

# Systems Analytic Strategies in Precision Medicine

**Kristel Van Steen, PhD<sup>2</sup> (\*)**

[kristel.vansteen@uliege.be](mailto:kristel.vansteen@uliege.be)

(\*) GIGA-R Medical Genomics, Systems Genetics Lab, University of Liège, Belgium

Systems Medicine Lab, KU Leuven, Belgium

## OUTLINE PART 1

- **Characteristics of Precision Medicine**
- **Characteristics of Systems Analytics**
  - **Integration**
  - **Interactions**
  - **Networks**



Archana Bhardwaj



Jestinah Mahachie-John



Elena Gusareva

## OUTLINE PART 2

- **Challenges and opportunities in systems**

**analytics for precision medicine**

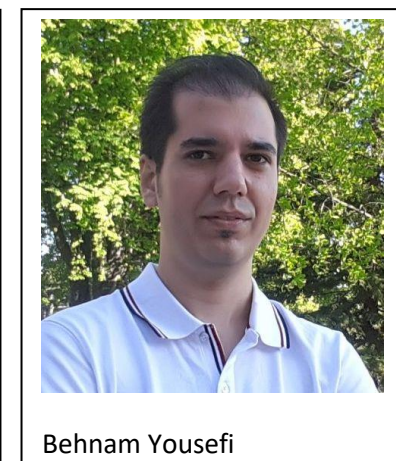
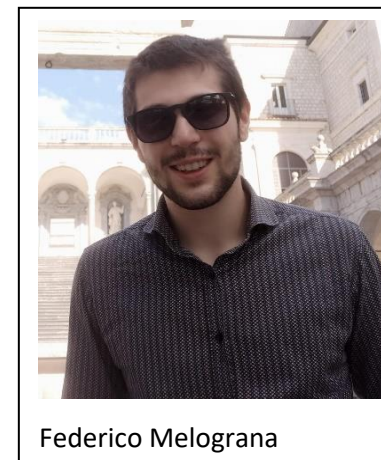
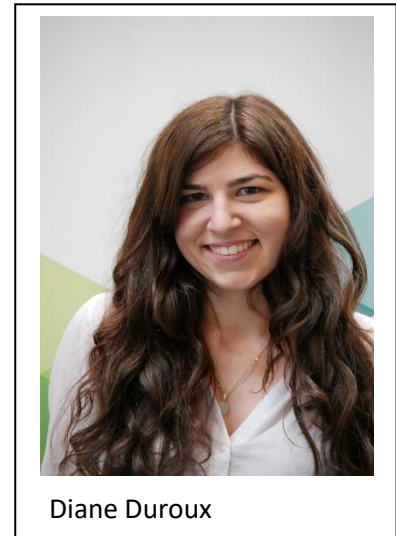
**by illustration:**

- **Population-based networks:**

**post-GWAS**

- **Individual-specific networks:**

**microbiome & transcriptome**



# **PART 1**

# **Characterisations**

# Characteristics of Precision Medicine

## Precision Medicine

“a medical model using characterization of individuals’ phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention.”

(HORIZON2020 Advisory Group; EU Health Ministers – December 2015)



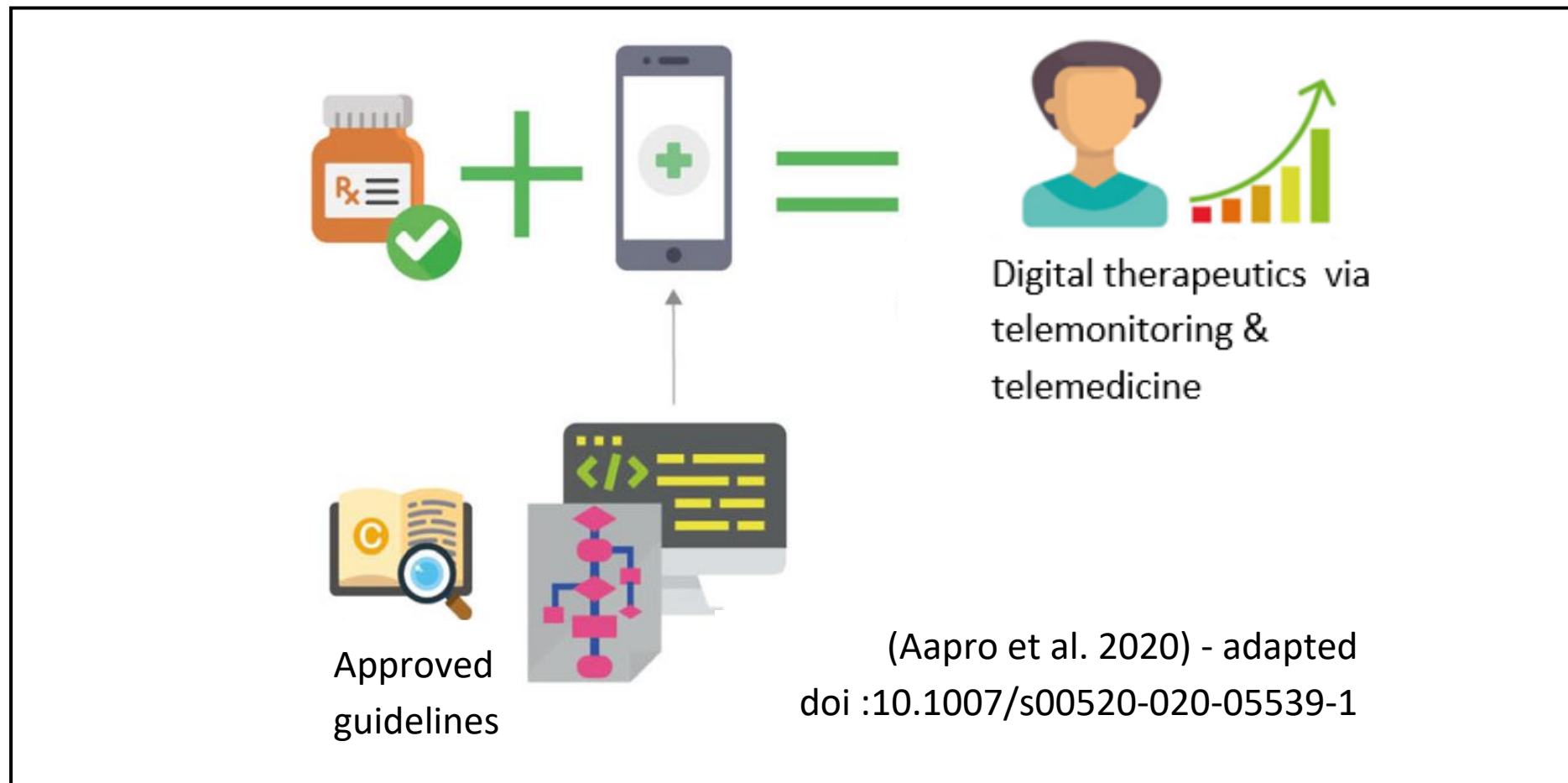
**Prevention**

**Diagnosis**

**Disease management**

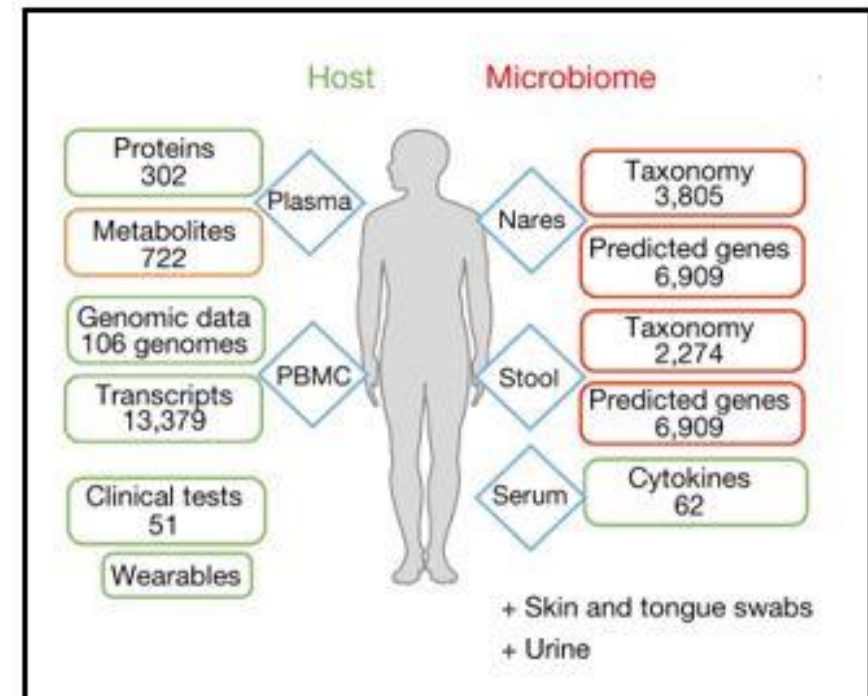
## Individual characterisation: comprehensive collection of information

- Digital health tools and solutions are redefining precision medicine and care



## Individual characterisation: Comprehensive collection of information

- What can integration bring us?
  - at a population level:
    - prediction of health outcomes (from samples to individual)
    - inter-personal variability (incl. identification of endotypes)



>> 1 individual

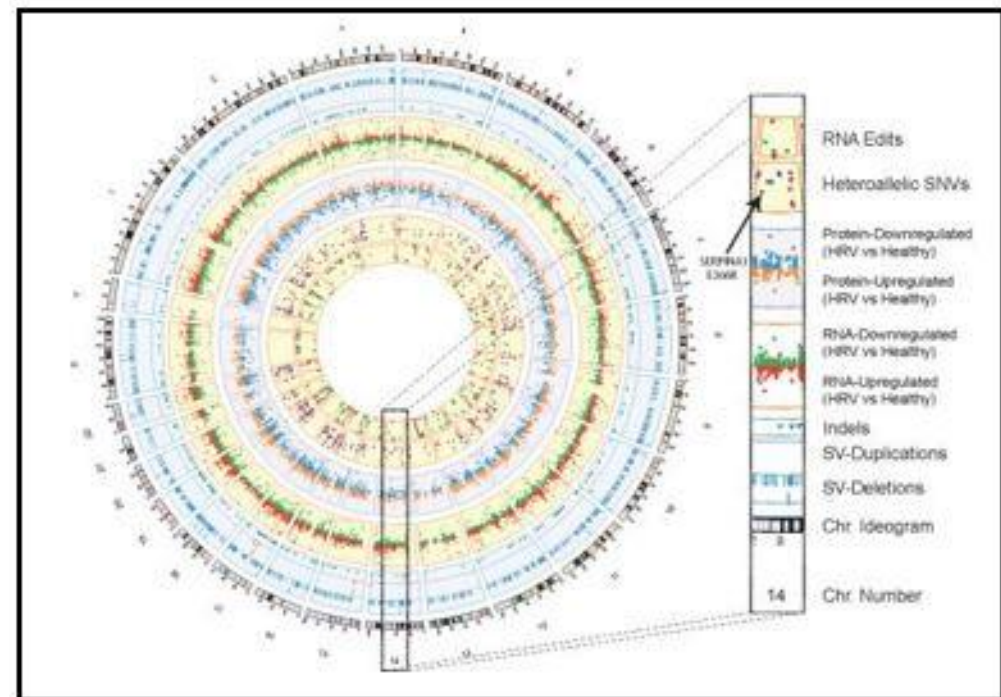
(Zhou et al. 2019)

doi : 10.1038/s41586-019-1236-x



## Individual characterisation: Comprehensive collection of information

- What can integration bring us?
  - at an individual level:
    - when measured over time, early detection of disease and forecasting (individual as internal control)
    - multi-view picture of the individual (informativity versus redundancy)



**1 individual**

(Chen et al. 2012)

doi: 10.1016/j.cell.2012.02.009

## *Interludium*

### *Endotypes and Phenotypes*

- Phenotype comes from the Greek words *φαίνω, φαίνô*, meaning “to show” and *τύπος, τυπος* meaning “type”: “observable properties of an organism that are produced by the interaction of the genotype and the environment.”
- Endotype is combination of the prefix *endo-*, from the Greek *ἔνδον, endon*, meaning “within,” and *τύπος, τυπος*. First occurred in 2008, in a review by Anderson on the pathogenetic mechanisms in asthma.

(Berdine et al. 2020)

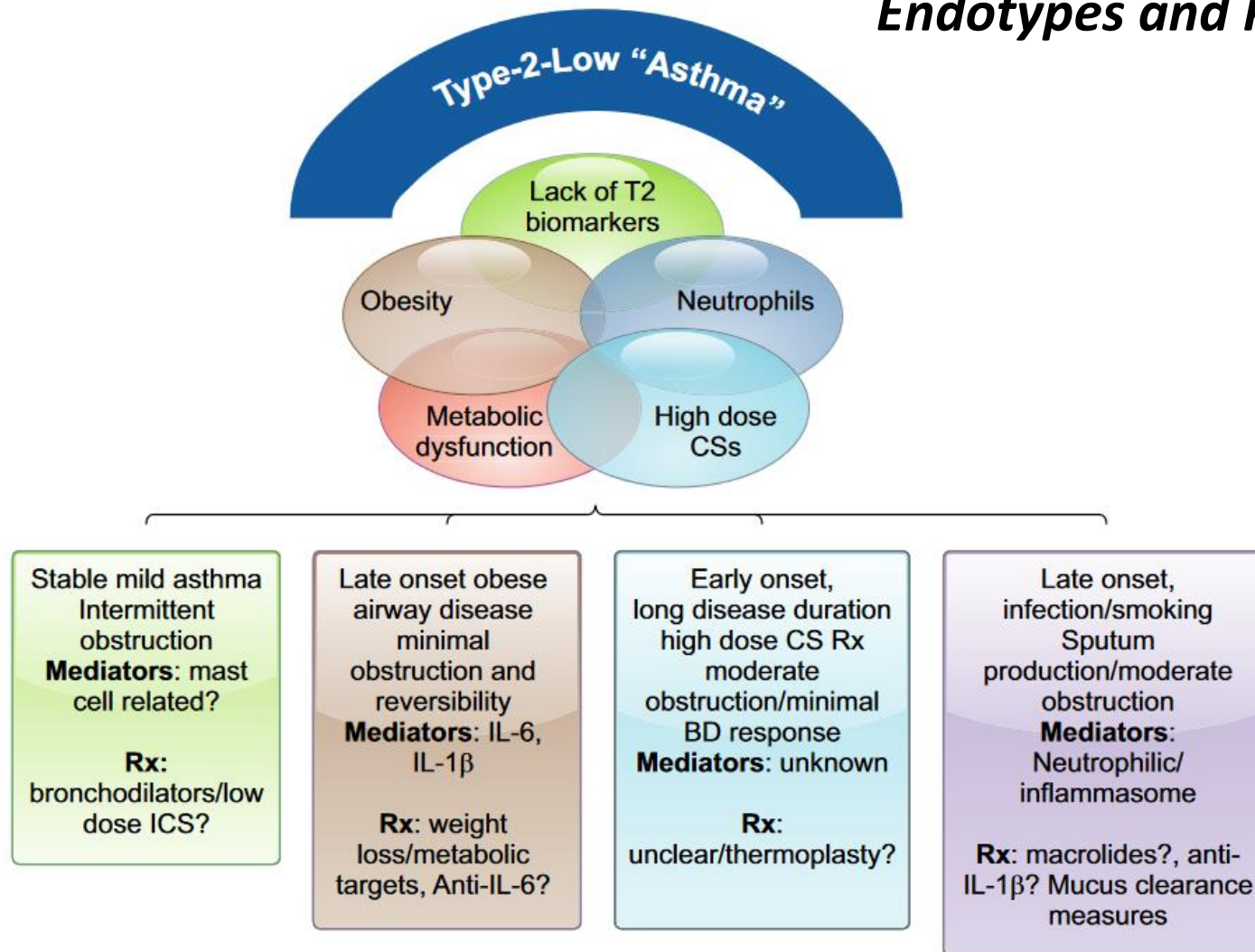
doi: 10.1080/08998280.2020.1793444

(Ray et al. 2020)

doi:10.1152/physrev.00023.20

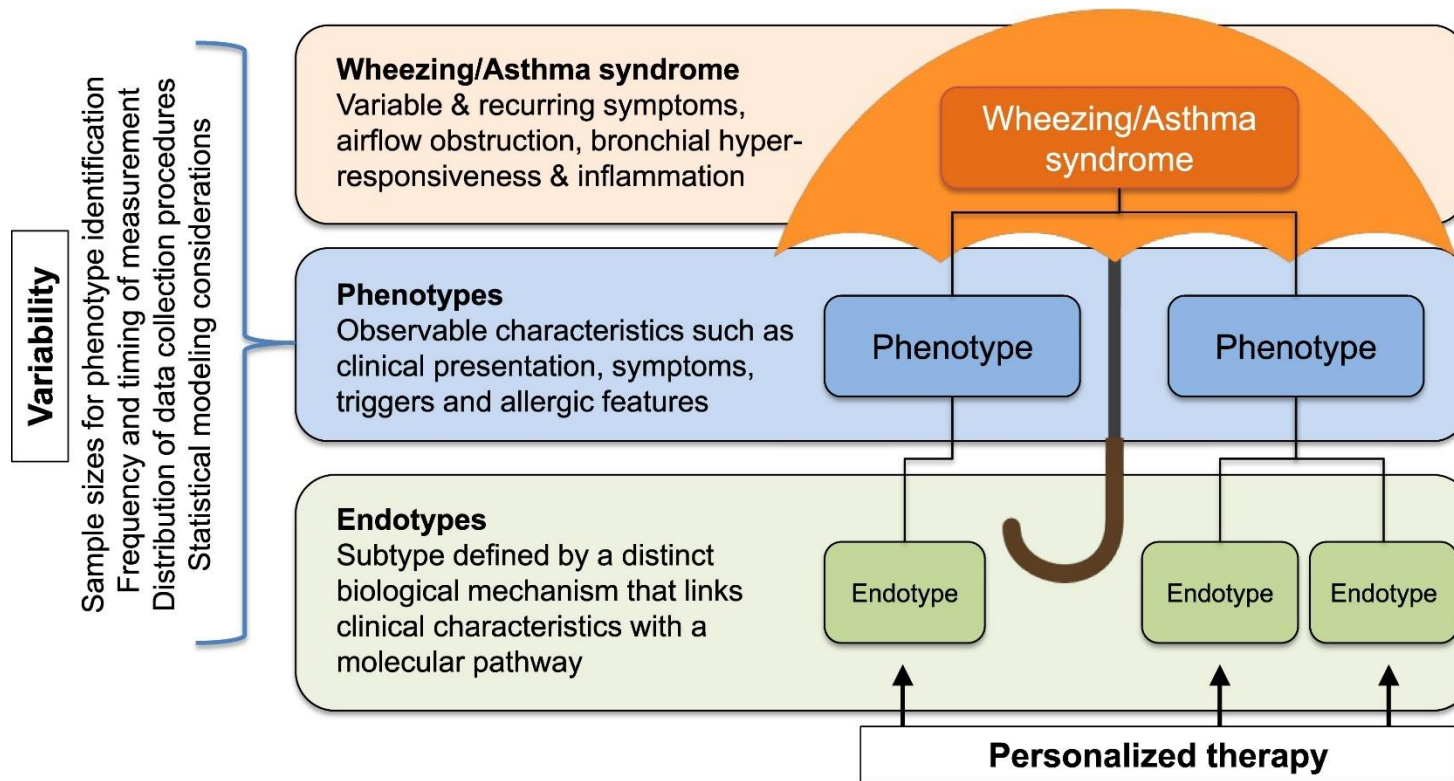
# Interludium

## Endotypes and Phenotypes



# Interludium

## Endotypes and Phenotypes



(Berdine et al. 2020)

doi: 10.1080/08998280.2020.1793444

## *Interludium*

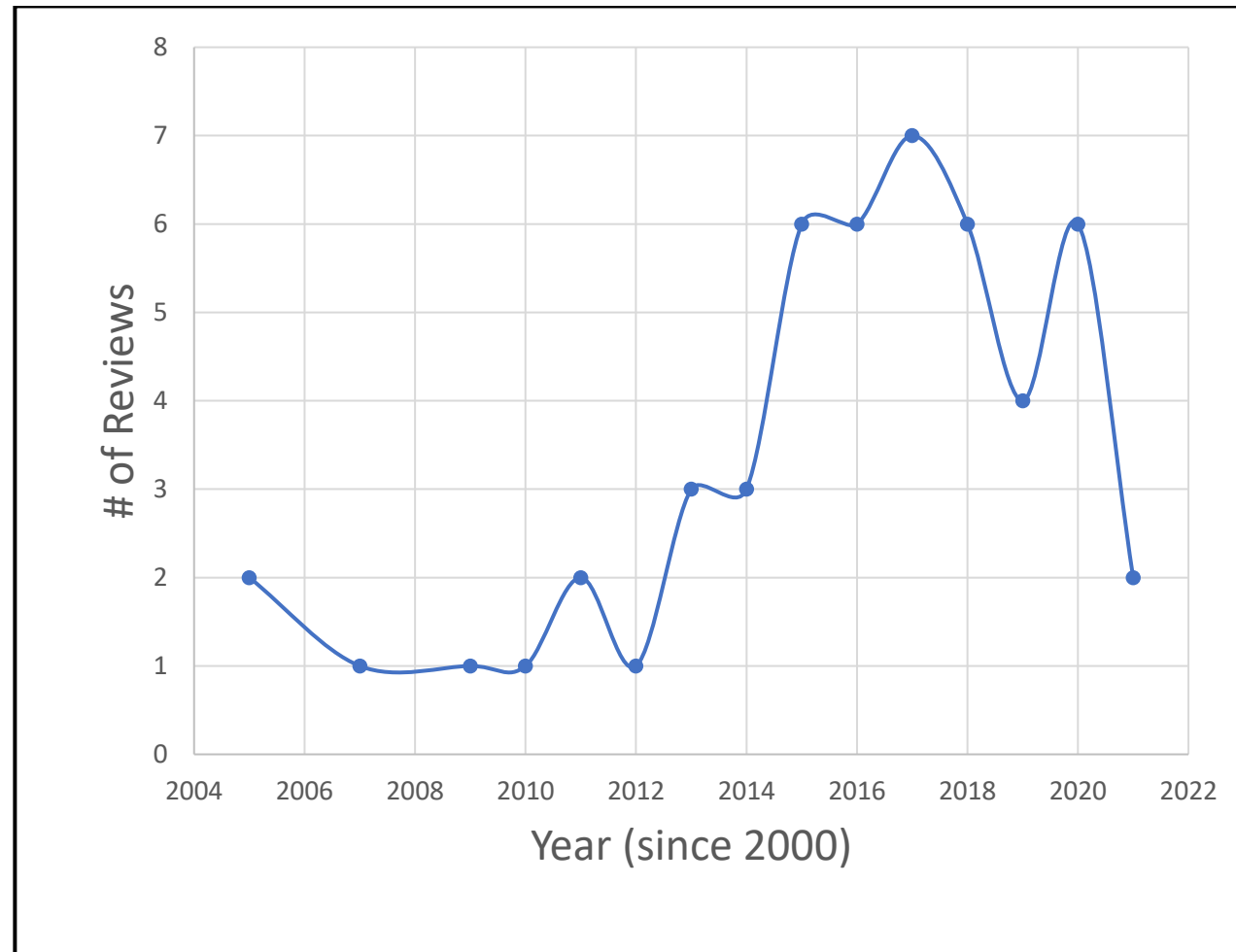
### *Endotypes and Phenotypes*

- The identification of endotypes for biomarker development is challenging
- Analytic methods need to embrace the possibility that different disease manifestations may involve partially overlapping or interconnected “systems”
- Obtaining an increased impact of endotypes on precision medicine will involve detailed investigations of the dynamics or stability of endotypes, and the relationship between disease endotypes and drug endotypes.

## *Interludium*

### *“Integration” in the literature*

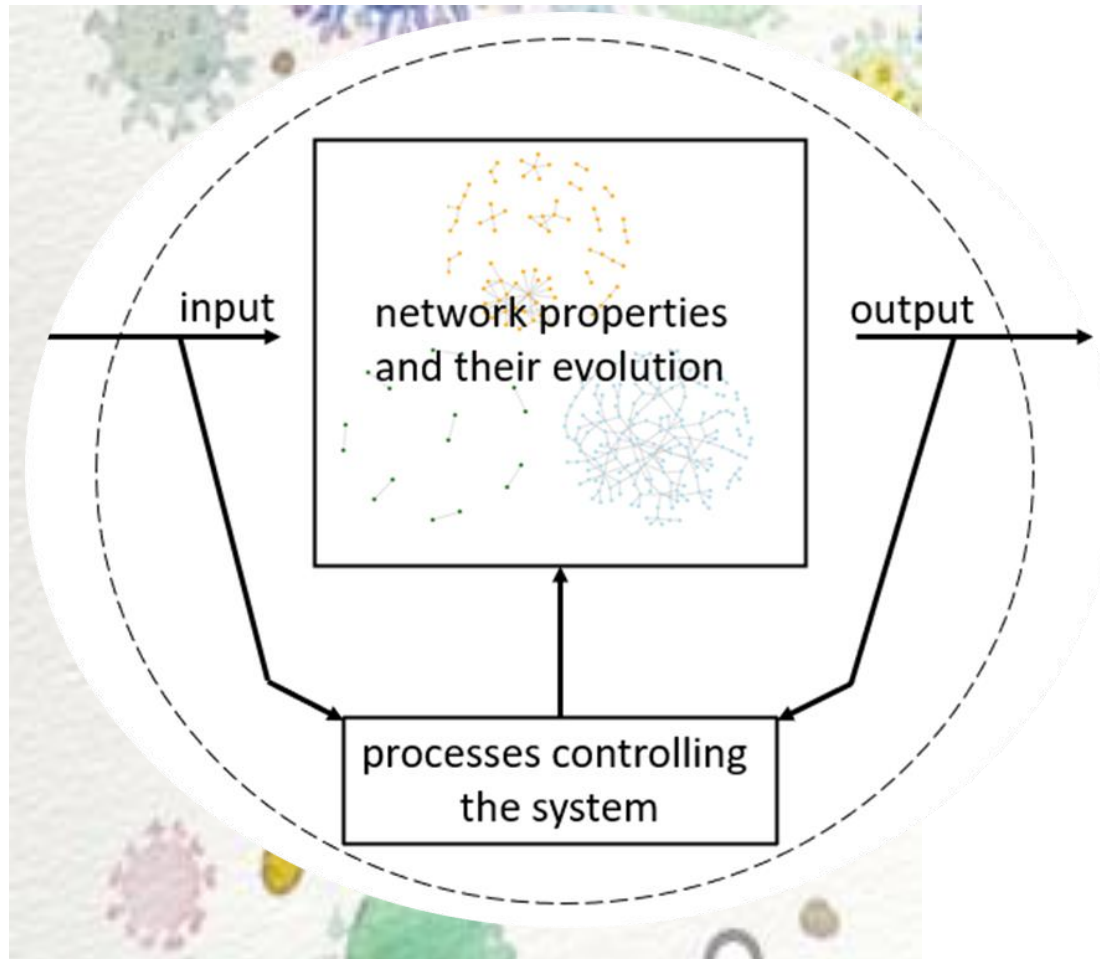
Search on PubMed 19  
September 2021:  
(integration[Title])  
AND  
(omics[Title/Abstract])  
Filters: Review,  
Humans, English, from  
2000/1/1 - 2021/9/11  
Sort by: Publication  
Date



## Individual characterisation: systems view

(Ackoff, 1971)

[doi.org/10.1287/mnsc.17.11.661](https://doi.org/10.1287/mnsc.17.11.661)

