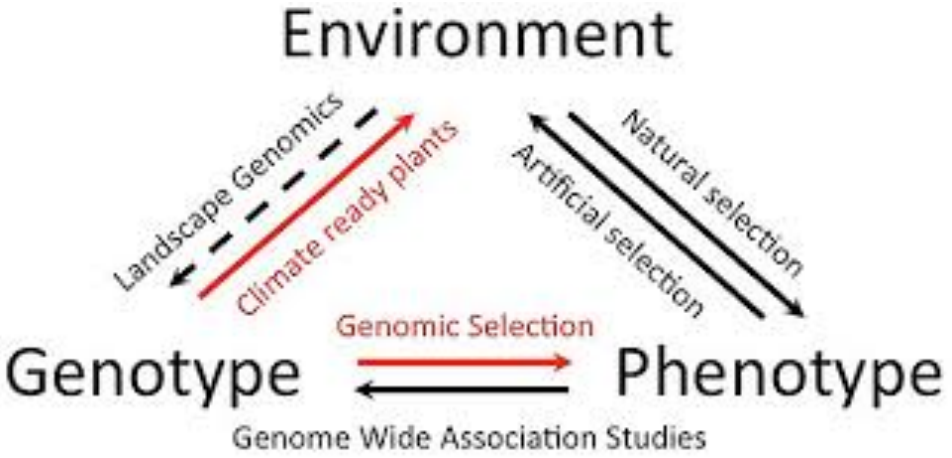
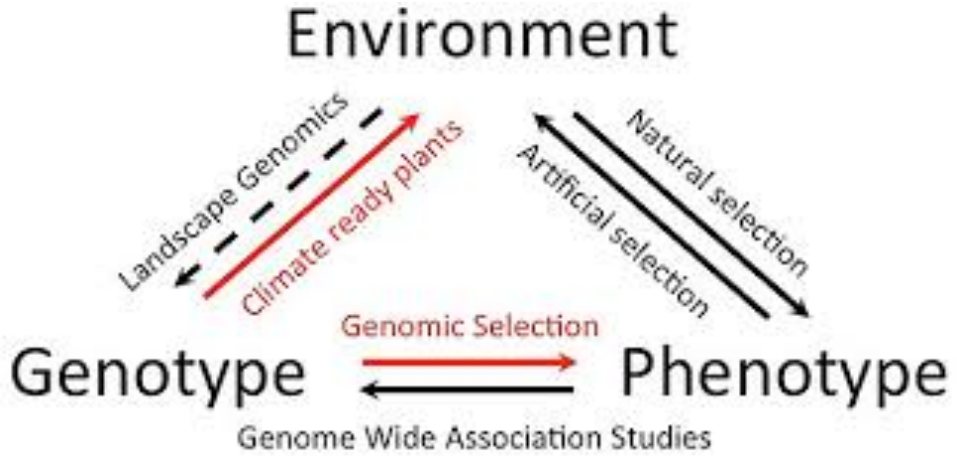
# A multi-focal point of view: Integrated analyses of multi-omics data

# Genotype-Phenotype axis

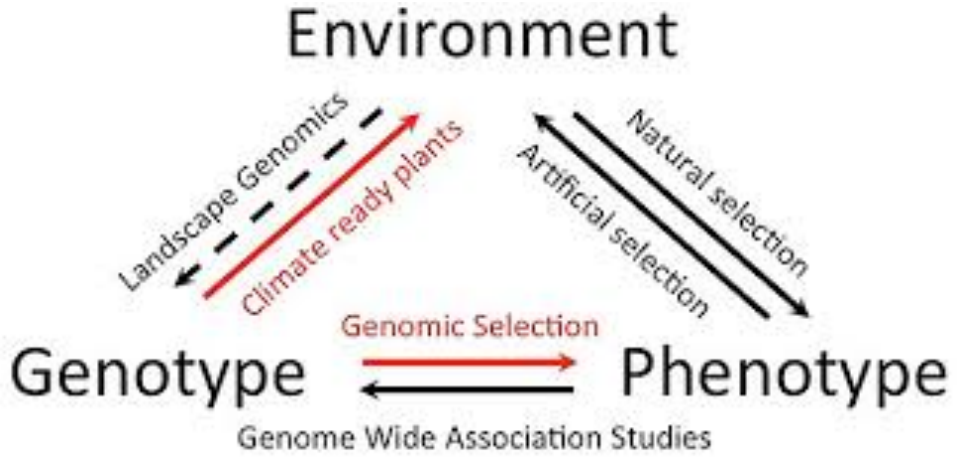


# Genotype-Phenotype axis




Classic model of how phenotypes are shaped by genotypes and environment

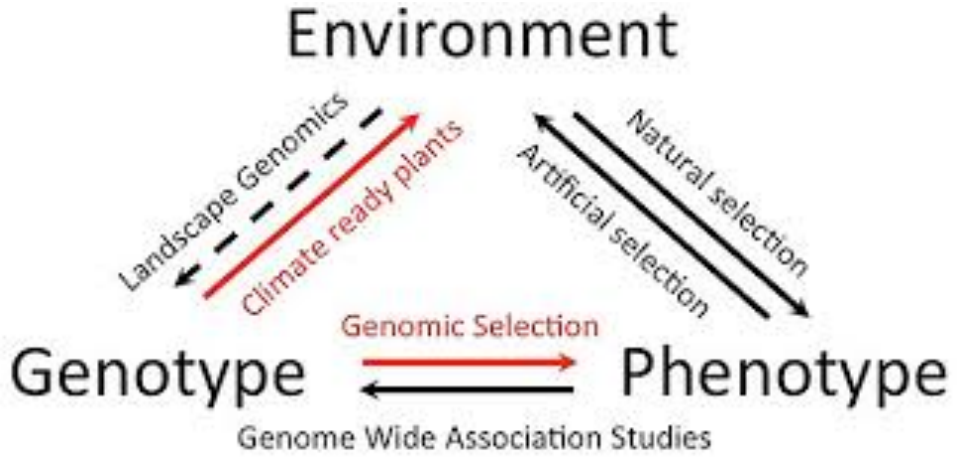
# Genotype-Phenotype axis




Classic model of how phenotypes are shaped by genotypes and environment

The labels on the arrows show processes OR statistical methods between any two measurements.

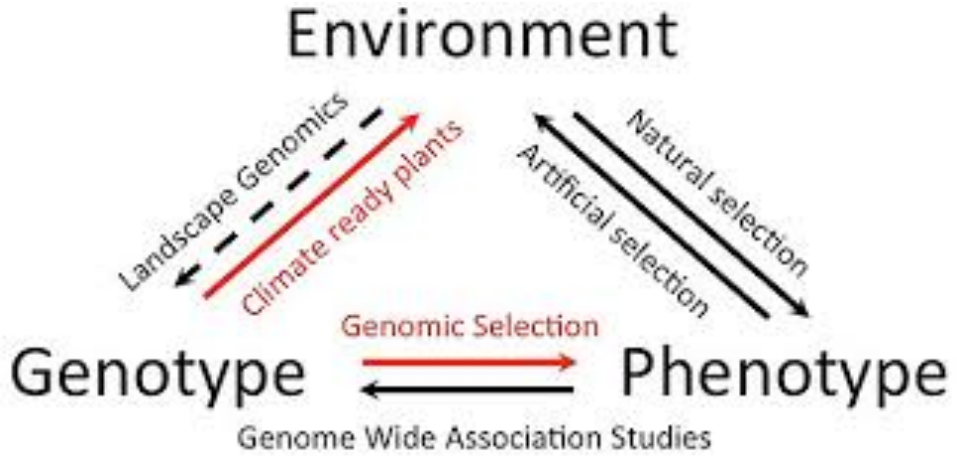
# Genotype-Phenotype axis




Classic model of how phenotypes are shaped by genotypes and environment

Is this a good model for phenotypic variance?

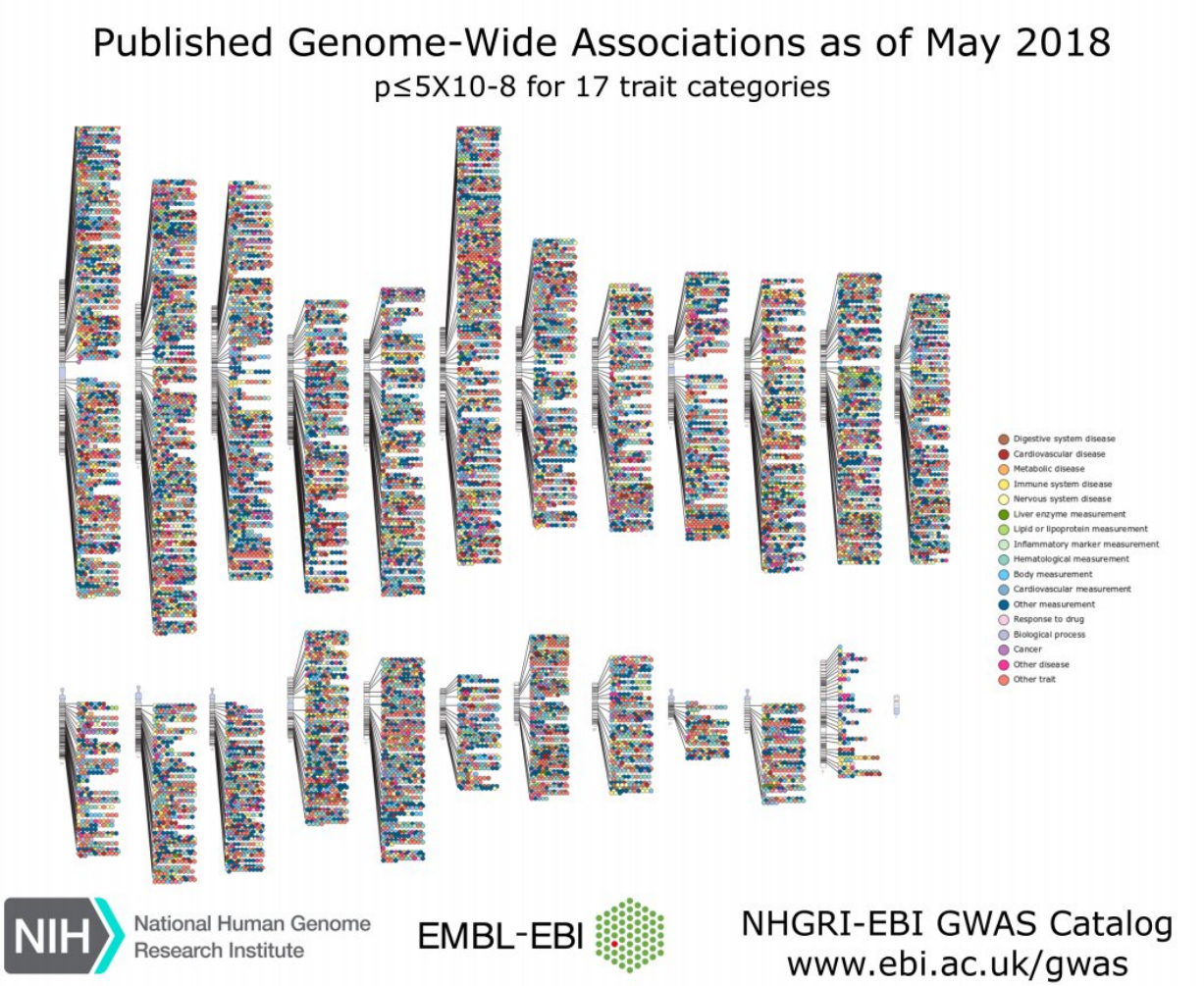
# Genotype-Phenotype axis




Classic model of how phenotypes are shaped by genotypes and environment

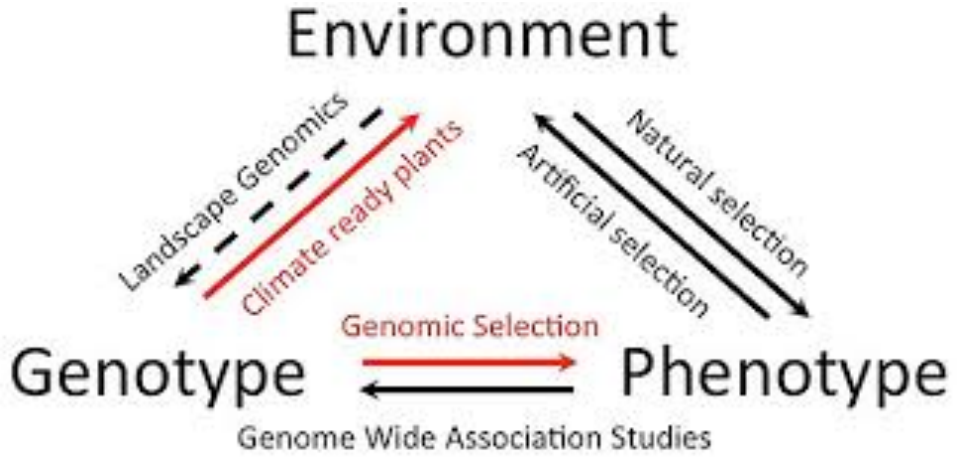
Is this a good model for phenotypic variance?
**Yes, easy to measure the components and fit models!**

