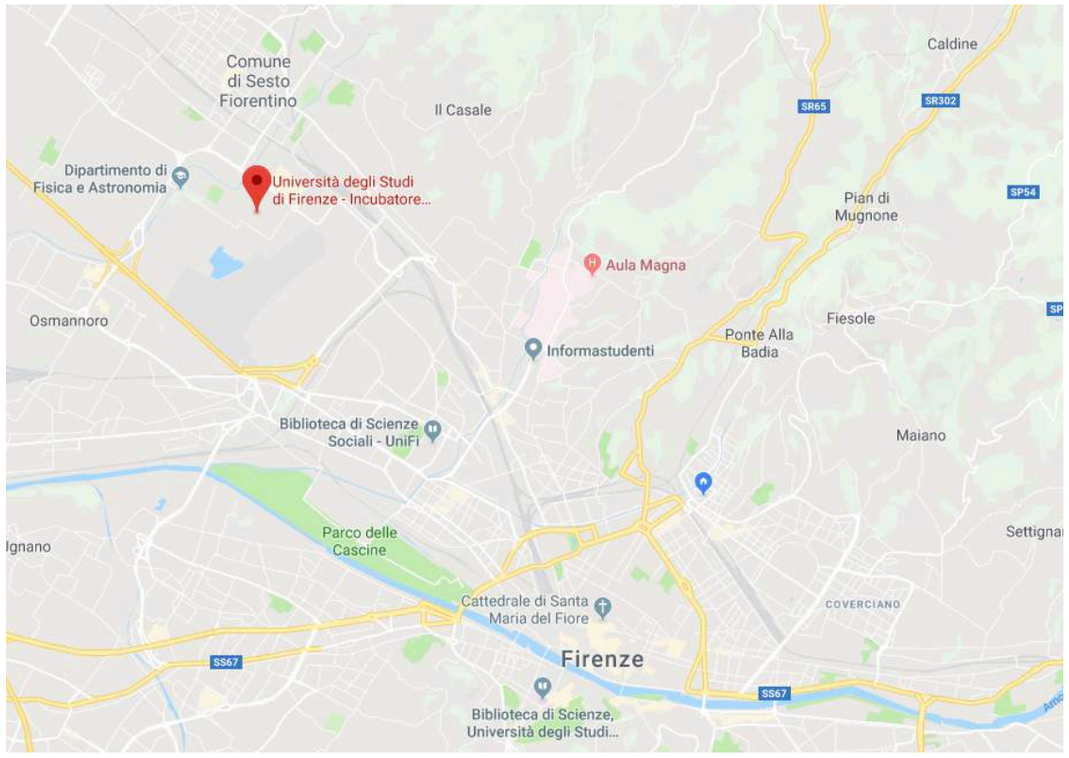
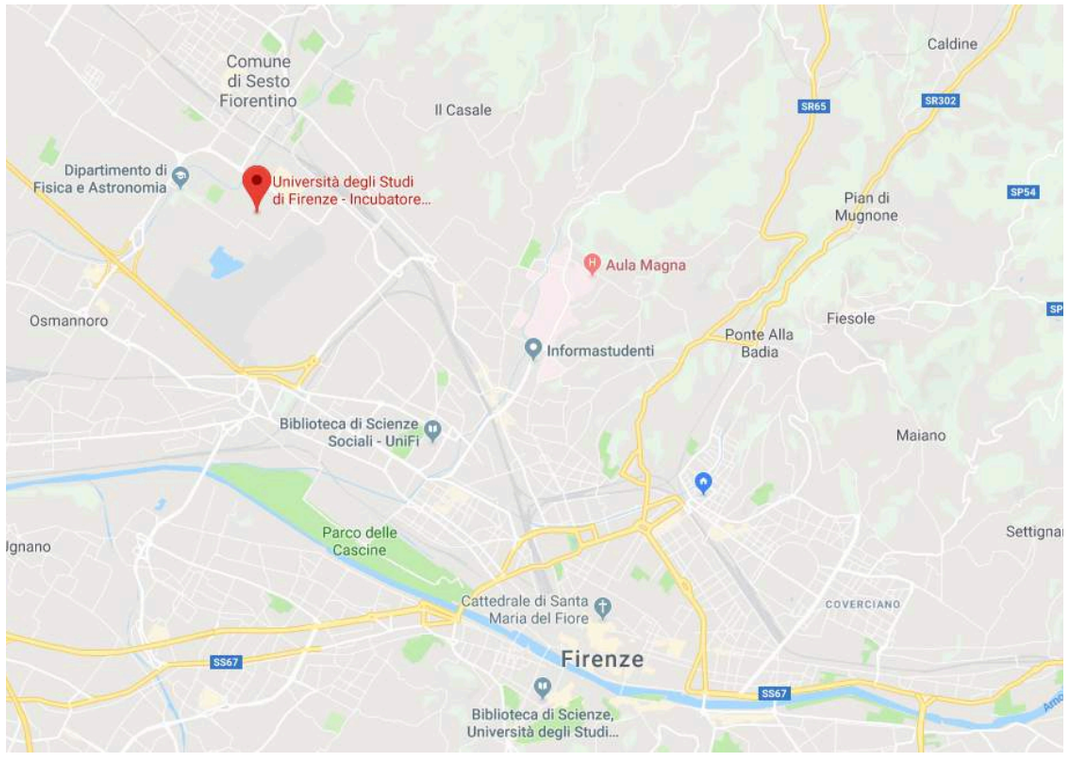
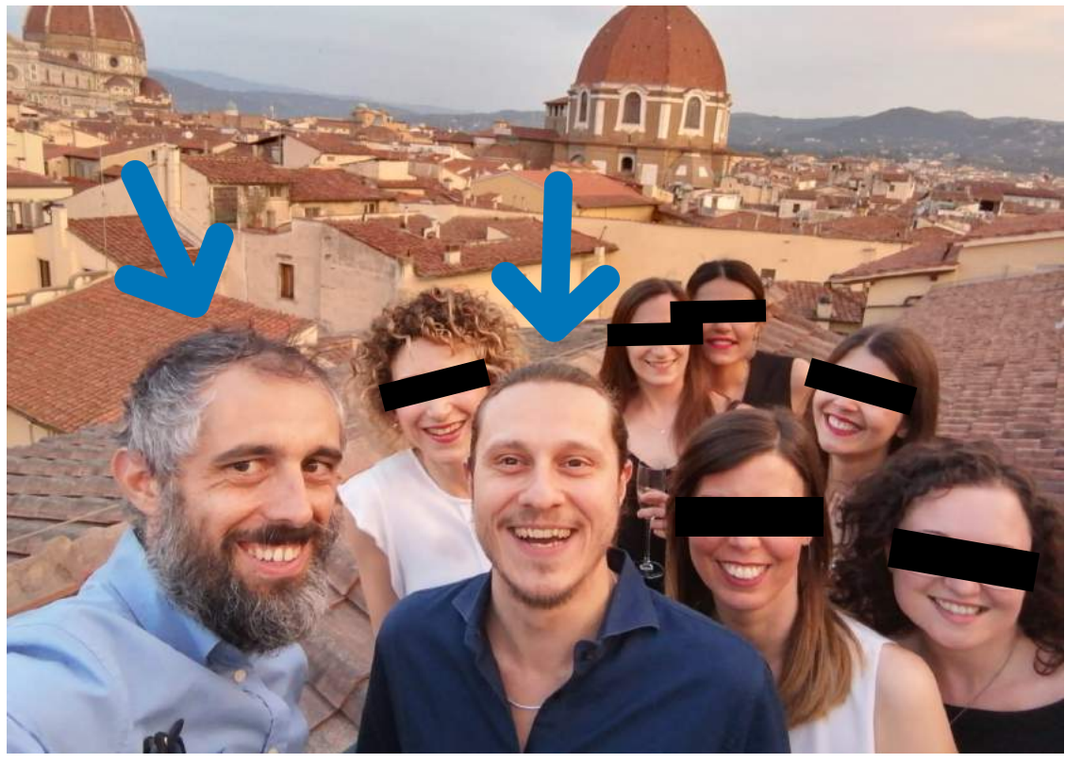
In-silico prediction of microbial assemblages

from metagenomics to community modelling

Giovanni Bacci - giovanni.bacci@unifi.it

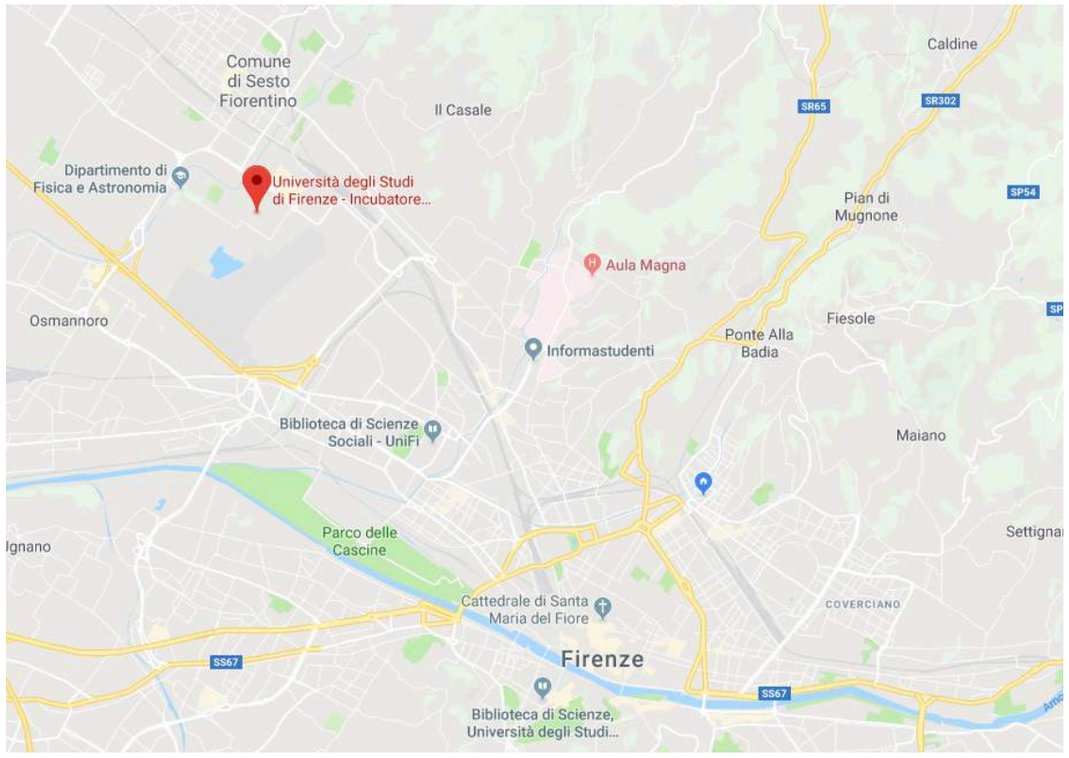

Comune
di Sesto
Fiorentino
Il Casale
Caldine
SR65
SR302
Dipartimento di
Fisica e Astronomia
Università degli Studi
di Firenze - Incubatore...
Pian di
Mugnone
SP54
Aula Magna
Osmannoro
Fiesole
Ponte Alla
Badia
Informastudenti
Biblioteca di Scienze
Sociali - UniFi
Maiano
Parco delle
Cascine
Cattedrale di Santa
Maria del Fiore
COVERCIANO
Firenze
SS67
SS67
Biblioteca di Scienze,
Università degli Studi...

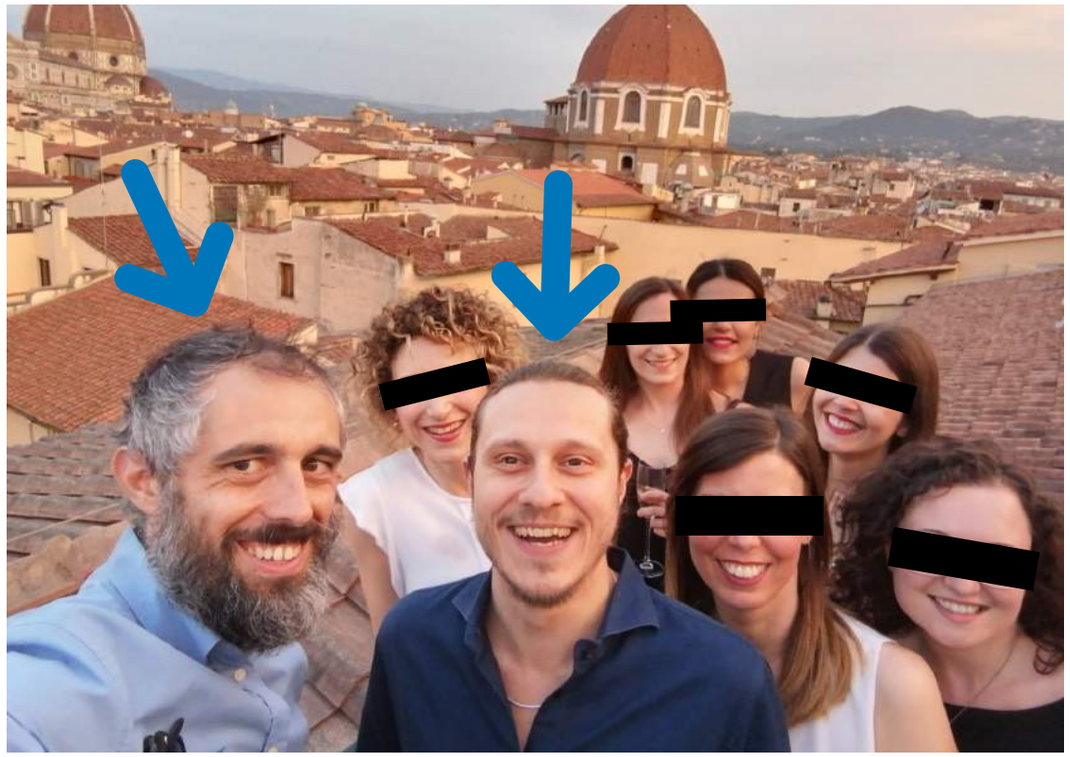

Comune
di Sesto
Fiorentino
Il Casale
Caldine
SR65
SR302
Dipartimento di
Fisica e Astronomia
Università degli Studi
di Firenze - Incubatore...
Pian di
Mugnone
SP54
Aula Magna
Osmannoro
Fiesole
Ponte Alla
Badia
Informastudenti
Biblioteca di Scienze
Sociali - UniFi
Maiano
Parco delle
Cascine
Cattedrale di Santa
Maria del Fiore
COVERCIANO
Firenze
SS67
SS67
Biblioteca di Scienze,
Università degli Studi...