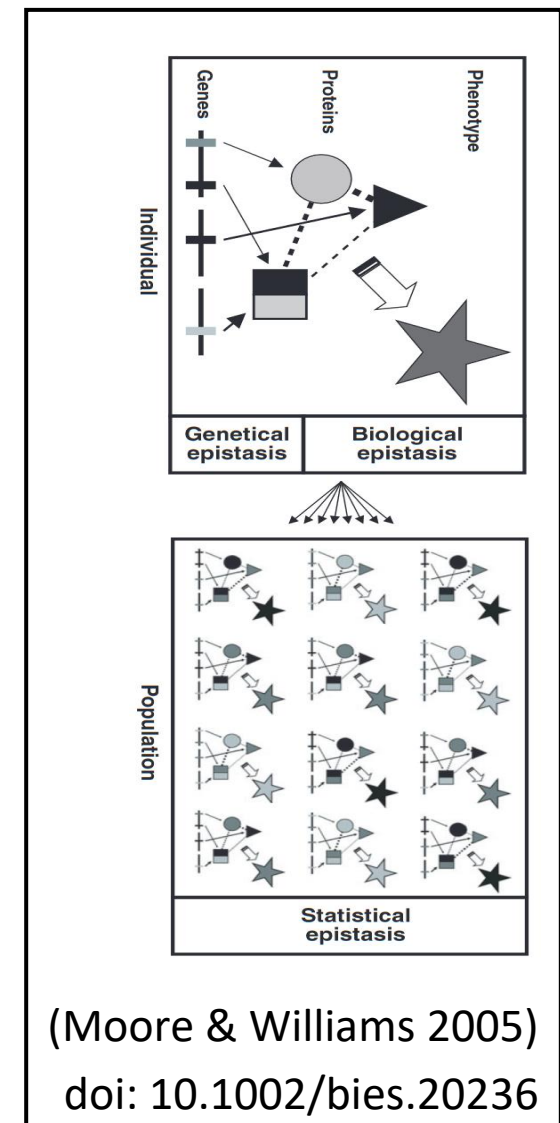


Elements of a system (a set of  $\geq 2$  interrelated items):

- Boundary
- Environment
- Observable interactions
- Subsystems
- Control mechanisms

## Systems view

- What can interactions bring us?
  - at a population level:
    - risk scores (e.g., MB-MDR)
    - understanding disease underlying complex mechanisms (e.g., mGWAS)
  - at an individual level:
    - individual-specific relevant biology
    - improved understanding about intra- and inter-individual heterogeneity

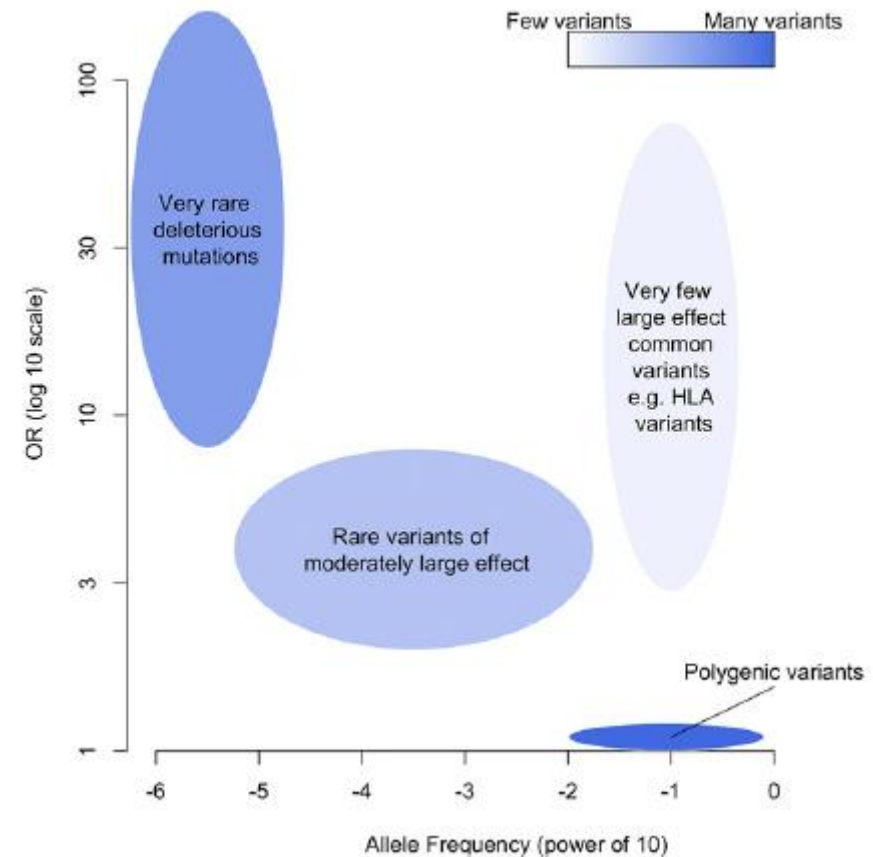
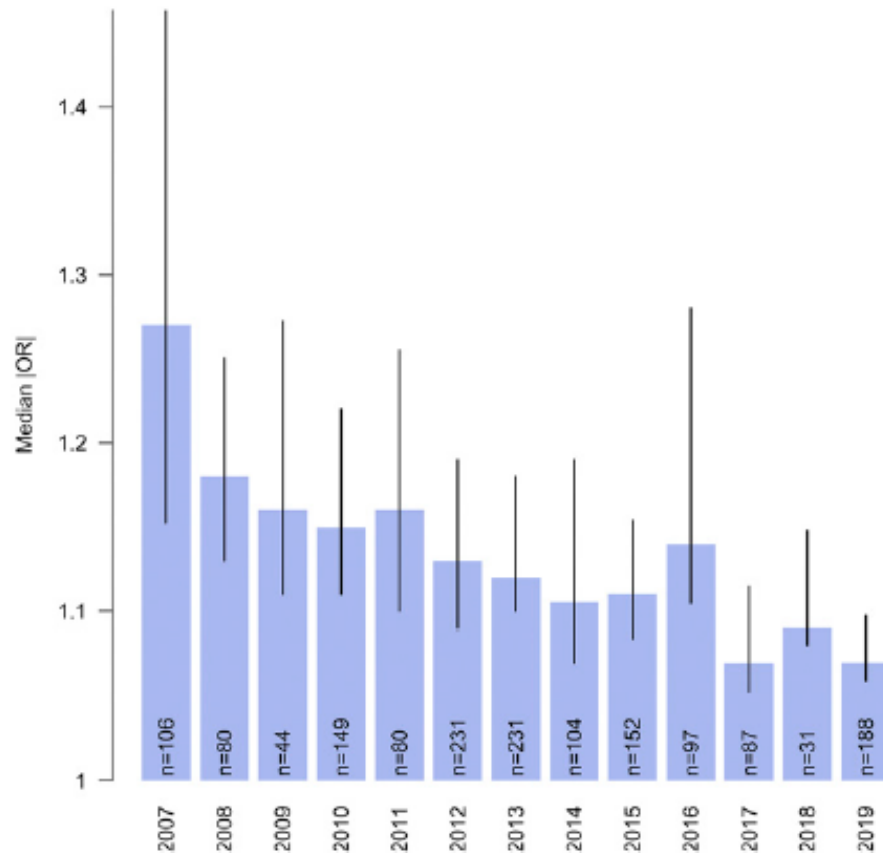




## Interludium

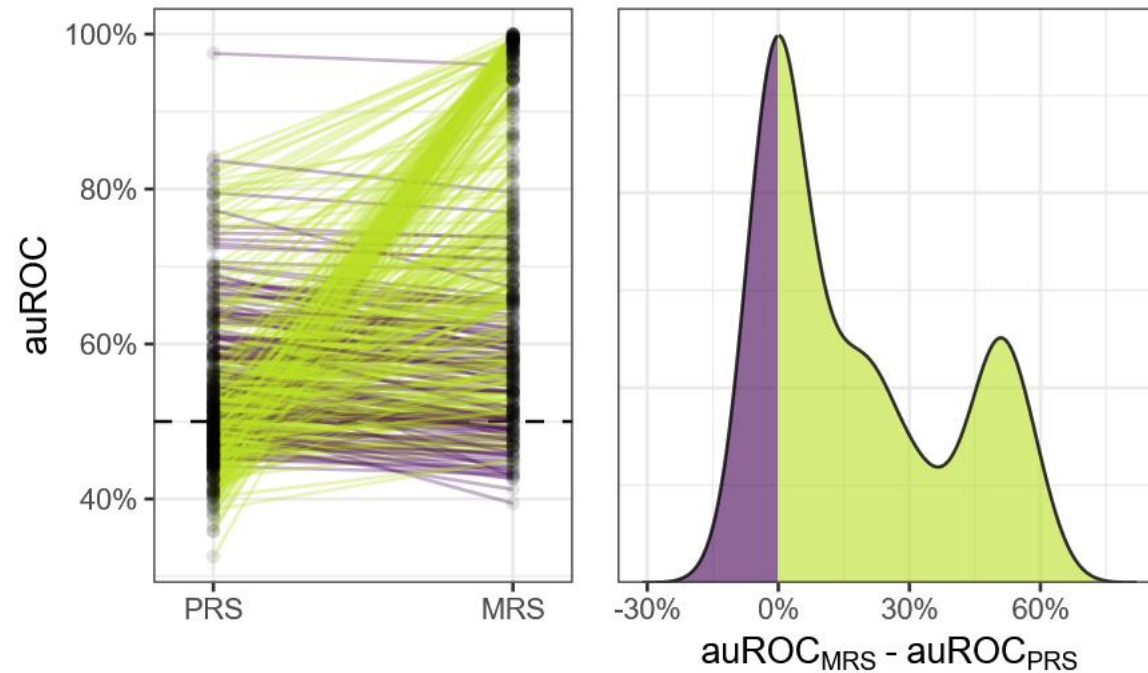
### Polygenic risk scores

$$PRS(i) = \sum_{j=1}^k \beta_j \times SNP_{ij}$$



(Crouch and Bodmer 2020)

doi: [10.1073/pnas.2005634117](https://doi.org/10.1073/pnas.2005634117)



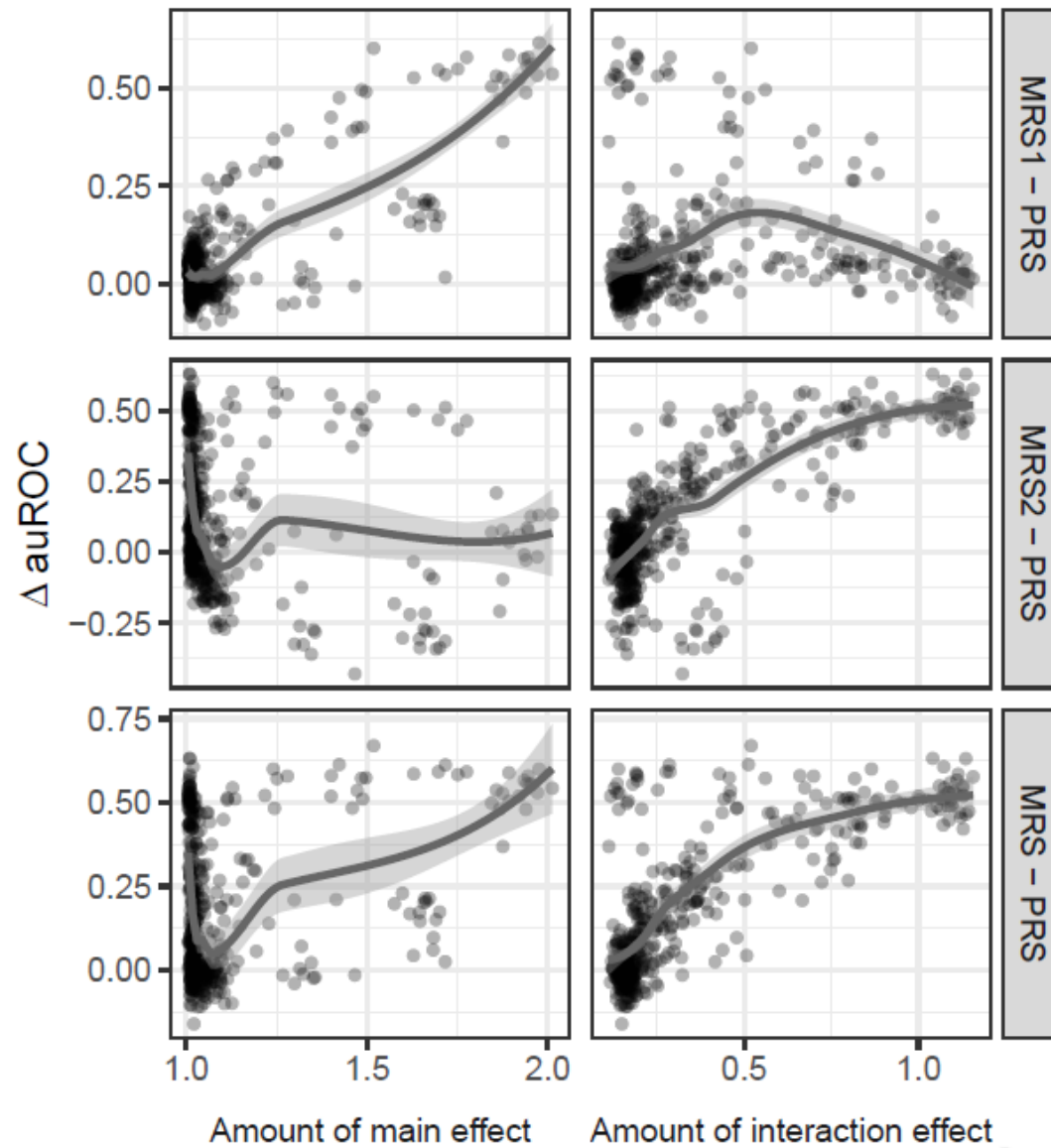
## *Interludium*

### *Polygenic risk scores*

(Le et al 2020)

doi : 10.5220/0008869700790084

“MRS produces improved auROC in the majority (335 green lines) of the 450 simulated datasets (each line represents a dataset). In many datasets, the standard PRS method performs poorly ( $auROC < 60\%$ ) while the new method yields auROC over 90%. This improvement in performance can be seen at the second peak ( $\sim 50\%$  auROC increase) in the density of the difference between the auROCs from the two methods (right)”



## Interludium

### Polygenic risk scores

(Le et al 2020)

doi : 10.5220/0008869700790084

### MB-MDR

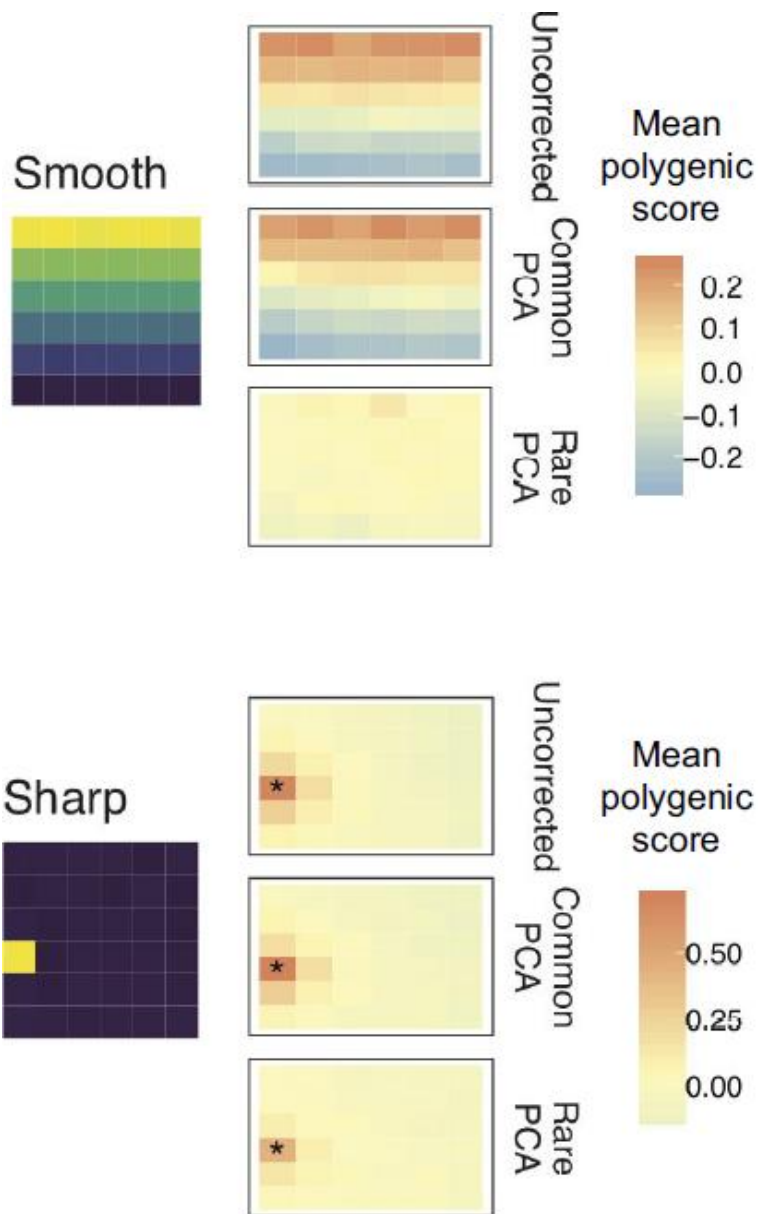
(Van Lishout et al. 2015)

	$SNP_1 = 0$	$SNP_1 = 1$	$SNP_1 = 2$
$SNP_2 = 0$	<i>O</i>	<i>O</i>	<i>O</i>
$SNP_2 = 1$	<i>O</i>	<i>H</i>	<i>L</i>
$SNP_2 = 2$	<i>O</i>	<i>L</i>	<i>H</i>

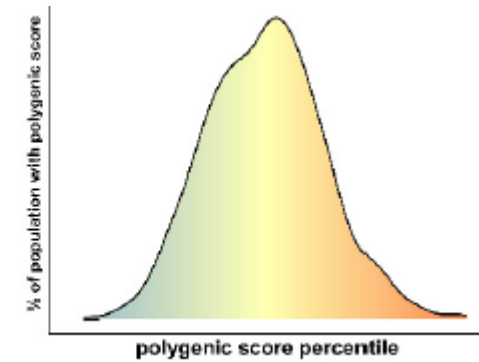
$$MRS_d(i) = \sum_{j=1}^{k_d} \gamma_j \times HLO_j(X_{ij})$$

# Interludium

## Polygenic risk scores

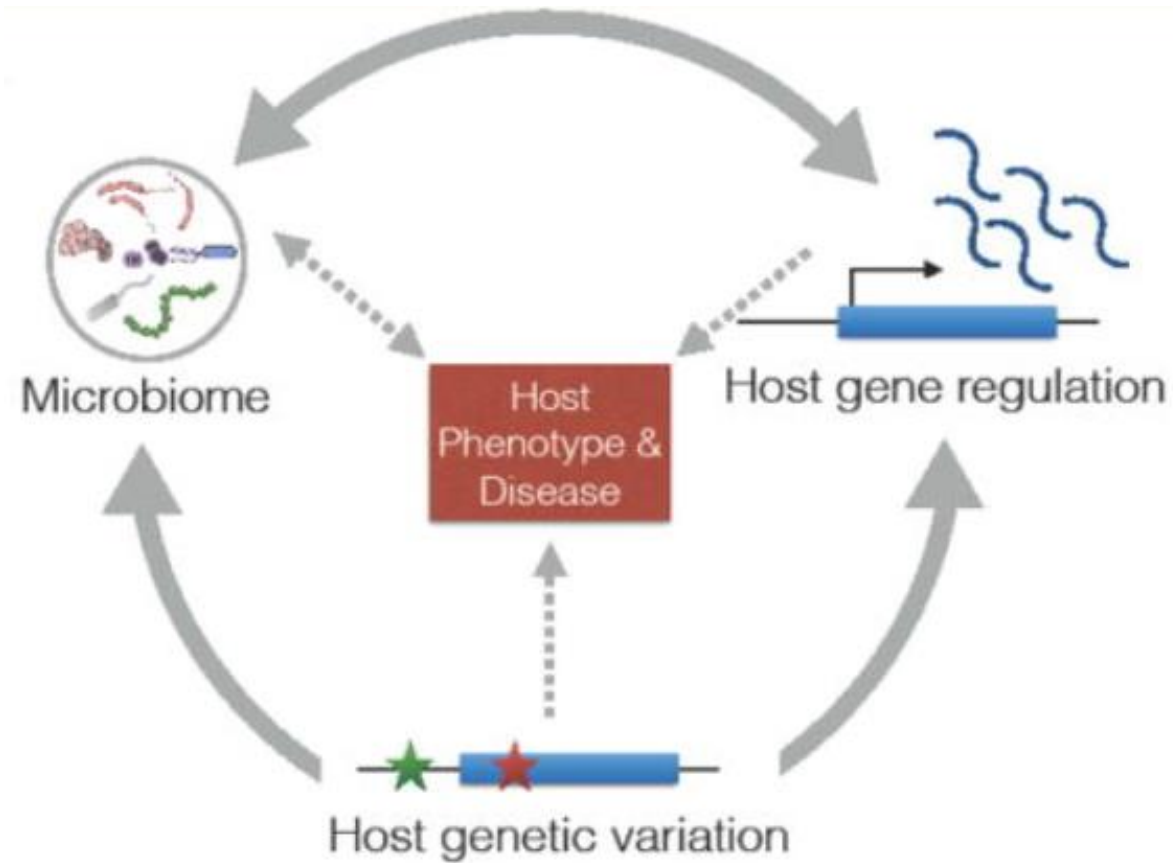


Measure genotypes in a new sample and weight by GWAS effect sizes to compute polygenic scores



(Blanc and Berg 2020)  
doi :10.7554/eLife.64948

# Interludium mGWAS



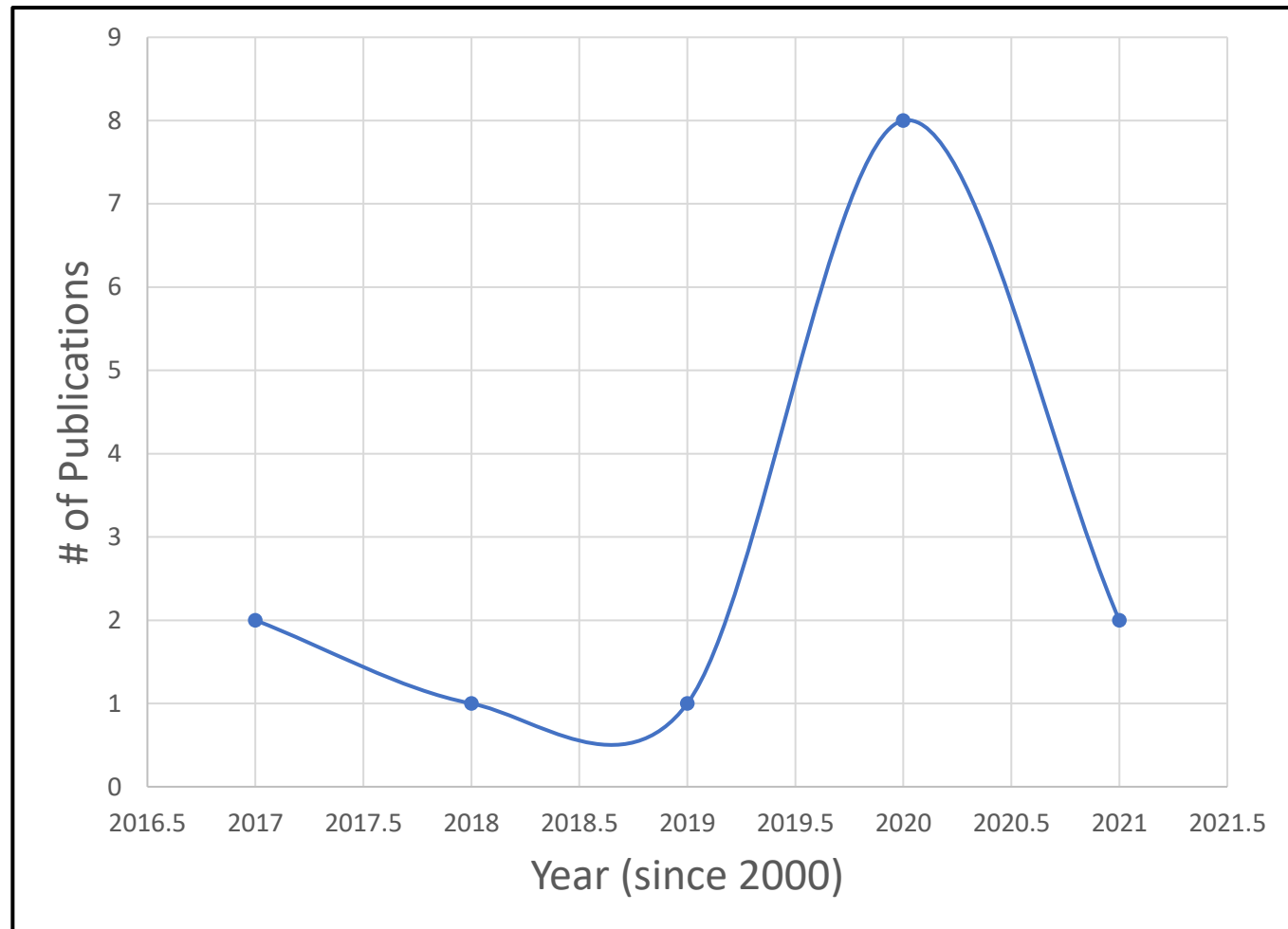
(Luca et al. 2018)

doi: 10.1016/j.tig.2017.10.001

## *Interludium*

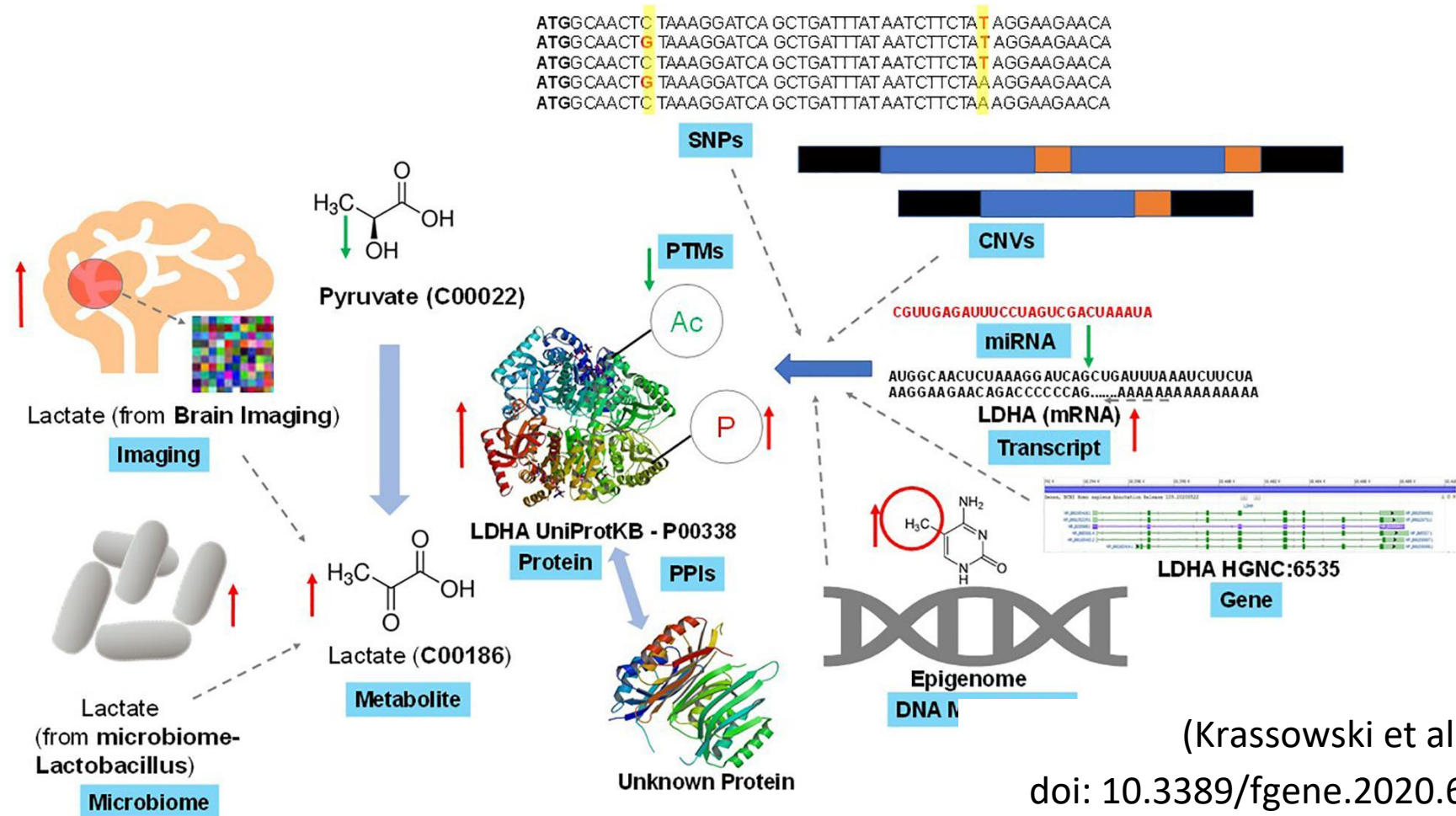
### *“Interaction” in the literature*

Search on PubMed  
11 September 2021:  
("multi-omic"[Title] AND  
"interaction"[Title/Abstract])  
AND ((humans[Filter] AND  
(2000/1/1:2021/9/11[pdat])  
AND (english[Filter]))) Filters:  
Humans, English, from  
2000/1/1 - 2021/9/11 Sort  
by: Publication Date



