INTEGROMICS

Boiling the ocean?

Kristel Van Steen, PhD[2] (*)

kristel.vansteen@ulg.ac.be

(*) Systems and Modeling Unit, Montefiore Institute, University of Liège, Belgium

(*) Bioinformatics and Modeling, GIGA-R, University of Liège, Belgium

OUTLINE

1. What is INTEGROMICS ?
2. What are the corner stones of an analysis pipeline ?
3. Why doing INTEGROMICS ?
4. Which analytic routes lead to INTEGROMICS ?
5. What are “obvious” methodological challenges ?
6. What are “non-obvious” methodological challenges ?
7. Will dimensionality reduction reduce too much ?
8. Is heterogeneity a nuisance or a relevant piece of info ?
9. Can we learn from cross-disciplinary marriages ?
10. Concluding remarks

Boiling the Ocean

– Ten expressions related to « boiling the ocean » :

exaggerate - excessive - impossible - needing more actionable steps - overkill - overreacting - pie in the sky - overdoing - plowing water - overly ambitious

– Looking at integromics *without* boiling the ocean **… in 10 STEPS**

STEP 1: What is INTEGROMICS?

- INTEGROMICS = integration + omics

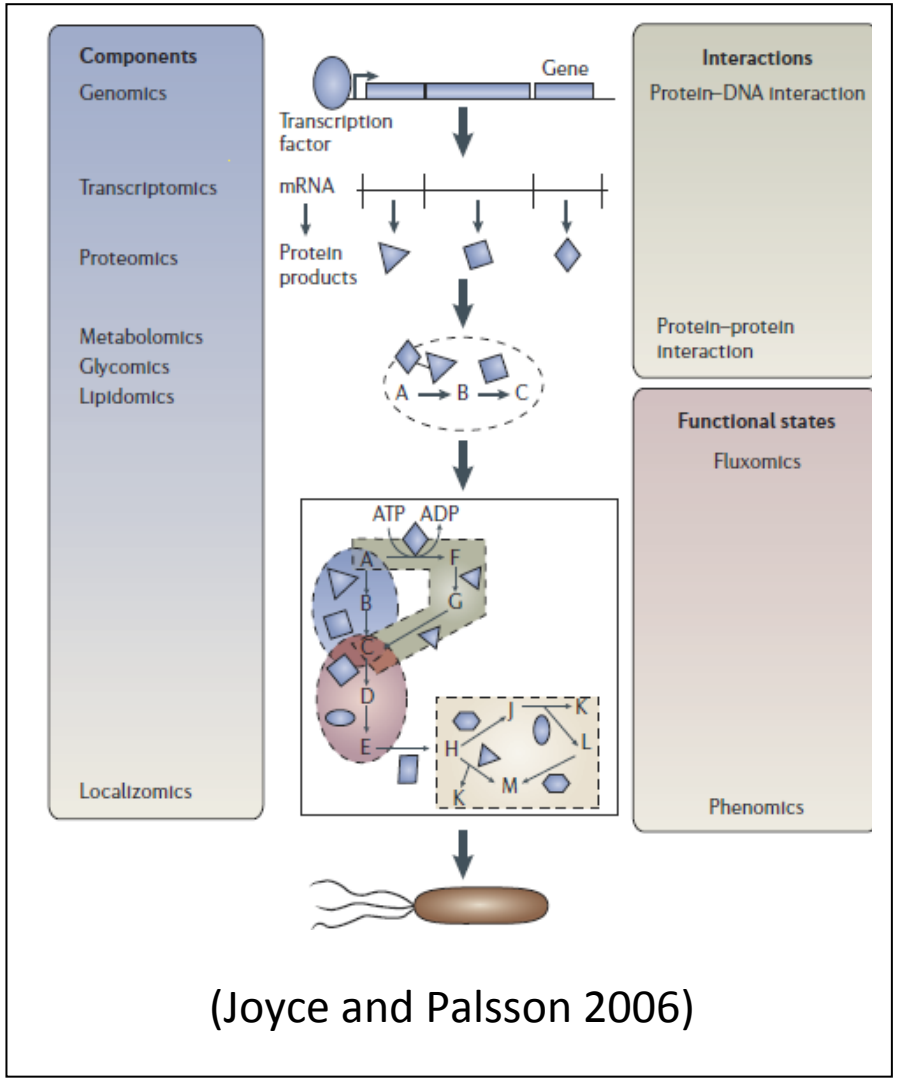
Integration

- Although some data integration efforts will rely on data fusion processes, data **fusion** and data integration are not equivalent.
  - Data fusion refers to fusing records on the same entity into a single file, and involves putting measures in place to detect and remove erroneous or conflicting data (Wang et al., 2014).
  - In this sense, data fusion is linked to data **concatenation**; mapping several objects into a single object (Oxley & Thorsen, 2004)
- **Integration** is the process of connecting systems (which may have fusion in them) into a larger system (Oxley & Thorsen, 2004)

… + omics

- Omics data is a generic term that describes genome-scale data sets that emerge from high-throughput technologies
  (e.g., whole genome DNA sequencing data [genomics], microarray-based genome-wide expression profiles [transcriptomics]

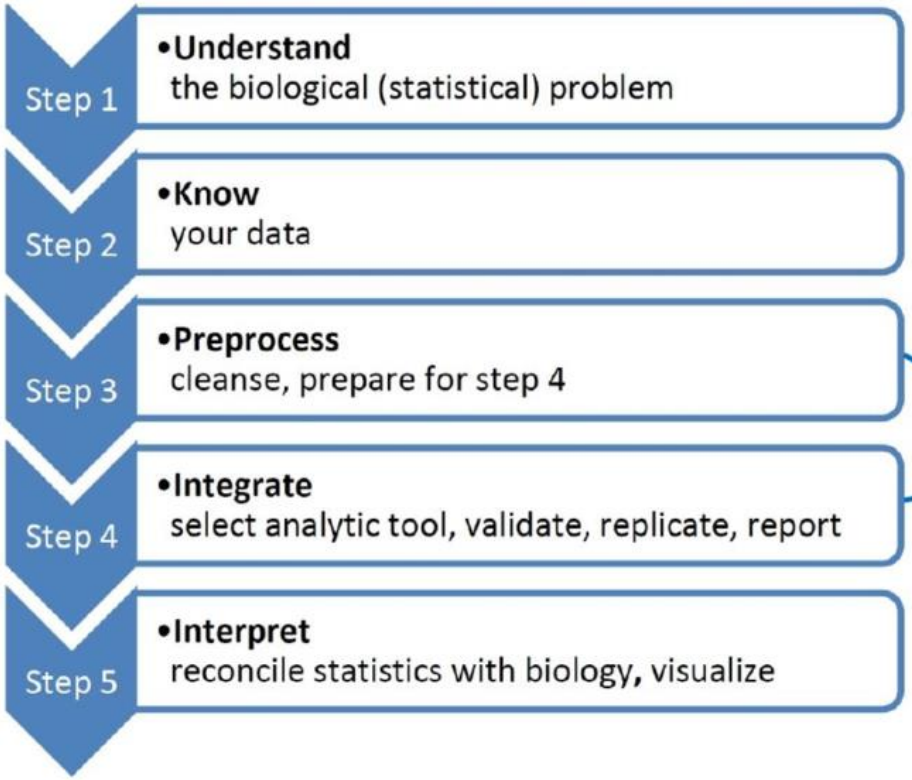
- These data describe virtually all biomolecules in a cell (e.g., proteins, metabolites)



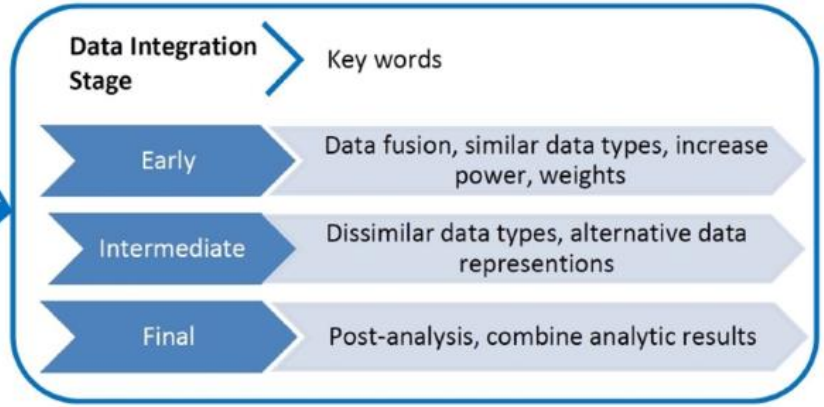
(Joyce and Palsson 2006)

STEP 2: What are the corner stones?

- The building blocks of an data integrative analysis pipeline



(Hamid et al. 2009)



Systems information by integration (Joyce and Palsson 2006)

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[71,103] • Biomarkers[125]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[20]	• Enzyme annotation[109]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[89,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[89]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

Step 1

- Formulating the biological (statistical) problem

Step 2

- Identifying the (characteristics of the) data types

- Data characterization (in my opinion) refers to finding first evidences for
  - intrinsic properties (e.g., small sample sizes, standard formats)
  - layers of information; hierarchies; dimensionality
  - noise patterns (related to technology, platform, the lab; systematic and random errors)
- EDA / Weighting: quality + information

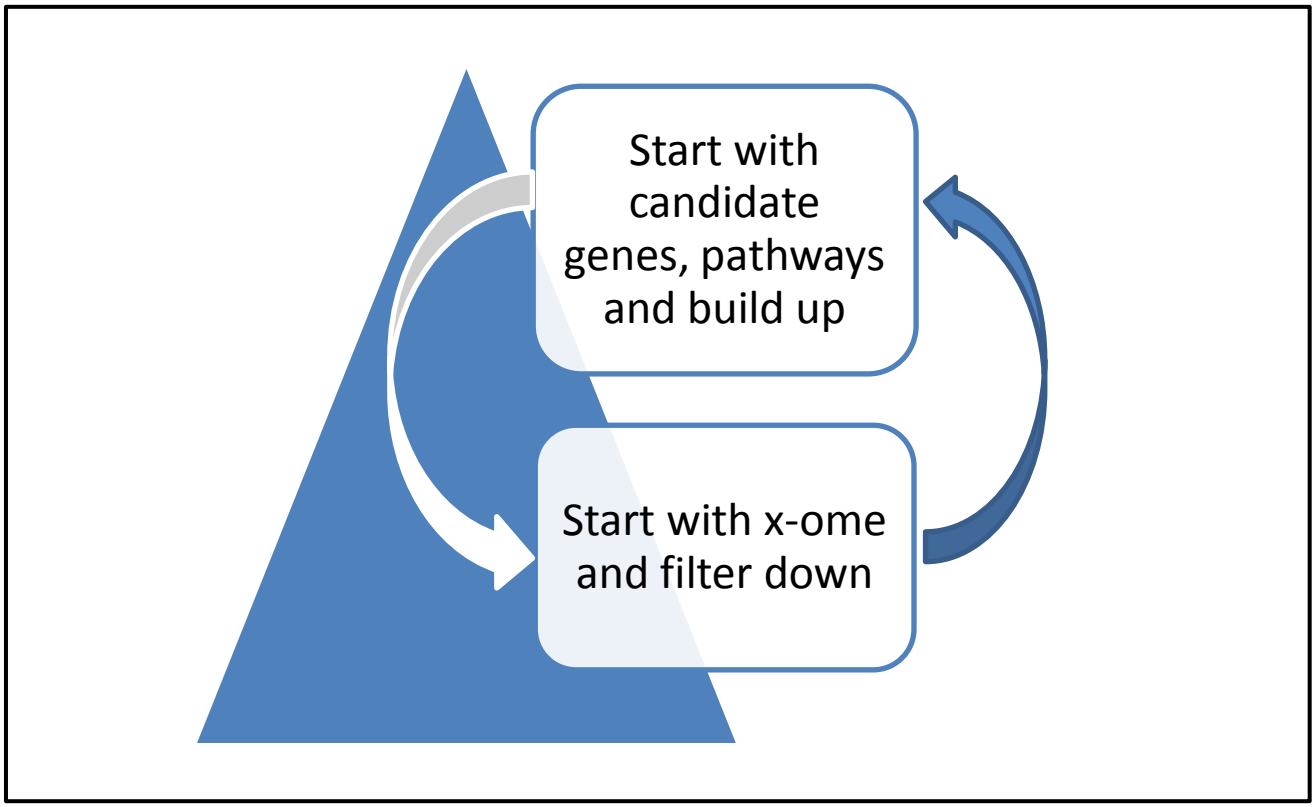
Step 3

- Data preprocessing

- Approaches for preprocessing vary depending on the type and nature of data:
  - e.g., arrays: background correction, normalization, quality assessment, which may differ from one platform to another
- Data (pre)processing can be done **at any step of the data integration** process:
  - e.g., at the **initial stage**
  - e.g., **prior to statistical analysis** (related to model assumptions)