Florence
computational
biology group

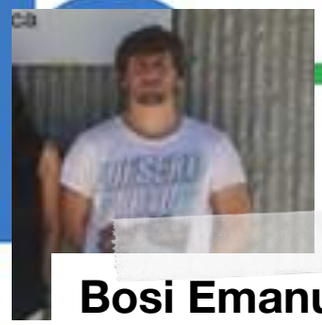
cmbg





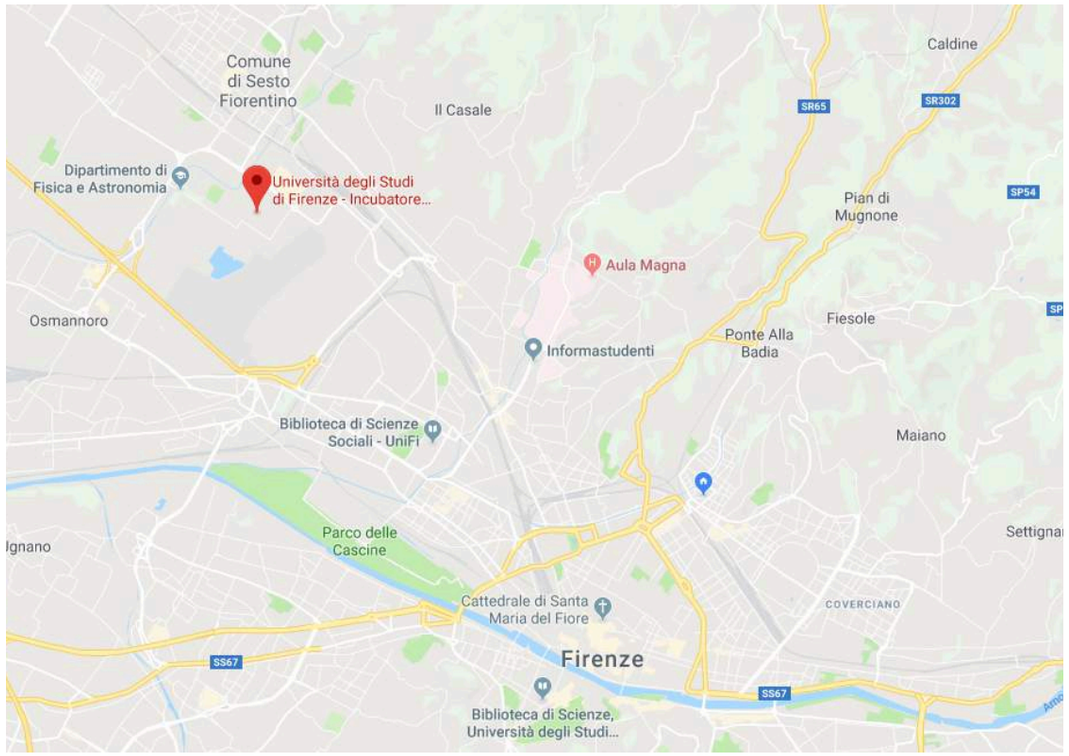
Luana Presta
University of Luxembourg

Florence computational biology group

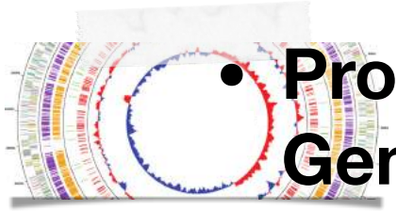


Bosi Emanuele
University of Pisa

Marco Galardini
Boston University

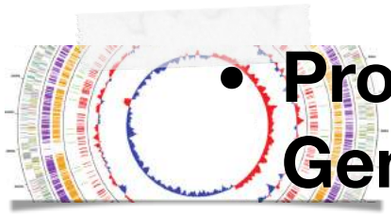


Research fields

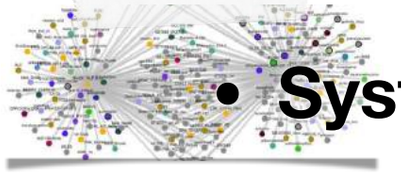


- **Prokaryotic / Eukaryotic Genomics**

Research fields

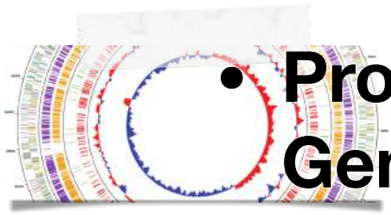


- **Prokaryotic / Eukaryotic Genomics**

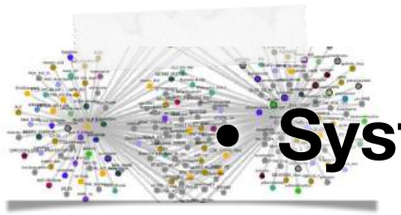


- **Systems Biology**

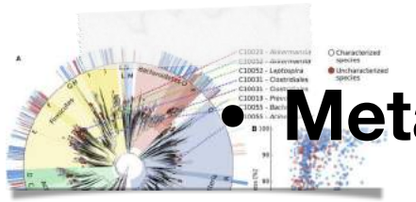
Research fields



- **Prokaryotic / Eukaryotic Genomics**

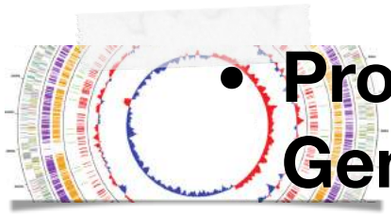


- **Systems Biology**

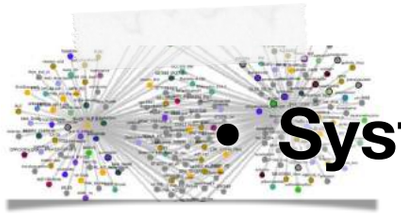


- **Metagenomics**

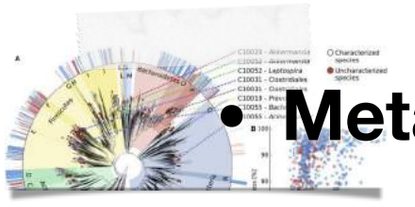
Research fields



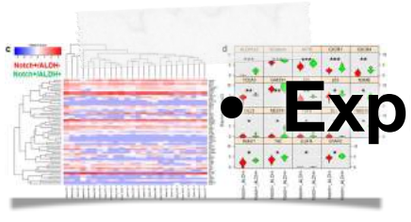
- **Prokaryotic / Eukaryotic Genomics**



- **Systems Biology**

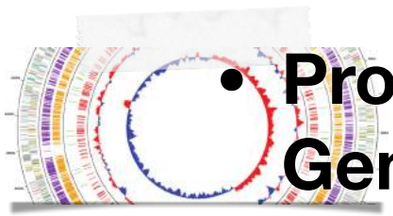


- **Metagenomics**

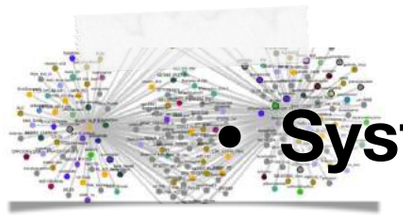


- **Expression data**

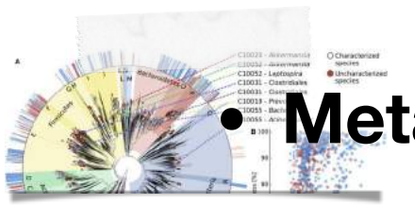
Research fields



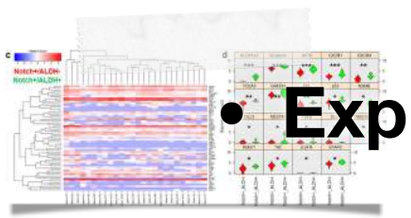
- **Prokaryotic / Eukaryotic Genomics**



- **Systems Biology**

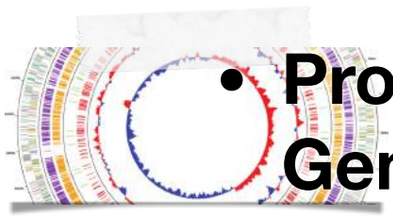


- **Metagenomics**

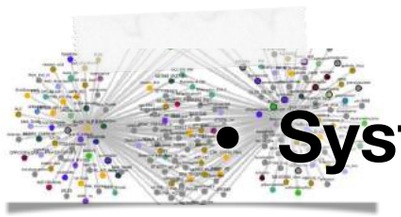


- **Expression data**

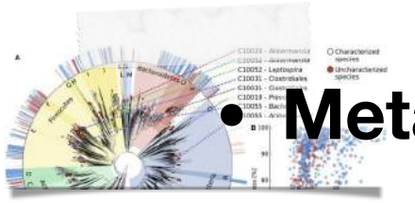
Research fields



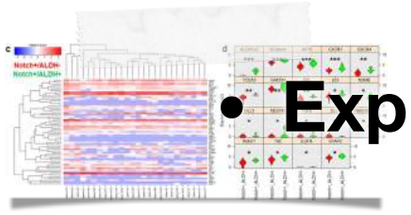
- **Prokaryotic / Eukaryotic Genomics**



- **Systems Biology**

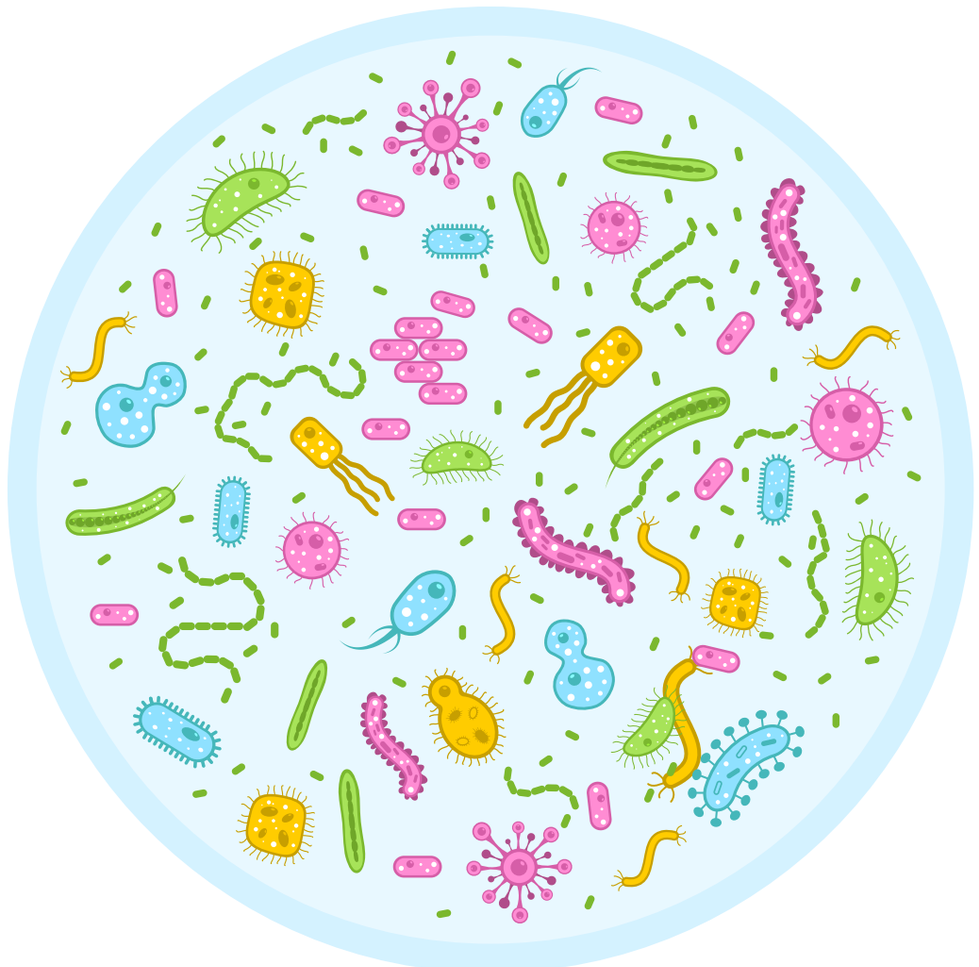


- **Metagenomics**

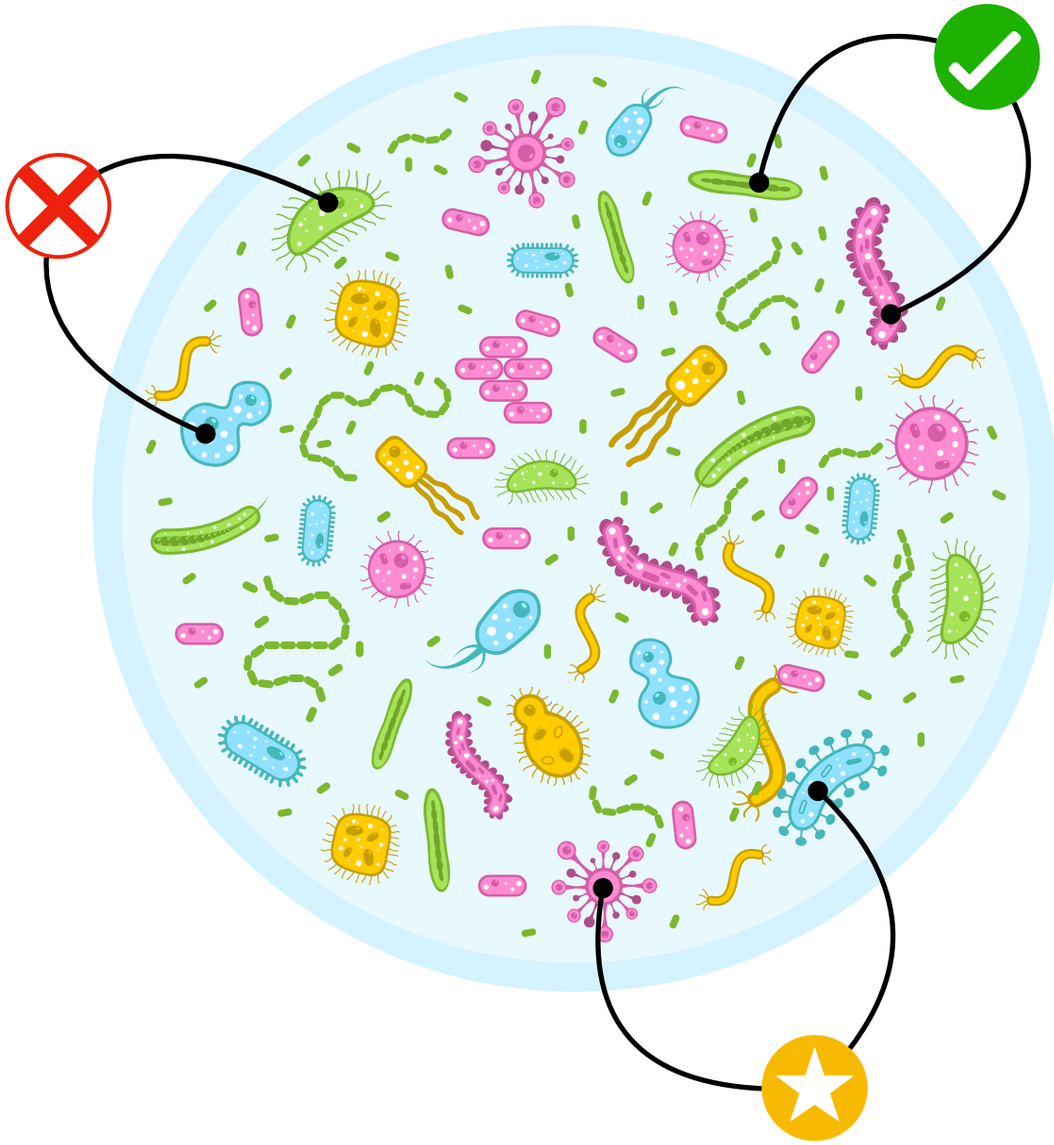


- **Expression data**

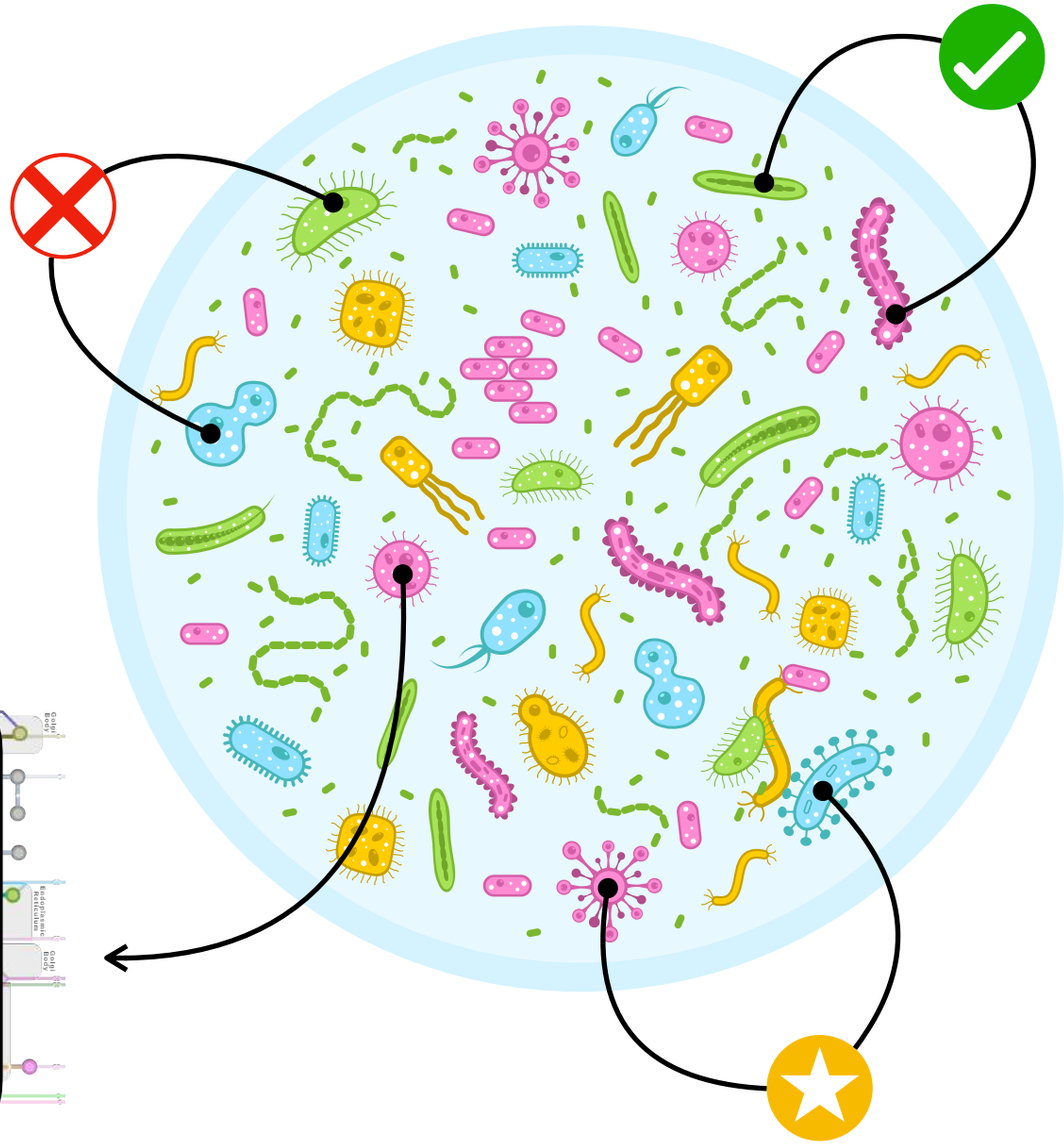
Bacterial roommates

