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: 10 steps to SUCCESS

1. Understanding “integromics”
2. Developing an integrative analysis pipeline
3. Motivating “integromics”
4. Working out the analytics
5. Dealing with the obvious
6. Dealing with the non-obvious
7. Reducing dimensions …with caution
8. Acknowledging (sub-)structures
9. Networking
10. Recognizing that “omics” provide one side of the story

Boiling the Ocean

– Ten words 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 **… 10 questions**

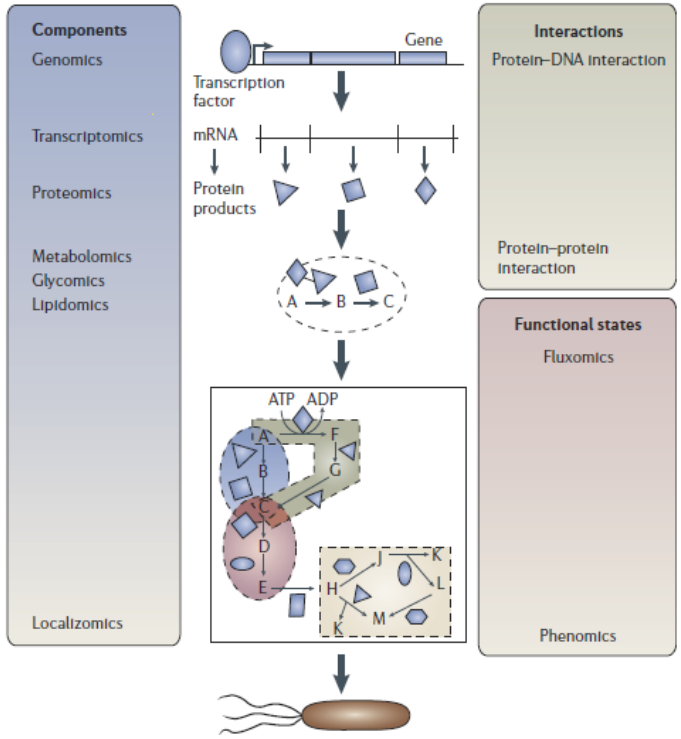
STEP 1: What is INTEGROMICS?

- INTEGROMICS = integration + omics
- Data integration may mean different things in different contexts.

- 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 data as a starting point

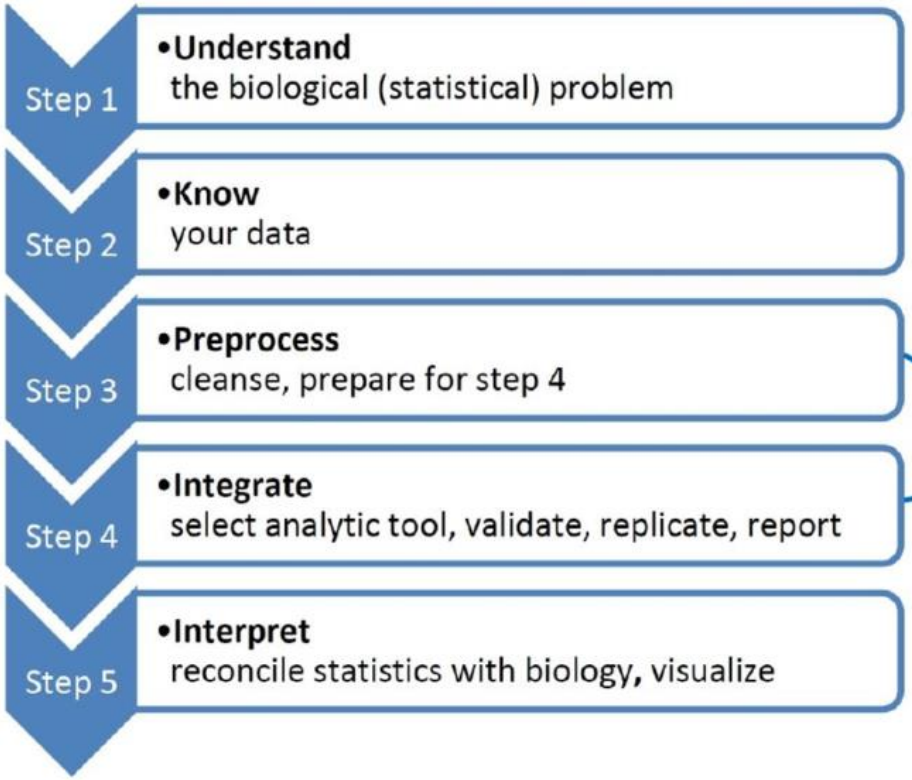
- These data can be classified into three categories: components, interactions and functional-states data:
  - **components** data detail the molecular content of the cell or system,
  - **interactions** data specify links between molecular components,
  - **functional-states** data provide an integrated readout of all omics data types by revealing the overall cellular phenotype.



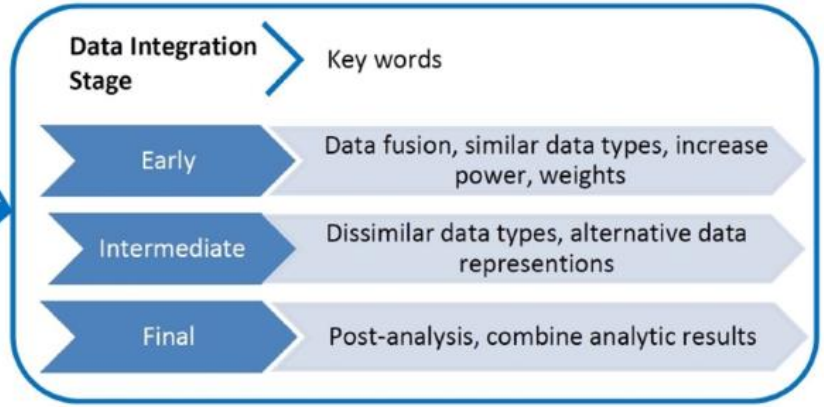
(Joyce and Palsson 2006)

STEP 2: What are its 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

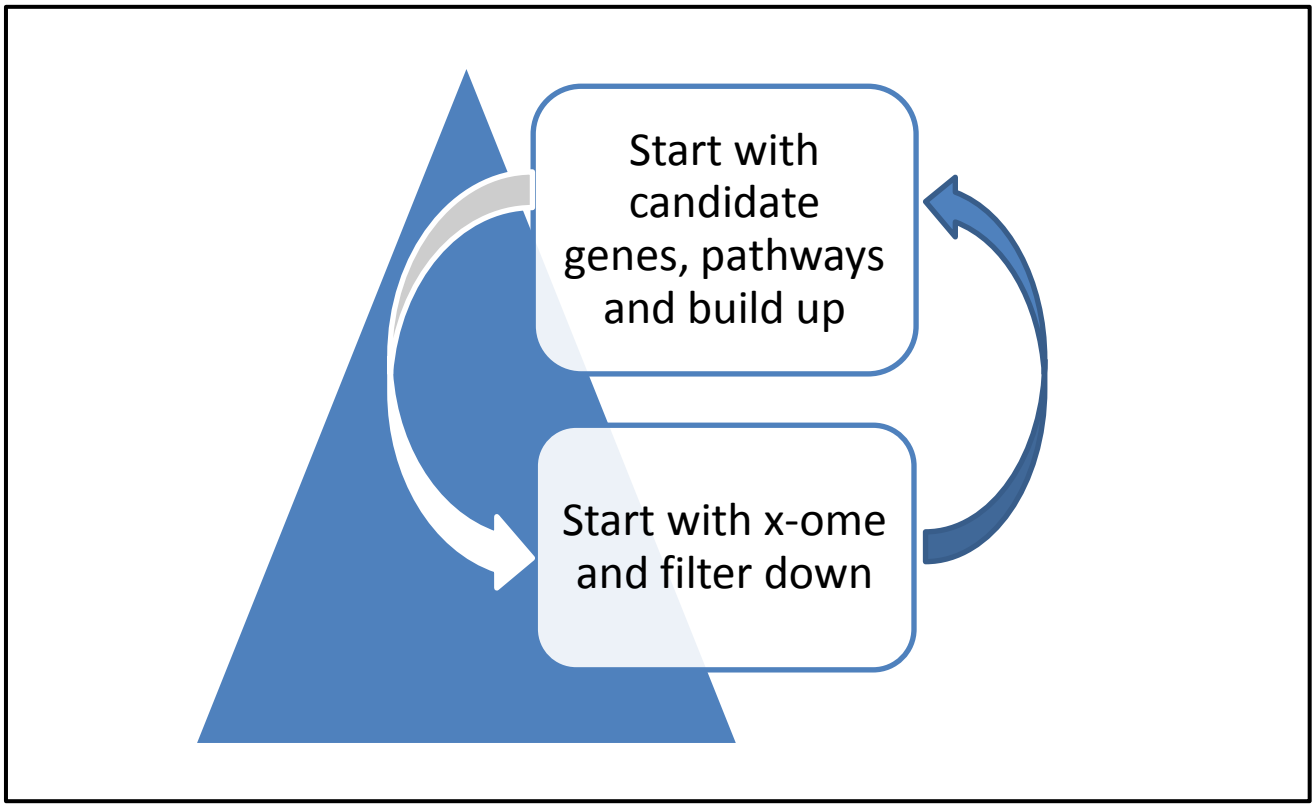
Step 3

- 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



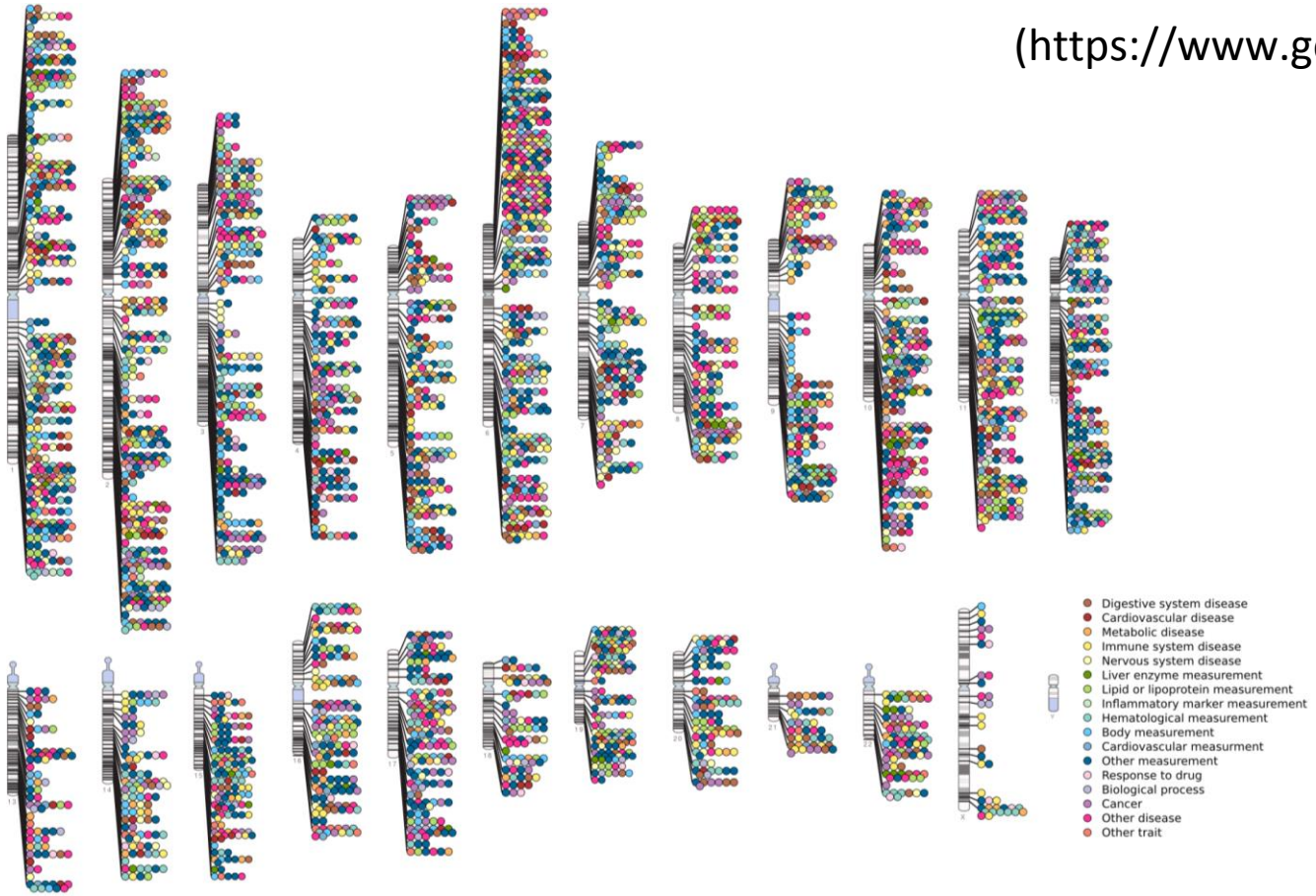
Step 5

- 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?

GWA replication successes



(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

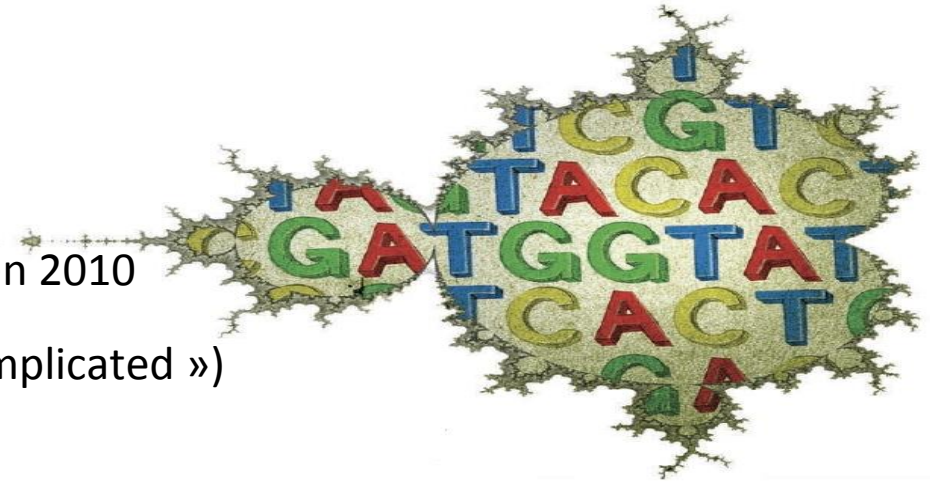
Explanation	Rationale	Comments
Overestimated heritability estimates	These estimates are typically performed in the absence of gene-gene or gene-environment interactions (Young et al 2014)	Limiting pathway modeling suggests that epistasis could account for missing heritability in complex diseases (Zuk et al 2012)
Common genetic variants	More common variants are likely to be found in GWAs with larger sample sizes (drawback:)	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)

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)
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.

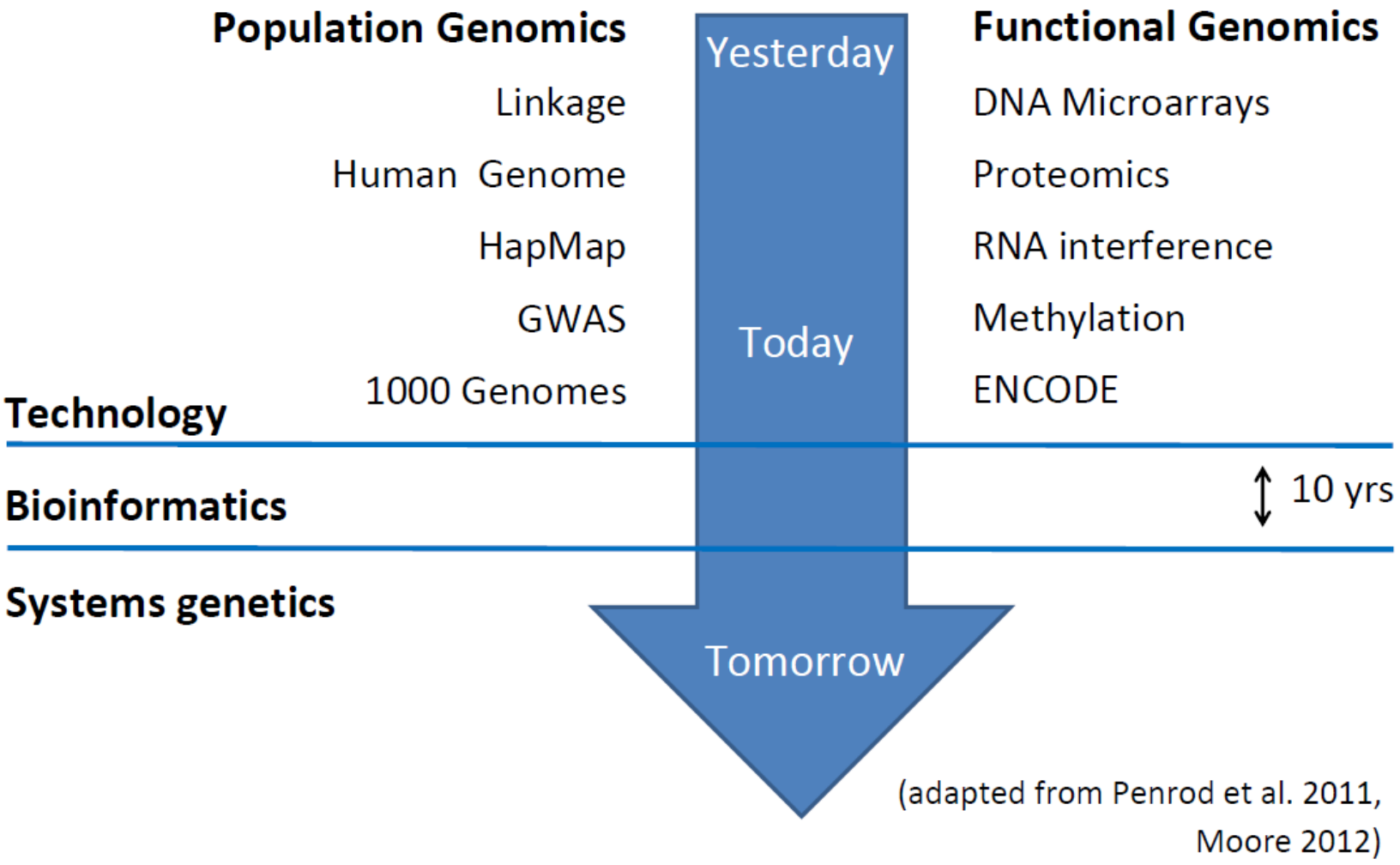
(adapted from Silverman et al 2012)

(pic : Hayden 2010

« Life is Complicated »)



Modeling systems genetics …



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

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. "

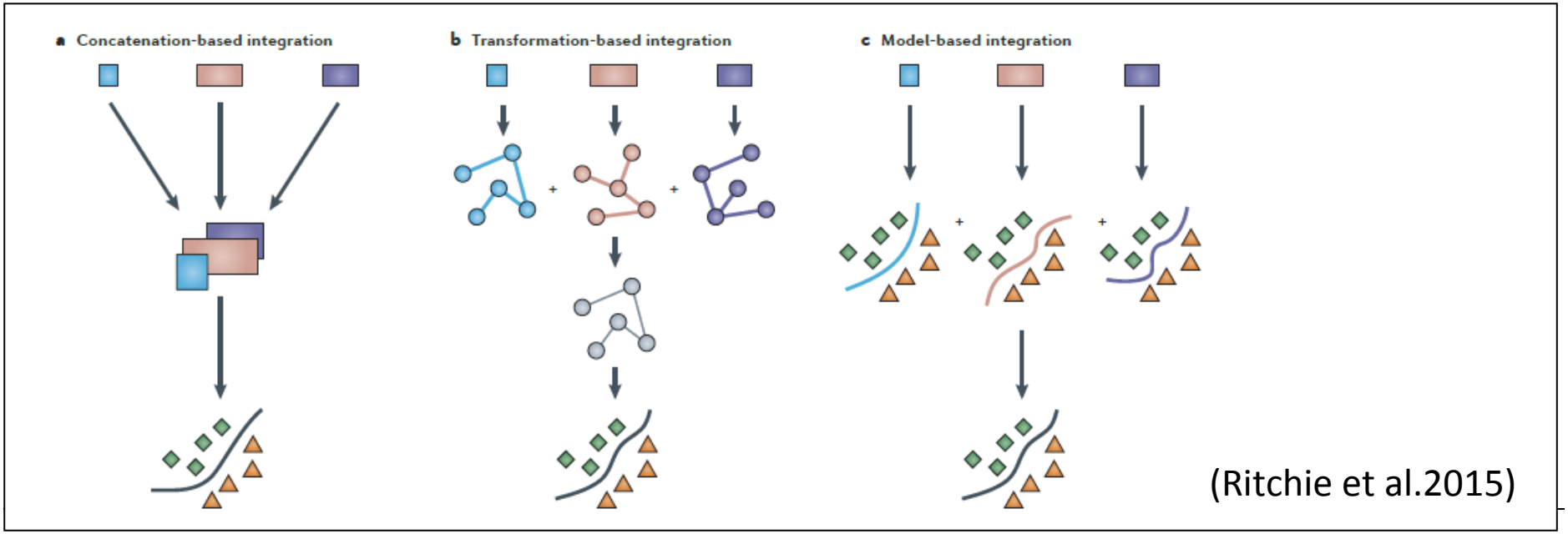
(Guserava et al., Van Steen 2015 – submitted)

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

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

STEP 4: Which routes lead to INTEGROMICS?