Step 4

- Integration analytics



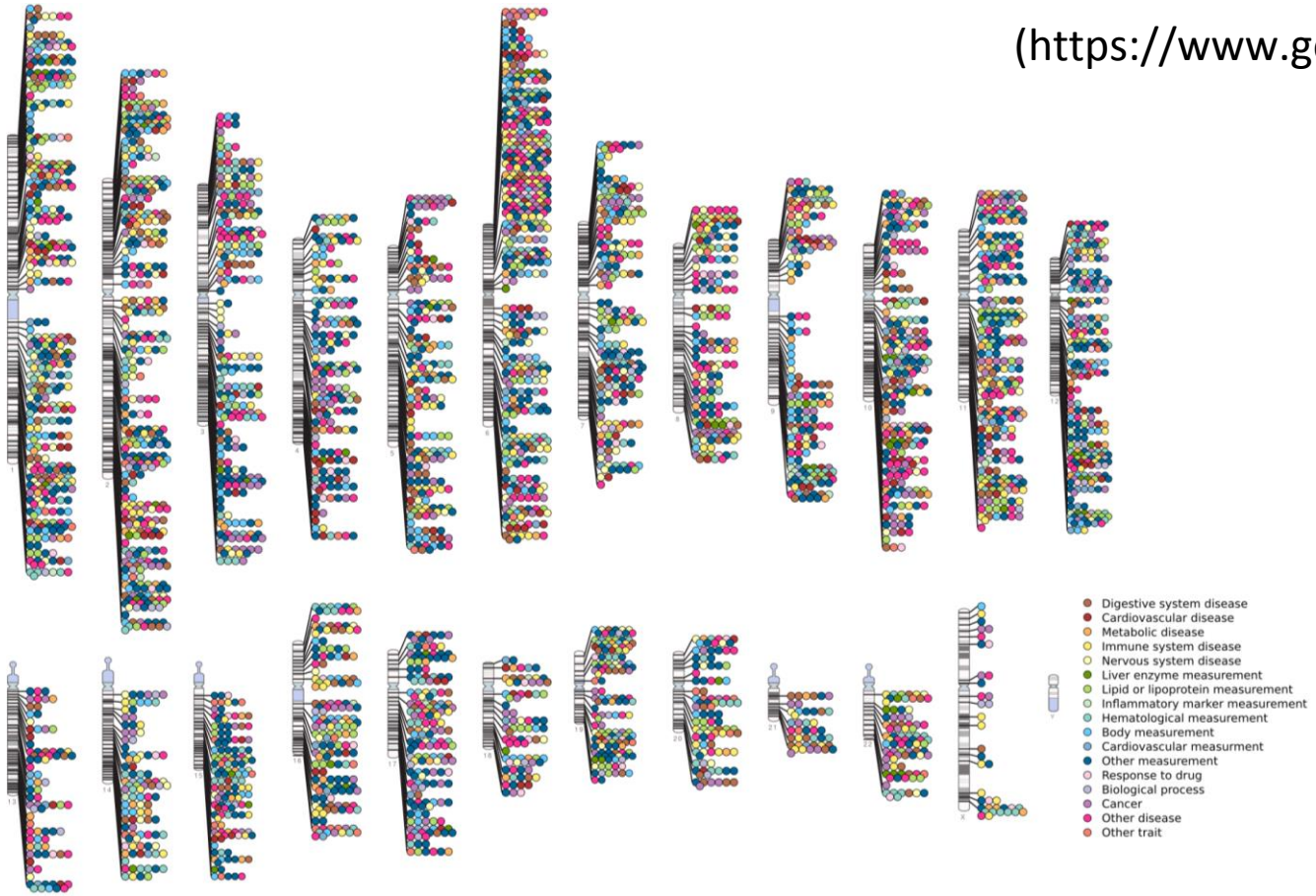
Step 5

- Interpretation (after integrative analytics)

- Is about “understanding” the problem that was initially posed and providing a “functional explanation”
- (Experimental) **validation** helps in the “understanding”, but becomes cumbersome in integromics settings/ simulations?
- What about **replication**?
- Challenges and opportunities for **visual analytics**
- Be aware of pitfalls when post-linking to biological knowledge data bases with black-box tools

STEP 3: Why doing INTEGROMICS?

From baby steps to leapfrog



(https://www.genome.gov/GWAStudies)

Published GWAs

through 12/2013

at p≤5X10-8

for 17 trait categories

Université de Liège

GWAs inability to explain heritability

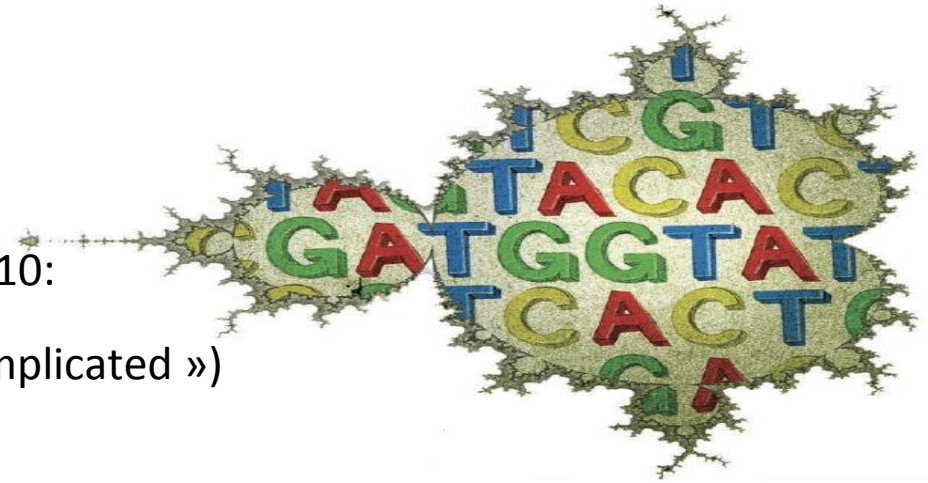
Explanation	Rationale	Comments
Overestimated heritability estimates	These estimates are typically performed in the absence of gene-gene or gene-environment interactions (Young et al. 2014)	Limiting pathway modeling suggests that epistasis could account for missing heritability in complex diseases (Zuk et al. 2012)
Common genetic variants	More common variants are likely to be found in GWAs with larger sample sizes (drawback: more is less?)	Effect sizes of known GWAs loci may be underestimated since functional variants have often not yet been found
Rare genetic variants	Resequencing studies (e.g., WES) could identify rare genetic determinants of large effect size (Zuk et al. 2014)	Limited evidence for rare variants of major effect in complex diseases accounting for large amount of genetic variation – most rare variants analysis methods currently suffer from increased type I errors (Derkach et al. 2014)

Phenotypic and genetic heterogeneity	Most complex diseases are like syndromes with multiple potentially overlapping disease subtypes	Improvements in phenotyping of complex diseases will be required to understand genetic architecture.
Interaction	Gene-gene and gene-environment interactions are likely to be important for complex diseases (Moore et al. 2005)	Limited evidence for statistical interactions in complex diseases; network-based approaches may be helpful (Hu et al. 2011)

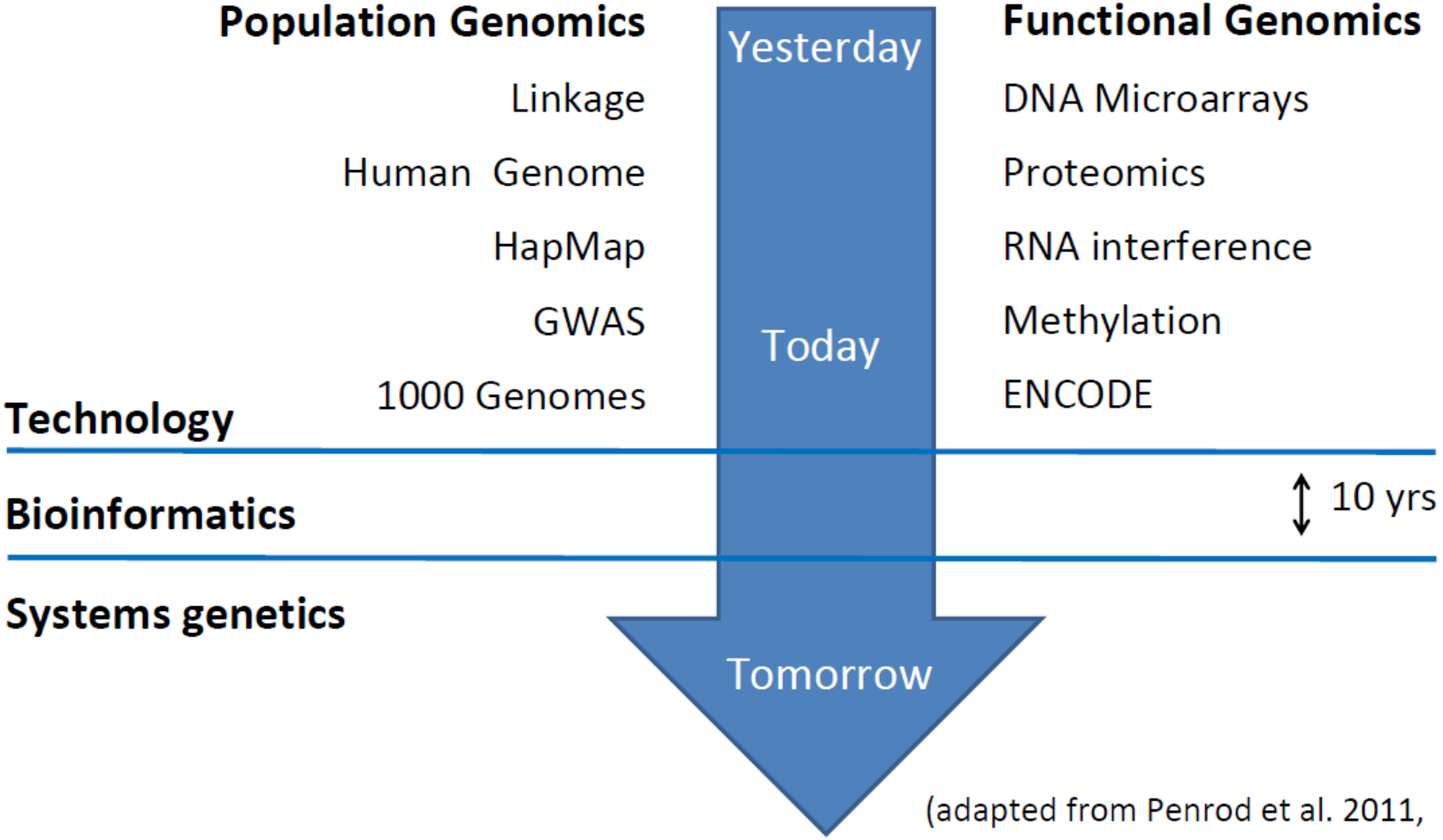
(adapted from Silverman et al. 2012)

