Ecological niche / lifestyle

Functional traits / metabolism

Genomic

Morphology and structural features

Interaction with other microbes

Clinical or industrial relevance

- **Carbon source utilization:** sugars, proteins, lipids.
- **Energy generation:** respiration, fermentation, photosynthesis, chemolithotrophy.
- **Nitrogen/sulfur cycling capabilities:** nitrate reduction, sulfate reduction, ammonia oxidation.
- **Secondary metabolite production:** antibiotics, bacteriocins, signaling molecules.

Bacterial species

main characteristics

Ecological niche / lifestyle

Functional traits / metabolism

Genomic

Morphology and structural features

Interaction with other microbes

Clinical or industrial relevance

- **Genome size** and GC content.
- **Plasmid presence** or mobile genetic elements.
- **Virulence genes** or toxin production.
- **Antibiotic resistance genes**.

https://gtdb.ecogenomic.org

Welcome to GTDB

GENOME TAXONOMY DATABASE

732,475 genomes

Release 10-RS226 (16th April 2025)

Bacterial species

main characteristics

Ecological niche / lifestyle

Functional traits / metabolism

Genomic

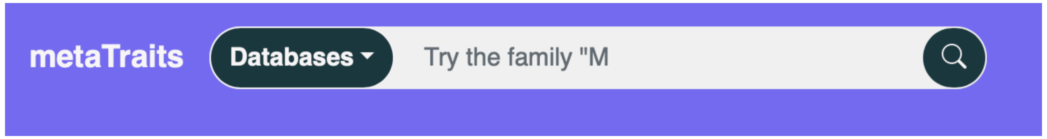
Morphology and structural features

Interaction with other microbes

Clinical or industrial relevance

- **Cell shape:** cocci, rods, spirals.
- **Motility structures:** flagella, pili.
- **Surface structures:** capsule, S-layer, biofilm-forming ability.
- **Sporulation ability.**

https://metatraits.embl.de



Bacterial species

main characteristics

Ecological niche / lifestyle

Functional traits / metabolism

Genomic

Morphology and structural features

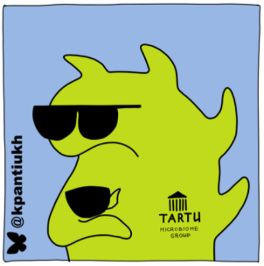
Interaction with other microbes

Clinical or industrial relevance

- **Symbiosis or antagonism:** production of inhibitory compounds, mutualistic relationships.
- **Biofilm formation:** ability to form communities on surfaces.
- **Quorum sensing / communication:** signaling mechanisms.

https://pubmed.ncbi.nlm.nih.gov

https://www.biorxiv.org



Bacterial species

main characteristics

Ecological niche / lifestyle

Functional traits / metabolism

Genomic

Morphology and structural features

Interaction with other microbes

Clinical or industrial relevance

- Pathogenicity to humans, animals, or plants.
- Probiotic potential.
- Industrial applications: fermentation, bioremediation, enzyme production.

https://pubmed.ncbi.nlm.nih.gov

PubMed®

https://www.biorxiv.org

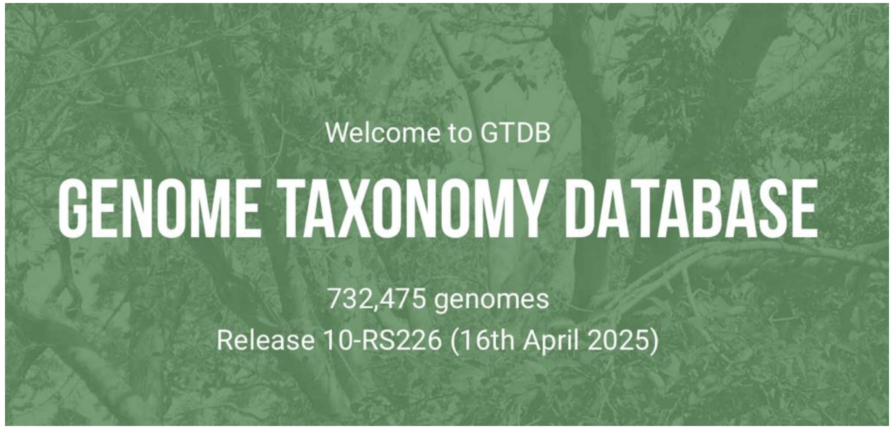
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