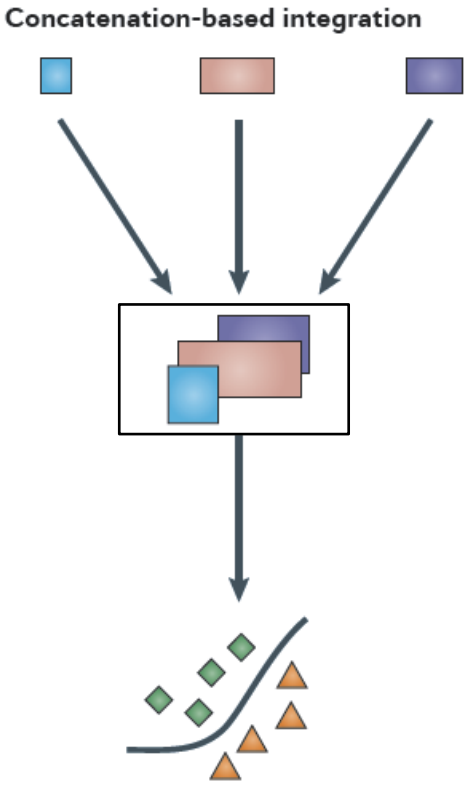
- In correspondence with the description of the Hamid "stages", Ritchie et al. (2015) refer to concatenation-based (left), transformation-based (middle) or model-based integrative (right) approaches
- The Hamid view and the Ritchie view are essentially two faces of the same coin

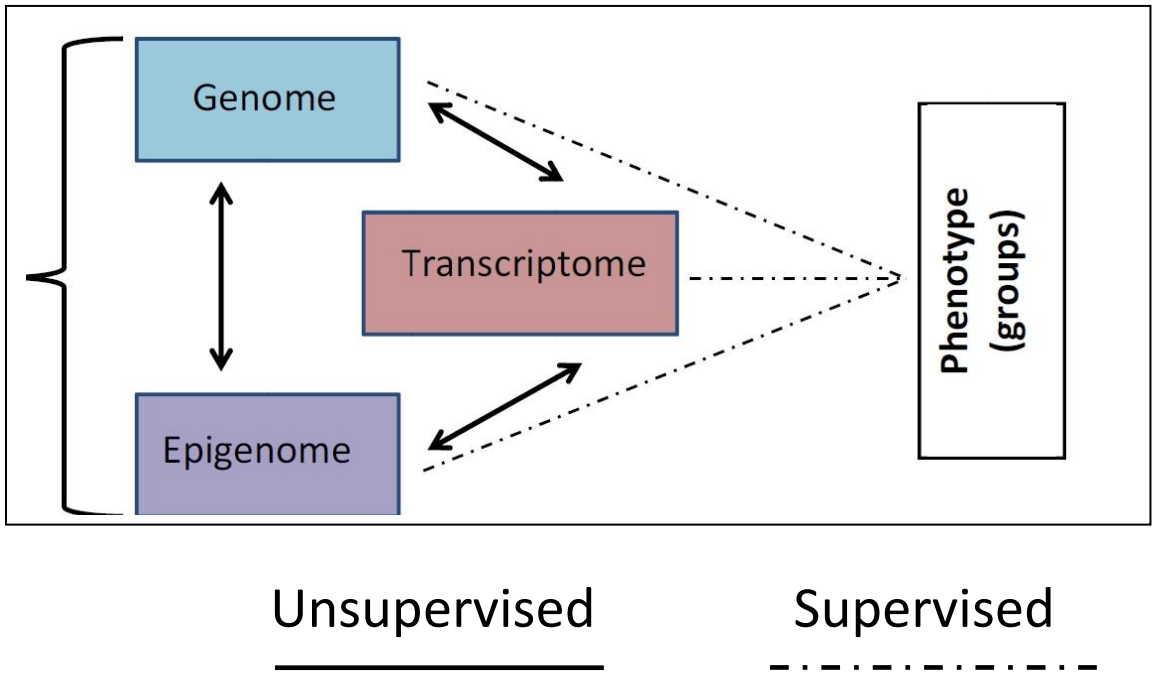


(Ritchie et al.2015)

STEP 5: What are "obvious" methodological challenges?

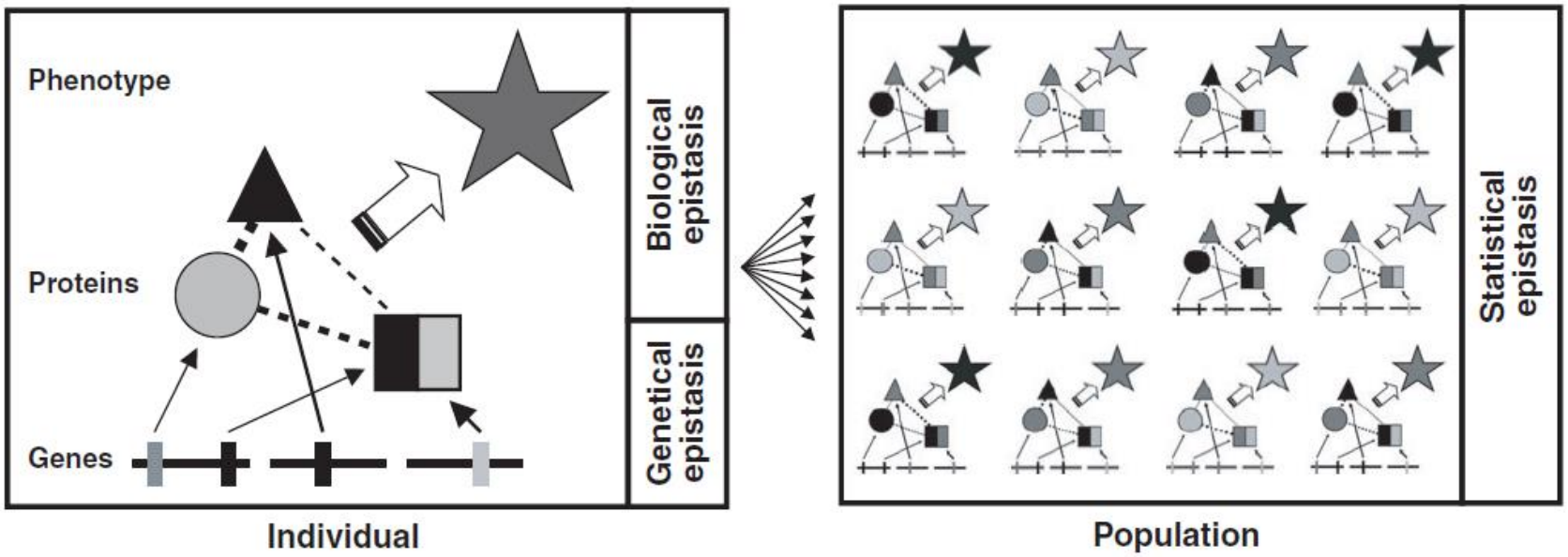
- It is obvious that only by concatenating, one is able to account for "relationships" between different omics data sources





Omics data are related

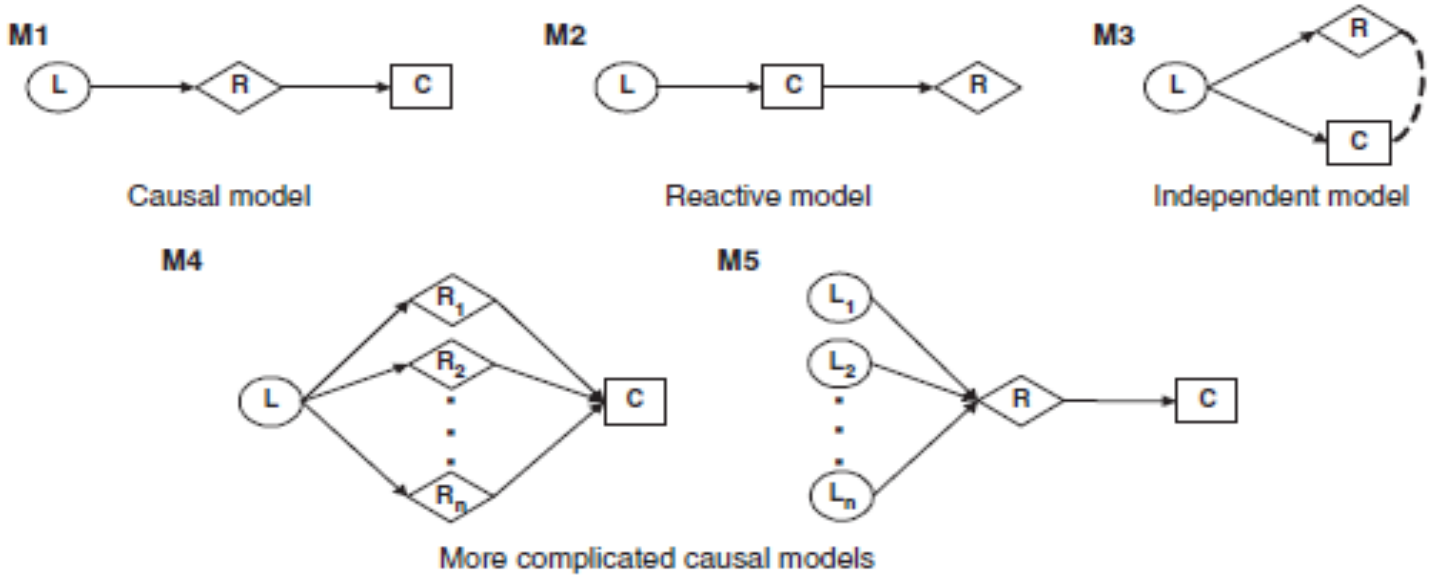
- Two or more DNA variations may “interact” either directly to change transcription or translation levels, or indirectly by way of their protein product (to alter disease risk separate from their independent effects)



(Moore 2005)

Omics data are related

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



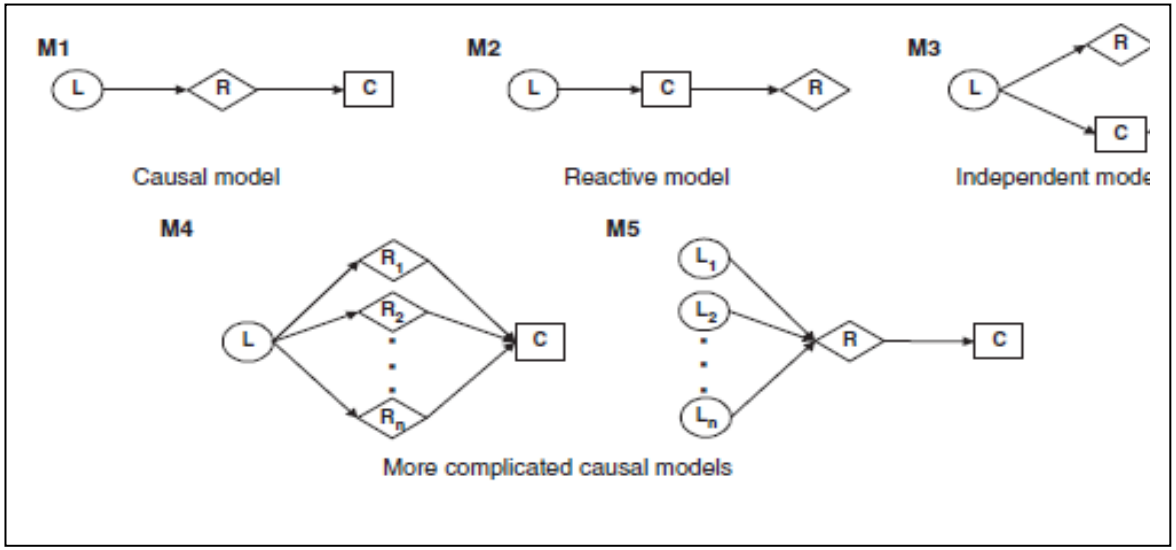
Graphical models for relationships between QTLs, RNA levels and complex traits, assuming gene expression (R) and complex trait (C) are under the control of a common QTL (L) - Schadt et al. (2005)

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

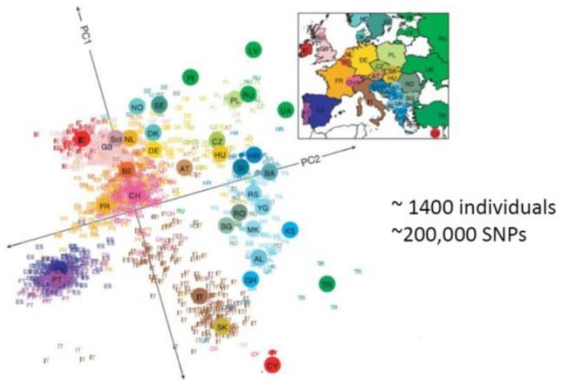
QTL (L), gene expression (R), protein (P), environment or epigenetic marker (E)



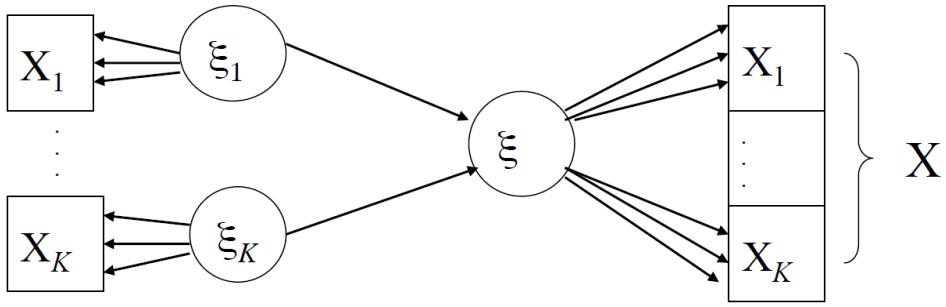
- **Genomic background** will remain playing a crucial role in complex traits, but not *the* only role.

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**

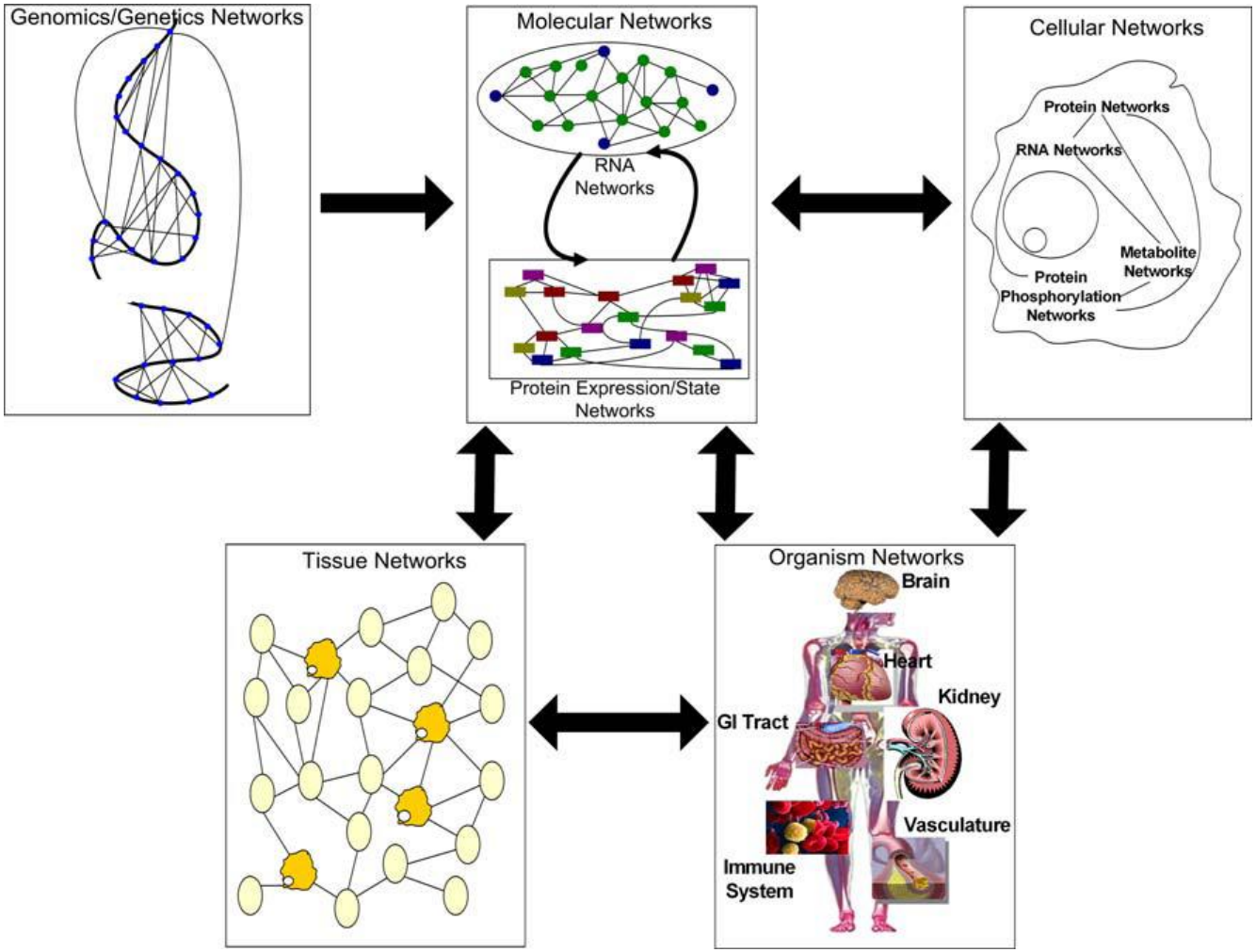
- Multivariate dimension reduction

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.

Flow of information in biological systems – a hierarchy of networks



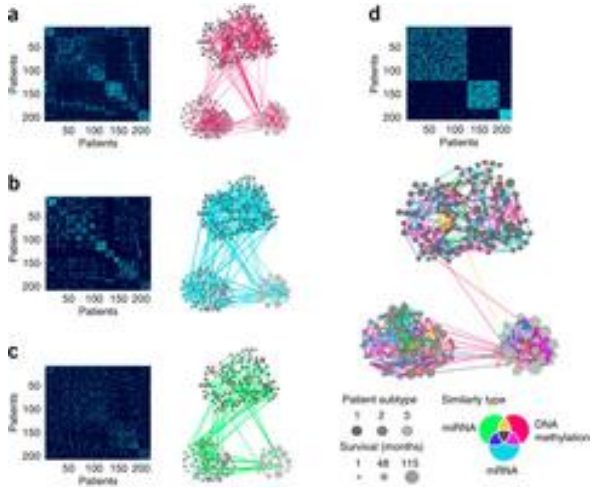
(Sieberts et al. 2007)

Methodological areas III

- Networks / graphical models

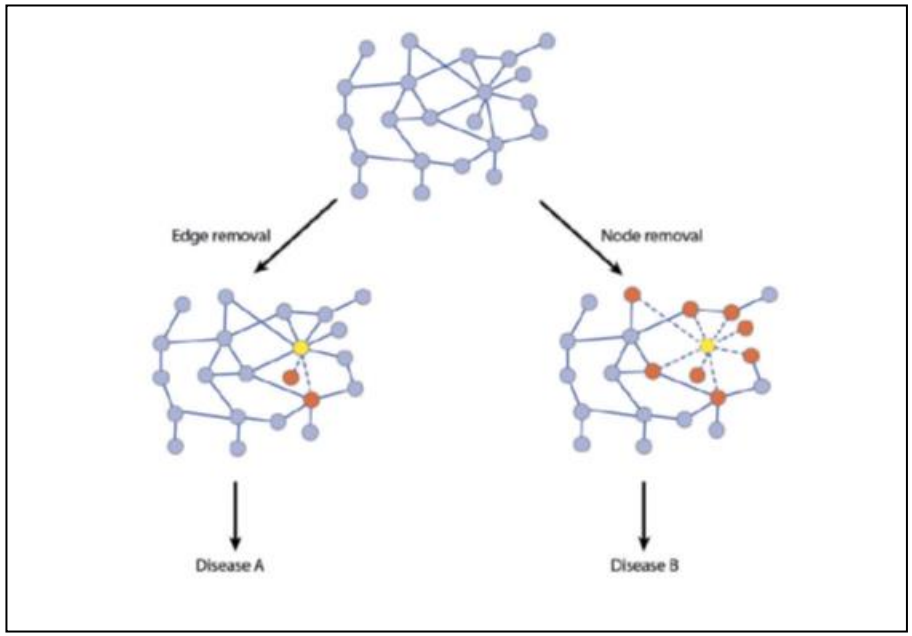
  - **Nodes:**
    - Original feature (Q: essential or redundant?)
    - Aggregate (Q: construction within a single omics data set or in the context of other sets as well)
  - **Edges**:
    - Biological vs statistical definition (cfr. statistical epistasis networks – supervised network construction)
    - 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