# Great success




Catalog of GWAS findings in humans

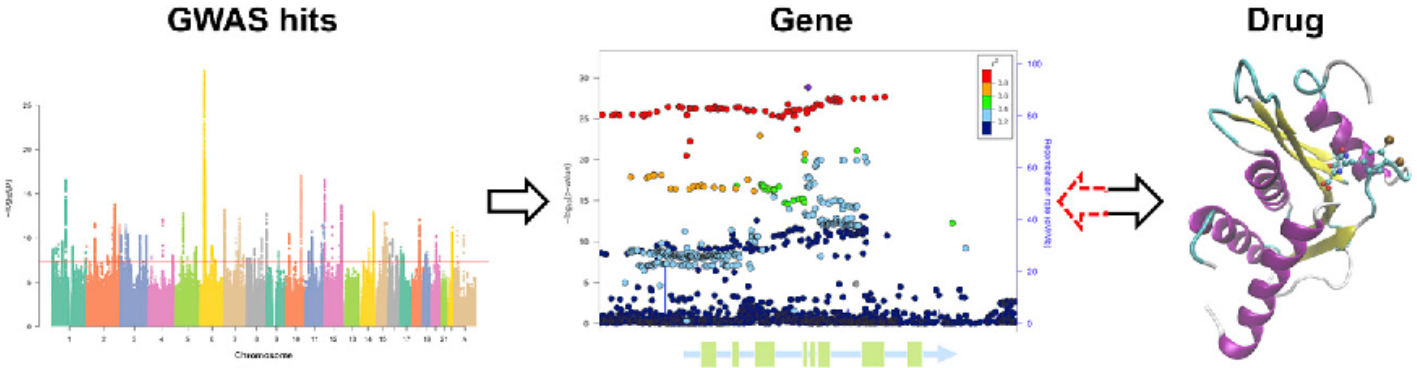
# Genotype-Phenotype axis




Classic model of how phenotypes are shaped by genotypes and environment

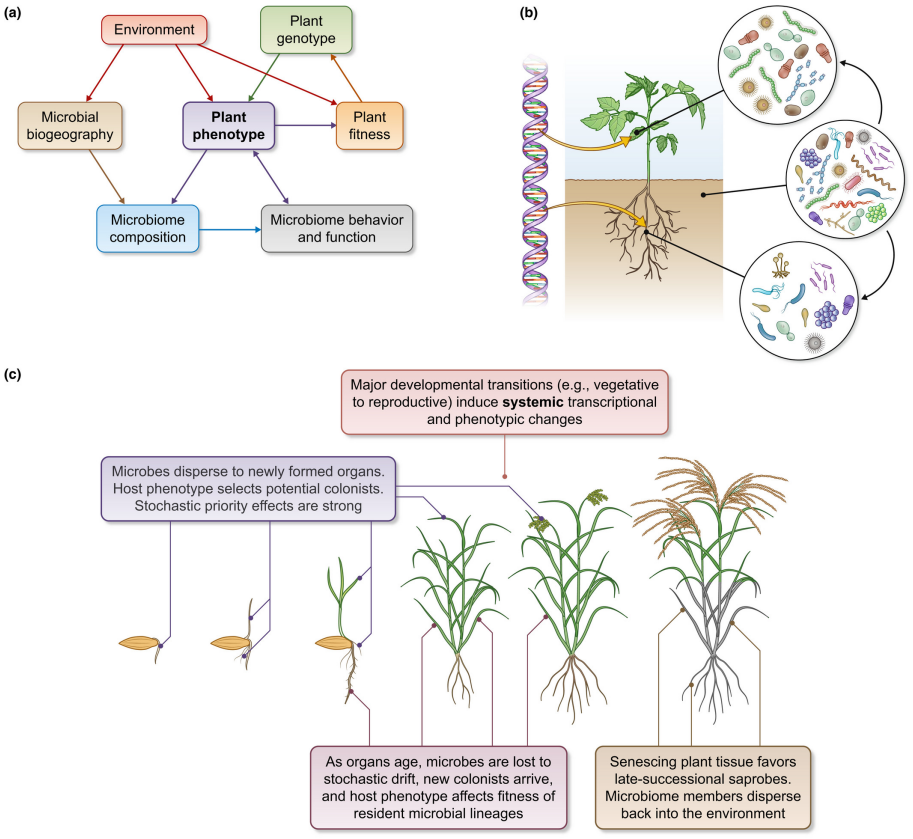
Is this a good model for phenotypic variance?
**No, glosses over biological complexity, and provides very little mechanistic insight.**

# Limited translational research




| Trait | Gene with GWAS hits | Known or candidate drug |
|---|---|---|
| **Type 2 Diabetes** | *SLC30A8/KCNJ11* | ZnT-8 antagonists/Glyburide |
| **Rheumatoid Arthritis** | *PADI4/IL6R* | BB-Cl-amidine/Tocilizumab |
| **Ankylosing Spondylitis(AS)** | *TNFR1/PTGER4/TYK2* | TNF-inhibitors/NSAIDs/fostamatinib |
| **Psoriasis(Ps)** | *IL23A* | Risankizumab |
| **Osteoporosis** | *RANKL/ESR1* | Denosumab/Raloxifene and HRT |
| **Schizophrenia** | *DRD2* | Anti-psychotics |
| **LDL cholesterol** | *HMGCR* | Pravastatin |
| **AS, Ps, Psoriatic Arthritis** | *IL12B* | Ustekinumab |

# Understanding biological processes – beyond the target organism




Wagner et al. 2021

# The human microbiome affects our health and disease

**The gut microbiome is the largest and richest microbial community of the human body**

1000+ bacterial species in the human gut

100 times more microbial than human genes

Microbiomes between individuals differ by 80-90%

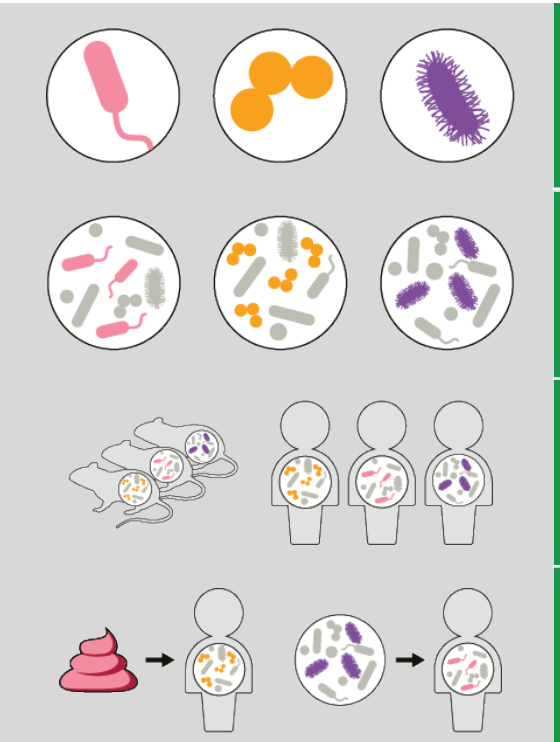
Environmental and host factors shape microbiome