MetaTraits demo

https://metatraits.embl.de

GTDB demo

https://gtdb.ecogenomic.org



Welcome to GTDB

GENOME TAXONOMY DATABASE

732,475 genomes

Release 10-RS226 (16th April 2025)

Bioinformatic analysis

Although MWAS can identify significant associations, the results

- may show **low reproducibility** and
- do not provide evidence of **causality**

MWAS limitations

Low reproducibility can result from:

- Technical differences between studies or laboratories
- Population-specific variation
- Small cohort sizes limiting statistical power
- Complex data structure, such as compositionality, that complicates analysis

MWAS limitations

Evidence of causality can be obtained by:

- Designing controlled laboratory experiments to test mechanistic effects
- Developing microbiome-focused approaches analogous to Mendelian randomization

MWAS limitations

Although MWAS can identify significant associations, the results may show **low reproducibility** and do not provide evidence of **causality**

Frustration?
Am I confident in my data? Does it make sense?

Bioinformatic analysis

Bioinformatics … this is the way

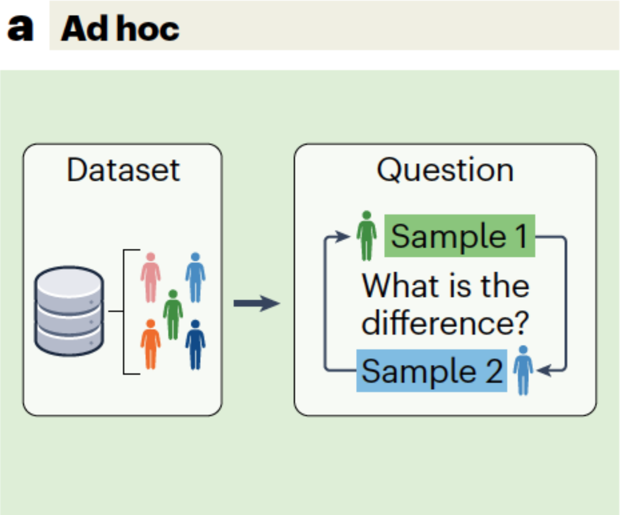
Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

Ad hoc

Literally: *“for this”* or *“for this purpose”*
From *ad* (to, for) + *hoc* (this).

- straightforward, problem-driven pattern recognition
- for hypothesis testing and/or early-stage hypothesis generation
- to answer specific questions
- easy to implement and adequate for preliminary or exploratory analyses



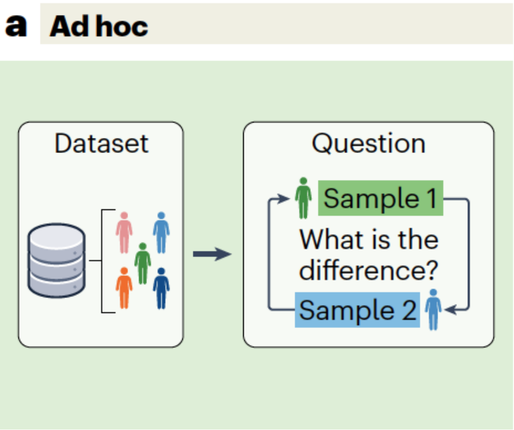
https://doi.org/10.1038/s41587-025-02852-0