(Hayden 2010:

« Life is Complicated »)



A partial picture …



(adapted from Penrod et al. 2011, Moore 2012)

Modeling systems genetics … (http://eupancreas.com)

WG2: “integration of omics data”
(work group leader: K Van Steen)

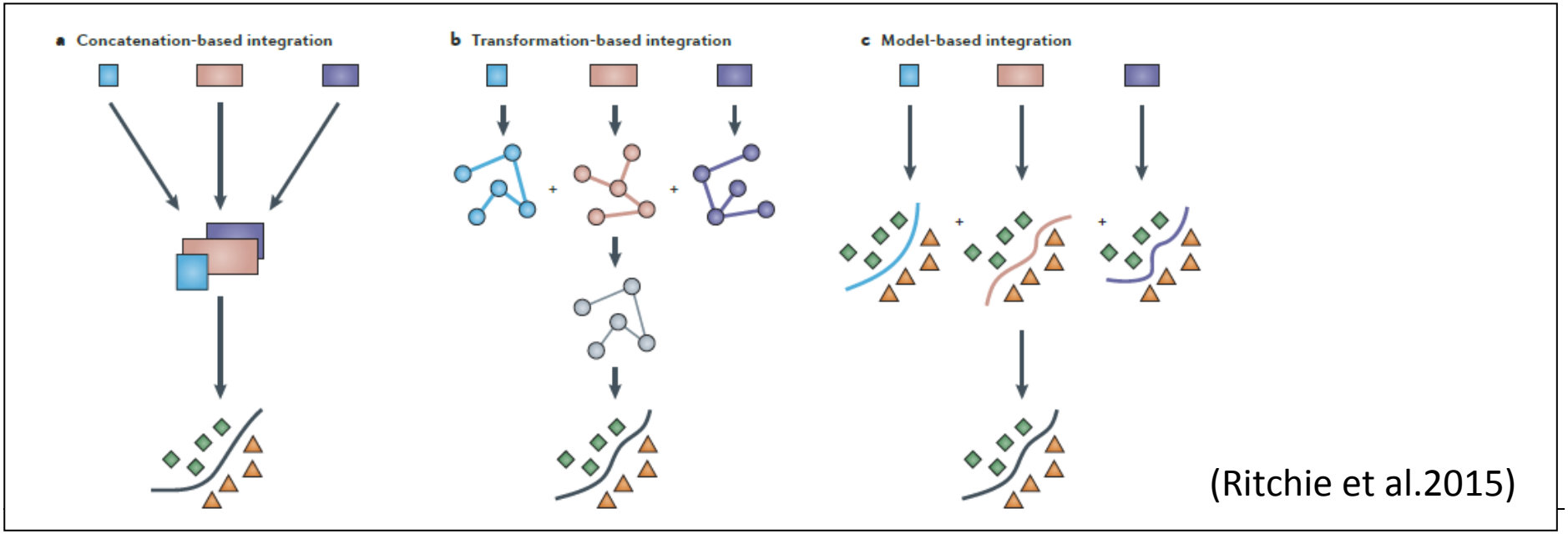
Modeling additional complexities – the GWAI story

" …just **adding one extra level of complexity** to a well-investigated data analysis type, such as when moving from genome-wide main effects SNP-based analyses to genome-wide interaction SNP-SNP analyses, offers **a sobering lesson** in what a lack of data (problem) acknowledgement can provoke. "

(Guserava et al., Van Steen 2015 – submitted)

STEP 4: Which routes lead to INTEGROMICS?

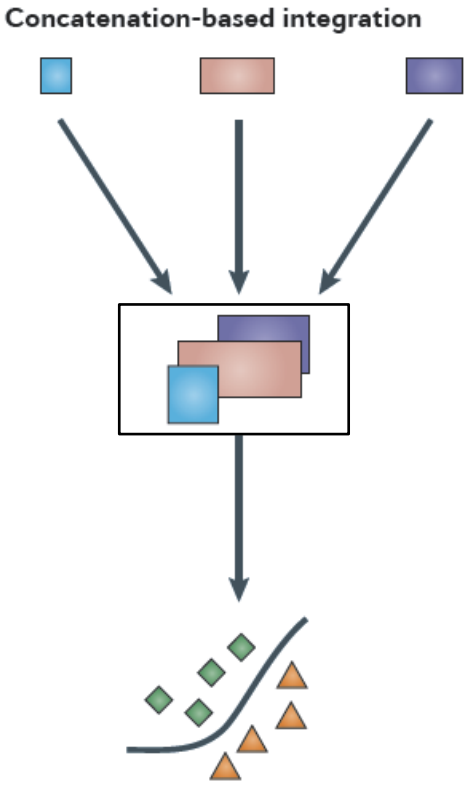
- In correspondence with the description of the Hamid "stages", Ritchie et al. (2015) refer to concatenation-based (left), transformation-based (middle) or model-based integrative (right) approaches
- The Hamid view and the Ritchie view are essentially two faces of the same coin

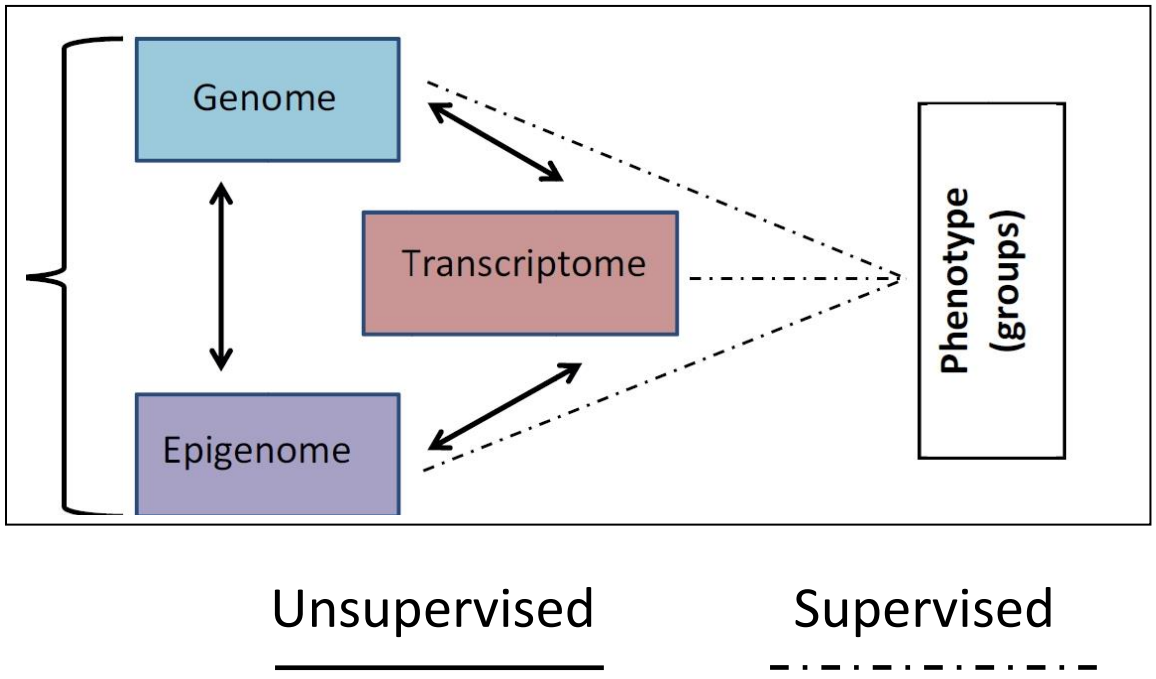


(Ritchie et al.2015)

STEP 5: What are "obvious" methodological challenges?

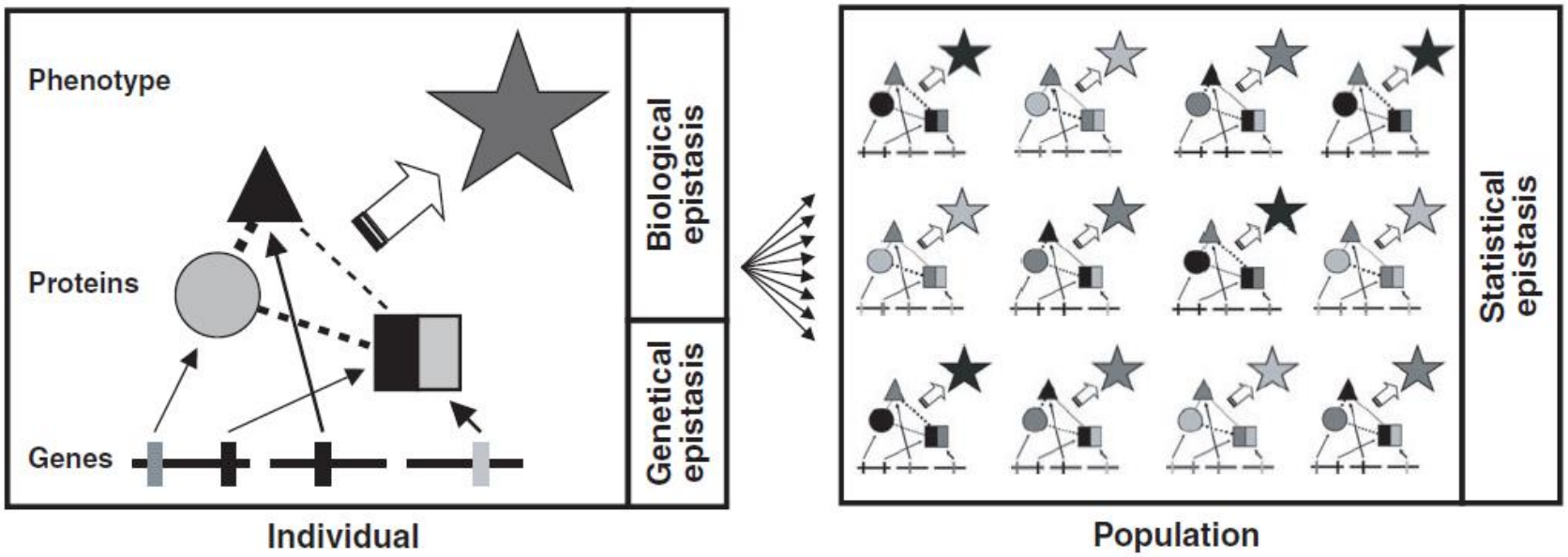
- It is obvious that only by concatenating, one is able to account for "relationships" between different omics data sources





