Systems Health

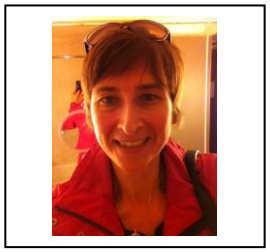
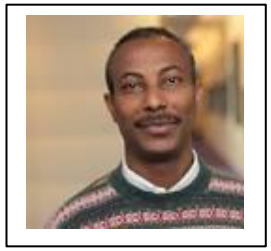
Challenges and Opportunities

Kristel Van Steen, PhD[2] (*)

kristel.vansteen@ulg.ac.be

(*) WELBIO, GIGA-R, Medical Genomics, University of Liège, Belgium

Systems Medicine Lab, KU Leuven, Belgium

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




OUTLINE

- Systems Health

  Precision medicine

- 10 challenges and opportunities

  - Tier 1: study design

  - Tier 2: analytics (interactions)

- Conclusion

Systems Health

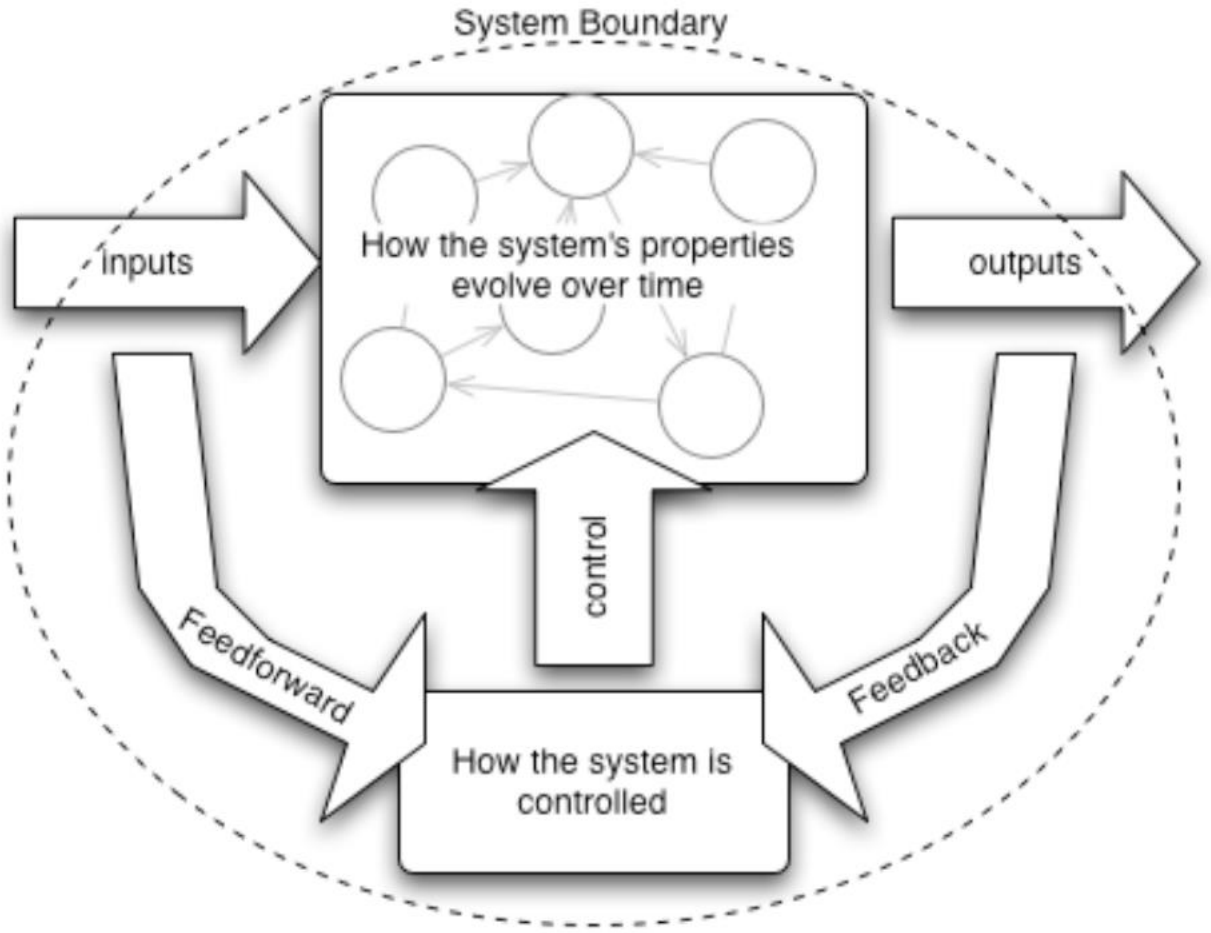
Systems

What is a system?

- A system is a set of two or more elements that satisfies the following conditions:
  - The behavior of each element has an effect on the behavior of the whole
  - The behavior of the elements and their effect on the whole are interdependent
  - Subgroups of elements can be formed, in which case each has an effect on the behavior on the whole and none has an independent effect on it.