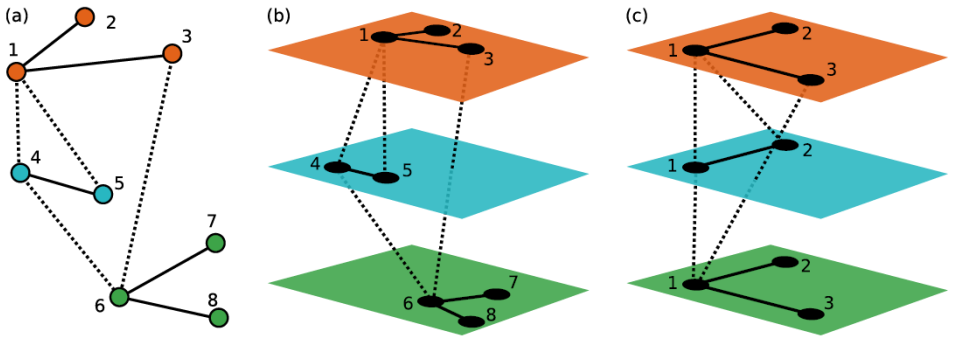
Biological networks

- 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. "

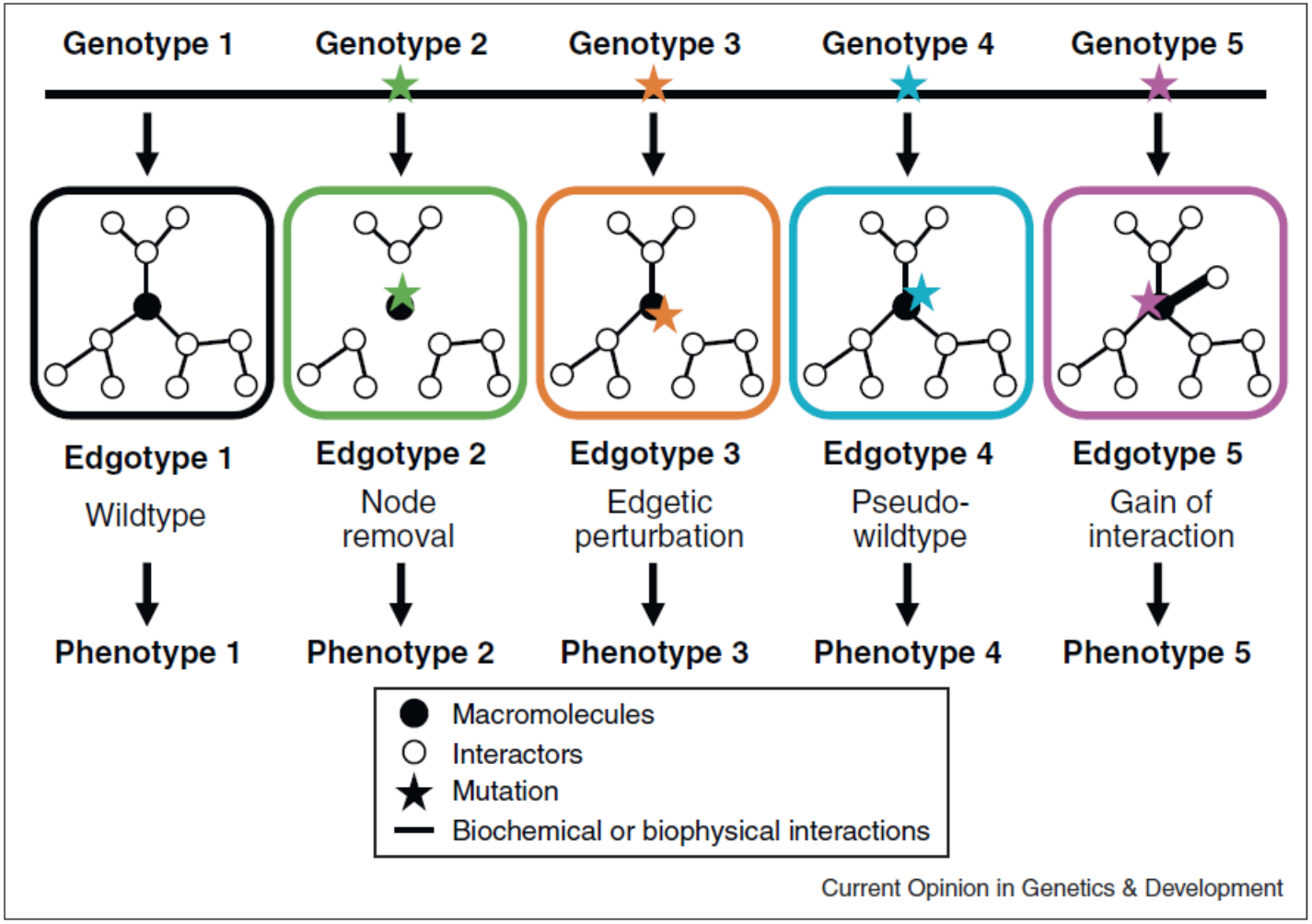
(Guserava et al., Van Steen 2015 – submitted)

- 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")



(Sahni et al. 2013)

Replication – the GWAI story

"Leaving aside for the moment **what replication means** or should mean in the context of GWAIS, even for the currently so-called replicated genetic interactions it is unclear to what extent **a false positive has been replicated** due to the adopted methodological strategy itself or whether the replication of epistasis is not solely attributed to main effects (such as HLA effects) not properly accounted for."

"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).

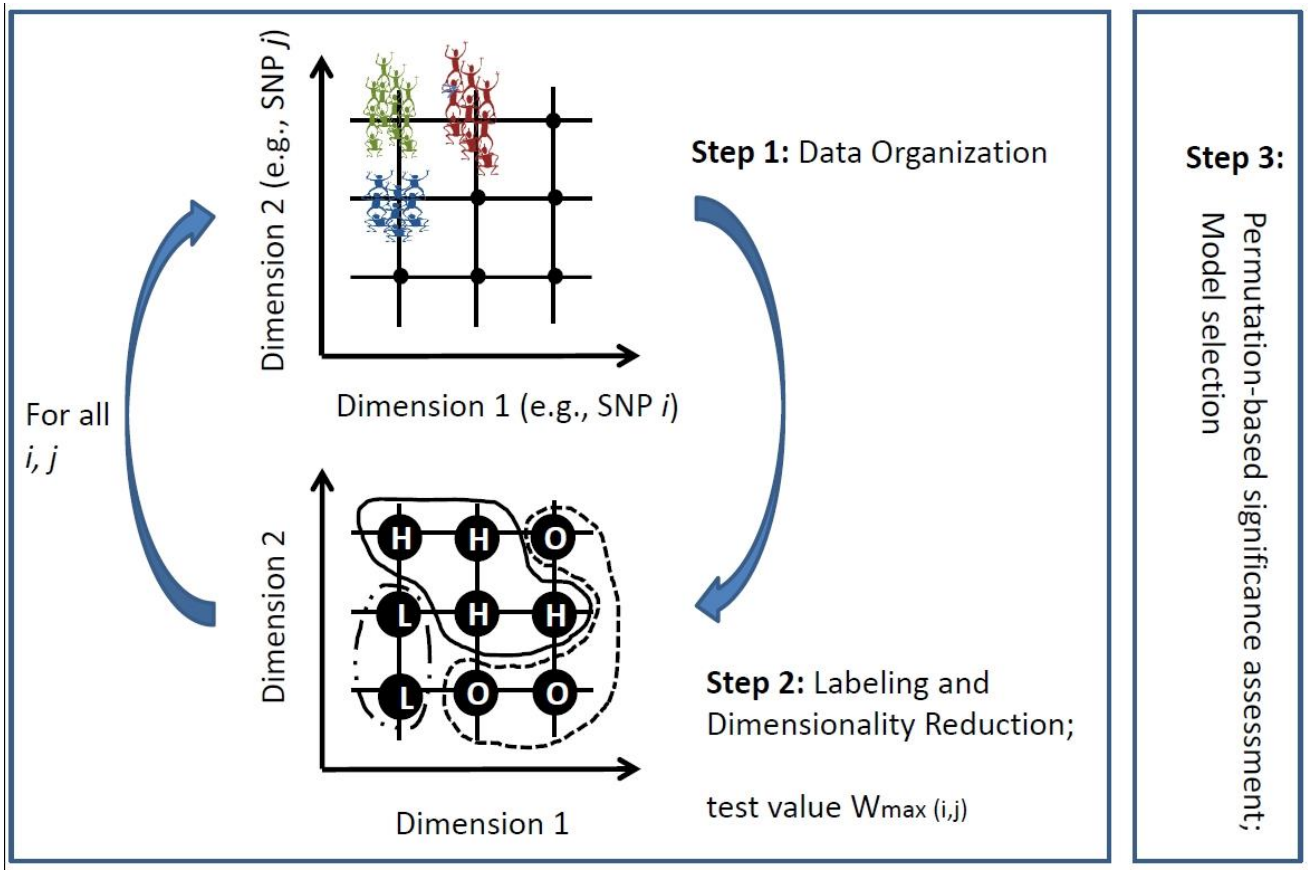
Meta-analysis – the GWAI story

- 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 **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 from data (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!]**

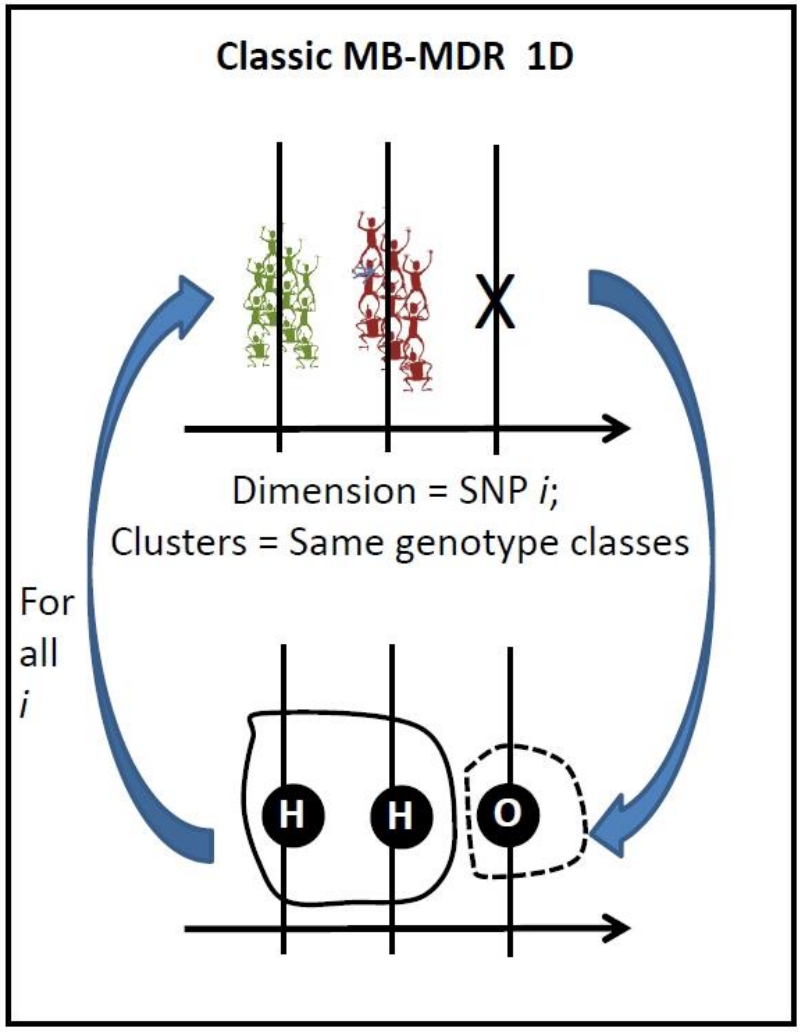
- **Gusareva ES,** Van Steen K (2014) Practical aspects of genome-wide association interaction analysis. Hum Genet 133(11):1343-58 **[GWAI analysis protocol]**
- **Van Lishout F,** Gadaleta F, Moore JH, Wehenkel L, Van Steen K (2015) gammaMAXT: a fast multiple-testing correction algorithm – submitted **[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]**

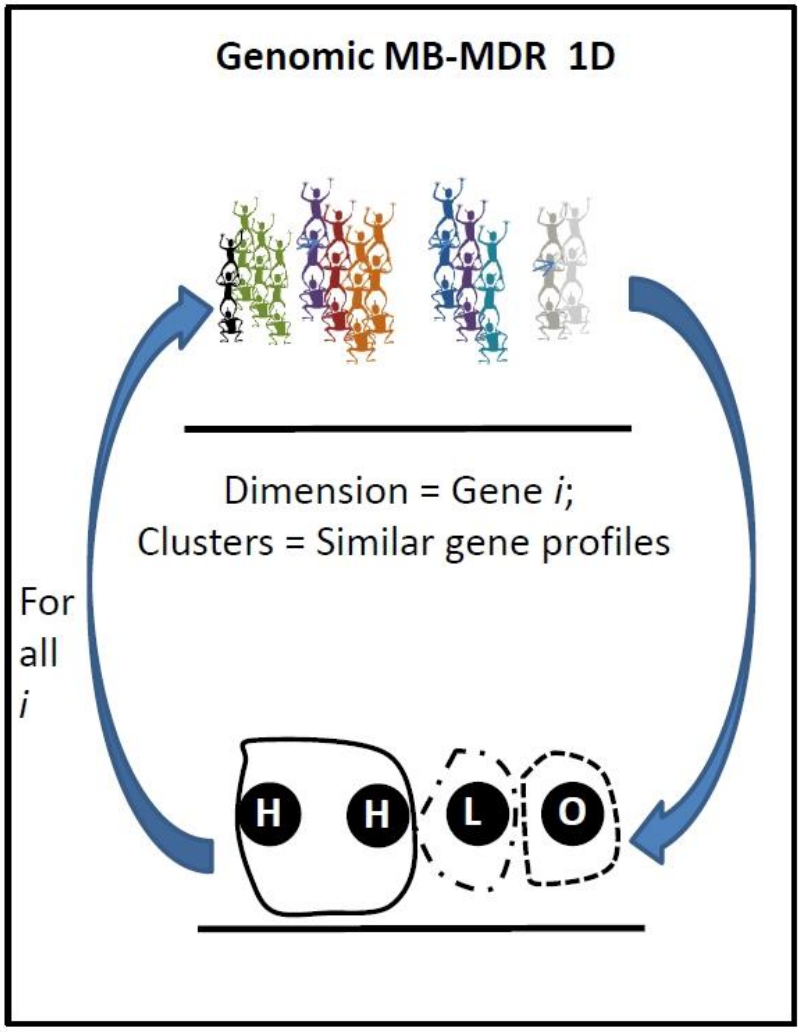
Learning by data summary

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

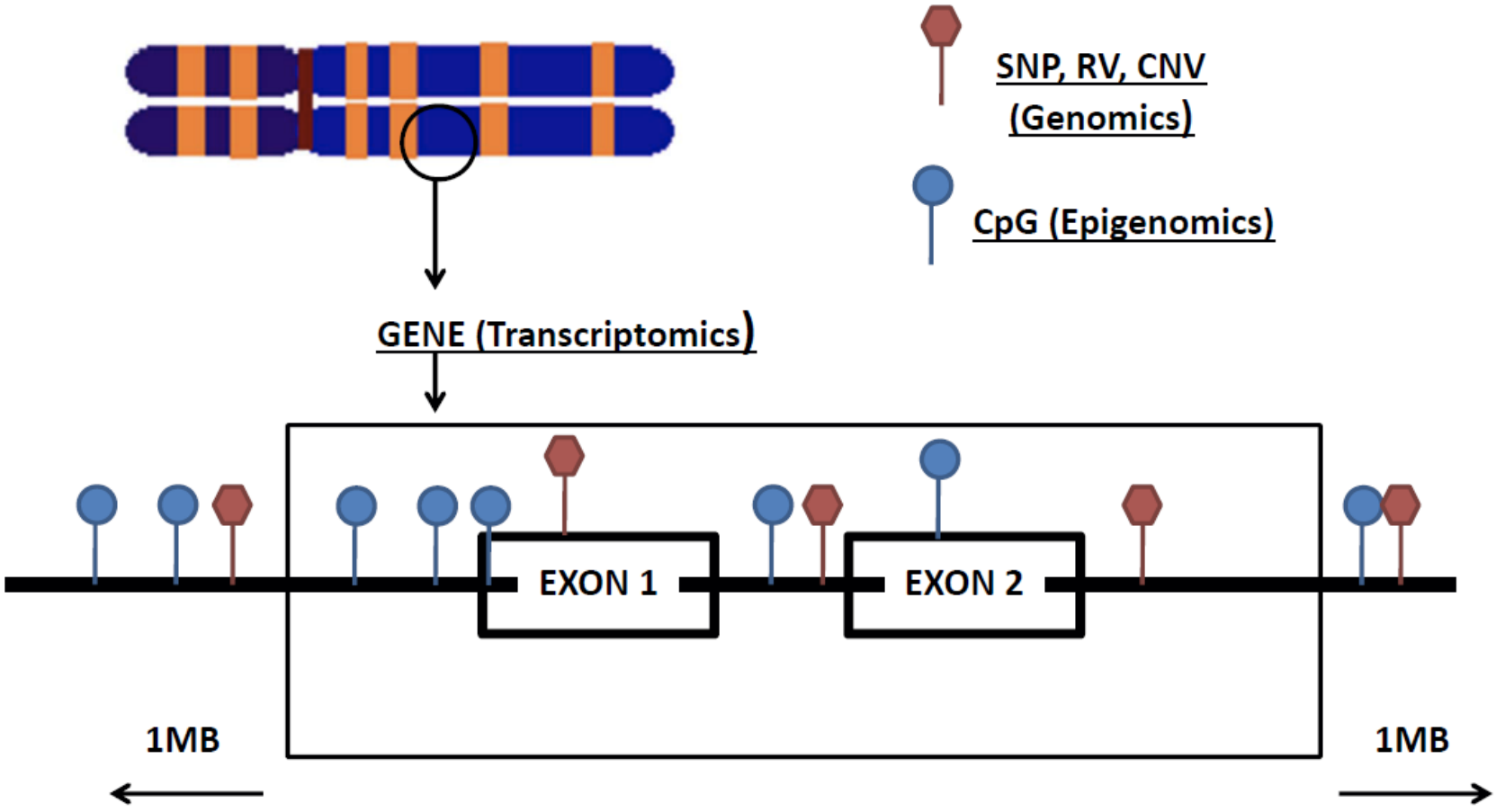
Item	Our label
Speed controller	Gamma MaxT (Van Lishout et al.2014 – submitted)
Population / patient substructure or (cryptic) relatedness chart	MB-MDR for structured populations (Van Lishout et al. 2013 – poster ASHG, manuscript in preparation)
Heterogeneous and correlated input features map	Component-based 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)





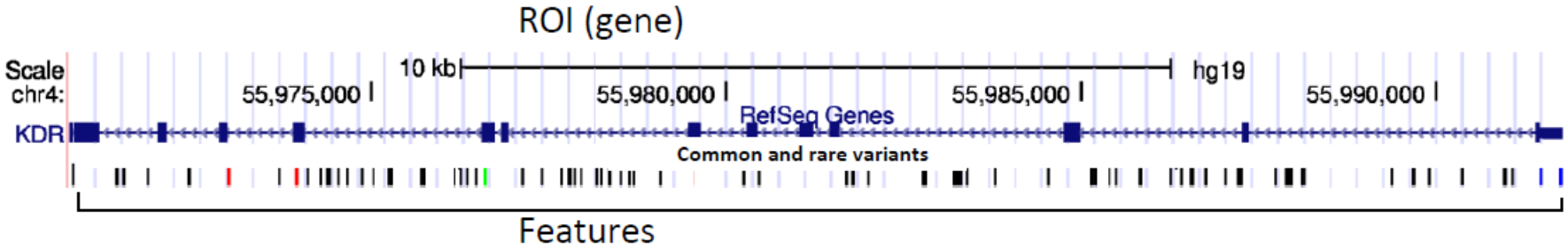
Genes have different faces



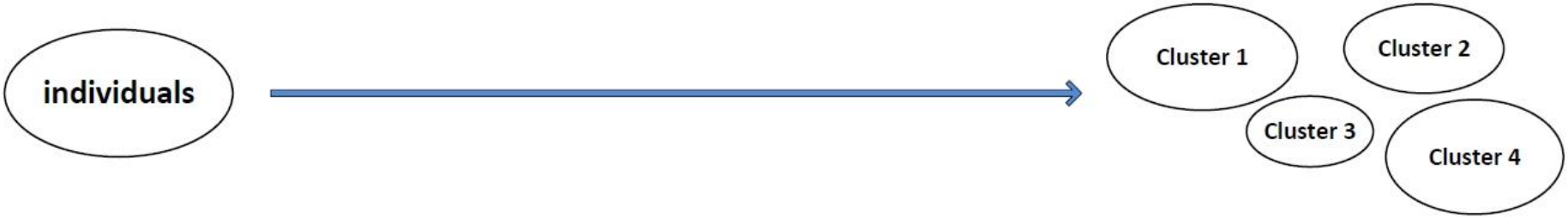
(Slide S Pineda – lab meeting 2014)