Omics data are related

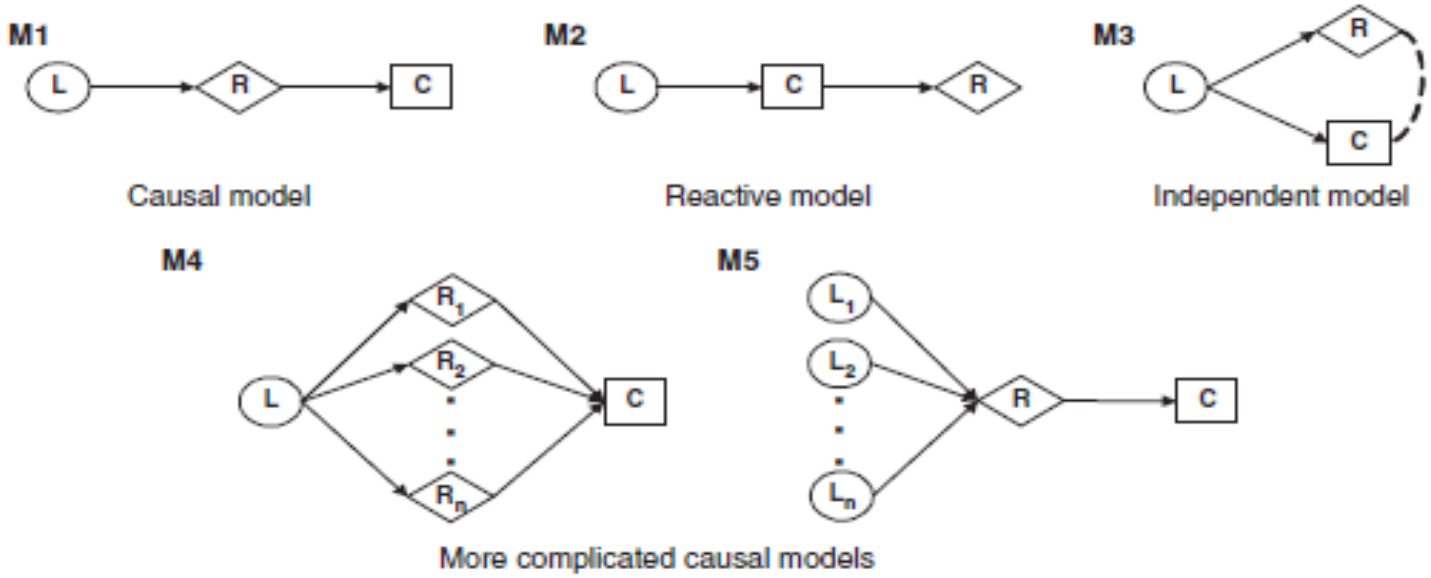
- Two or more DNA variations may “interact” either directly to change transcription or translation levels, or indirectly by way of their protein product (to alter disease risk separate from their independent effects)



(Moore 2005)

Omics data are related

- The road from SNPs to phenotype is complex; **multiple roads** may lead to the same phenotype



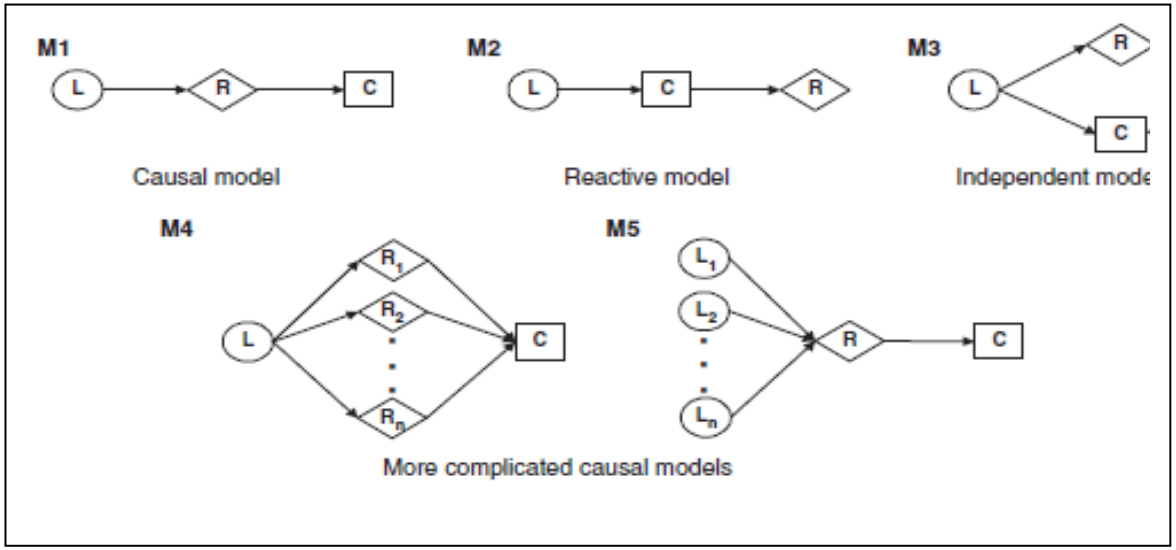
Graphical models for relationships between QTLs, RNA levels and complex traits, assuming gene expression (R) and complex trait (C) are under the control of a common QTL (L) - Schadt et al. (2005)

Omics data are related

- Extra complexities can be added, as features that belong to the same omics data source may jointly be involved in **non-independent or non-linear** relationships

  - L_1 x L_2
  - R_1 x R_2
  - P_1 x P_2
  - E_1 x E_2

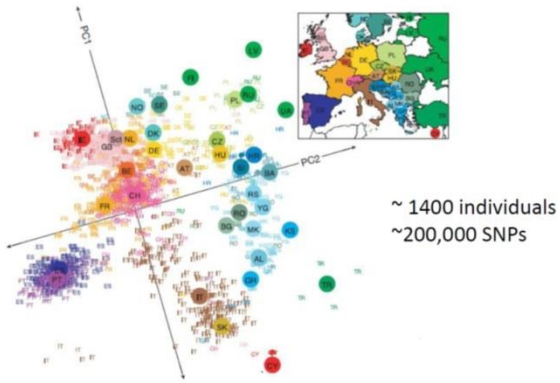
QTL (L), gene expression (R), protein (P), environment or epigenetic marker (E)



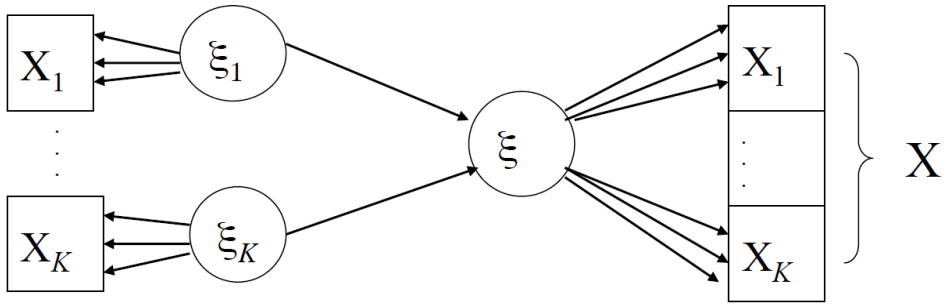
- **Genomic background** will remain playing a crucial role in complex traits, but not *the* only role.

Methodological areas I

- Multivariate dimension reduction

Unsupervised	
1 omics	
 (Novembre et al. 2008)	– **Principal component analysis** (PCA) when individuals are described by quantitative variables; – **Correspondence analysis** (CA) when individuals are described by two categorical variables that leads to a contingency table; – **Multiple correspondence analysis** (MCA) when individuals are described by categorical variables – **Non-linear (kernel) components analysis**

- Multivariate dimension reduction

Unsupervised + Supervised	
>1 omics	
 **component-based path-modelling** (Vinzi): PCA, CCA, GCCA (CarrolI), multiple factor analysis (Escofier and Pagès 1988), …, PLS regression of traits	– **Canonical** is the statistical term for analyzing latent variables (which are not directly observed) that represent multiple variables (which are directly observed) - extended to **more than two** sets as **generalized canonical analysis (GCA).** - **Different measurement scales** and high-dimensional **intra-correlated:** combine GCA with **optimal scaling,** with **sparsity** (Waaijenborg et al. 2009) and **regularization criteria** (Tenenhaus and Tenenhaus 2011) or co-inertia analysis techniques (Chessel & Hanafi 1996)

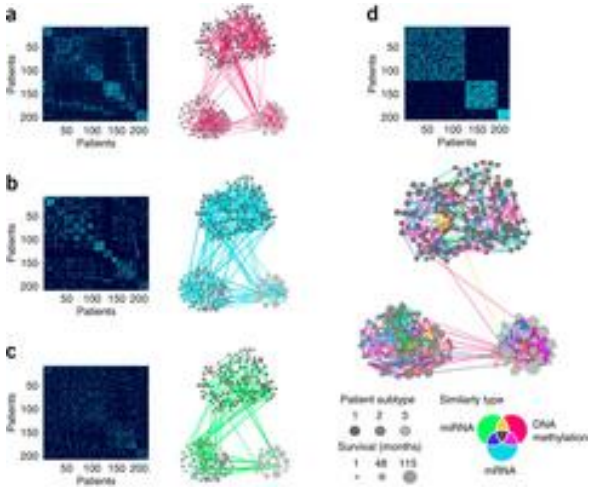
Methodological areas II

- Kernel-based statistical methods
  - Quite often kernel versions of data compression and de-noising algorithms exist (e.g., for supervised Fisher's discriminant analysis, unsupervised PCA)
  - At the basis lies a kernel matrix, which essentially constitutes similarity measures between pairs of entities (Q: genes, proteins, patients?)
  - The choice of kernel depends on the application field (research questions) and therefore flexibility is needed to accommodate the true nature of each omics data set.