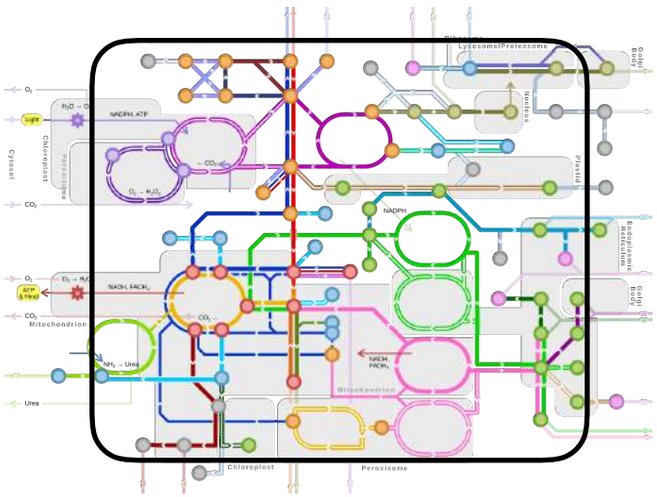


Bacterial roommates

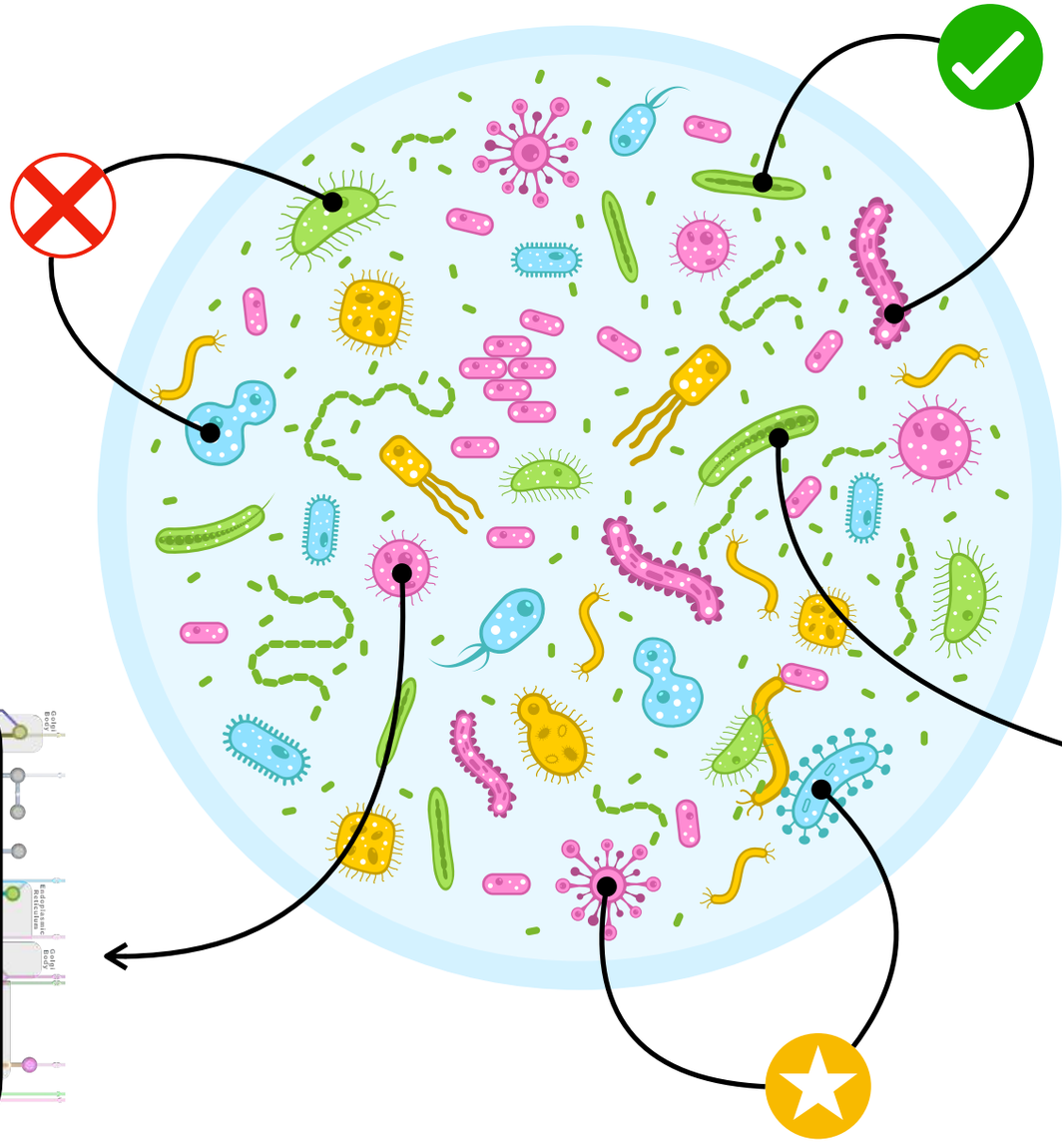


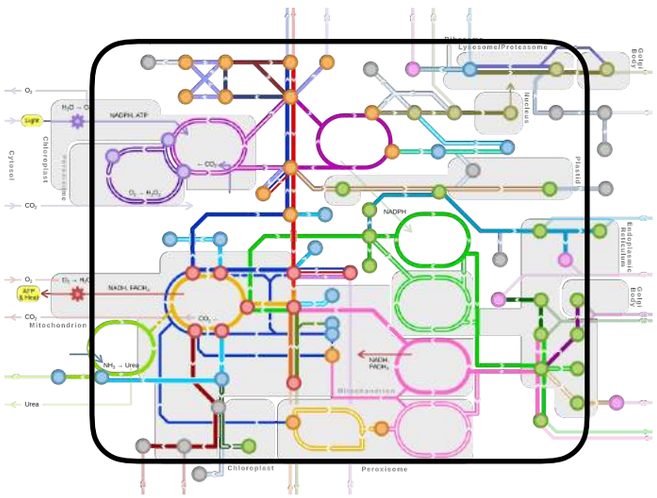
Bacterial roommates



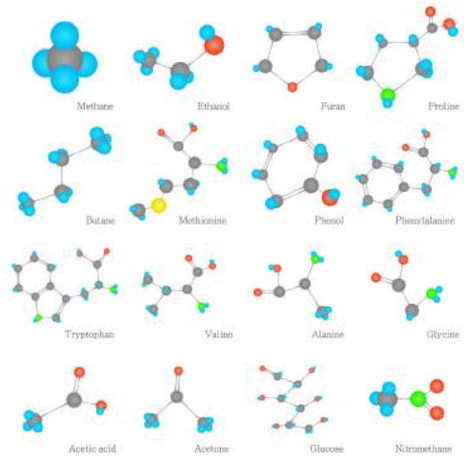


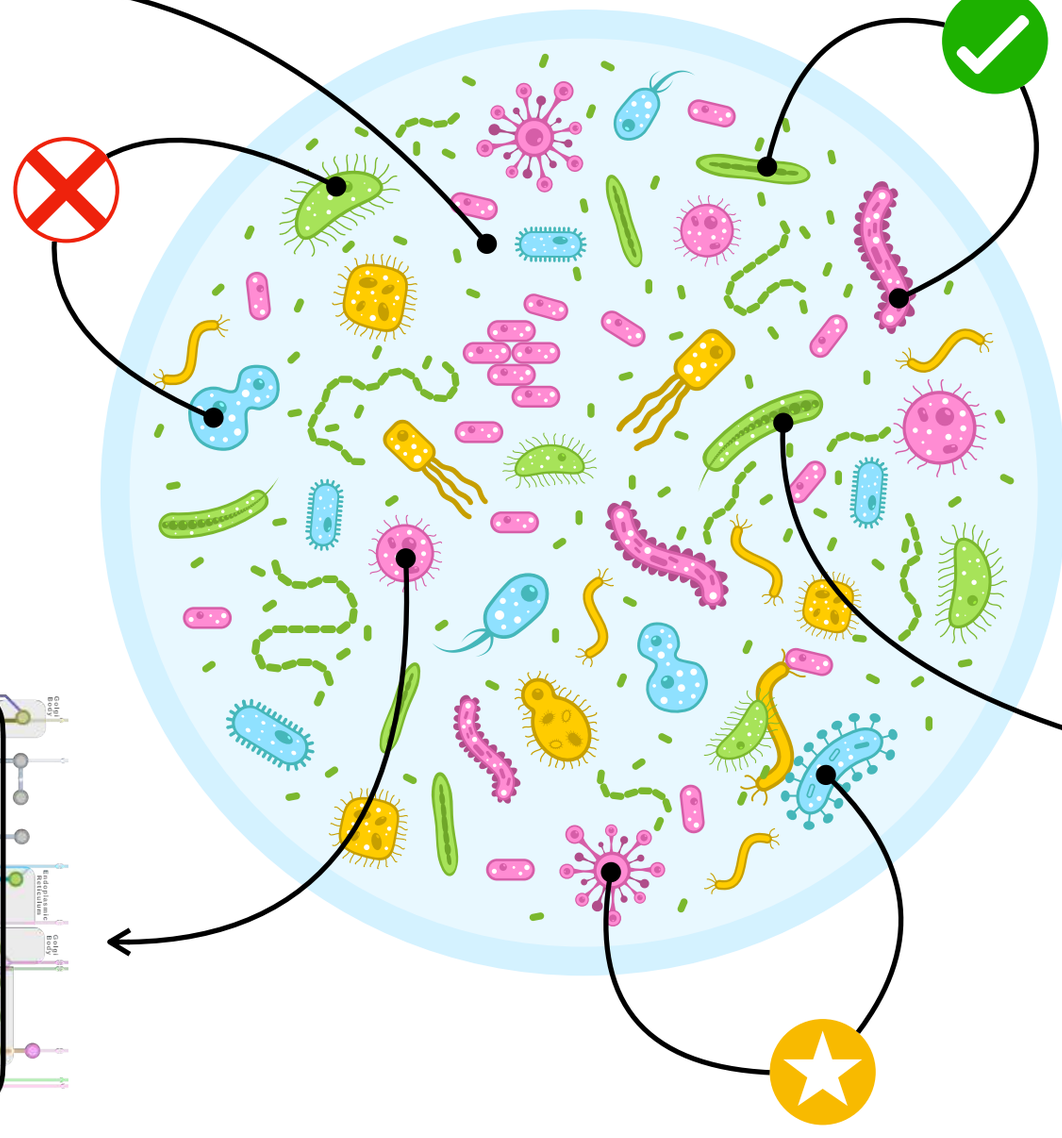
Bacterial roommates

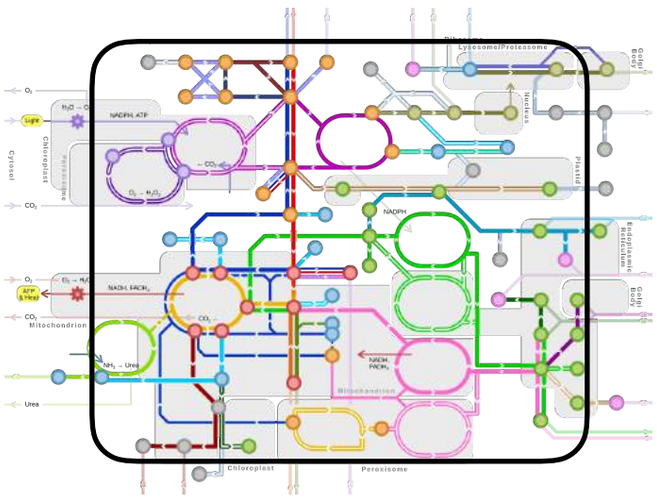


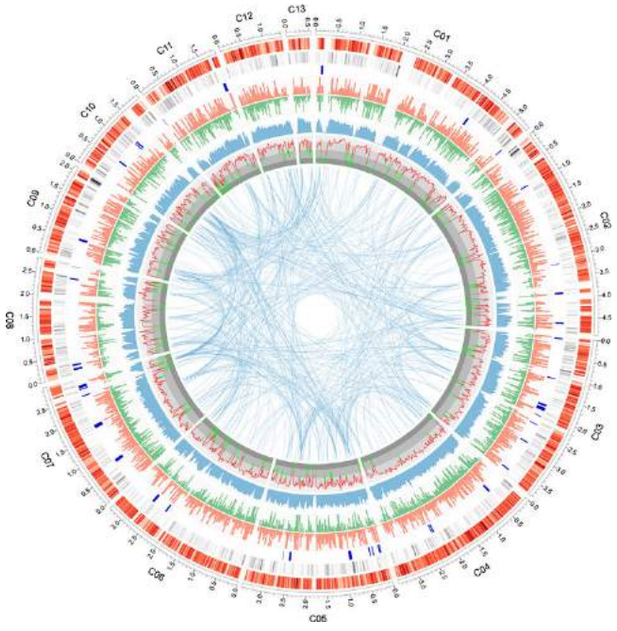


Bacterial roommates

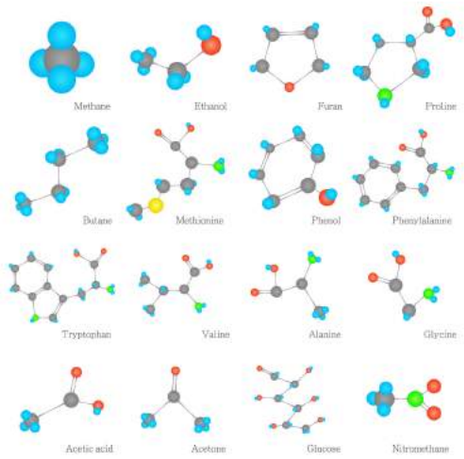


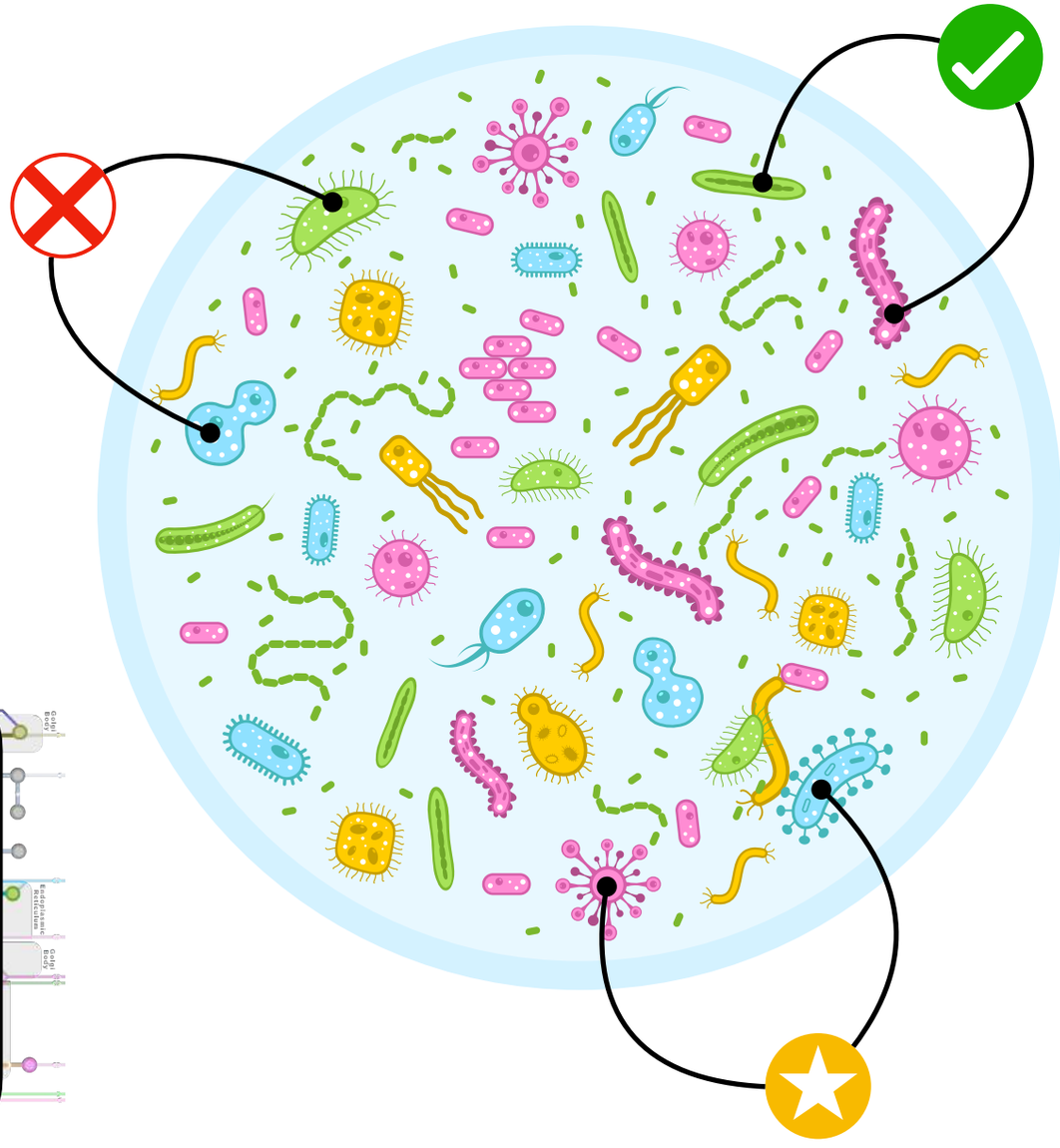


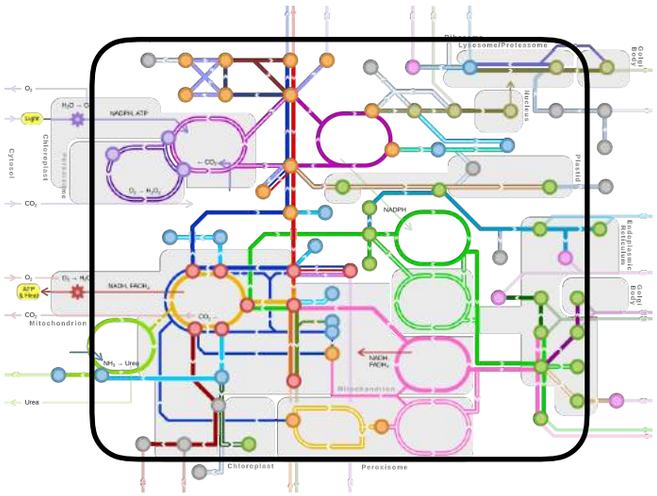




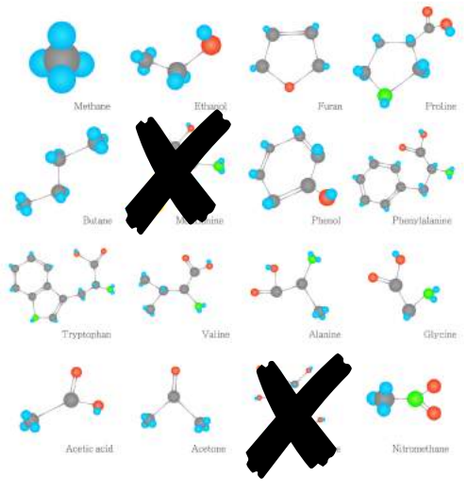
Bacterial roommates

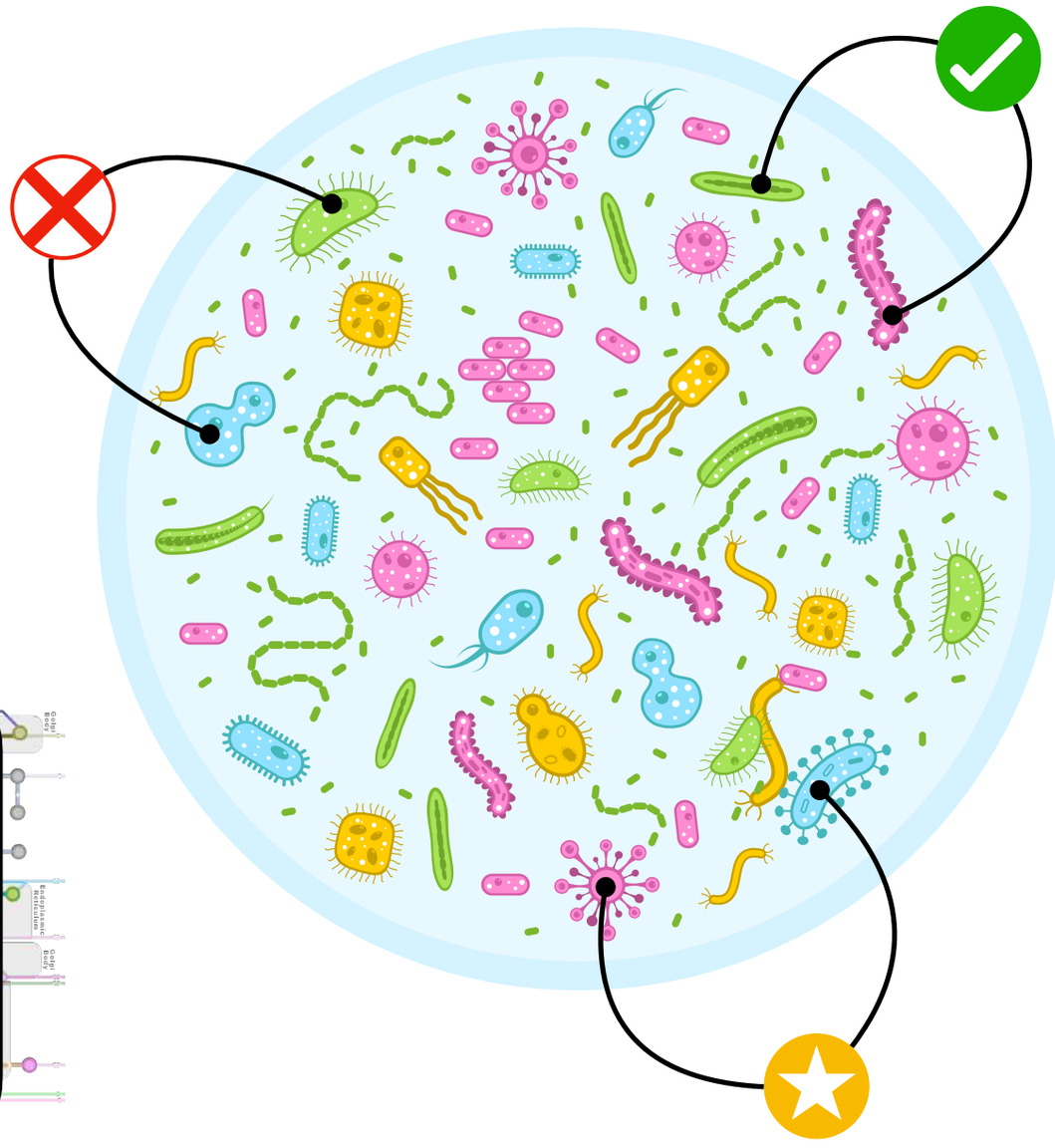


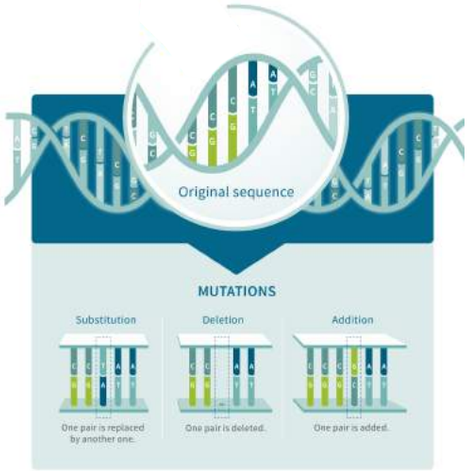


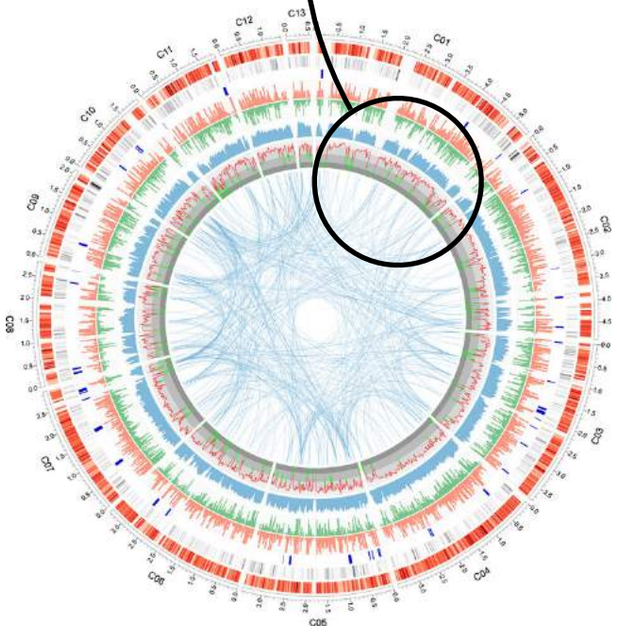


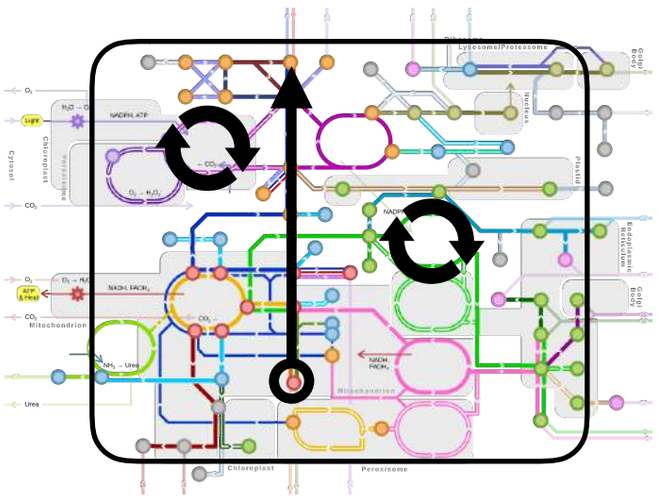
Bacterial roommates



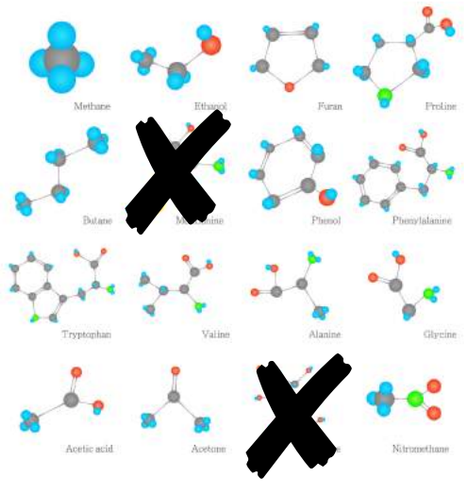


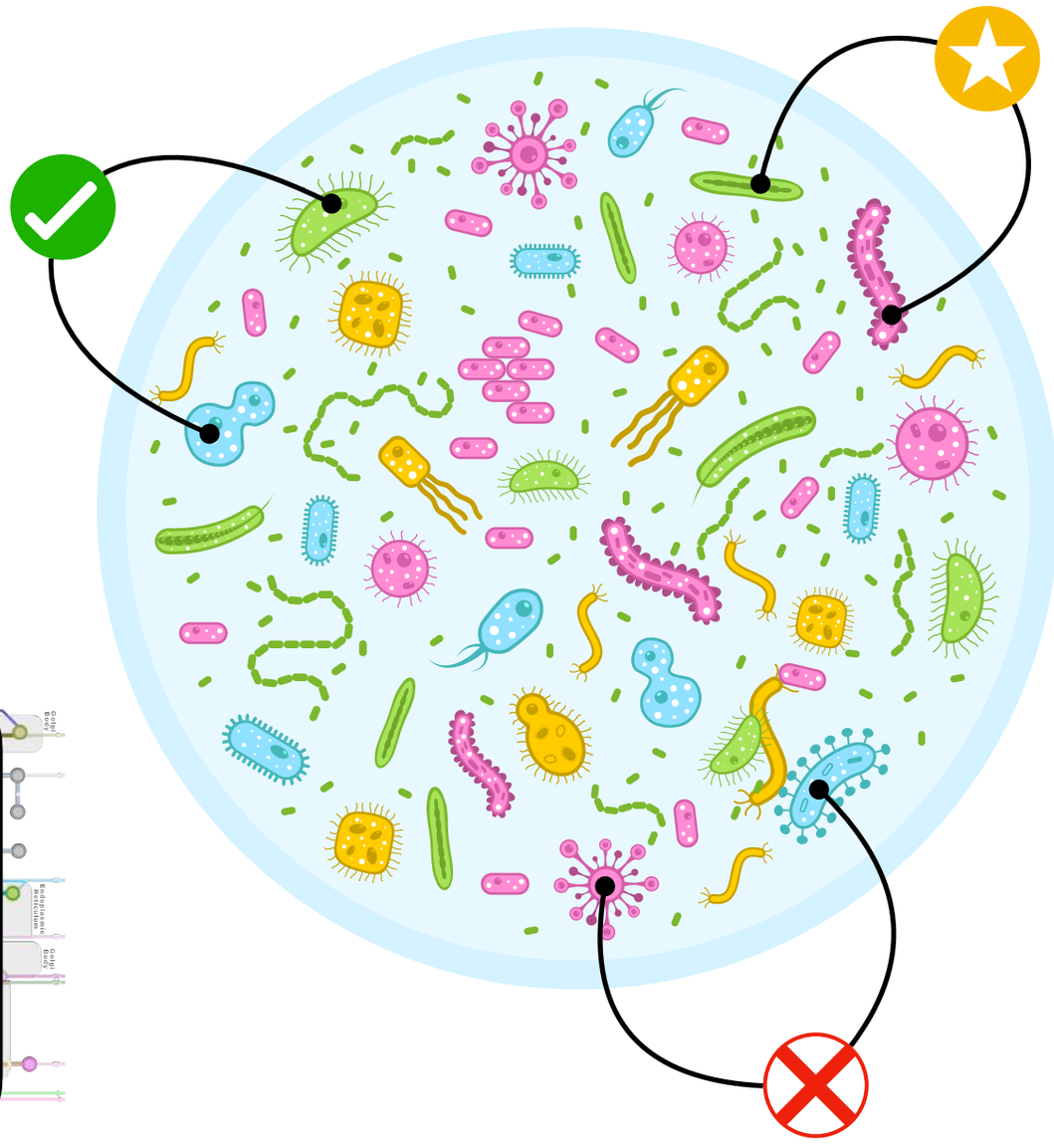


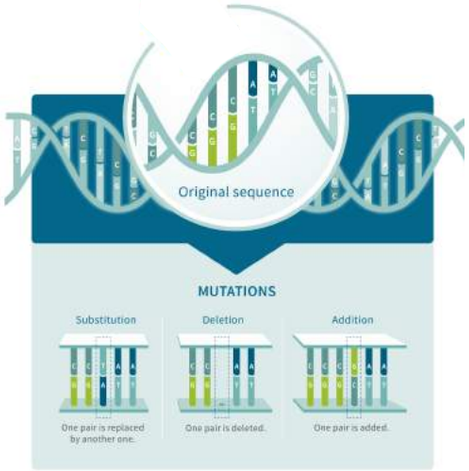


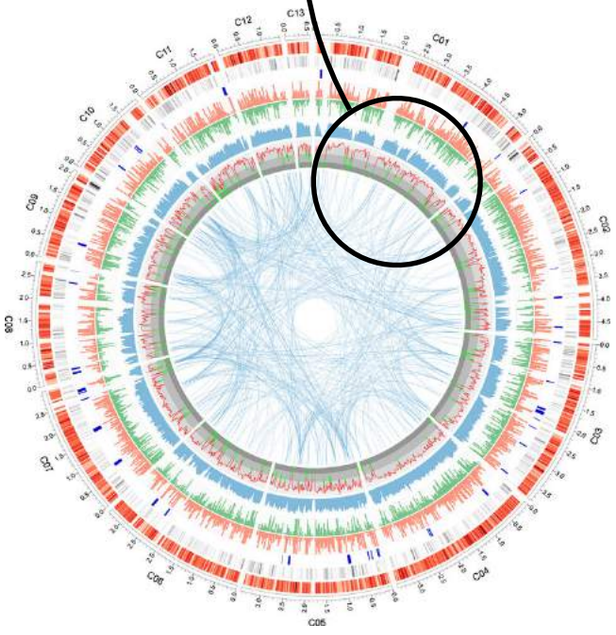


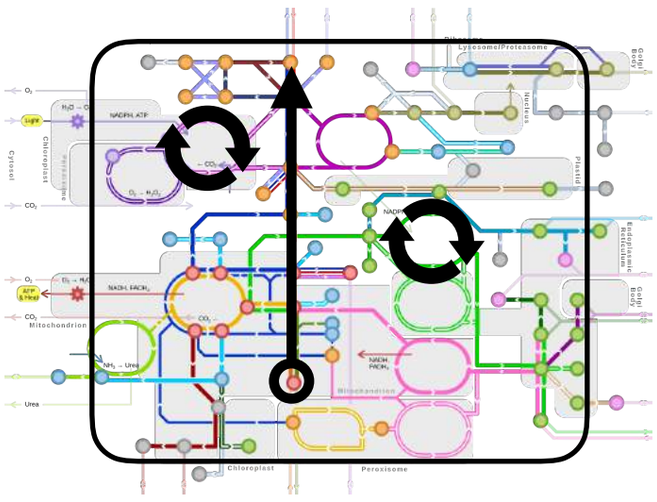
Bacterial roommates









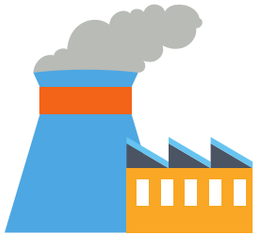


Microbes matter

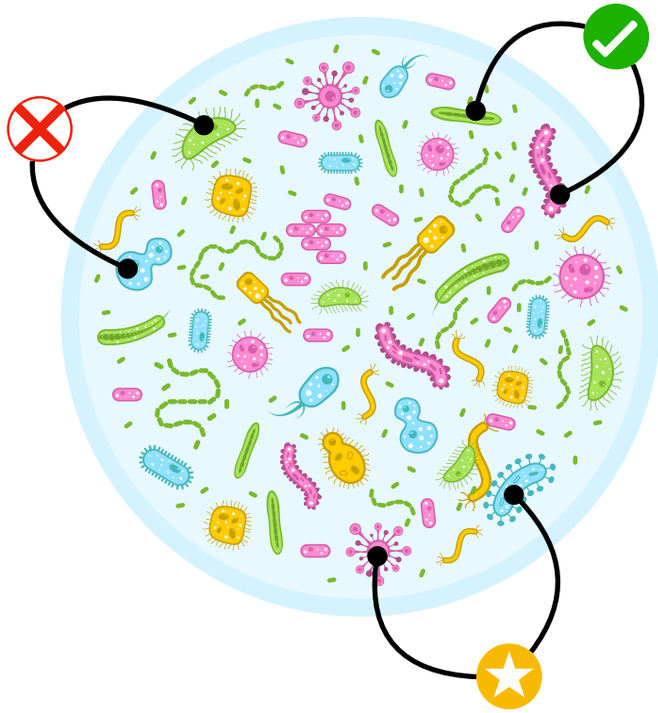
Microbes are a lot!
They are (almost) everywhere
They do a lot of things

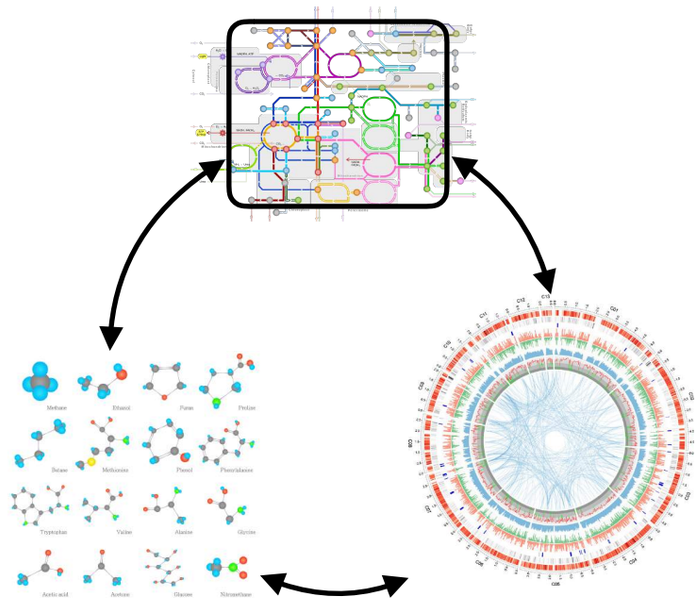
Biogeochemical cycles

Host genome integration
(Hologenome)

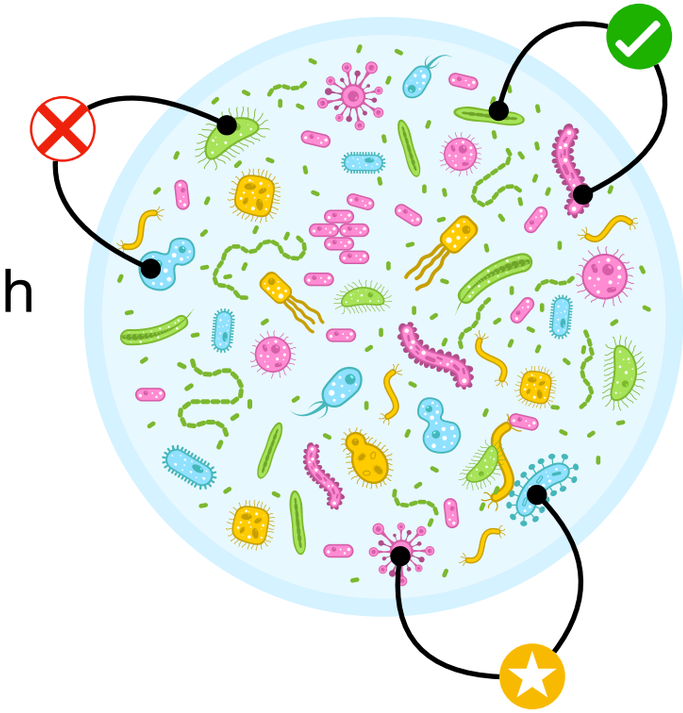


Useful compounds

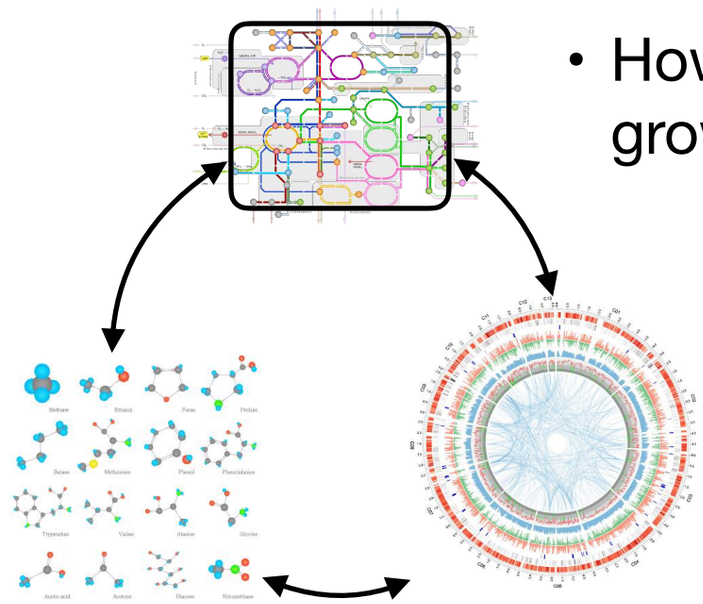



Methane
Ethanol
Furan
Proline
Butane
Methionine
Phenol
Phenylalanine
Tryptophan
Valine
Alanine
Glycine
Acetic acid
Acetone
Glucose
Nitromethane

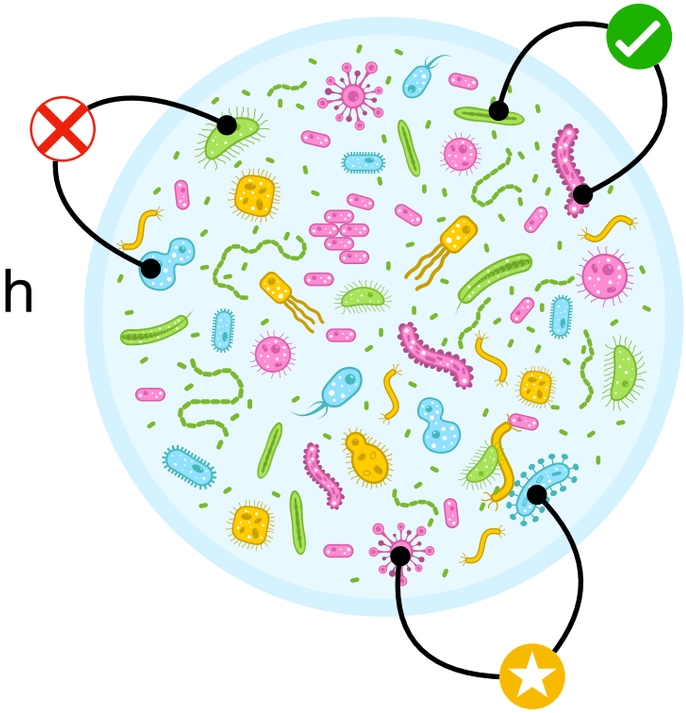
- Who is there?
- How are they influencing each other?



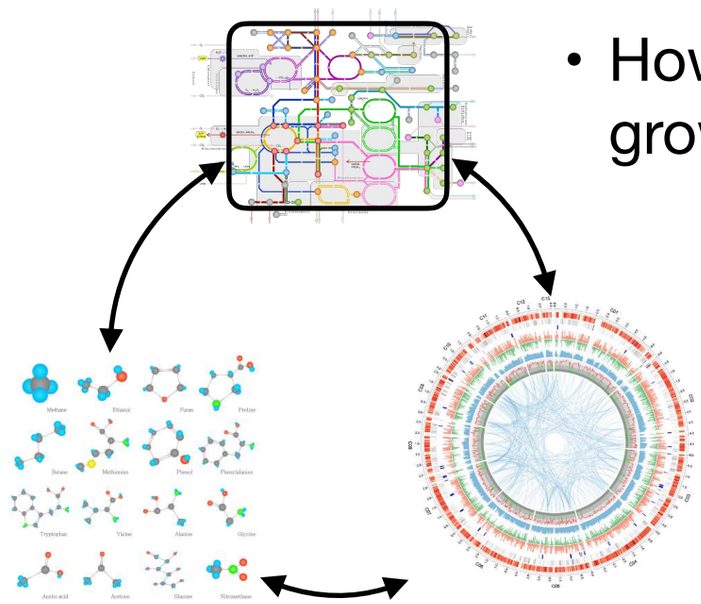
- What are they doing?
- How fast are they growing?

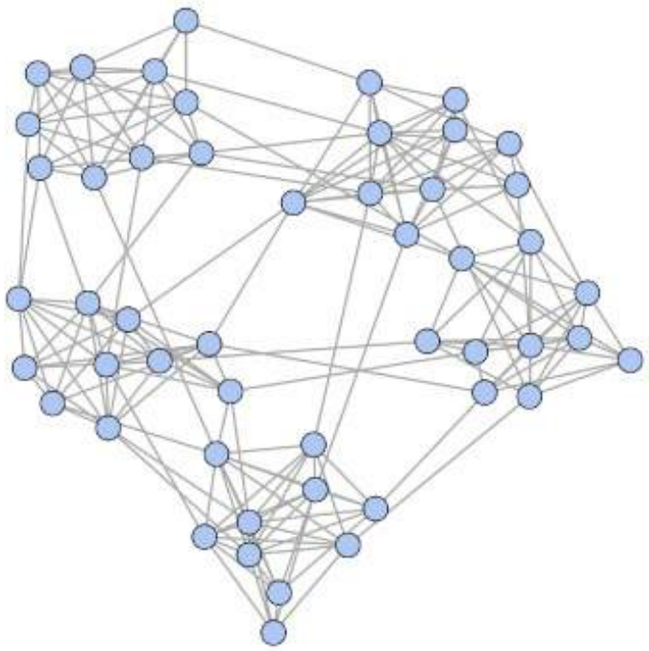


- Who is there?
- How are they influencing each other?