Bioinformatic analysis

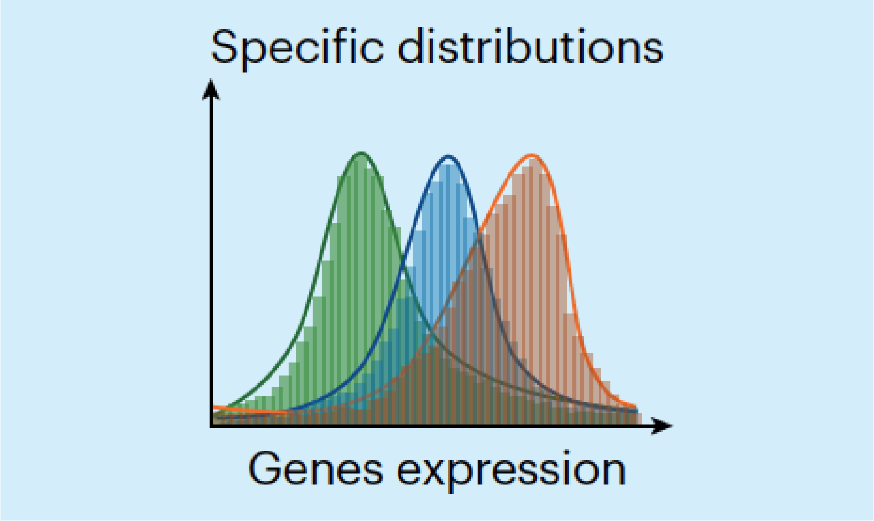
3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

Ad hoc

Literally: *“for this”* or *“for this purpose”*
From *ad* (to, for) + *hoc* (this).



Analytical strategies



... community diversity, MWAS

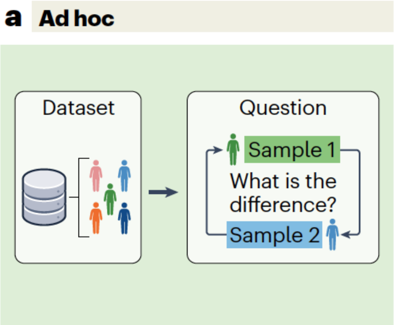
https://doi.org/10.1038/s41587-025-02852-0

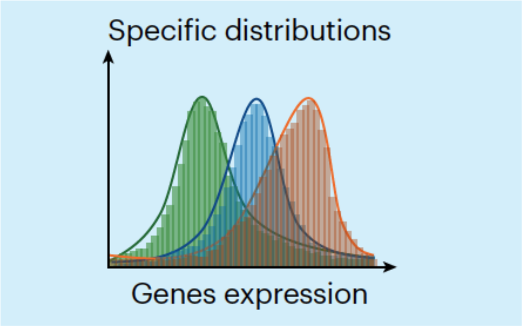
Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

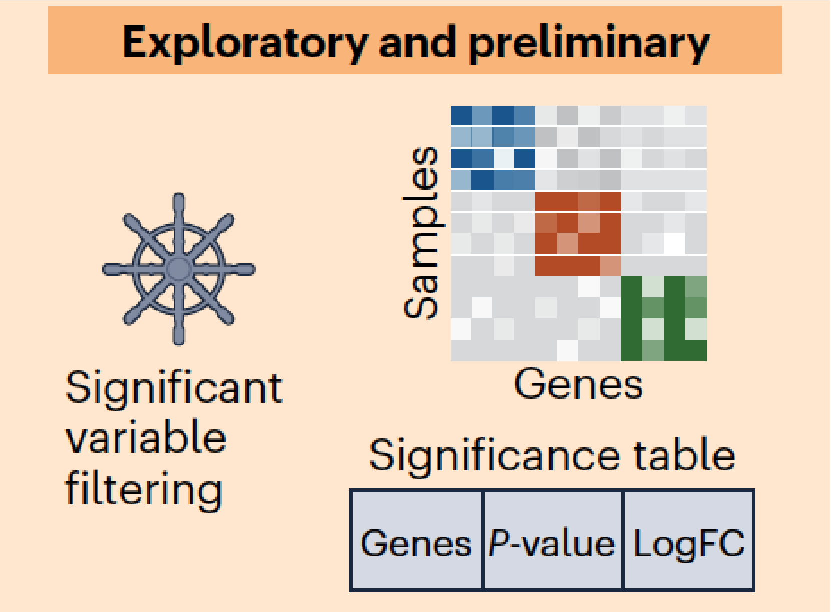
Ad hoc

Literally: *“for this”* or *“for this purpose”*
From *ad* (to, for) + *hoc* (this).