(Ackoff, 1970)

A System's Eco-system



(@2004-5 Steve Easterbrook)

Health → Individual health

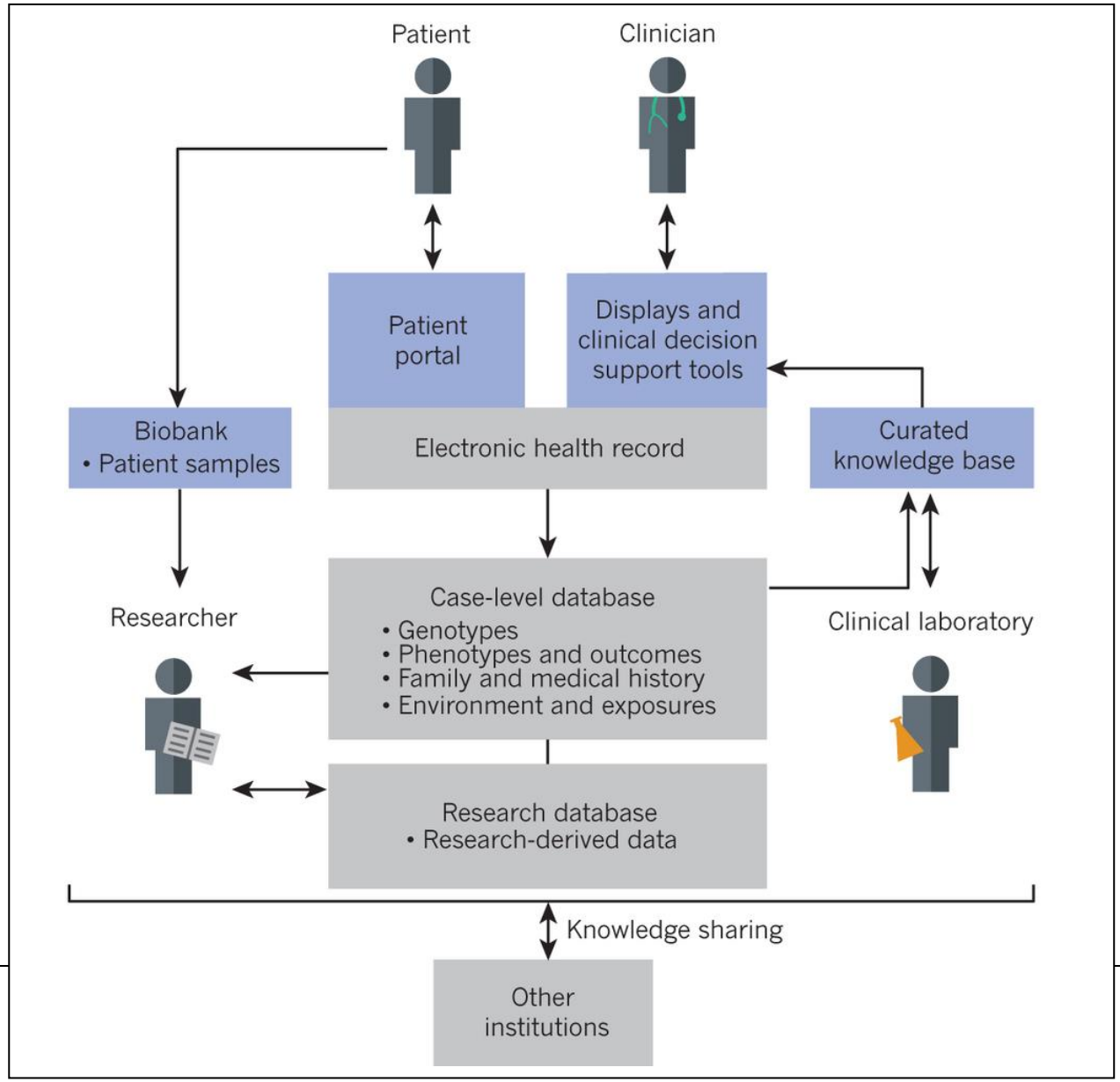
What is precision medicine?

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

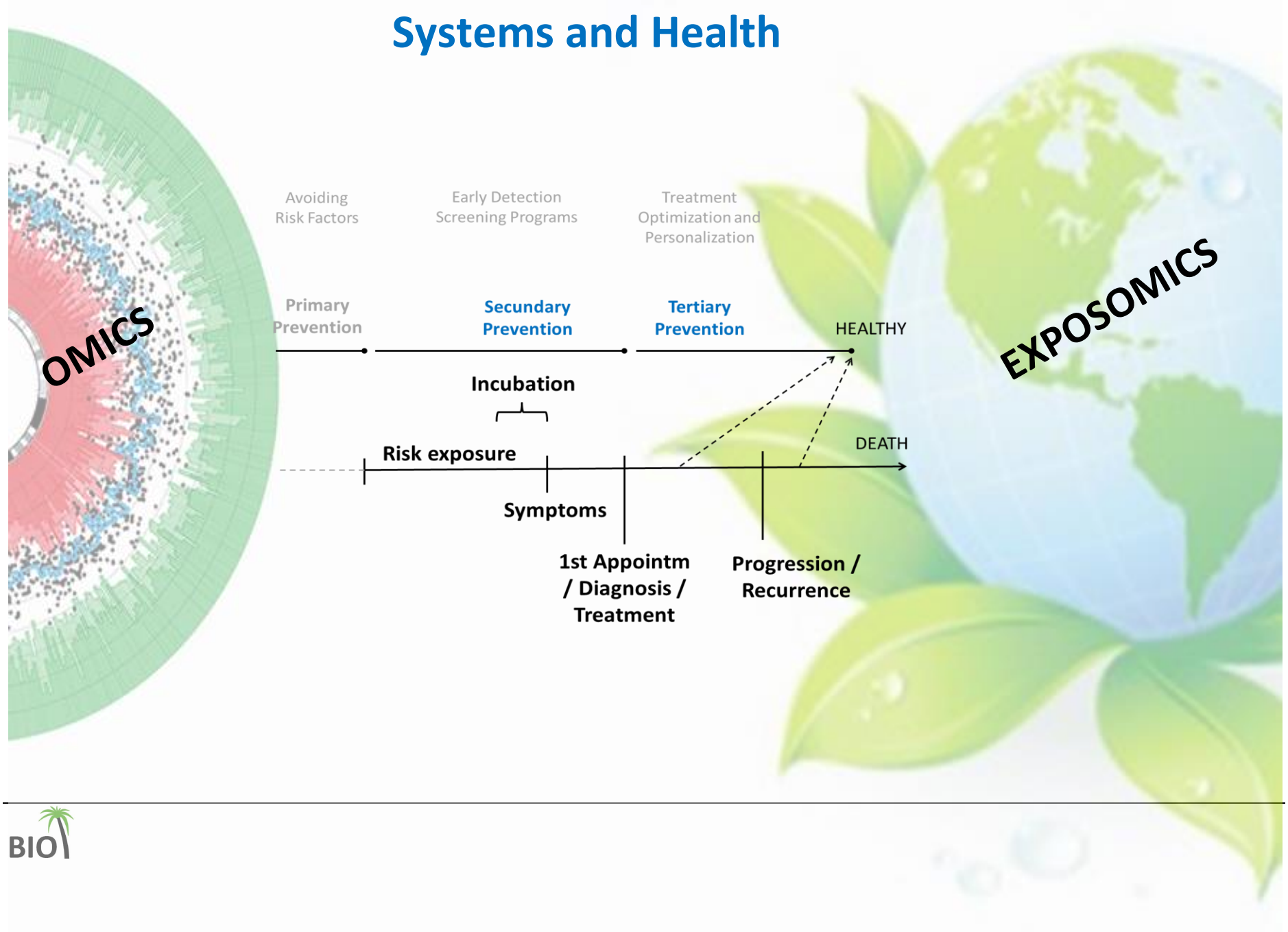
(HORIZON2020 Advisory Group)