Methodological areas III

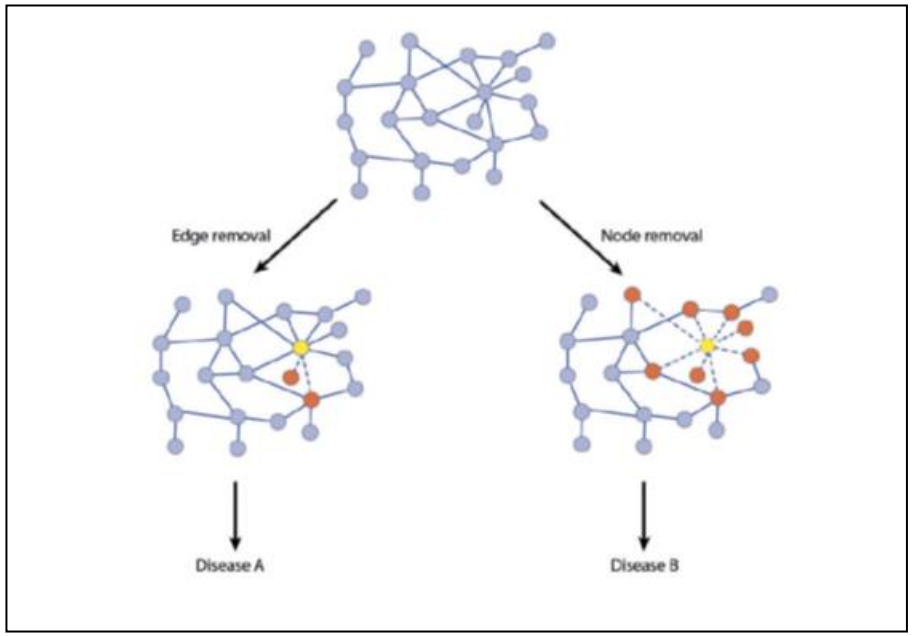
- Networks / graphical models
  - **Nodes:**
    - Original feature (Q: essential or redundant?)
    - Aggregate (Q: construction within a single omics data set or in the context of other sets as well)
  - **Edges**:
    - Biological vs statistical definition (cfr. statistical epistasis networks – supervised network construction)
    - Directed vs undirected
  - **Network comparison** (between different samples, e.g., cases and controls):
    - descriptive vs formal hypothesis testing

- Networks / graphical models

1 omics at a time	
 (cfr. Wang et al. 2014 applied to derive omics-based clusters of individuals)	– Combine / **fuse** different **gene networks**, preferentially accounting for different degrees of granularity, informativeness and precision. Comments: – **Inter-relationships** (signaling between omics layers) are ignored – **Nodes** appear in all segments – Still needs **post-linking** to phenotypes

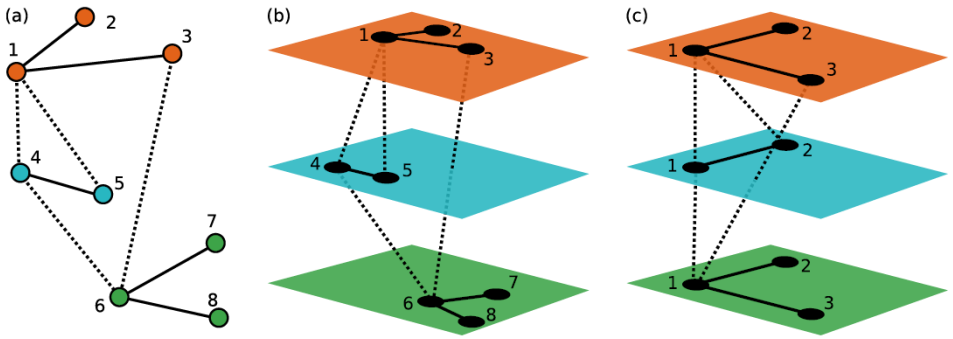
Biological networks

- From an evolutionary biology perspective, for a phenotype to be buffered against the effects of mutations, it must have an underlying genetic architecture that is comprised of networks of genes that are redundant and robust.
- The existence of these networks creates dependencies, realized as gene-gene interactions.
- omics-specific **intra-relationships** can be modified by another omics data types (e.g., genetic background / mutations)



(Wang et al.2011)

- Networks / graphical models

Multiple omics at once	
 Left: Example of a node-colored network (i.e., an interconnected network, a network of networks). Middle + Right: Same network represented by a multilayer network formalism. Right: The identity of the layer is needed to uniquely identify each node – Kavelä et al. 2013	– **Multi-layer networks** (Kavelä et al. 2013; Sánchez-García et al. 2013). Comments: – Allows for **inter-layer and intra-layer** edges. – Layers may exhibit different node/edge definitions

STEP 6: What are "non-obvious" methodological challenges?

" …just **adding one extra level of complexity** to a well-investigated data analysis type, such as when moving from genome-wide main effects SNP-based analyses to genome-wide interaction SNP-SNP analyses, offers **a sobering lesson** in what a lack of data (problem) acknowledgement can provoke. "

(Guserava et al., Van Steen 2015 – submitted)

- Population/patient heterogeneity: allow for non-linearity
- Replication: aggregate - micro-macro
- Meta-analysis: go non-parametric

Population substructure – the GWAI story

- Mixed models with (robust) genomic kinship estimates competes with determining (a number of) linear axes of genetic variation
  - Consider **non-linearity** (kernel PCA - ongoing)
- Structured Association
  - Improved clustering (generalized PCA, iterative PCA) (ongoing)
- Genomic control: one factor to deflate "all" statistical tests
  - Adapt the factor according to the particular test setting (MAF, …) (ongoing)

(FNRS PDR grant on "Foresting in integromics")

Replication – the GWAI story

"Genome-wide SNP genotyping platforms consist predominantly of **tagSNPs** from across the genome. Most of these SNPs are not causal and have no functional consequences. **When two or more tagSNPs are combined in a genetic interaction** model, is it reasonable to assume that the same combination of tagSNPs interacts in an independent dataset?"

(Ritchie and Van Steen 2015 – under review)

- Define the **(higher) level** that is common to studies (e.g., gene-level).

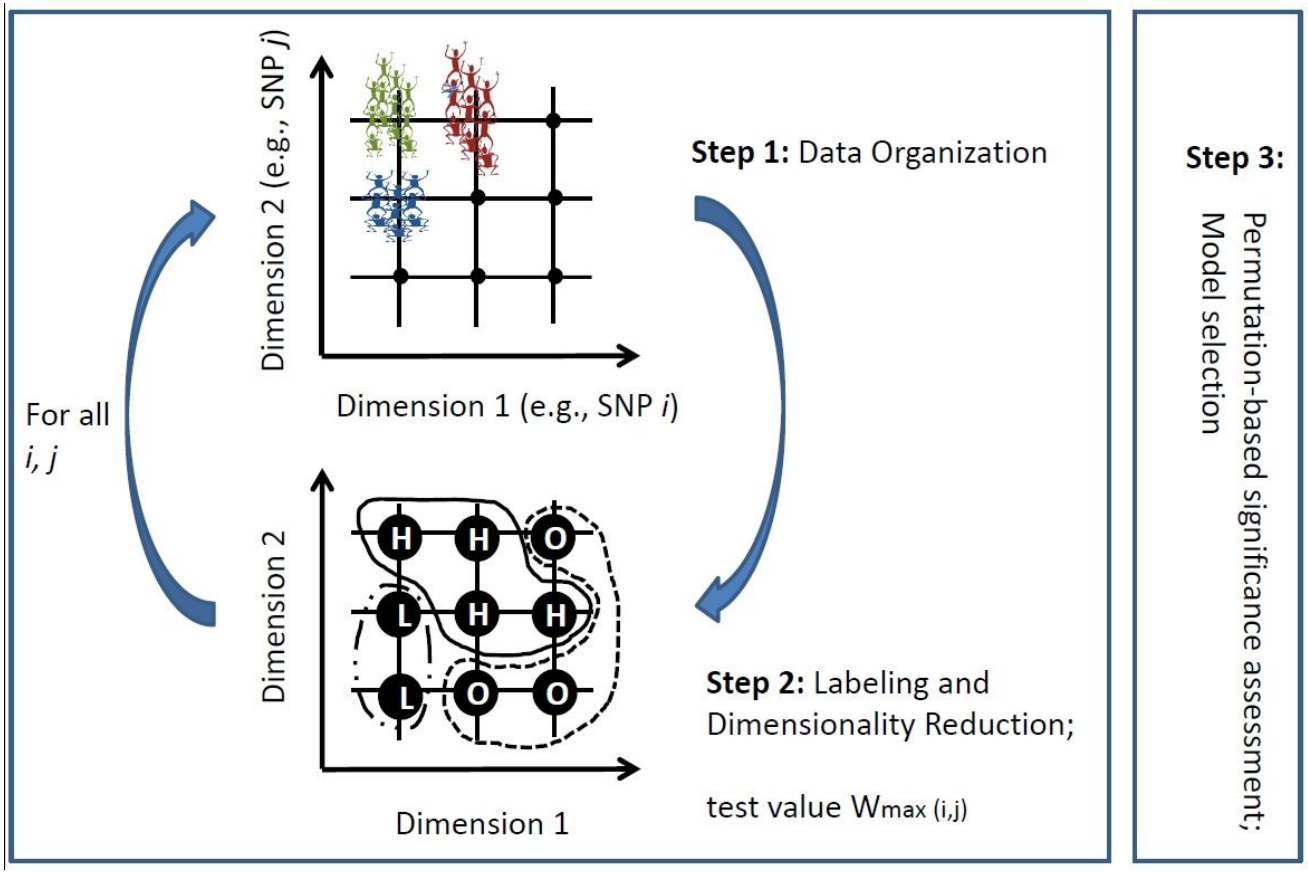
Meta-analysis – the GWAI story

- A multitude of analytic tools for GWAI analysis exist (Van Steen 2011)
  - Some give effect sizes → fixed or random-effects meta-analysis
  - Some give p-values → Fisher's combined p-value
  - New methods are needed to properly account for **analytic heterogeneity**
- As complexity increases, some model assumptions are expected to be too restrictive and too distinct from what is really going on in nature (Pereira et al. 2011)
  - Expose the field to **model-free / non-parametric meta-analysis** techniques

(FNRS grant on "Meta-analysis in GWAIs")

STEP 7: Will dimensionality reduction "keep the baby in the bathtub"?

Genomic MB-MDR ← MB-MDR (SNPxSNP) (Van Steen 2014)