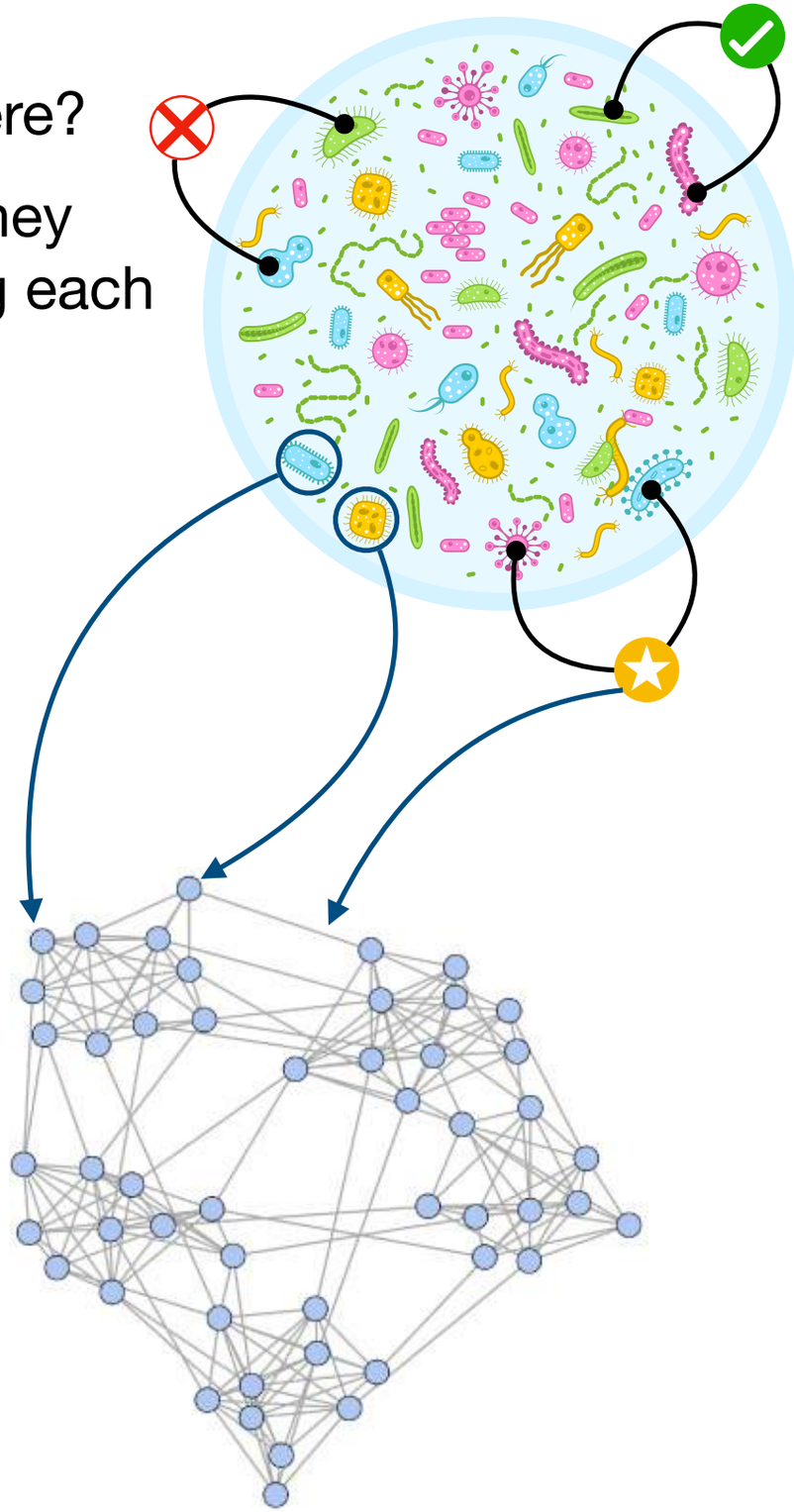


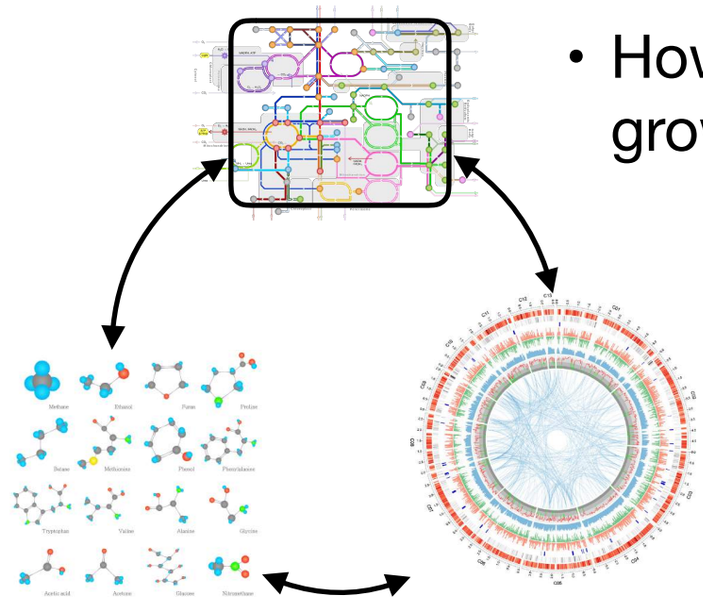
- What are they doing?
- How fast are they growing?





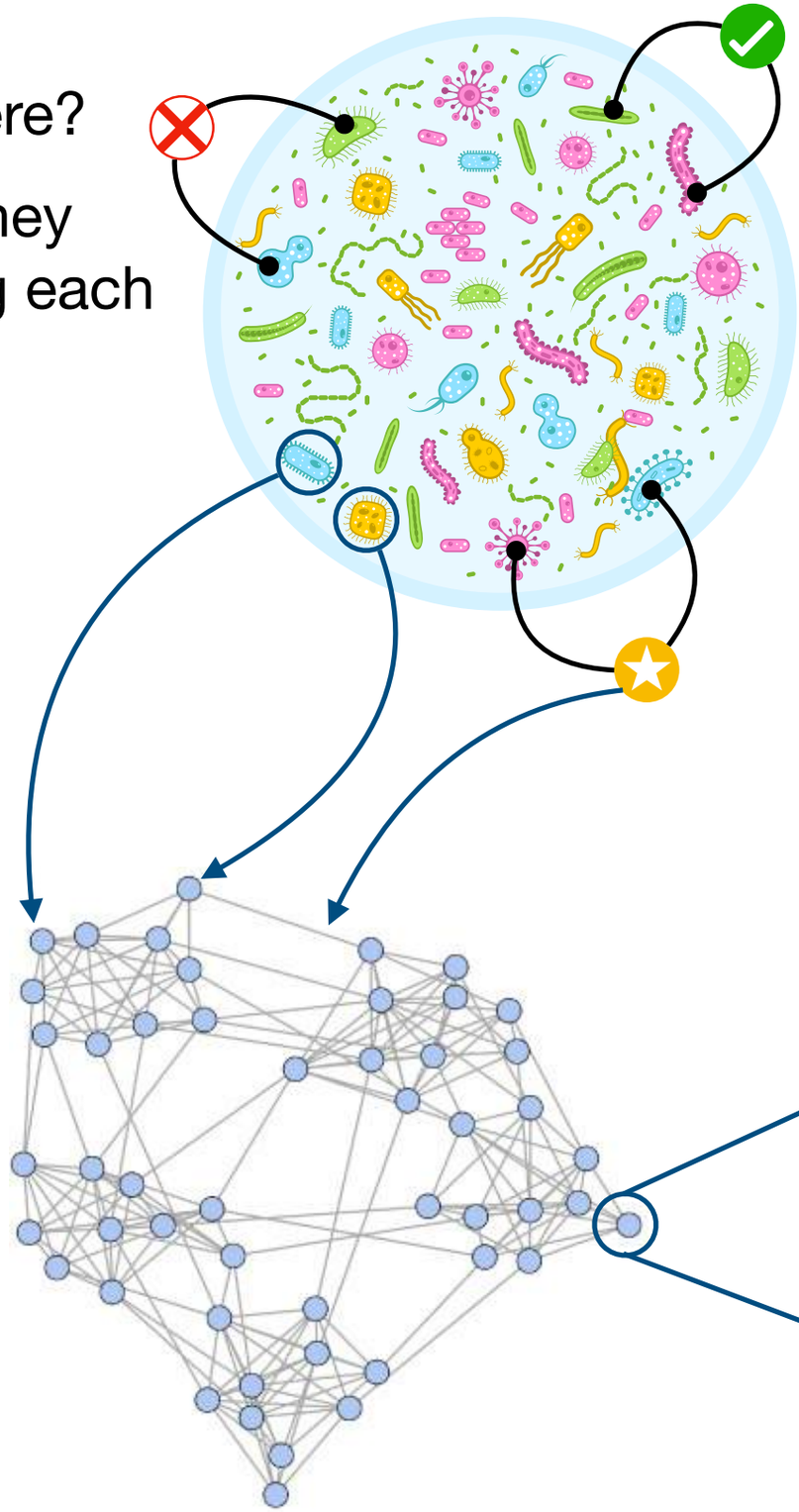
- Who is there?
- How are they influencing each other?



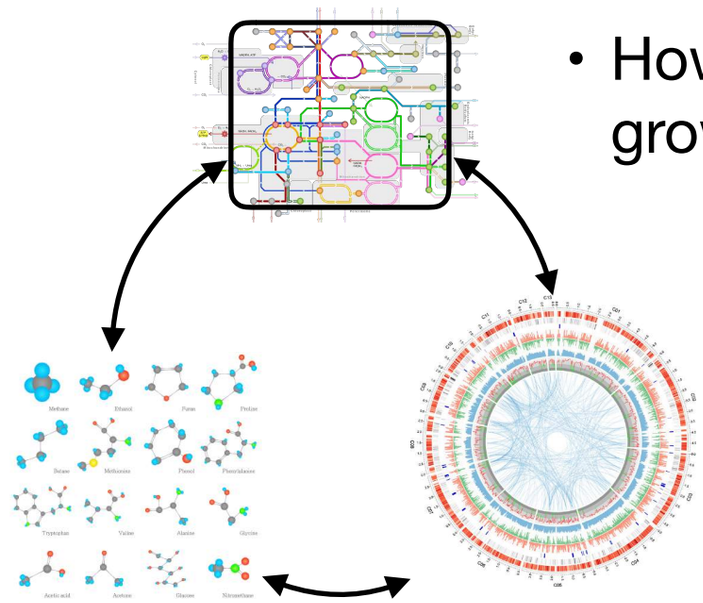


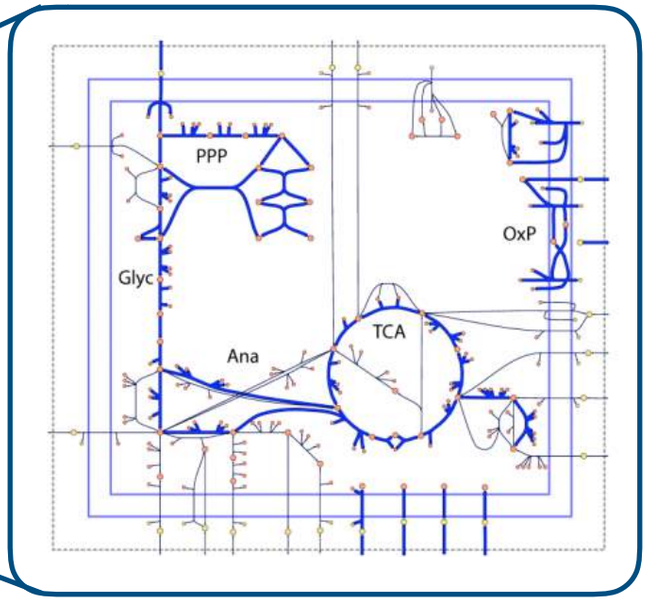
- What are they doing?
- How fast are they growing?

- Who is there?
- How are they influencing each other?



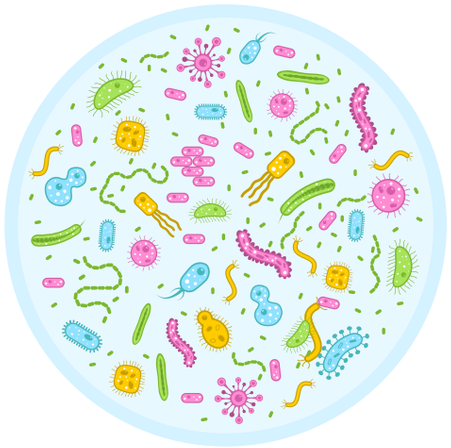
- What are they doing?
- How fast are they growing?





How can we model such complex interactions?

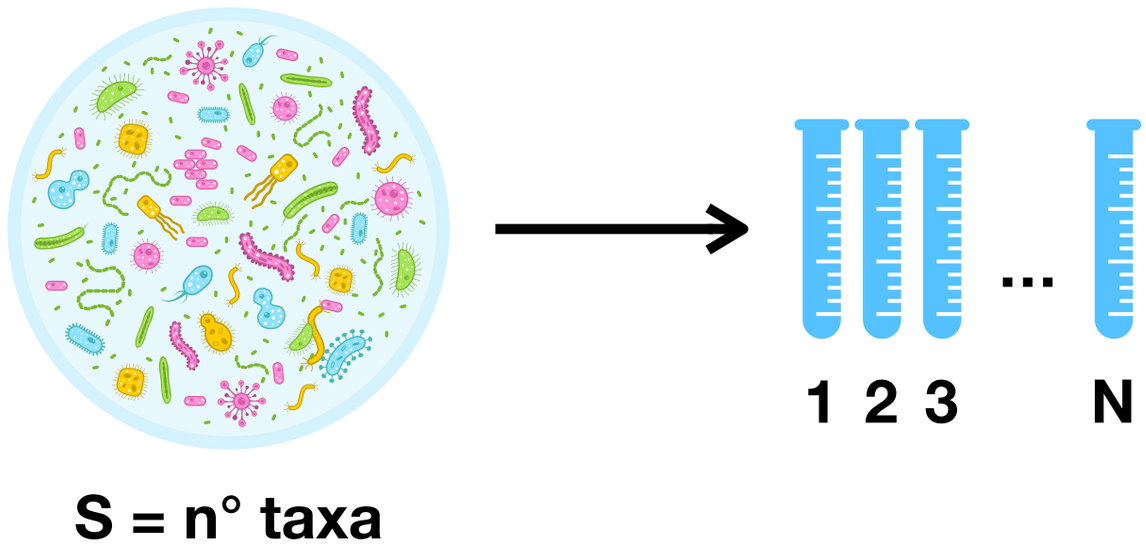
Community profiling



S = n° taxa

Community profiling

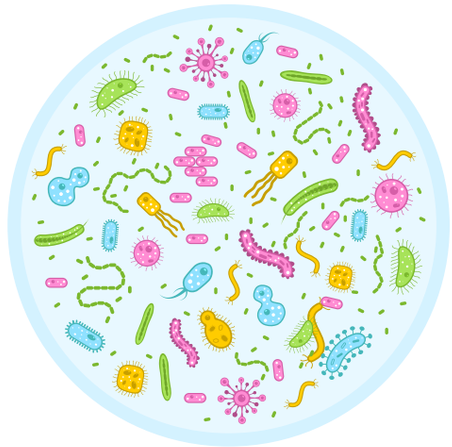
Community sampling



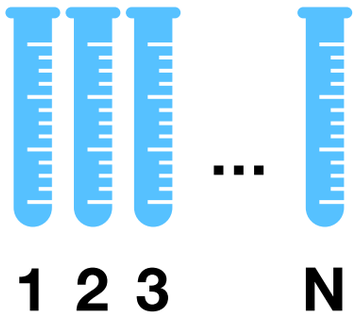
S = n° taxa

Community profiling

Community sampling



S = n° taxa



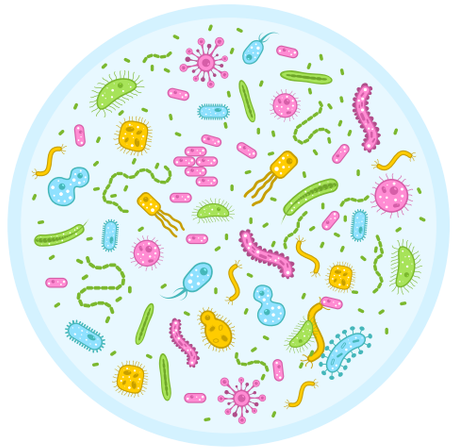
N x S matrix

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,S} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,S} \\ X_{3,1} & X_{3,2} & \cdots & X_{3,S} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \cdots & X_{N,S} \end{pmatrix}$$

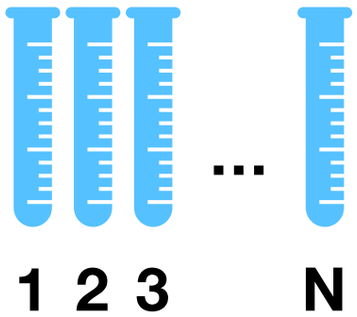
Taxonomic profiling

Community profiling

Community sampling



S = n° taxa



N x S matrix

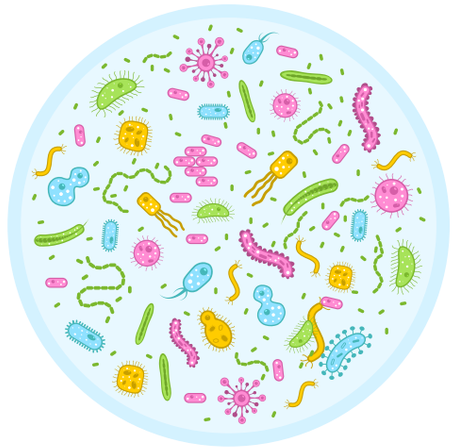
$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,S} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,S} \\ X_{3,1} & X_{3,2} & \cdots & X_{3,S} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \cdots & X_{N,S} \end{pmatrix}$$

Taxonomic profiling

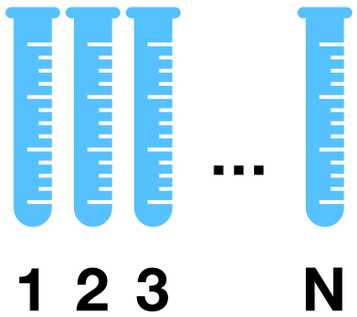
target gene (taxonomic marker)
shotgun metagenomics
T-RFLP
DGGE
...

Community profiling

Community sampling



S = n° taxa

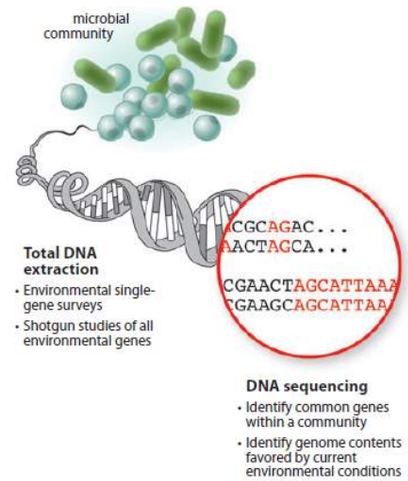


N x S matrix

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,S} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,S} \\ X_{3,1} & X_{3,2} & \cdots & X_{3,S} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \cdots & X_{N,S} \end{pmatrix}$$

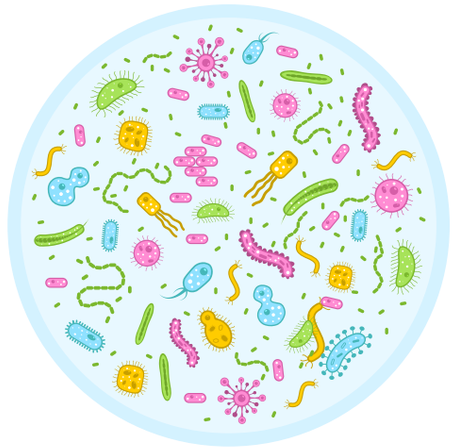
Taxonomic profiling

target gene (taxonomic marker)
shotgun metagenomics
T-RFLP
DGGE
...

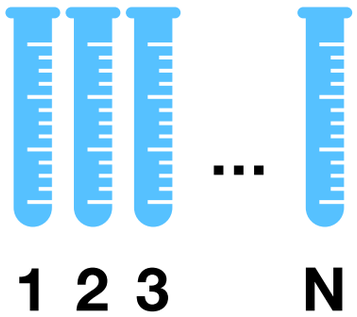


Community profiling

Community sampling



S = n° taxa



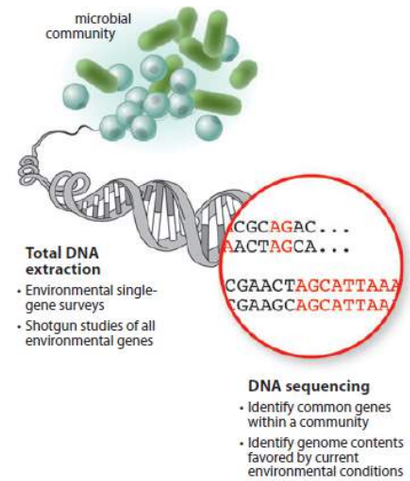
N x S matrix

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,S} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,S} \\ X_{3,1} & X_{3,2} & \cdots & X_{3,S} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \cdots & X_{N,S} \end{pmatrix}$$

How can we use quantitative information to infer community structure?

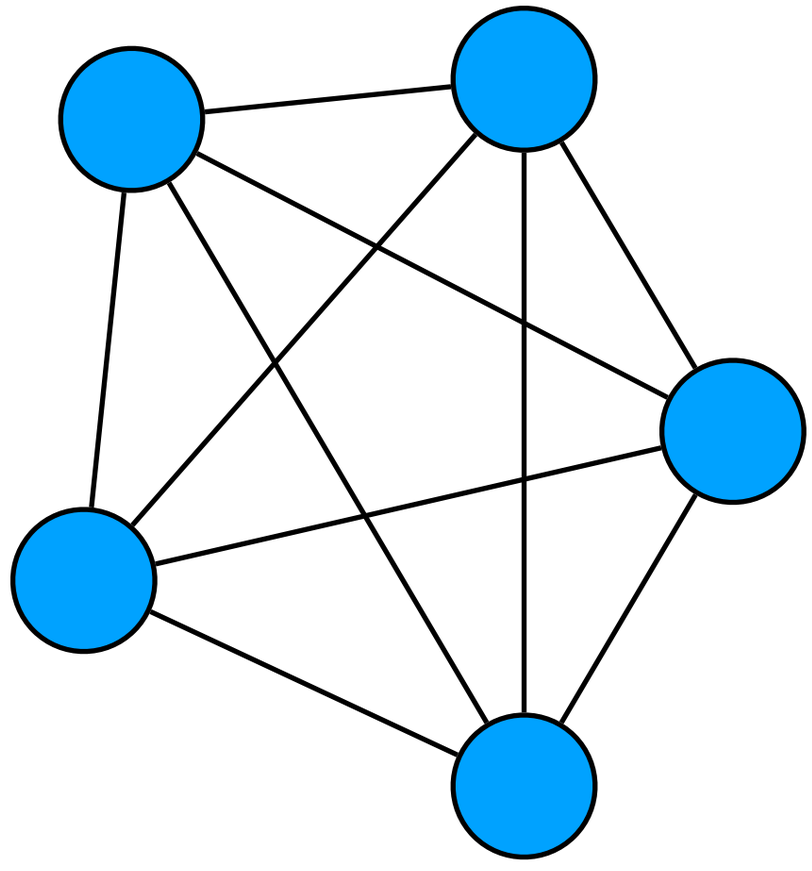
Taxonomic profiling

target gene (taxonomic marker)
shotgun metagenomics
T-RFLP
DGGE
…



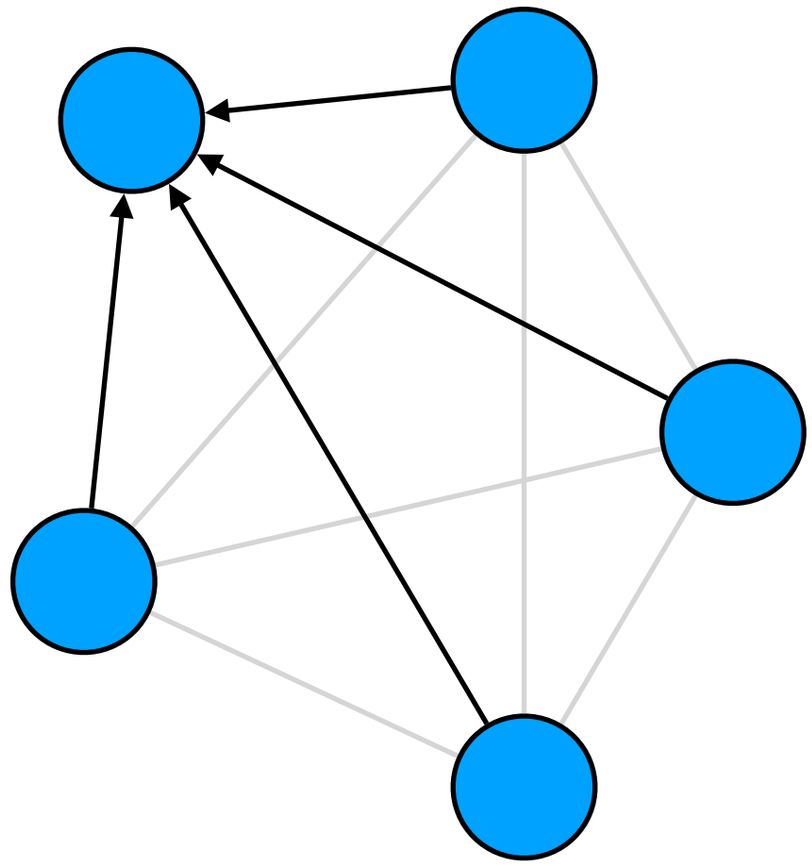
Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network



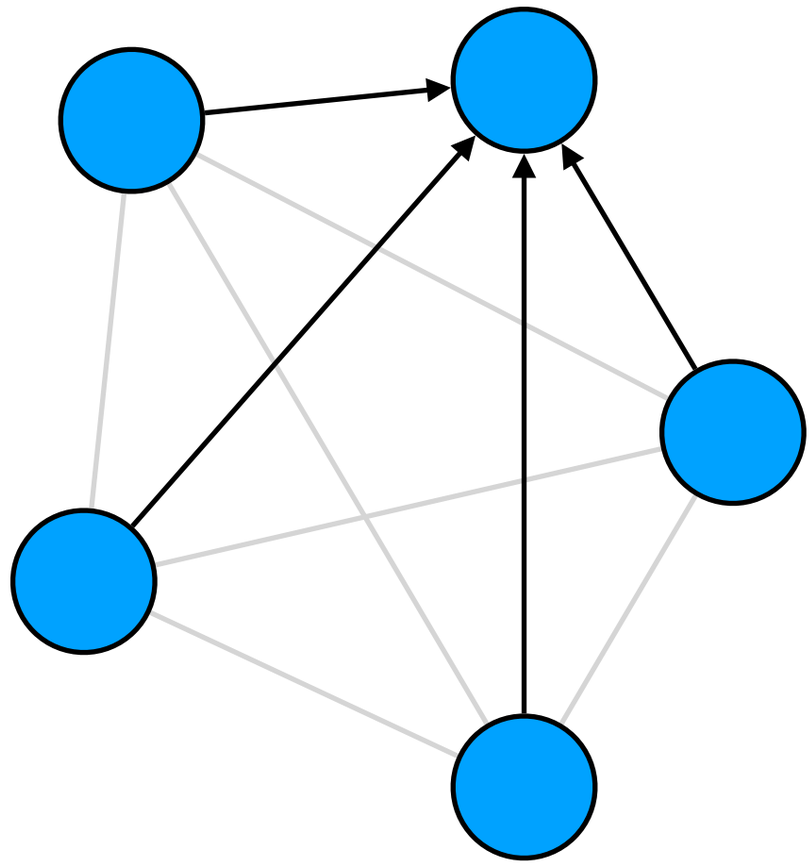
Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network