A Patient's Eco-System

(Aronson and Rehm 2015)



Systems and Health



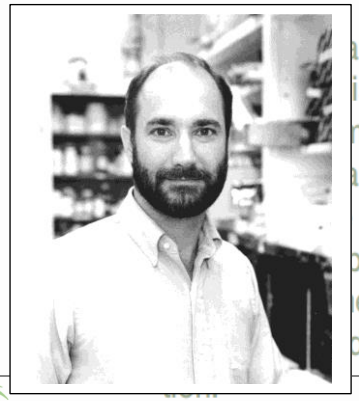
Data deluge allows precise individual-level characterizations

Cell PRESS

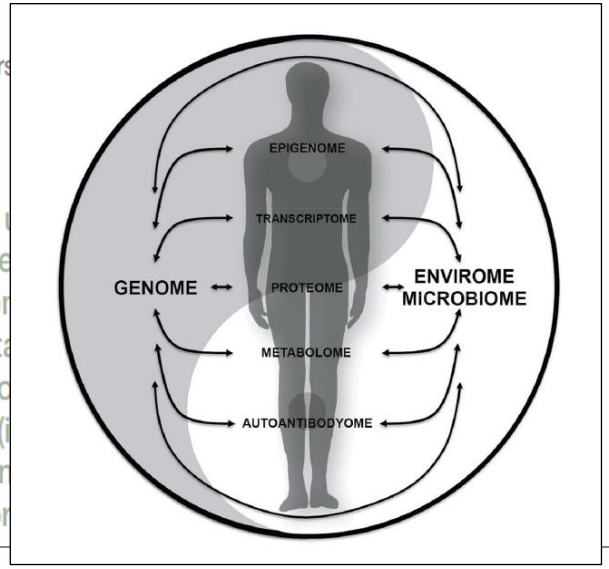
Chemistry & Biology
Review

iPOP Goes the World: Integrated Personalized Omics Profiling and the Road toward Improved Health Care

Jennifer Li-Pook-Than[1] and Michael Snyder[1,*]
[1]Department of Genetics, Stanford University School of Medicine, Stanford Univers
*Correspondence: mpsnyder@stanford.edu
http://dx.doi.org/10.1016/j.chembiol.2013.05.001



an individual depends upon their DNA as well as
is expected that although the genome is the blue
nes such as the DNA methylome, the transcriptor
amic assessment of the physiology and health sta
current progress of omics analyses and how c
believe that integrative personal omics profiling (i
ealth care and may improve disease risk assessr
d treatments, and understanding the biological pr