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

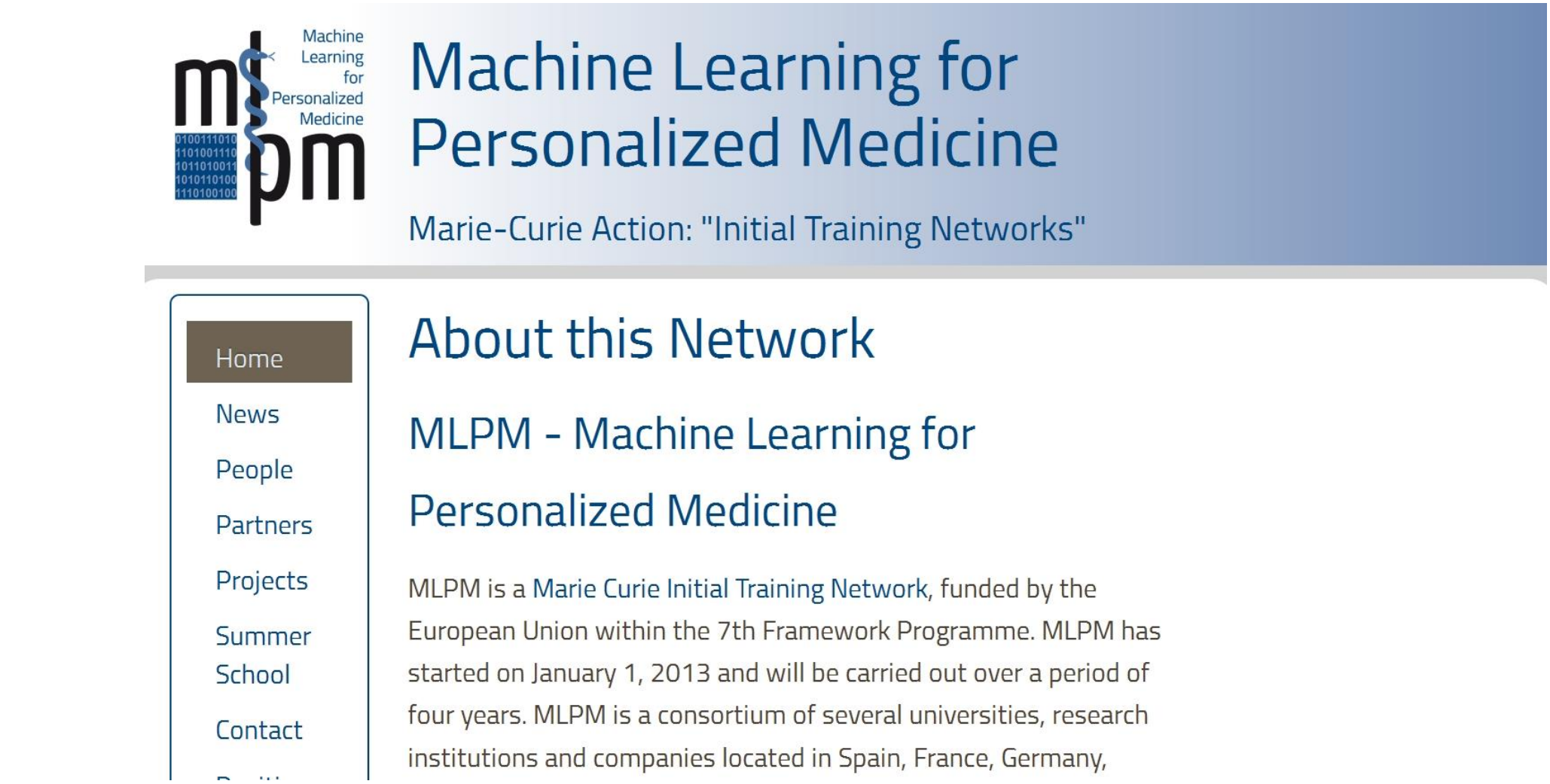
- **Phase 1:** Select sets of interest (ROI) / Prepare the data



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

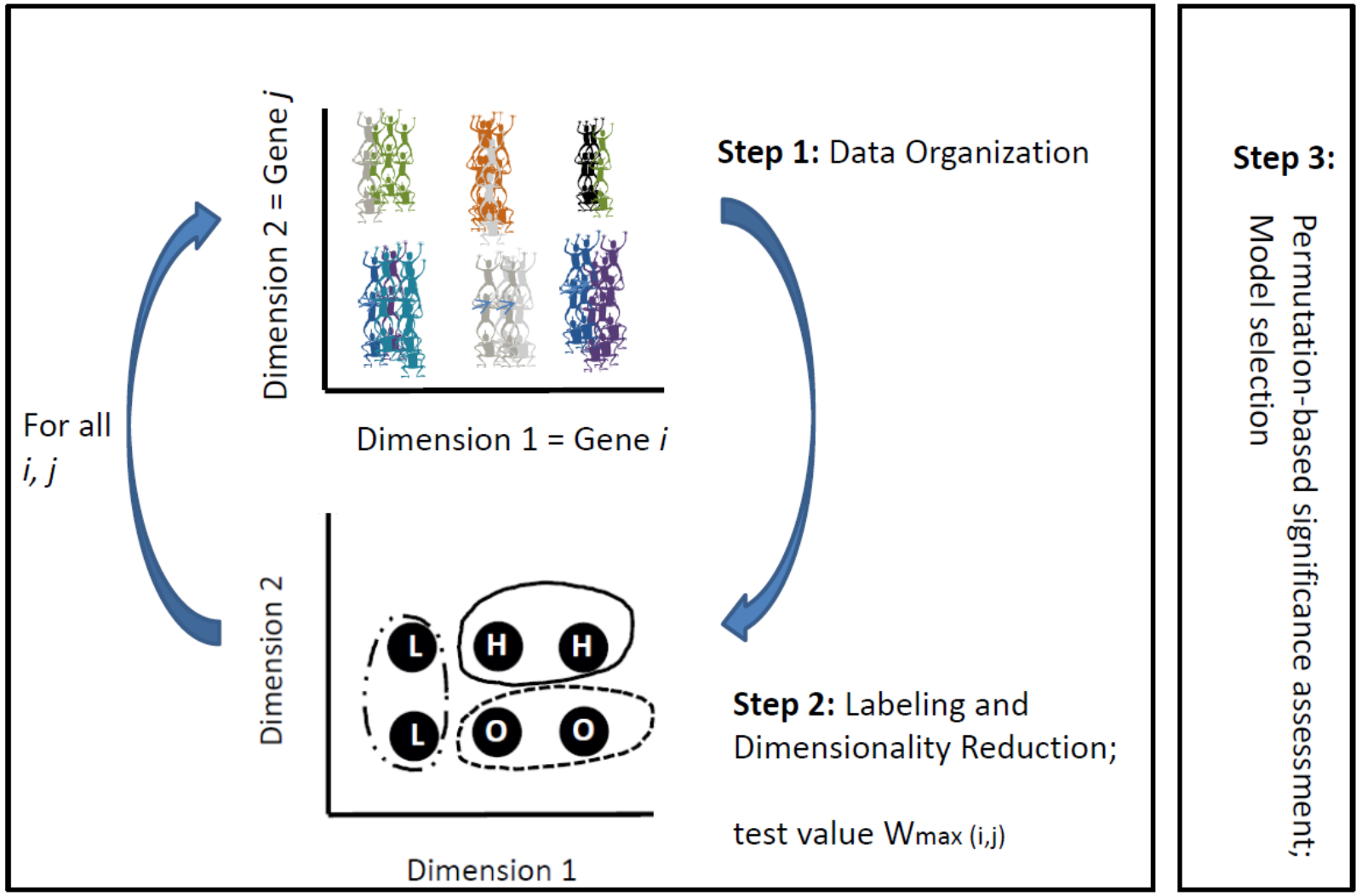


- **Phase 3:** Application of classic MB-MDR on new constructs


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/)

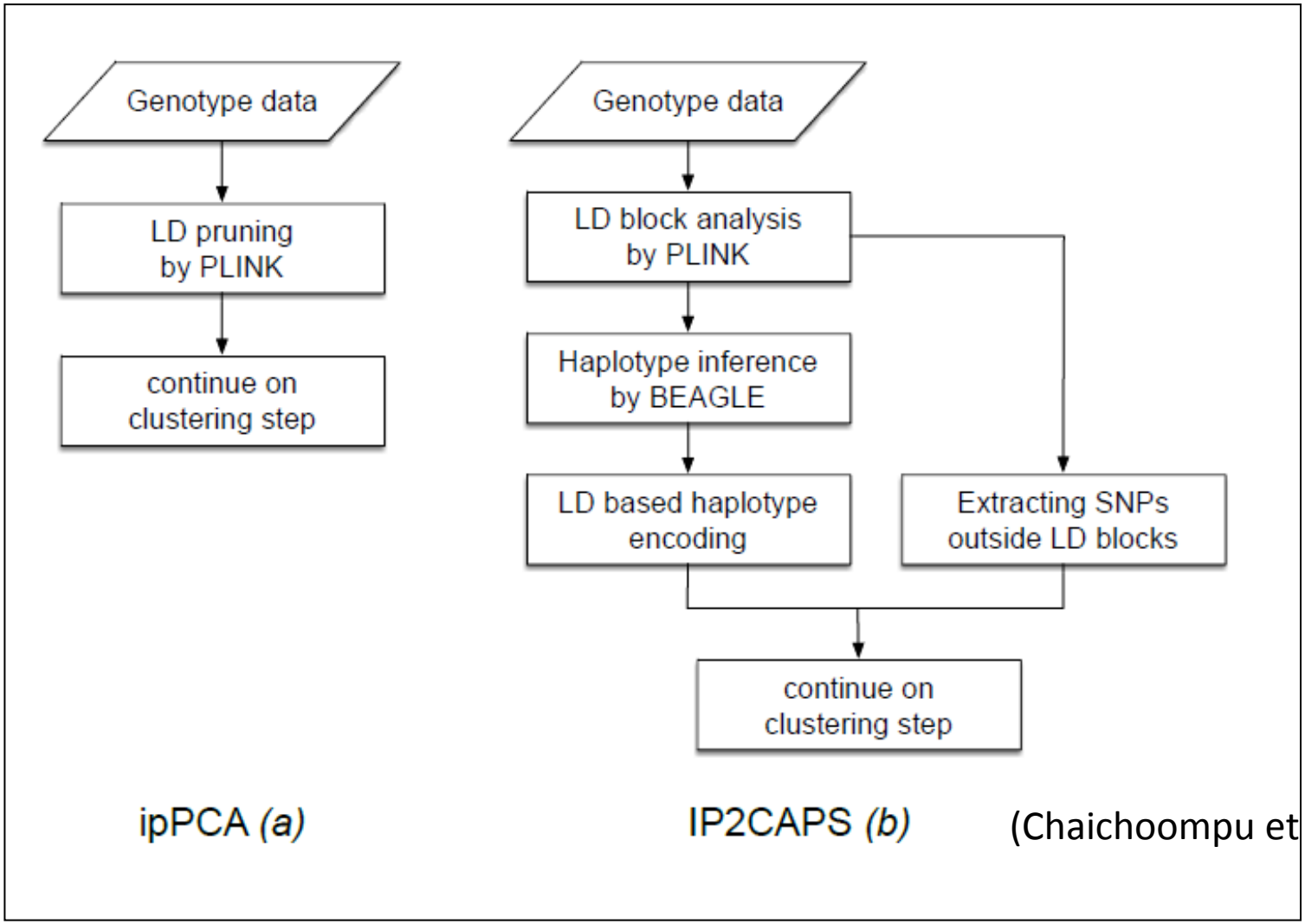
Bonus: gene-based statistical interaction networks



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)?

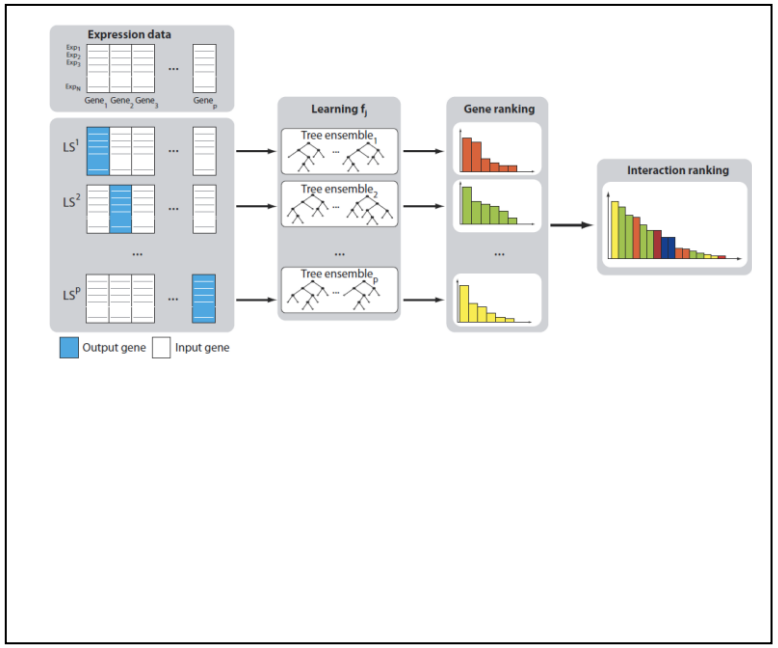
IP2CAPS → integrative fine structure detection



(Chaichoompu et al.2015 - submitted)

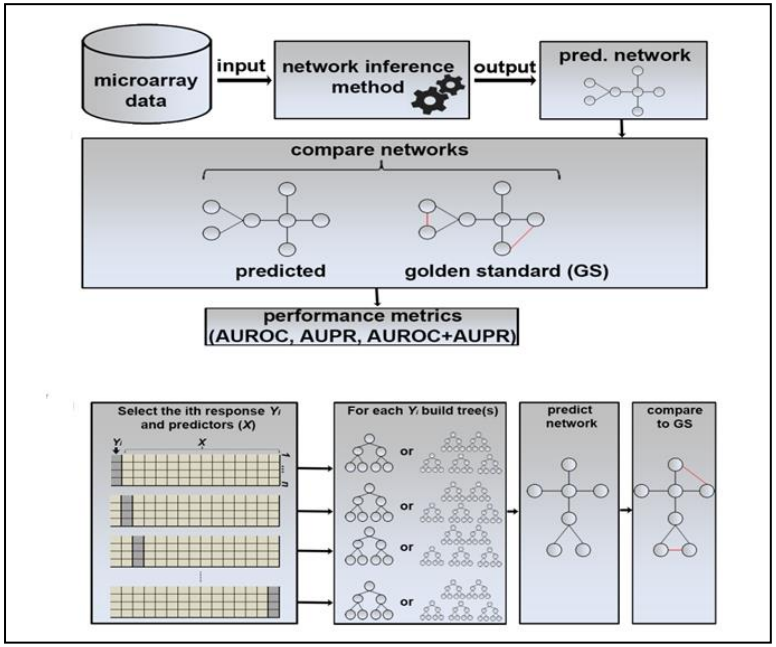
STEP 9: Can we learn from cross-disciplinary marriages?

- 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)



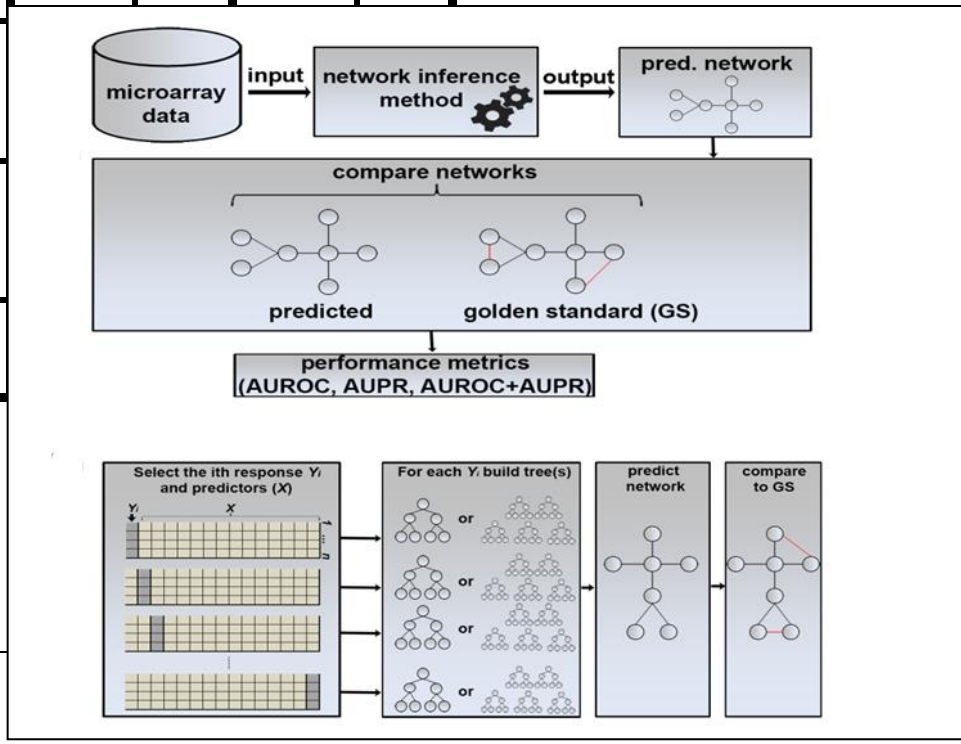
A Dialogue for Reverse Engineering Assessments and Methods

(http://dreamchallenges.org/)

Unbiased Gene Regulatory Network Inference via CIF

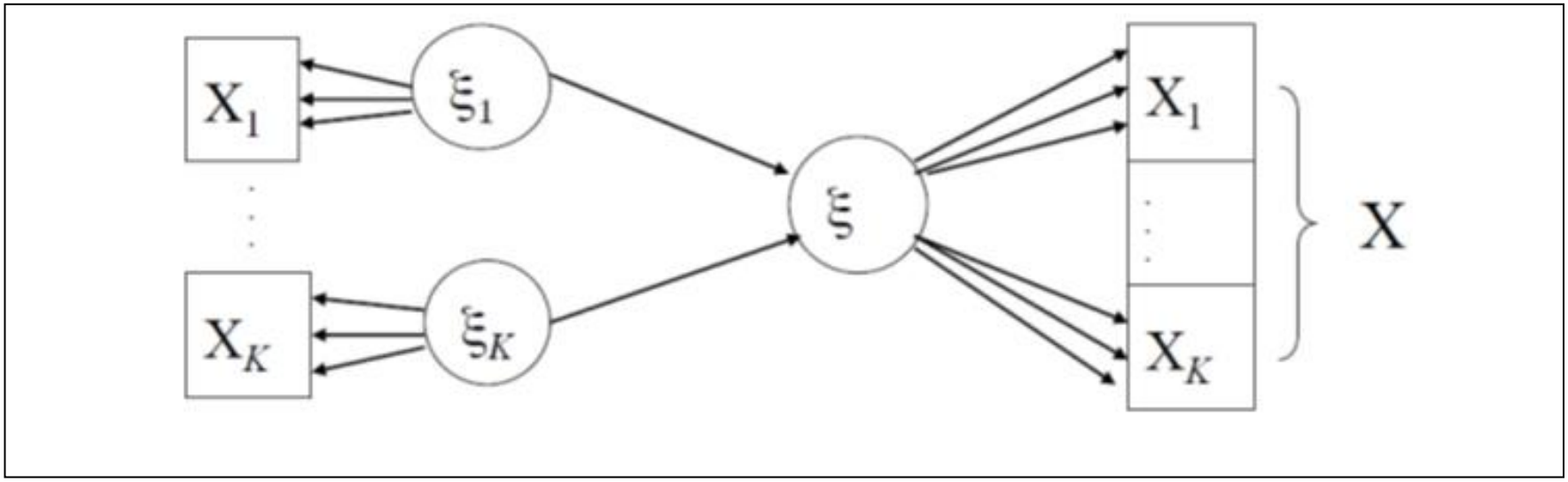
Methodology	DREAM4 Overall score	Network 1		Network 2		Network 3		Network 4		Network 5	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
CIT	7.961	0.554	0.080	0.546	0.063	0.566	0.065	0.565	0.074	0.577	0.069
CIF	30.323	0.677	0.132	0.699	0.135	0.743	0.194	0.7315	0.182	0.755	0.185
CIFcond	30.861	0.665	0.130	0.692	0.136	0.740	0.201	0.7306	0.188	**0.763**	**0.202**
CIFmean (test stat)	28.111	0.704	0.106	**0.708**	**0.151**	**0.753**	0.175				
CIFmean (*p*-value)	28.926	**0.723**	**0.138**	0.702	0.148	0.732	0.184				
RF	**30.927**	0.659	0.142	0.682	0.133	0.735	**0.229**				

(5 synthetic networks ; 100 genes;
overall score: 1/2 * (AUROC score + AUPR score) - Marbach et al. 2010)



CIF-based networks → integrative gene networks

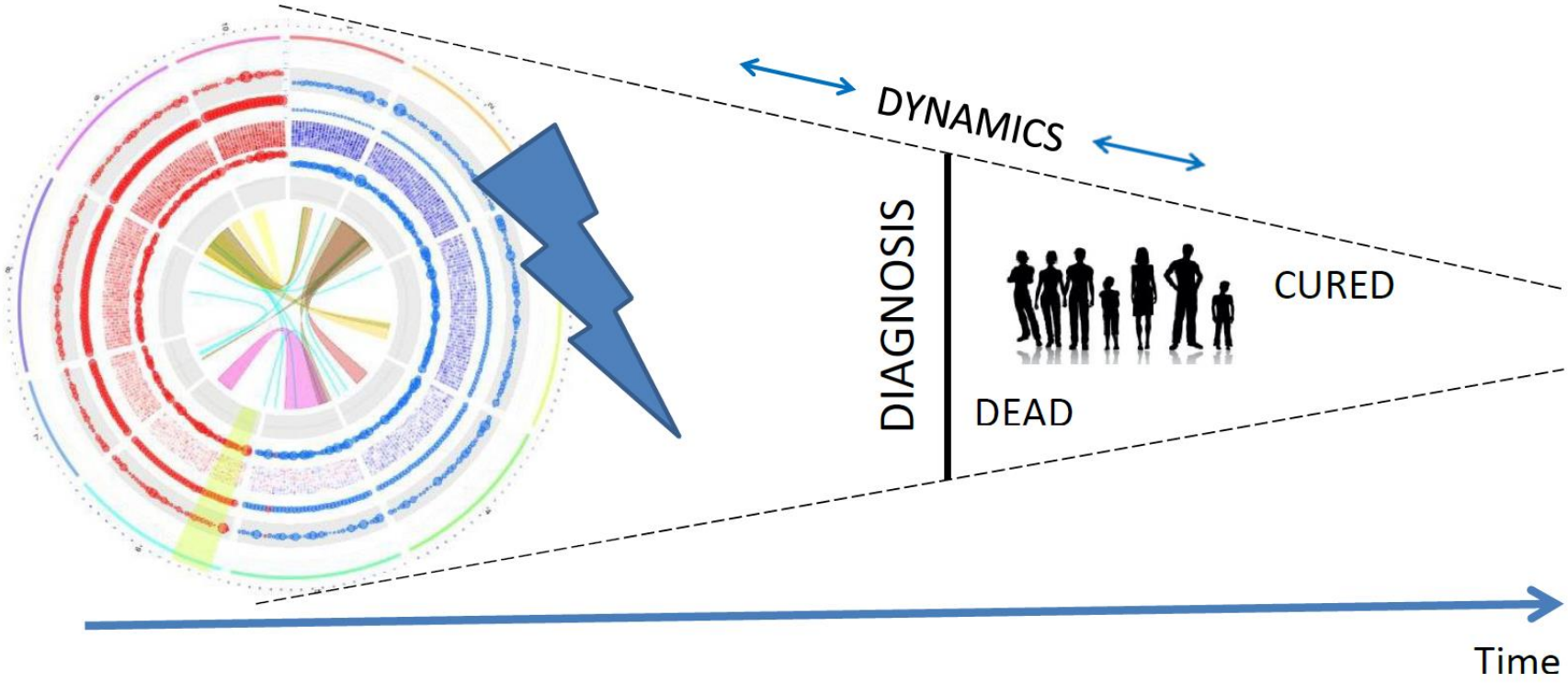
- Reduce computation complexity by building multiple CIF-based networks and fusing them into a single integrated network
- Acknowledge within-"gene" structures using component-based path modelling or kernel theory



Work in progress …..