# Characteristics of Systems Analytics

# Integration & interactions





## Reviews on integration (multi-omics)

- Subramanian et al. (2020): overlay six groups of methods with targeted application contexts
- Labory et al. (2020): how methods “use” the data – feature selection, clustering, “fusion”
- Rappoport and Shamir (2018): methods classification into early, intermediate, late integration
- Huang et al. (2017): unsupervised, supervised and semi-supervised algorithms

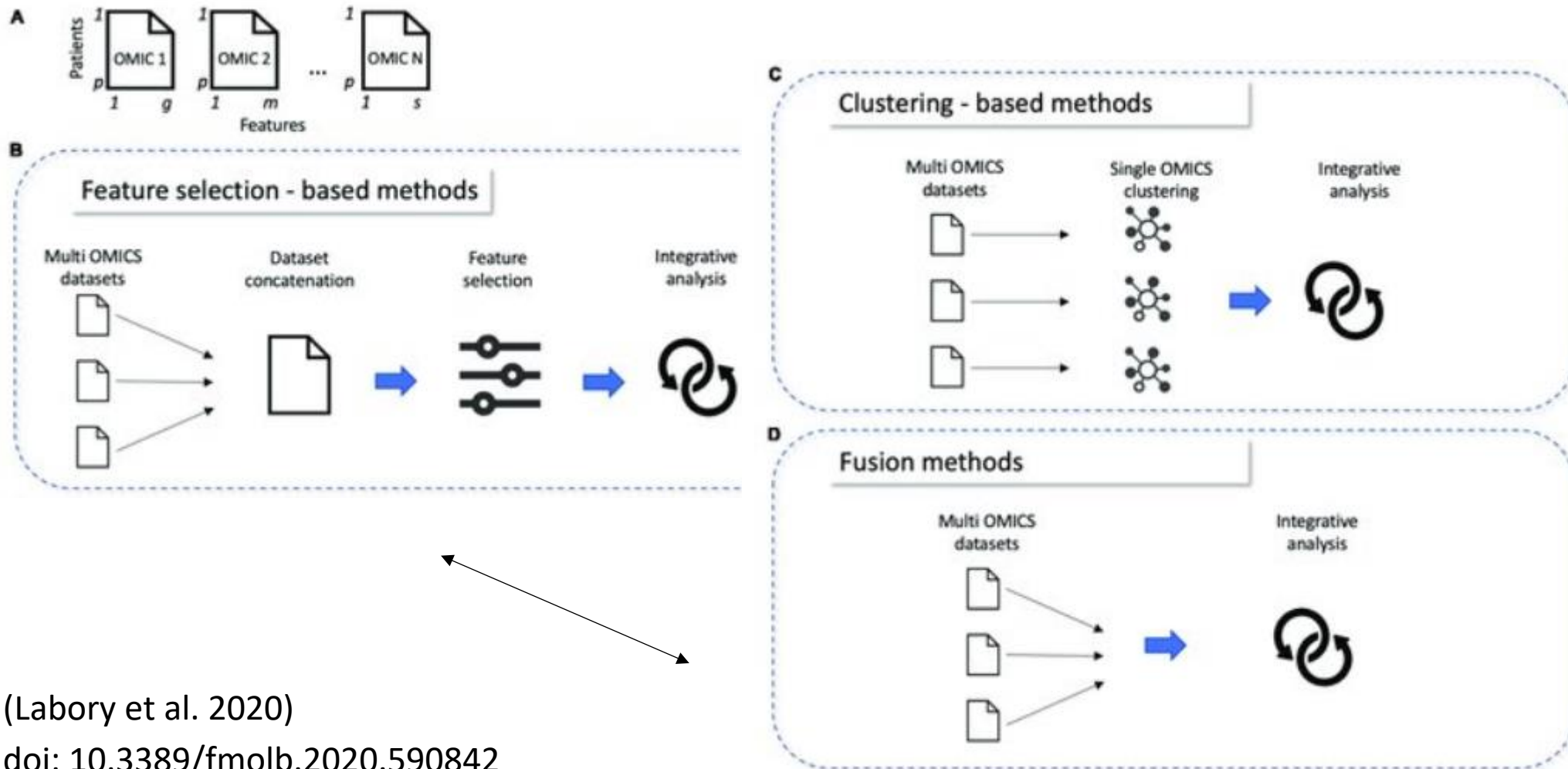
(Labory et al. 2020)

doi: 10.3389/fmolb.2020.590842

- Yu and Zeng (2018): bottom-up and top-down integration

# Interludium

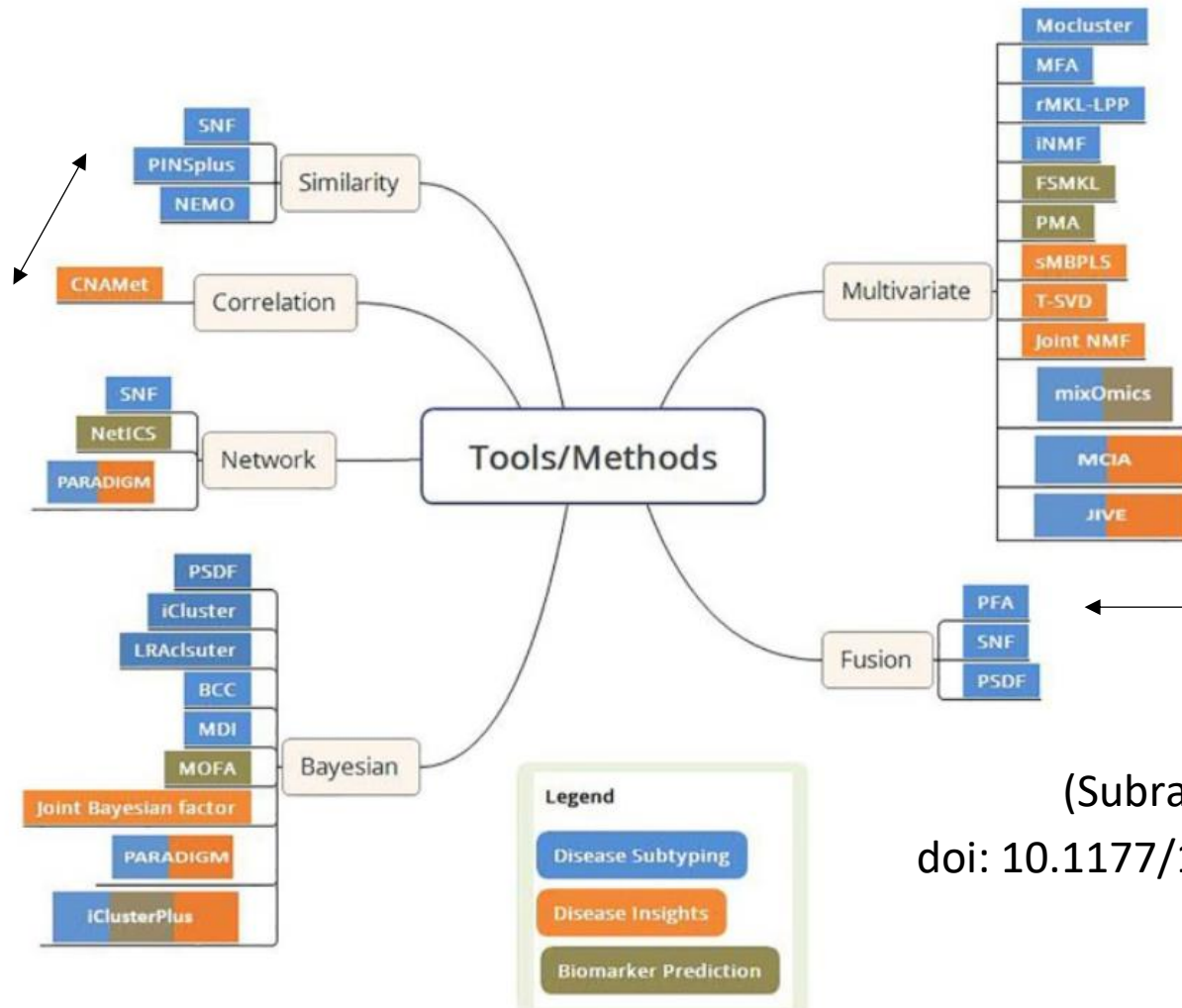
## Data use



(Labory et al. 2020)

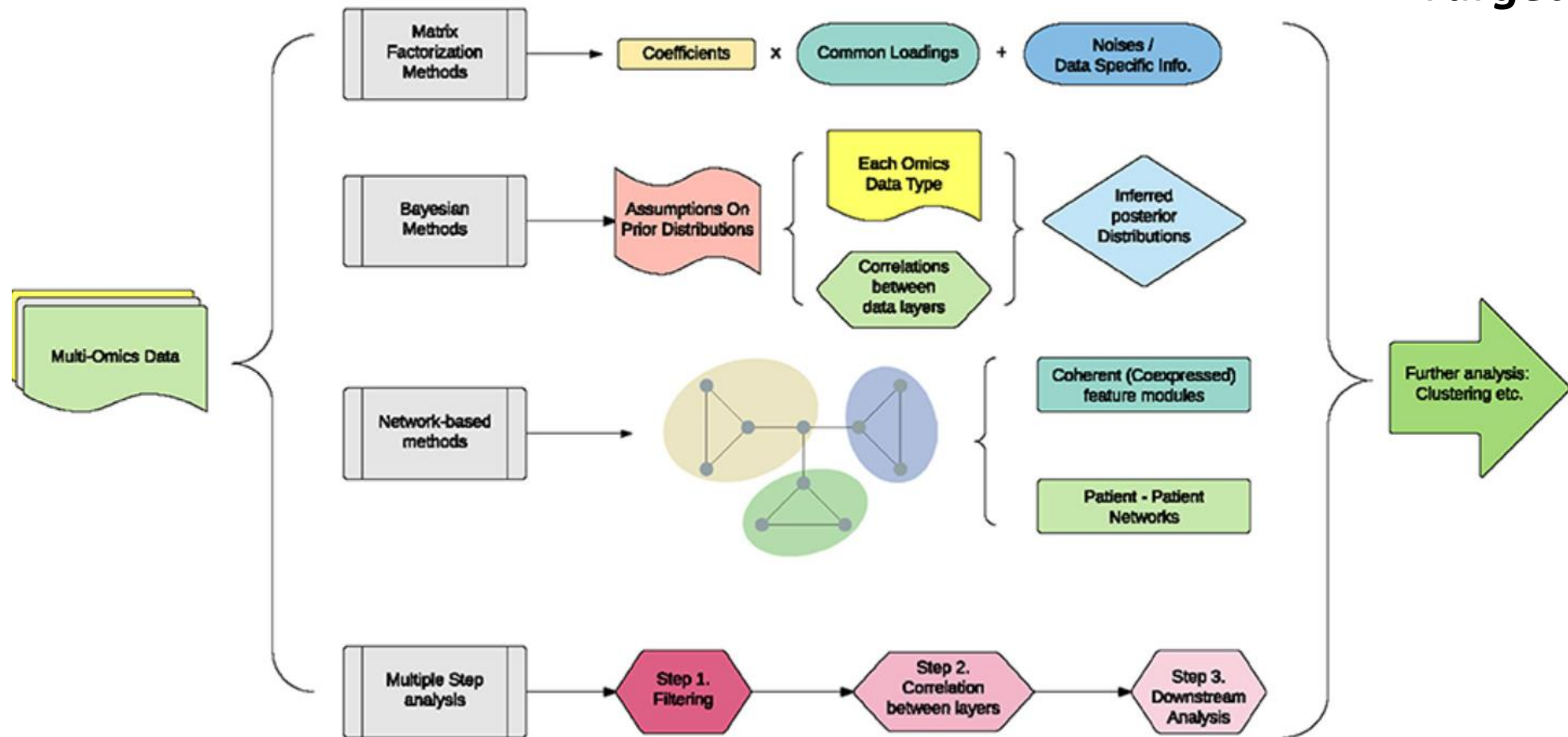
doi: 10.3389/fmolb.2020.590842

# Interludium Methods



(Subramanian et al. 2020)  
doi: 10.1177/1177932219899051

# Interludium Target



(Huang et al. 2017)

10.3389/fgene.2017.00084

## *Interludium*

### *Bull's eye*

- Non-omics



**Clinical trans-omics: an integration of clinical phenomes with molecular multiomics**

(Wang 2018)

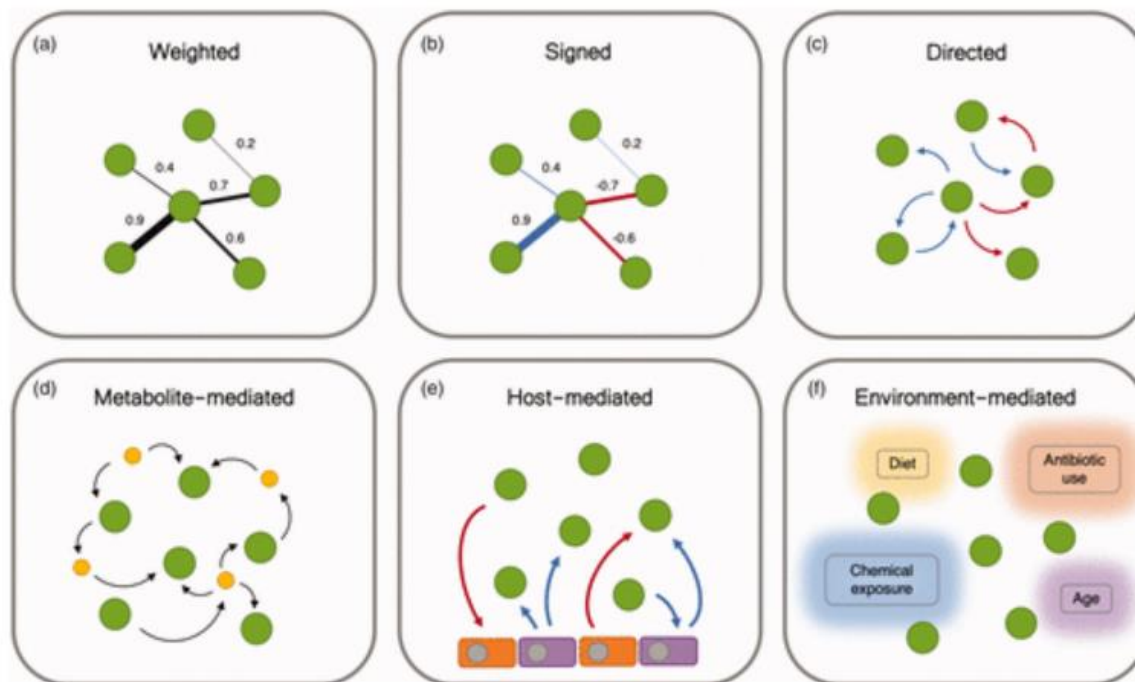
doi: [10.1007/s10565-018-9431-3](https://doi.org/10.1007/s10565-018-9431-3)

(Jansen et al. 2015) PMID: 26029010;  
PMCID: PMC4445433.

## *Interludium*

### *Bull's eye*

- Other organisms: microbiome



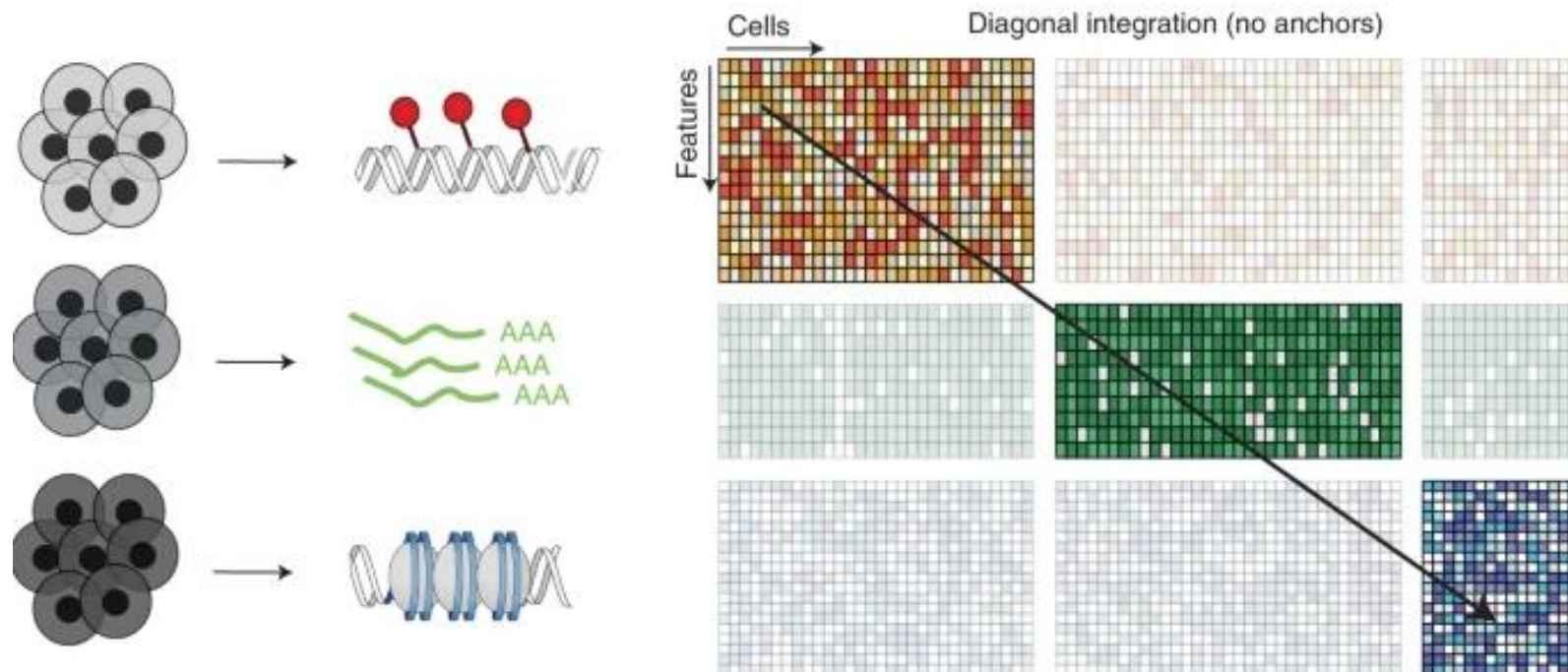
(Dohlman and Shen 2019)

doi: [10.1177/1535370219836771](https://doi.org/10.1177/1535370219836771)

## Interludium

### Bull's eye

- Other levels of granularity: single cells



(Argelaguet et al. 2021)

doi : 10.1038/s41587-021-00895-7

# *Interludium*

## *Tower of Babel*

Term	Definition
Multi-omics/panomics/ integromics/integrated omics polyomics/transomics cross-omics	An approach aiming to improve the understanding of systems regulatory biology, molecular central dogma and genotype-phenotype relationship by combining 3 or more different omics data.
Multi-table, Multi-block	Terms focusing on the format of the data rather than its nature, popular in chemoinformatics (among other fields); can (but does not have to) imply a larger number of features than observations in the integrated tables/blocks
Multi-view	Method often used in the field of ML for learning heterogeneity in the data and identification of patterns. By comparison to multiple cameras viewing an object from different angles, in omics context, the object can vary – whether it's "cell," "organism," or just "genome" viewed via different seq* techniques
Multi-source	This term encompasses datasets that are derived from multiple sources of molecular assays. This terminology is used, for example by the joint and individual variation explained (JIVE) tool (O'Connell and Lock, 2016) during EDA.
Multi-modal	A term often used in omics in reference to multiple measurements methods done at molecular level to gain holistic insights of cellular machinery (e.g., one cell at a time). It is also popular in drug repositioning that involves integration of more nuanced electronic health record (EHR) data integration

(Krassowski et al. 2020)

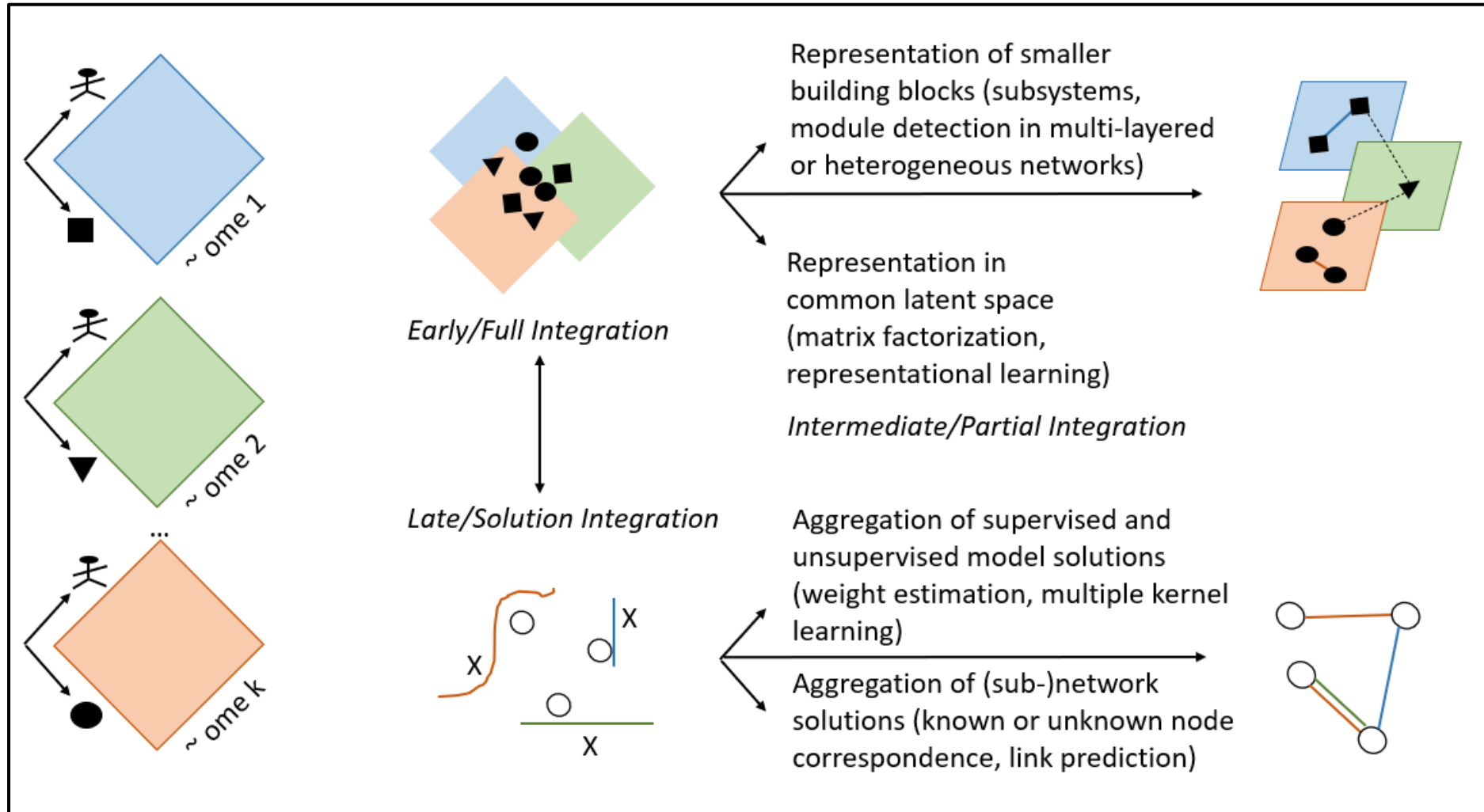
doi: 10.3389/fgene.2020.610798



## Reviews on integration (multi-omics)

- Subramanian et al. (2020): overlay six groups of methods with targeted application contexts
- Labory et al. (2020): how methods “use” the data – feature selection, clustering, “fusion”
- Rappoport and Shamir (2018): methods classification into early, intermediate, late integration
- Huang et al. (2017): unsupervised, supervised and semi-supervised algorithms  
(Labory et al. 2020)  
doi: 10.3389/fmolb.2020.590842
- Yu and Zeng (2018): bottom-up and top-down integration
- Argelaguet et al. (2021): single cells integromics, anchor-driven

# Integration & interactions



Inspired by Ritchie et al. 2015 → Rappoport and Shamir (2018)

## *Interludium*

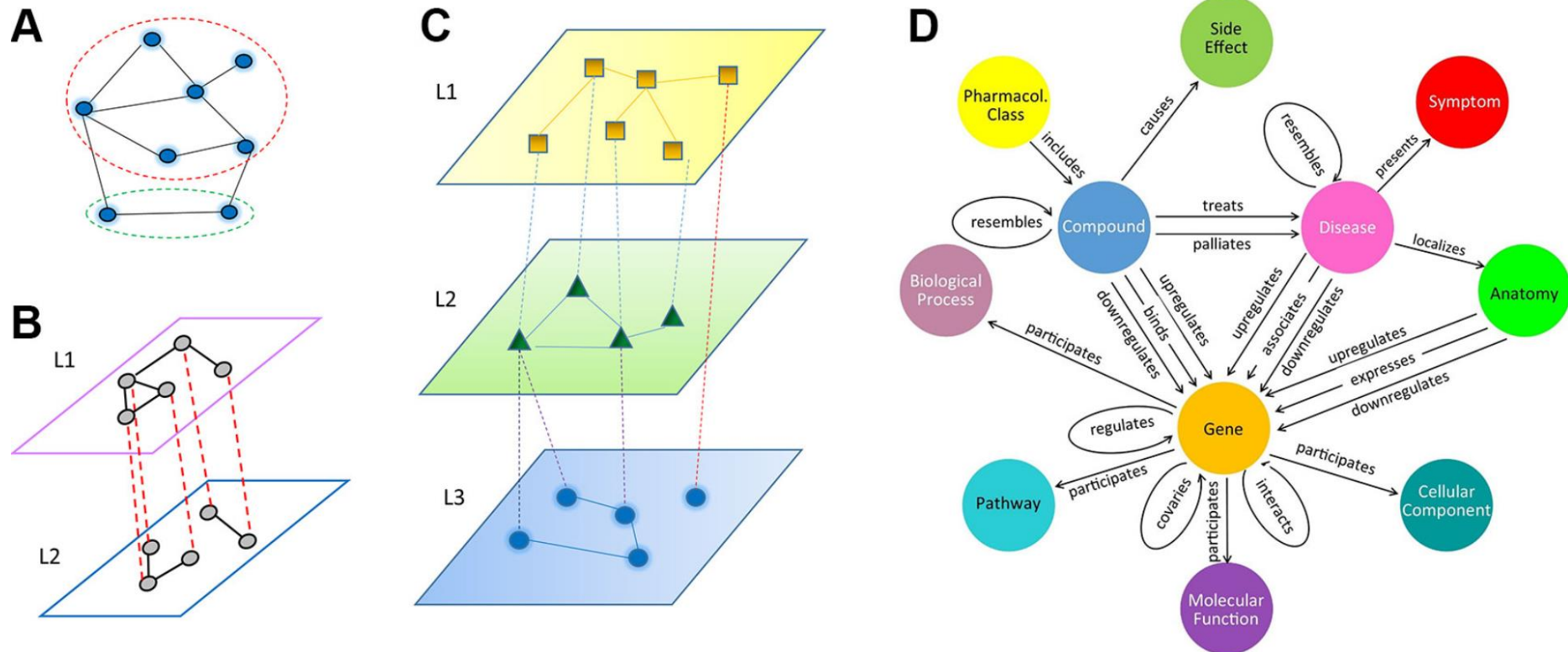
### **Networks – unifying integration and interactions**