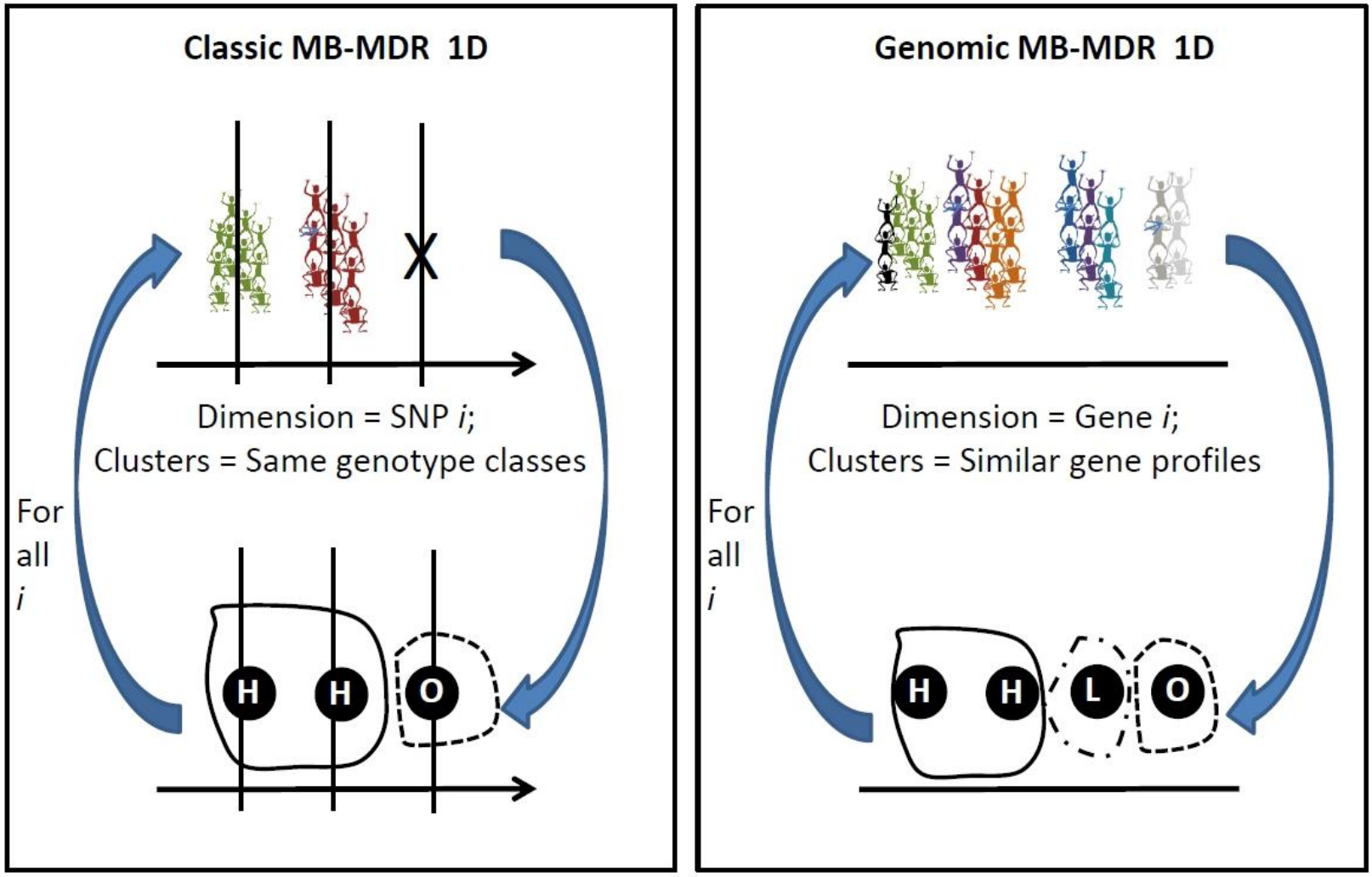


Learning by data summary

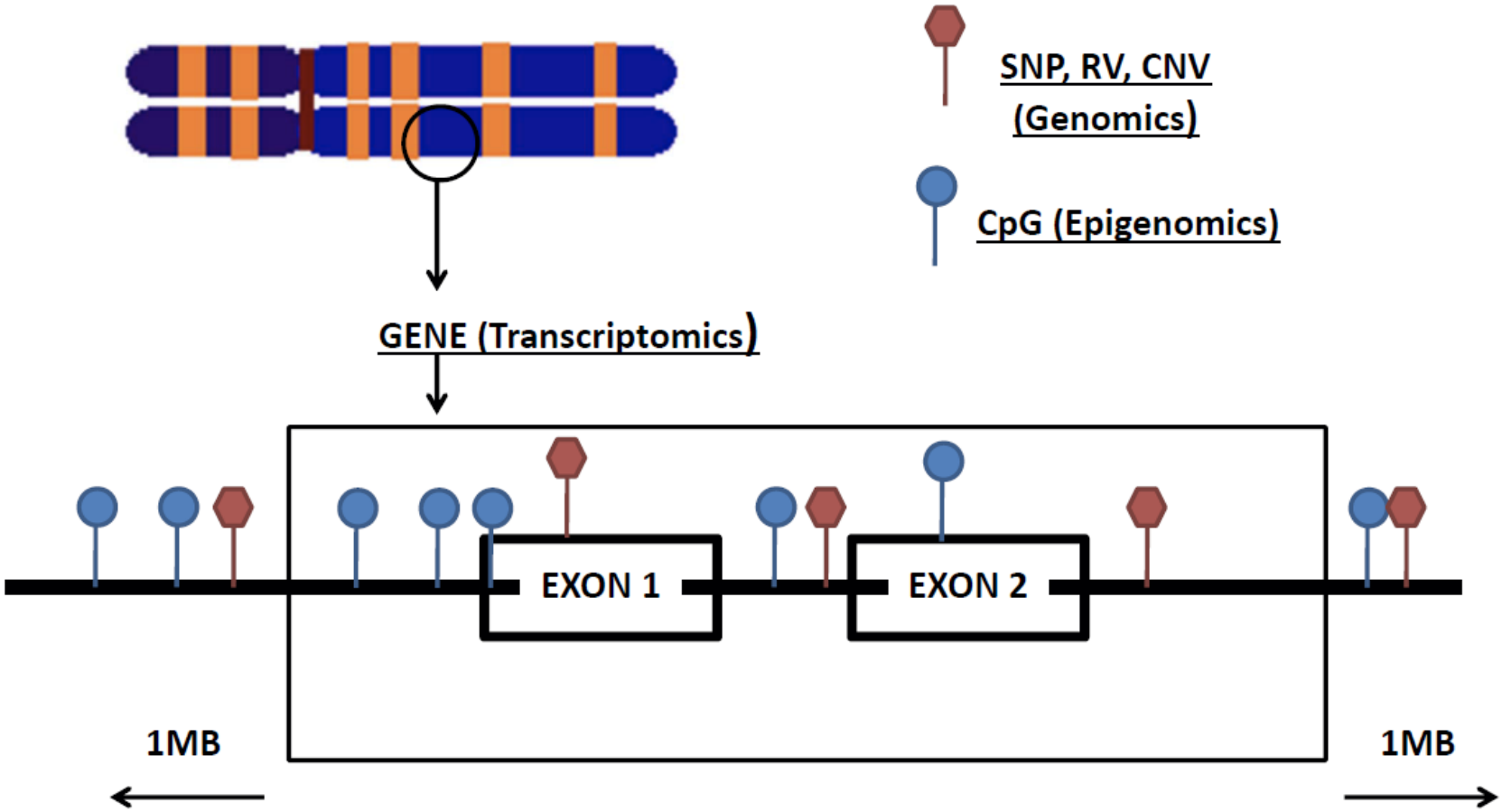
- Backpack items on the integromics road less travelled by, include:

Item	Our label
Speed controller	Gamma MaxT (Van Lishout et al.2015 – submitted)
Population / patient substructure or (cryptic) relatedness chart	MB-MDR for structured populations (Van Lishout et al. 2013 – poster ASHG, manuscript in preparation)
Heterogeneous and correlated input features map	Component-based Path Modeling (PLS-PM; Esposito Vinzi @ ERCIM2014 short course)
Replication / Meta-analysis tools	Easier to do when units of analysis are at a higher level (such as genes instead of {SNPs, epigenetic markers, miRNAs, …}) (Gusareva et al. 2014 – GWAI protocol)

MB-MDR (SNPxSNP) → Genomic MB-MDR (gene) (Fouladi et al. 2015)



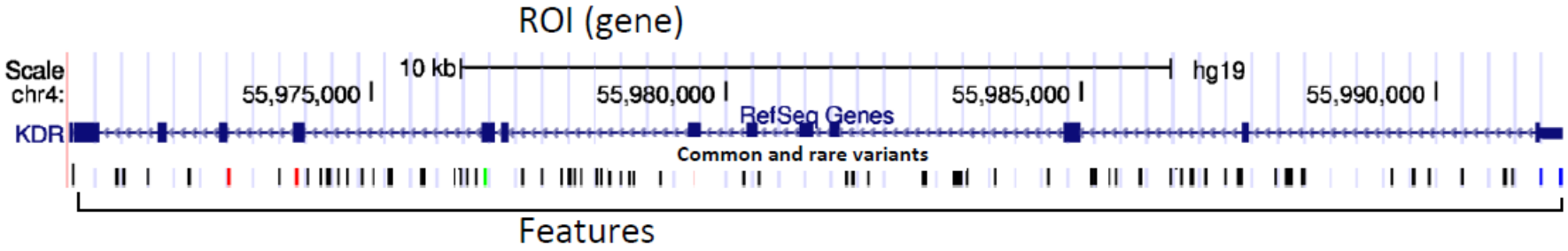
Genes have different faces



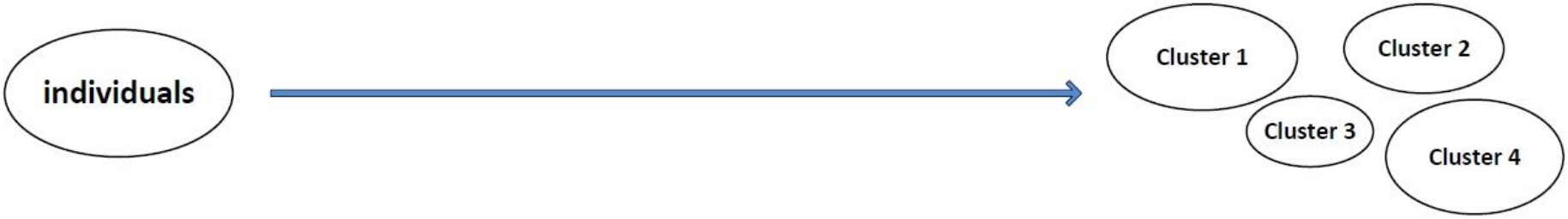
(Slide S Pineda – lab meeting 2014)

The genomic MB-MDR framework (Fouladi et al. 2015 – DNA-seq)

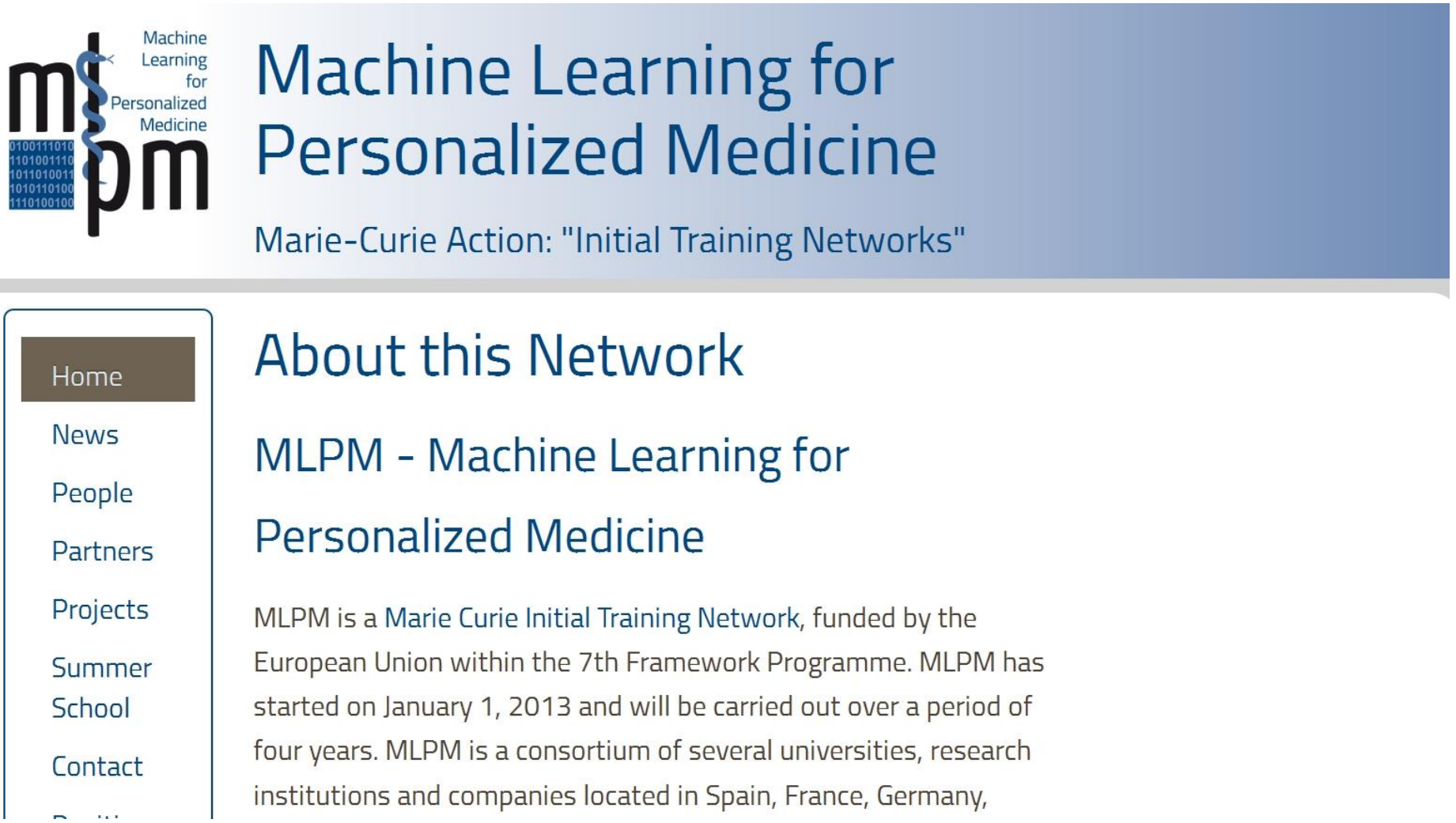
- **Phase 1:** Select sets of interest (ROI) / Prepare the data