Result interpretation



Exploratory and preliminary

Significant variable filtering

Significance table

Genes	*P*-value	LogFC

https://doi.org/10.1038/s41587-025-02852-0

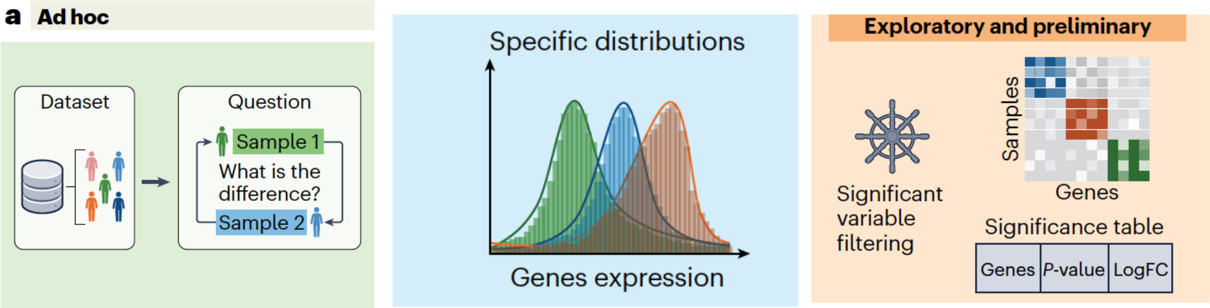
Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

Ad hoc

Literally: *“for this”* or *“for this purpose”*

From *ad* (to, for) + *hoc* (this).



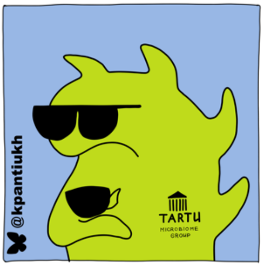
Pros

- Simple and fast
- Transparent statistical outputs
- Ideal for exploratory analyses

Cons

- Highly sensitive to parameters
- Limited robustness across datasets
- Reproducibility issues

https://doi.org/10.1038/s41587-025-02852-0



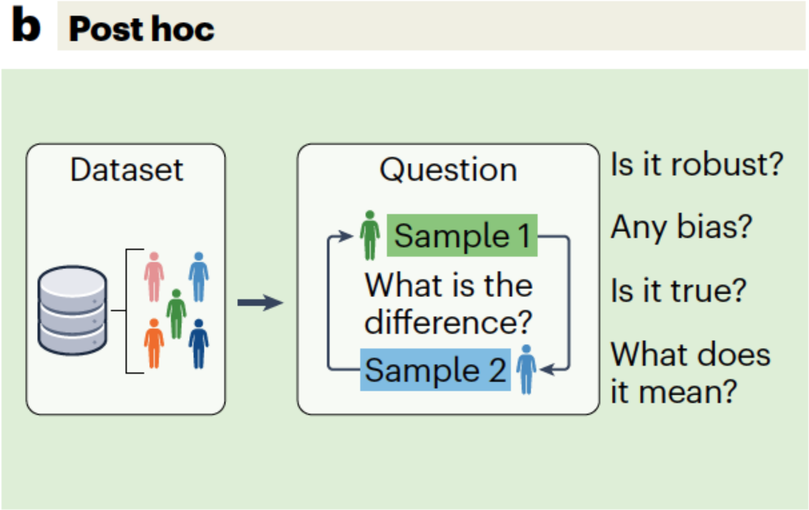
Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

Post hoc

Literally: *“after this”*
From *post* (after) + *hoc* (this).

- overcome the limitations of direct, single-target analyses by addressing additional questions, such as whether a result is robust or biased according to the experimental design or dataset selection



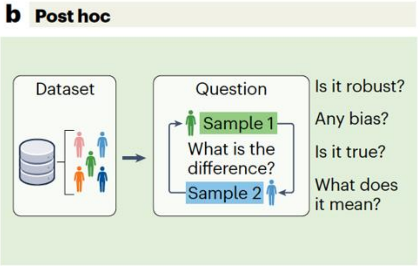
https://doi.org/10.1038/s41587-025-02852-0