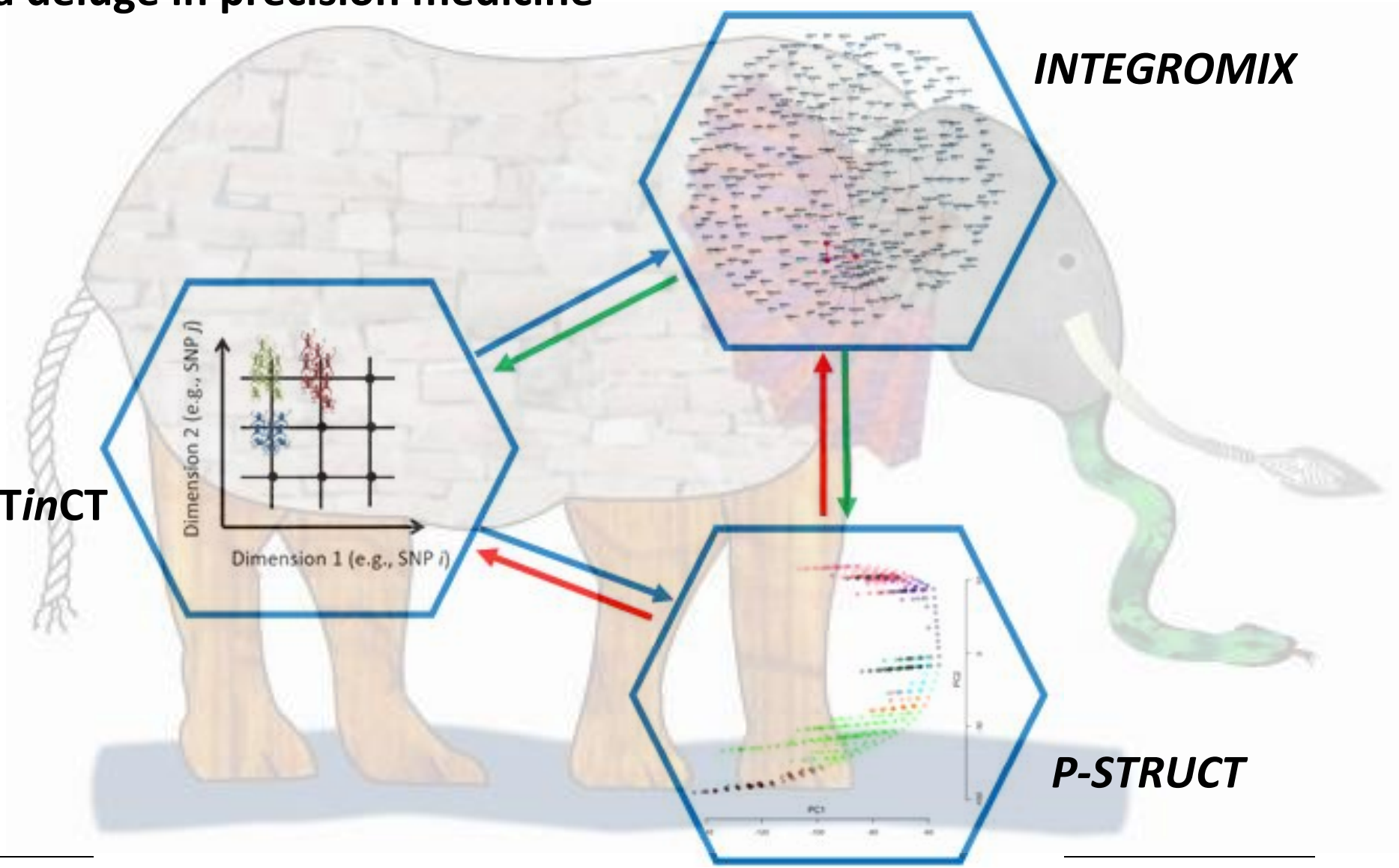


Data deluge shows extensive disease heterogeneity (IBD)

	Observation in subgroups of patients	Disease	Refs
Genetic	Variants in autophagy genes (*ATG16L1, IRGM*)	CD	[14]
	NOD2 polymorphisms	CD	[15,16]
	HLA-DRA polymorphisms	UC	[20]
	IL10 polymorphisms	UC>>CD	[20]
	IL2/IL21 polymorphisms	UC>>CD	[14]
	Variants in Th1 genes (*STAT1, STAT4, IL12B, IFN, IL18RAP*)	CD, UC	[13,14]
	Variants in Th17 genes (*IL23R, STAT3, RORC*)	CD, UC	[14,23]
Immunological	Great inter- and intra-individual variability in mucosal proinflammatory cytokine production	CD, UC	[32,33]
	↑ IFN-γ production by lamina propria T cells	CD>UC	[34]
	↑ IL-5 production by lamina propria T cells	UC>CD	[34]
	↑ mucosal IL-12, STAT4, T-bet	CD>>UC	[35,36]
	↑ IL-13 production by lamina propria NK T cells	UC>CD	[37]
	↑ mucosal IL-17A, Th17 and Th1/Th17 cells compared to controls	CD, UC	[32,40]
	↑ IFN-γ production by lamina propria T cells in early but not late disease	CD	[46]
	↑ mucosal IL-17A, IL-6, IL-23 before endoscopic recurrence but not in established lesions	CD	[47]
	Transcriptional signatures in circulating $CD8^+$ T cells associated with different prognosis	CD, UC	[57]
Clinical	Inflammatory/penetrating/fibrostenosing phenotype	CD	[48]
	Inter-individual variability in disease extension	CD, UC	[3,50]
	Great inter-individual variability in prognosis	CD, UC	[50]
	Young age at diagnosis, current smoking, presence of perianal and/or extensive disease, initial requirement for steroids: associated with worse prognosis	CD	[50,55]
	Young age at diagnosis, pancolitis, no appendectomy in childhood: associated with worse prognosis	UC	[50]
	Great inter-individual variability in need for surgical intervention	CD, UC	[50]

(Biancheri et al. 2013)