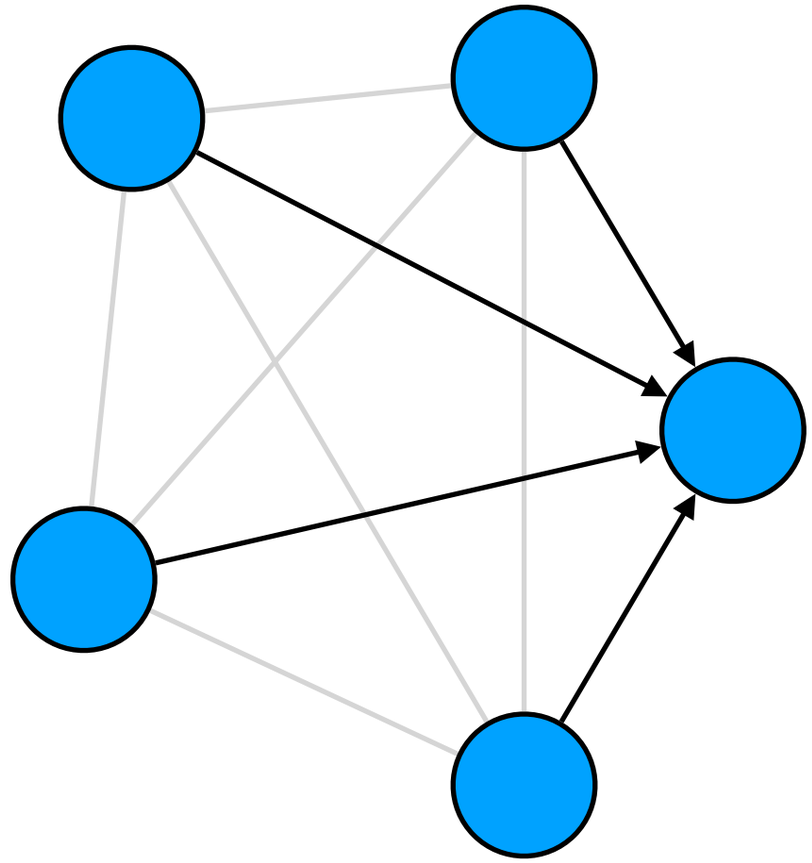
Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network



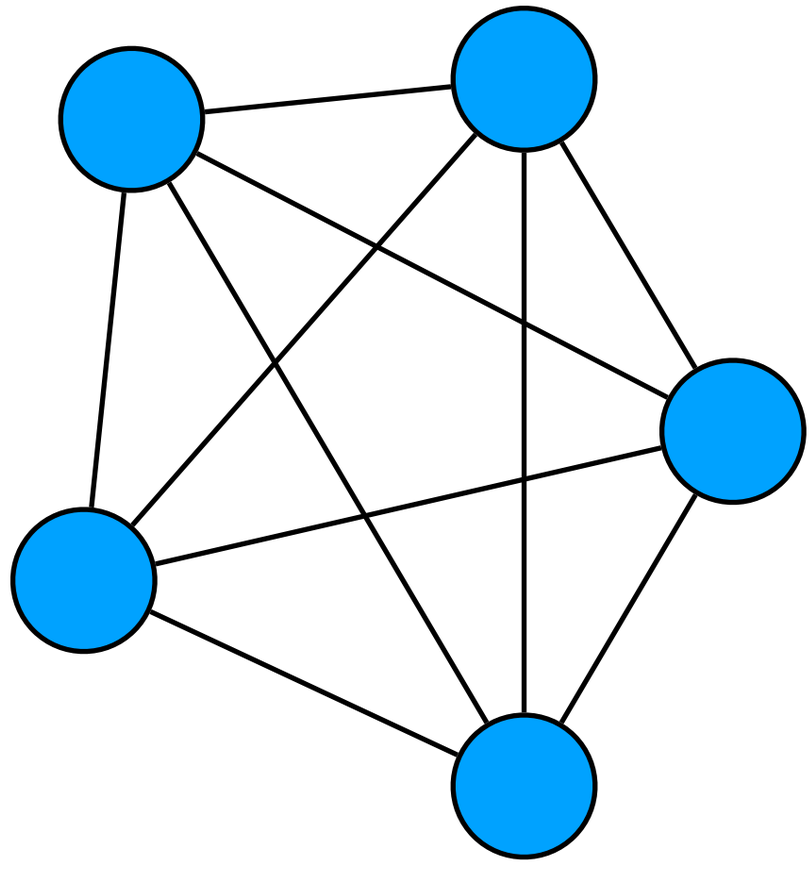
Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network



Machine Learning

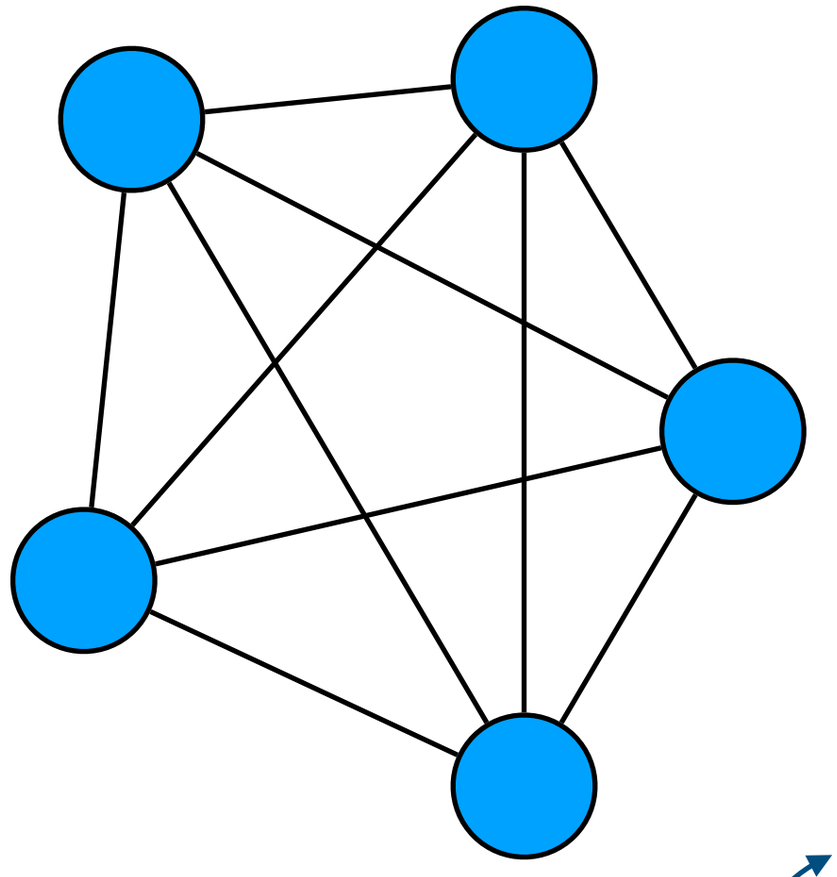
- All nodes (taxa) are predicted using all other nodes in the network
- The effect of each predictor is estimated using root mean squared error



$$\mathbb{R}^{NxN} = \begin{pmatrix} - & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & - & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & - \end{pmatrix}$$

Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network
- The effect of each predictor is estimated using root mean squared error

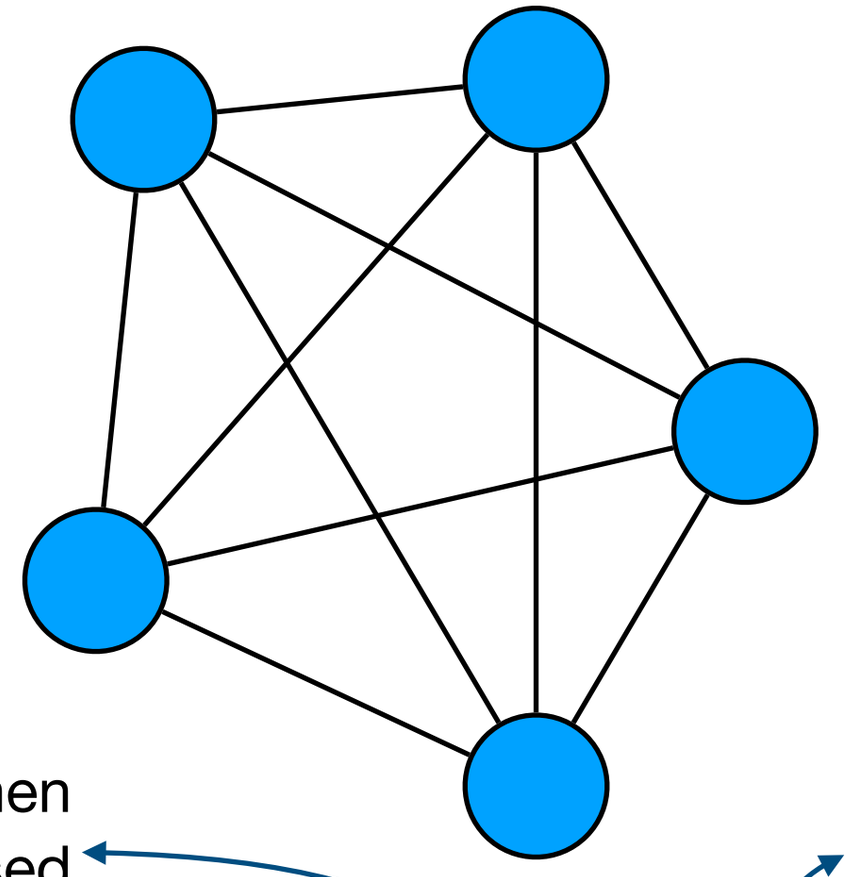


#nodes

$$\mathbb{R}^{NxN} = \begin{pmatrix} - & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & - & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & - \end{pmatrix}$$

Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network
- The effect of each predictor is estimated using root mean squared error



Δ(RMSE) when node 2 is not used in predicting node 1

#nodes

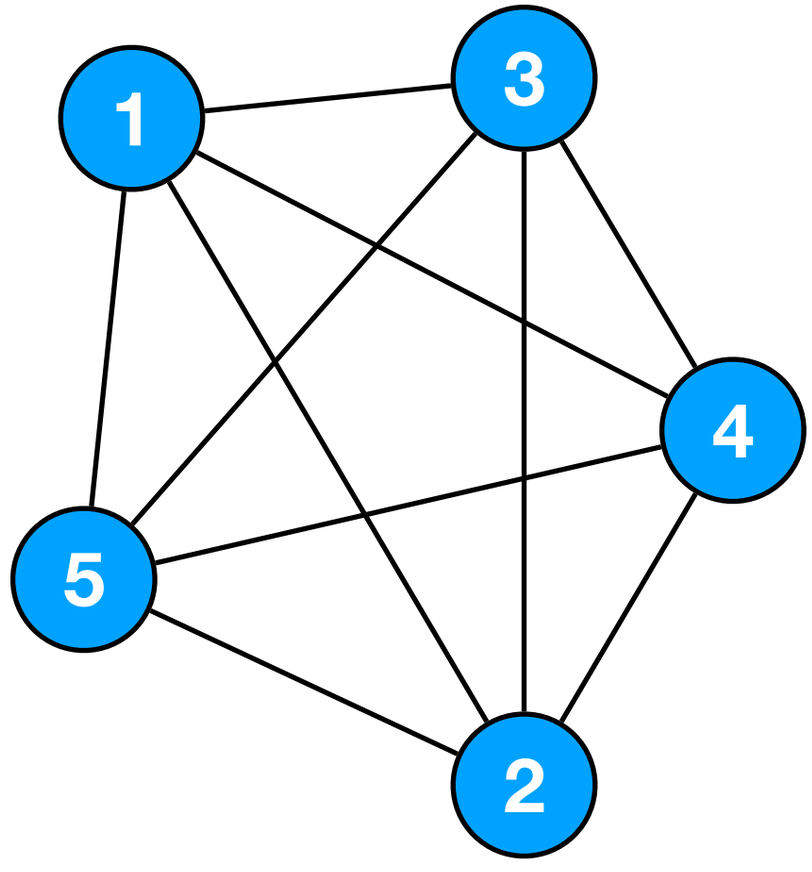
$$\mathbb{R}^{NxN} = \begin{pmatrix} - & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & - & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & - \end{pmatrix}$$

Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network

- The effect of each predictor is estimated using root mean squared error

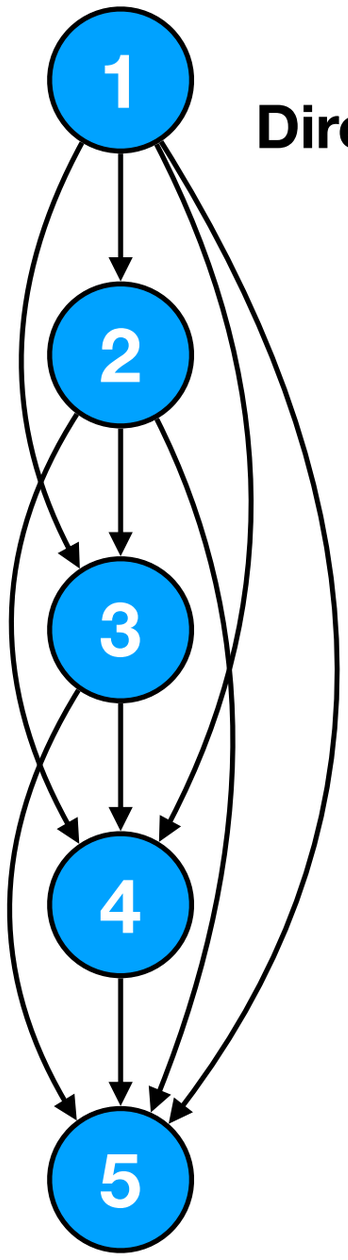
- Nodes are ranked according to their prediction power



$$\mathbb{R}^{NxN} = \begin{pmatrix} - & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & - & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & - \end{pmatrix}$$

Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network
- The effect of each predictor is estimated using root mean squared error
- Nodes are ranked according to their prediction power
- The system is optimised using genetic algorithm



Directed Acyclic Graph
DAG

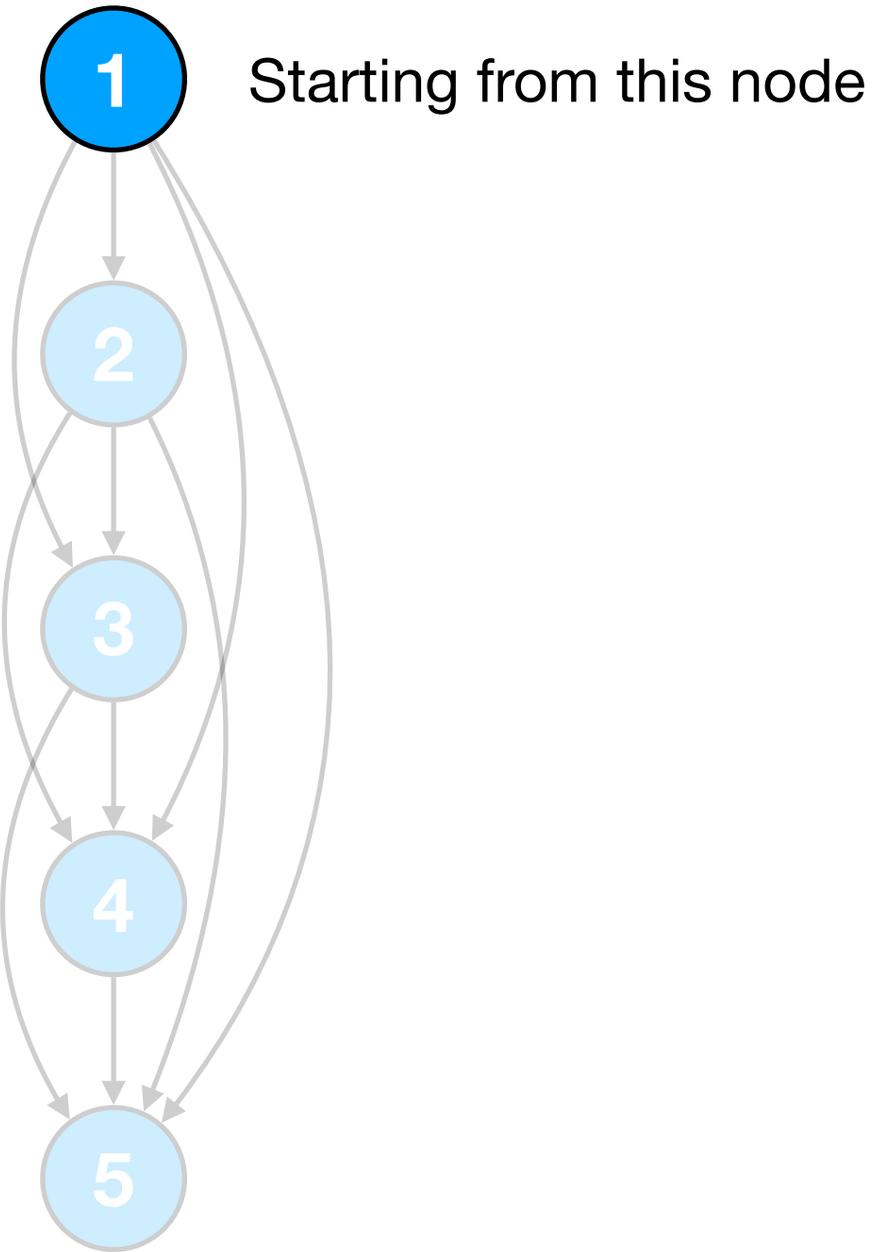
Machine Learning

- All nodes (taxa) are predicted using all other nodes in the network

- The effect of each predictor is estimated using root mean squared error

- Nodes are ranked according to their prediction power

- The system is optimised using genetic algorithm



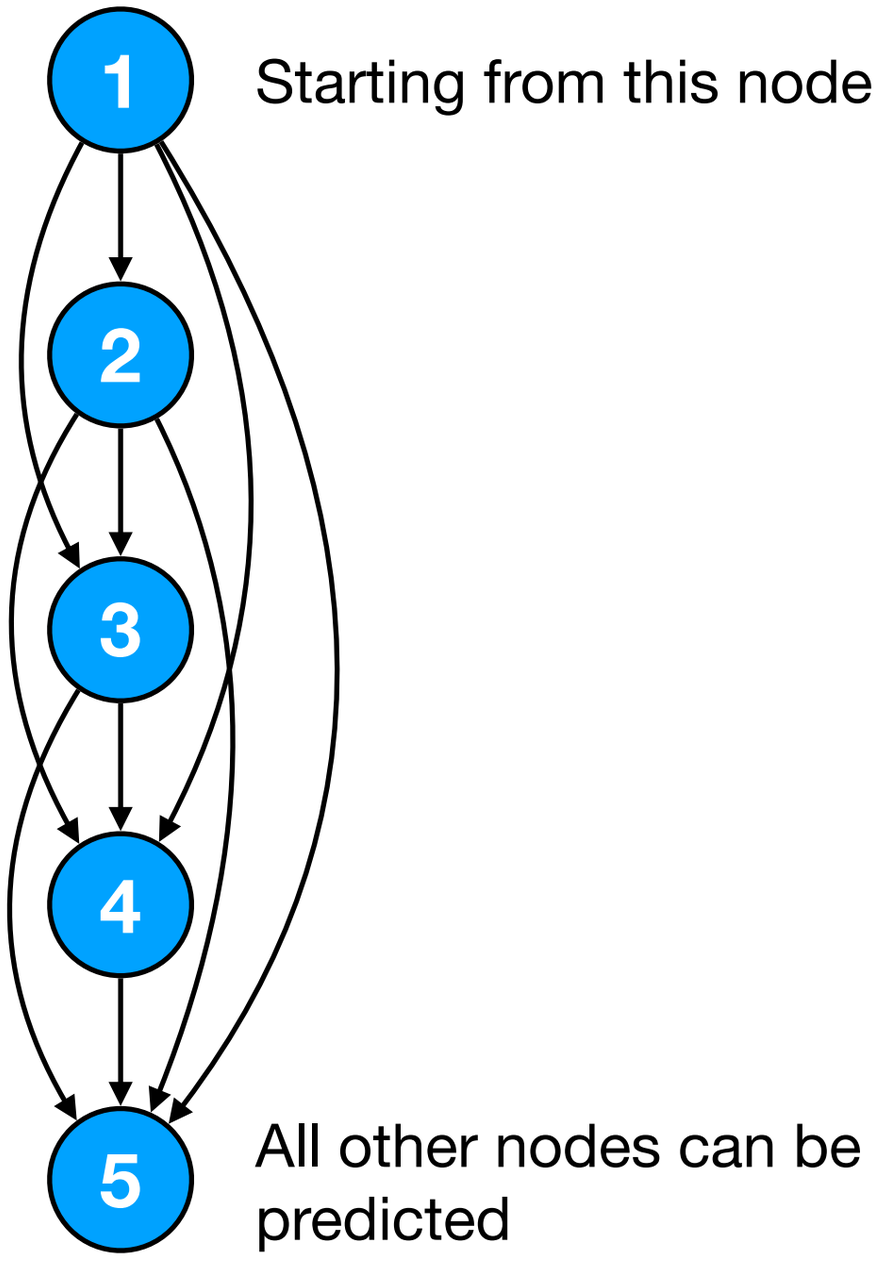
Machine Learning

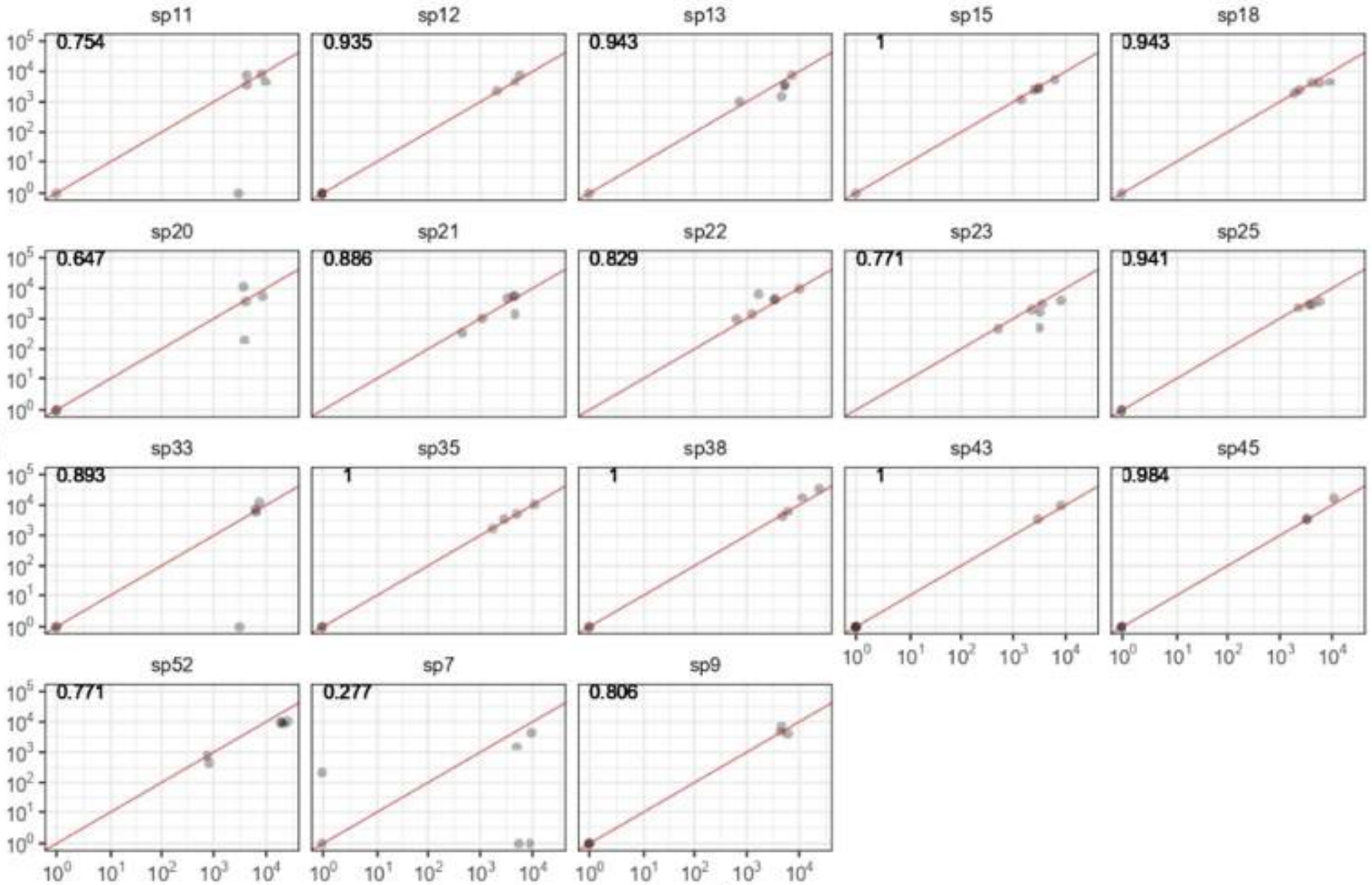
- All nodes (taxa) are predicted using all other nodes in the network

- The effect of each predictor is estimated using root mean squared error

- Nodes are ranked according to their prediction power

- The system is optimised using genetic algorithm





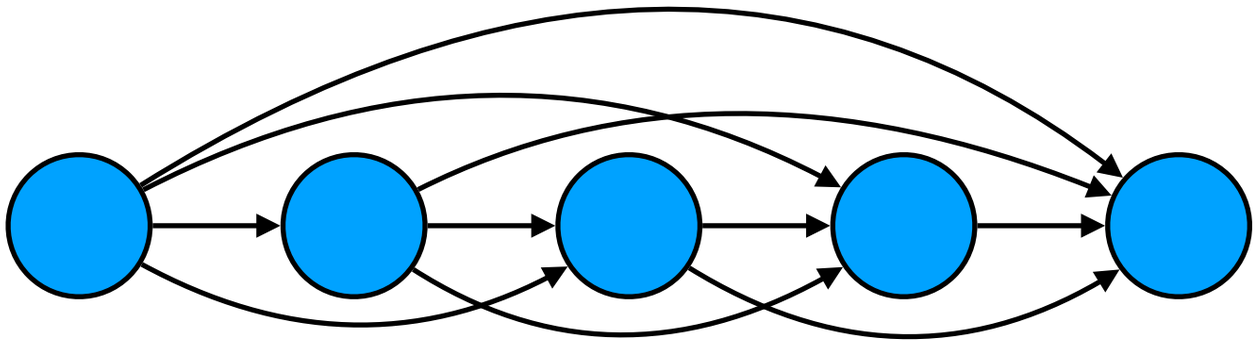
Preliminary results

R^2 value ranging from 0.65 to 1 (except for sp7)

Thaiss et al 2016 - Persistent microbiome alterations modulate the rate of post-dieting weight regain

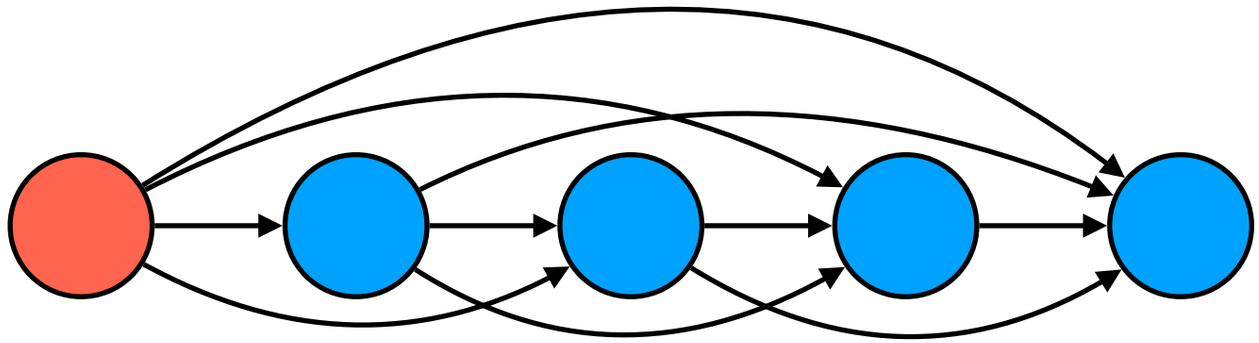
What about metabolism?