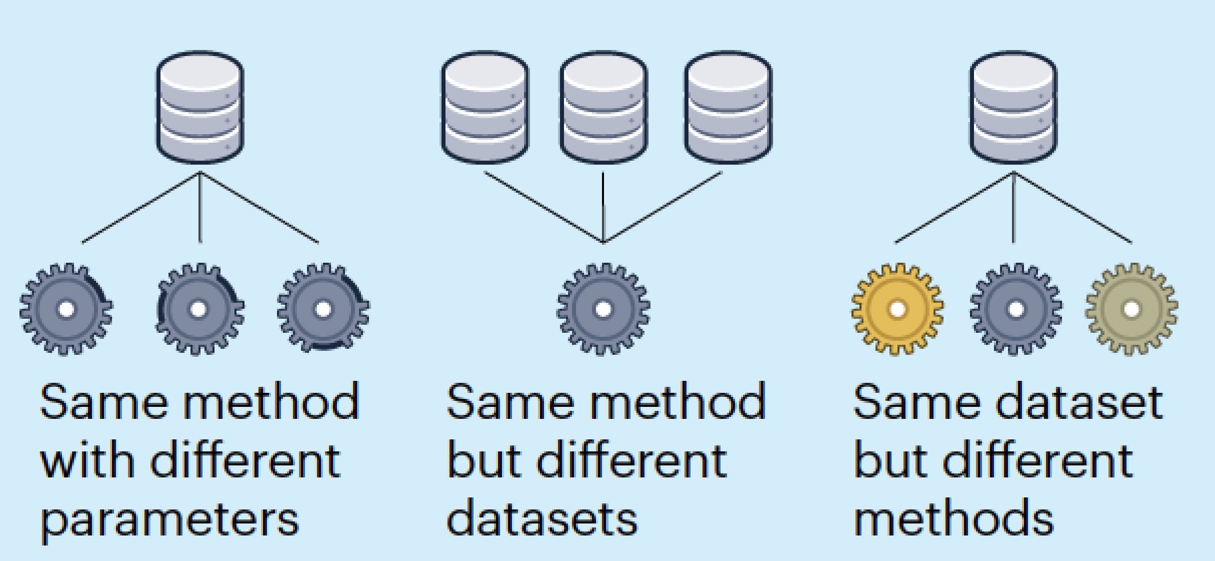
Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

Post hoc

Literally: *"after this"*
From *post* (after) + *hoc* (this).




Same method with different parameters
Same method but different datasets
Same dataset but different methods

- integrate outputs from multiple analyses
- applying the same method to different datasets
- or combining different analytical methods in one study

https://doi.org/10.1038/s41587-025-02852-0

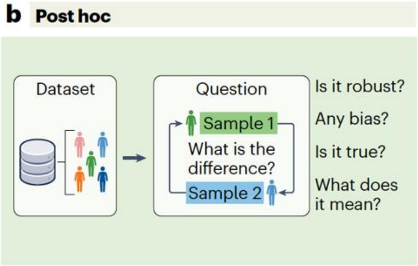
Bioinformatic analysis

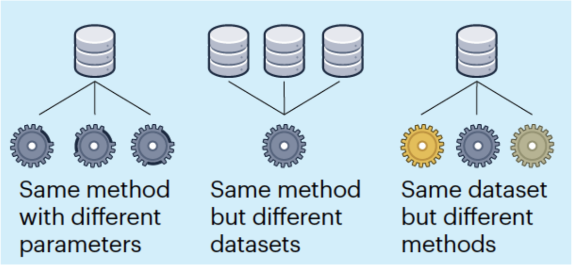
3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