**Gut microbiome is a prime example to study microbial ecosystems**



EMBL

# The grand challenge of human microbiome research



Understand functional diversity of microbial species and strains

Dissect the interactions and properties of microbial communities

Investigate the interplay of microbial communities and their environment

Modulate microbiomes and their interactions

**From microbes to ecosystems and their modulation.**



EMBL

# Thousands of unknown species discovered

## The unified human gut genome (UHGG) catalogue

~12,000 human gut metagenomic samples

↓ *Assembly and binning*

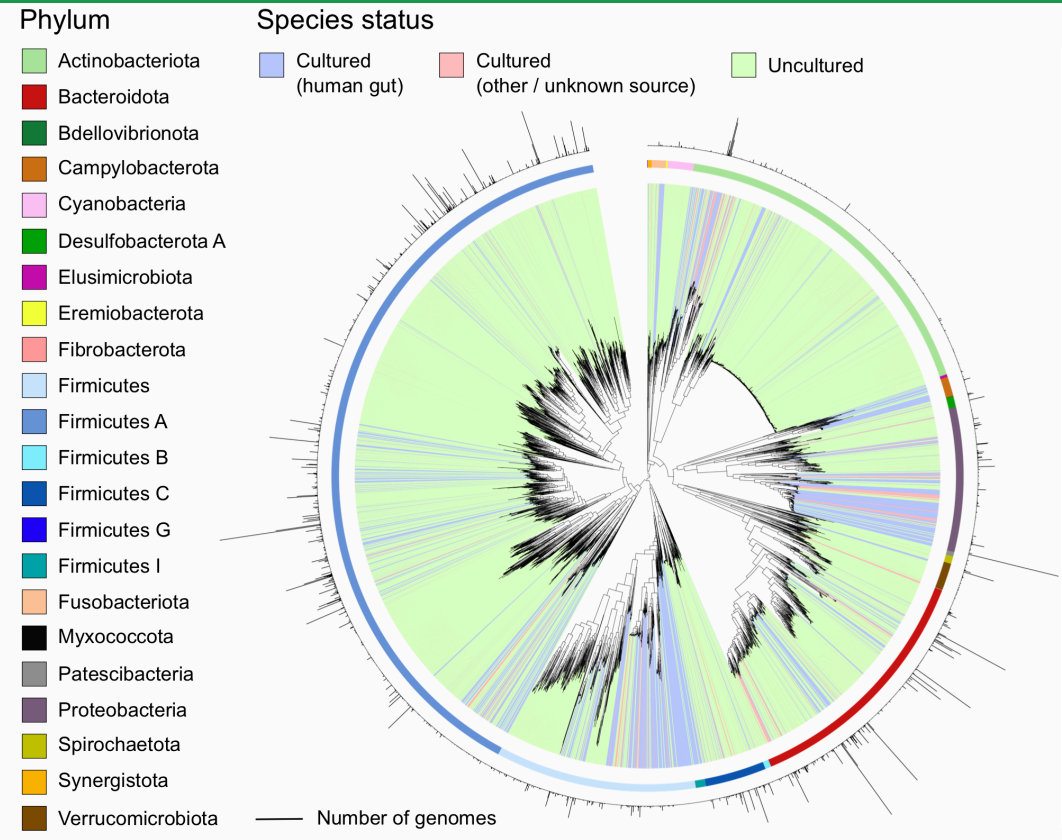
**1,952 species**

↓ *Merging Human Gut Collections*

**4,644 species** (~80% uncultured)

The most comprehensive catalogue of human gut bacteria to date

*210,000 genomes,170 million proteins*




*Almeida et al. Nature 2019*
*Almeida et al. Nature Biotech 2020*

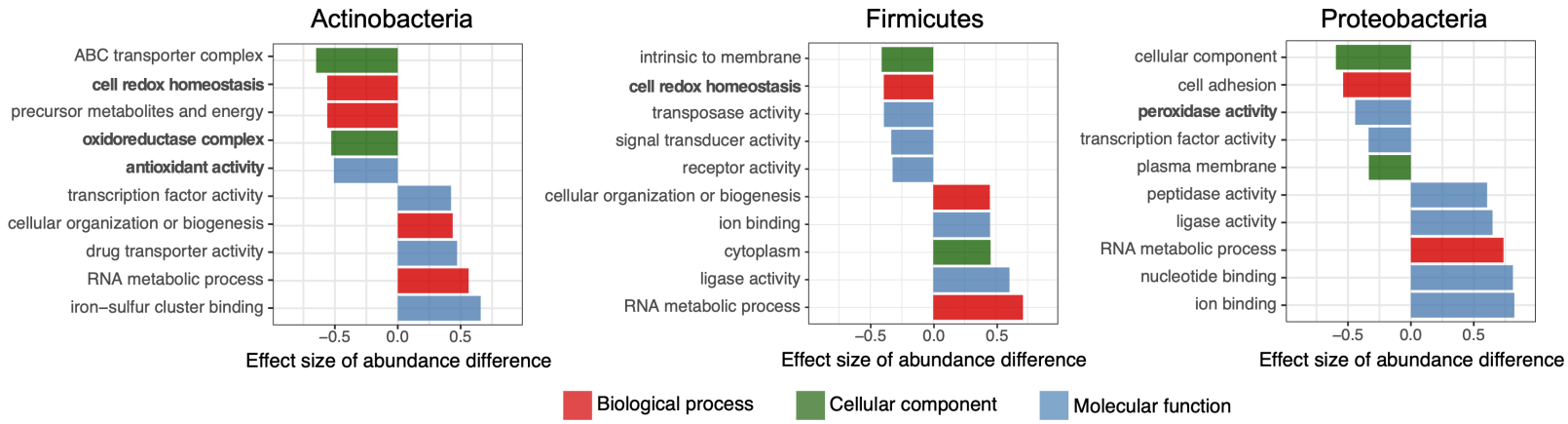


EMBL

# Why have we been missing these MAG?

## Functional analysis gives us some broad clues

Uncultured species vs Cultured species

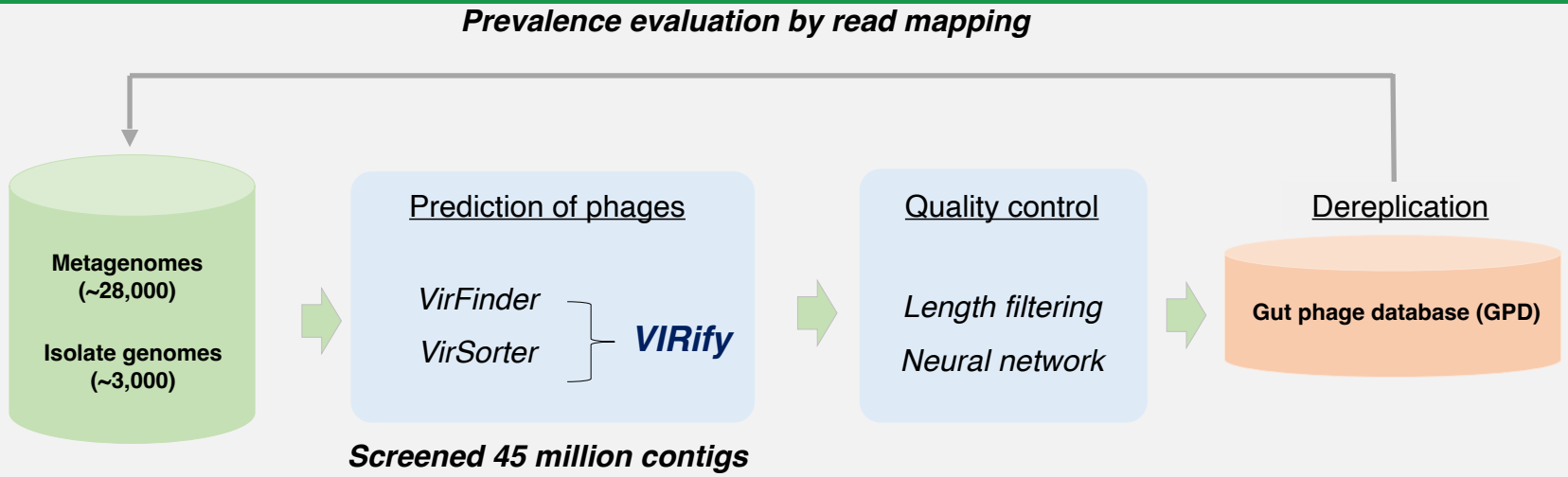



Genes involved in **oxygen tolerance** (higher sensitivity to oxygen)

EMBL-EBI

# What else do these (meta-)genomes contain?

## Pipelines for identifying and taxonomically annotating phage




We generated >**142,000** high-quality genomes of human gut bacteriophages

*Camarillo-Guerrero et al, Cell 2021*

**VIRify: an integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden Markov models**

Guillermo Rangel-Pineros, Alexandre Almeida, Martin Beracochea, Ekaterina Sakharova, Manja Marz, Alejandro Reyes Muñoz, Martin Hölzer, Robert D. Finn

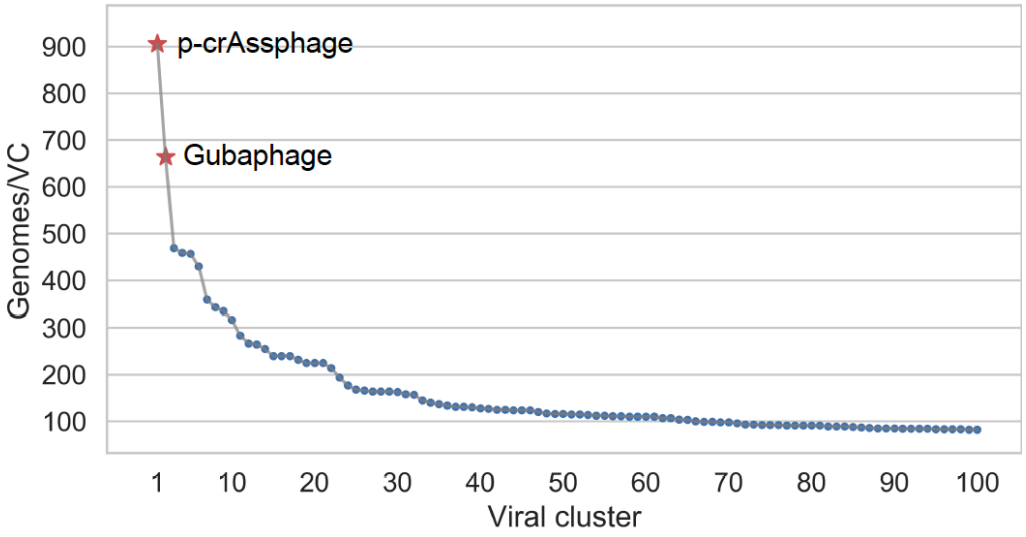
**doi:** https://doi.org/10.1101/2022.08.22.504484

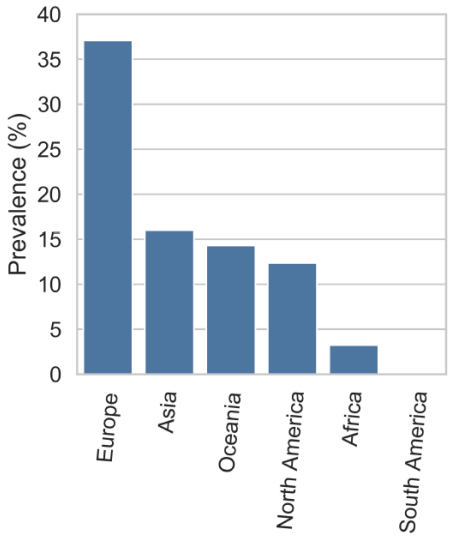


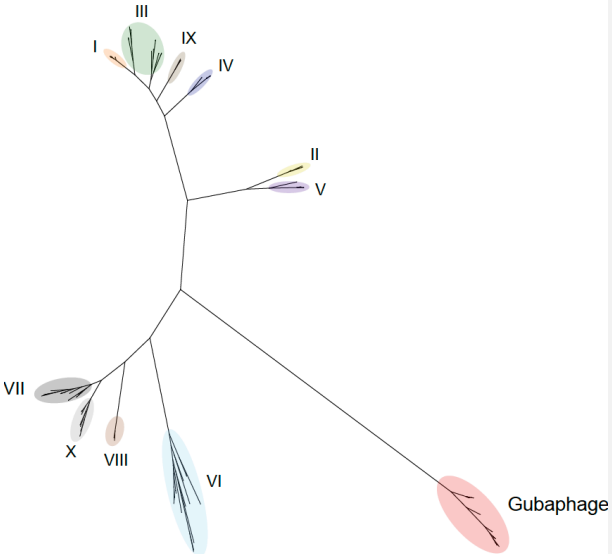
EMBL

# Discovery of a highly prevalent phage

## Gubaphage found at a comparable prevalence to crAssphage




Top 100 viral clusters (prevalence)



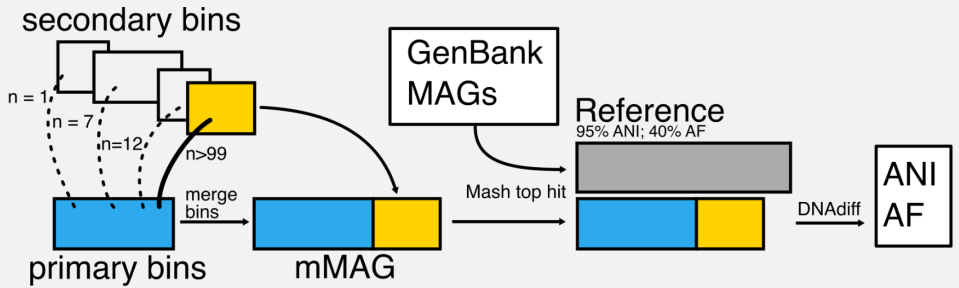
EMBL

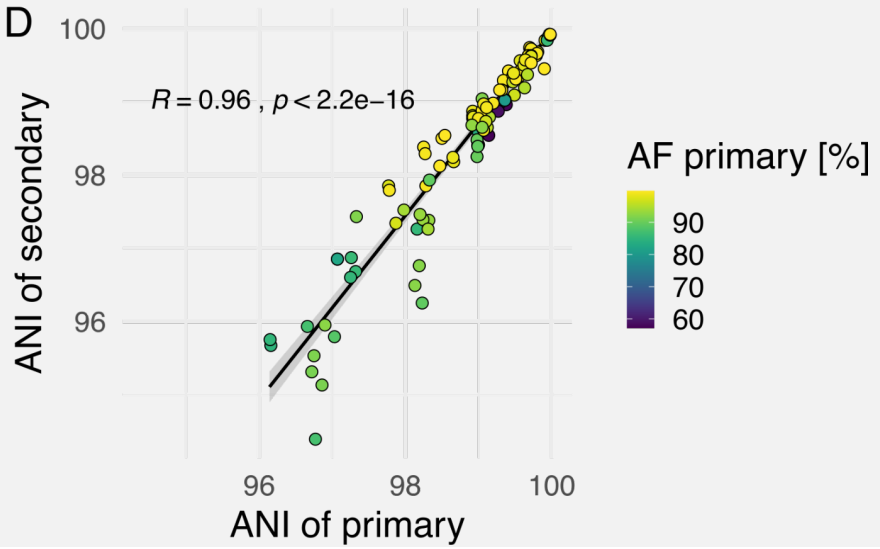
# EukCC - bin merging functionality

EukCC can :

(i) Accurately estimate completeness and contamination of Eukaryotic MAGs
(ii) detect split genomes and create merged MAGs (mMAG)

Only bins combine when increase in completeness is greater than 5 * contamination & increase completeness >10%