- Nodes: biological features such as **microbial taxa** (abundance of a microbial taxon), **genes** (expression level), metabolites (concentration), and proteins (concentration); environmental (exposures) or other host features (demographics)
- Edges (connections between nodes):
  - empirically or statistically derived interactions;  
more generally: **association** between nodes, s.a. correlation between the abundance of two taxa/dependencies
  - weights to reflect association strength
  - directions to reflect “cause and effect”



# Interludium

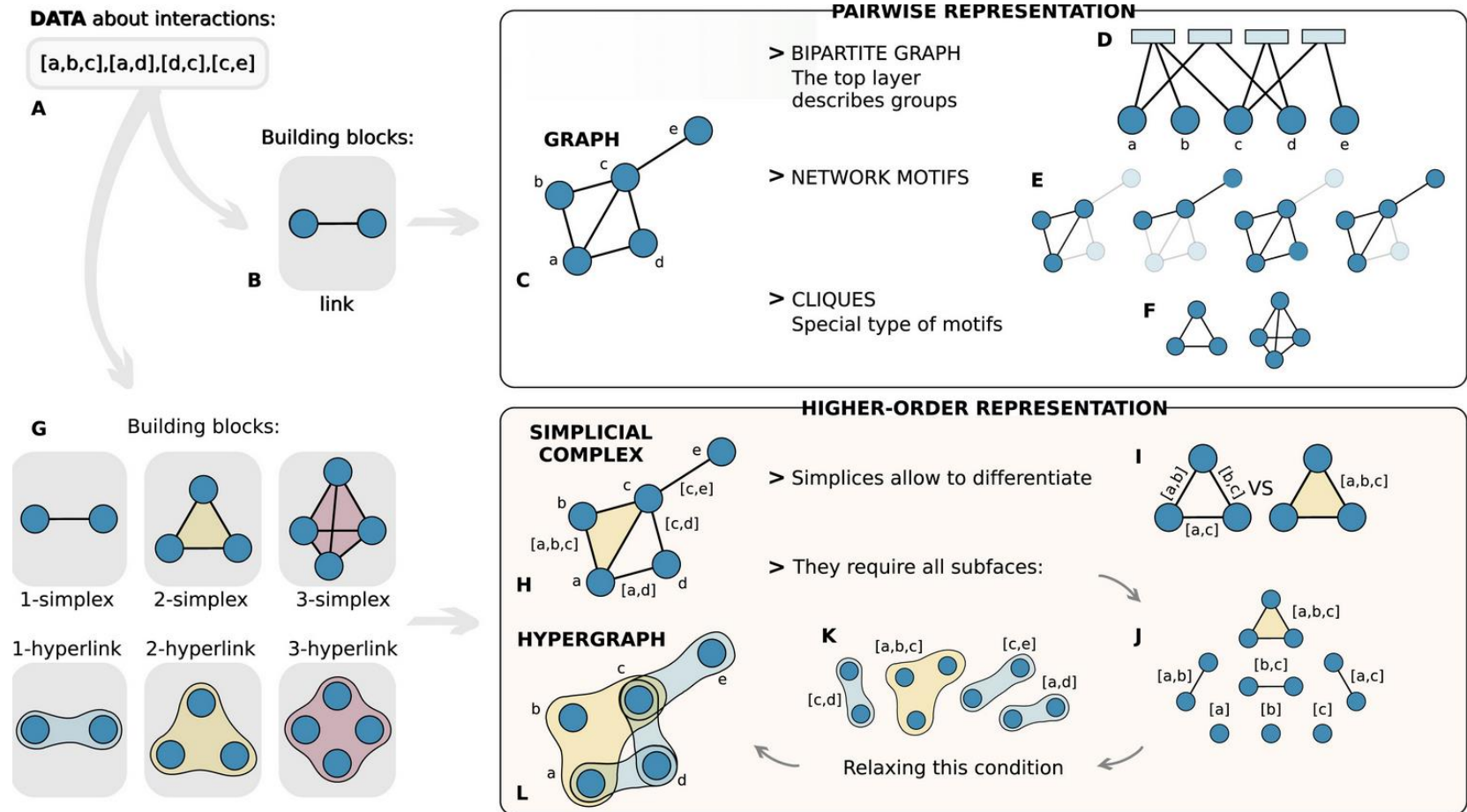
## Networks – organisation of network building blocks



(Lee et al. 2020)

doi: [10.3389/fgene.2019.01381](https://doi.org/10.3389/fgene.2019.01381)

# Networks – beyond pairwise interactions

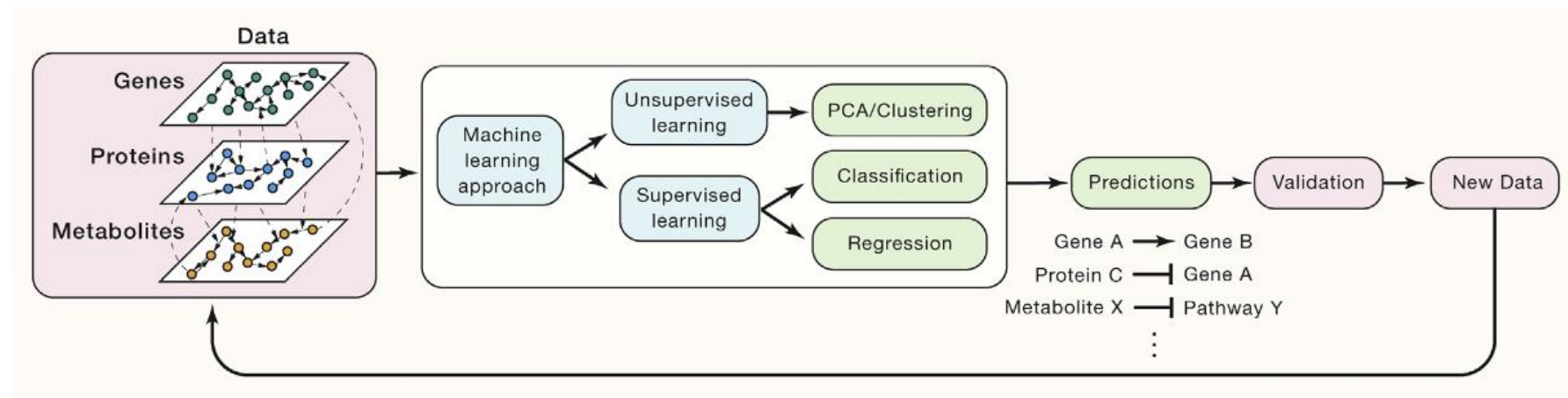


(Battiston et al. 2020)

doi: 10.1016/j.physrep.2020.05.004

## Networks – increasing role of machine learning

- Review 1: Increasing role of machine learning in biological network analysis



(Camacho et al. 2018)

doi: 10.1016/j.cell.2018.05.015

- Review 2: role of deep learning (graph neural networks GNNs)

(Muzio et al. 2021)

doi: 10.1093/bib/bbaa257

## Networks – increasing role of machine learning

Future developments:

- Interaction heterogeneity: often GNNs proposed for the learning of heterogeneous networks do not particularly distinguish between- and within-layer interactions
- designs for GNNs: often functions too heavily rely on heuristics; which conditions should they satisfy, given the (biological) data at hand
- interpretation: underperformance of GNNs in this sense; which nodes/edge contribute to the results

(Lee et al. 2020)

doi: [10.3389/fgene.2019.01381](https://doi.org/10.3389/fgene.2019.01381)

# Network construction – edge definition

## LabNet

```

1: procedure LASSO2NET( $X_i, X, B, fanout, best$ )
2:  $fit \leftarrow lasso.cv(X_i, X)$ 
3:  $\lambda_{cv} \leftarrow fit.lambda$ 
4:  $S \leftarrow fit.coeffs$ 
5:  $S \leftarrow sort(S, decreasing)[1 : best]$ 
6: while  $r < B$  do
7:  $X_i^{perm} \leftarrow permute(X_i)$ 
8:  $permfit \leftarrow lasso(X_i^{perm}, X, \lambda_{cv})$ 
9:  $update(counter[S])$  update counters of selected variables
10:  $r \leftarrow r + 1$ 
11:  $sel \leftarrow sort(counter[S], increase)[1 : fanout]$  order and select first fanout
12: return  $sel$ 

```

(Gadaleta & Van Steen 2014)

doi: 10.1371/journal.pone.0110451

	True	Pred	TP	FP	TN	FN	MCC	TPR	FPR	ACC	Time[sec]
<b>fused</b>	48	252	5	247	2205	43	0.001	0.10	0.100	0.884	177.66
<b>hier</b>	48	224	21	203	2249	27	0.170	0.4375	0.08	0.908	544.71
<b>group</b>	48	414	21	393	2059	27	0.102	0.4375	0.16	0.832	17.84
<b>LABnet</b>	48	98	16	82	2370	32	0.212	0.33	0.03	0.9544	206.2
<b>ridge perm</b>	48	0	0	0	2452	48	NA	0	0	0.9808	202.60
<b>enet perm</b>	48	86	10	76	2376	38	0.133	0.20	0.030	0.9544	201.93
<b>lasso</b>	48	254	8	246	2206	40	0.030	0.16	0.10	0.8856	1.345
<b>ridge</b>	48	254	8	246	2206	40	0.030	0.16	0.10	0.8856	1.28
<b>oneenet</b>	48	254	8	246	2206	40	0.030	0.16	0.10	0.8856	1.52

- Nine “Lasso’s” compared incl new LABNet for gene expression networks, addressing multicoll. and high-dimensionality (Gadaleta 2015)



# **PART 2**

## **Challenges & Opportunities**

### **by examples**

# **GWAS Systems Analysis for Precision Medicine**

*global networks*

## Previous summer school – post-GWAS

- Principles of GWAS:
  - Imputation
  - Meta-analysis
- Towards interpretation & Translation:
  - Functional annotation of disease loci through eQTLs and epigenetic data
  - Integrating GWAS outcomes and pathway information
  - Network analysis of GWAS outcomes: identification of important gene modules
  - Polygenic risk scores enhancing disease risk prediction

## Challenge 1: *How to define/construct edges?*

“Interaction is a kind of action that occurs as two or more objects have an effect upon one another.

The idea of a two-way effect is essential in the concept of interaction, as opposed to a one-way causal effect.”

(en.wikipedia.org; 14 Febr 2021)

## Masking

Genotype at locus B	Genotype at locus G		
	<i>g/g</i>	<i>g/G</i>	<i>G/G</i>
<i>b/b</i>	White	Grey	Grey
<i>b/B</i>	Black	Grey	Grey
<i>B/B</i>	Black	Grey	Grey

Genotype at locus A	Genotype at locus B		
	<i>b/b</i>	<i>b/B</i>	<i>B/B</i>
<i>a/a</i>	0	0	1
<i>a/A</i>	0	0	1
<i>A/A</i>	1	1	1

(Cordell et al. 2002)

doi: 10.1093/hmg/11.20.2463

### Compositional epistasis:

If the genotype at locus G is not *g/g* then the effect at locus B is not observable, it is masked. Allele G at locus G is epistatic to allele B at locus B.

### Mathematical heterogeneity model:

If we define the 'effect' of locus B to be a recessive disease model (B/B causes disease) then having A/A is sufficient to mask this effect.

## According to (statistical) genetics literature ...



(Photo: J Murken)

Compositional epistasis

Mechanistic interaction

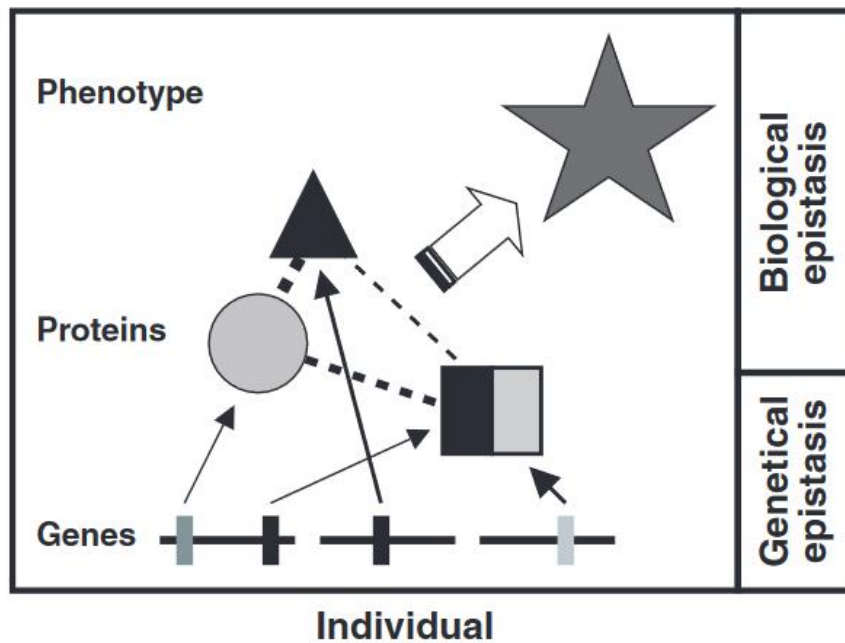
Fisher's epistasis

Statistical epistasis

Essential epistasis

...

# What is the problem we wish to solve?



## Precision Medicine

Prevention

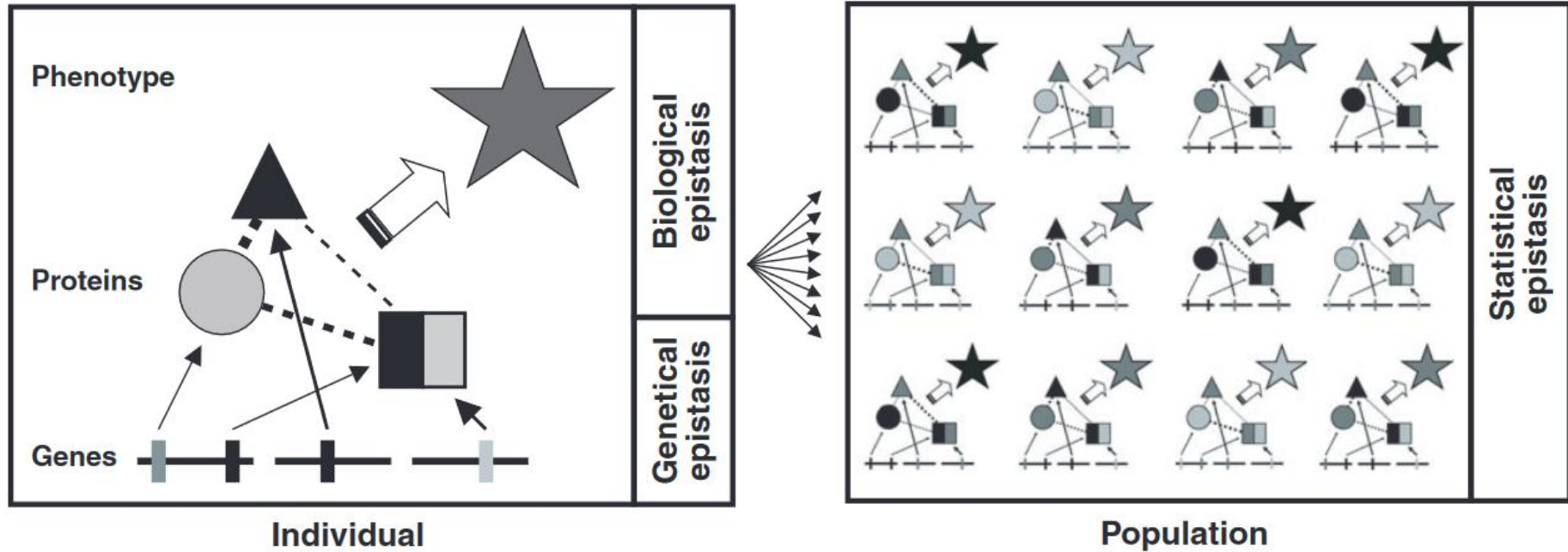
Diagnosis

Disease management

(Moore 2005)

doi : [10.1002/bies.20236](https://doi.org/10.1002/bies.20236)

## How to solve the problem?



(Moore 2005)

doi : 10.1002/bies.20236



# Traveling the world of gene-gene interactions

Kristel Van Steen

*Briefings in Bioinformatics*, Volume 13, Issue 1, January 2012, Pages 1–19, <https://doi.org/10.1093/bib/bbr012>



Review | [Open Access](#) | Published: 06 March 2019

## How to increase our belief in discovered statistical interactions via large-scale association studies?

[K. Van Steen](#)  & [J. H. Moore](#)

*Human Genetics* **138**, 293–305(2019) | [Cite this article](#)

**1945** Accesses | **4** Citations | **2** Altmetric | [Metrics](#)

# Traveling the world of gene-gene interactions

Review Paper | Published: 04 July 2012

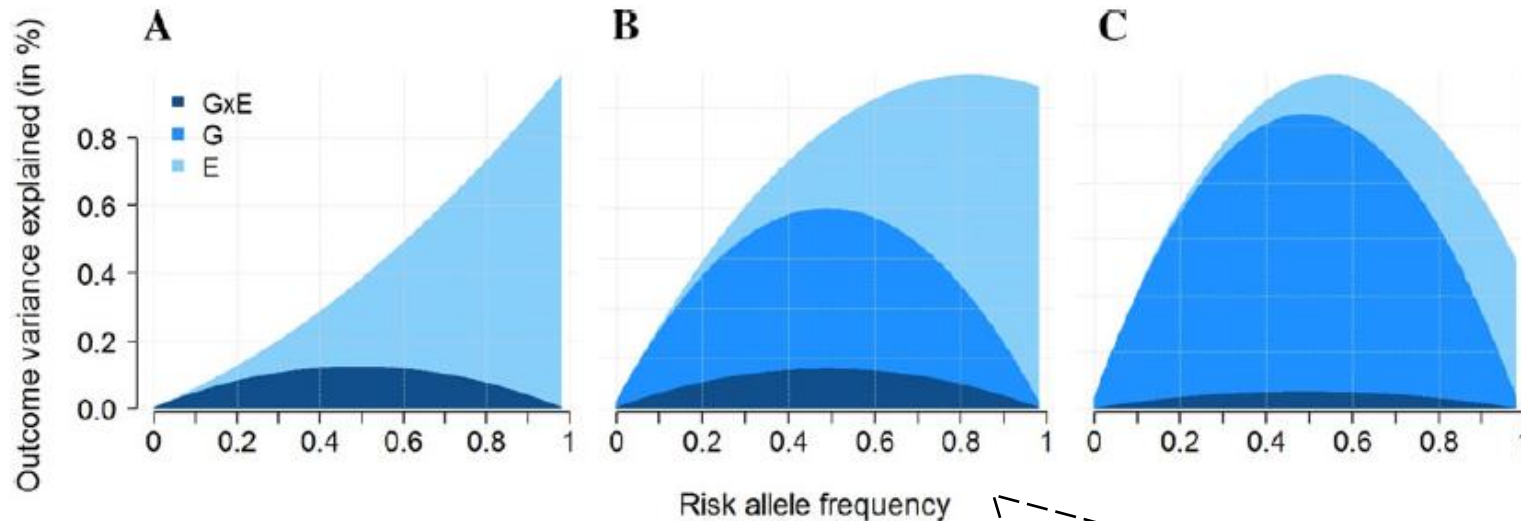
## Challenges and opportunities in genome-wide environmental interaction (GWEI) studies

[Hugues Aschard](#) , [Sharon Lutz](#), [Bärbel Maus](#), [Eric J. Duell](#), [Tasha E. Fingerlin](#), [Nilanjan Chatterjee](#), [Peter Kraft](#) & [Kristel Van Steen](#)

[Human Genetics](#) **131**, 1591–1613(2012) | [Cite this article](#)

GWAIS and GWAIS “via the common genetic component they involve, share quite a number of challenges. ... The most interesting types of G x E interactions are those that are coined “non-removable”, in the sense that the evidence of (statistical) interaction exists when no obvious monotone transformation of the trait exists (i.e., rescaling of the trait) that removes the interaction.”

## Regression paradigm – easy, too easy?



“Examples of attribution of phenotypic variance explained by an interaction effect.

“Proportion of variance of an outcome  $Y$  explained by a genetic variant  $G$ , an exposure  $E$  and their interaction  $G \times E$  in a model harboring a pure interaction effect only ( $Y = \beta_{GE} \times G \times E + \epsilon$ ). The exposure  $E$  follows a normal distribution with a standard deviation of 1 and mean of 0 (A), 2 (B), and 4 (C). The genetic variant is biallelic with a risk allele frequency increasing from 0.01 to 0.99. The interaction effect: maximum of the variance explained by the model equals 1%.” (Aschard 2016) 10.1002/gepi.21989

## Traveling the world of gene-gene interactions

Review | [Open Access](#) | Published: 06 March 2019

### How to increase our belief in discovered statistical interactions via large-scale association studies?

[K. Van Steen](#)  & [J. H. Moore](#)

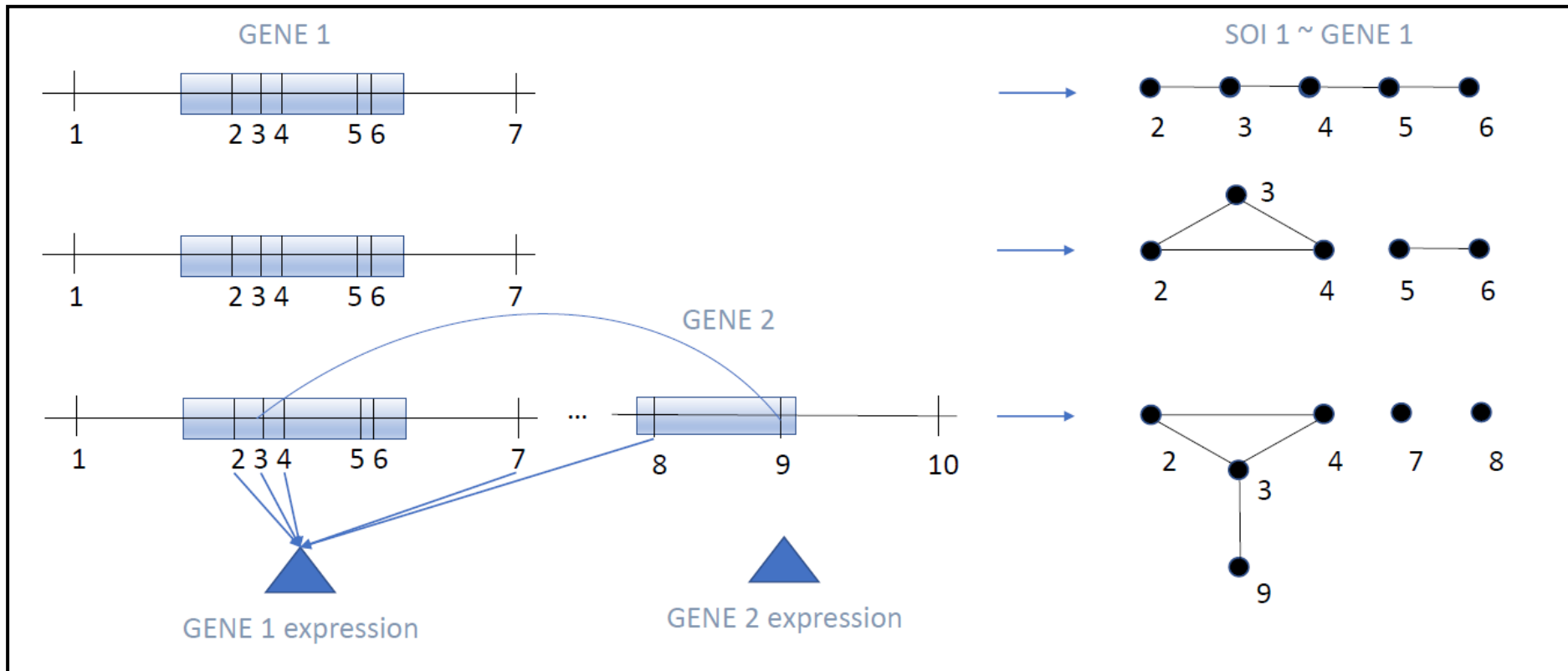
[Human Genetics](#) **138**, 293–305(2019) | [Cite this article](#)

**1945** Accesses | **4** Citations | **2** Altmetric | [Metrics](#)

“Finally, **moving from localization to function** will be essential to explain molecular mechanisms playing a synergetic role in human complex diseases. Although there is still a long way to go before epistasis findings can be brought into clinical practice, our practical and theoretical experience has shown that, by taking advantage of various methodologies and by examining data from different angles, it is feasible to reveal strong evidence for biological gene interactions derived from genome-wide SNP panels.”

## Challenge 2a: *Which unit of analysis?* – “unsupervised” sets

Seek a compromise between holistic and reductionist approach: mini-systems  
(Fouladi 2018, PhD thesis manuscript)



- Define Laplacian for each gene's graph

$$L_{ij} = \begin{cases} W_{ij}, & \text{for } i \neq j \\ -\sum_{l=1}^n W_{il}, & \text{for } i = j. \end{cases}$$

- Define the diffusion kernel

$$K_L = e^{\beta L},$$

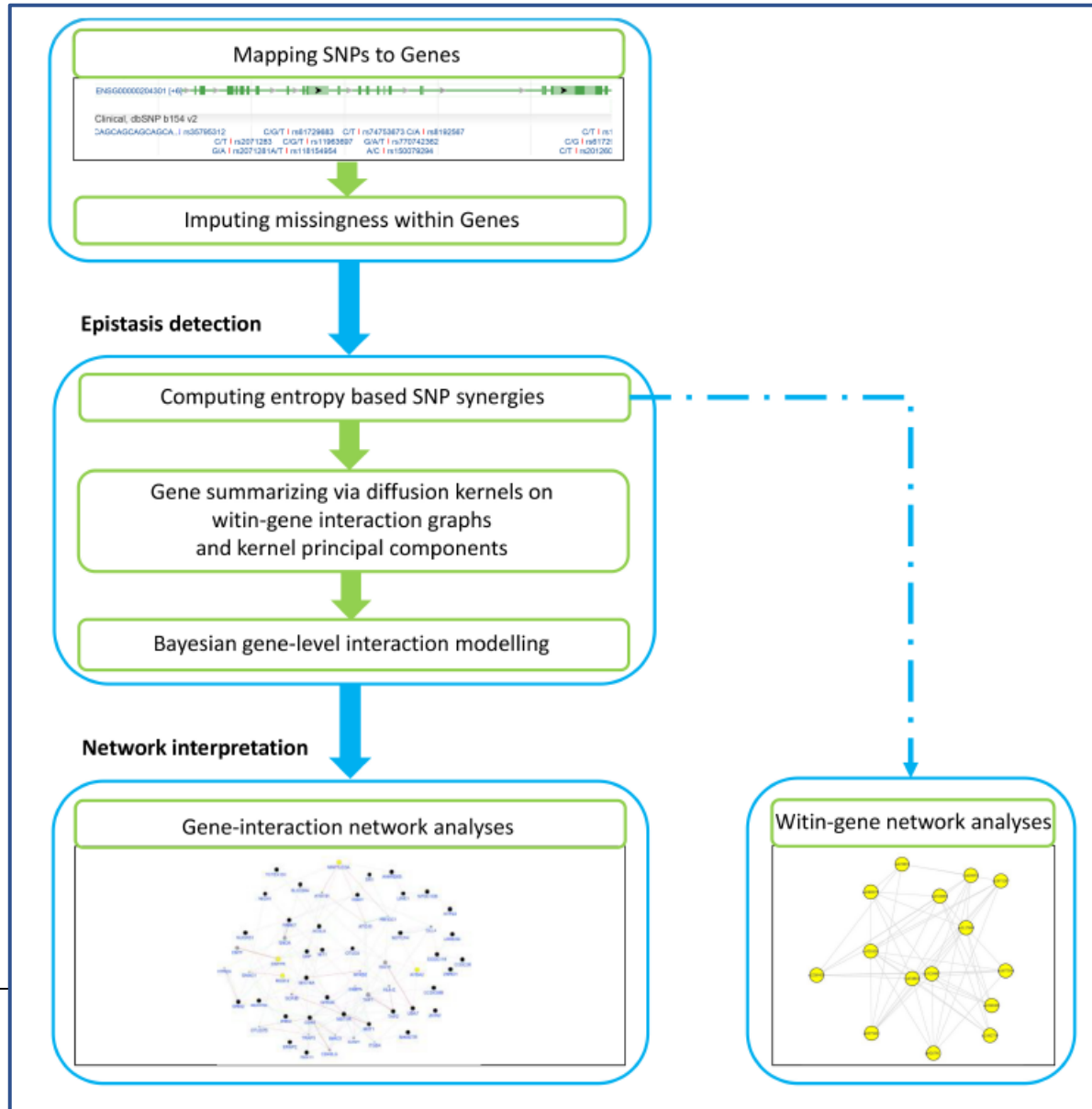
which is a power series

$$e^{\beta L} = \sum_{k=0}^{\infty} \frac{1}{k!} (\beta L)^k,$$

- Potentially reduce dimensionality via kernel PCA and the kernel K

$$K = GK_L G^T.$$

# Challenge 2b: Which unit of analysis? – “supervised” sets



Gene-based epistasis model:

Bayesian semiparametric regression and sparsity inducing priors (Antonelli et al. 2020)

Workflow in Walakira et al. (2021 – submitted)

## Challenge 3: *What about “reproducibility”?*

- Methods reproducibility:

“A method is reproducible if reusing the original code leads to the same results [focus in machine learning; setting seeds, same hardware, ...]”