Data deluge in precision medicine

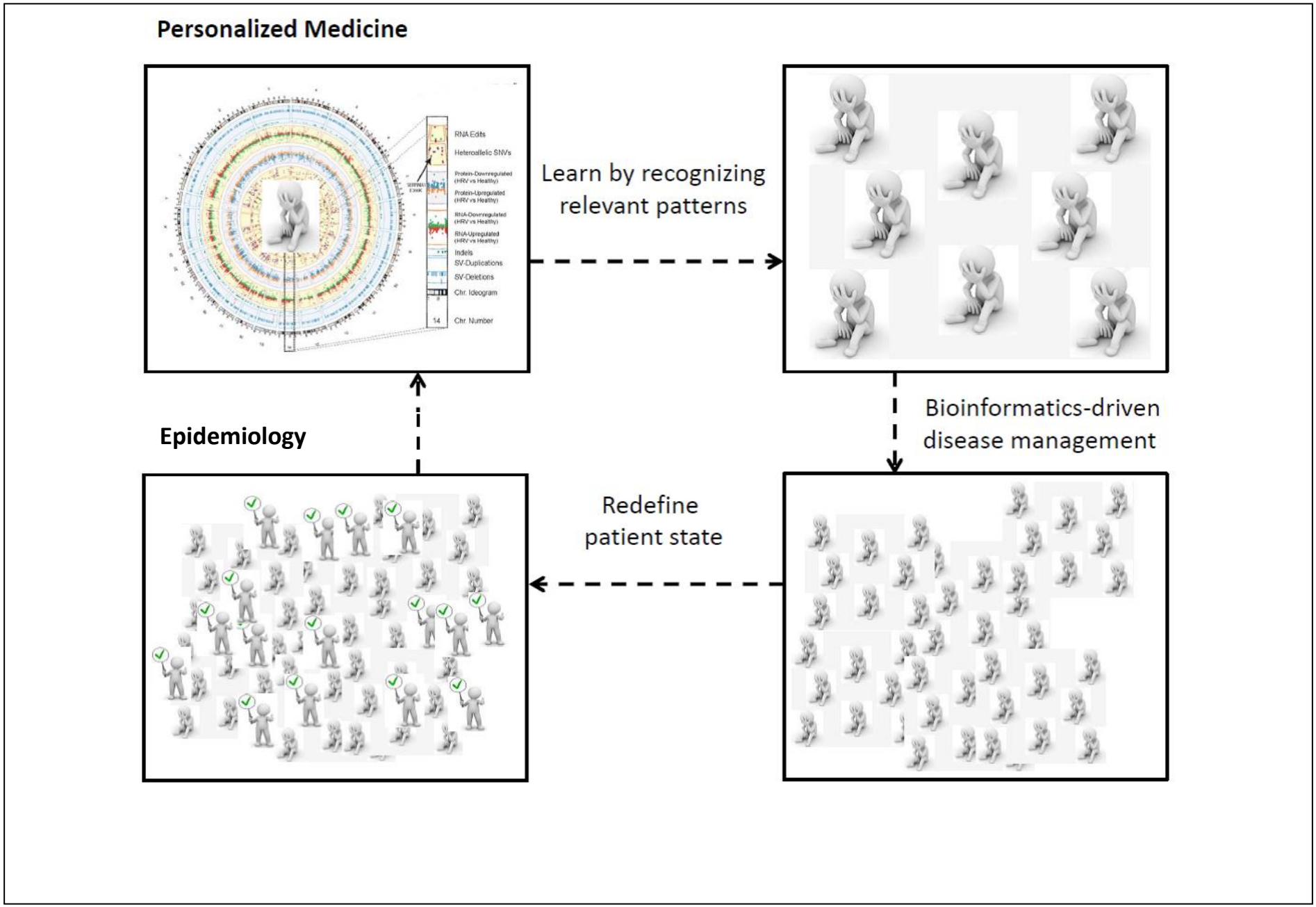


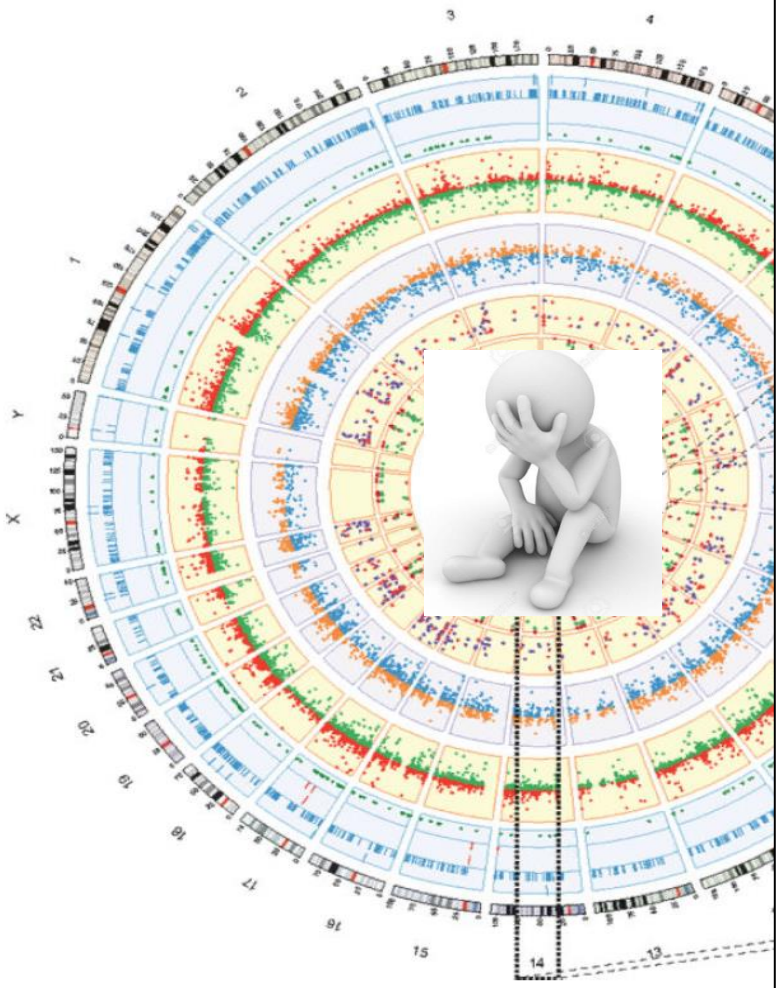
Tier 1: Systems thinking in study design