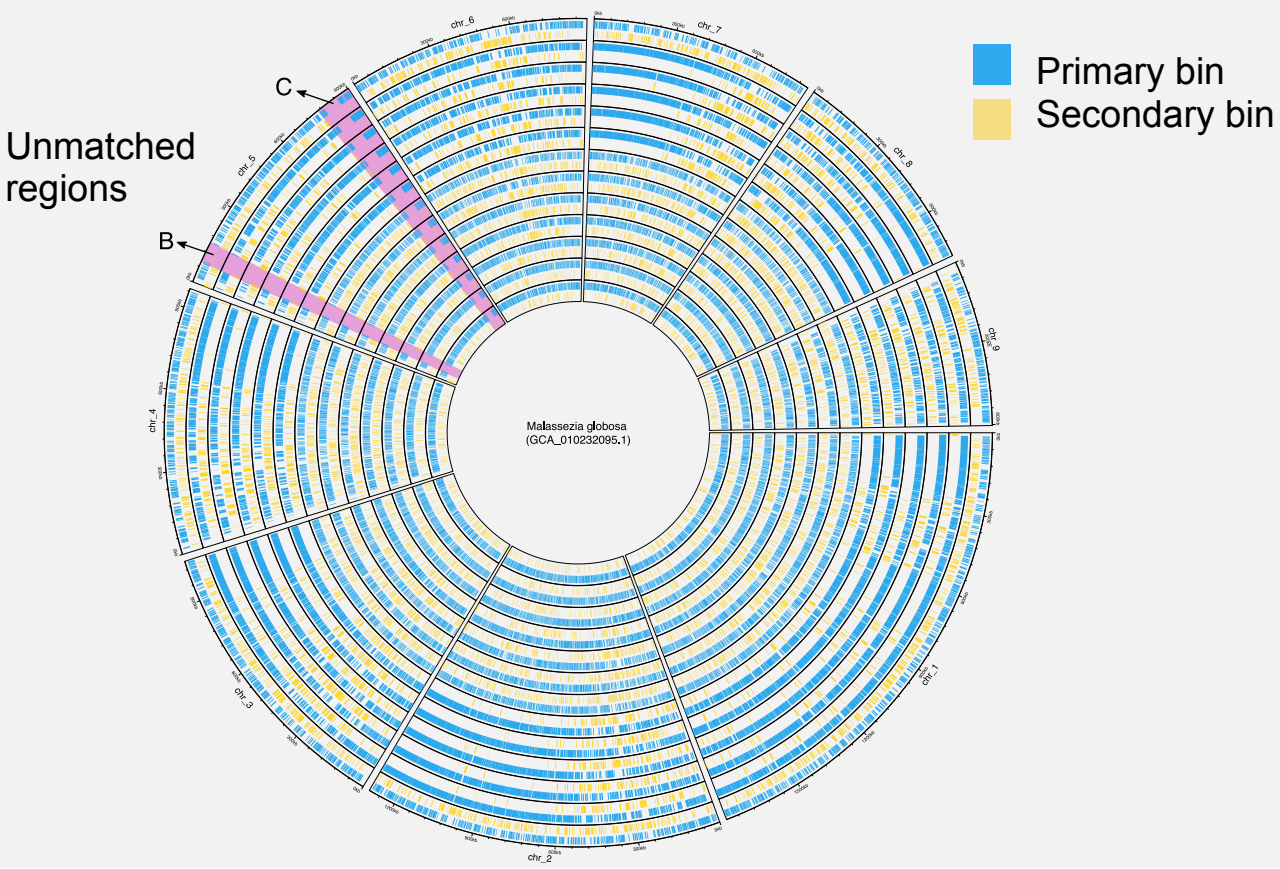





EMBL

# Referene *M. globosa* genome vs mMAGs

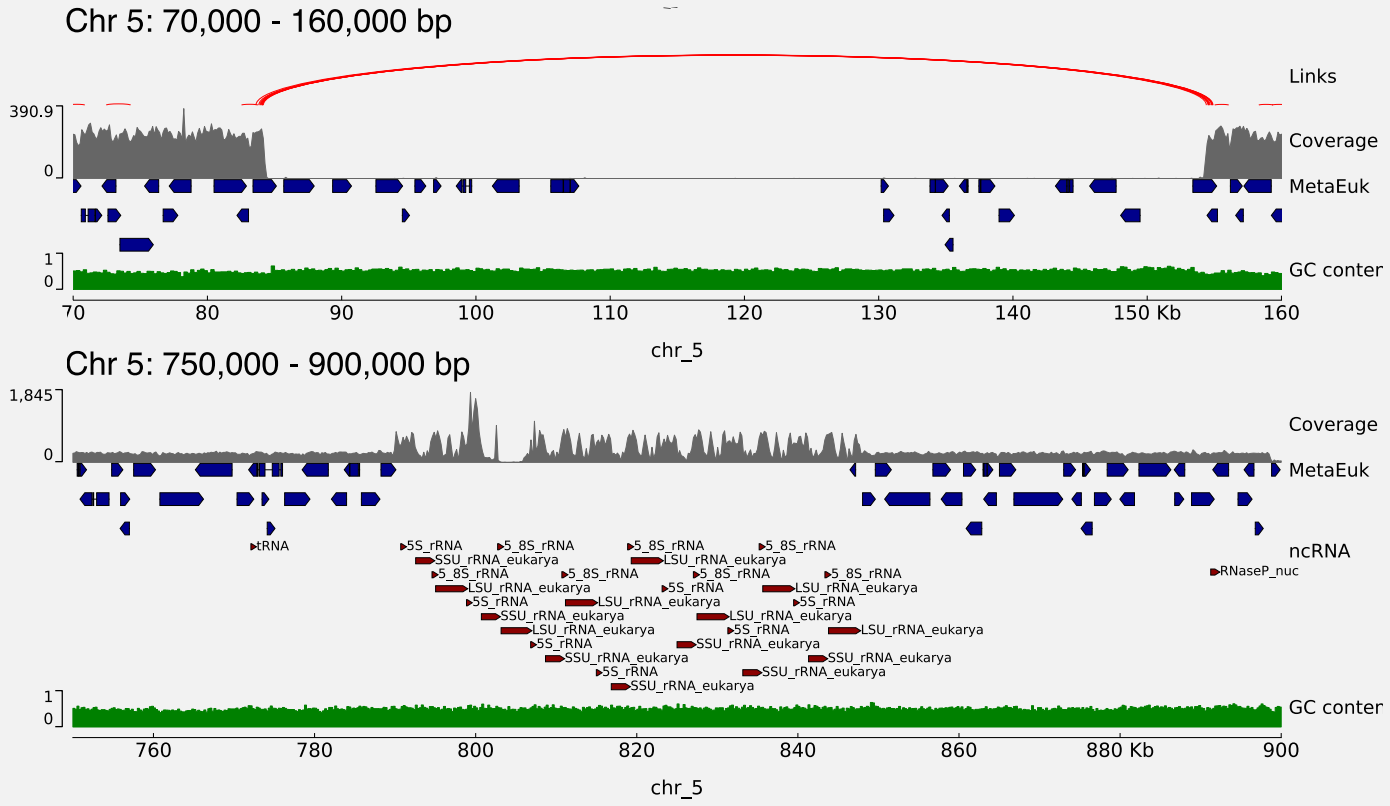



- 13 mMAGs compared to the reference genome
- Primary and secondary bins interleaved, with consistency
- Two consistent gaps

MGnify

EMBL-EBI

# “Missing” genomic features in detail

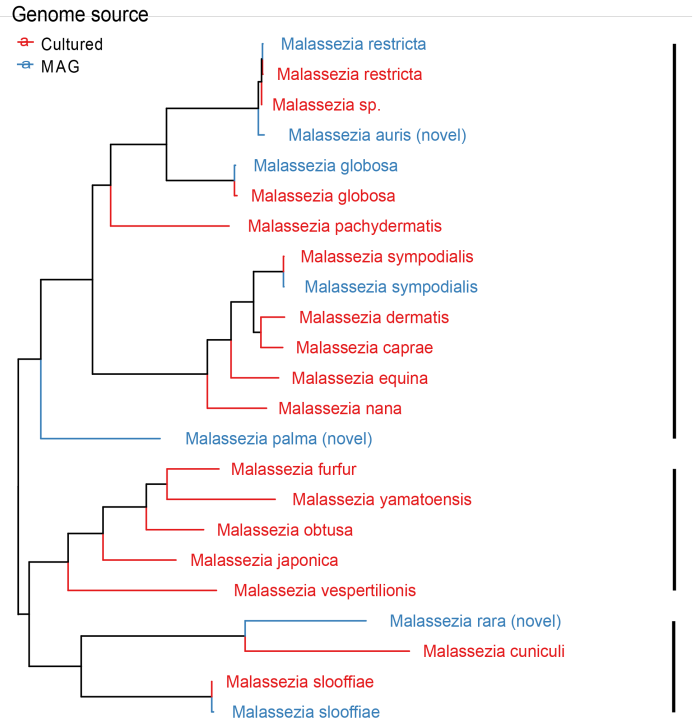


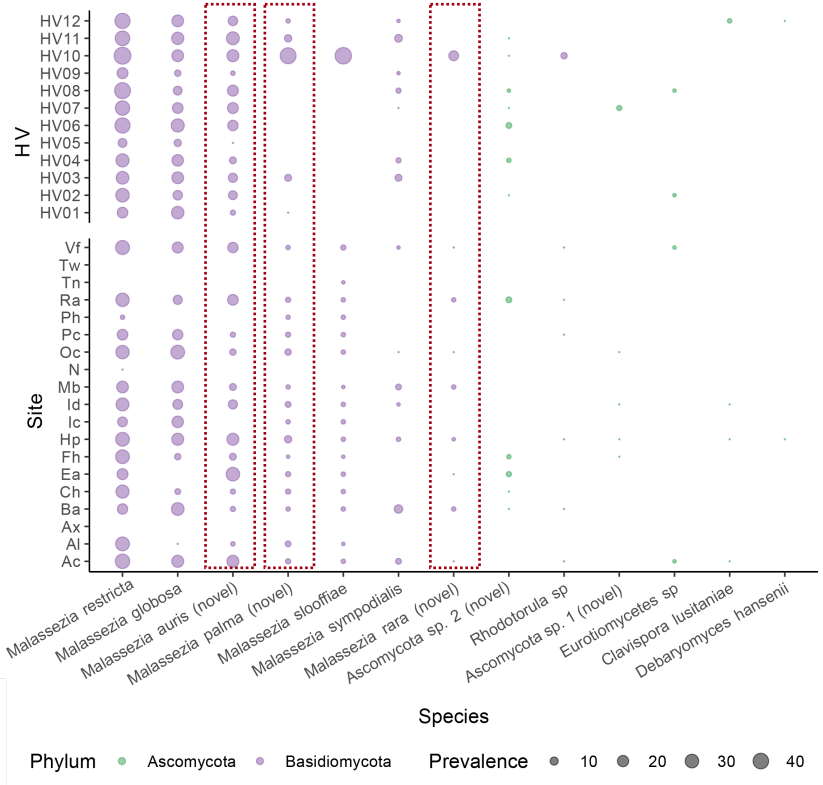

MGnify

EMBL-EBI

# Novel *Malassezia* species on the skin

## Excellent environment for comparing MAGs to references

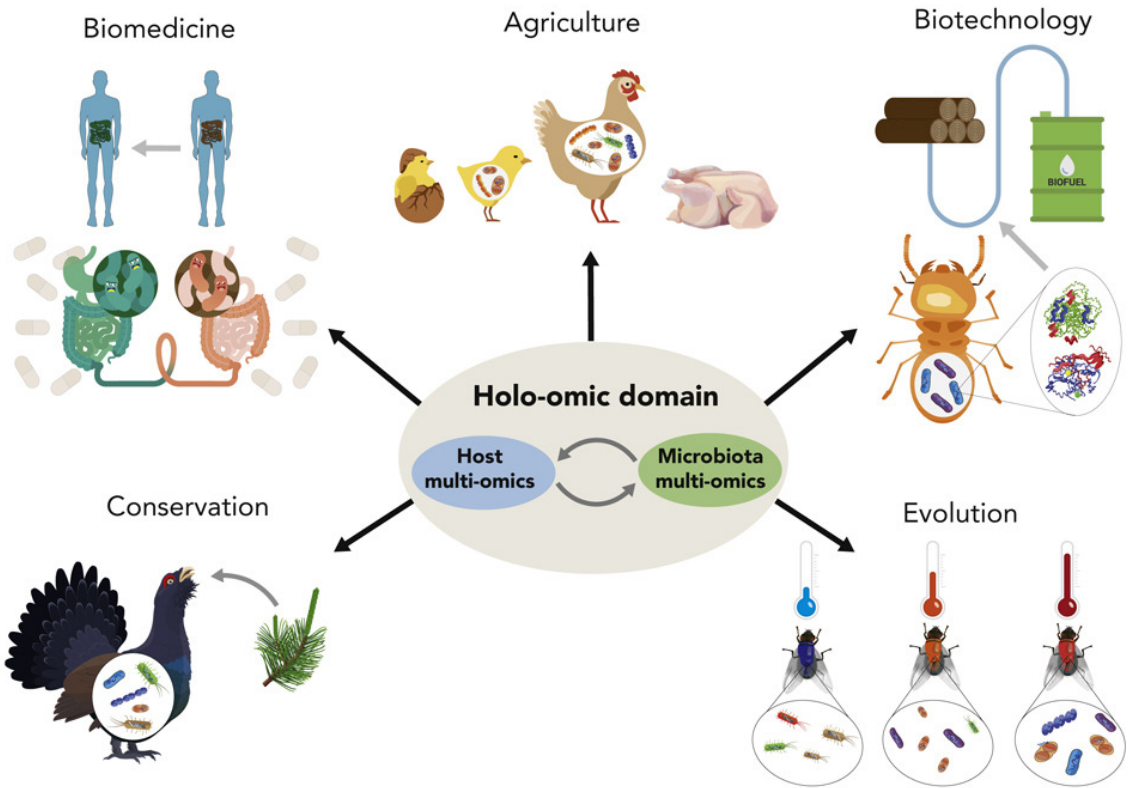



*Kashaf et al, Nature Micro, 2022*

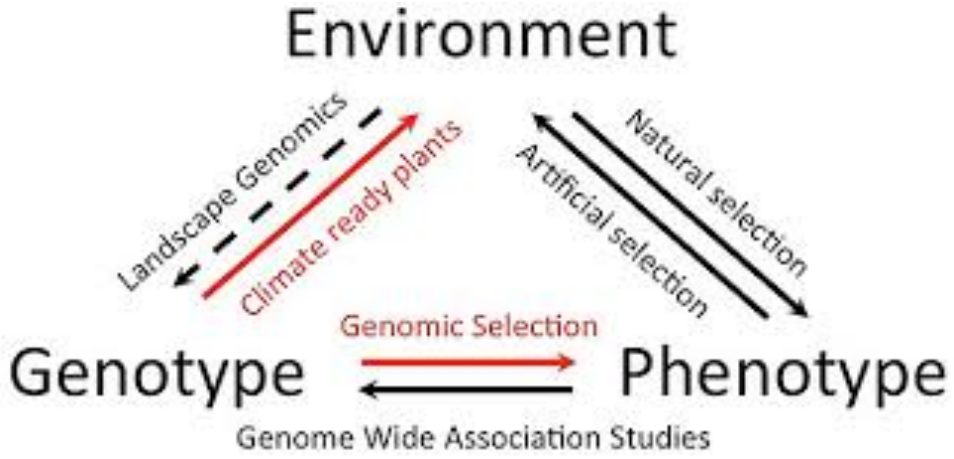
*Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions.*