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)

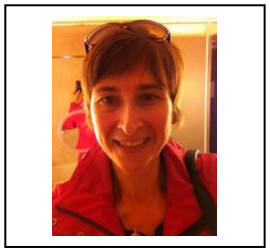
In conclusion

- 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

Appendix A

Interaction network approaches

for complex diseases

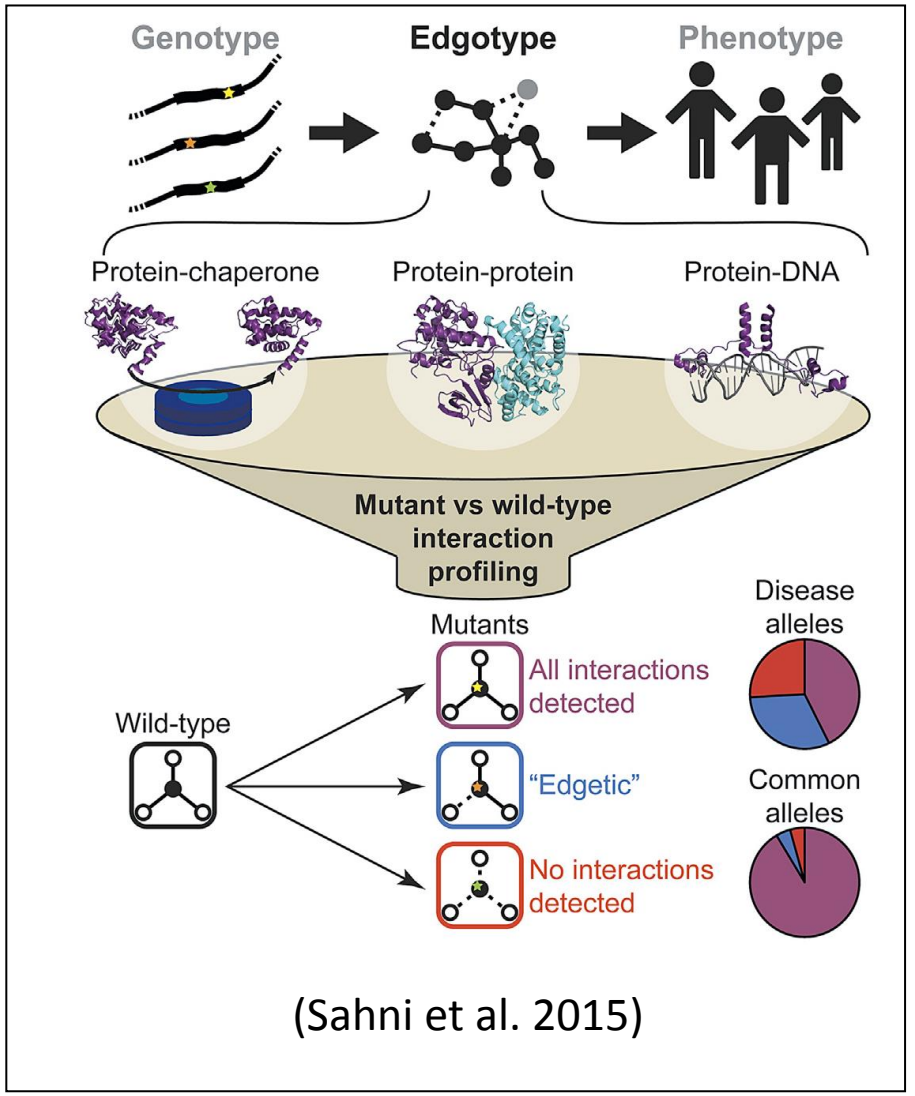
The interactome

- Classical ‘one-gene/one-disease’ models cannot fully reconcile with the increasingly appreciated prevalence of complicated genotype-to-phenotype associations in human disease.
- Genes and gene products function not in isolation but as components of intricate networks of macromolecules (DNA, RNA, or proteins) and metabolites linked through biochemical or physical interactions.
- These interactions are represented in **interactome network models** as ‘nodes’ and ‘edges’, respectively.

(Sahni et al. 2013)

The interactome

- Most missense disease mutations appear not to impair protein folding or stability
- Interaction profiling helps distinguish disease mutations from non-disease variants
- Distinct interaction perturbations underlie distinct disease phenotypes
- Integrative interaction networks enhance genotype-to-phenotype understanding



(Sahni et al. 2015)

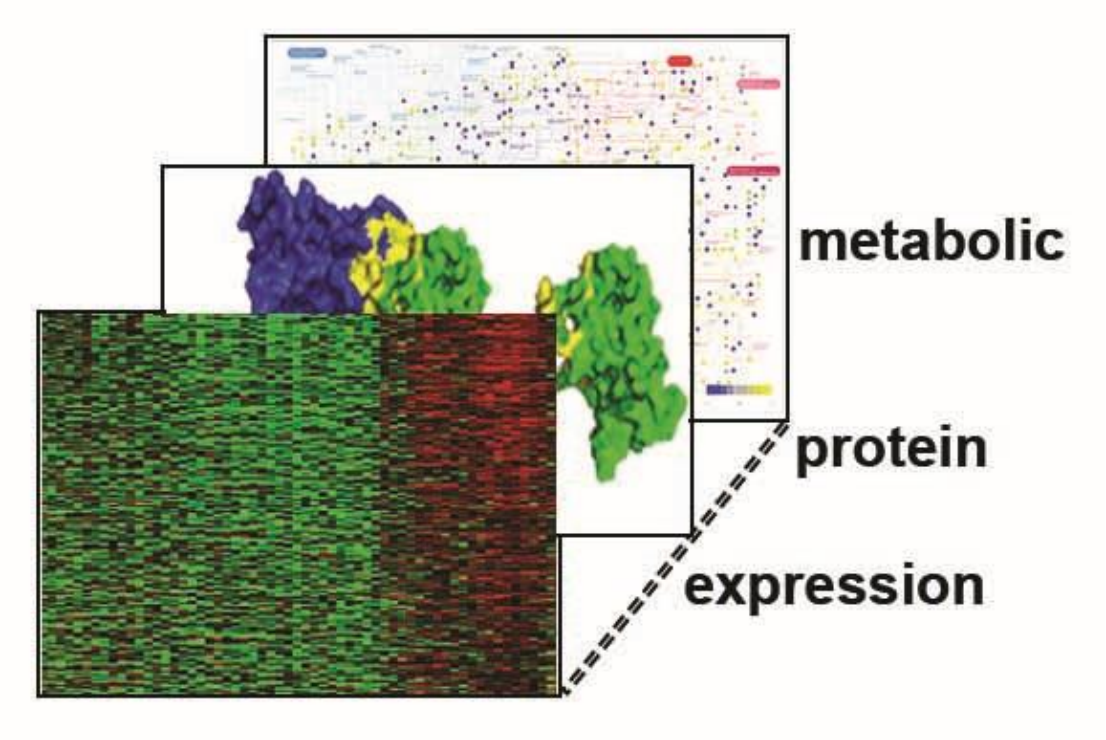
Biological networks

- One of the basic properties observed in many biological networks is the scale-free property (Albert 2005).
- A scale free network is defined as a network whose node degree distribution follows a power law: $P(k)= \alpha k^{-\gamma}$

  [P(k) is the fraction of nodes interacting with k other nodes in the network, α is a normalization constant, the degree exponent γ usually satisfies $2< \gamma< 3$]
- In biological networks the scale free property holds only approx. and practically the most important implication of this observation is the fact that these networks are characterized by a small number of highly connected nodes while most nodes interact with only a few neighbors. These “hubs” have been proposed to play important roles in biological processes (Jeong et al. 2001)

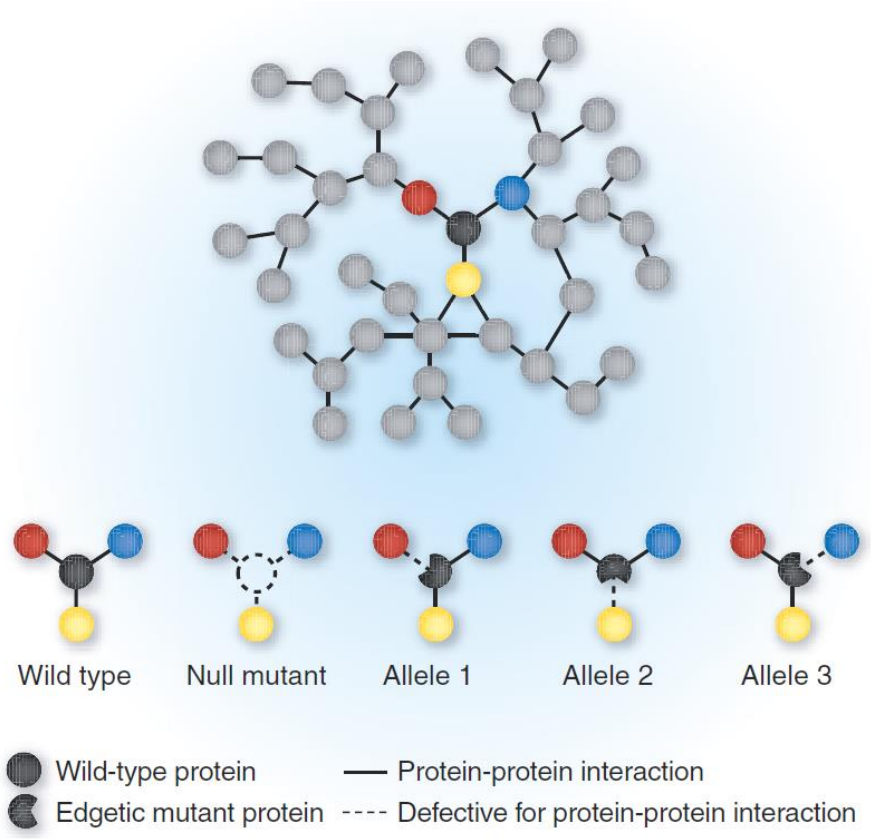
Biological networks

- Pair-wise interaction of biological entities:



Biological networks – "wild-type" networks

- Example: protein-protein interactions



(Costanzo et al. 2009)

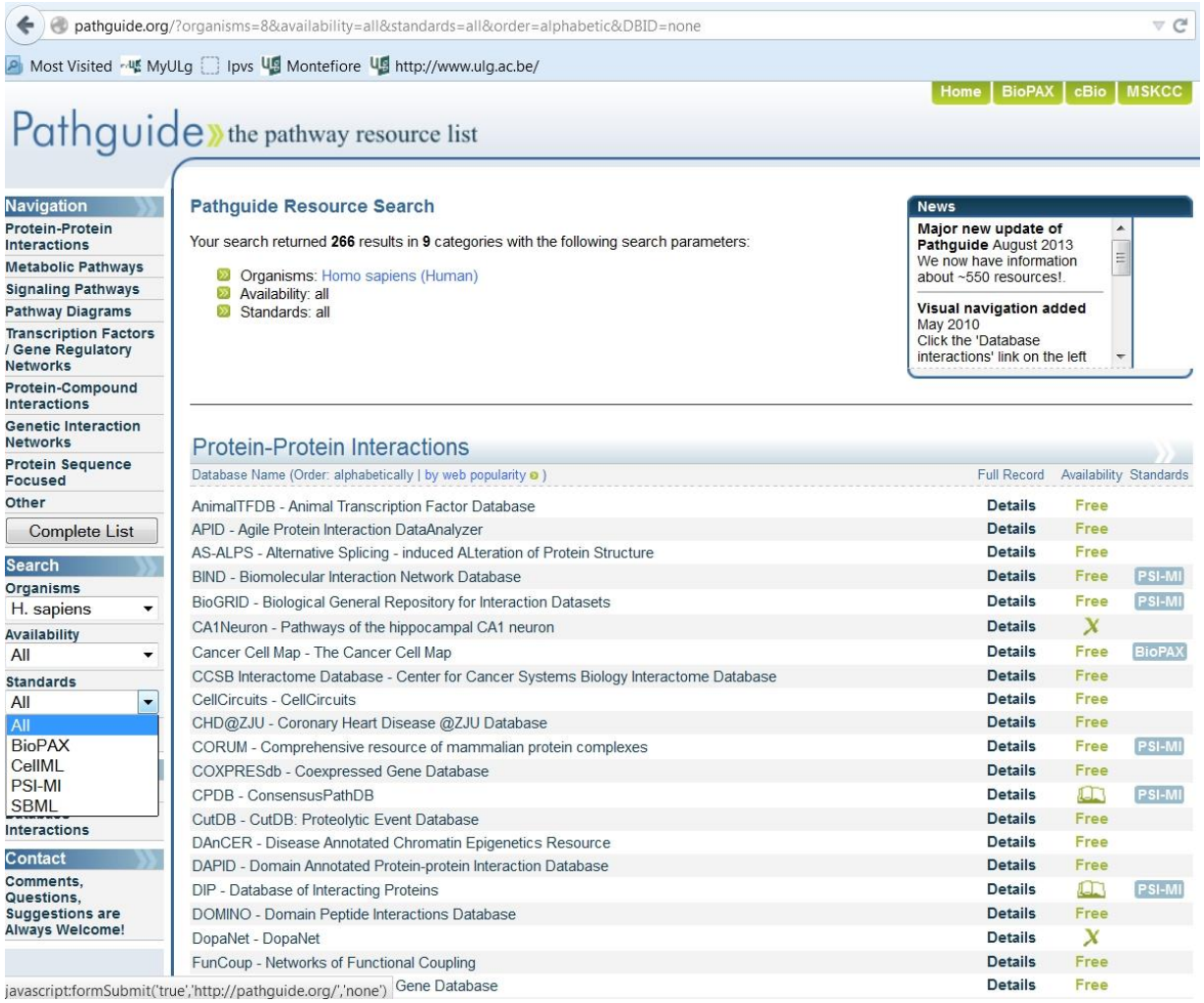
Physical interaction versus functional interaction

- **Physical interaction networks** obtained by high-throughput techniques are found to include numerous non-functional interactions (Levy et al 2009), at the same time many true interactions may be missed

- **Functional networks** aim to connect genes with similar or related functions even if they do not necessarily physically interact
- Similarly, **functional regulatory networks** are constructed so that the interactions depict direct or indirect regulatory relationships

(Cho et al. 2012)

Resources



(http://www.pathguide.org/)

The algorithmic landscape

Published in final edited form as:
Nat Rev Genet. 2013 May ; 14(5): 333–346. doi:10.1038/nrg3433.

Computational solutions for omics data

Bonnie Berger[1,2], **Jian Peng**[2], and **Mona Singh**[3]

Bonnie Berger: bab@mit.edu

[1]Department of Mathematics and Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[3]Department of Computer Science and the Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08542, USA

Hypotheses in network medicine (Barabási et al. 2011)

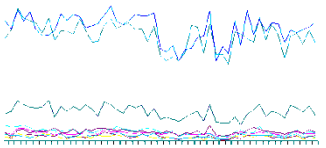
- *Hubs:* Non-essential disease genes (representing the majority of all known disease genes) tend to avoid hubs and segregate at the functional periphery of the interactome. *In utero* essential genes tend to associated with hubs
- *Disease module hypothesis:* Cellular components associated with a specific disease phenotype show a tendency to cluster in the same network neighborhood
- *Network parsimony principle*: Causal molecular pathways often coincide with the shortest molecular paths between known disease-associated components
- *Shared components hypothesis:* Diseases that share disease-associated cellular components (genes, proteins, metabolites, miRNAs) show phenotypic similarity and comorbidity

Analytics 1: Co-expression networks – gene expression

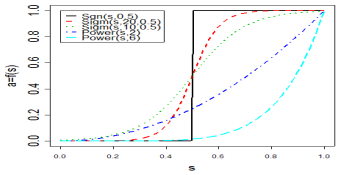
- Since functionally related genes are likely to show mutual dependence in their expression patterns (Eisen et al. 1998), it is natural to look at gene expression to detect functional relationships
- This has led to a variety of methods to construct gene co-expression networks, differing in their way they define mutual dependence, or carry out follow-up analyses

Weighted Gene Cco-expression Network Analysis

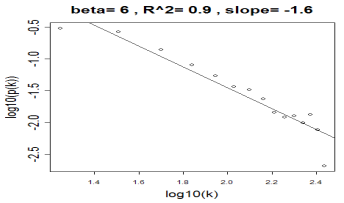
Define a Gene Co-expression Similarity



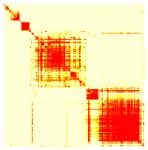
Define a Family of Adjacency Functions



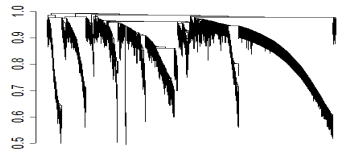
Determine the AF Parameters



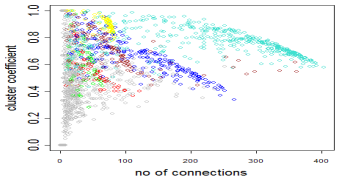
Define a Measure of Node Dissimilarity



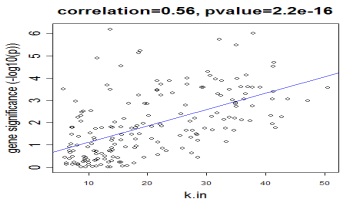
Identify Network Modules (Clustering)