Challenges and opportunities

5 Challenges and Opportunities

- Continuum – range of disease presentations (dozens of IBD? what are outliers?)
- Informativity versus redundancy – not all data are relevant for a particular data problem (definition of relevance)
- Multiple data sources in a system – not available to all patients (missing data)
- Heterogeneity – a target and a nuisance (corrections for confounding)
- Replication and validation – translation to the clinic (finding “similar” independent data)


Personalized Medicine
RNA Edits
Heteroallelic SNVs
Protein-Downregulated (HRV vs Healthy)
Protein-Upregulated (HRV vs Healthy)
RNA-Downregulated (HRV vs Healthy)
RNA-Upregulated (HRV vs Healthy)
Indels
SV-Duplications
SV-Deletions
Chr. Ideogram
14
Chr. Number
Learn by recognizing relevant patterns
Bioinformatics-driven disease management
Redefine patient state
Epidemiology

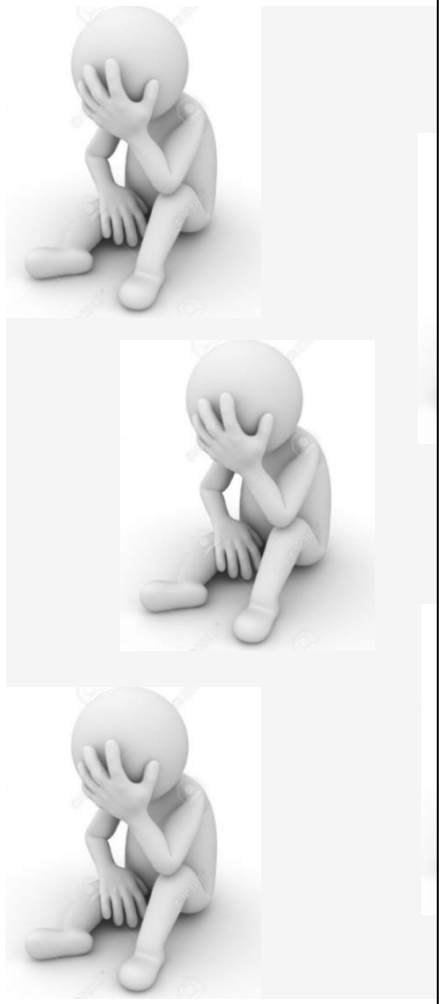


Do you think that omics profiling will be routinely used in the clinic in future?

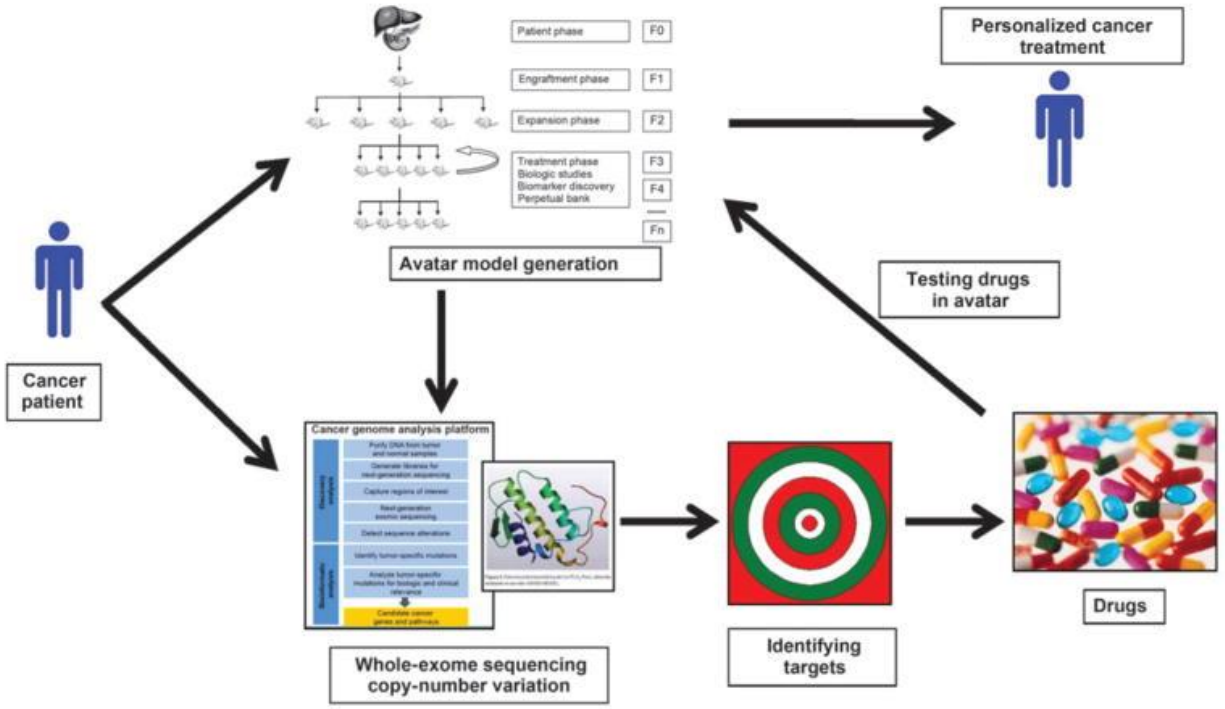
"Not in the form we are doing it. At the moment we have a very incomplete picture of what's going on, whereas if we were able to make thousands of measurements we would have a much better feeling. We just don't know, for the clinical tests, which thousand measurements are going to be most useful. We'll need certain measurements for diabetes, others for cancer, and specific tests will probably reveal themselves useful for different diseases."