# Understanding biological processes – hologenomic framework

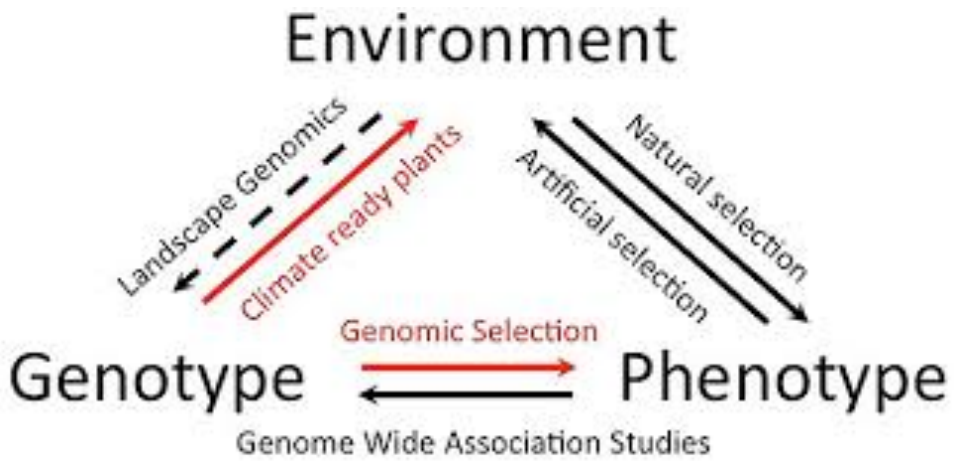


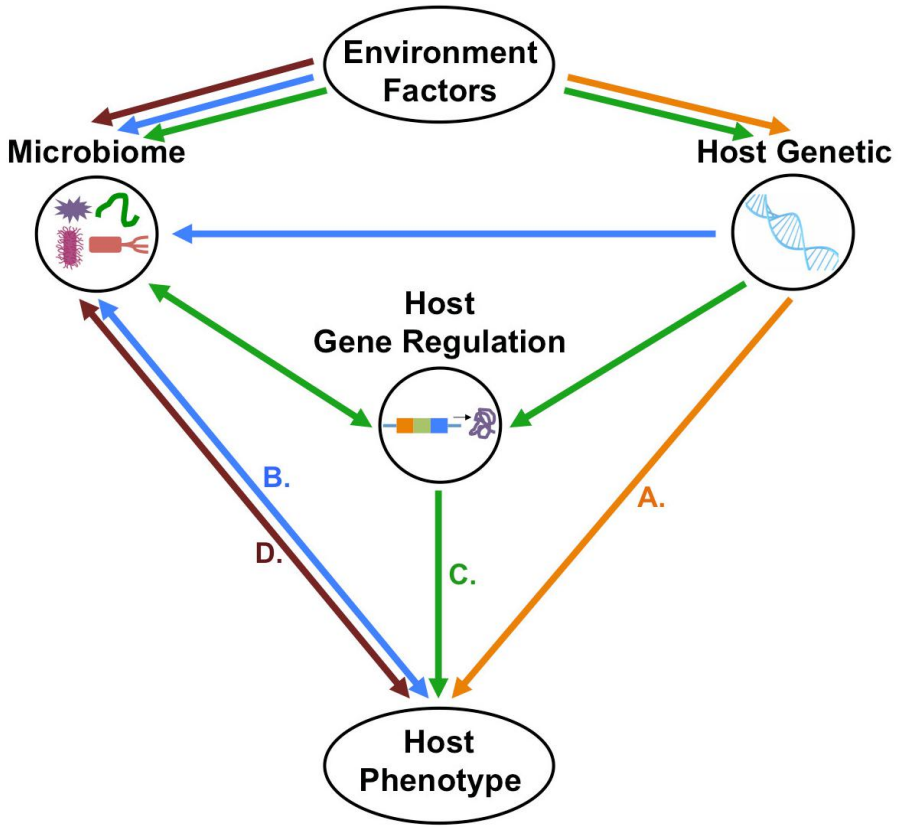

Nyholm et al 2020

# “Central dogma” of genotype-phenotype association



# ~~"Central dogma" of~~ genotype-phenotype association



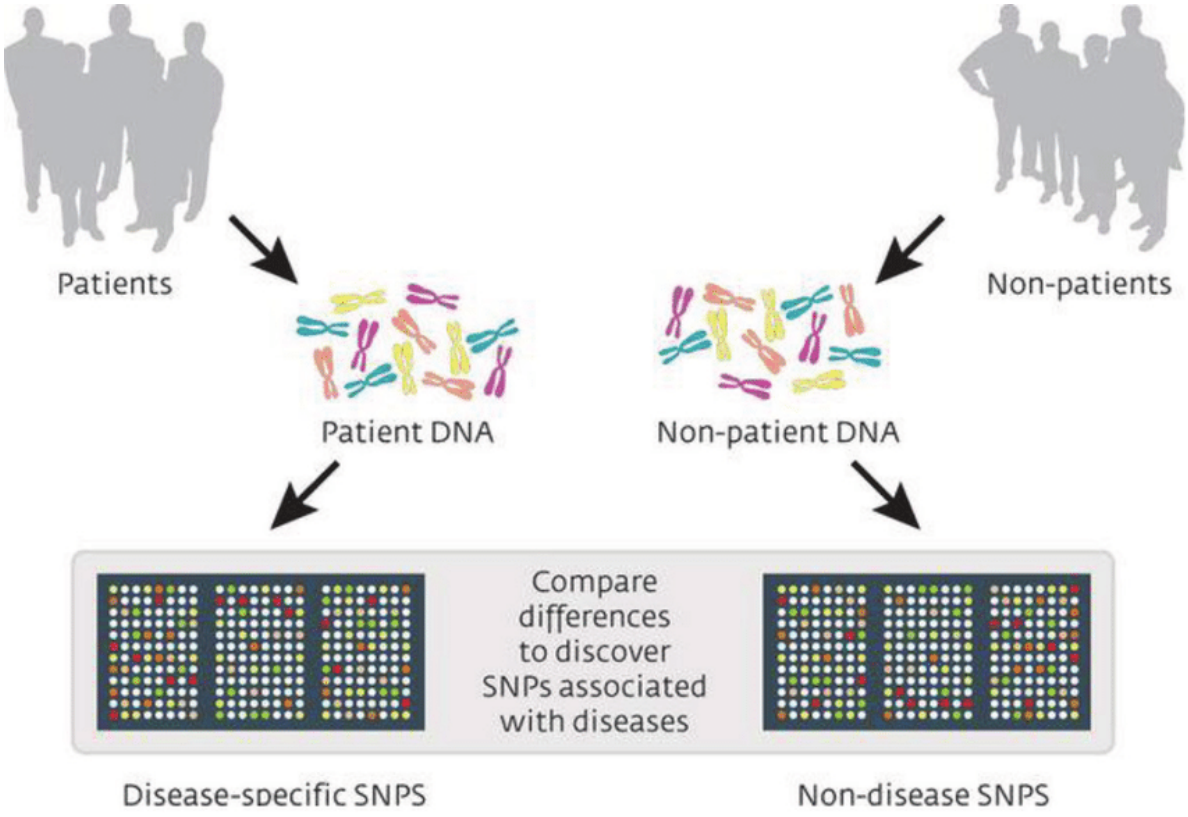




Awany et al. 2019

# Classical association studies

- How to find genotypes that change phenotypes?
  - Most successful are association studies
  - Correlation between genotypes and phenotype

# Closer look at association studies



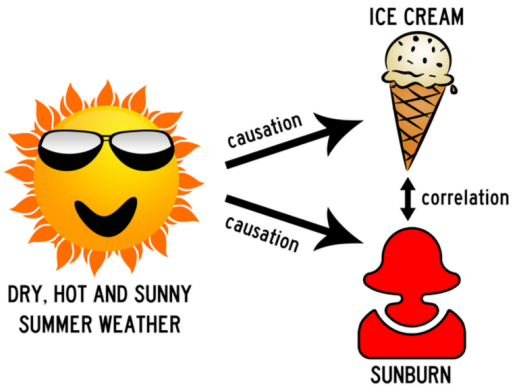

Find the variants that have segregate with the phenotype

- Case control (disease) status
- Anthropometric traits – height, weight etc.

Book: Genetic Diversity and Disease Susceptibility

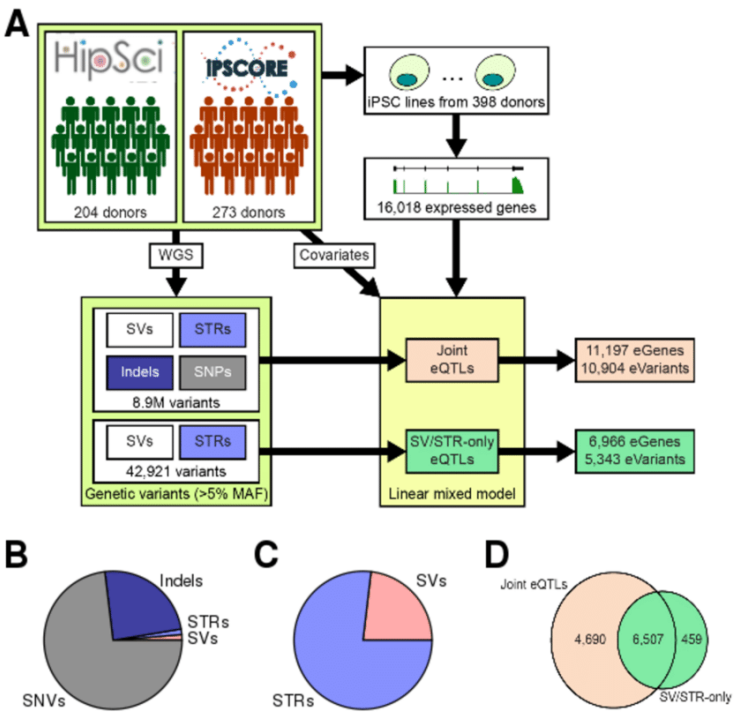
# Shortcomings of association studies

- Mostly correlation based
  - 1-on-1 analysis – GWAS, eQTL, MWAS, MGWAS ...
  - No mechanistic models
  - “Correlation is not causation”