Community modulation

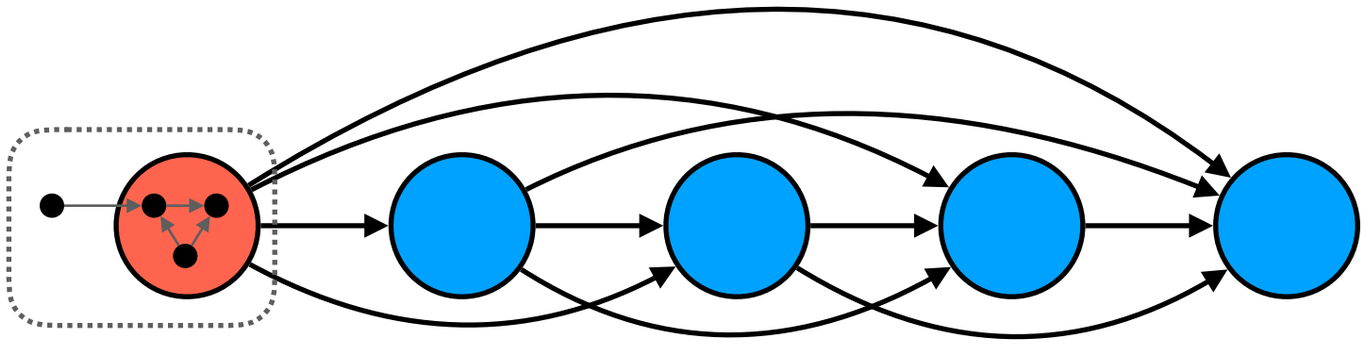


Community modulation

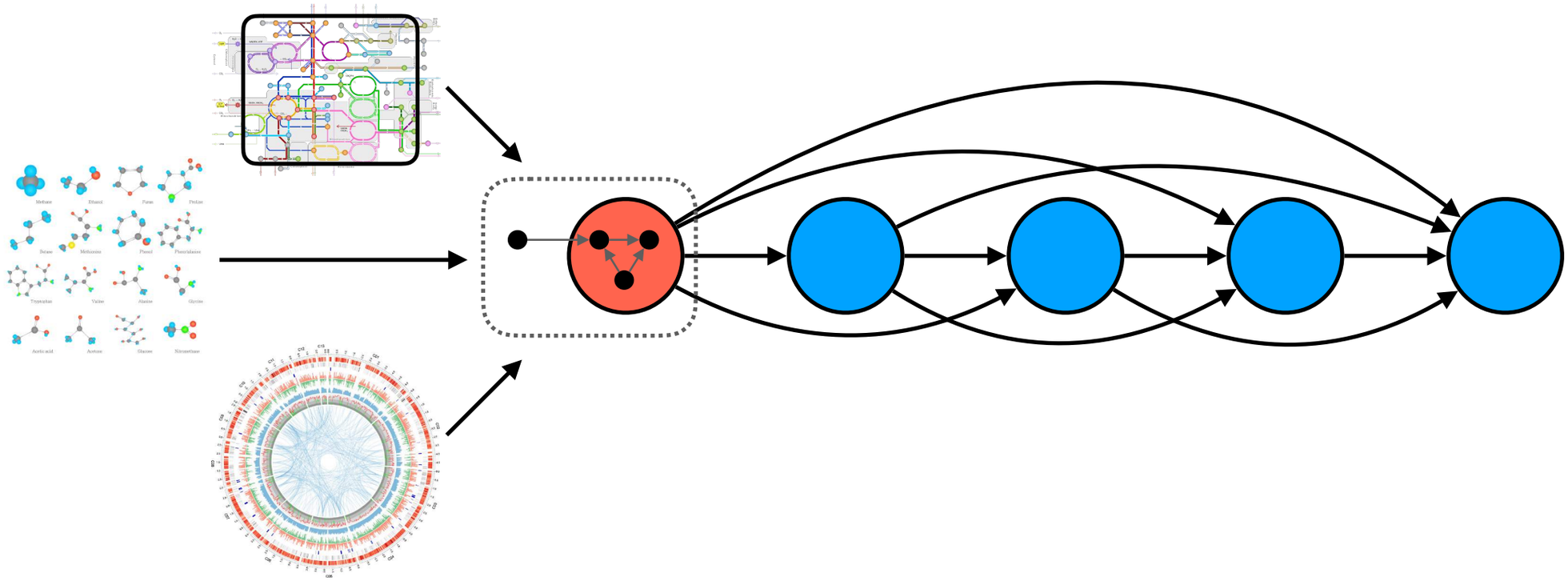
Not predicted by the model



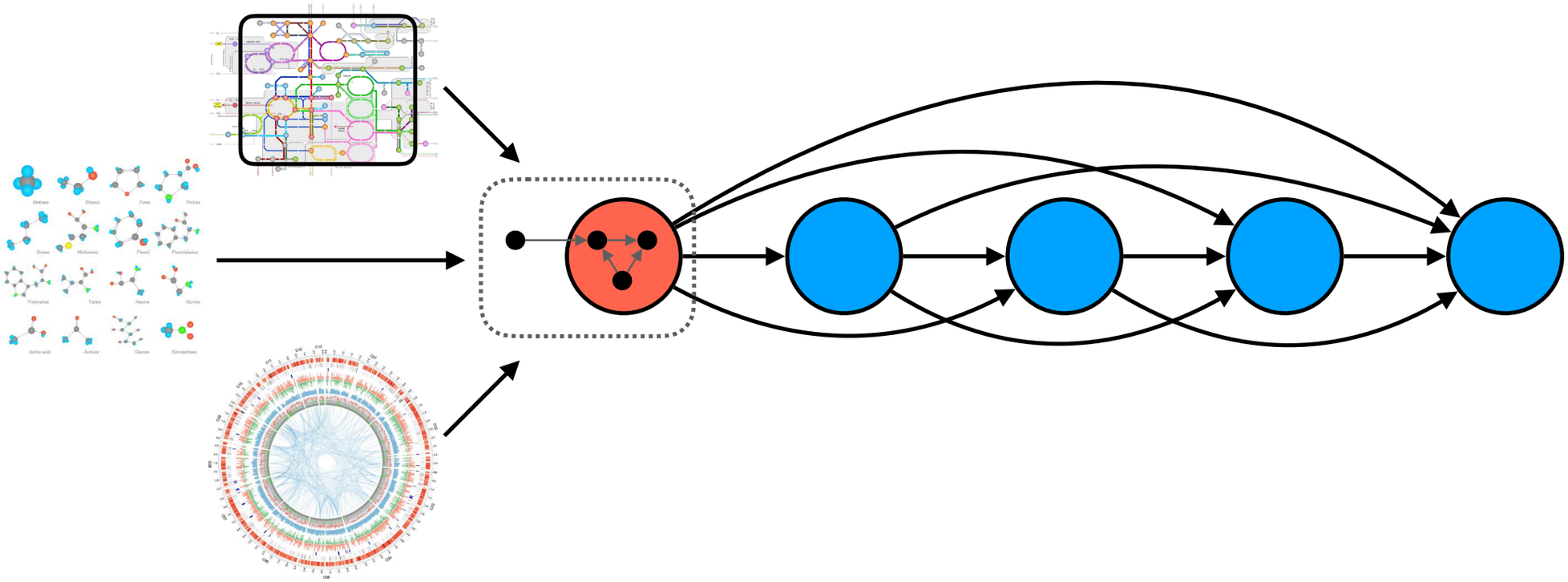
Community modulation



Community modulation

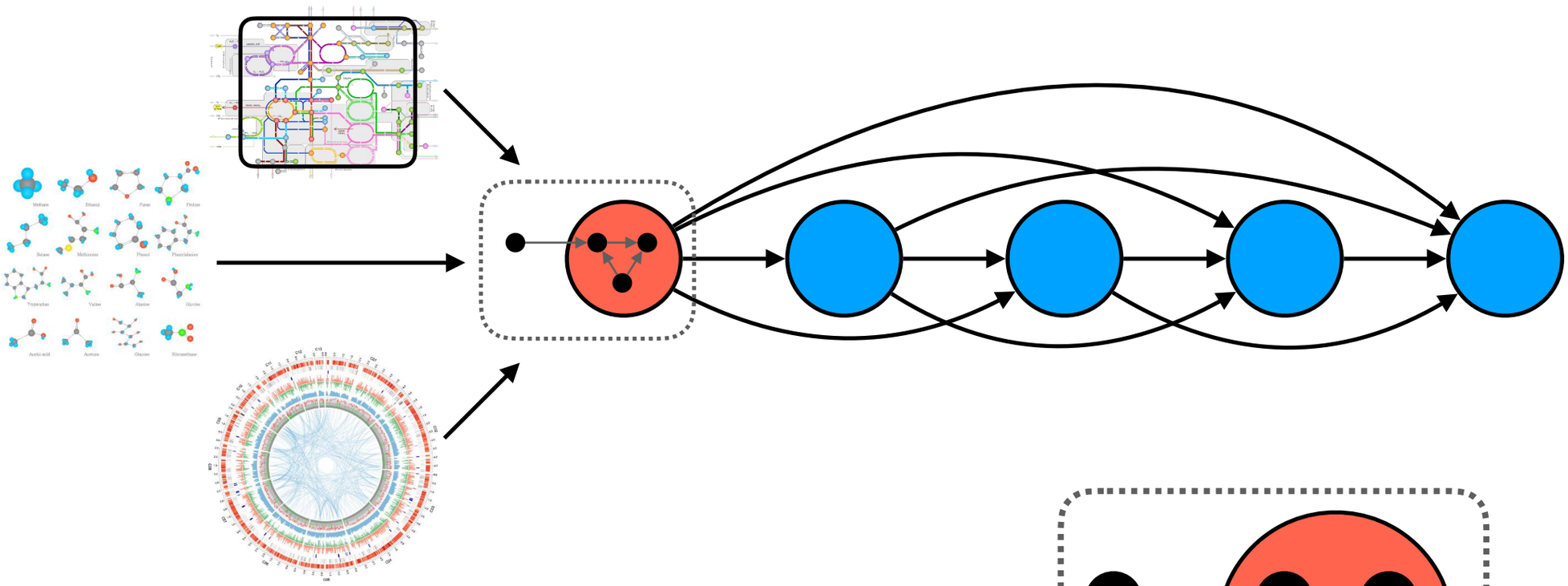


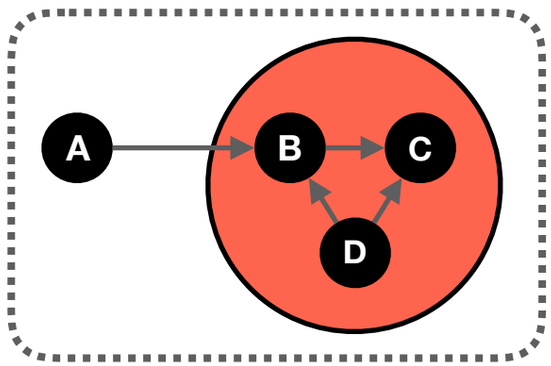
Community modulation



- Nutrient effect simulation
- Effect of antibiotic molecules
- Changes in metabolic assets

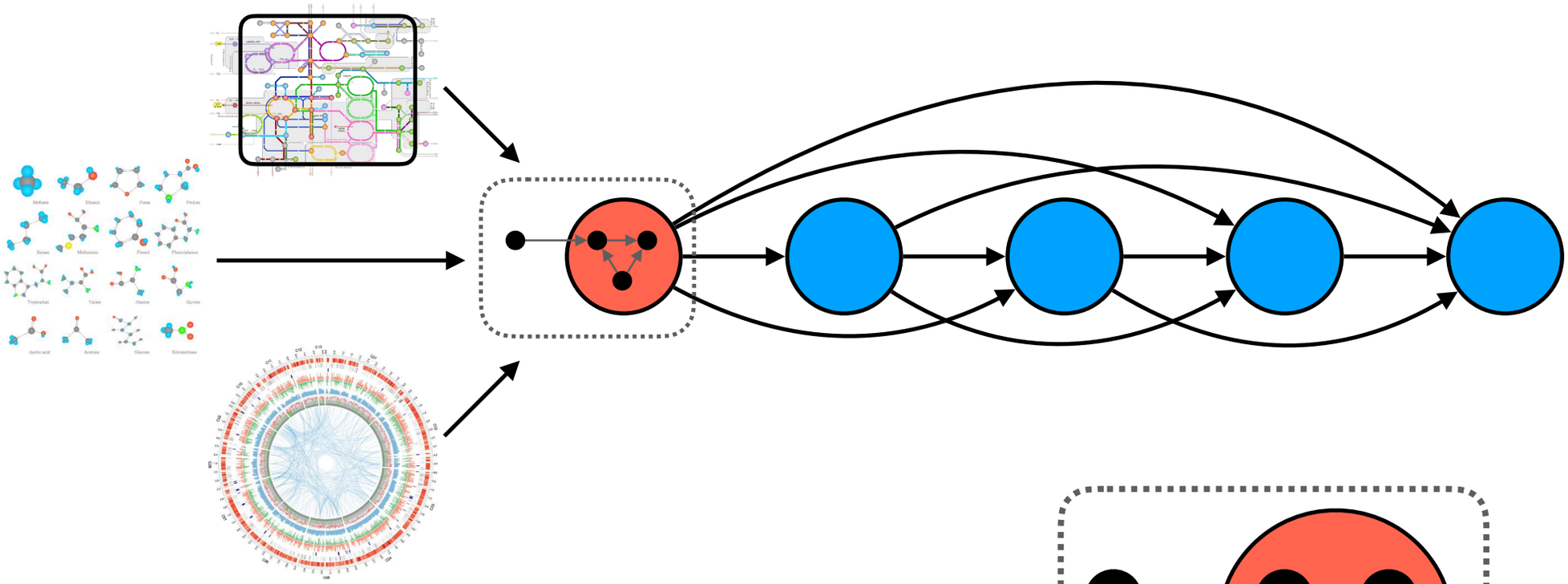
Community modulation

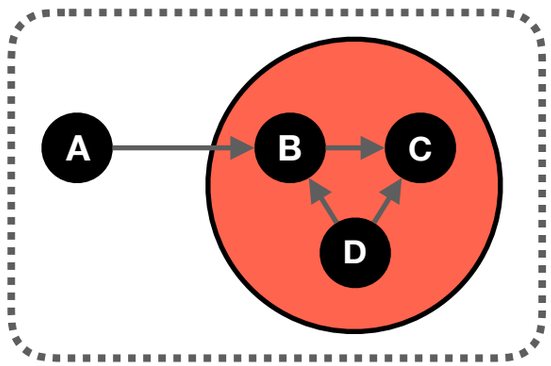




- Nutrient effect simulation
- Effect of antibiotic molecules
- Changes in metabolic assets

Community modulation

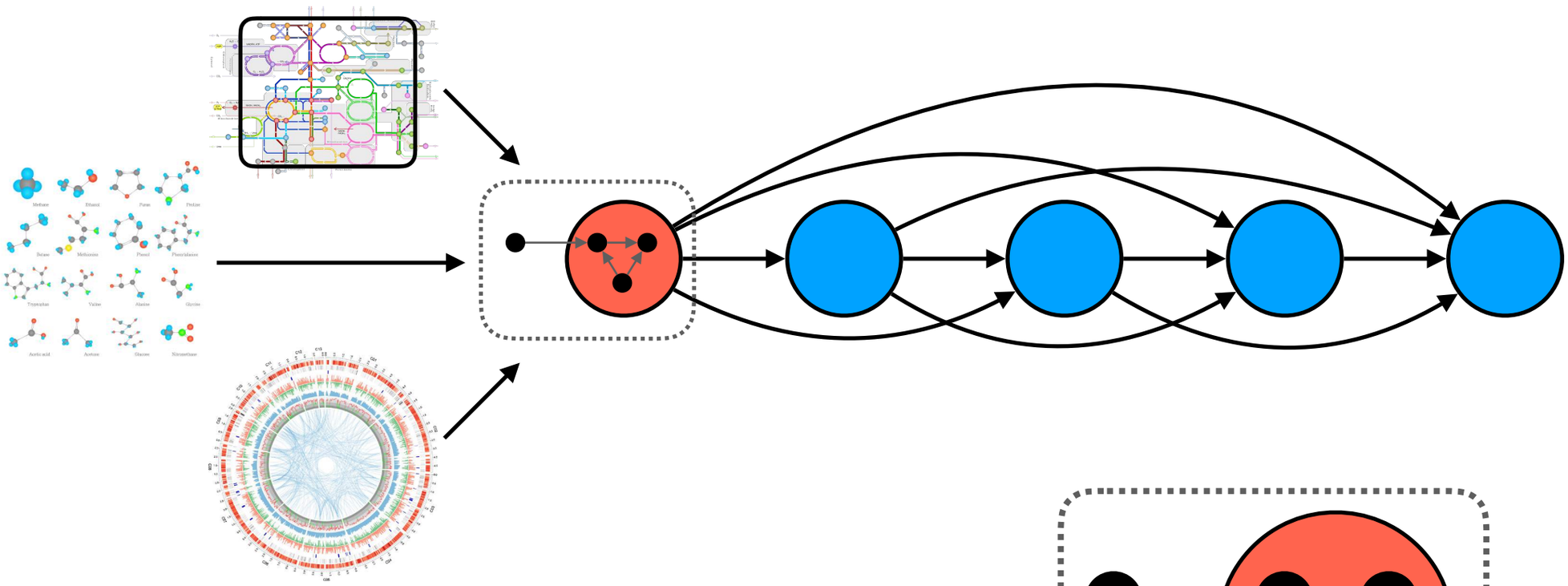




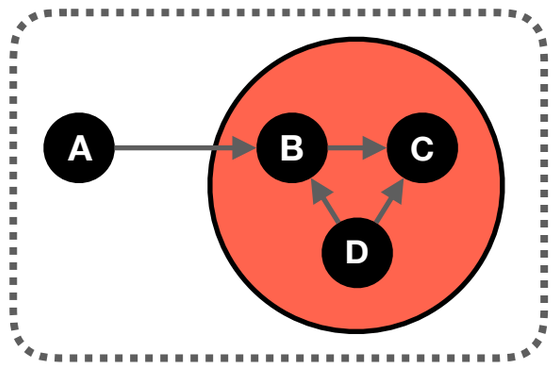
$$Biomass = f(A) = \mu(h^{-1})$$

- Nutrient effect simulation
- Effect of antibiotic molecules
- Changes in metabolic assets

Community modulation



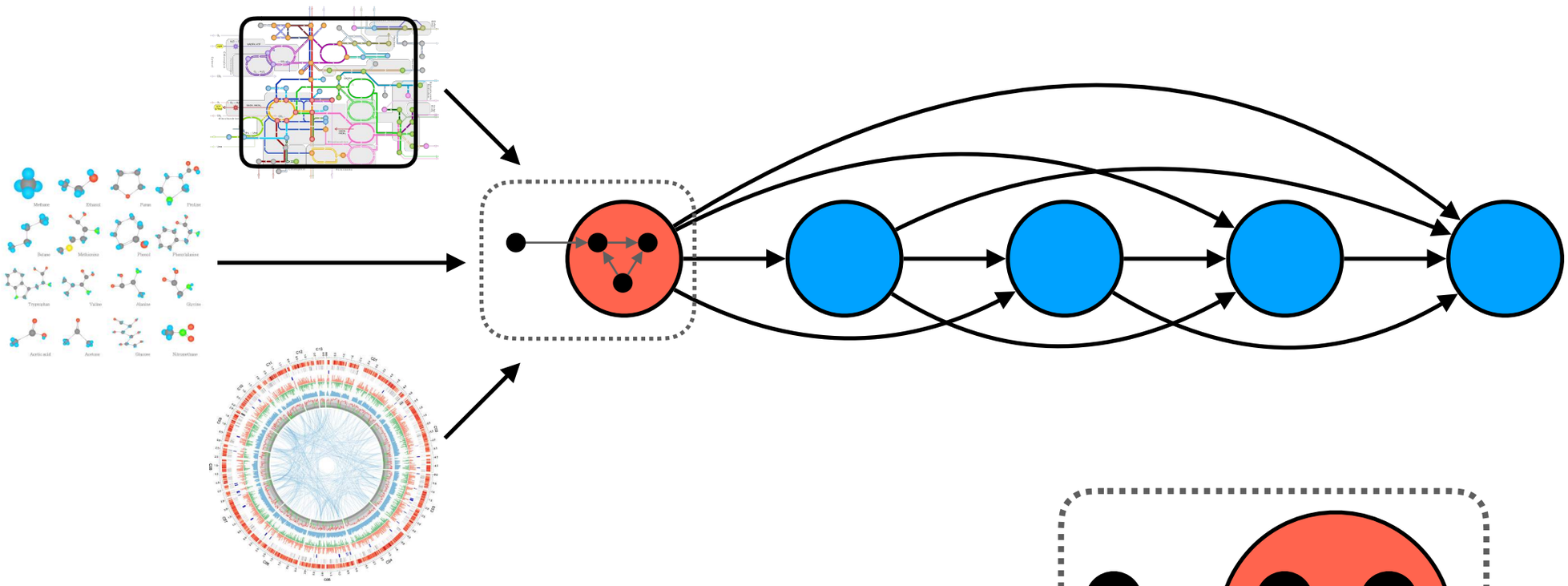
- Nutrient effect simulation
- Effect of antibiotic molecules
- Changes in metabolic assets

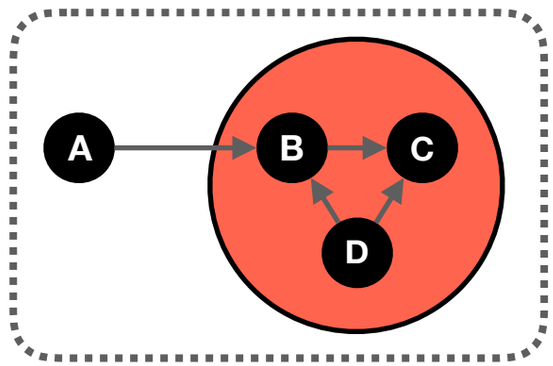


$$Biomass = f(A) = \mu(h^{-1})$$

$$Coverage = n_{reads}$$

Community modulation



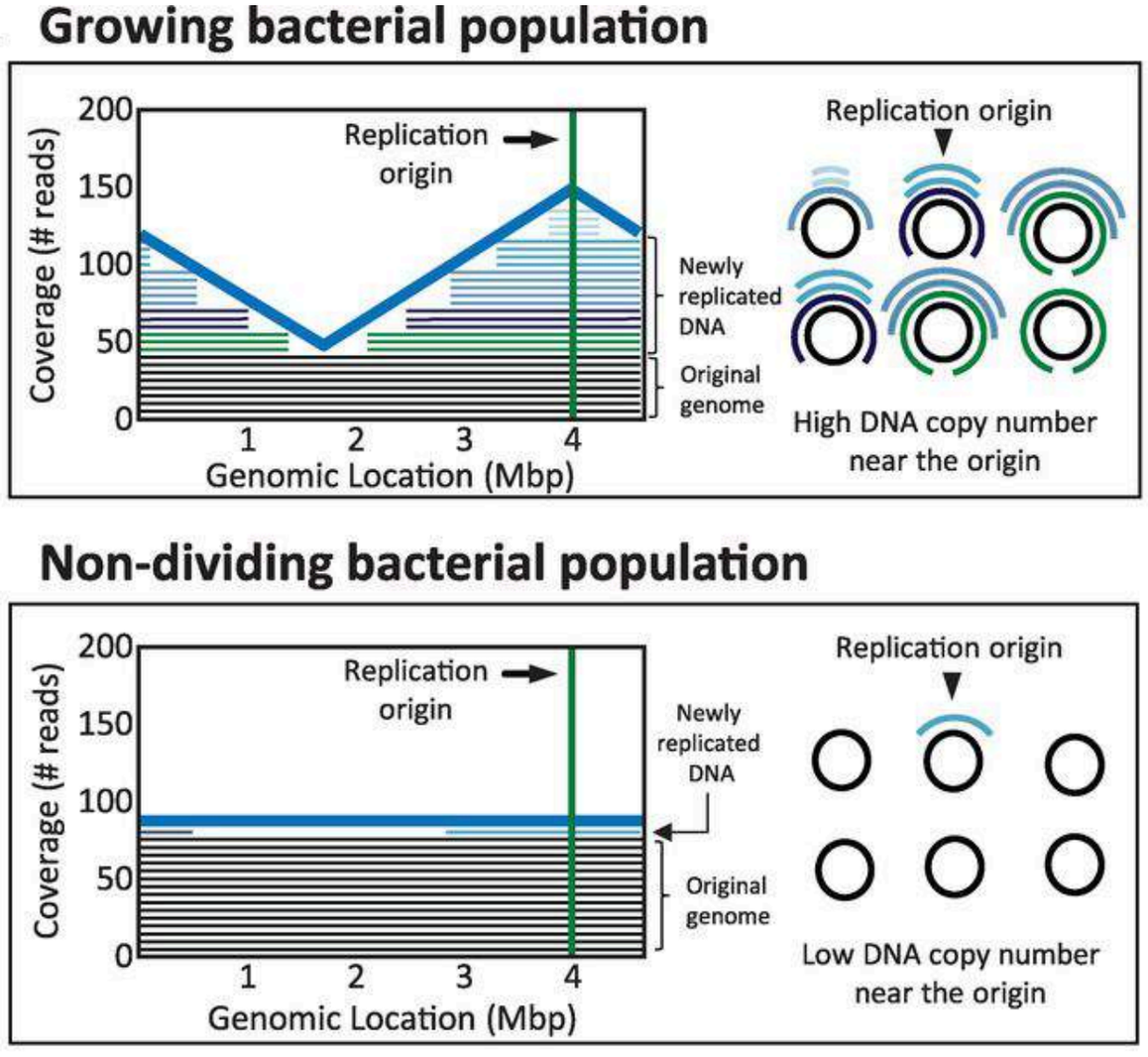


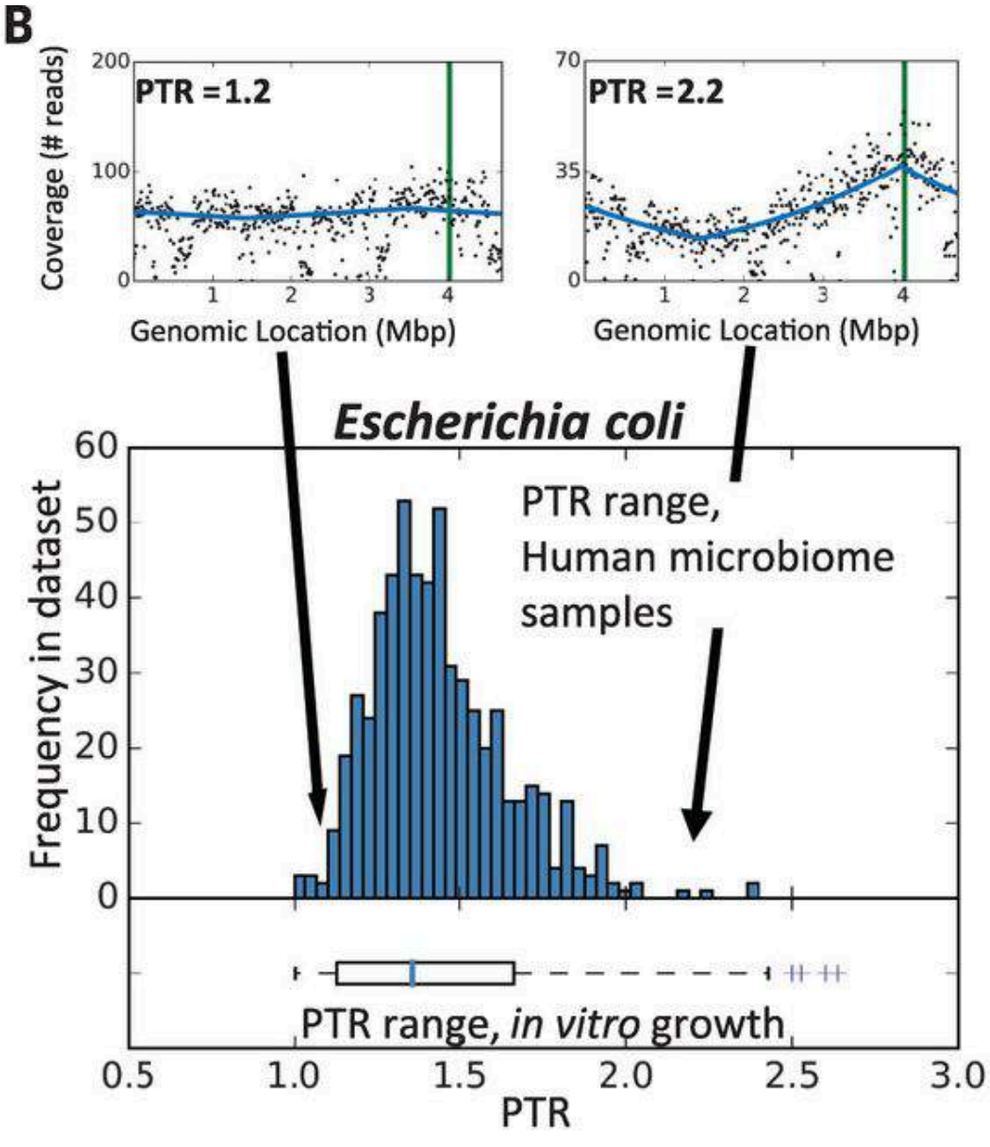
- Nutrient effect simulation
- Effect of antibiotic molecules
- Changes in metabolic assets