- **Phase 2:** Clustering individuals according to features (e.g., common and rare variants, epigenetic markers, … and kernel PCA)

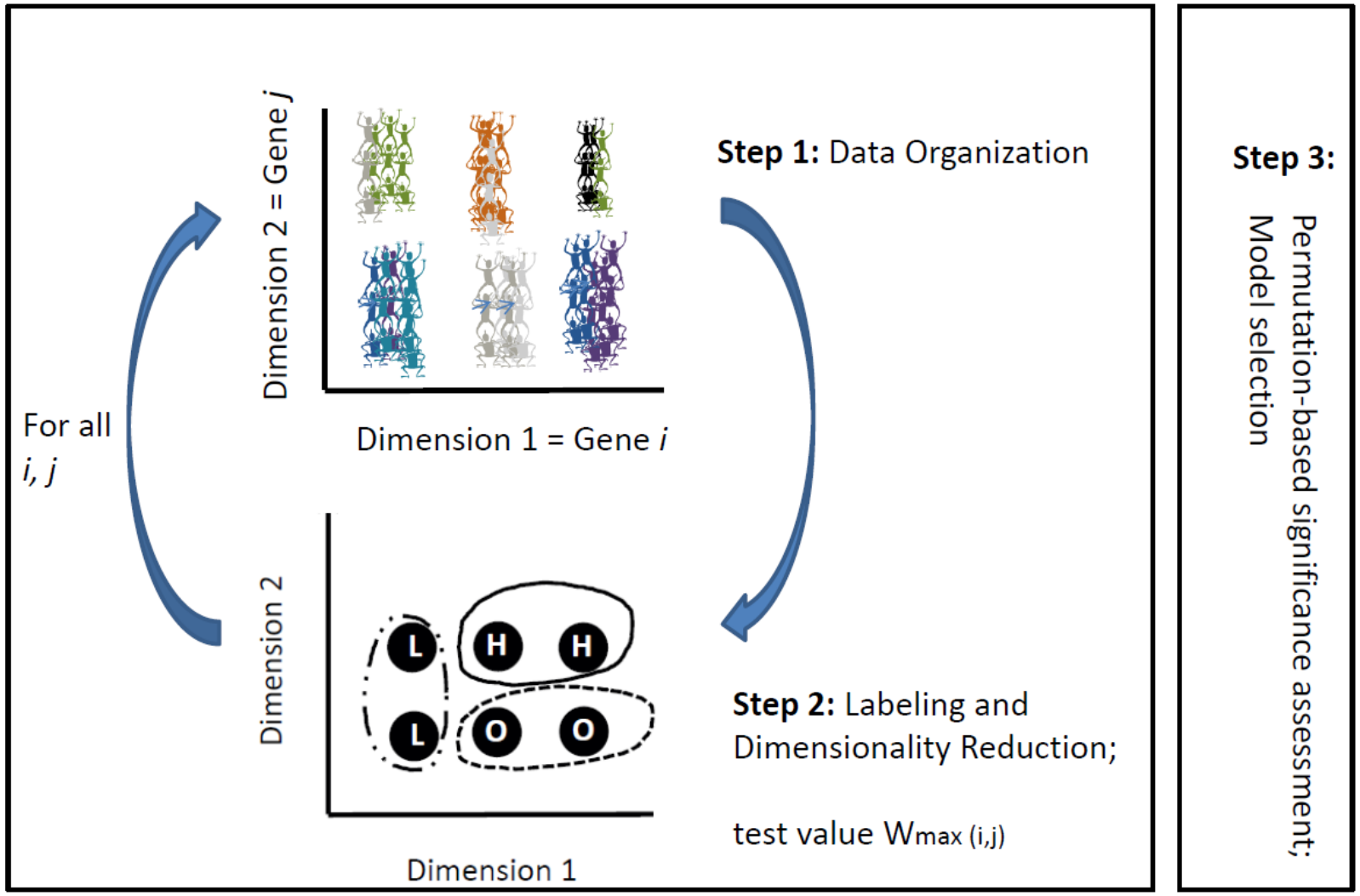


- **Phase 3:** Application of classic MB-MDR on new constructs


Machine Learning for Personalized Medicine
Marie-Curie Action: "Initial Training Networks"
Home
News
People
Partners
Projects
Summer School
Contact
About this Network
MLPM - Machine Learning for Personalized Medicine
MLPM is a Marie Curie Initial Training Network, funded by the European Union within the 7th Framework Programme. MLPM has started on January 1, 2013 and will be carried out over a period of four years. MLPM is a consortium of several universities, research institutions and companies located in Spain, France, Germany,

(http://mlpm.eu/)

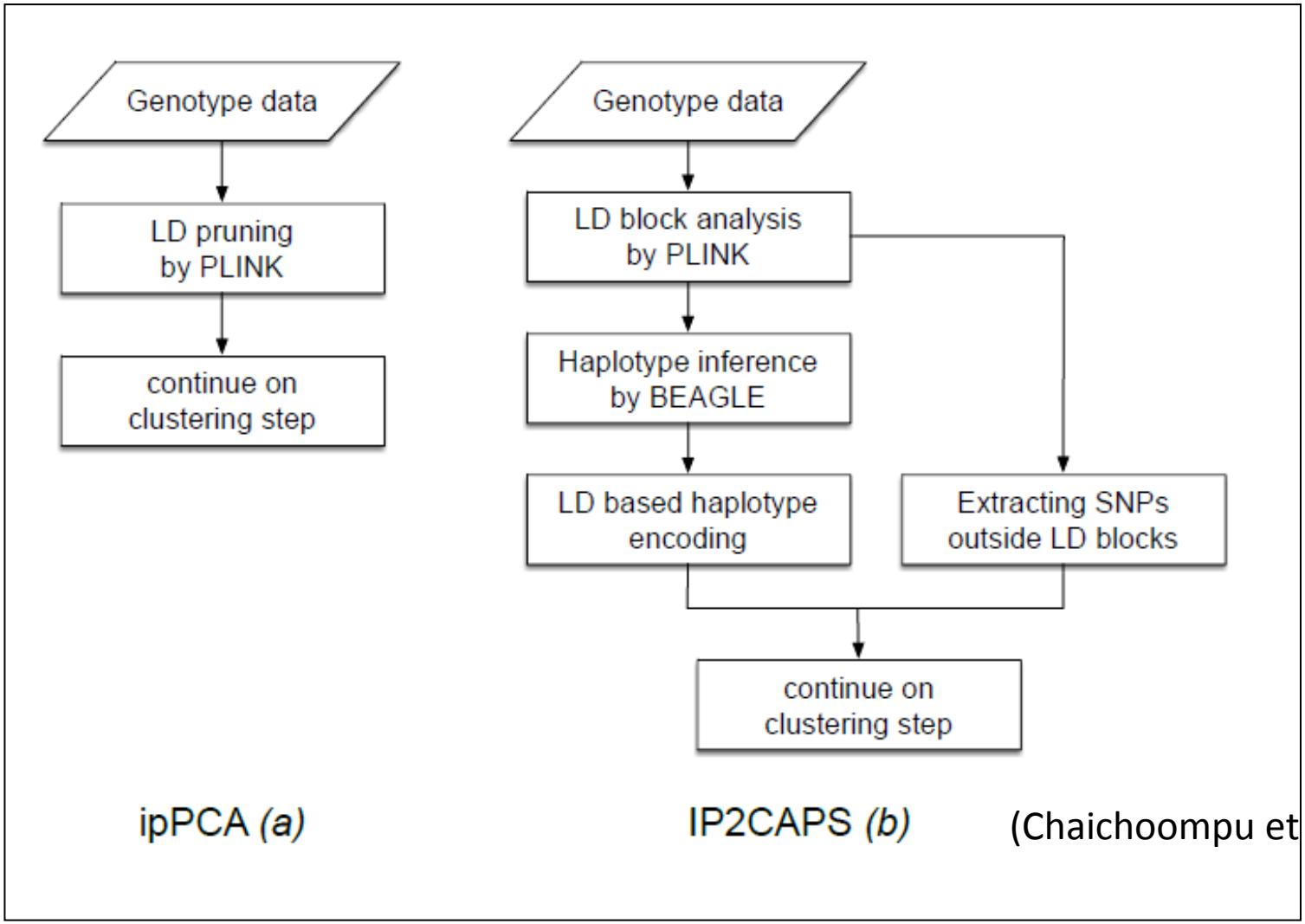
Bonus: gene-based statistical interaction networks



STEP 8: Is heterogeneity a nuisance or a relevant piece of information?

- With multiple omics data, chances increase to unravel very fine substructures in population or patient groups
- Emerging questions:
  - Are these structures “important”?
  - How to detect them?
  - How to optimally “use” this information in the integrative analysis (which is an analysis addressing a specific research question)?