- Results reproducibility:

“A result is reproducible if a reimplementation of the method generates statistically similar values. [confid. intervals seldomly provided with deep learning]”

- Inferential reproducibility:

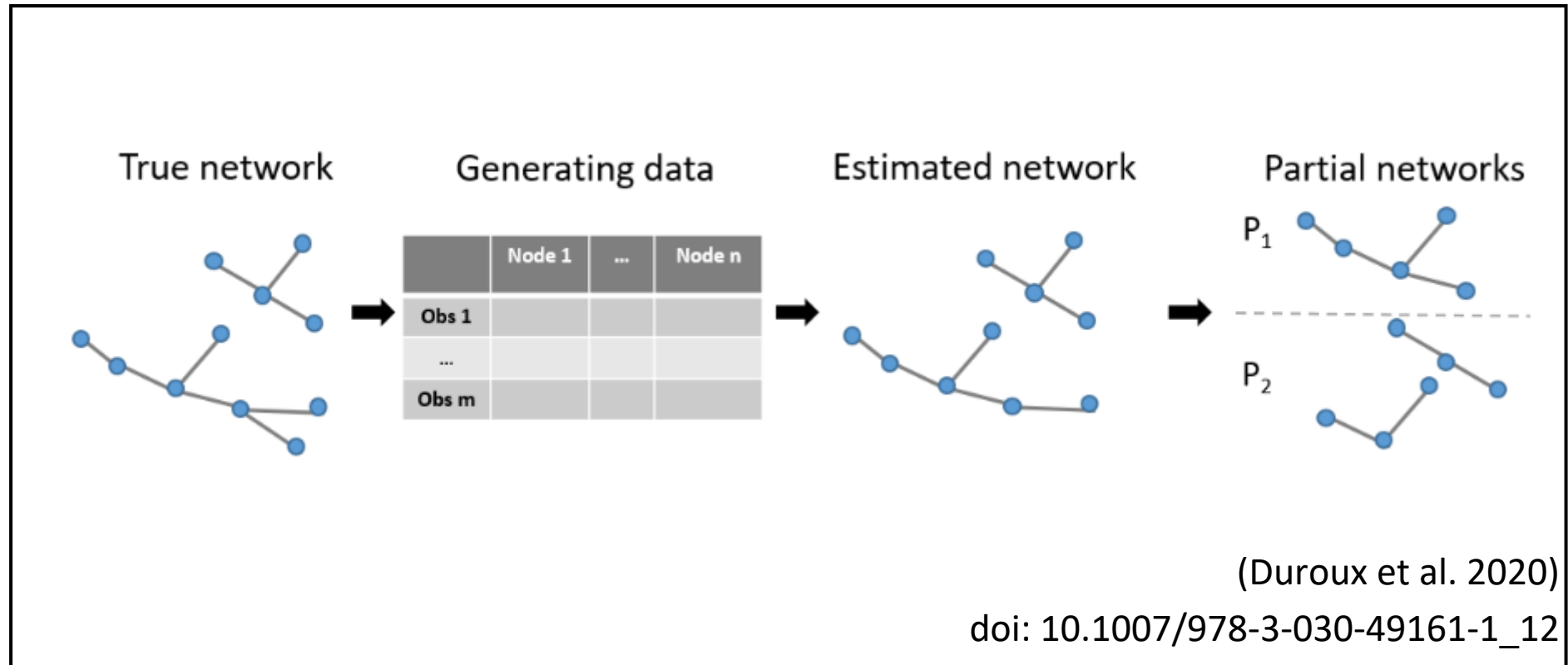
“A drawn conclusion is reproducible if one can draw it from a different experimental setup [not about numerical results; experimental design variations and susceptibility of methods].”

(Bouthillier et al. 2019; Goodman et al. 2016)



## Network based model aggregation

- Simulate partial representations of a true network




[IFIP International Conference on Artificial Intelligence Applications and Innovations](#)

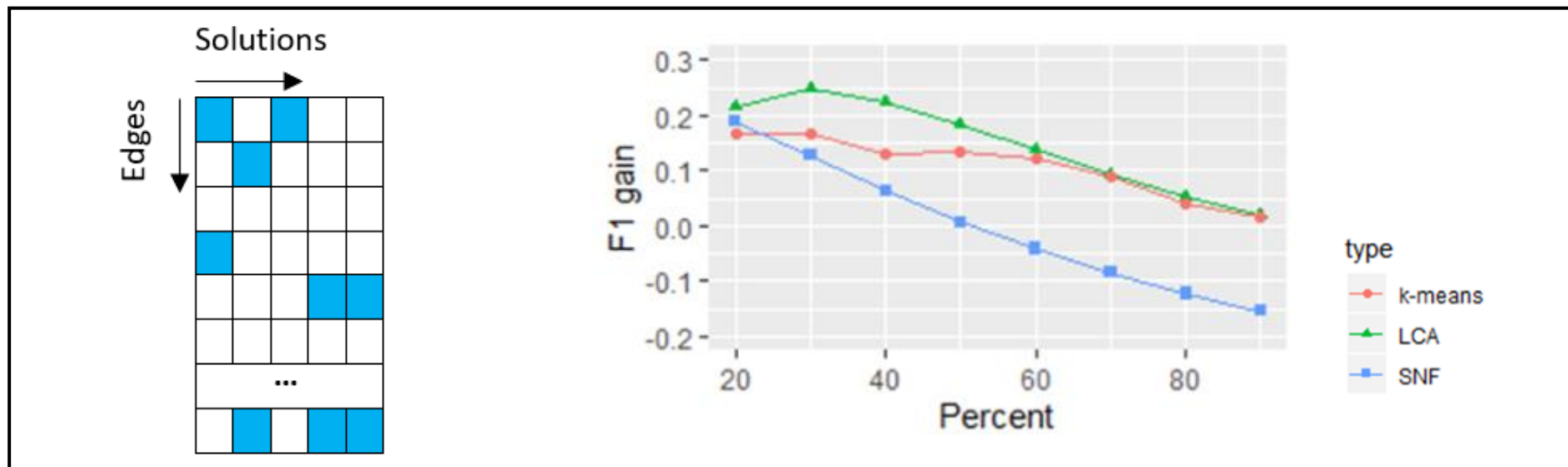
AIAI 2020: [Artificial Intelligence Applications and Innovations](#) pp 128-140 | [Cite as](#)

# Network Aggregation to Enhance Results Derived from Multiple Analytics

Authors

[Authors and affiliations](#)

Diane Duroux , Héctor Climente-González, Lars Wienbrandt, Kristel Van Steen



## *Interludium*

### Reviews on network similarities

(Tandardini et al. 2019)

doi : 10.1038/s41598-019-53708-y

- Comparison of comparing networks
  - Known node-correspondence methods (e.g. DeltaCon)
    - Based on comparison of the similarities between all node pairs in the two graphs; Similarity between node pair depends on all r-paths connecting  $(i,j)$ , which is more sensitive than measuring overlap between two edge sets
    - In weighted graphs, the bigger the weight of a removed edge, the bigger the impact on the (Matusita) distance between  $(i,j)$
  - Unknown node-correspondence methods (e.g. spectral methods)
    - Based on spectrum of network representation (e.g. Laplacian) and e.g. taking Euclidean distance

## Interludium

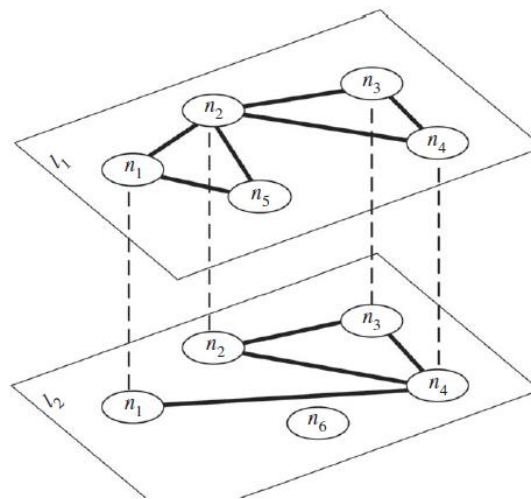
(Bródka et al. 2018)

doi: 10.1098/rsos.171747

### Reviews on network similarities

- Quantifying layer similarity in multiplex networks

- Tower of Babel in layer similarity measures; Relationship between circulating measures?
- How to choose the appropriate measure given a specific dataset?



	$n_1$	$n_2$	$n_3$	$n_4$	$n_5$	$n_6$
$l_1$	2	4	2	2	2	n.a.
$l_2$	1	2	2	3	n.a.	0

property matrix

# Microbiome Systems Analysis for Precision Medicine

*individual-specific networks*

## Microbiome systems analytics

*or how components of a microbiome system interact*

“Novel associations between the human microbiome and health and disease are routinely emerging, and important host–microbiome interactions are targets for new diagnostics and therapeutics. Understanding how broadly host–microbe associations are maintained across populations is revealing individualized host–microbiome phenotypes that can be integrated with other ‘omics’ data sets to enhance precision medicine.”

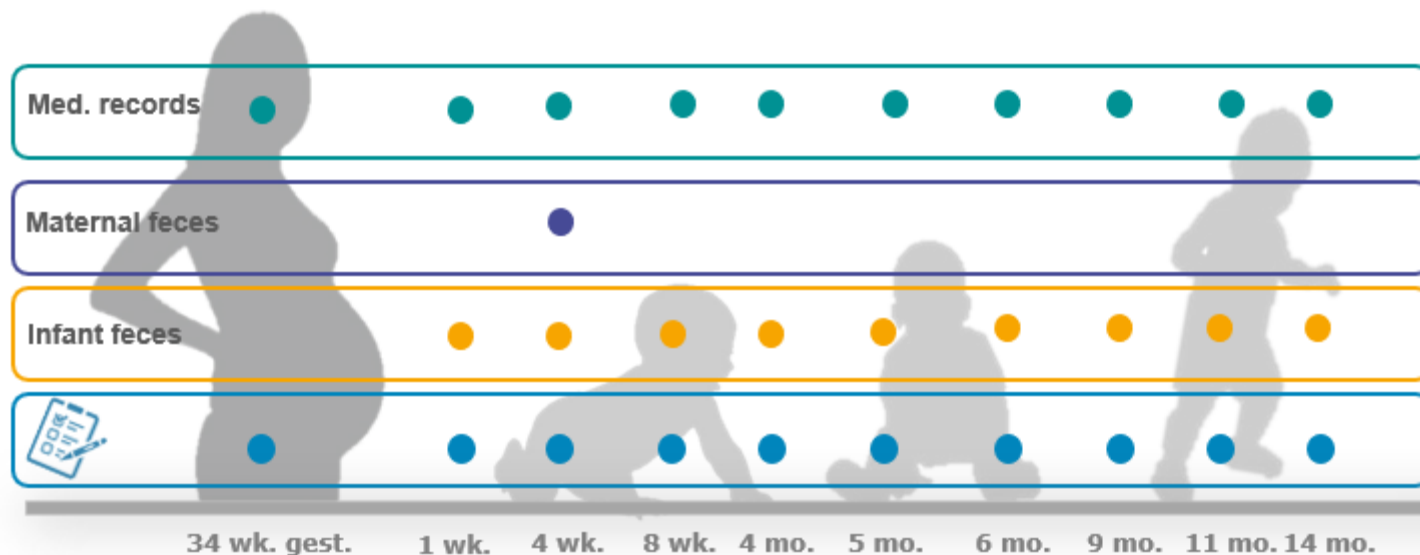
(Petrosino 2018)

doi: [10.1186/s13073-018-0525-6](https://doi.org/10.1186/s13073-018-0525-6)

## Data from Lucki Cohort Study

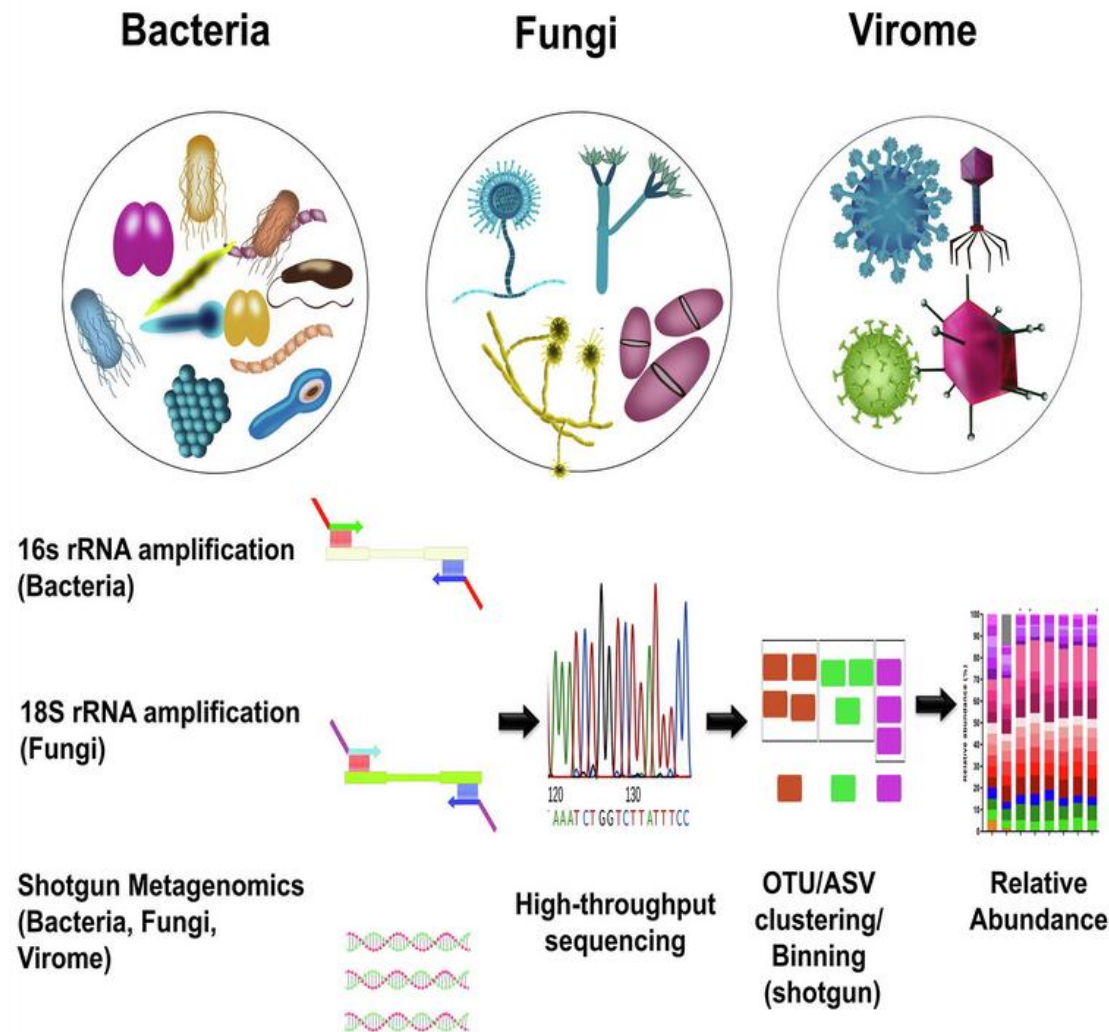
- Participants enrolled in the South Limburg area (the Netherlands): professionals involved in mother and childcare + internet
- Currently, ~140 newborns and their parents

Microbial profiling by next-generation sequencing of 16S rRNA V3-V4 hypervariable gene region; Amplicon Sequence Variants identified using DADA2-based pipeline; clr-transformation of data to account for compositionality (ALDEx2 package)



# Interludium

## Nature of the data

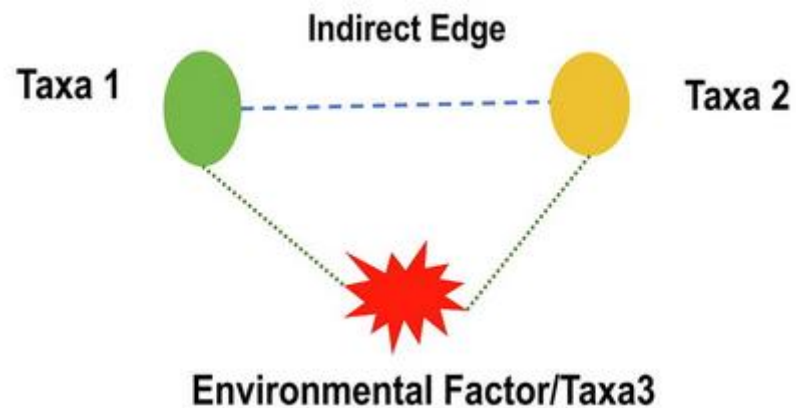
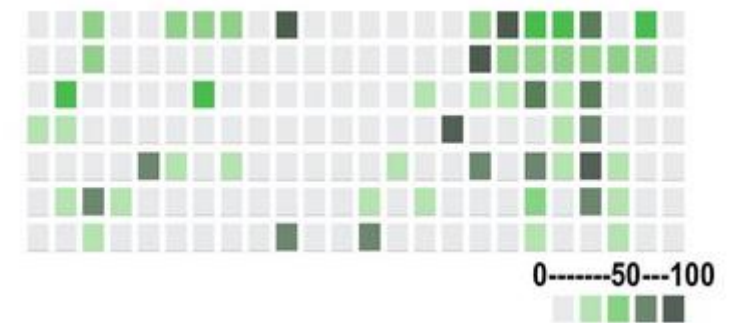
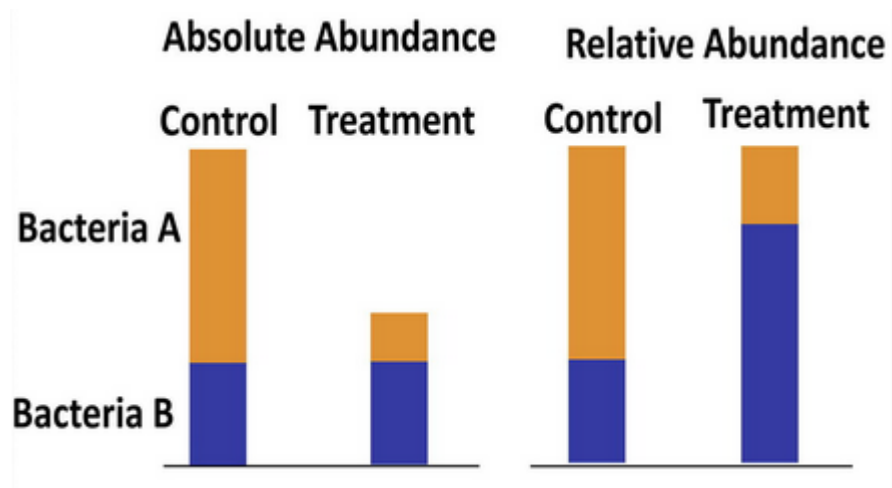


(Matchado et al. 2021)  
doi : 10.1016/j.csbj.2021.05.001



## Interludium

### Biases



- Compositionality
- Sparsity
- Spurious associations

(Matchado et al. 2021)

doi : 10.1016/j.csbj.2021.05.001

# Interludium

## Consequences

Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi $\phi$ $\rho$
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

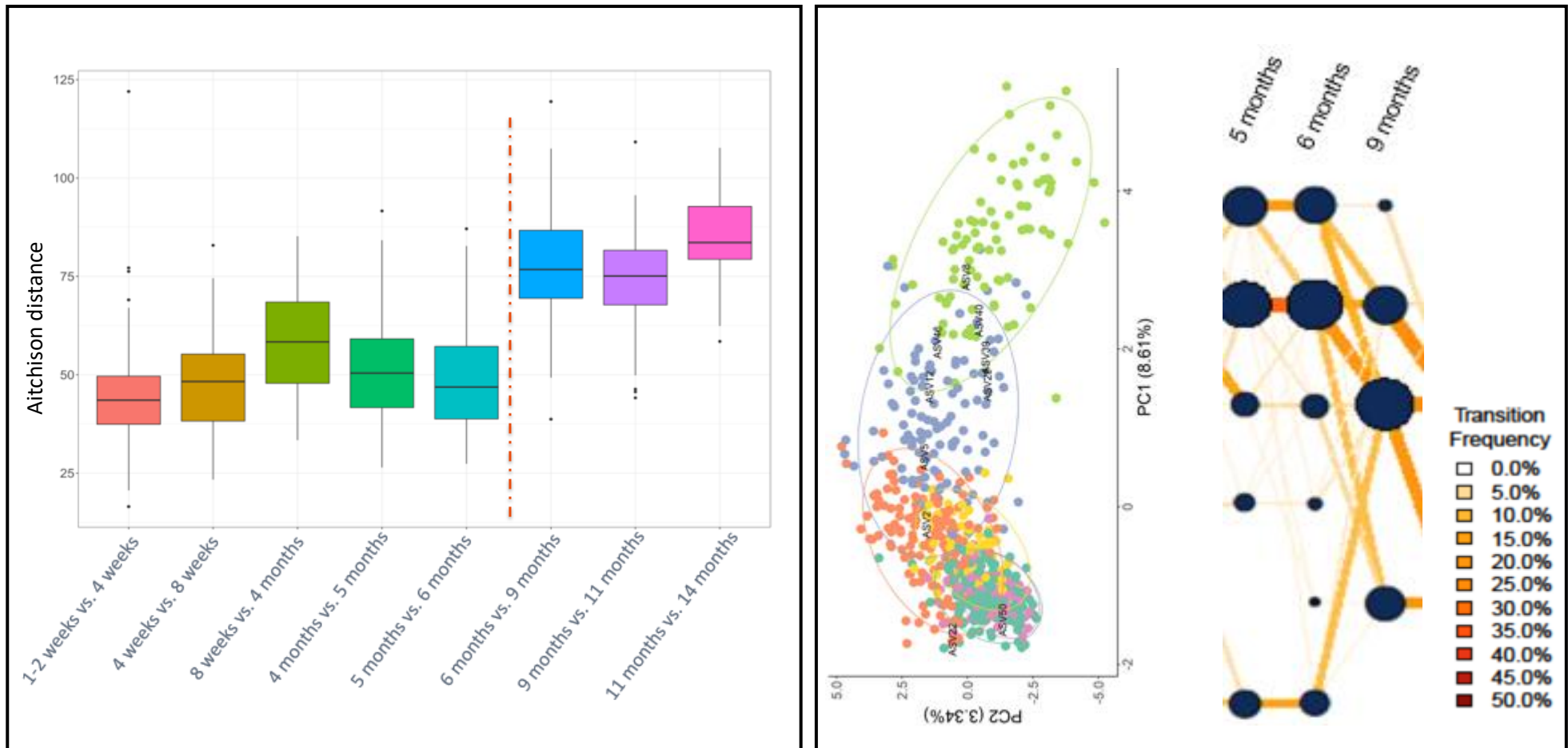
$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{g=1}^D \left[ \ln \frac{x_{gi}}{g(\mathbf{x}_i)} - \ln \frac{x_{gj}}{g(\mathbf{x}_j)} \right]^2}$$

CoDA (Aitchison, 1986)

(Gloor et al. 2017)

doi : 10.3389/fmicb.2017.02224

## Milestone in microbial community maturation: 6 to 9 months



## Challenge: *How to construct edges?*

- Modelling conditional dependence:

- differentiate indirect from direct associations
- infer a sparse inverse covariance matrix for a network
- account for confounders

(Matchado et al. 2021)

doi : 10.1016/j.csbj.2021.05.001

mLDM – metagenomic Lognormal-Dirichlet-Multinomial (Yang et al. 2017)

SPIEC-EASI – SParse Inverse Covariance Estimation for Ecological Association Inference (Kurtz et al. 2014)

MAGMA – Microbial Association Graphical Model Analysis (Cougoul et al. 2019)

## Challenge: edge construction

- Correlation as measure of association (co-occurrence)

### SparCC – Sparse Correlations for Compositional data

(Friedman and Alm 2012)

- First method to infer correlations between latent absolute abundances
- Pos. definiteness of var-cov matrix not guaranteed;  $r^2$  may be  $>1$

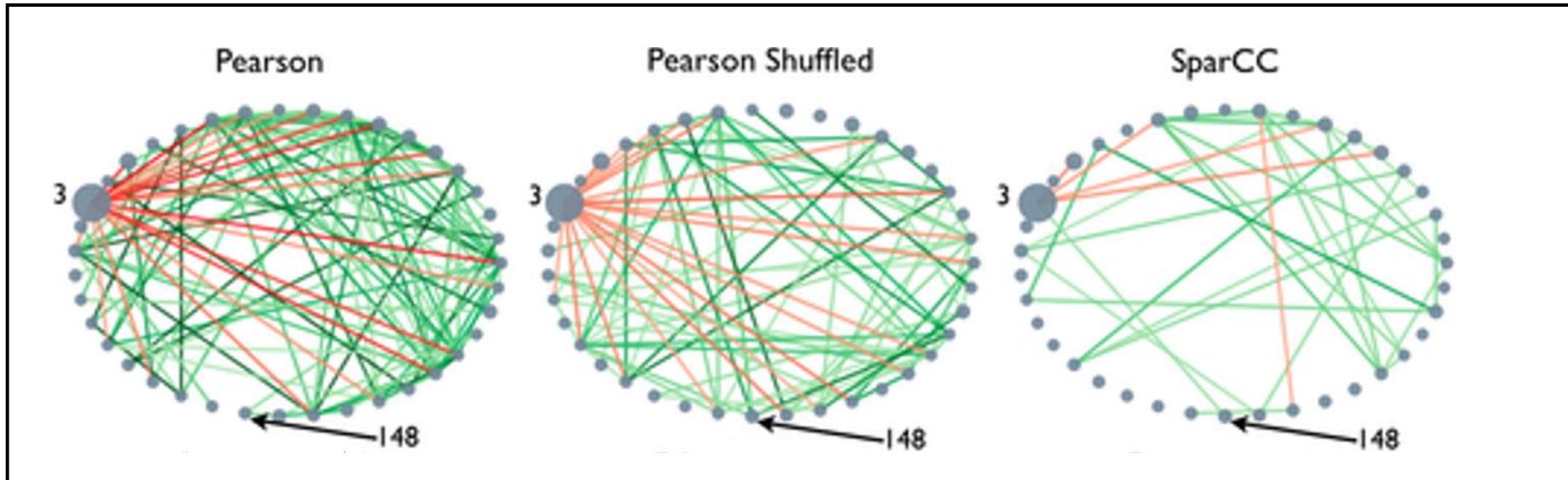
Tools	Principle/Models	Advantages	Limitation
<b>Correlation based Methods</b>			
SparCC (2012) python r-sparcc	<ul style="list-style-type: none"> <li>• Pearson correlations from log-transformed abundance</li> <li>• Bayesian approach to differentiate true fractions from the observed counts and to handle sparsity</li> <li>• Log-ratio transformed abundance/count matrix</li> </ul>	<ul style="list-style-type: none"> <li>• Handles compositionality bias and sparsity</li> </ul>	<ul style="list-style-type: none"> <li>• High computational complexity due to the iterative approximation approach</li> <li>• Nonlinear relationships cannot be detected</li> </ul>
CCLasso (2015) R package	<ul style="list-style-type: none"> <li>• Latent variable model with l1-norm shrinkage method</li> <li>• simple pseudo count implementation</li> <li>• Log-ratio transformed abundance/count matrix</li> </ul>	<ul style="list-style-type: none"> <li>• Faster than SparCC</li> <li>• Handles Compositionality bias</li> </ul>	<ul style="list-style-type: none"> <li>• Nonlinear relationships cannot be detected</li> <li>• Study only pairwise correlations between microbiomes [107]</li> </ul>