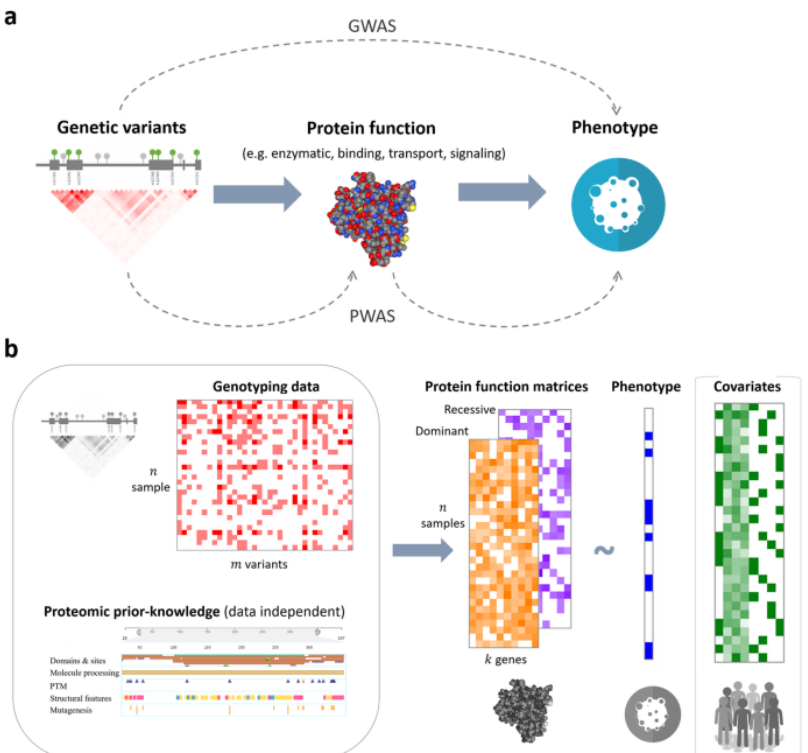


# Getting to causal pathways

- Collect intermediate biological data

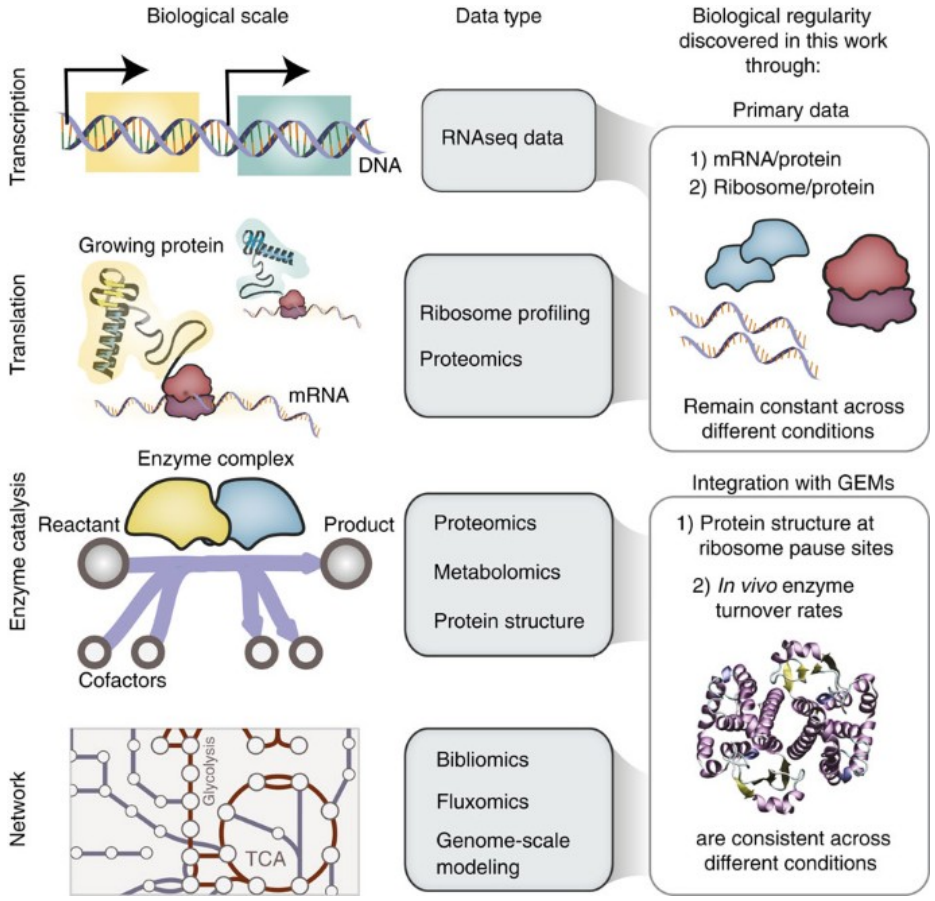



Jakubosky et al 2010




Brandes et al. 2020

# The multi-omics scheme

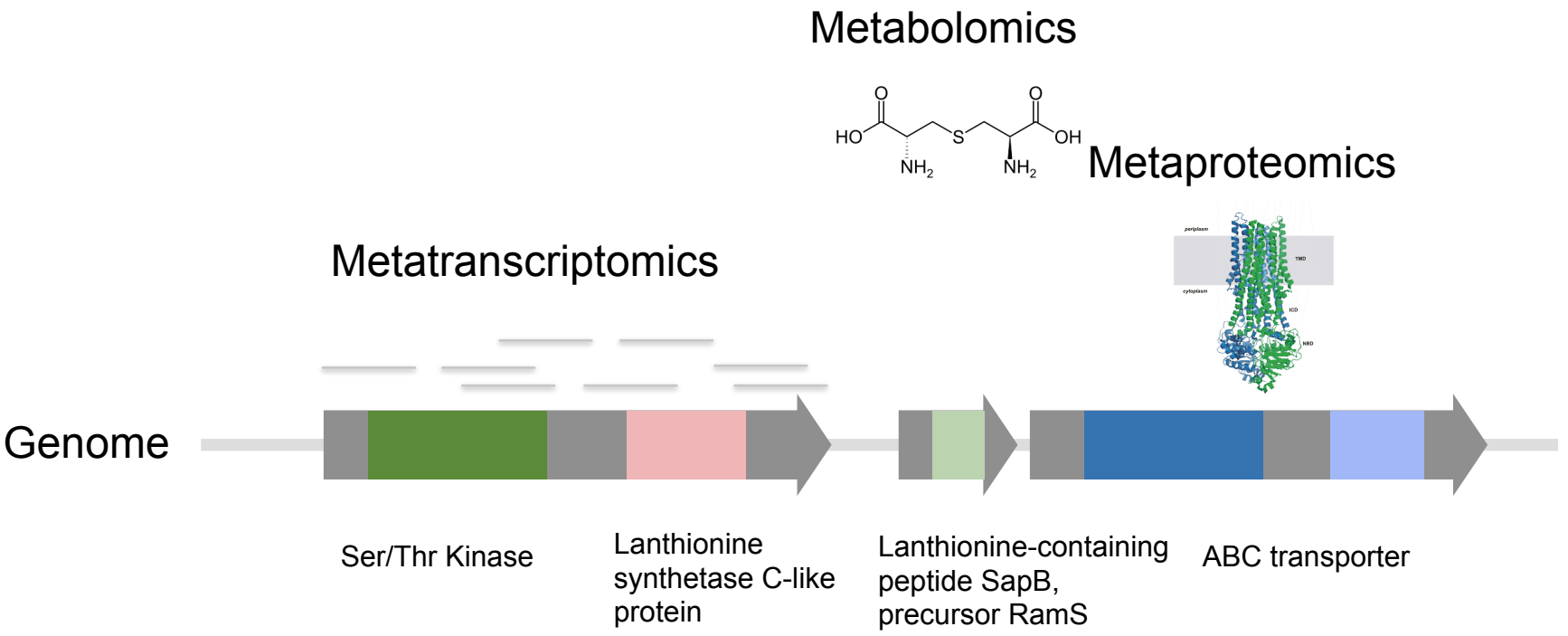



Still missing one whole aspect of an organism's biology

The microbiome!

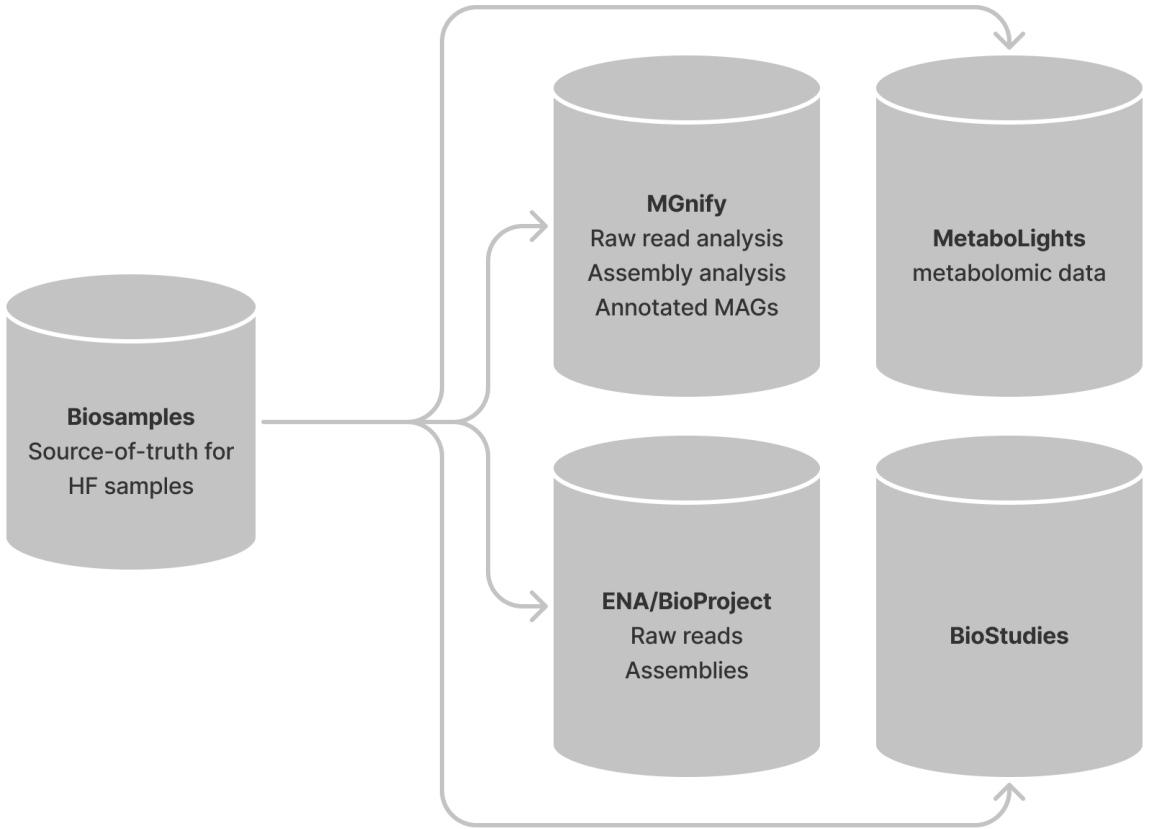
Ebrahim et al. 2016

# Genomes are foundational for understanding multi-‘omics




EMBL-EBI

# Challenges of mutliomics datasets




Each database assigns their own accession

Different data types often submitted by different centres

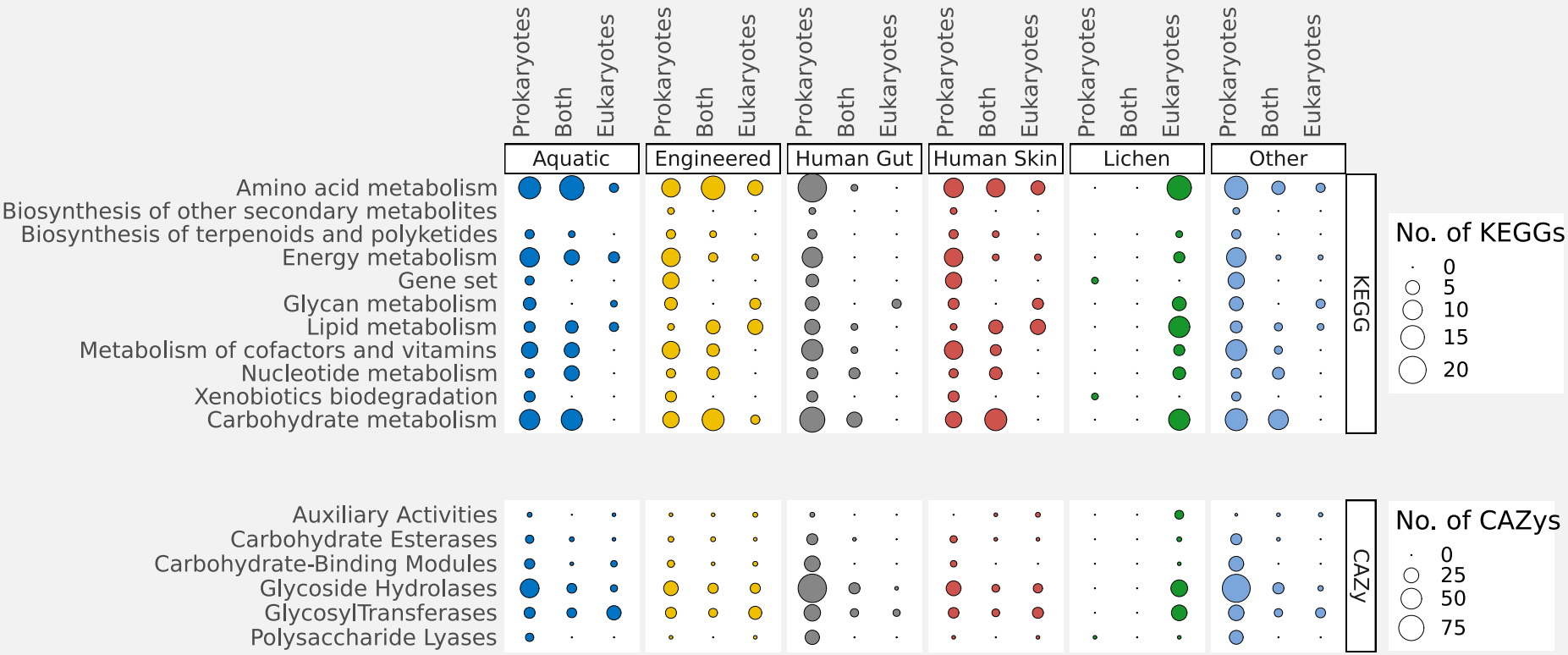
MGnify generates additional data types and derived data

Not all datatypes have an archive, so BioStudies is a catch all

The key is a registration of a BioSample identifier

EMBL-EBI

# Eukaryotes contribute a significant functional repertoire

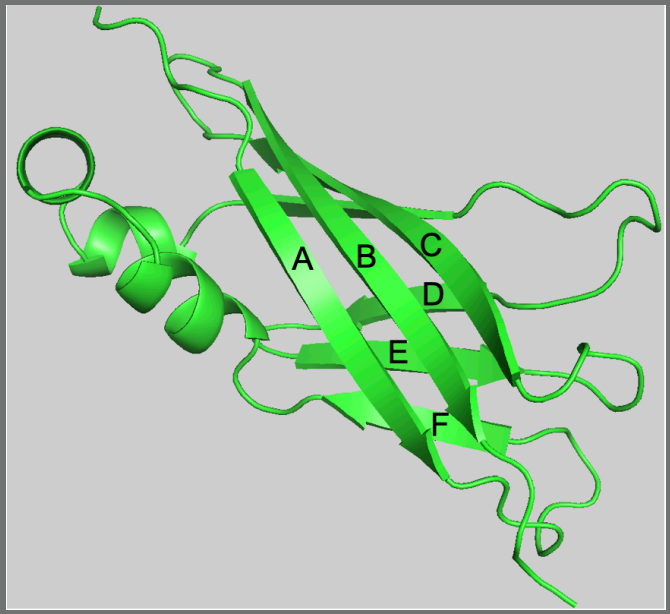



MGnify

EMBL-EBI

# Future challenges

- 50% of all sequences lack any functional annotation
- Expanding protein annotations with latest ML tools
  - ProtENN - Google AI
  - AlphaFold2 - DeepMind
- Prokaryotic and Eukaryotic MAGs are a treasure trove of new functional information, e.g. biosynthetic gene clusters, KEGG modules and carbohydrate utilisation
- With these genomes, metabolic models becomes possible
- How human microbiomes interact with the host requires cultivation of many more microbes to allow in vitro and in vivo studies