Relate Network Concepts to Each Other



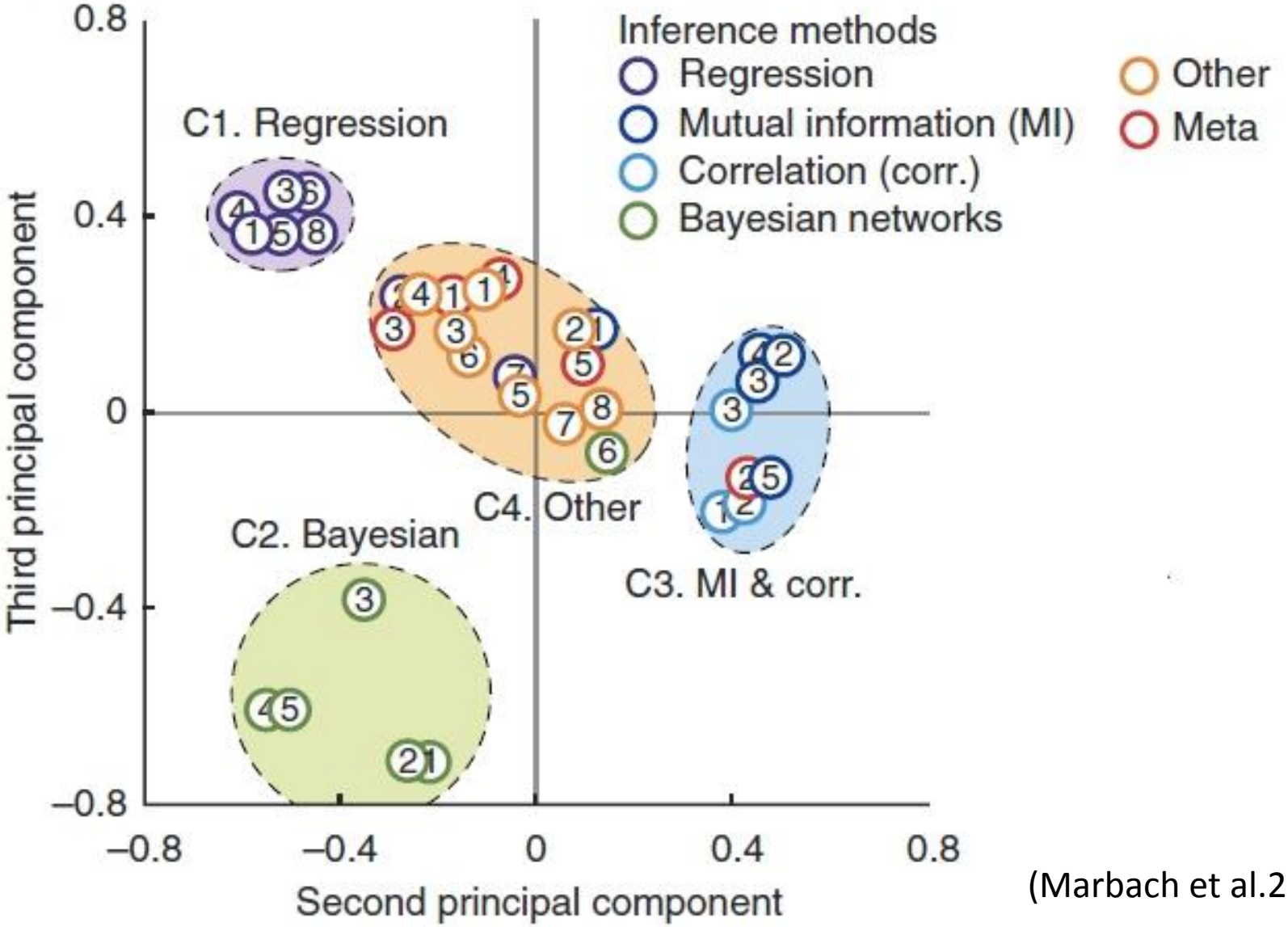
Relate the Network Concepts to External Gene or Sample Information



(Horvath)

Analytics 2: Gene regulatory networks – homogeneous samples

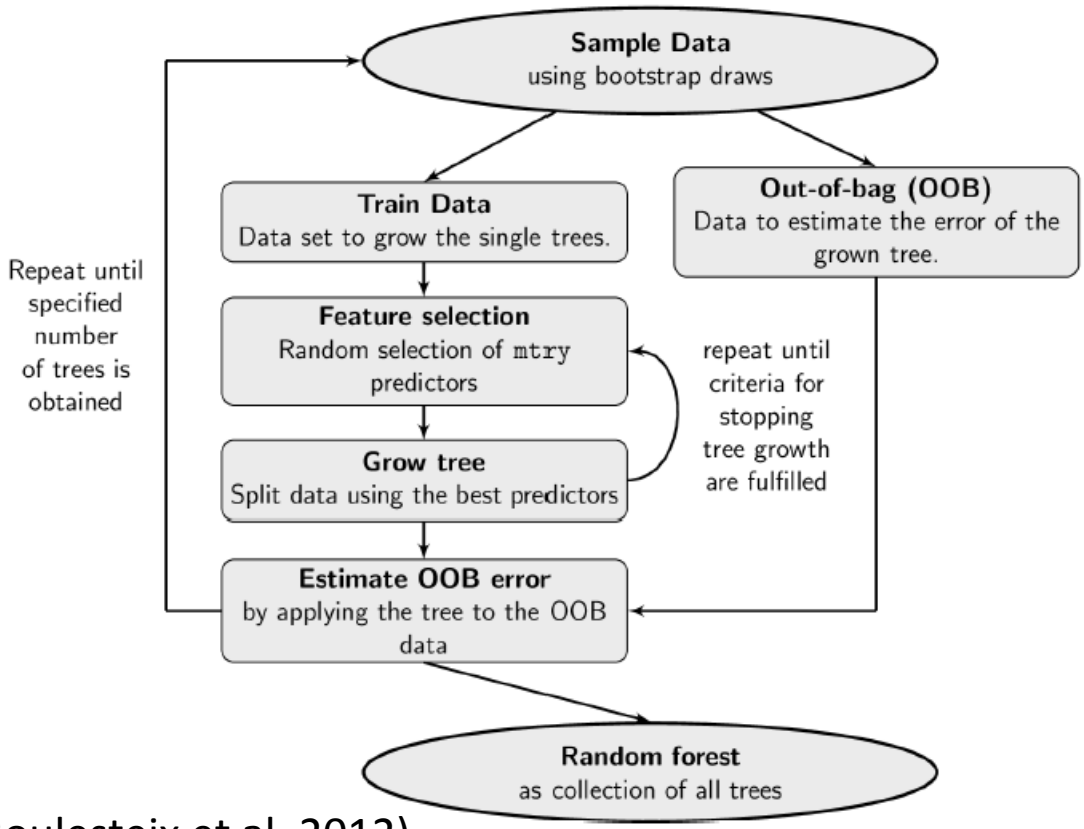
- **Regression**: transcription factors are selected by target gene–specific (i) sparse linear-regression and (ii) data-resampling approaches
- **Mutual information & correlation**: edges are (i) ranked based on variants of mutual information and (ii) filtered for causal relationships, or are ranked based on variants of correlation [see before]
- **Bayesian networks**: optimize posterior probabilities by different heuristic searches
- **Other approaches**: network inference by heterogeneous and novel methods such as RF-based edge prediction often lead to similar prediction results based on (i) applying multiple inference approaches and (ii) computing aggregate scores (i.e., **meta-predicting**)



(Marbach et al.2012)

Analytics 2: "Other" - Conditional Inference Forests

- Based on the Random Forests algorithm



(Boulesteix et al. 2012)

- In the original RF method suggested by Breiman et al. (1984), trees are built using the Decrease of Gini Impurity (DGI) as a splitting criterion.
- The splitting predictor comes from a randomly selected subset (different at each split).
- In particular, each tree is constructed from a bootstrap sample drawn with replacement from the original data set; all predictions are aggregated through majority voting.

Measures of importance in RFs

- **Gini VIM** is subject to the same **bias in favor of variables with many categories and continuous variables** or **highly unbalanced variables**, that affects variable selection in single trees, and also to a new source of bias induced by the **resampling scheme** (Strobl et al. 2007)

(sum of the DGI criteria of the splits that are based on the predictor, scaled by the total number of trees in the forest)

- The **permutation VIM** is a reliable measure of variable importance for uncorrelated predictors when subsampling without replacement – instead of bootstrap sampling – and unbiased trees are used in the construction of the forest. For **correlated** predictors it will **overestimate** (Strobl et al. 2007 + technical report, Party on!)

(difference between the OOB error resulting from a data set obtained through random permutation of the predictor of interest and the OOB error resulting from the original data)

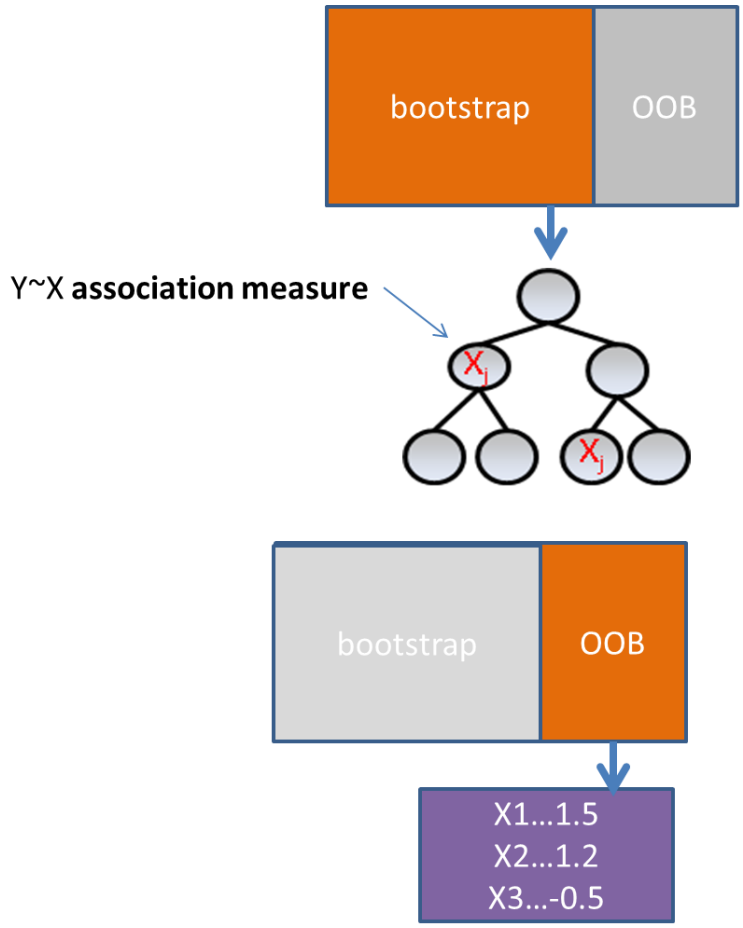
Conditional Inference Forests (short: CIFs)

(Hothorn et al. 2006 and Strobl et al. 2007)

- The CIF algorithm, overcomes a few "issues" with classic RF
- Its primary difference with the RF algorithm is that it **does not use the decrease of Gini impurity as a splitting criterion**
- Rather it uses principles of "conditional" hypothesis testing (explaining its name); at each split, each candidate is globally tested for its association with the response and a **(conditional) p-value** is computed
- Variable importance in CIF: **Permutation-based VIMs**

CIFs

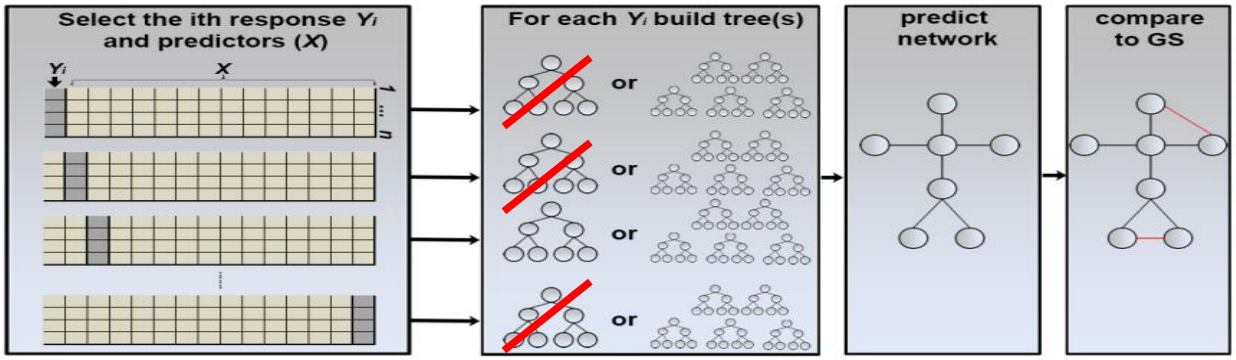
- Build tree(s) based on boot-strap sample (~0.632% of samples):
  For node *i*
  - decide if data is split further based on global independence test
  - select X_j based on global independence test
  - Split X_j based on maximally ranked statistics
- Ensemble prediction over the forest
- OOB samples to find **VIM** for $X_{1..n}$
  - Normal permutations of *X* values **(CIF)**
  - Conditional permutation of *X* values **(CIF_{cond})**



(Benjamin Dizier 2014)

CIFs to build functional regulatory networks (K Bessonov 2014+)

- **1 omics**: gene expression
- Let Y_i be a response of the i-th gene
- Let X_j be a predictor $X_j = \{X_1..X_{k-1}\}$ (k: total number of genes, $j \neq i$)



- Decision rule to connect or not to connect two genes (e.g., based on permutation-based VIMs; CIF and CIF_{cond}; directed or undirected)

(FNRS PDR grant “foresting in integromics”)

$\mathbf{CIF_{mean}}$

- Direct generalization of Conditional Inference "Trees" (in which you only have p-values attached to a node, derived from the "global association test")
- For each tree (predicting Y_i) and node X_j, extract the node's CIF p-value (when X_j appears multiple times in a tree, take the p-value based on the largest sample size)
- For each pair (Y_i,X_j), compute the element a_{ij} of the adjacency matrix (a_{ij} in [0,1]):

$$a_{ij} = \frac{\sum_{j=1}^{k} p_{X_j}}{\#\ X_j}$$

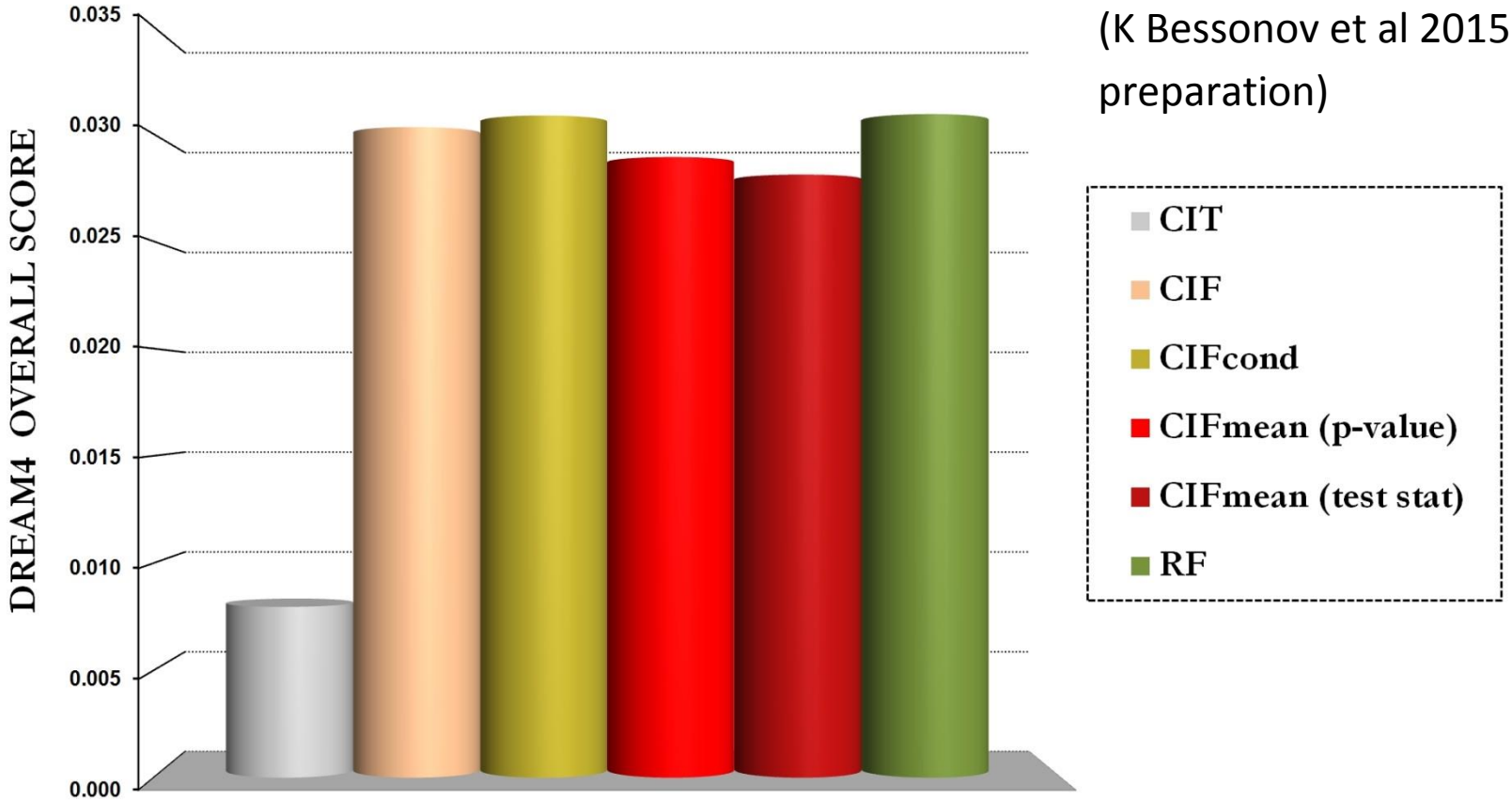
(a_{ij} is edge weight; a_{ij} =0 is equivalent to no edge)

- $\mathbf{CIF_{mean}} < \mathbf{CIF_{cond}}$, but significant computational advantage

Advantages of CIFs

- Splitting is based on an unbiased splitting criterion that automatically adjusts for different marginal distributions of the predictors and thus does not share the above pitfall:
  - No bias in selection that would be the result of having candidate predictors with large numbers of candidate splits, different scales or unbalancedness
- CIF has the option to use **permutation VIM** or **Conditional Permutation VIM** (cond) (see discussion about RF VIMs):
  - No issues related to correlated predictors
- In addition to standard regression and classification, problems, the CIF methodology also directly addresses the case of censored survival response variables

Performance of CIF-based functional regulatory networks



(K Bessonov et al 2015 – in preparation)

(DREAM 4, 100 nodes; overall score: 1/2 * (AUROC score + AUPR score);Marbach et al. 2010)

- CIF_{cond} very close to RF performance; but CIF-based ~ increased stability

Disadvantages of CIFs include

- Tends to miss non-linear associations since CIF p-values per node reflect linear association only
  (key statistic in original implementation of CIF: Strasser and Weber 1999)
- Relies on outcome variable transformations for splitting, resulting in a tendency towards splits of linearly associated variables at the median
- Conditional permutation scheme is computationally intensive

Possible extensions (B Dizier 2015+)

Decision	CIF	New method
Is there enough evidence of association to split data further?	Global independence test	Goeman's Global Test (Goeman et al.2004)
Which variable to use?		MaxT
Which split in this variable?	Maximally ranked statistics	MaxT

- First simulation results look promising in terms of power and false positive control … (work in progress)

Note

- Approach 1: Build a gene-gene expression network and analyse the modules, linking them to phenotypes quite similar as in a WGCNA.
- Approach 2: Use phenotype information directly during the edge construction, hence before module identification.
- Question 1: What is the complementarity of both approaches
- Question 2: What is the impact on results of ignoring population substructure?
- Question 3: What is the impact on results of ignoring clinically relevant patient subphenotypes?