Post hoc

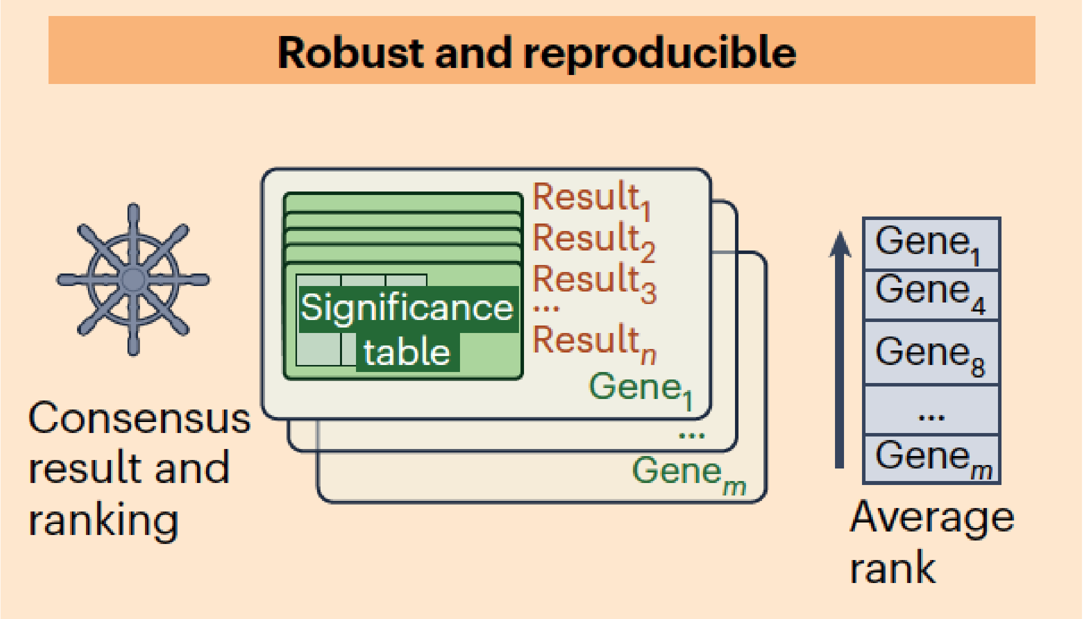
Literally: *“after this”*

From *post* (after) + *hoc* (this).

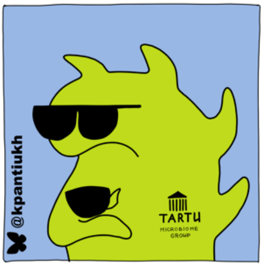




Same method with different parameters
Same method but different datasets
Same dataset but different methods



https://doi.org/10.1038/s41587-025-02852-0



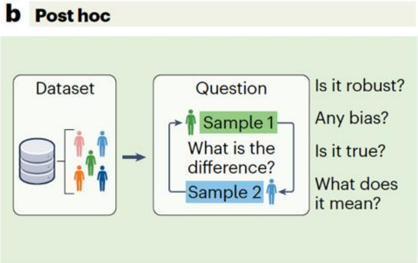
Bioinformatic analysis

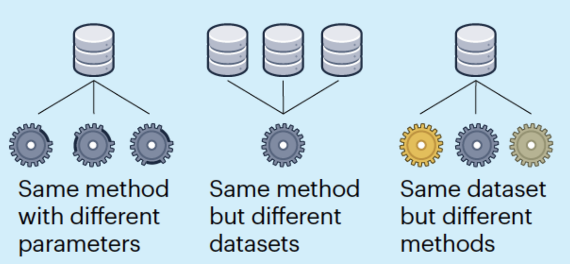
3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

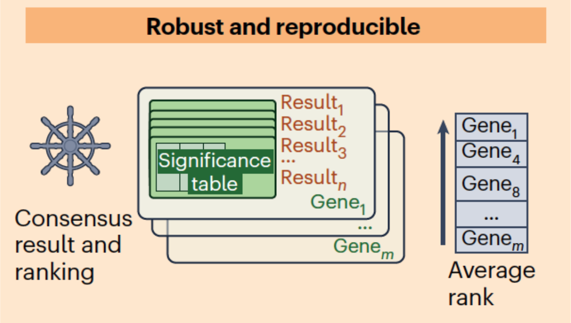
Post hoc

Literally: *“after this”*

From *post* (after) + *hoc* (this).




Same method with different parameters
Same method but different datasets
Same dataset but different methods



Pros

- Integrates results from multiple analyses
- Enhances robustness and reproducibility