(Snyder 2014)

Redundancy - Informativity

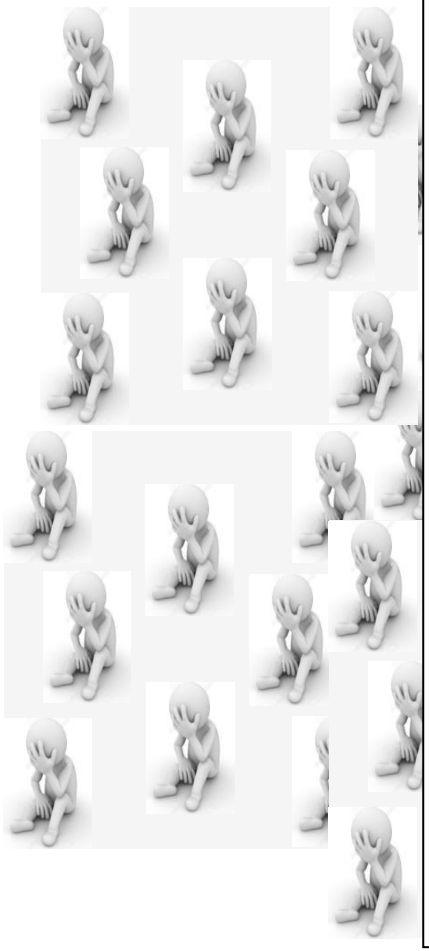


Integrating sequencing and avatar mouse models

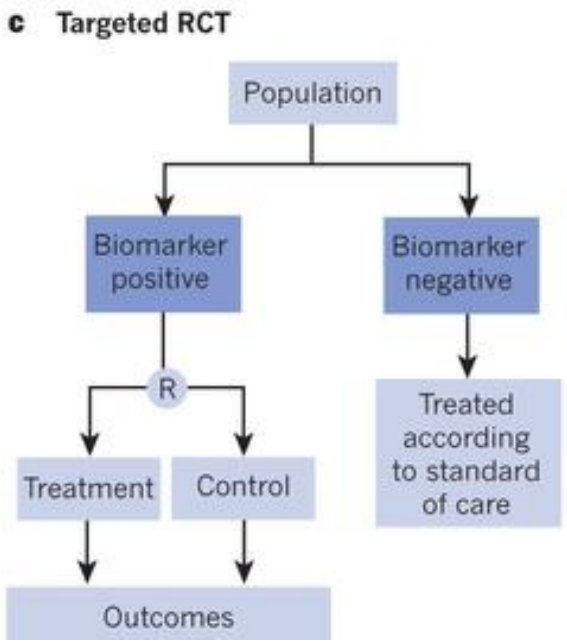


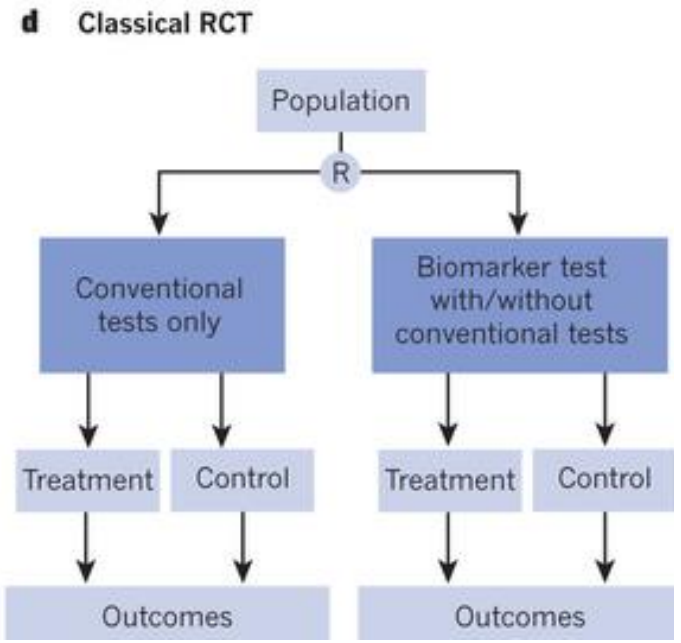
Missingness

(Garralda et al. 2014)