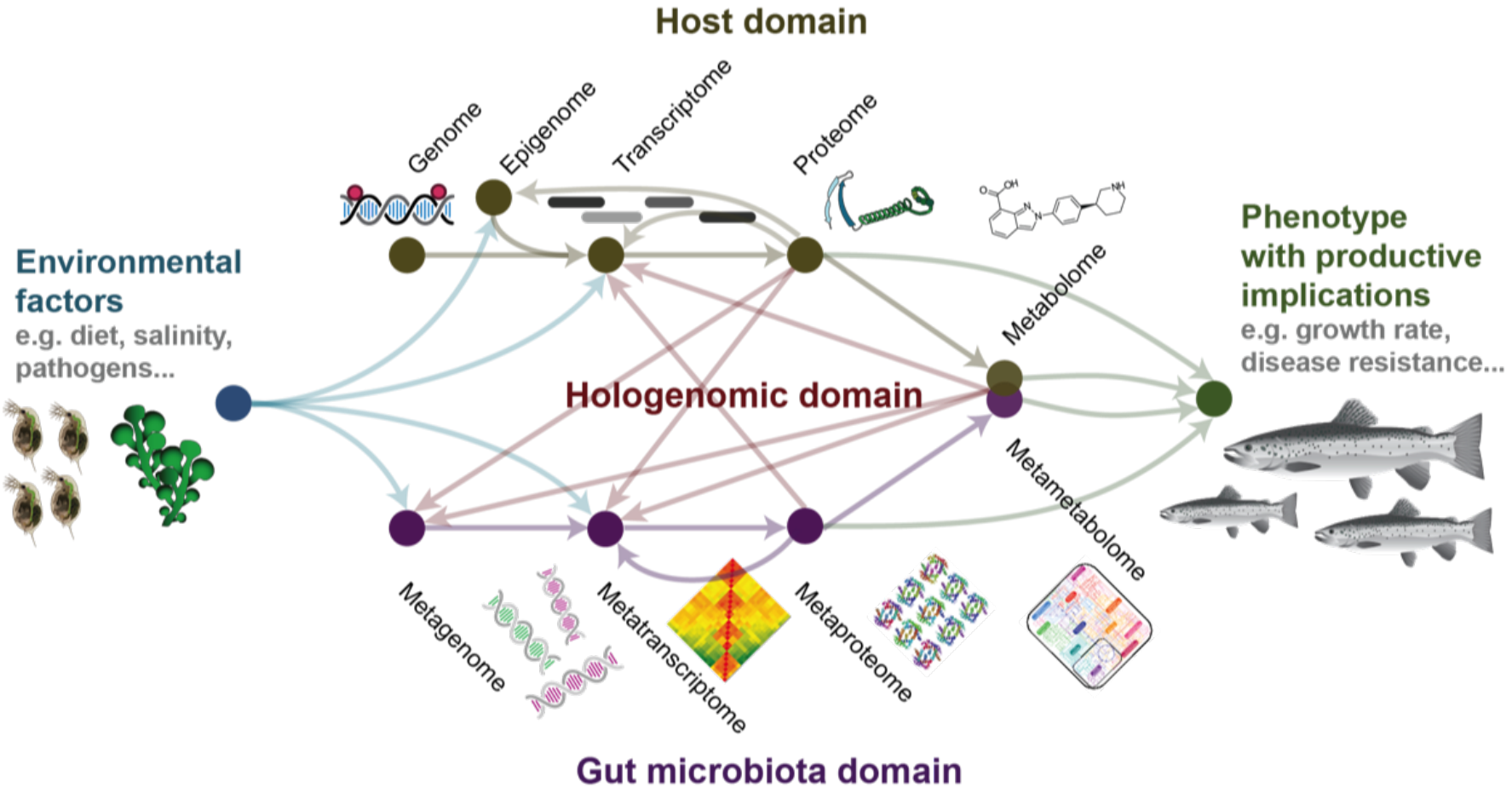



New approaches will start unravelling the "unknowns"



EMBL

# The hologenomics scheme



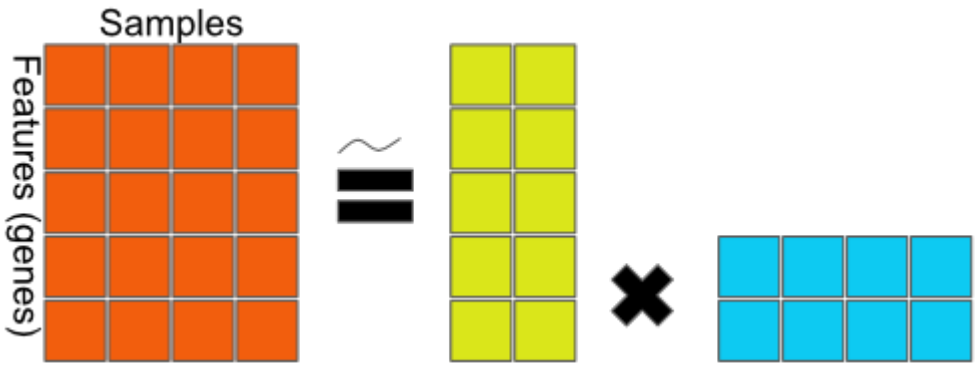
# Why use multi-omics?

- Integrate information on biological pathways
  - Mechanistic information embedded in multi-omics measurement
  - Ability to model causal pathways

# Challenges of multi-omics

- Curse of dimensionality
  - Millions of genomic variants
  - 1000s of gene expression and metabolites
  - 100s-100,000s of microbial genes

# Matrix factorization




Reduce a matrix of samples and features (any measurements) to two components
- one corresponding to samples
- one corresponding to features

Can represent correlation in sample space and feature space

Courtesy: compgenomr.github.io

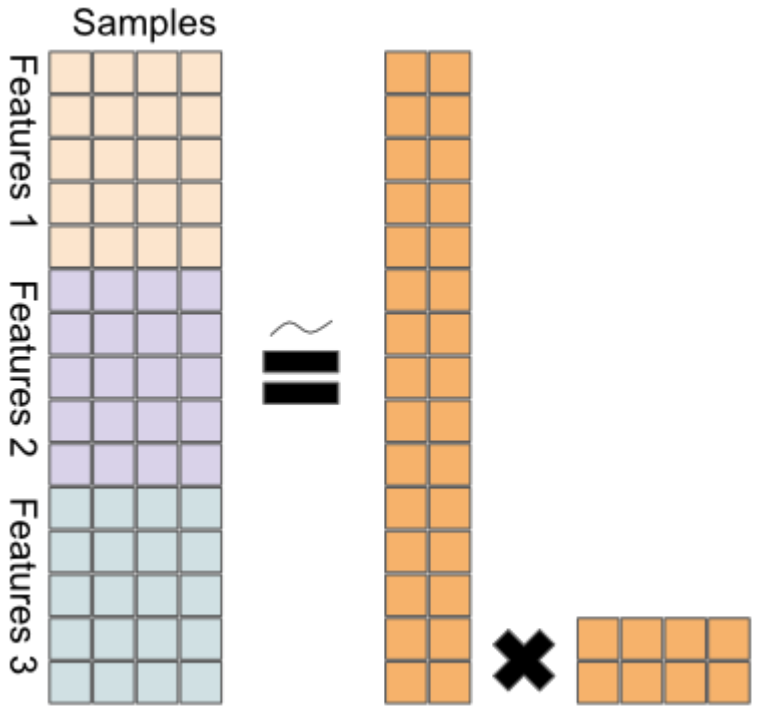
# Matrix factorization

How to extend it to multi-omics?