Cons

- Increased computational complexity
- Dependent on retrospective data integration
- May inherit biases from initial analyses

https://doi.org/10.1038/s41587-025-02852-0

Bioinformatic analysis

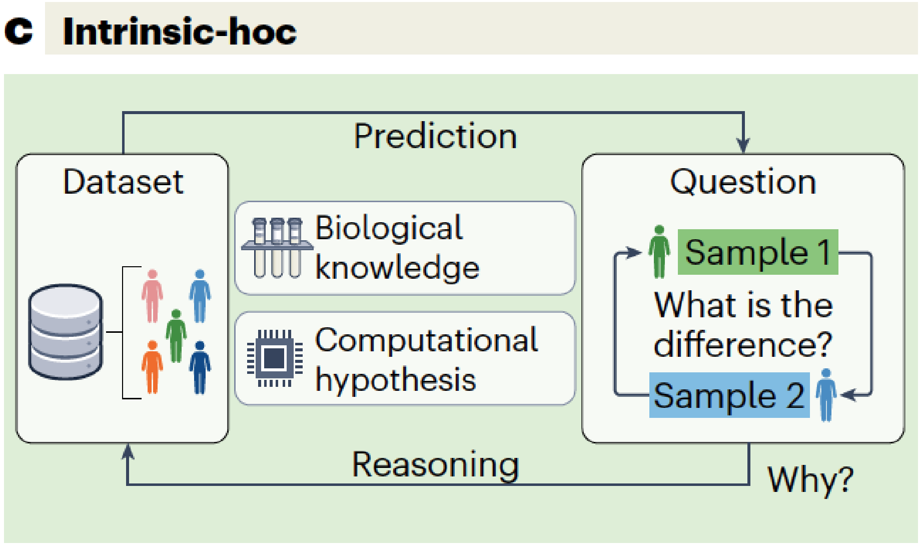
3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

intrinsic-hoc

modern, invented term, inspired by Latin expressions “ad hoc” and “post hoc”

Intrinsic-hoc strategies prioritize **understanding over raw predictive power**

Biology aware models



https://doi.org/10.1038/s41587-025-02852-0

Bioinformatic analysis

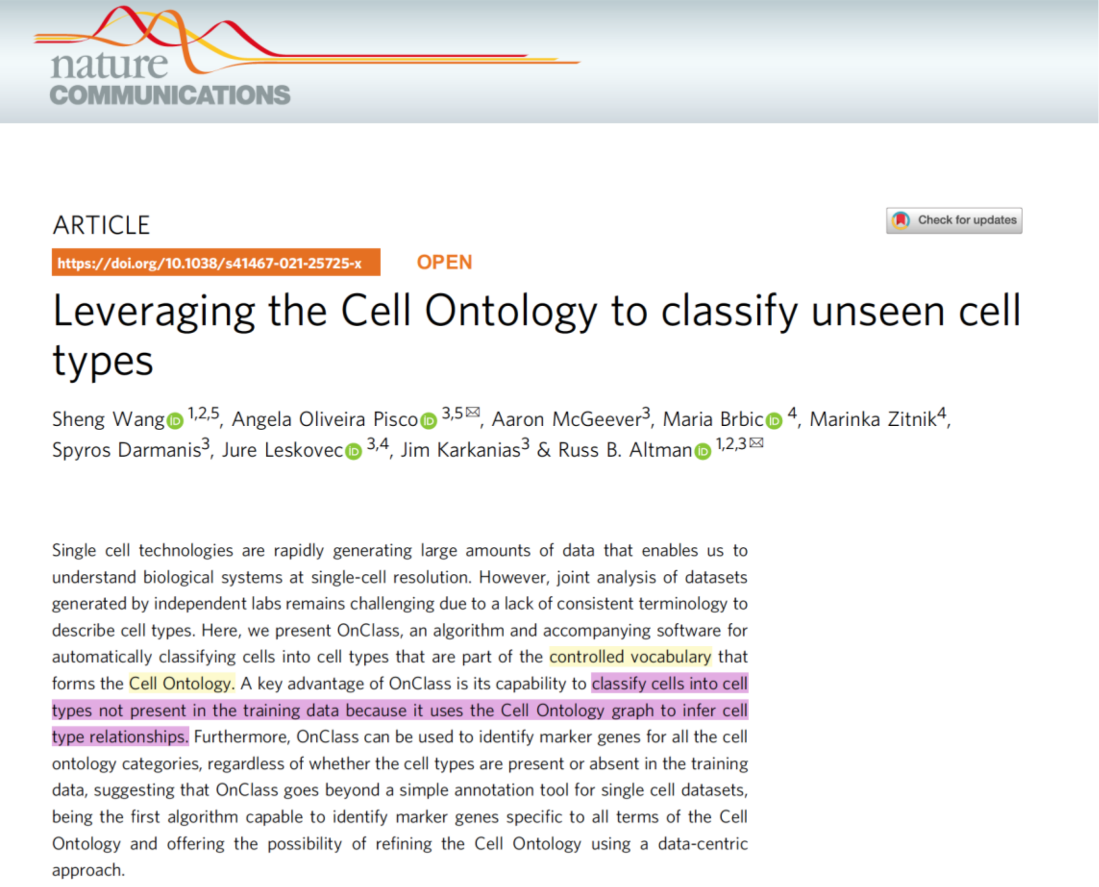
3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