(Matchado et al. 2021)

doi : 10.1016/j.csbj.2021.05.001

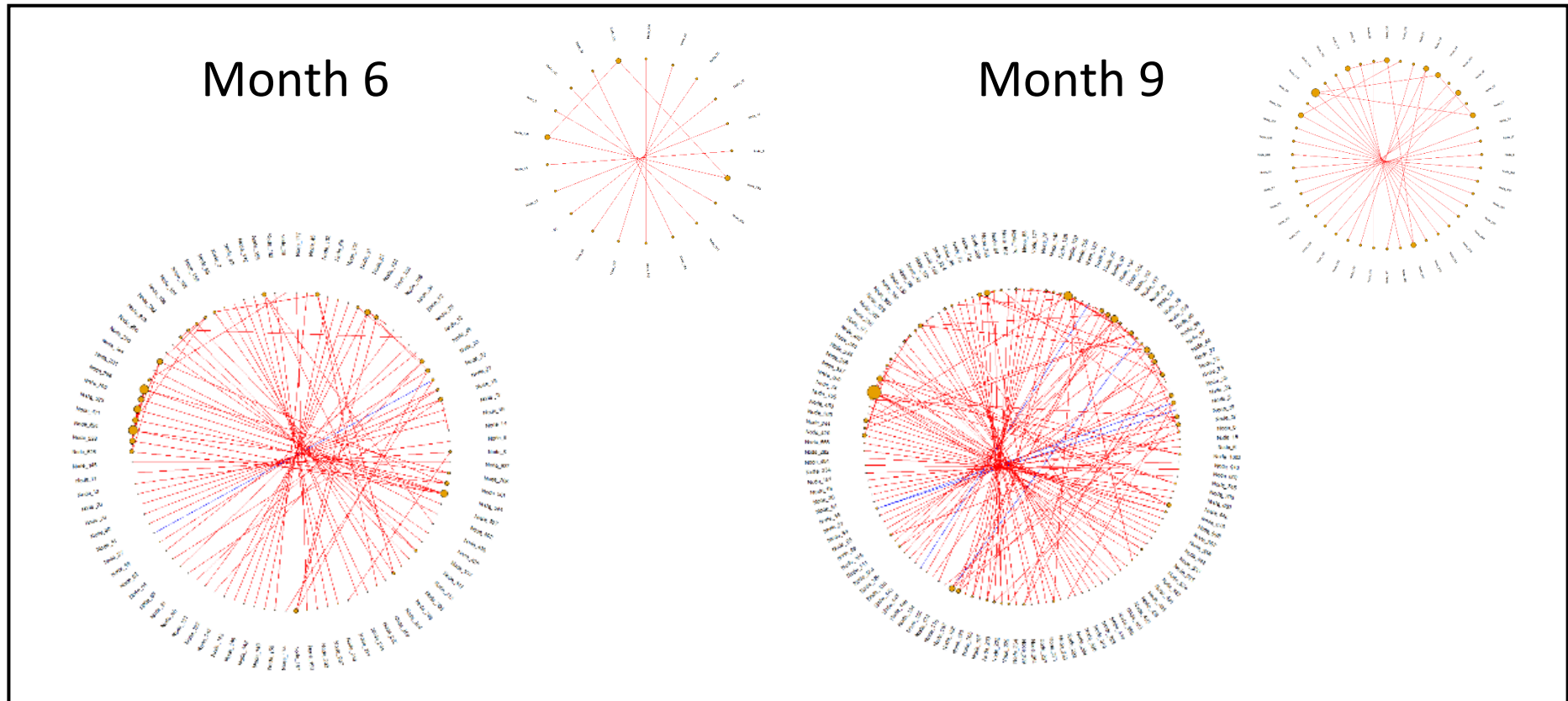
### FastSpar (Watts et al. 2019)

## Challenge: edge construction



(Friedman and Alm 2021)  
doi: [10.1371/journal.pcbi.1002687](https://doi.org/10.1371/journal.pcbi.1002687)

## Global network construction



For clarity, only edges corresponding to correlations with magnitude  $>0.4$  or  $>0.5$  (bottom, resp. top) are drawn. Red (blue): positive (negative) association

## Individual-specific interaction modelling

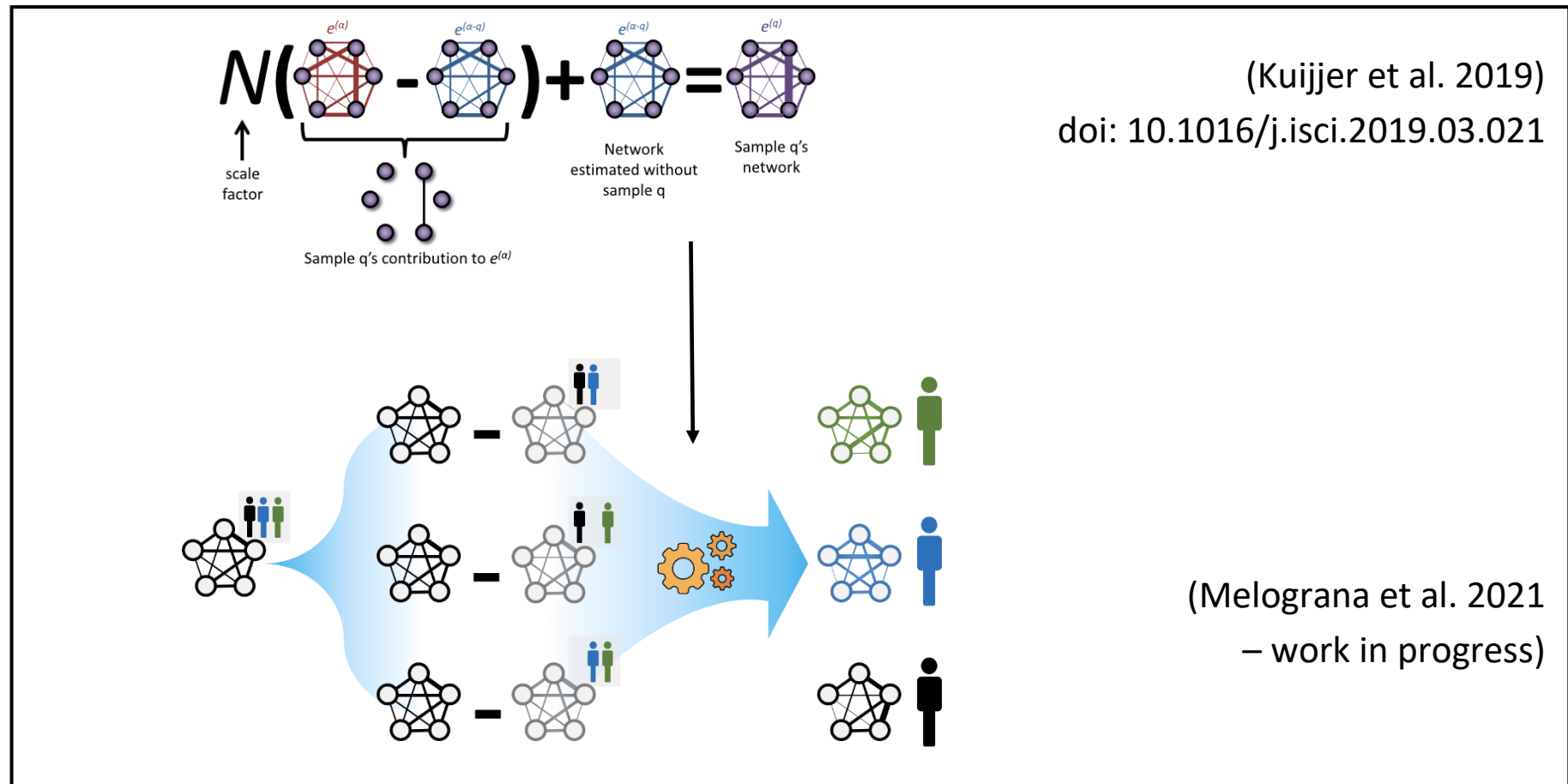
### • What?

- Networks that refer to co-occurrence, association, interaction
- In the literature:
  - Usually based on multiple measurements for the same individual (e.g. neurosciences)
  - Often individual-specific nodes on a fixed edge template common to all individuals (e.g., protein-protein interaction network, gene regulatory network)
- Recently:
  - **Individual-specific edges** (individual-specific node information available or not)



# Individual-specific interaction modelling

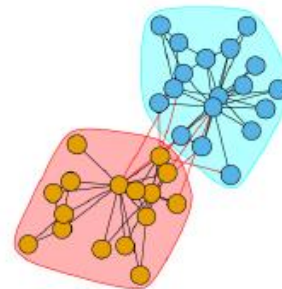
## • How?



- **Why?**

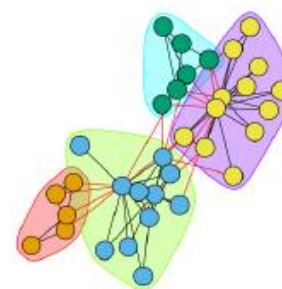
- Networks derived from a collection of individuals are often seen as models for an “average” individual
- Translating network interpretation strategies from pop. to indiv. assumes “deductions” can be made to the level of the individual
- Pop network interpretation strategies apply to ISNs:

a) Ground Truth

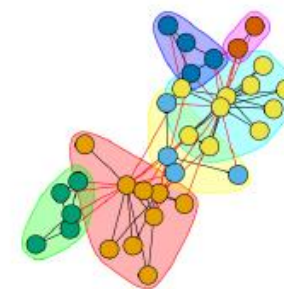


Modularity: 0.37

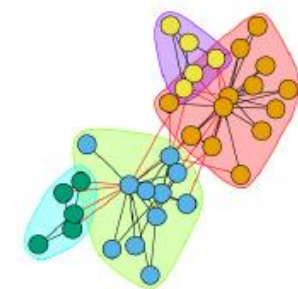
b) Louvain

Modularity: 0.45  
Similarity: 0.77

c) Girvan-Newman

Modularity: 0.34  
Similarity: 0.68

d) Walktrap

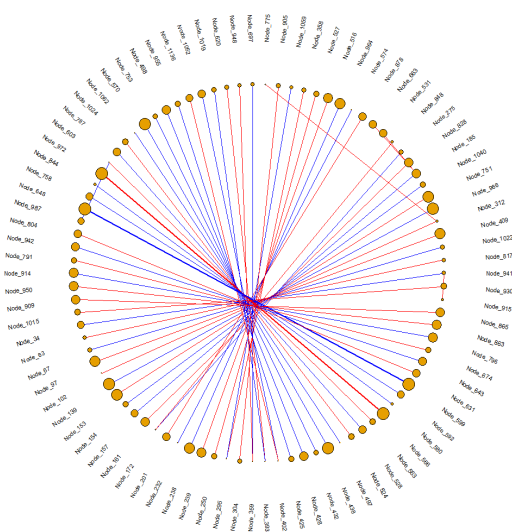
Modularity: 0.44  
Similarity: 0.79

## Edge-oriented approach

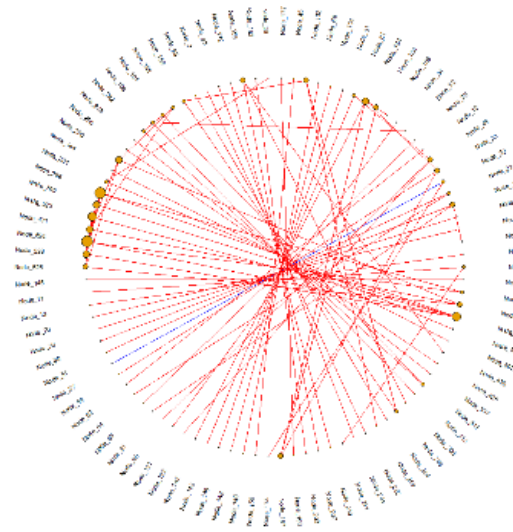
	Time point 1			Time point 2		
	Edge			Edge		
Indiv.	1	...	K	1	...	K
1						
...						
N						

## Node-oriented approach

	Time point 1			Time point 2		
	Node			Node		
Indiv.	1	...	L	1	...	L
1						
...						
N						

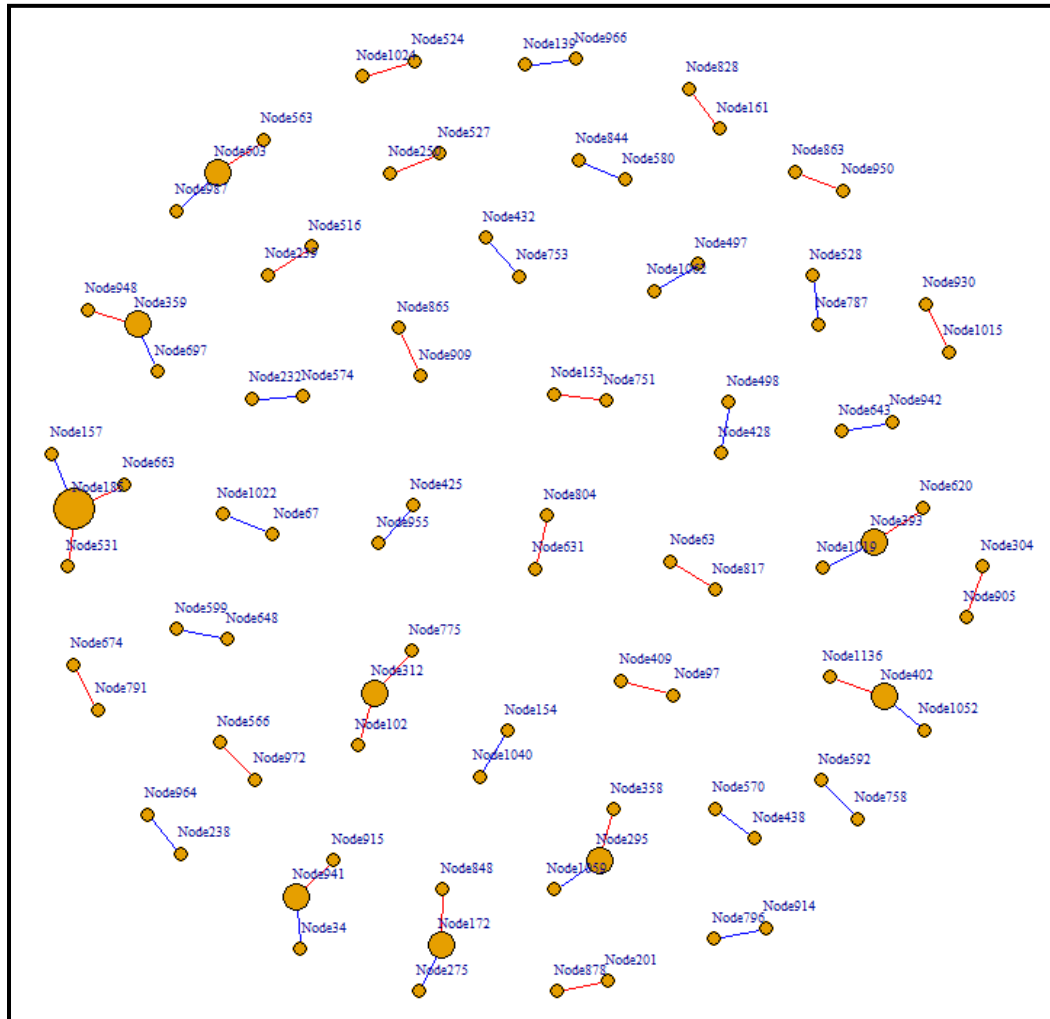


Red (blue): positive (negative) assoc.  
 Right: Abs correlations >0.4  
 Left: Top 50 differential edges



Lucki cohort

## Edge-wise filtering based on dynamics

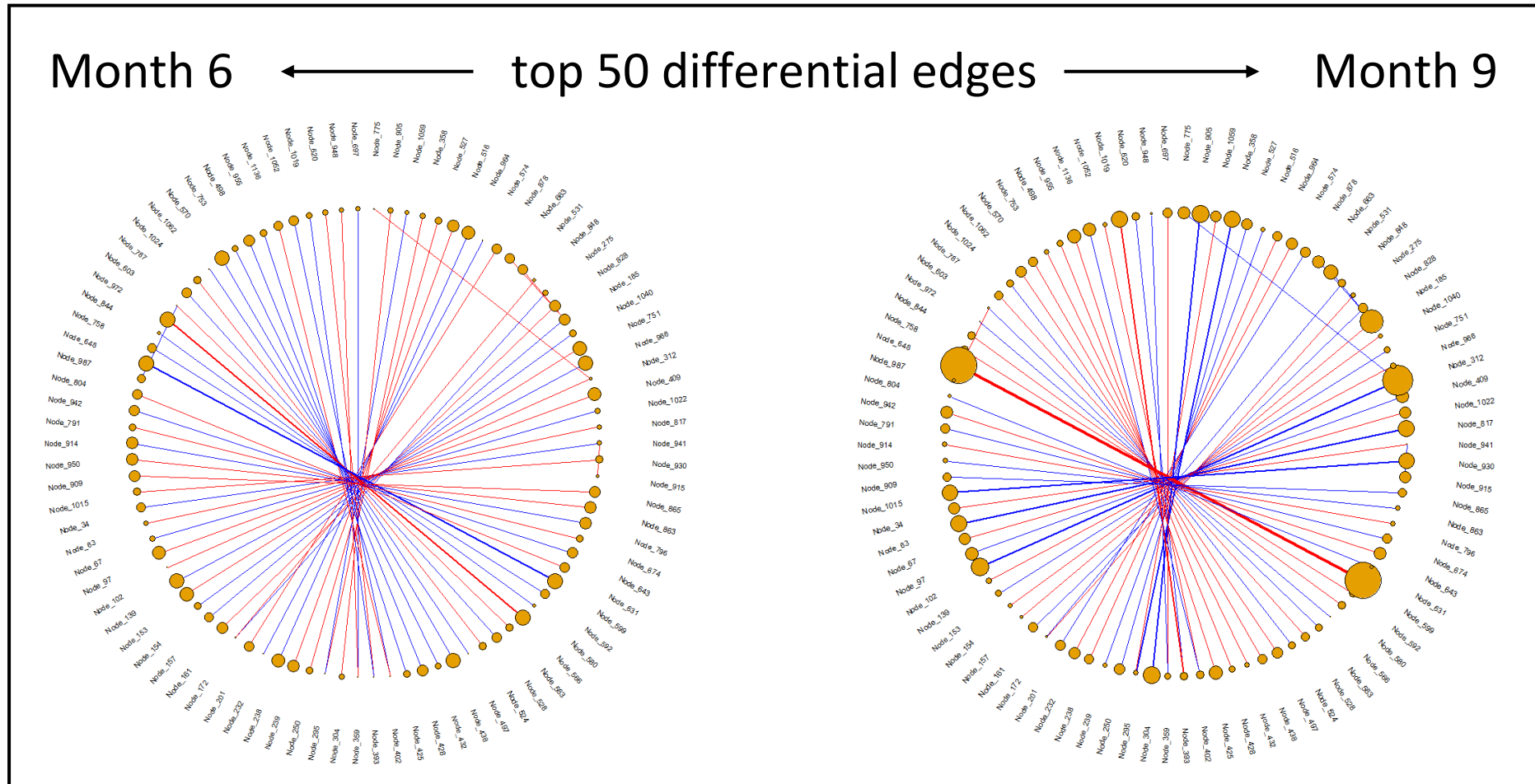


- LIMMA paired moderated t-test
- Multiple testing adjustment for 654,940 pairs with Benjamini-Yekutieli False Discovery Rate (FDR) control and Bonferroni

Shown:

binary significance network; top 50 differential edges; node size: degree; color code: association strength (red:  $m_9 > m_6$ )

# Edge-wise filtering based on dynamics



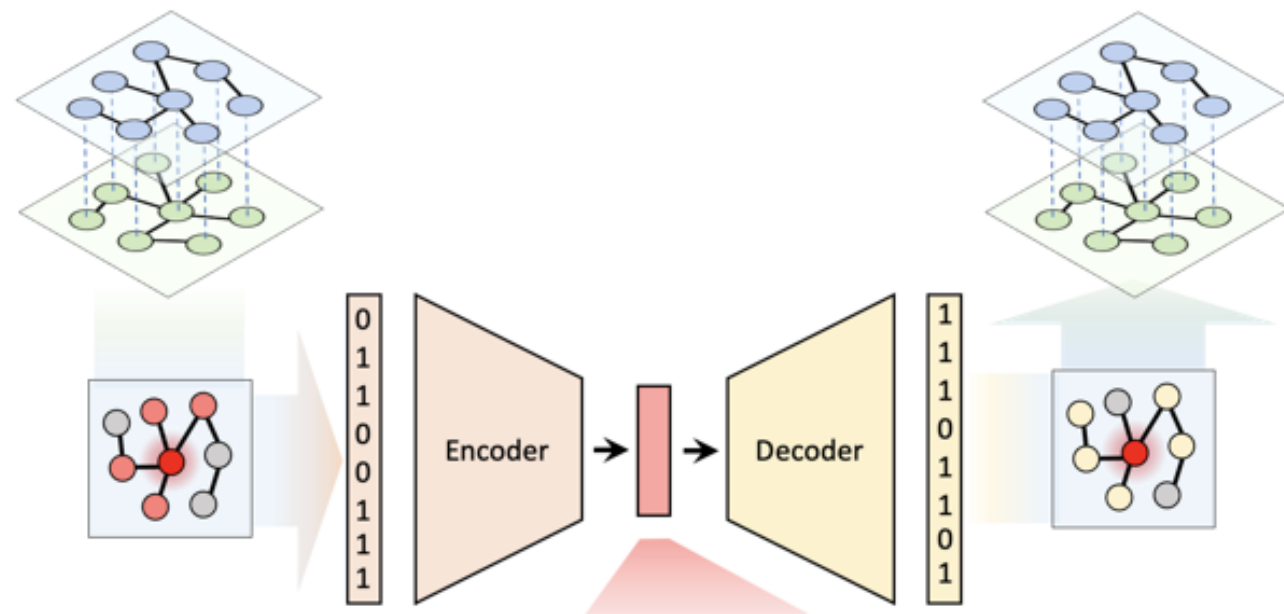
LIMMA paired moderated t-test; multiple testing corrected; blue (red): neg (pos) corr



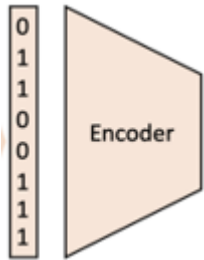
## Opportunity 1: Systems assessment of time-course similarity between individuals

Our solution:

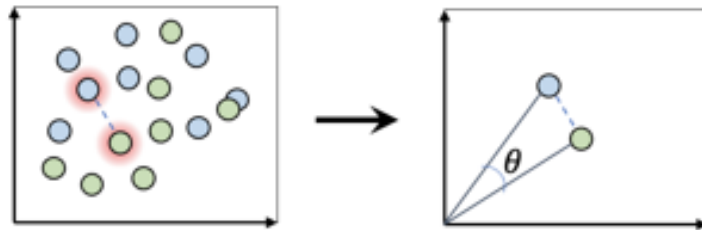
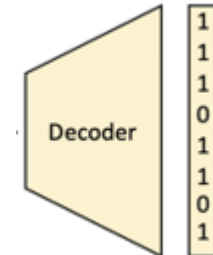
- Embedded analysis of multiplex ISNs (one individual, multiple conditions or time points) in a low-dimensional space with an Encoding-Decoding Neural Network (EDNN) algorithm



## Random walker approach



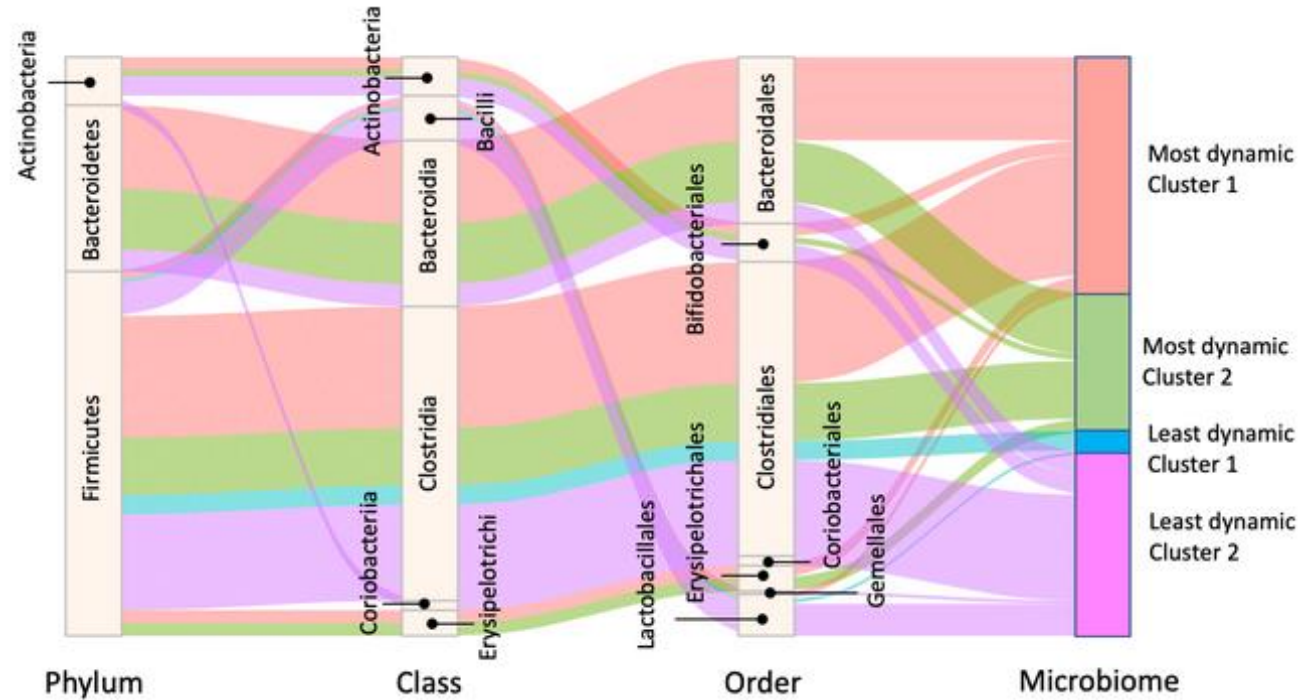
- Input: 1<sup>st</sup> grade neighbors
  - Output: Random walker visited neighbors
- Embedded space: (cosine) distance between conditions or time points