  (E.g., in WGCN a group of patients is used to build a gene co-expression network and patient's phenotypic variability is associated to particular modules)

Analytics 3: Network perturbation – genetic background / causality

- Causal analysis of expression looks at how levels of mRNA predict each other and therefore directs edges.
- However, a directed edge from node 1 → node 2 does not require node 2 to be causally dependent on node 1 (Steyvers et al. 2003)
- So-called "rooted genes" have known causes of variation (e.g., SNPs and/or other genes) and therefore can be used to construct the rest of the **causal network**.
- The same holds for "modules": sets of variables that vary coordinately across samples

Analytics 3: Network perturbation – genetic background / causality

- Because modules are comprised of many nodes, the causal association may be more significant than if we study only the association of one locus with expression of one gene.
- If we can identify sites where a defined part of the variability of the module is attributed to specific QTL, then we have the beginnings of a causal network (Lee et al. 2009 - LIRNET)

  LIRNET: A hypothesis based systematic way of identifying the loci that determine expression of a module

- If the module also correlates with a genetic disease phenotype, then it is reasonable to offer hypothesis that the disease ALSO relates from the same causal network

Our novel contribution

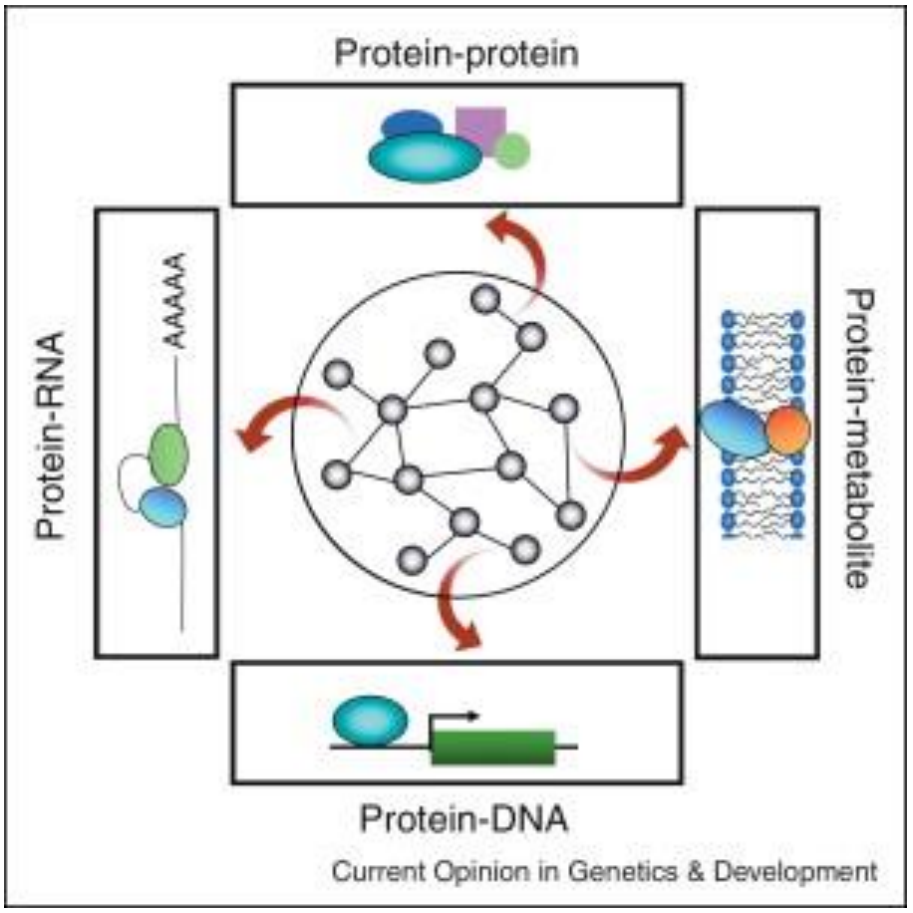
- Use genetical genomics, in particular eQTL (epistasis) analysis techniques to find groups of interacting determinants of a module

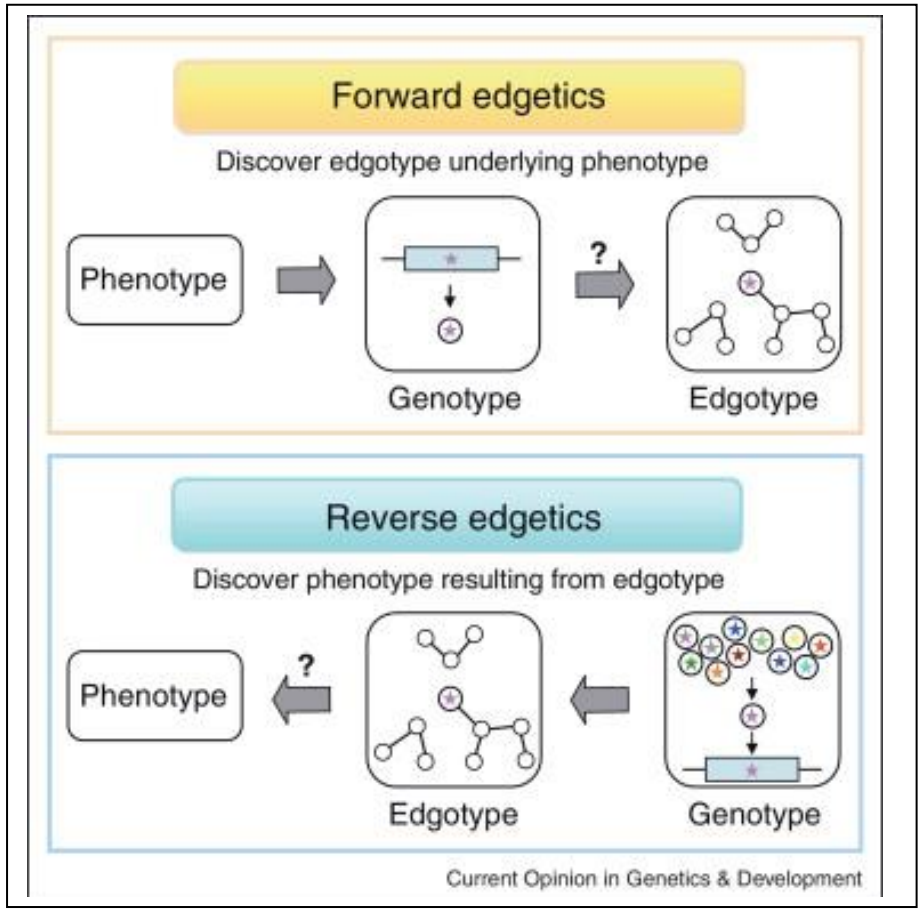
  Example of large-scale eQTL epistasis analysis using MB-MDR (Asthma, K Bessonov in collaboration with B Raby from Harvard Medical School)

- Instead of genes, look at modules (eigengenes) of a co-expression network and perform eQTL (epistasis) analysis to find genetic drivers (modifiers)
- Application of MB-MDR in the context of finding cis/trans modifier genes to target areas (windows around a significant eQTL): collaboration with B Raby (Harvard Medical School) – manuscript in preparation

Note

- What can we learn from edgetics? (Sahni et al.2013, Charloteaux et al. 2011)





Note

- Does causal mediation analysis have a role to play in the aforementioned context?

Statistical Methods for
Causal Mediation Analysis

A dissertation presented

by

Linda Valeri

to

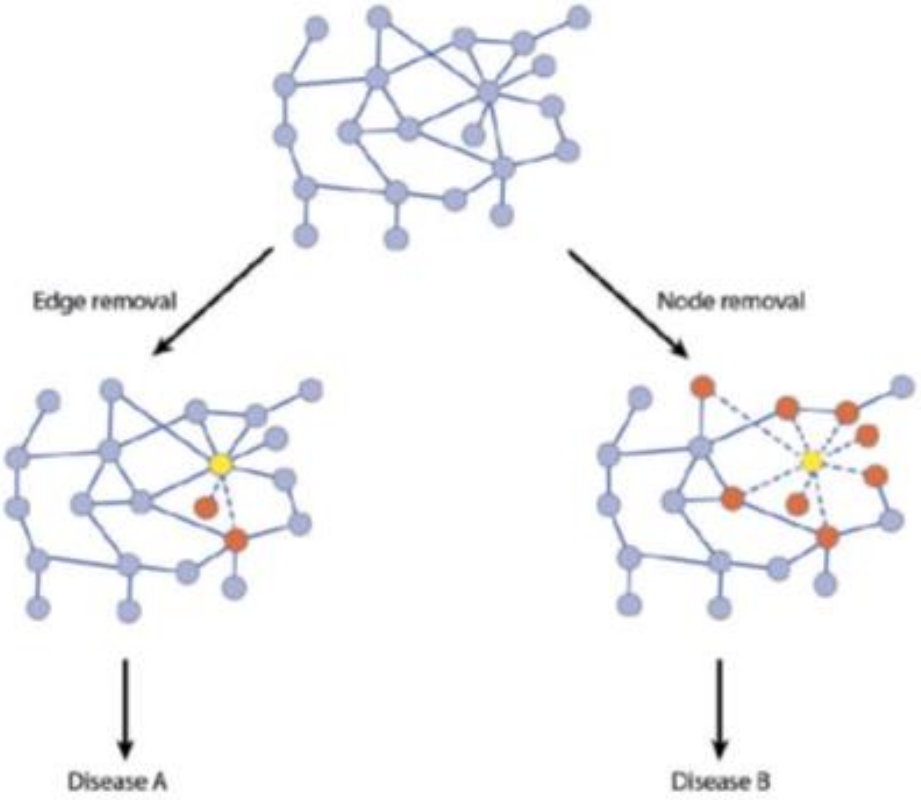
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

December 2012

Genetic background affecting network nodes or edges



(Wang et al. 2011)

Our novel contribution - complementing the physical network with a functional network

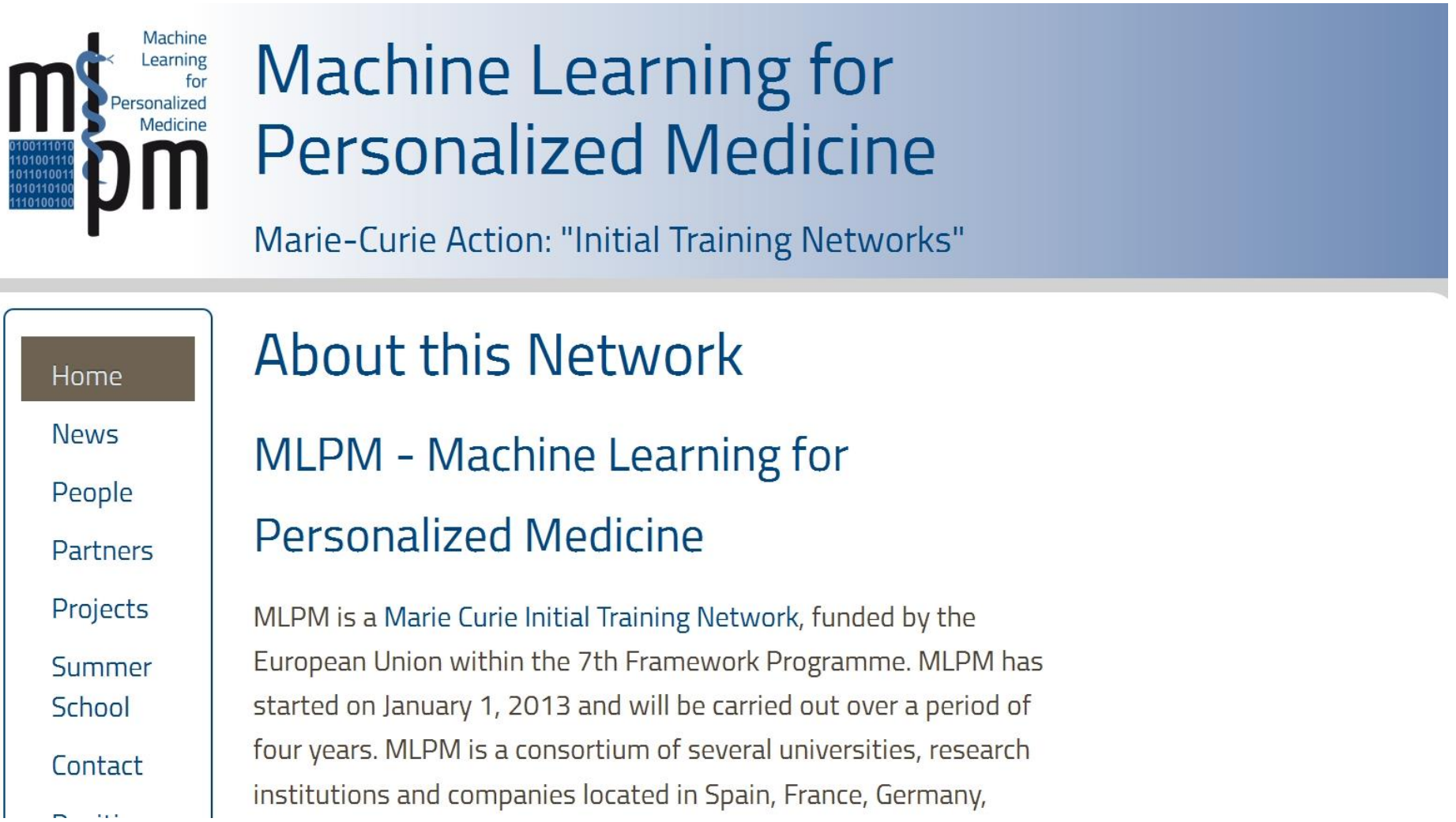
- Initially, we fused functional regulatory networks (K Bessonov, F Gadaleta)
  - Gene expression ← gene expression (e.g., CIFs, LabNet)
  - Gene expression ← SNPs (e.g. CIFs, LabNet)
- Assumes that SNPS can be mapped to genes + needs extra measures when multiple SNPs are mapped to the same gene (other solution – see later)
- Questions:
  - Can/should physical interaction networks be used as prior knowledge?
  - Can/should the fusing-based network be linked to the physical interaction network of Wang et al. 2011?

Statistical epistasis networks

- SNPs
  - Physically mapped to genes (the usual)
  - Functionally mapped to genes (the unusual but most intuitive)
- Genes
  - Aggregates based on features mappable to the gene
  - Can harbor detailed graph structure via appropriate kernels, such as diffusion kernels (Kondor and Lafferty 2002)
- MB-MDR can accommodate both as units of (statistical interaction) analysis (Cattaert et al 2011 - SNPS, Fouladi et al 2015- Genes)
- Considering a third order interaction with SNP profiles allows to investigate whether that particular background causes a loss/gain of a statistical SNP x SNP / gene x gene interaction (~Wang et al 2011)

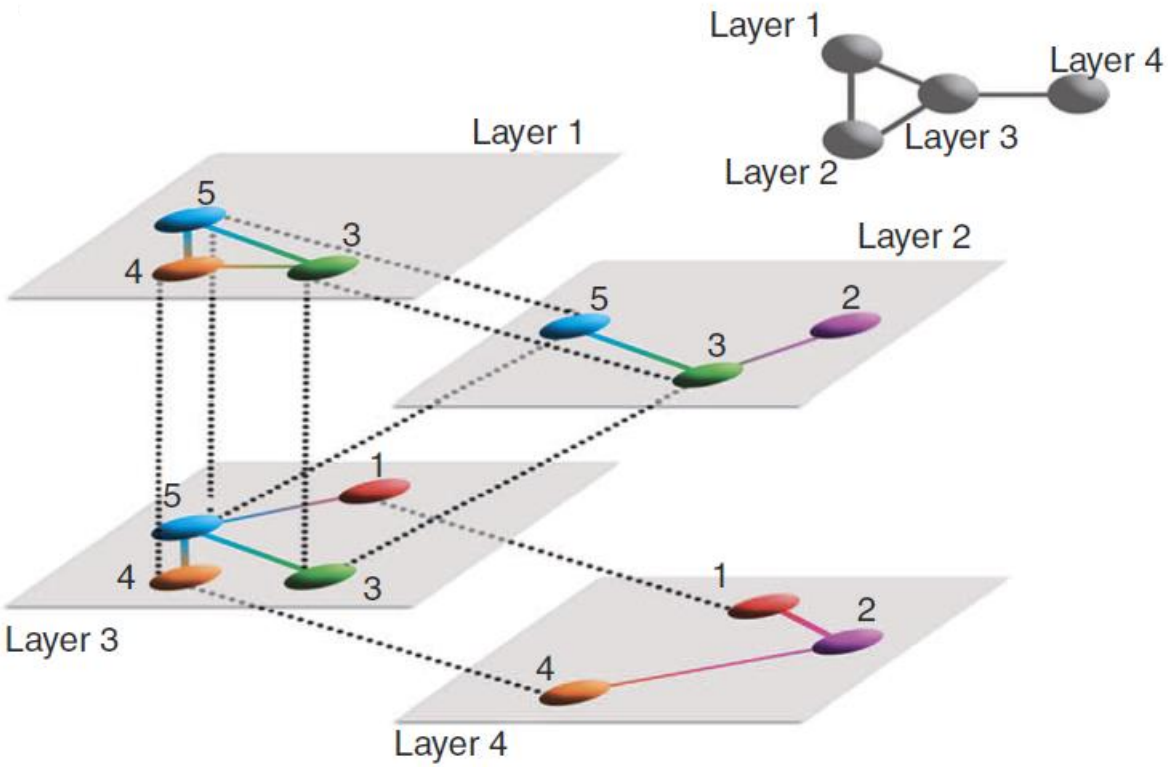
Our novel contribution

- Fusion – in progress
  - Gene expression ← gene expression (e.g., CIFs)
  - Gene x Gene based on SNPs and/or epigenetic markers (e.g., gene-based MB-MDR, incorporating disease info from the start)
  - Note:
    - Like epigenetic markers and SNPs, also (co-expressed) genes can be mapped to a gene, directly leading to a statistical gene x gene statistical interaction network without the need to "fuse".
    - Via appropriate kernels, the within-gene structure can be fully exploited (work in the pipeline – context of MLPM network coordinated by Karsten Borgwardt, in which K Van Steen is a node leader)


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/)

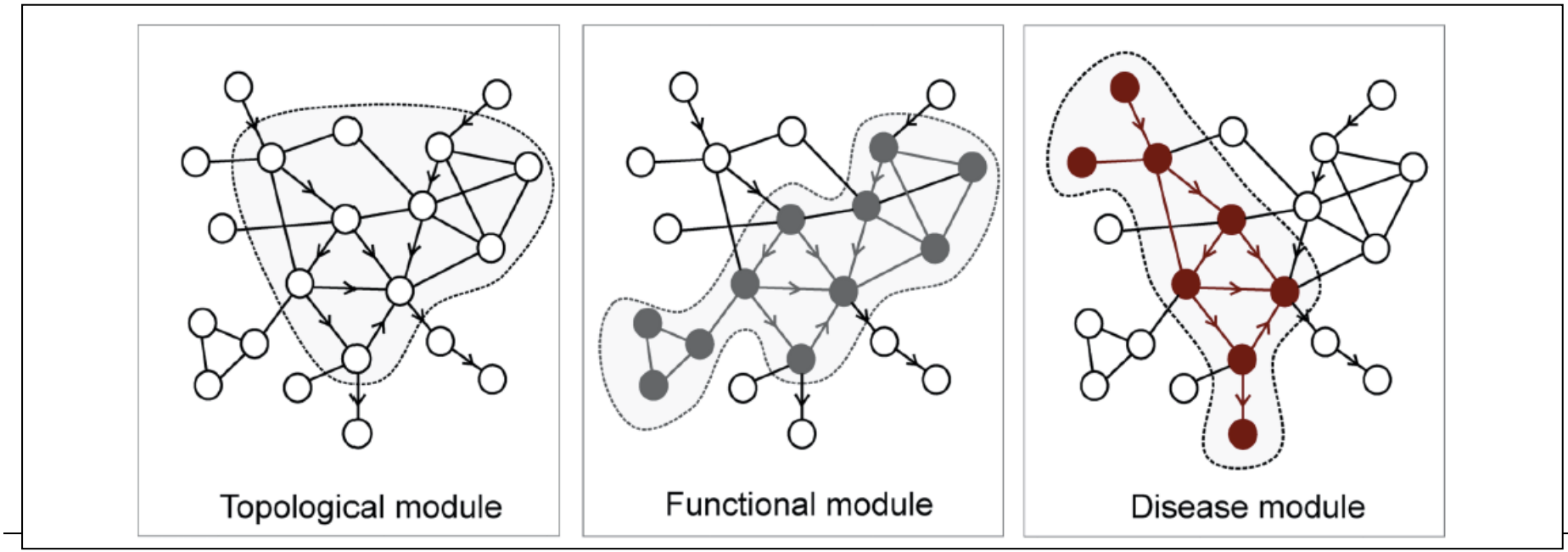
- Multilayer – in progress
  - Layer 1: gene co-expression based network (e.g., WGCN)
  - Layer 2: SNP based interaction network (MB-MDR)
  - Connection between layers: SNP-gene mapping (functional: via eQTL analyses)



(De Domenico et al. 2013)

Network medicine

- The **underlying assumption** in network medicine is that the topological, functional, and disease modules overlap so that functional modules correspond to topological modules and a disease can be viewed as the breakdown of a functional module (Barabási et al. 2011)



Network medicine

- **Topological modules** correspond to locally dense neighborhoods (i.e., a pure network property), such that the nodes of the module show a higher tendency to interact with each other than with nodes outside of the module.
- **Functional modules** correspond to network neighborhoods in which there is a statistically significant segregation of nodes of related function (e.g., via pathway analysis)
- **Disease modules** represent groups of nodes whose perturbation (mutations, deletions, copy number variations, or expression changes) can be linked to a particular disease phenotype

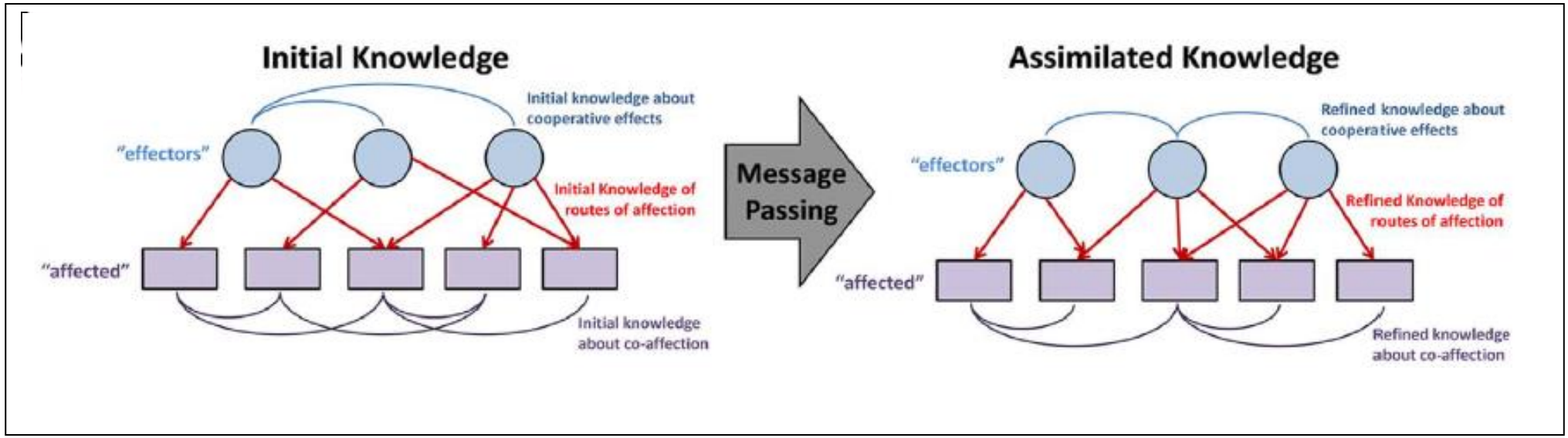
(Barabási et al. 2011)

Note 1

- De Domenico et al. (2013) developed a mathematical framework that allows computing centrality in multi-layer networks
  - Finding the ones that play the most central roles in the cohesion of the whole structure
  - Bringing together different types of relationships
- It is important to realize that each layer represents a different level of "interaction" or "relationship" between nodes and that nodes may or may not exist in all layers (in contrast to what is needed in the fusion methodology of)

Note 2

- Glass et al. (2013) developed a method that, at convergence, provides harmonized expression and interaction modules specific to a biological condition of interest, as well as the output regulatory network controlling those modules in each condition



(Glass et al.2013)

Note 2 (continued)

- The passing-messages-between-data-sources approach can be used to harmonize our integrated multilayer network (previous slides)
- Question:
  - What are the key conceptual differences between De Domenico approach and the Glass approach in identifying key players?