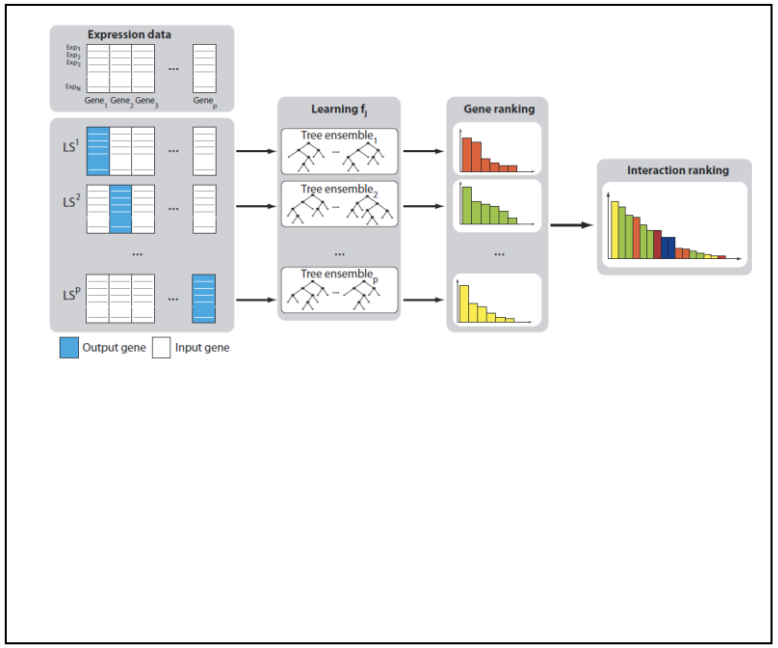
IP2CAPS → integrative fine structure detection



(Chaichoompu et al.2015 - submitted)

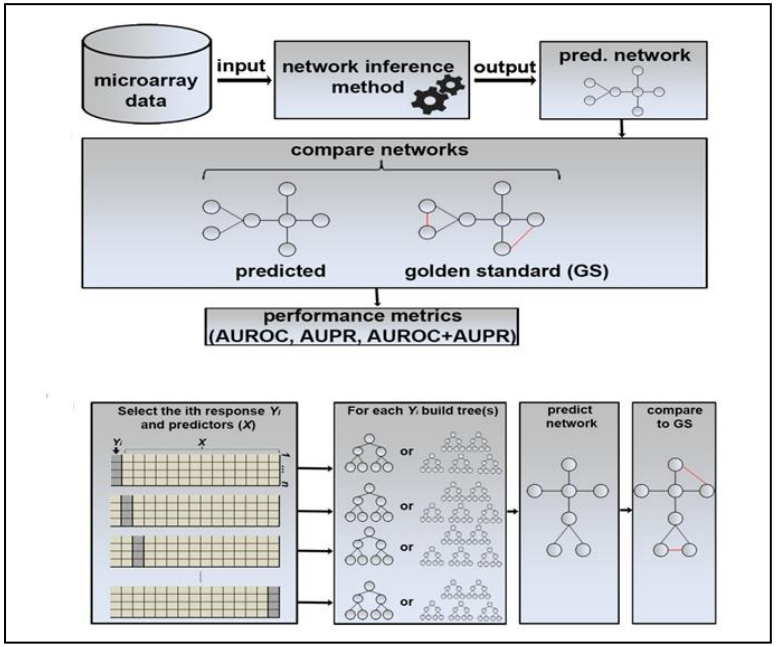
STEP 9: Can we learn from cross-disciplinary marriages?

- Huynh-Thu et al. (2010) had the clever idea to use Random Forests to infer regulatory networks (from expression data – genie3)



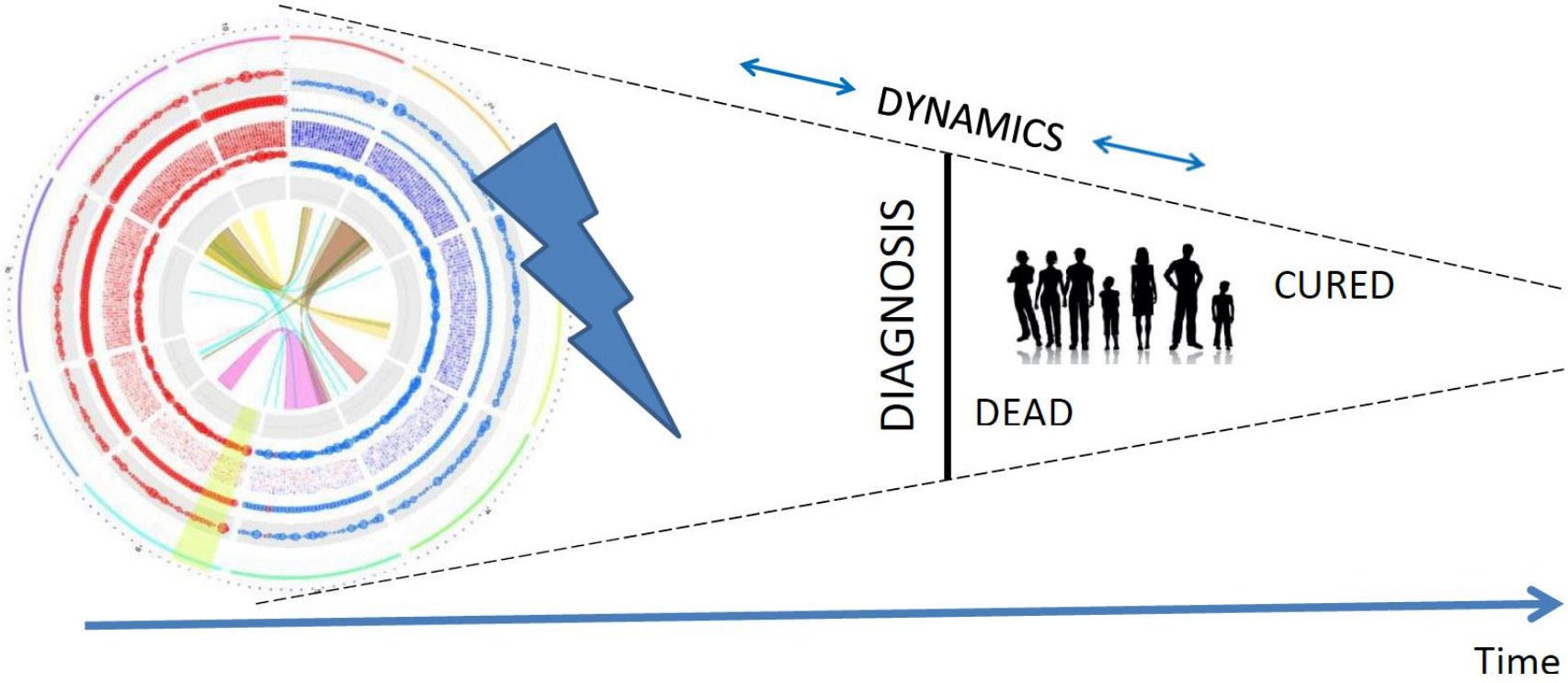
- Using Conditional Inference Forests" (CIFs) instead, has a few interesting advantages:

  Flexible integration of multiple **correlated** and/or **differently scaled** features (networks of networks)



STEP 10: Don't forget about …

- the fact that complex phenotypes are determined by multiple factors, both omics and non-omics, possibly modified over time



(Van Steen K and Malats N 2015)

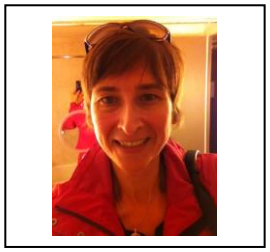
In conclusion

- Global genome-wide studies (e.g. GWAs, GWAIs) describe systems of a size that cannot be modeled to the detailed level of biological systems
- Integrative studies and systems genetics may help in providing functional interpretations
- To date, both are still too high level to provide full functional explanations at a molecular or even atomic level
- There is a niche for combined statistical modeling and machine learning (deep learning), as well as mathematical modeling

Acknowledgements

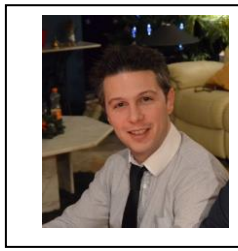
Biostatistics, Biomedicine, Bioinformatics

E. Gusareva



F. Gadaleta

F. Van Lishout



BIO

B. Dizier

Bio^3: **Bio**statistics – **Bio**medicine - **Bio**informatics

K. Bessonov

R. Fouladi

S. Pineda

K. Chaichoompu

Learning from data with MB-MDR (synthetic + real-life)

- **Calle** ML, Urrea V, Van Steen K (2010) mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. Bioinformatics Applications Note 26 (17): 2198-2199 **[first MB-MDR software tool]**
- **Cattaert** T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards T, Van Steen K. (2010) FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals, PLoS One 5 (4). **[first implementation of MB-MDR in C++, with improved features on multiple testing correction and improved association tests + recommendations on handling family-based designs]**
- **Cattaert** T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K (2010) Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case-control data in the presence of noise (*invited paper*). Ann Hum Genet. 2011 Jan;75(1):78-89 **[detailed study of C++ MB-MDR performance with binary traits]**
- **Mahachie John** JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, Van Steen K (2011) Comparison of genetic association strategies in the presence of rare alleles. BMC Proceedings, 5(Suppl 9):S32 **[first explorations on C++ MB-MDR applied to rare variants]**

- **Mahachie John** JM, Cattaert T, Van Lishout F, Van Steen K (2011) Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. European Journal of Human Genetics 19, 696-703. **[detailed study of C++ MB-MDR performance with quantitative traits]**
- **Van Steen** K (2011) Travelling the world of gene-gene interactions *(invited paper).* Brief Bioinform 2012, Jan; 13(1):1-19. **[positioning of MB-MDR in general epistasis context]**
- **Mahachie John** JM , Cattaert T , Van Lishout F , Gusareva ES , Van Steen K (2012) Lower-Order Effects Adjustment in Quantitative Traits Model-Based Multifactor Dimensionality Reduction. PLoS ONE 7(1): e29594. doi:10.1371/journal.pone.0029594 **[recommendations on lower-order effects adjustments]**
- **Mahachie John** JM, Van Lishout F, Gusareva ES, Van Steen K (2012) A Robustness Study of Parametric and Non-parametric Tests in Model-Based Multifactor Dimensionality Reduction for Epistasis Detection. BioData Min. 2013 Apr 25;6(1):9**[recommendations on quantitative trait analysis]**
- **Van Lishout** F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, Theâtre E, Charloteaux B, Calle ML, Wehenkel L, Van Steen K (2012) An efficient algorithm to perform multiple testing in epistasis screening. BMC Bioinformatics. 2013 Apr 24;14:138 **[C++ MB-MDR made faster!]**

- **Gusareva ES,** Van Steen K (2014) Practical aspects of genome-wide association interaction analysis. Hum Genet 133(11):1343-58 **[GWAI analysis protocol]**
- **Van Lishout F,** Gadaleta F, Moore JH, Wehenkel L, Van Steen K (2015) gammaMAXT: a fast multiple-testing correction algorithm – submitted **[C++ MB-MDR made SUPER-fast]**
- **Fouladi R**, Bessonov K, Van Lishout F, Van Steen K (2015) Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis. Human Heredity – accepted **[aggregating based on similarity measures to deal with DNA-seq data]**
- **Bessonov K, Gusareva ES, Van Steen K (2015)** A cautionary note on parameter impact in Genome-Wide Association gene-1 gene Interaction protocols exemplified in ankylosing spondylitis. Hum Genet - accepted **[non-robustness of GWAI analysis protocols]**
- **Chaichoompu K,** Fouladi R, Pongsakorn W, Wangkumhang, Wilantho A, Chareanchim W, Sakuntabhai A, Shaw PJ, Tongsima S, Van Steen K (2015) IP2CAPS: Iterative pruning to capture population structure – submitted **[dealing with fine population substructure]**