Via local neighborhoods, node distances reflect instability in individual specific networks across conditions (f.i. the smaller  $\theta$ , the higher the stability)

# Our results: Subtypes based on ISN stability

	Time point 1-2		
	Angle $\theta \rightarrow$ similarity/distance		
Indiv.	1	...	L
1			
...			
N			



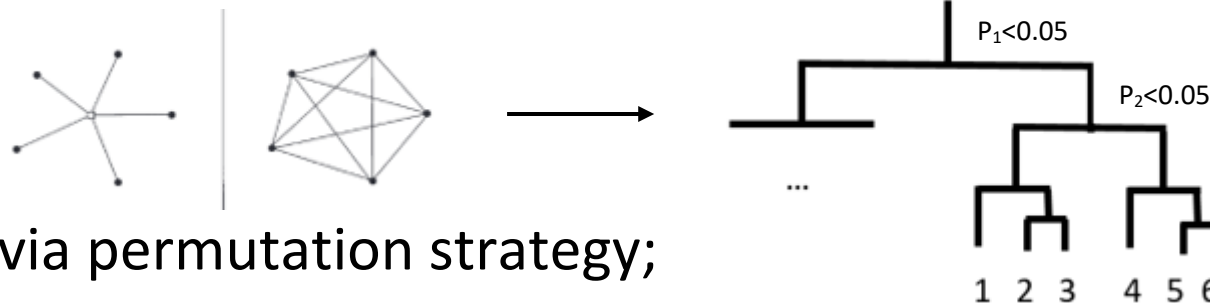
Yousefi, Melograna et al. 202  
(2021 – in preparation)



## Opportunity 2: systems assessment of similarity between individuals

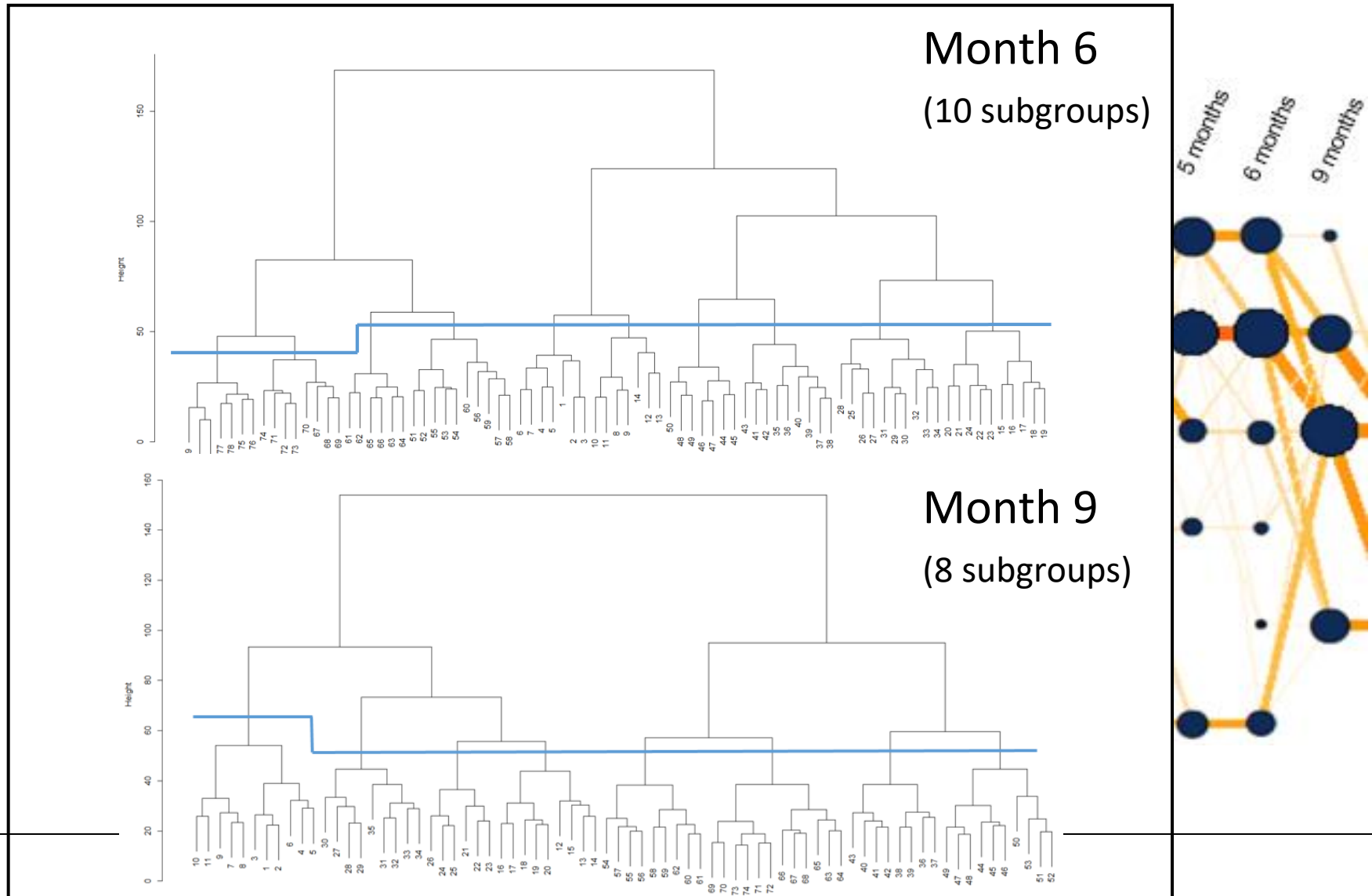
Our solution:

- Creating a similarity matrix between networks:
  - graph kernels: shortest path distance, graph diffusion distance, ...
- Unsupervised hierarchical algorithm to identify latent classes of similar individual-specific networks via edge difference distance
- Recursive strategy to determine an optimal number of clusters via an adapted distance-based ANOVA inspired by ecology;

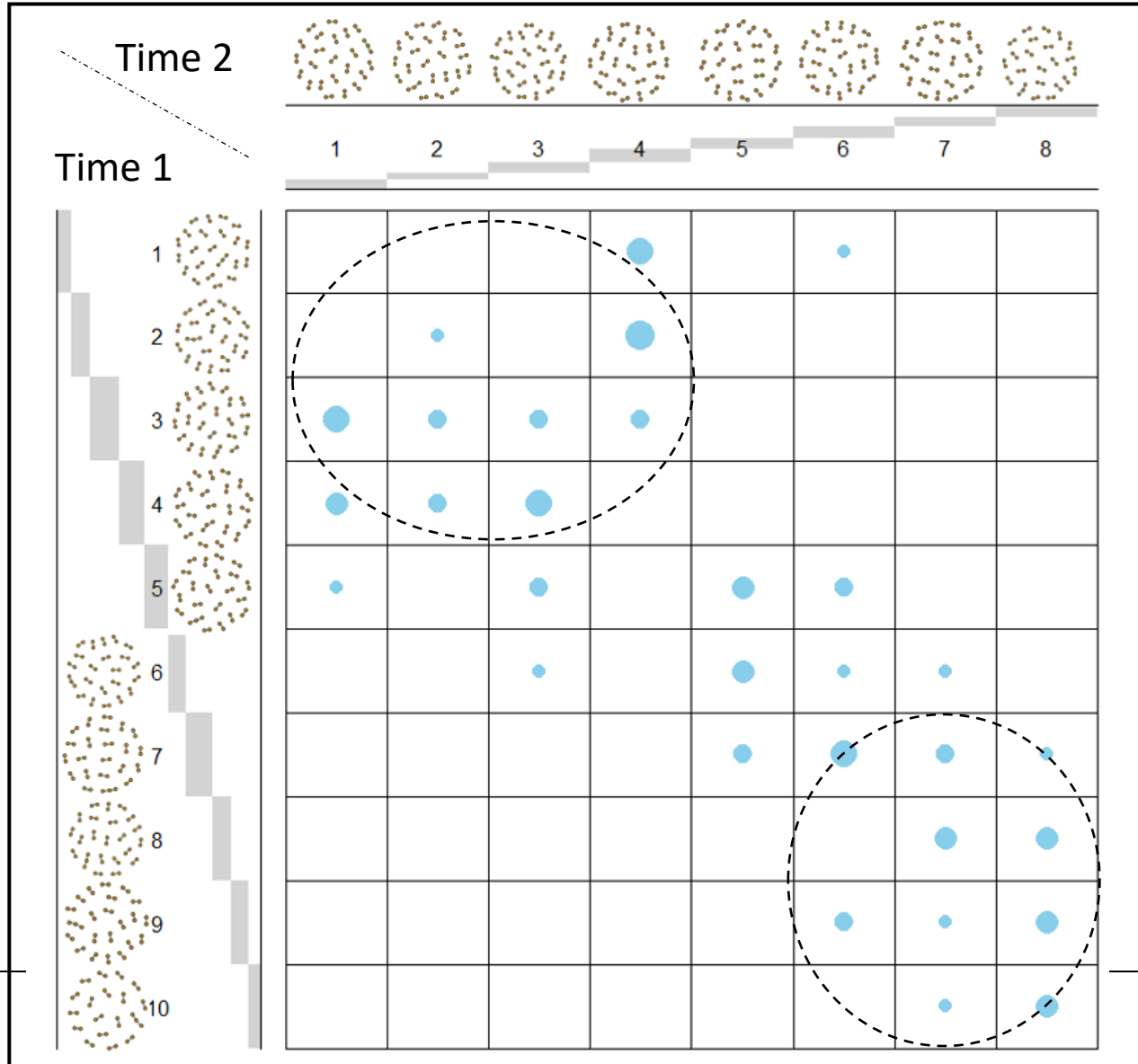


p-values via permutation strategy;  
multiple testing correction

# Our results: subtypes via individual-specific networks

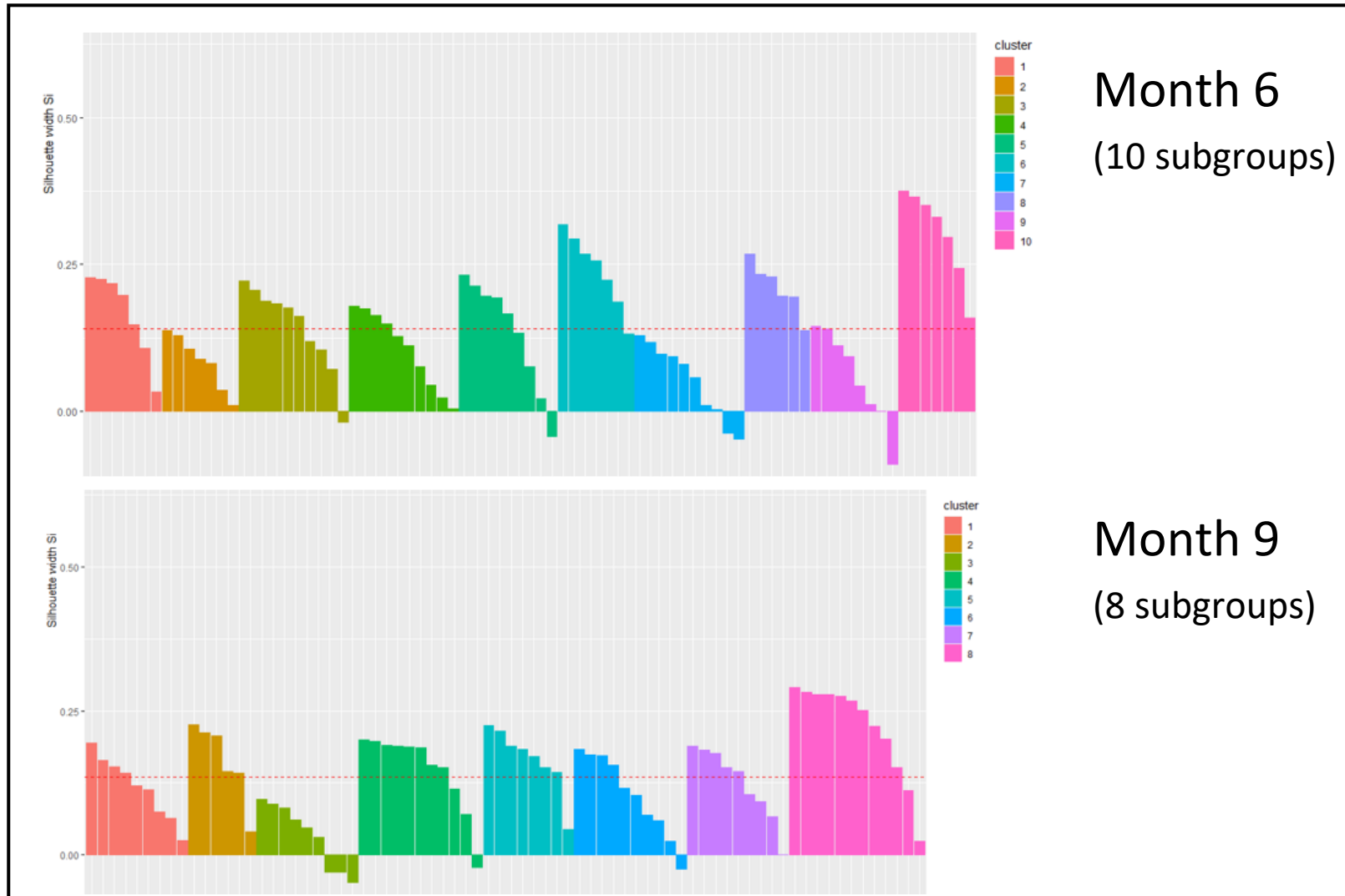


# Our results: subtypes via individual-specific networks & transitions



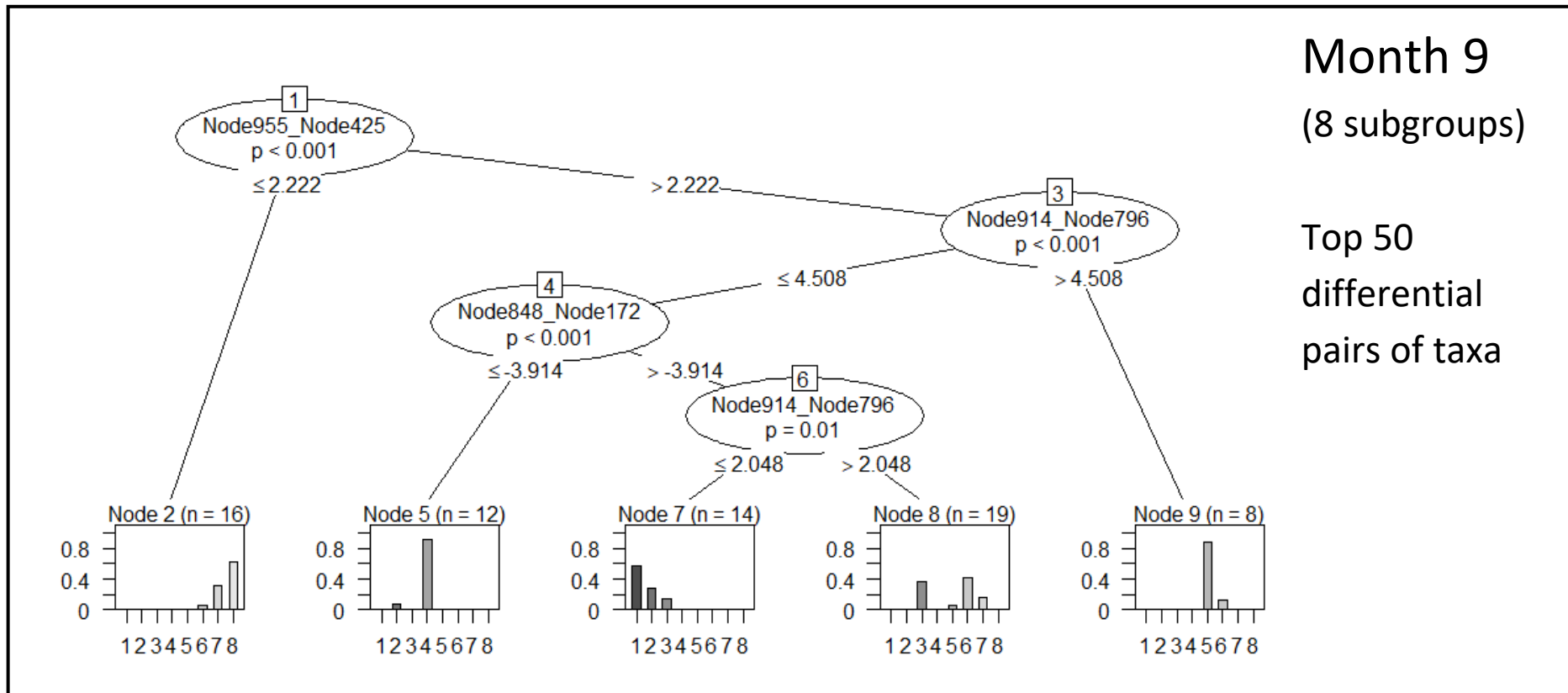
- Jaccard : 0.15  
(p : 0.0009\*)
  - Variation of information (VI) :  
3.22 (p=0.0009\*)
  - compared to  
VI(block 1): 0.2  
VI(block 2): 0.18
- \* 1000 permutations

## Microbiome individual-specific networks: clustering quality



# Microbiome individual-specific networks: digging in deep

## Link with cluster membership

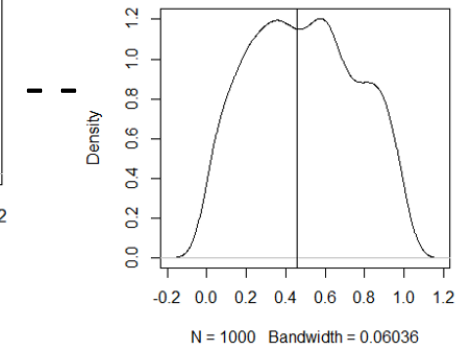
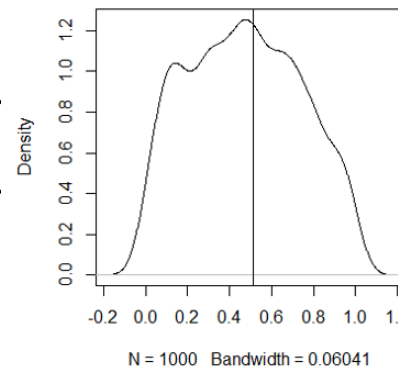


LIMMA paired moderated t-test; multiple testing corrected (FDR -BY; Bonf.); *ctree*

## Microbiome individual-specific networks: digging in deep

Link with clinical data: mode of delivery

- Highly unbalanced data (10 C-section only out of 69)
- ReliefF to prioritize top 10,000 edges
- Random Forest prediction model, with downsampling to ensure similar number of observations in mode of delivery classes
- Acceptable AUC:
  - Month 6: 78% - - - - -
  - Month 9: 70% - - - - -



## Network representation learning (graph classification)

dataset	graphs	classes	[min,max] nodes	[min,max] edges	[min,max] avg-deg
D&D	1178	2	[30, 5748]	[63, 14267]	[7.22, 17.87]
ENZYMES	600	6	[2, 125]	[1, 149]	[2.00, 10.46]
PROTEINS	1113	2	[4, 620]	[5, 1049]	[3.43, 10.14]
NCI109	4127	2	[4, 111]	[3, 119]	[2.50, 5.54]
COLLAB	5000	3	[60, 492]	[60, 40120]	[13.94, 952.02]
RDT-M12K	11929	11	[2, 3782]	[1, 5171]	[4.00, 26.37]

Average accuracy on graph classification.

Baselines	D&D	ENZYMES	PROTEINS	NCI109	COLLAB	RDT-M12K
BaseLine [37]	78.07	61.72	75.16	66.95	55.65	23.58
GraphSAGE [3]	72.36	45.59	70.48	76.50	68.25	42.20
SortPool [8]	78.32	31.29	73.54	70.80	73.76	31.44
gPool [10]	75.01	48.33	71.63	74.52	71.12	OOB
SAGPool [9]	76.94	43.99	72.91	72.51	79.27	43.25
GIN [11]	75.57	48.32	71.65	75.44	<b>79.48</b>	47.22
DiffPool [6]	80.01	62.17	75.96	80.10	71.78	47.05
MxGNN (Ours)	<b>80.47</b>	<b>68.22</b>	<b>76.30</b>	<b>81.89</b>	74.43	<b>47.52</b>

(Liang et al. 2021)

doi: 10.1109/ACCESS.2021.3070690)

## Network representation learning (link prediction)

- Network embedding is a method for embedding a network to a lower dimensional space by converting nodes in the network to vectors in the embedding space
- One of the representative methods for multi-layered graphs is Multi-Task Network Embedding (MTNE): works well when network structures in the different layers are similar
- A solution == Multiplex network Embedding via Learning Layer vectors (MELL): embeds both undirected and directed networks

(Matsuno and Murata 2018)

doi : [10.1145/3184558.3191565](https://doi.org/10.1145/3184558.3191565)



## Microbiome individual-specific networks: digging in deep

- Edges ~ clinical data: acceptable discrimination performance
  - Discrimination versus calibration
  - Edge selection
- ISNs ~ clinical data: no significance so far
  - Clinical data at our disposal
  - Graph kernel choice
    - (Borgwardt et al 2020)
    - [dx.doi.org/10.1561/22000000076](https://doi.org/10.1561/22000000076)
  - Network representation learning (for classification of graphs)

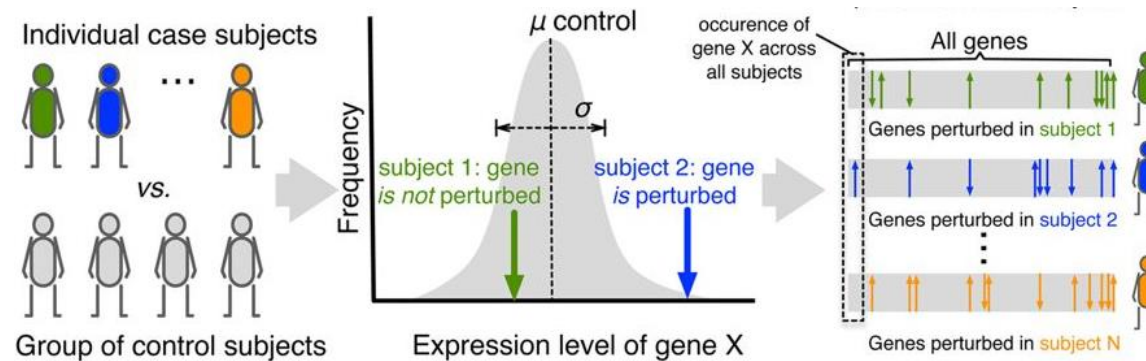
Common denominator: sparsification

# Transcriptome Systems Analysis for Precision Medicine

*individual-specific networks*

## Transcriptome systems analytics

“Gene expression data are routinely used to identify genes that on average exhibit different expression levels between a case and a control group. Yet, very few of such differentially expressed genes are detectably perturbed in individual patients. ... personalized perturbation profiles for individual subjects ...”



(Menche et al. 2017)

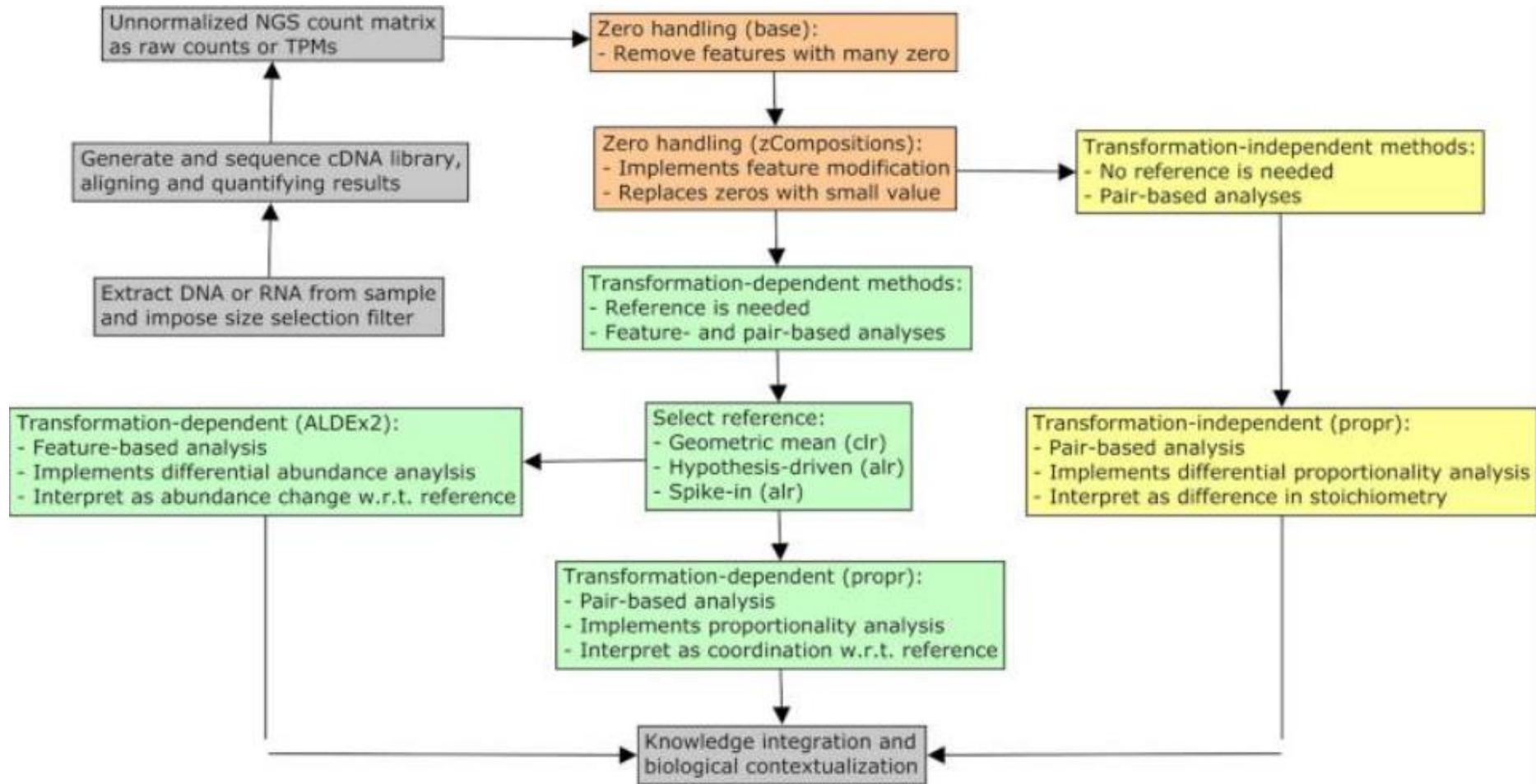
doi:10.1038/s41540-017-0009-0

(Quinn et al. 2019)

doi : 10.1093/gigascience/giz107

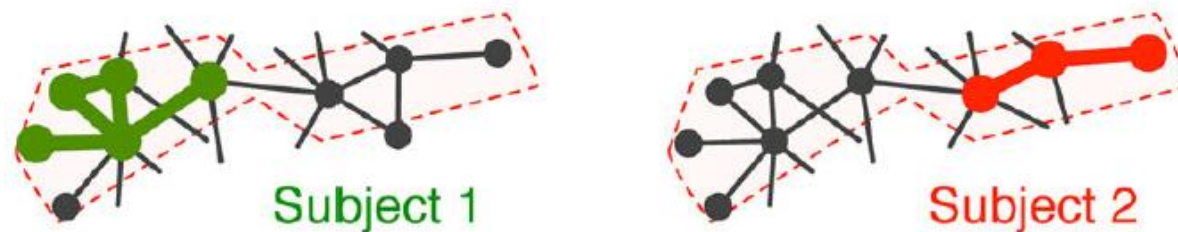
# Interludium

## Nature of the data



## Challenge: *How does molecular level heterogeneity between case subjects translate into precision medicine practices?*