$$Biomass = f(A) = \mu(h^{-1})$$

$$Coverage = n_{reads}$$

$$Coverage = f(Biomass)$$

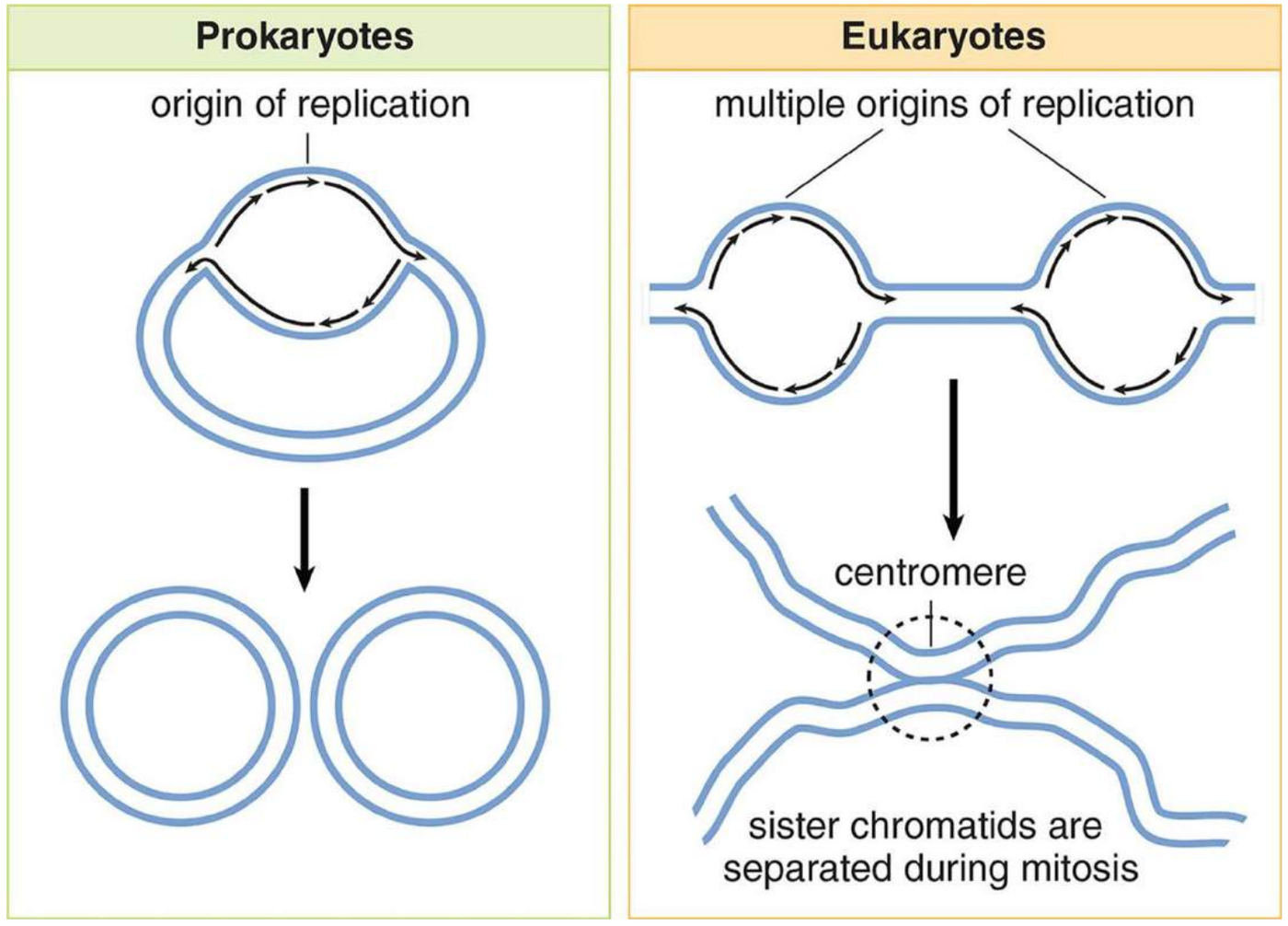




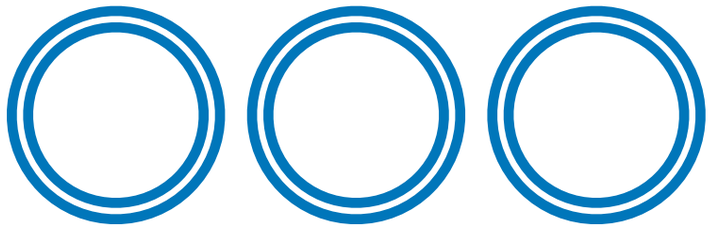
Growth rates estimation from metagenomes

Korem et al 2015 - Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples

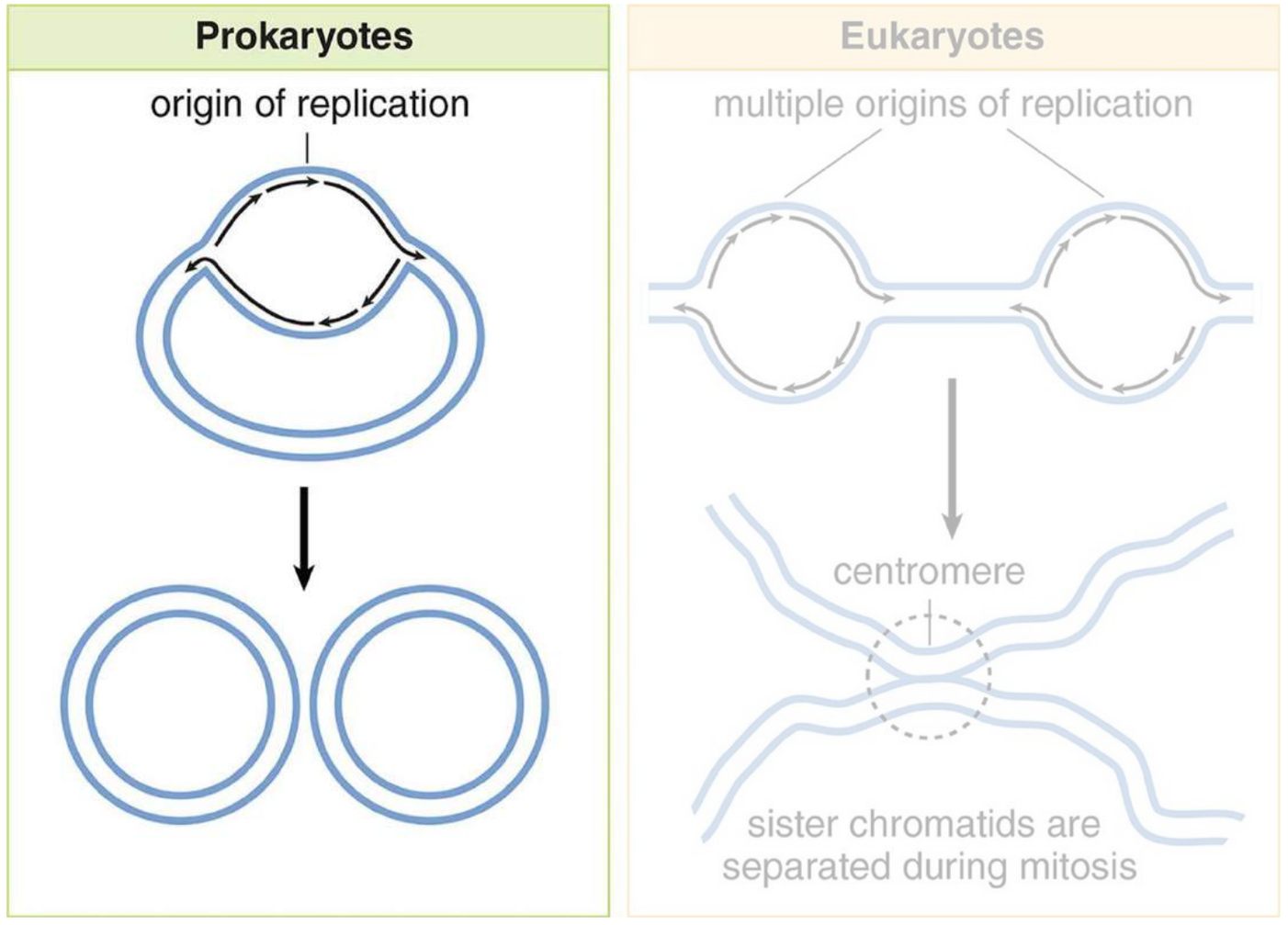
Prokaryotic DNA replication



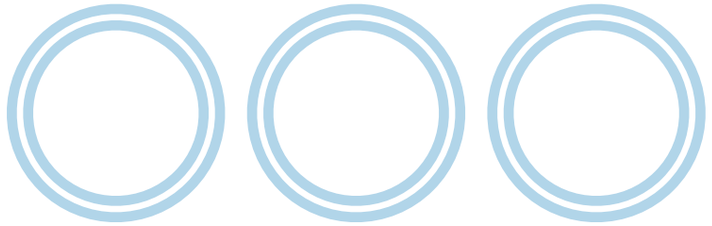
Prokaryotic DNA replication



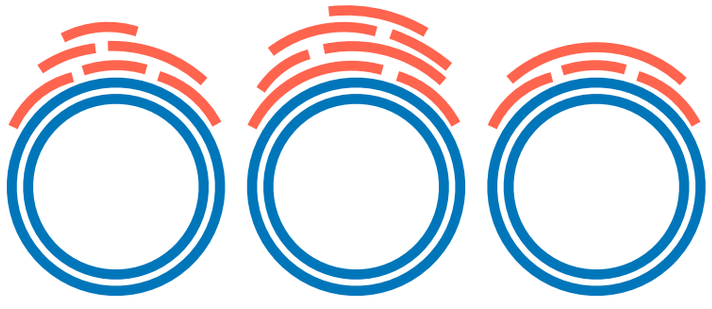
During stationary phase coverage is evenly distributed along the entire genome

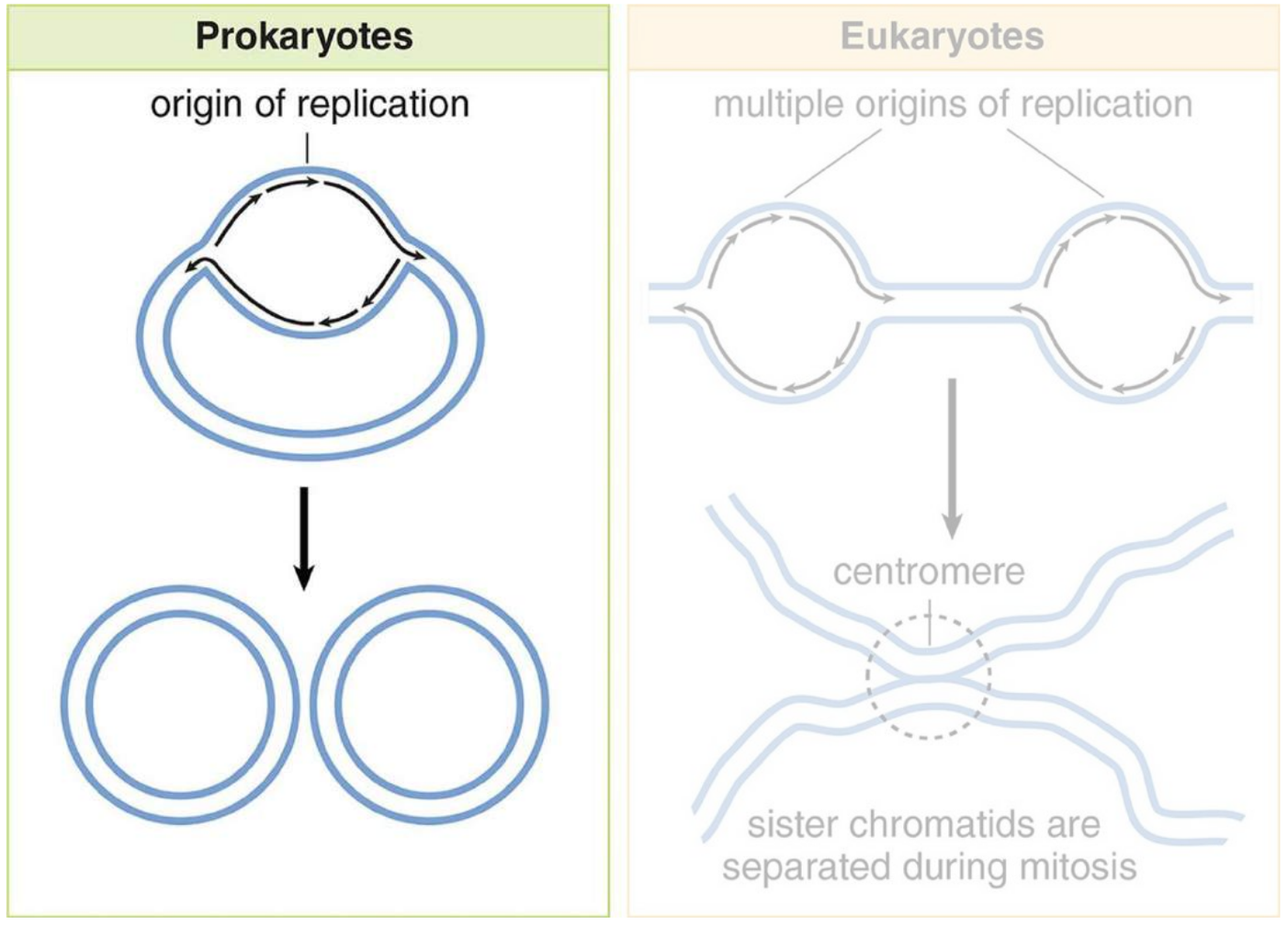


Prokaryotic DNA replication

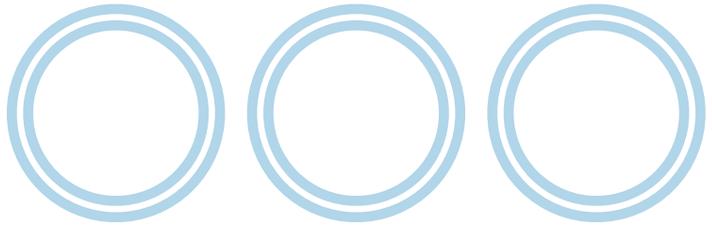


During exponential phase coverage increases near the origin of replication

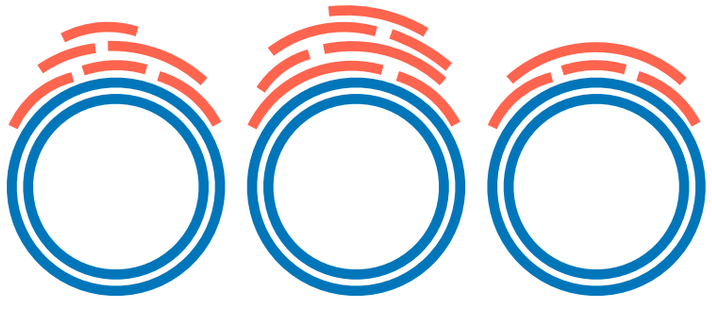




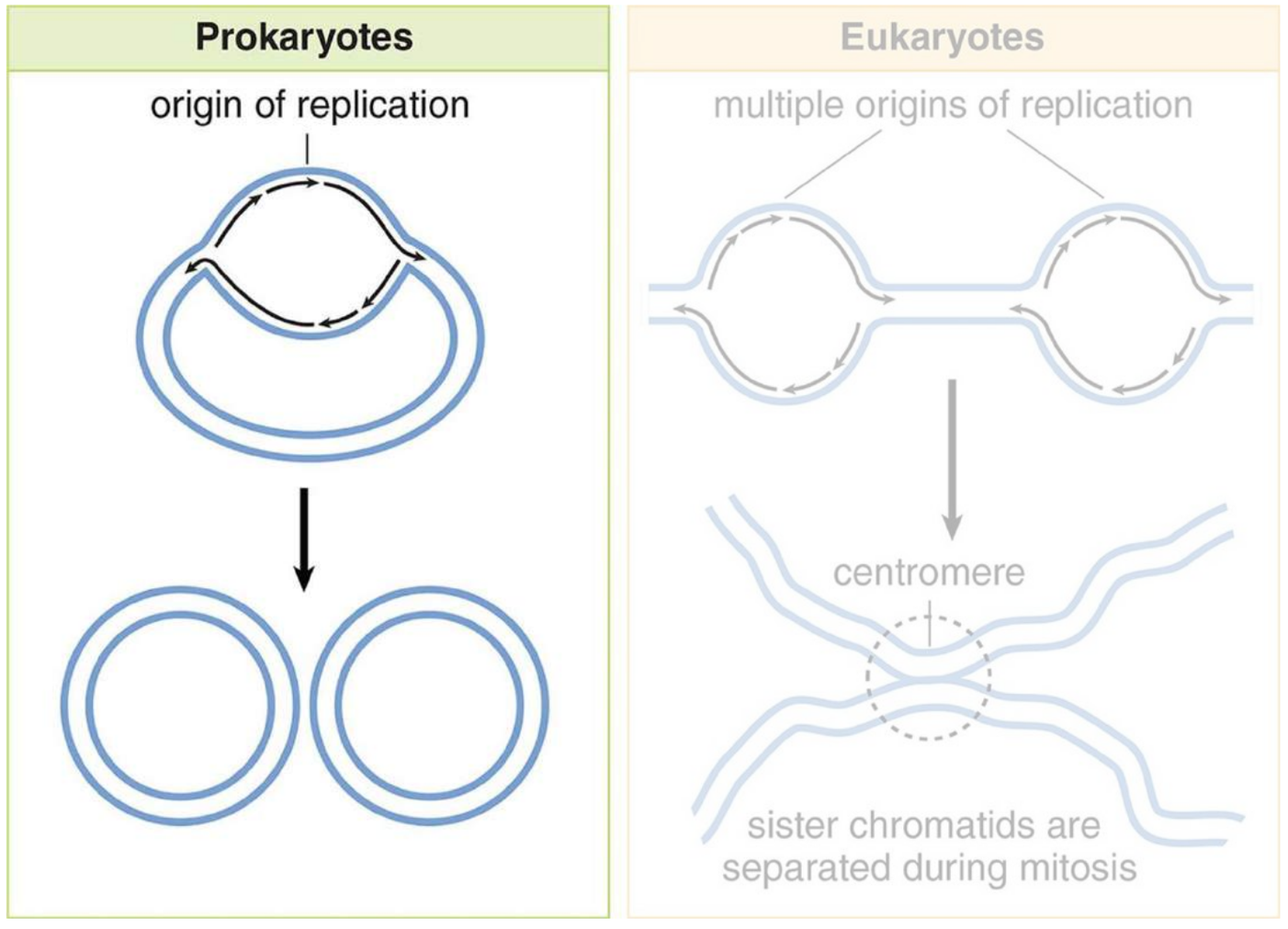
Prokaryotic DNA replication



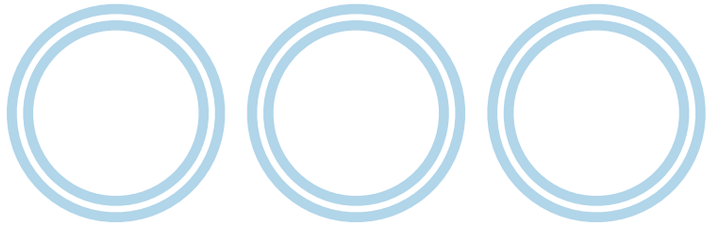
During exponential phase coverage increases near the origin of replication



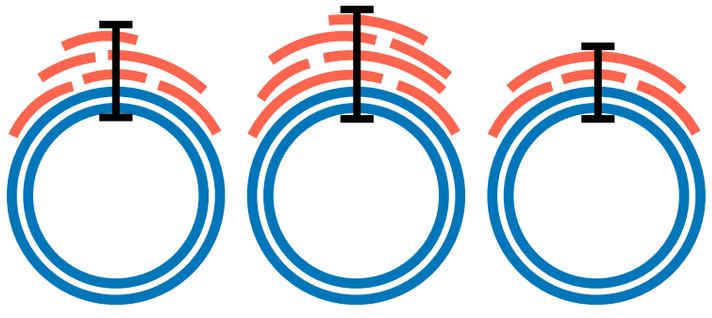
$$PTR = \frac{2^{ori}}{2^{ter}}$$

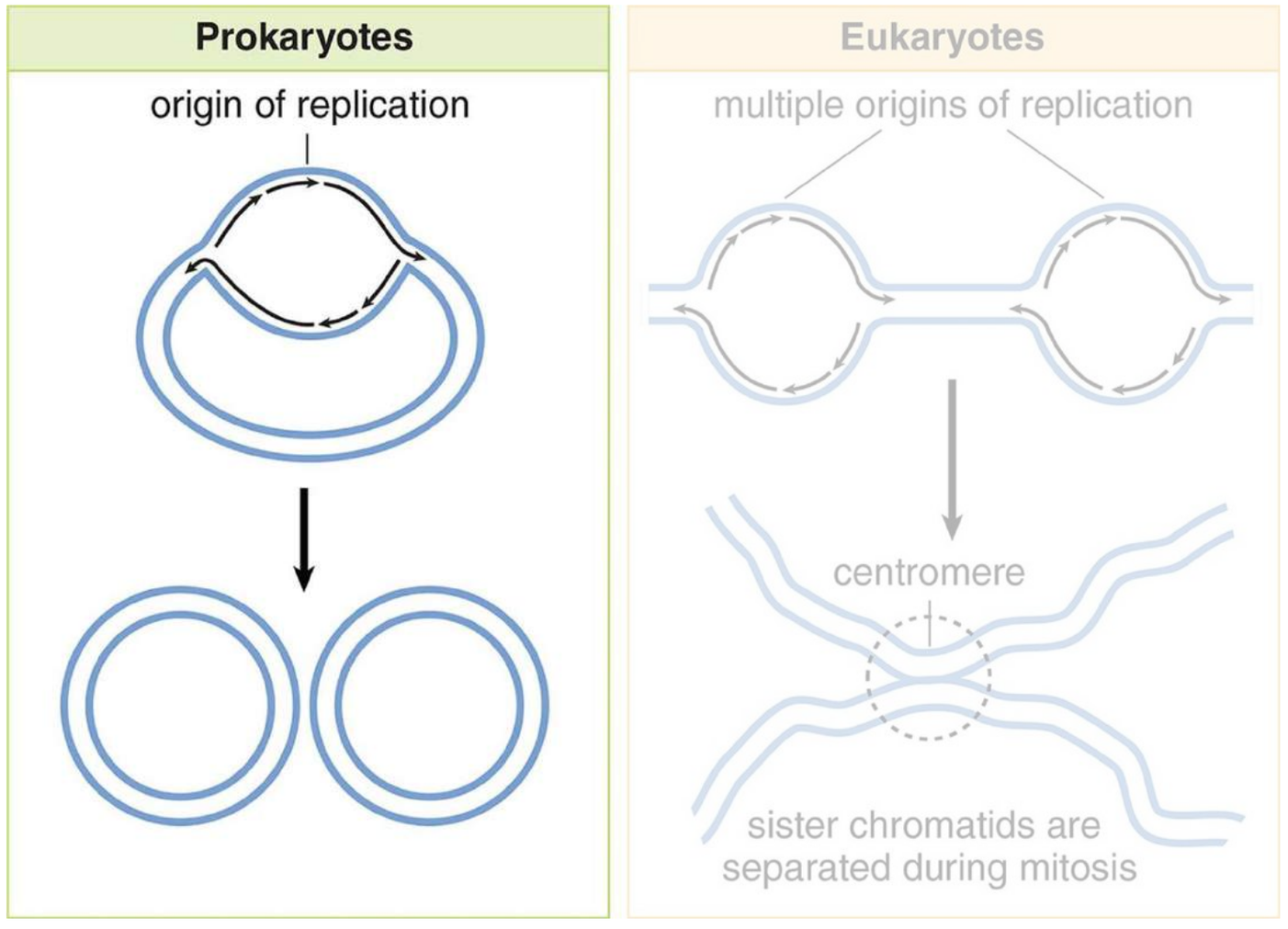


Prokaryotic DNA replication



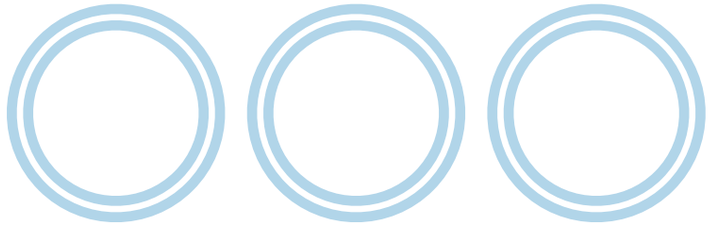
During exponential phase coverage increases near the origin of replication