Appendix B

Patient / Population substructure

Subphenotyping

OPEN ACCESS Freely available online PLoS one

Molecular Reclassification of Crohn's Disease by Cluster Analysis of Genetic Variants

Isabelle Cleynen[1]*, Jestinah M. Mahachie John[2,3], Liesbet Henckaerts[4], Wouter Van Moerkercke[1], Paul Rutgeerts[1], Kristel Van Steen[2,3], Severine Vermeire[1]

1 Department of Gastroenterology, KU Leuven, Leuven, Belgium, **2** Systems and Modeling Unit, Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium, **3** Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, **4** Department of Medicine, UZ Leuven, Leuven, Belgium

Abstract

Background: Crohn's Disease (CD) has a heterogeneous presentation, and is typically classified according to extent and location of disease. The genetic susceptibility to CD is well known and genome-wide association scans (GWAS) and meta-analysis thereof have identified over 30 susceptibility loci. Except for the association between ileal CD and *NOD2* mutations, efforts in trying to link CD genetics to clinical subphenotypes have not been very successful. We hypothesized that the large number of confirmed genetic variants enables (better) classification of CD patients.

Methodology/Principal Findings: To look for genetic-based subgroups, genotyping results of 46 SNPs identified from CD GWAS were analyzed by Latent Class Analysis (LCA) in CD patients and in healthy controls. Six genetic-based subgroups were identified in CD patients, which were significantly different from the five subgroups found in healthy controls. The identified CD-specific clusters are therefore likely to contribute to disease behavior. We then looked at whether we could relate the genetic-based subgroups to the currently used clinical parameters. Although modest differences in prevalence of disease location and behavior could be observed among the CD clusters, Random Forest analysis showed that patients could not be allocated to one of the 6 genetic-based subgroups based on the typically used clinical parameters alone. This points to a poor relationship between the genetic-based subgroups and the used clinical subphenotypes.

- There is a need to adjust for overall population substructure

Subphenotyping

- Accounting for general SNP-based (rough) substructure

OPEN ACCESS Freely available online PLOS ONE

Molecular Reclassification of Crohn's Disease: A Cautionary Note on Population Stratification

Bärbel Maus[1,2*], Camille Jung[3,4,5], Jestinah M. Mahachie John[1,2], Jean-Pierre Hugot[3,4,6], Emmanuelle Génin[7,8], Kristel Van Steen[1,2]

1 UMR843, INSERM, Paris, France, **2** Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, **3** UMR843, Institut National de la Sante et de la recherche Medicale, Paris, France, **4** Service de Gastroentérologie Pédiatrique, Hôpital Robert Debré, APHP, Paris, France, **5** CRC-CRB, CHI Creteil, Creteil, France, **6** Labex Inflamex, Université Paris Diderot, Paris, France, **7** UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies, INSERM, Brest, France, **8** Centre Hospitalier Régional Universitaire de Brest, Brest, France

- Setting: 450 CD patients with phenotypic measurements; 51 CD-reported SNPs; 30 ancestry informative markers

Subphenotyping

- Similar to Price et al. (2006) a correction for population stratification is enforced by projecting the standardized genotypes G onto the space orthogonal to the space spanned by the principal components:

$$G' = I_n - A\,(A^T A)^{-1} A^T\, G$$

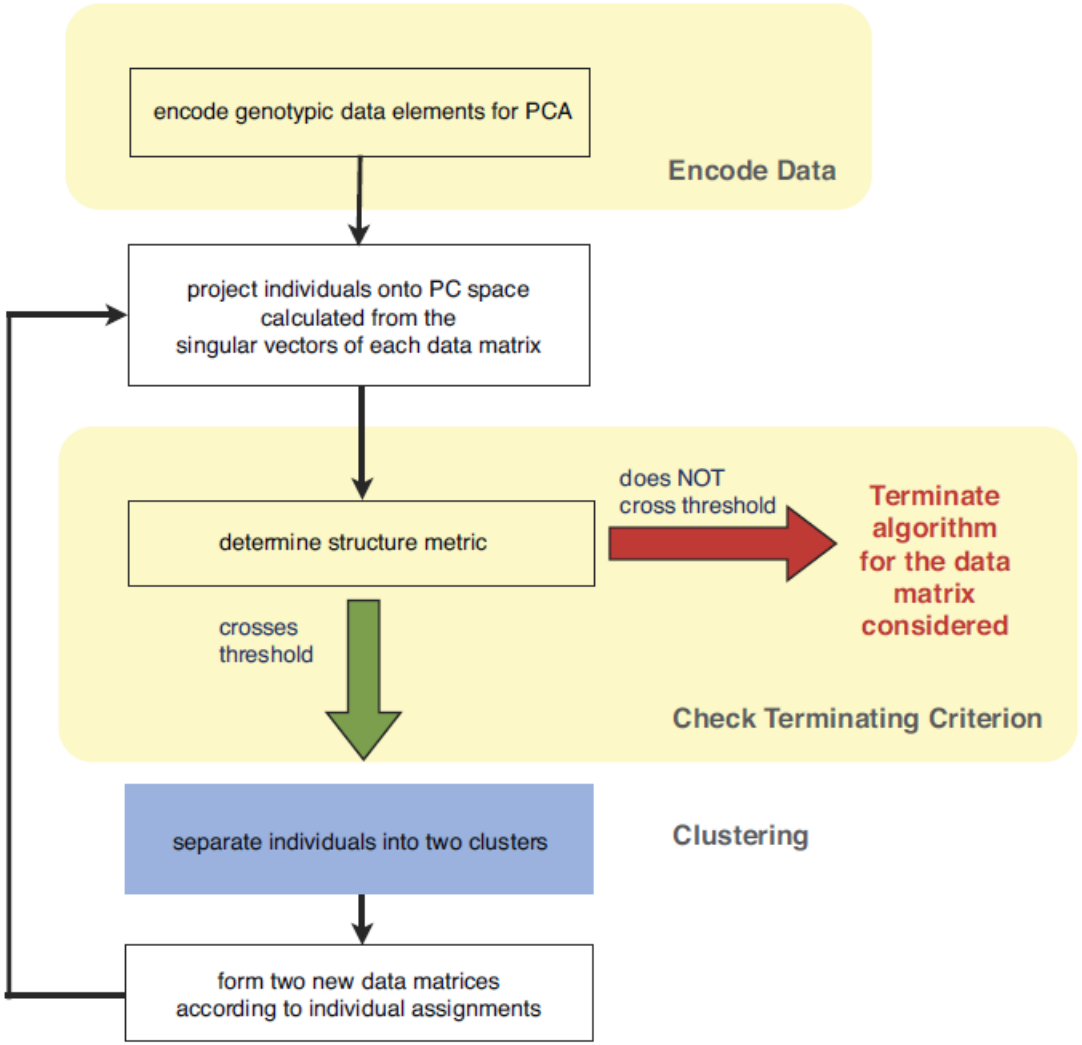
where I_n is the identity matrix of size n (number of individuals) and A is a $n \times k$ matrix containing (a subset of) k principal components

- Before ancestry correction, patient clusters were significantly associated with ethnicity, while after correction no significant association could be detected

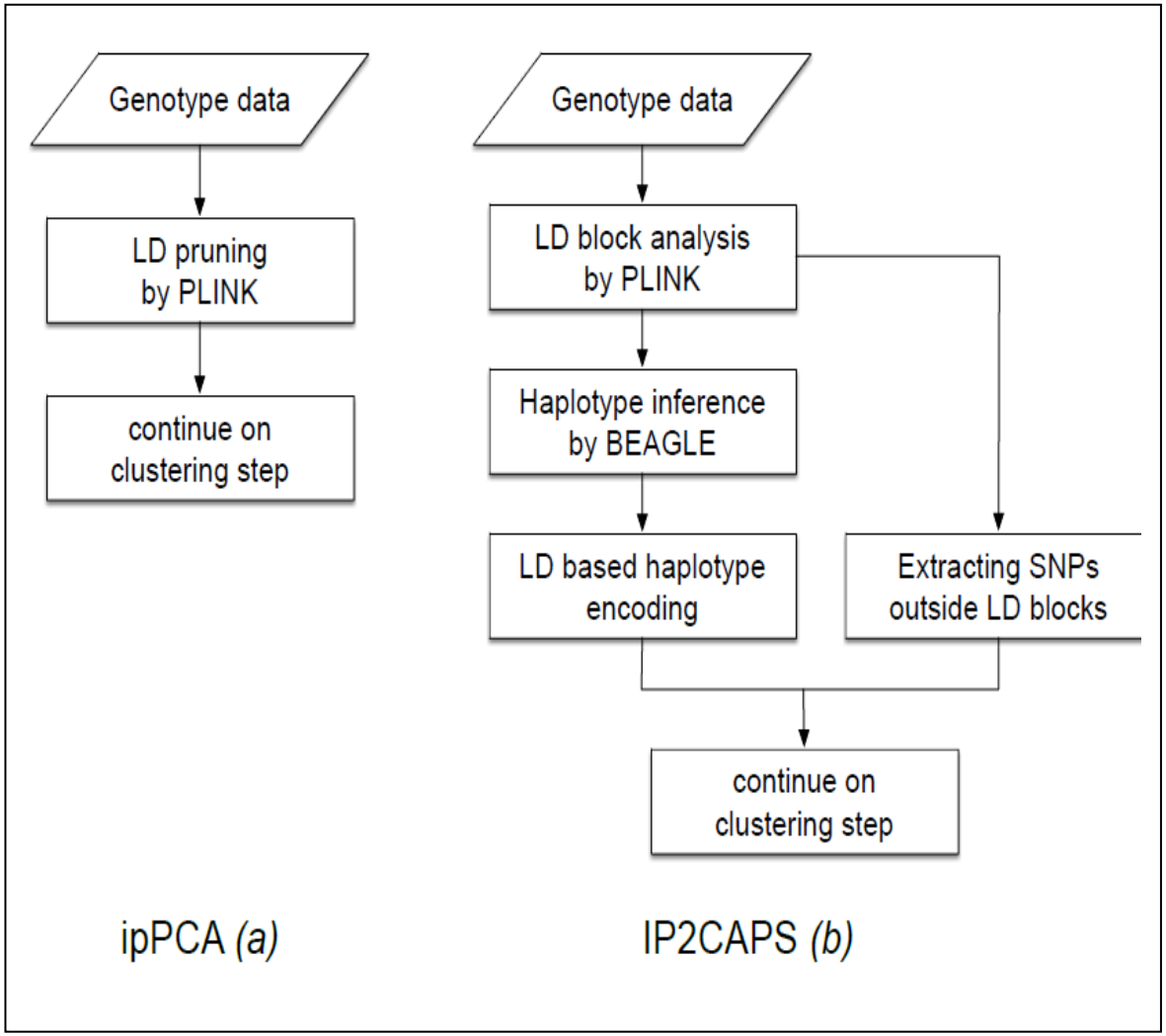
Detecting fine structure in general populations

- Conventional population structure analysis uses **Bayesian statistics** to show or incorporate in the association analysis between-individual relationship in terms of their admixture profiles (e.g., Zhu et al. 2013)
- The **iterative pruning Principal Component Analysis** (ipPCA) approach differs from this paradigm in that high-dimensional clustering is used to assign individuals to subpopulations without using assumptions of population membership or ancestry (Intarapanich, et al., 2009)
- Although computing **haplotypes** incurs extra computational effort, haplotypes may provide more power for genetic analysis, e.g. inferring population structure (Lawson, et al. 2012) and detecting disease association (Xu and Guan 2014). (Chaichoompu et al. 2015 – submitted)

ipPCA (Limpiti et al. 2011)



IP2CAPS (Chaichoompu et al. 2015 – submitted)

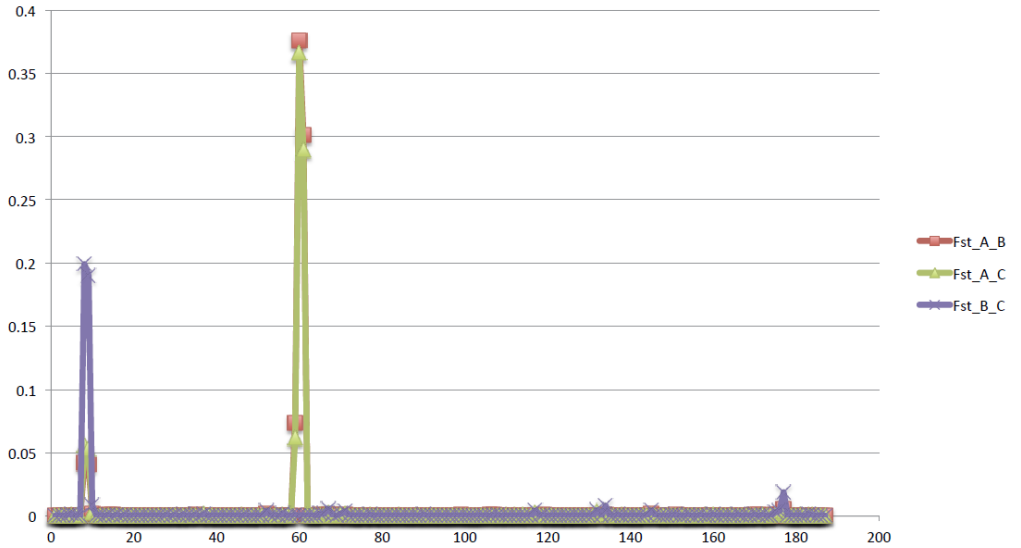


- Whereas ipPCA uses a pruned set of SNPs to prevent distortions of eigenvalues in PCA, IP2CAPS potentially utilizes all SNPs
- Haplotype information inferred from LD-blocks, as well remaning SNPS can be used

IP2CAPS extensions (Chaichoompu 2015+)

- When used as a fine structure detection tool in patients (molecular reclassification), regress out general population substructure
- General population structure can be captured using
  - ancestry informative markers,
  - genome-wide SNPs and/or
  - epigenetic markers
- Regressors may be classic PCs or generalized PCs (R Fouladi 2015+)
- IP2CAPS naturally treats input features as if continuous (ordinal categories); hence **SNPs adjusted for ancestry** more naturally fit in compared to unadjusted SNPs (ordinal, with only a few categories)
- Application: IIBDGC CD (SNPs)

F_{ST} - controls



CONTROL	UC	CD	IBD
SNP1	SNP1	SNP1	SNP1
SNP2	SNP2	SNP2	SNP2
SNP3	SNP4	SNP4	SNP5
SNP6	SNP3		
	SNP6		

	Normal – Group1	Normal – Group2
Recoded as 0/1 – Group1	21	19
Recoded as 0/1 – Group2	3	7
Recoded as 0/1 – Group3	106	147
Recoded as 0/1 – Group4	87	129
Recoded as 0/1 – Group5	97	114
Recoded as 0/1 – Group6	5550	6929

(Chaichoompu, 5 May 2015 IIBDGC TC)

IP2CAPS extensions (Chaichoompu 2015+)

- Categorical treatment of input features (i.e., when the assumption of ordered category levels is unacceptable or invalid)
- This implies replacing PCA in the IP2CAPS algorithm by **CATPCA** or a **generalized version of PCA**
- A clever choice of kernel can acknowledge complex relationships between different types of features (e.g., epigenetic markers and SNPs; and diffusion kernels to take into spatial relationships)
- In order for kernel PCA to be feasible, a heavy pruning will be needed.
  - Note that correlation patterns between epigenetic markers are distinct from those observed between SNPs (Barfield et al. 2012)

IP2CAPS extensions (Chaichoompu 2015+)

- Alternatively, “blocks” of epigenetic markers and SNPs are summarized into a novel categorical variable and/or complemented by remaining SNPs and epigenetic markers (similar to SNP-based IP2CAPS (b) before)
- Such blocks can be captured by the data preparation step adopted in (Fouladi et al. 2015)

  Phase 2 in genomic MB-MDR: Clustering individuals according to features (e.g., common and rare variants, epigenetic markers, …) *within a block,* followed by a kernel PCA, followed by hierarchical clustering and Dynamic Tree Cut)

Note

Smith *et al. BMC Genomics* 2014, **15**:145
http://www.biomedcentral.com/1471-2164/15/145

BMC Genomics

RESEARCH ARTICLE Open Access

Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type

Alicia K Smith[1,2*†], Varun Kilaru[1†], Mehmet Kocak[3], Lynn M Almli[1], Kristina B Mercer[2], Kerry J Ressler[1,4], Frances A Tylavsky[3] and Karen N Conneely[5]

Abstract

Background: Individual genotypes at specific loci can result in different patterns of DNA methylation. These methylation quantitative trait loci (meQTLs) influence methylation across extended genomic regions and may underlie direct SNP associations or gene-environment interactions. We hypothesized that the detection of meQTLs varies with ancestral population, developmental stage, and tissue type. We explored this by analyzing seven datasets that varied by ancestry (African American vs. Caucasian), developmental stage (neonate vs. adult), and tissue type (blood vs. four regions of postmortem brain) with genome-wide DNA methylation and SNP data. We tested for meQTLs by constructing linear regression models of methylation levels at each CpG site on SNP genotypes within 50 kb under an additive model controlling for multiple tests.

Results: Most meQTLs mapped to intronic regions, although a limited number appeared to occur in synonymous or nonsynonymous coding SNPs. We saw significant overlap of meQTLs between ancestral groups, developmental stages, and tissue types, with the highest rates of overlap within the four brain regions. Compared with a random group of SNPs with comparable frequencies, meQTLs were more likely to be 1) represented among the most associated SNPs in the WTCCC bipolar disorder results and 2) located in microRNA binding sites.

Combining it all ….

- Apply the IP2CAPS tool to identify fine **substructure** in patients (hence, finding levels of molecular heterogeneity)
- Construct a functional **network** on the original (unstratified) set of patients (possibly based on **statistical epistasis**)
- Investigate whether modules derived from the functional network above behave differently in different patient strata (molecularly defined)
- Find the molecular drivers of the “patient clusters”
- Investigate whether these drivers may perturb earlier constructed functional (integrated) networks on the initial set of patients and assess the **implications for future personalized healthcare**

Personalized healthcare implies targeted patient clustering

- Major drug companies are interested in identifying those patients that would benefit most from treatment
- One way of doing so is to make use of Random Forest (RF) methodology
- In particular, having omics characterization of patients in mind, Conditional Inference Forests have added value over RFs (see before), but needs adaption.

Adapted CIF methodology (Dizier 2015+)

- Base node splitting on Generalized Linear Models to allow covariates adjustment (e.g., population structure!) and interaction
- Instead of choosing the best split in the variable with the best linear association, choose the best variable and split when there is enough **evidence of association**
- Model-based targeted variable importance

Adapted CIF methodology (Dizier 2015+)

<table>
<tr><th>Decision</th><th>CIF</th><th>New method</th></tr>
<tr><td>Is there enough evidence of association to split data further?</td><td rowspan="2">Global independence test</td><td>Goeman's Global Test
(Goeman et al.2004)</td></tr>
<tr><td>Which variable to use?</td><td rowspan="2">MaxT</td></tr>
<tr><td>Which split in this variable?</td><td>Maximally ranked statistics</td></tr>
</table>

Adapted CIF methodology (Dizier 2015+)

- Globaltest fits a GLM random effect model with all variables considered at a node, possibly appended with covariates Z.

$$g(Y) = \beta_0 + \beta_Z Z + \sum_{i=1}^{m} \beta_i X_i$$

$H_0: \beta_1 = \cdots = \beta_m = 0 \leftarrow\rightarrow \beta_i \sim N(0, \tau^2); H_0: \tau^2 = 0$

- Only one test that does not require multiple testing correction

- Fit $g(Y) = \beta_0 + \beta_Z Z + \beta_G G + \beta_{Split} Split + \beta_{GSplit} GSplit$, with Z additional covariates, G a treatment label and Split a dummy variable based on splitting the variable in two groups

$$H_0 : \beta_{GSplit} = 0$$

- Number of tests = variables x splits and requires adequate multiple testing correction (work in progress)

In conclusion

- A wealth of **omics** information is available, including but not limited to genomics, epi-genomics, meta-genomics and proteomics
- One of the major challenges for the next decade is to determine when and how this wealth of omics can be usefully applied by both the public and private sectors for the development of **personalized /stratified approaches** in health promotion and disease prevention
- Personalized medicine not only involves taking into account the 'omics' characteristics of individuals, but also complementing these with information about **environmental and/or lifestyle** factors

(EU HORIZON2020 call: **"**translating 'omics' into stratified approaches")