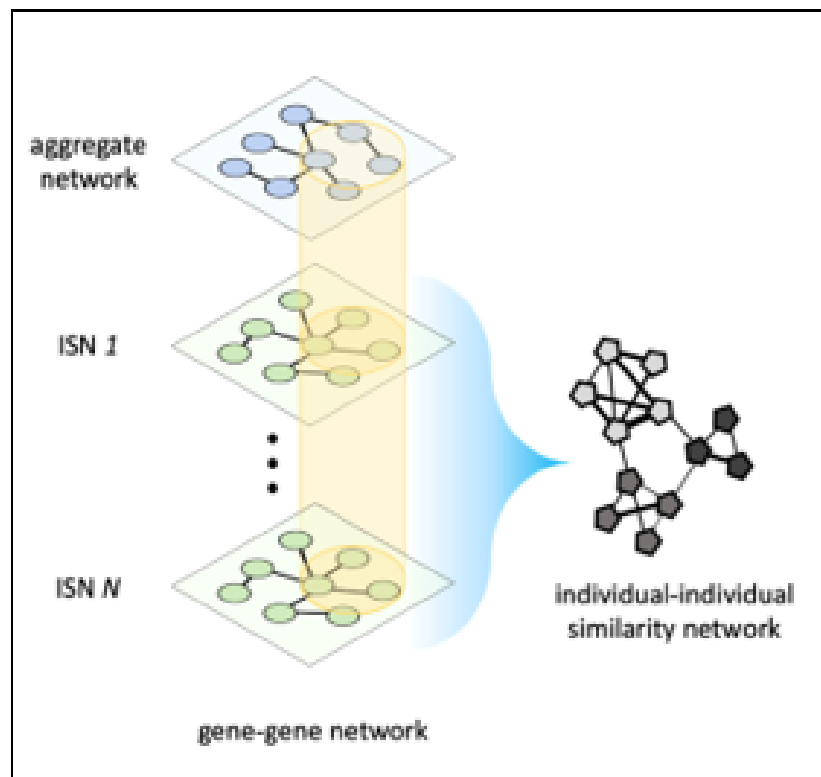
- Relatively low overlap between personalized perturbation profiles of case subjects
- Integrated personalized perturbation profiles with fixed edges (f.i. generic protein-protein interaction network)
- The same pathway associated with a specific function may be disrupted by perturbations at different locations in different subjects:



(adapted from Menche et al. 2017)

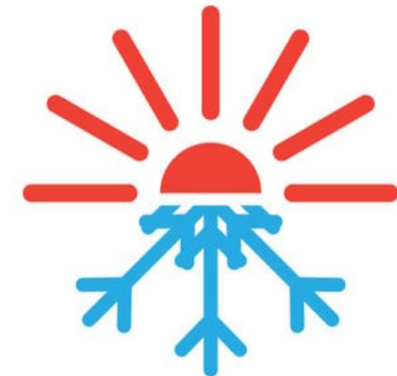
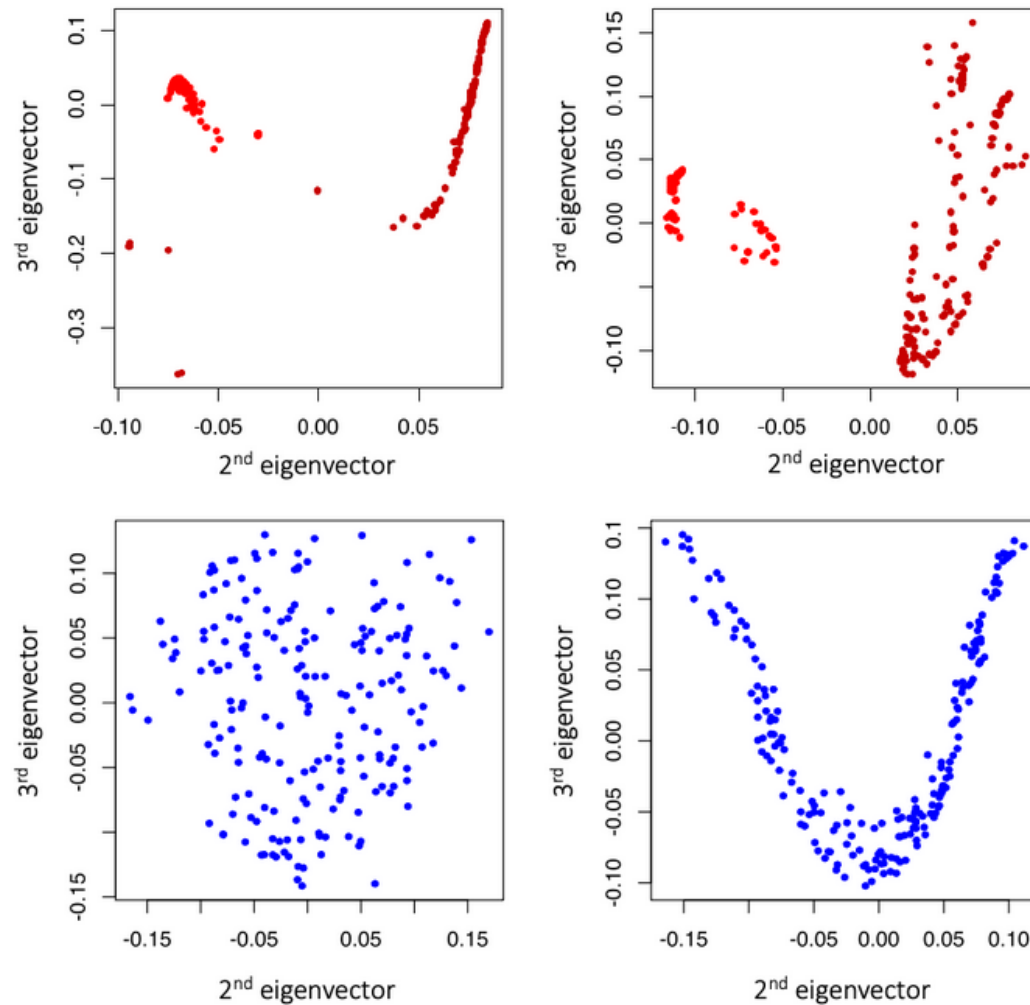
## Opportunity 3: HotZones – prioritising multi-sample based network modules with maximal between-individual heterogeneity

Our solution:



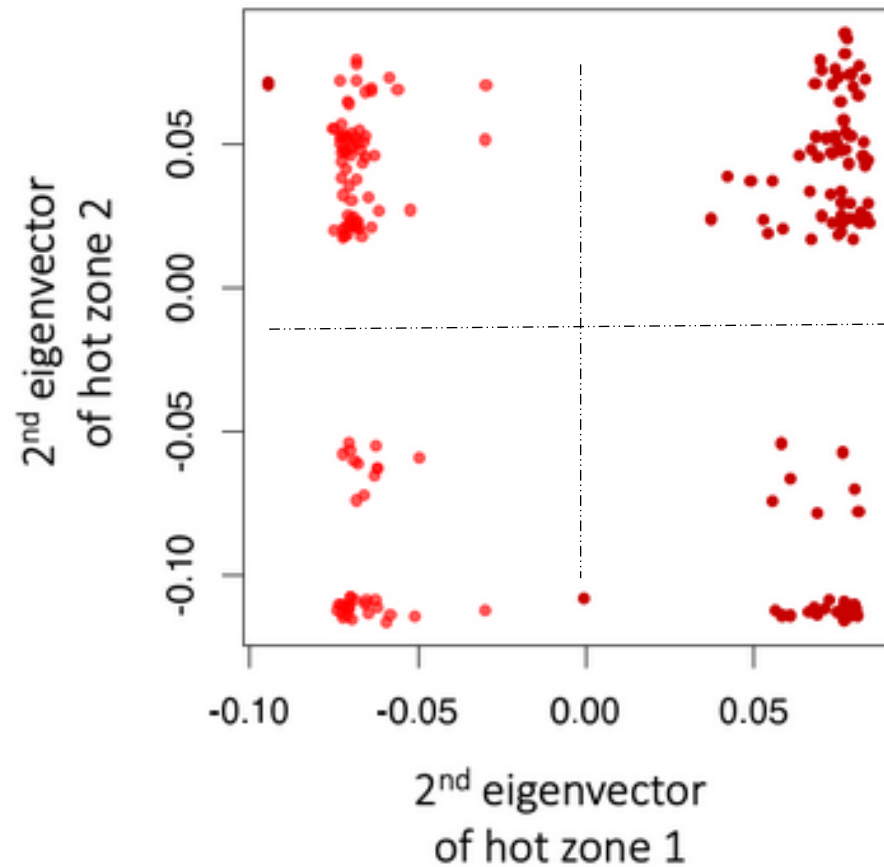
- Module derived on aggregate network (all samples); yellow
- ISNs restricted to the identified module
- Similarity metric on restricted-ISNs
- Statistical test assessing presence of clusters (i.e. heterogeneity); 3 shown

## Results: Gene modules with (in)variable personalized profiles



Yousefi, Melograna, et al.  
(2021 – in preparation)

## Results: Each hot zone clusters individuals from a different angle



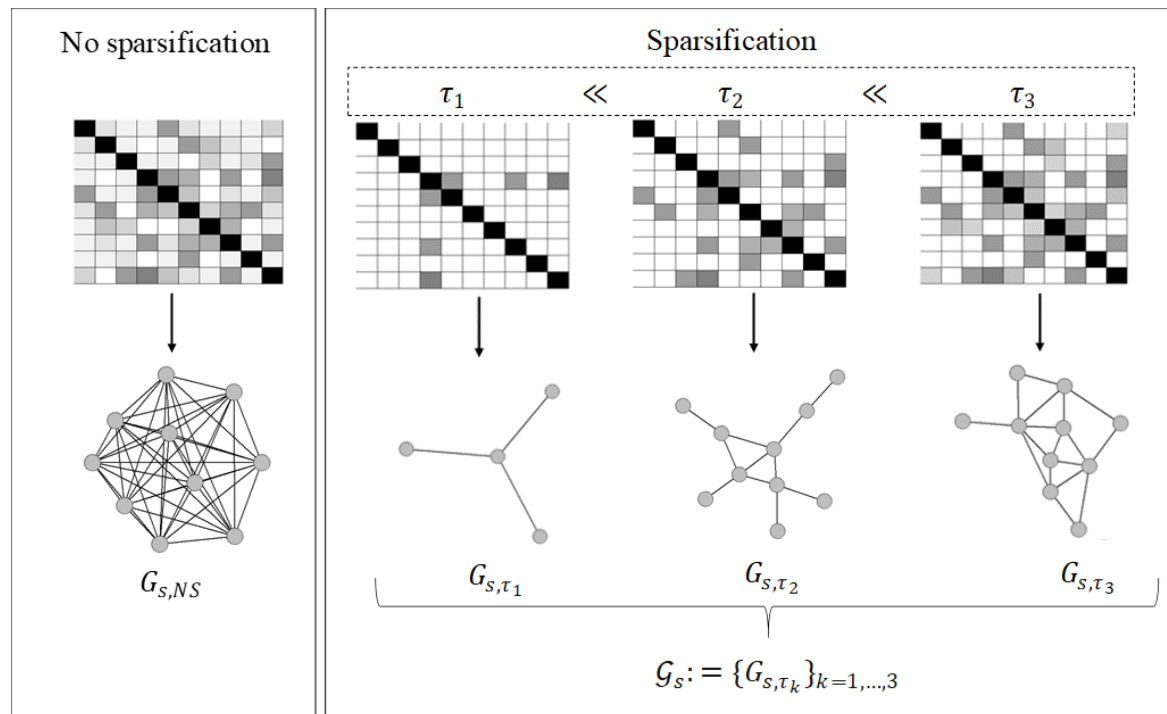
Yousefi, Melograna, et al.  
(2021 – in preparation)



## Interludium

### Sparsification of networks via edge (instead of node) selection

- Graph filtration



(Gregorich et al. 2021 – under review; adapted)

- ✓ Given a graph  $G=(V,E)$  with an edge weight function  $w: E \rightarrow R$ , a filtration  $F_G$  is a series of monotone-increasing subgraphs that defines a graph decomposition
- ✓ Naively, the edge weights themselves determine the edge weight function

(Edelsbrunner et al. 2002)

doi: 10.1007/s00454-002-2885-2

## Interludium

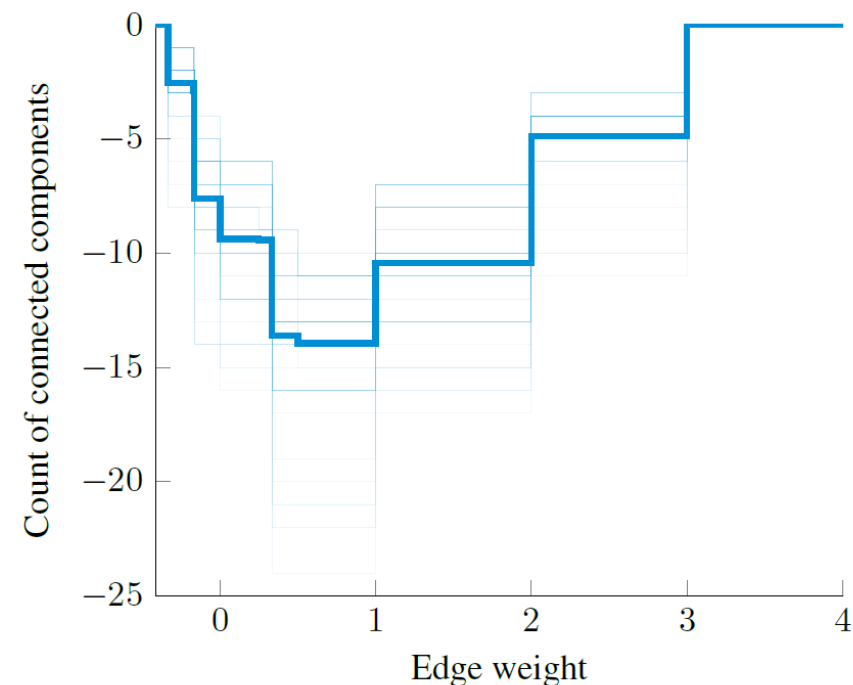
### Sparsification of networks via edge (instead of node) selection

- Filtration curves to complement commonly used graph comparison methods

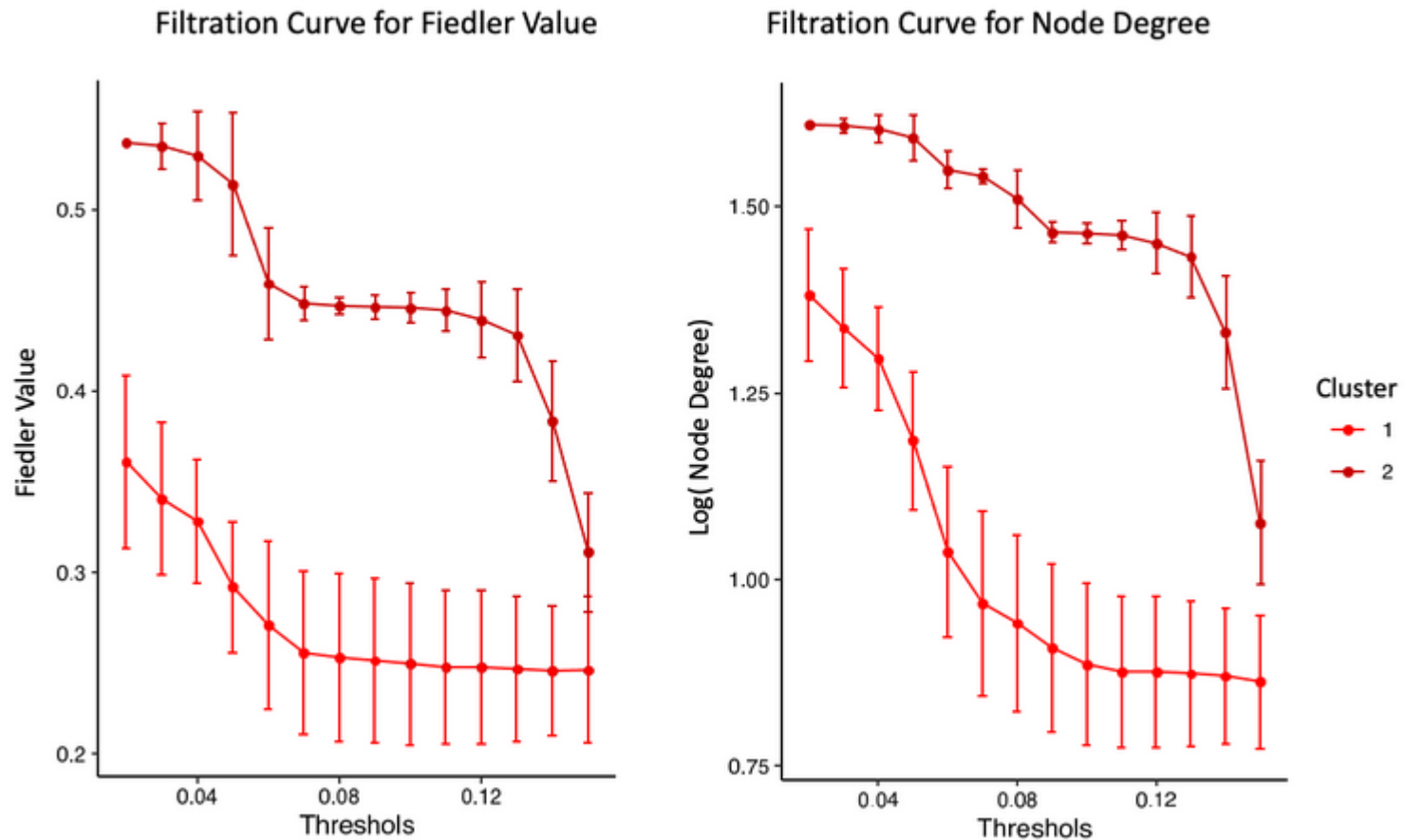
- ✓ Graph descriptor function  $f: G \rightarrow R^d$  that evaluates certain attributes of a graph and embeds them into a  $d$ -dimensional real-valued space.
- ✓ Example:  $f(G) = \#$  connected components in  $G$

(O'Bray et al. 2021)

doi: 10.1145/3447548.3467442

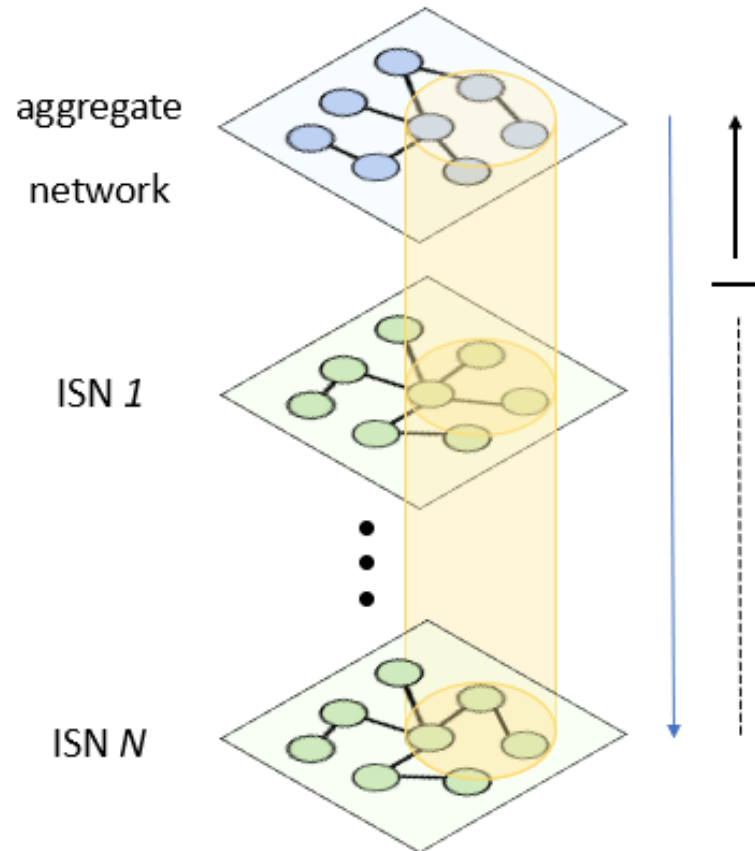


## From filtration curves to distance-based ANOVA (as before)



Yousefi & Melograna & Duroux, et al. (2021 – in preparation)

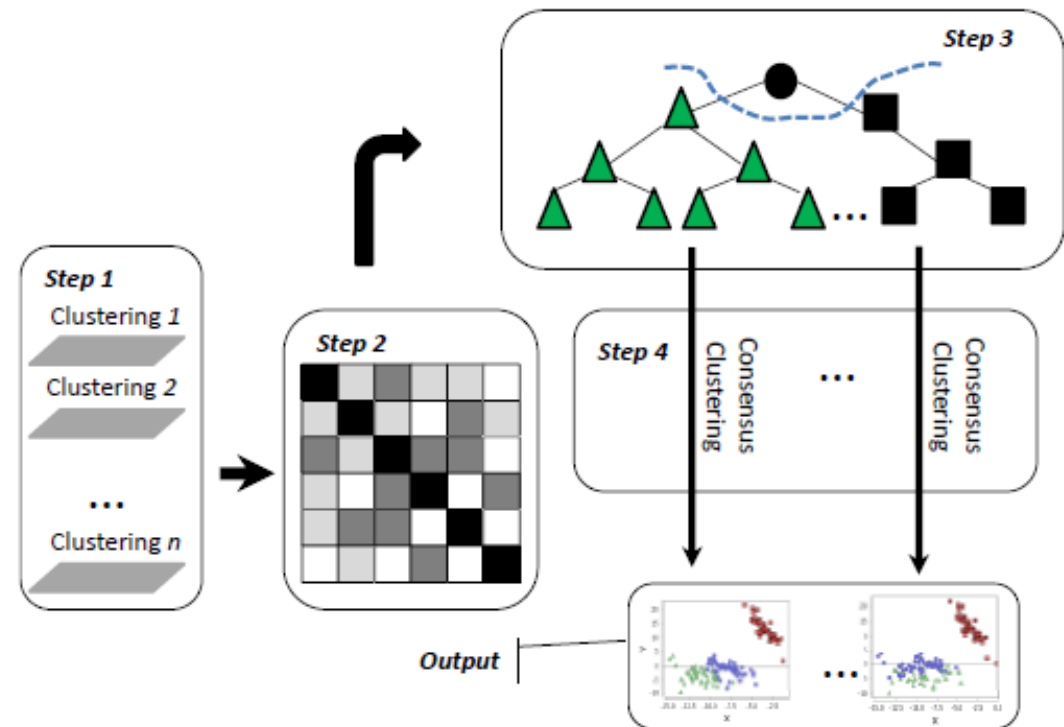
## From aggregate network modules to free module searches



- **Consensus clustering:** find a single clustering from different input clusterings with improved cluster separation
- Input clusterings:
  - Same data, different algorithms
  - Different subsets of the data
  - Different feature spaces

## Combine meta clustering and consensus clustering

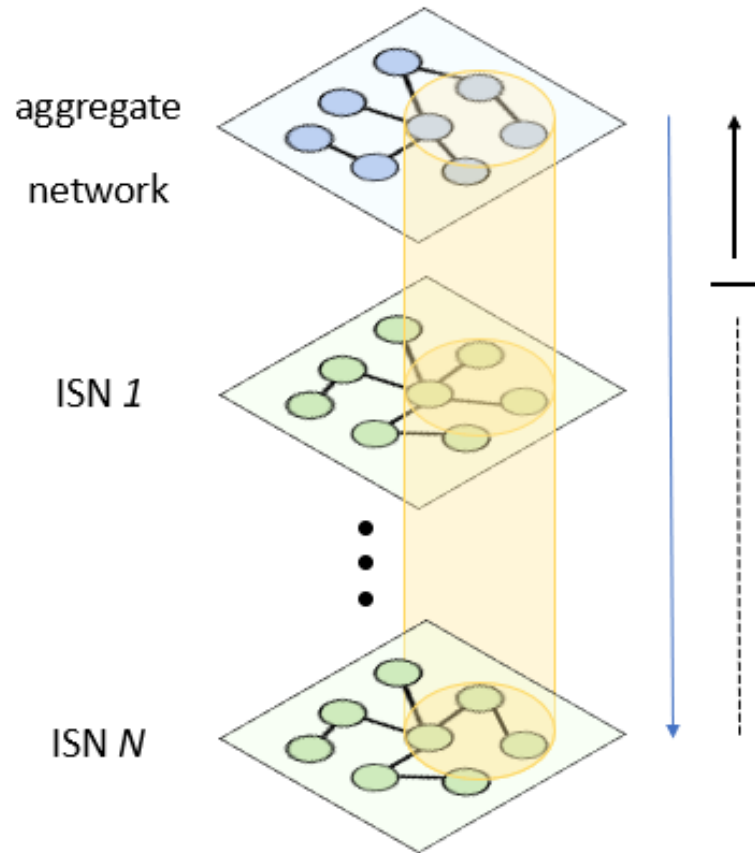
- Meta clustering: based on finding a variety of “reasonable” clusterings and grouping clusterings into meta clusters (Caruana et al. 2006)
- **Multiple Consensus Clustering (MCC)**



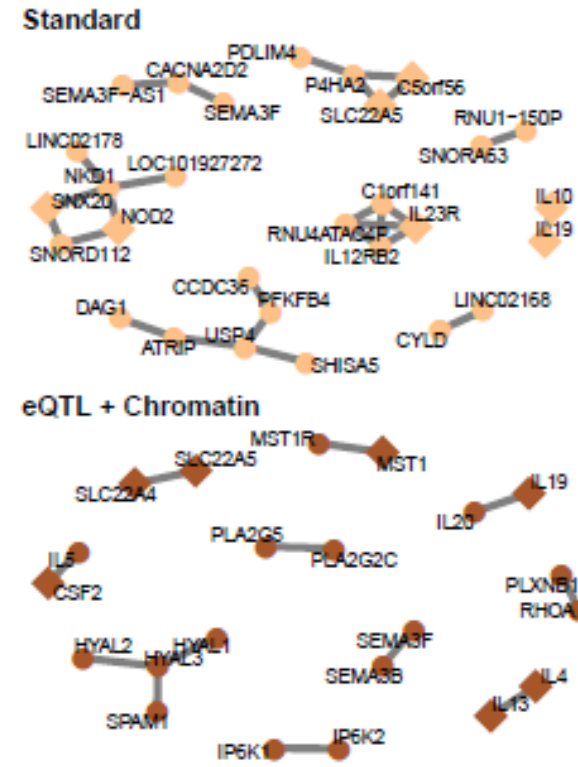
(Zhang and Li 2011)

doi: 10.1137/1.9781611972818.79

# From network aggregation to inferential reproducibility



Melograna & Yousefi et al.  
(2021 – manuscript in preparation)



Duroux & Climente et al.  
(2021 – under review)

# Take-home messages

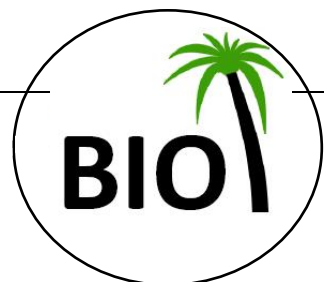
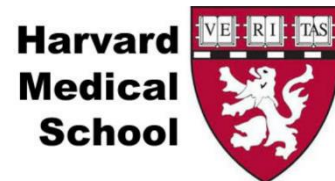
## Take-home messages

- Integration & interaction need to go hand in hand
- Precision medicine benefits from longitudinal follow-up; new avenues with machine learning should not be left unwalked; novel developments are needed
- Individual-specific networks are promising in the context of precision medicine and individual heterogeneity assessment and may complement standard analyses
- Determining causality is often a challenge as is moving from undirected to directed individual-specific networks



# Acknowledgements

## Collaborators include



[bio3.giga.ulg.ac.be](http://bio3.giga.ulg.ac.be)

## Funding includes



Presented work received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements N° 813533 and 860895.



Contact: [kristel.vansteen@uliege.be](mailto:kristel.vansteen@uliege.be)