intrinsic-hoc

modern, invented term, inspired by Latin expressions “ad hoc” and “post hoc”

Intrinsic-hoc strategies prioritize **understanding over raw predictive power**

Biology aware models
EXAMPLE: cell type classification using cell ontology graph *https://doi.org/10.1038/s41467-021-25725-x*



nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-25725-x OPEN

Leveraging the Cell Ontology to classify unseen cell types

Sheng Wang[1,2,5], Angela Oliveira Pisco[3,5]✉, Aaron McGeever[3], Maria Brbic[4], Marinka Zitnik[4], Spyros Darmanis[3], Jure Leskovec[3,4], Jim Karkanias[3] & Russ B. Altman[1,2,3]✉

Single cell technologies are rapidly generating large amounts of data that enables us to understand biological systems at single-cell resolution. However, joint analysis of datasets generated by independent labs remains challenging due to a lack of consistent terminology to describe cell types. Here, we present OnClass, an algorithm and accompanying software for automatically classifying cells into cell types that are part of the controlled vocabulary that forms the Cell Ontology. A key advantage of OnClass is its capability to classify cells into cell types not present in the training data because it uses the Cell Ontology graph to infer cell type relationships. Furthermore, OnClass can be used to identify marker genes for all the cell ontology categories, regardless of whether the cell types are present or absent in the training data, suggesting that OnClass goes beyond a simple annotation tool for single cell datasets, being the first algorithm capable to identify marker genes specific to all terms of the Cell Ontology and offering the possibility of refining the Cell Ontology using a data-centric approach.

https://doi.org/10.1038/s41587-025-02852-0

Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc

intrinsic-hoc

modern, invented term, inspired by Latin expressions “ad hoc” and “post hoc”

Intrinsic-hoc strategies prioritize **understanding over raw predictive power**

Biology aware models

What about microbiome?

- Taxonomy-aware models

Phylogeny-aware distance metrics (e.g., UniFrac) used as inputs to interpretable models

- Functional-group aggregation models

Grouping taxa by oxygen requirement, fermentation type, or bile tolerance

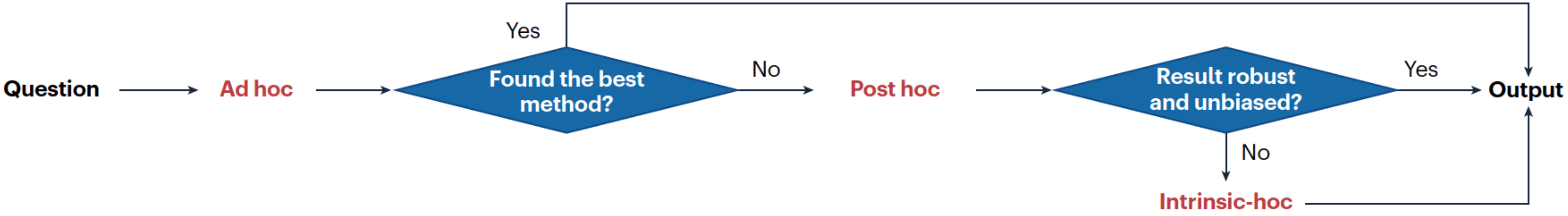
- Pathways structured models

Models with layers correspond to metabolic pathways

https://doi.org/10.1038/s41587-025-02852-0

Bioinformatic analysis

3 metodological strategies – ad hoc, post hoc, intrinsic-hoc



https://doi.org/10.1038/s41587-025-02852-0