- just append the features to each other

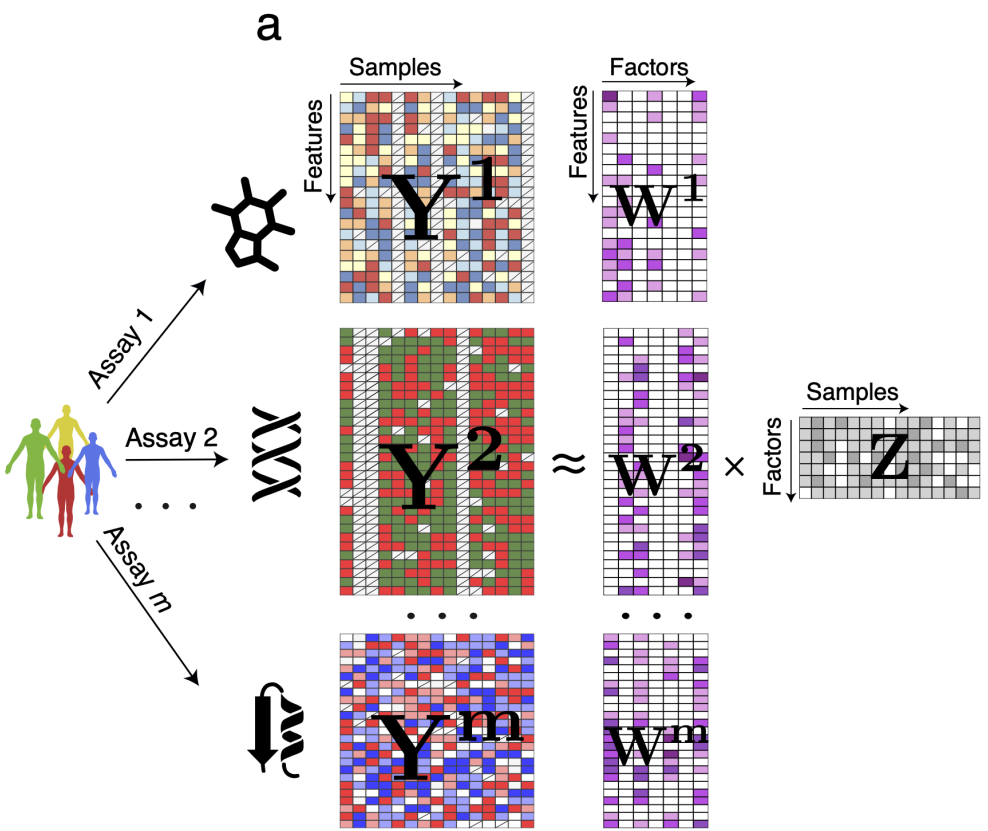
- A large matrix representing features

- Not a very good way to combine information across different -omics methods




Courtesy: compgenomr.github.io

# Multi-omics factor analysis (MOFA)

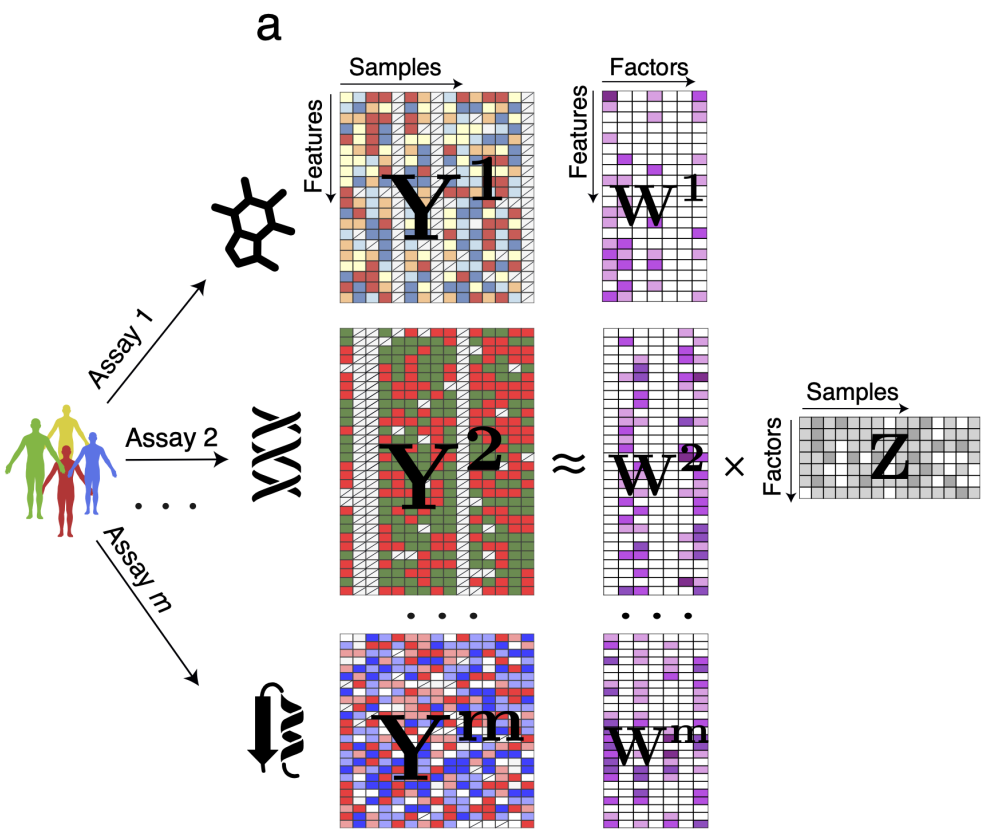


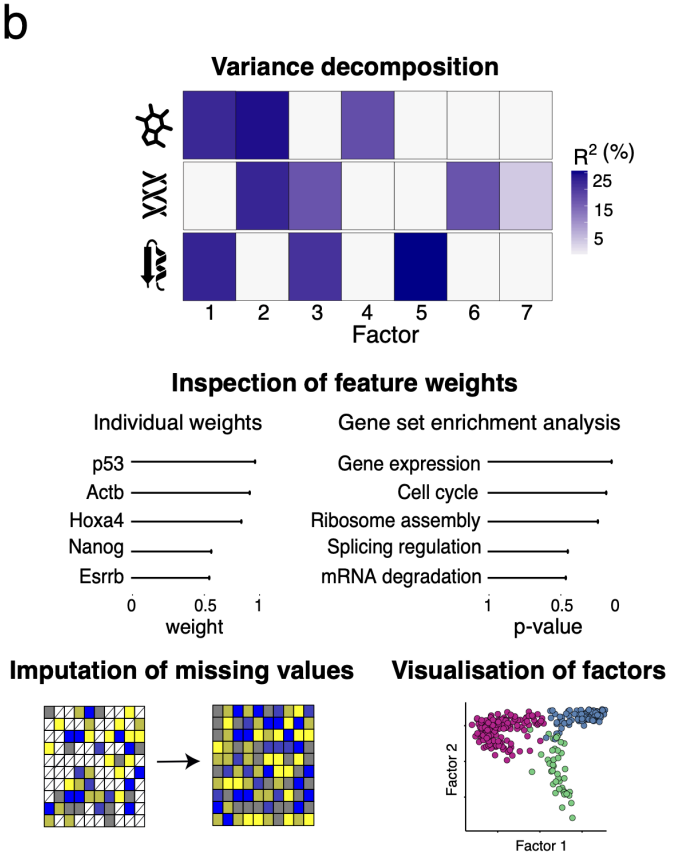

One of the many ways to use matrix factorization (also called latent variable analysis) to model multi-omics data

- Each feature gets its own reduced feature factor matrix
- Sample matrix is shared across all features
- Allows for variance decomposition on different factors
- Imputation for missing values

https://biofam.github.io/MOFA2/

# Multi-omics factor analysis (MOFA)

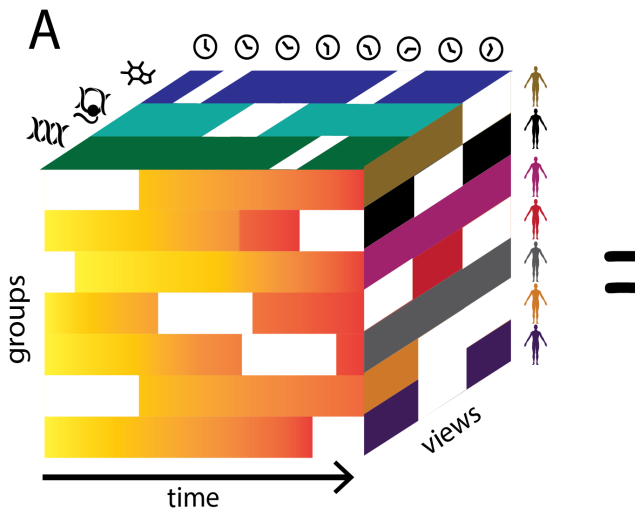


# What about time series?

- Until now, all we have talked about is single time measurements
- How can we use this framework for time-series data?

# MEFISTO (time-series MOFA)

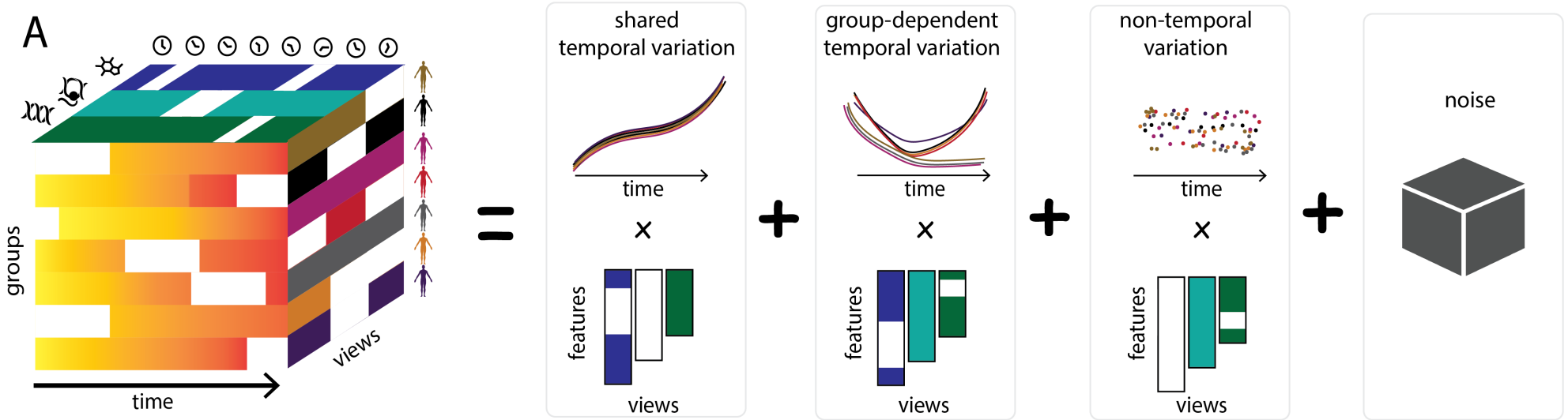



=

Full data
- the left column should is basically what we worked with in MOFA.

https://biofam.github.io/MOFA2/

# MEFISTO (time-series MOFA)

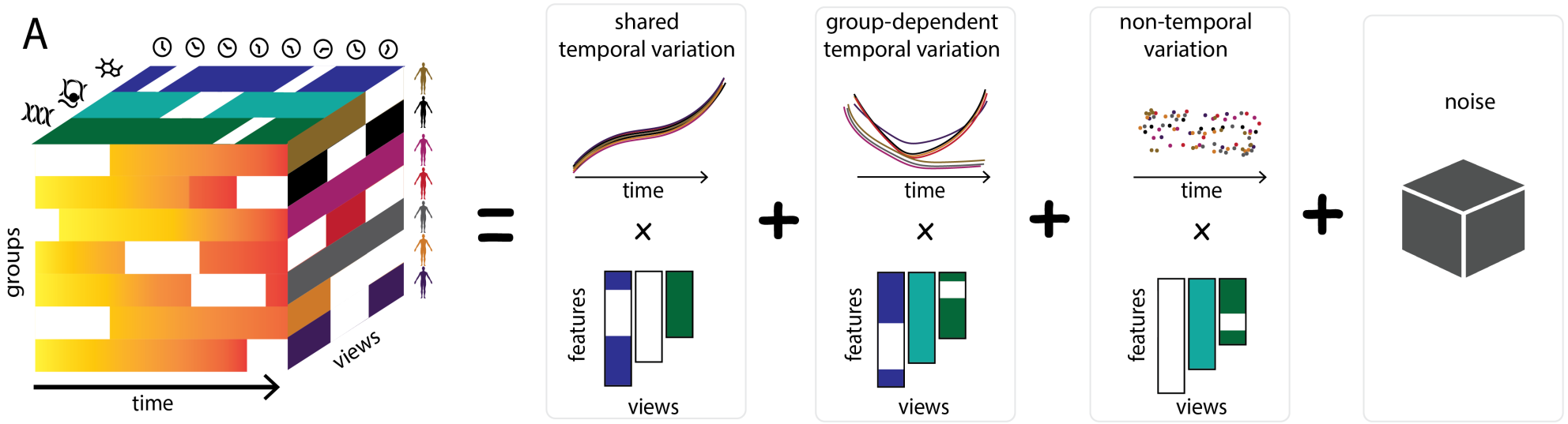



The variation is broken down into 4 pieces.
- time variation across all -omics and groups.
- time variation in groups
- variation not dependent on time but only groups
- noise

Each of the first 3 are modelled using latent factors

https://biofam.github.io/MOFA2/

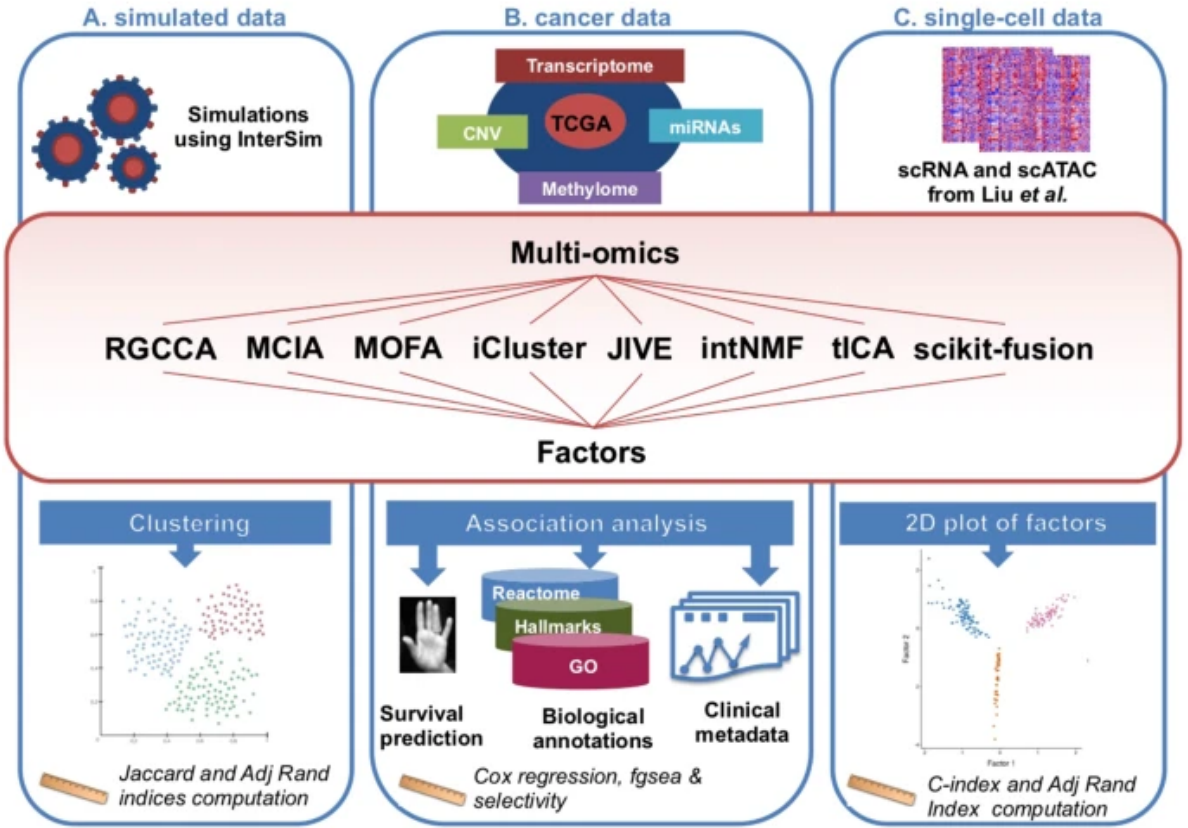
# MEFISTO (time-series MOFA)




Full workflow

https://biofam.github.io/MOFA2/

# A world of integration methods




Cantini et al. 2021

# Where do we go from here?

- Multi-omics methods
  - Beyond pairwise association
    **Factor analysis, networks – MOFA, iCluster ...**
  - Add temporal and spatial information
    **Extended latent factors – MEFISTO ...**
  - Incorporate *a priori* information
    **Biological knowledge graphs, interactome graphs**

# FindingPheno: Multi-omics for genotype-phenotype associations

UNIVERSITY OF COPENHAGEN

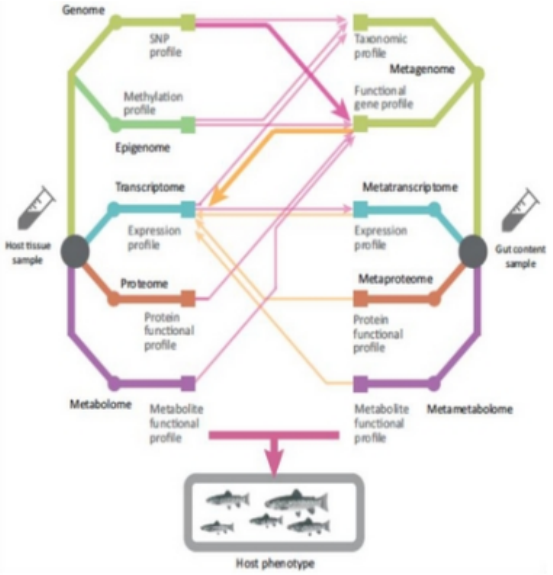
UNIVERSITY OF TURKU

ÖKOLÓGIAI KUTATÓKÖZPONT

EMBL-EBI

QIAGEN

CHR HANSEN

# Thank you!