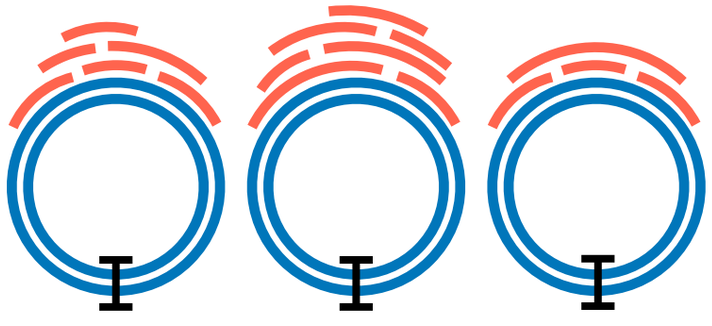


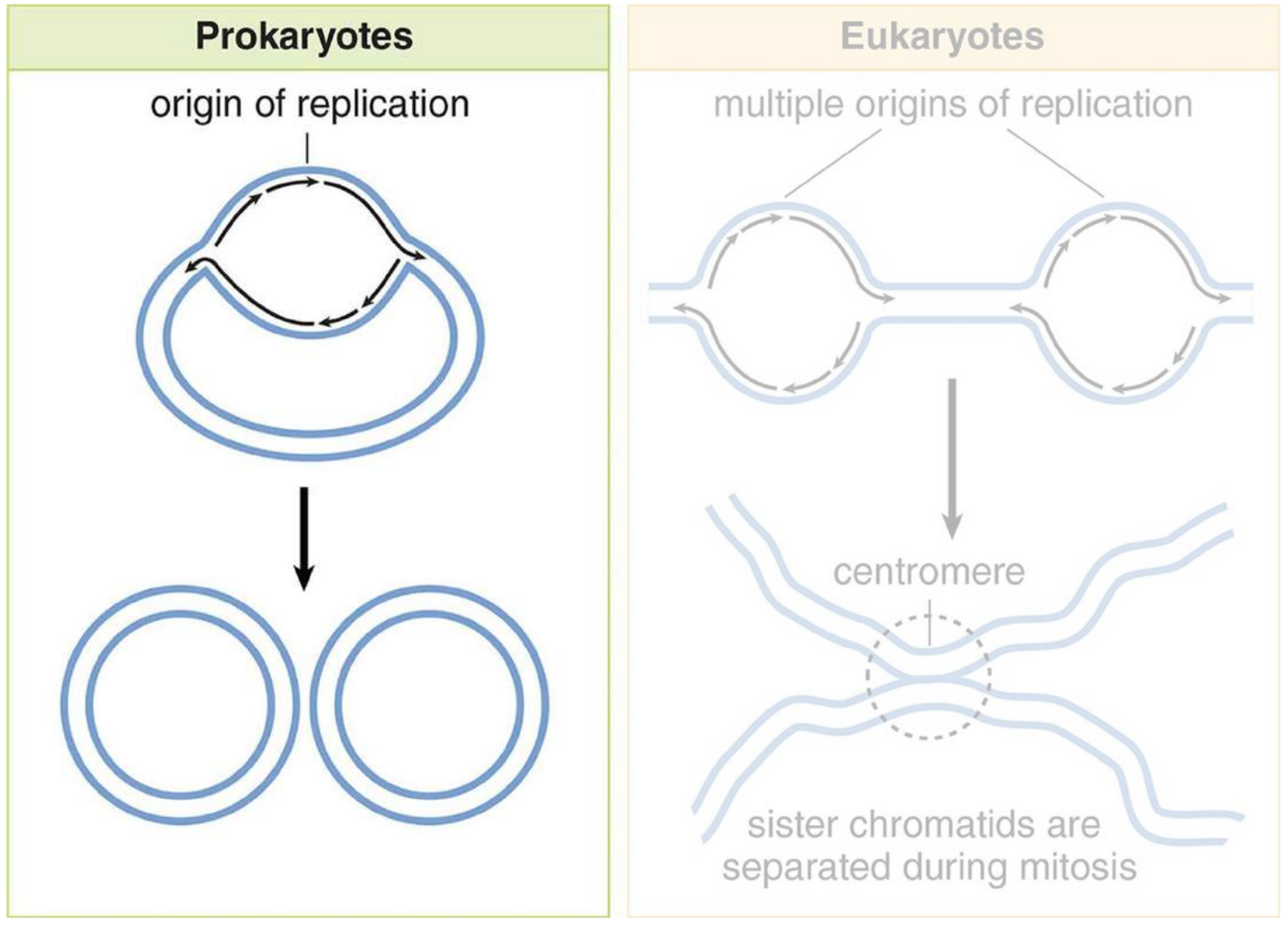
$$PTR = \frac{2^{ori}}{2^{ter}}$$ = coverage close to the origin of replication

Prokaryotic DNA replication



During exponential phase coverage increases near the origin of replication



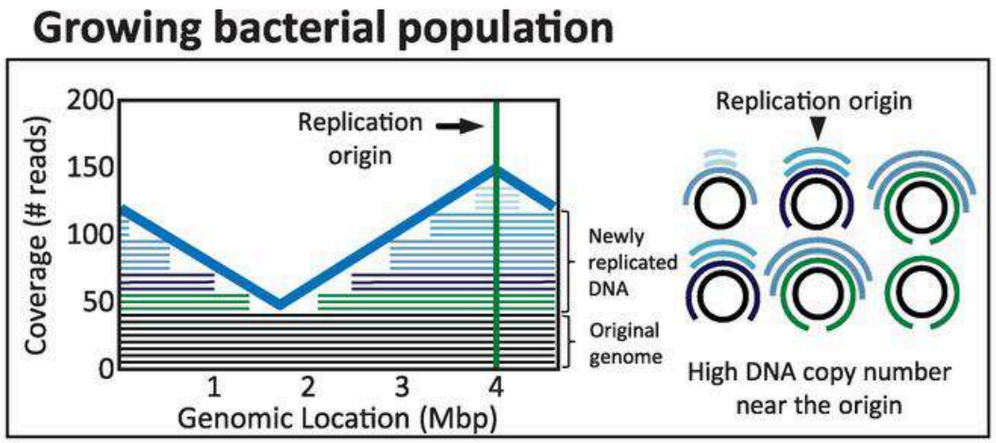


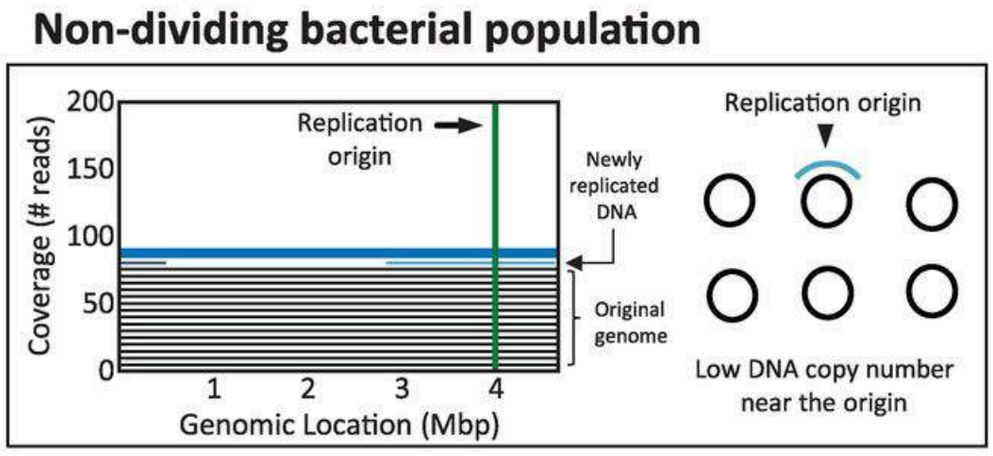
$$PTR = \frac{2^{ori}}{2^{ter}}$$

= coverage close to the origin of replication

= coverage close to the termination

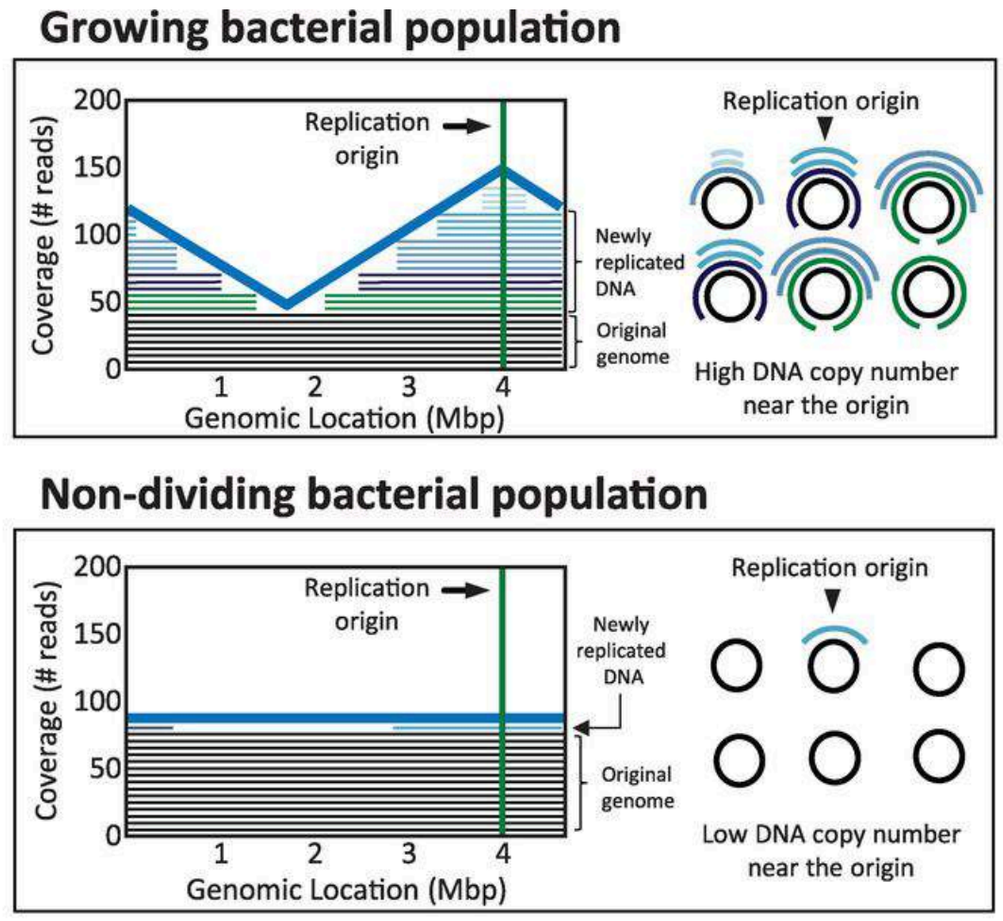
PTR and Growth rate





PTR and Growth rate

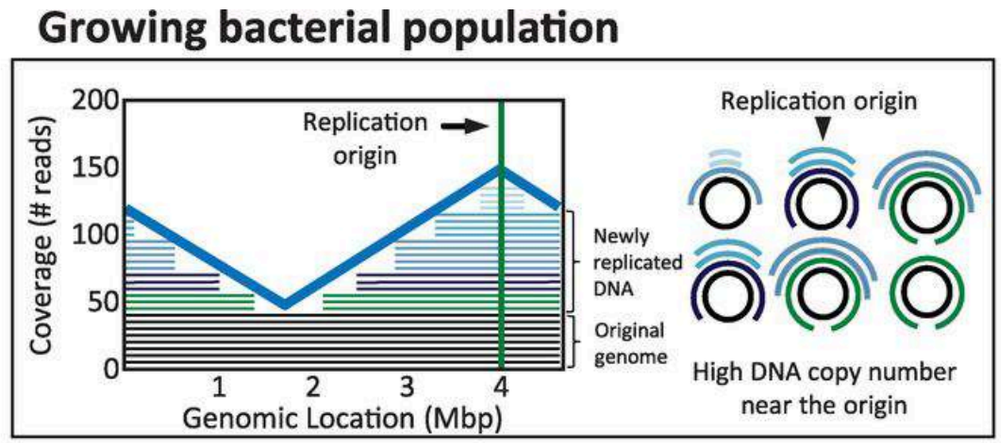
- PTR can be inferred from the same data used for community profiling



PTR and Growth rate

- PTR can be inferred from the same data used for community profiling

- Reference complete genome must be available

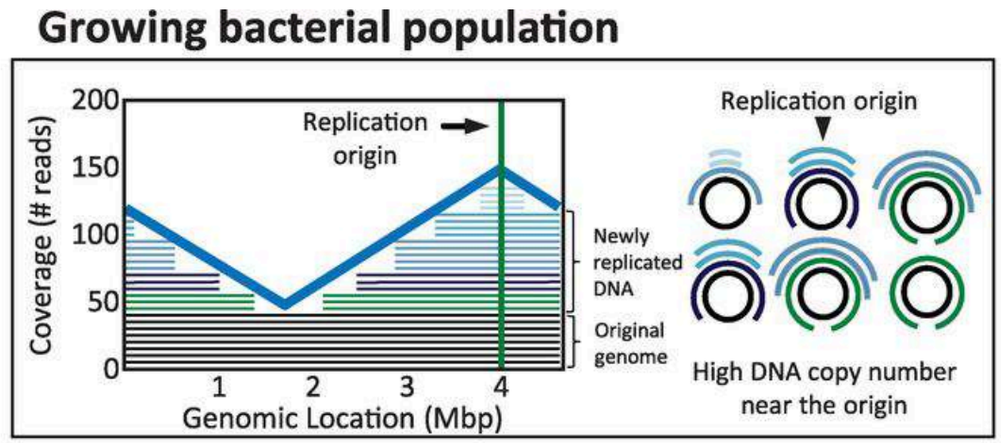


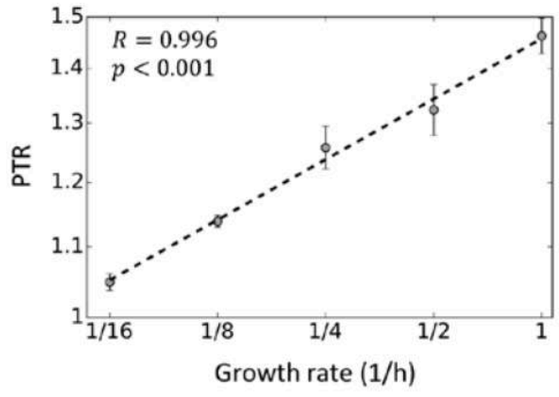
PTR and Growth rate

- PTR can be inferred from the same data used for community profiling

- Reference complete genome must be available

- PTR and Growth rate are highly correlated





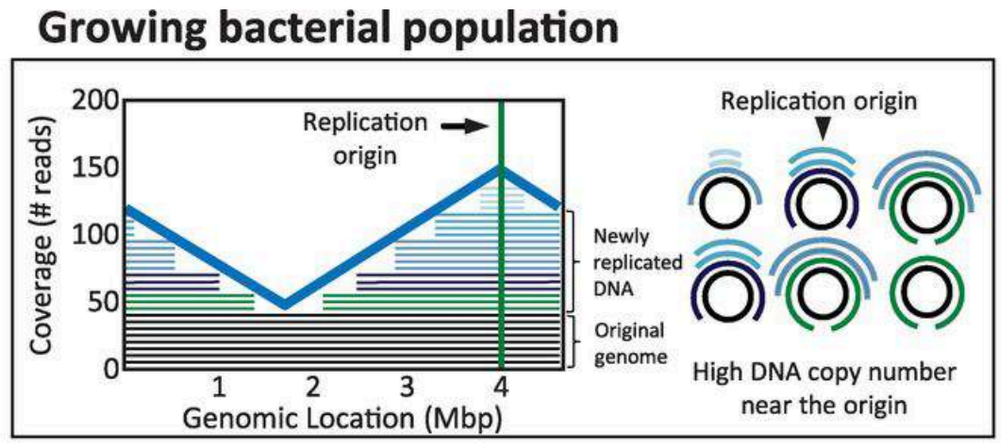
PTR and Growth rate

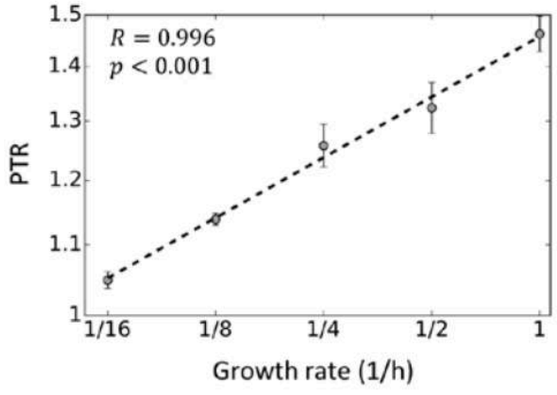
- PTR can be inferred from the same data used for community profiling

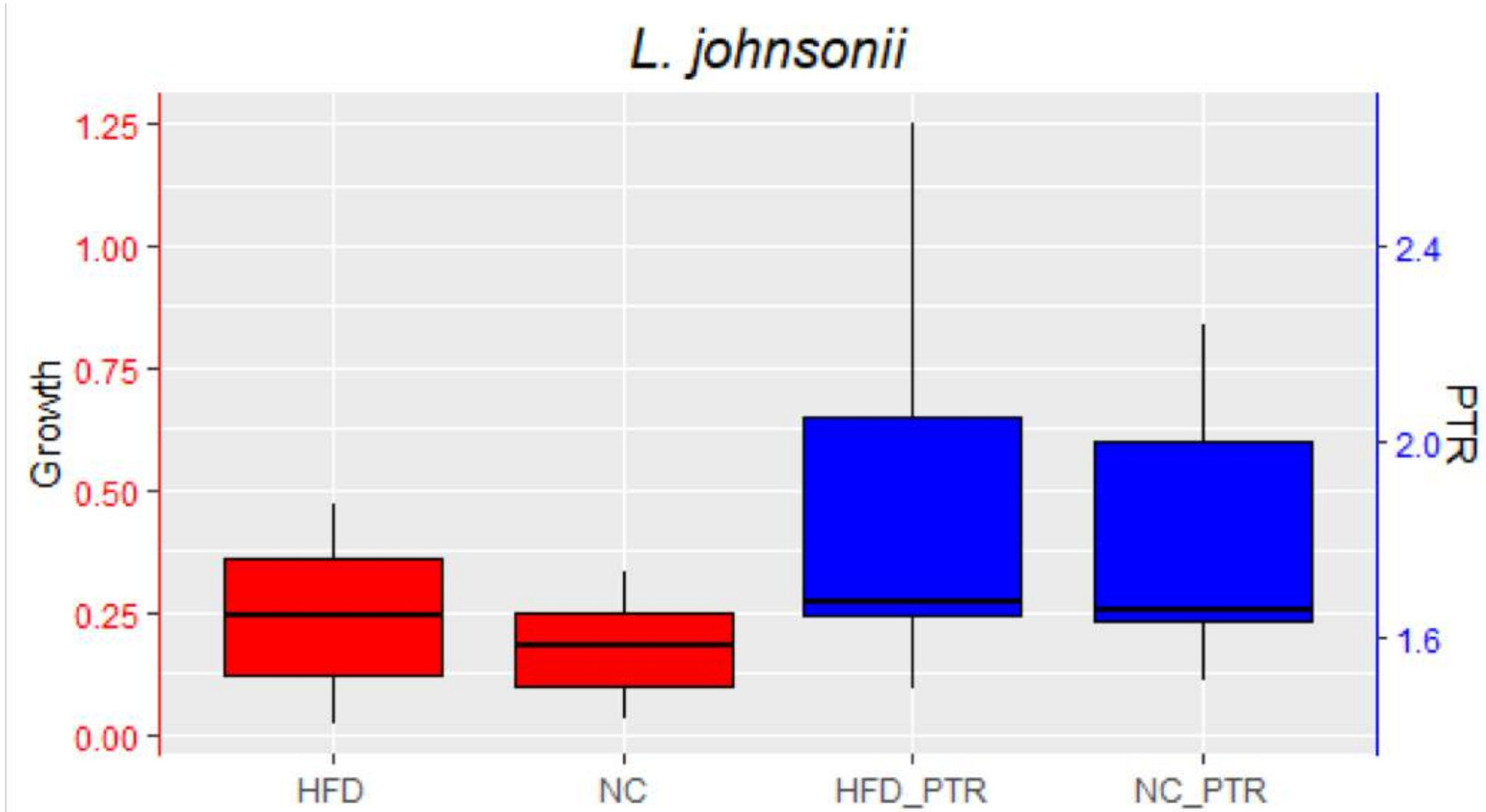
- Reference complete genome must be available

- PTR and Growth rate are highly correlated

- We need to find a relation between different growth rate indexes, PTR and coverage

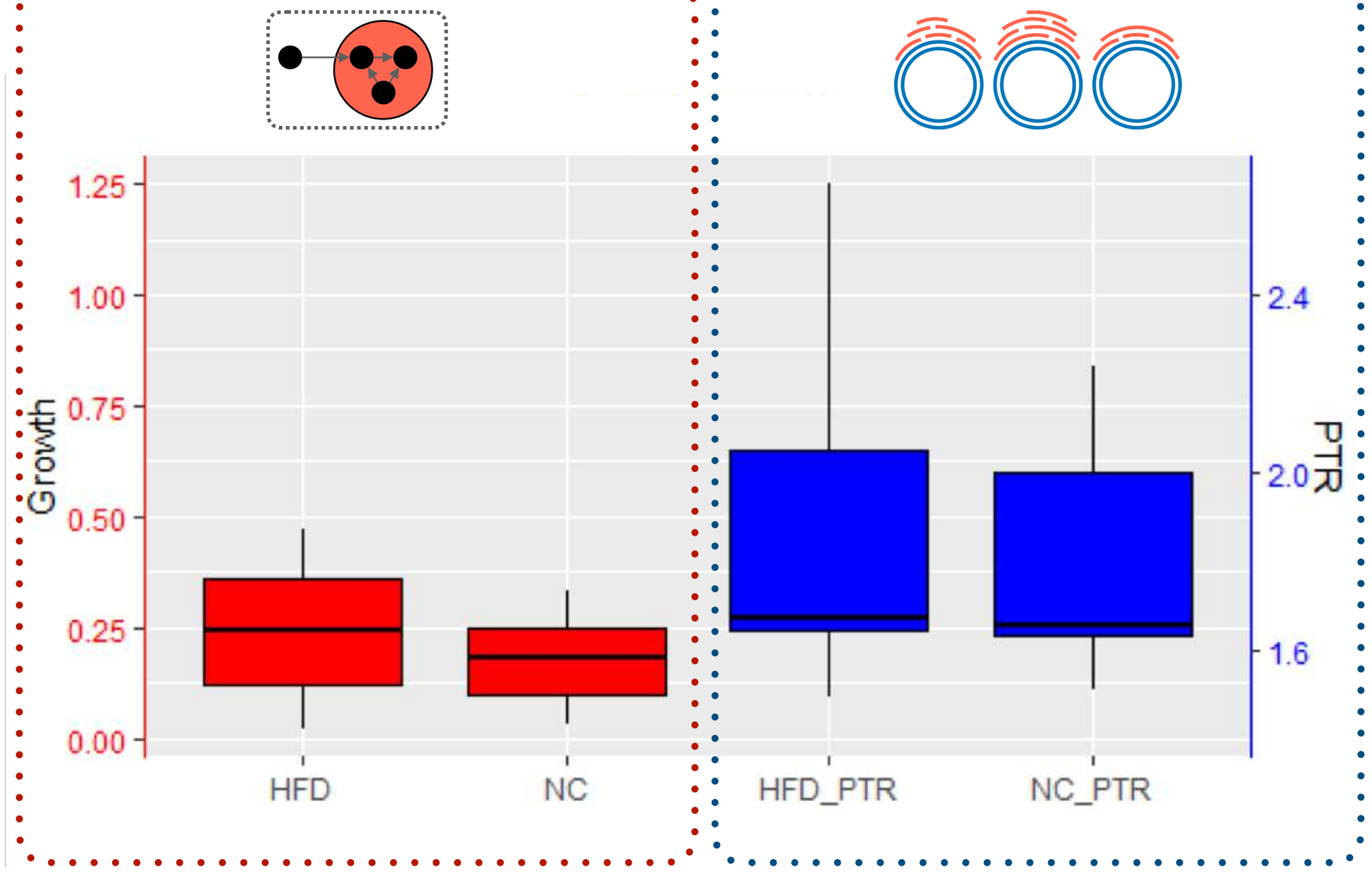






Other preliminary results

In two different conditions (NC and HFD) PTR and growth rate seem to be correlated



Other preliminary results

In two different conditions (NC and HFD) PTR and growth rate seem to be correlated

THANKS FOR THE ATTENTION!

THANKS FOR THE ATTENTION!

Questions, suggestions, and collaborations are welcome