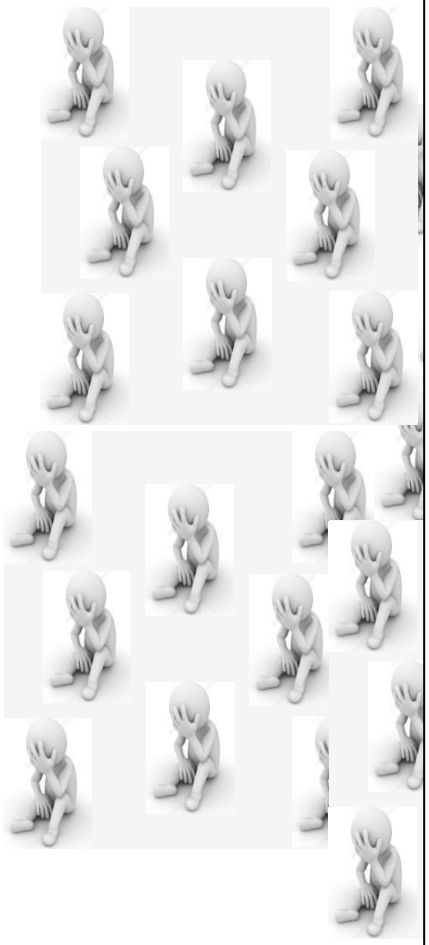
Testing precision-medicine strategies



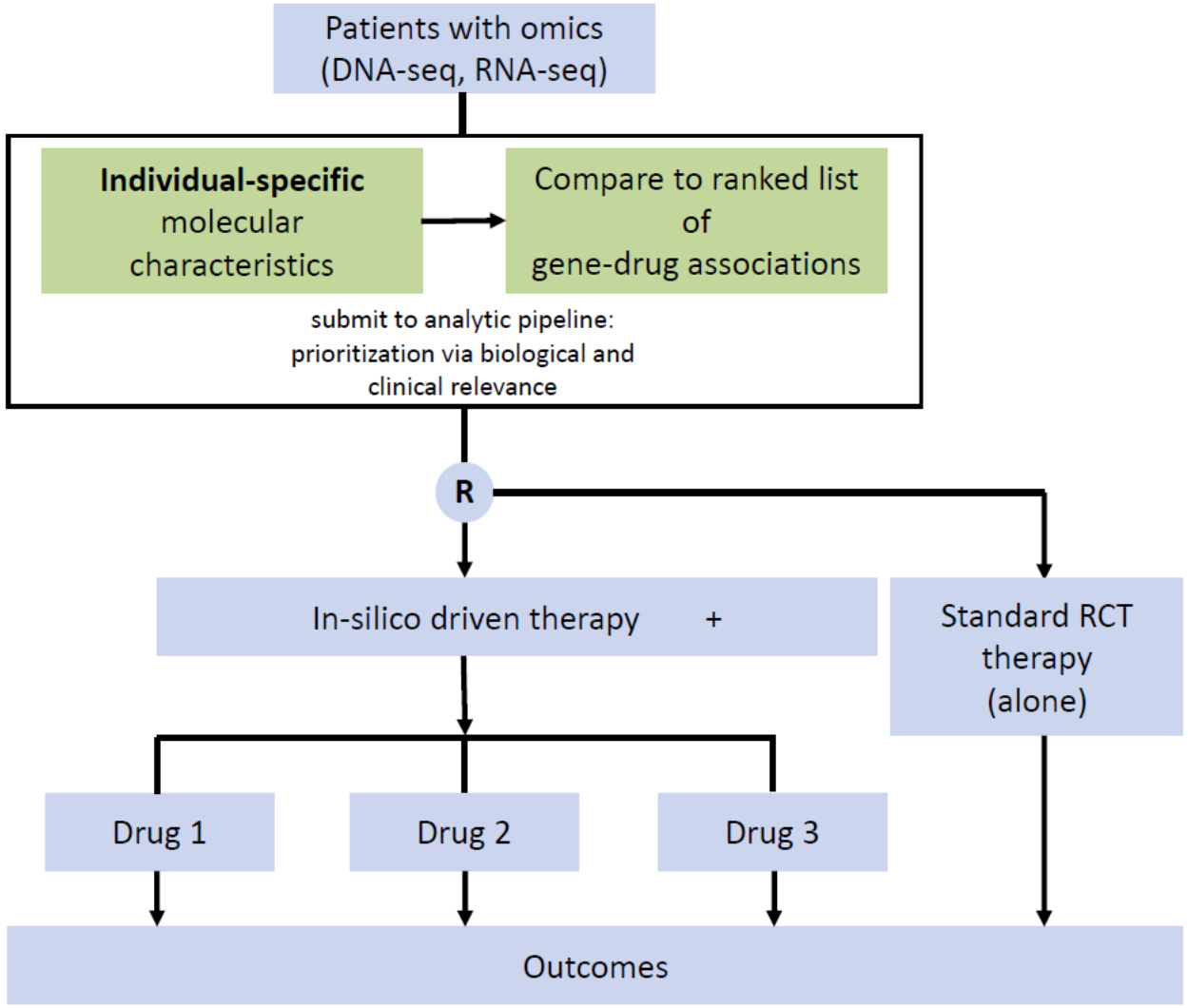


Replication and validation

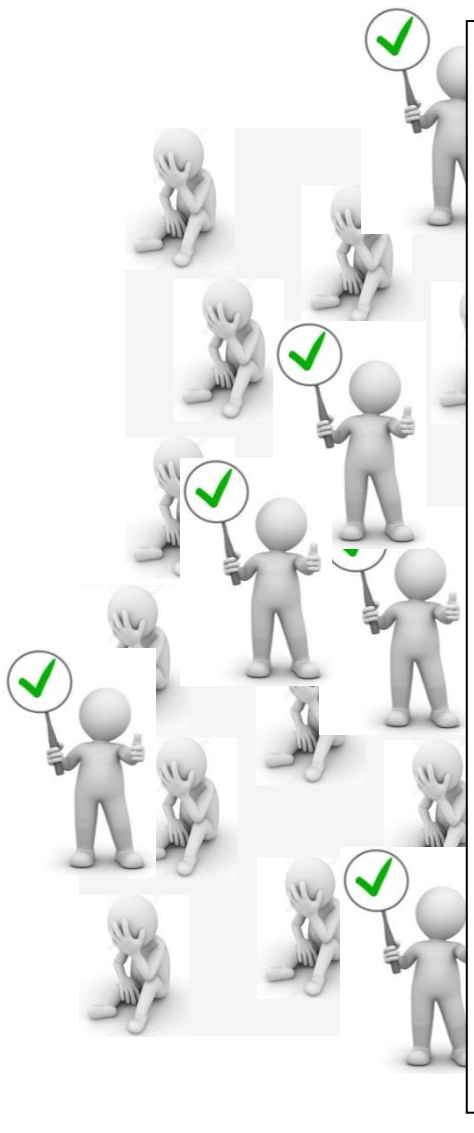
(Biankin et al. 2015)



Testing precision-medicine strategies



BIO



Molecular profiling; What does it mean to be „Diseased“?

OPEN ACCESS Freely available online

PLOS ONE

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

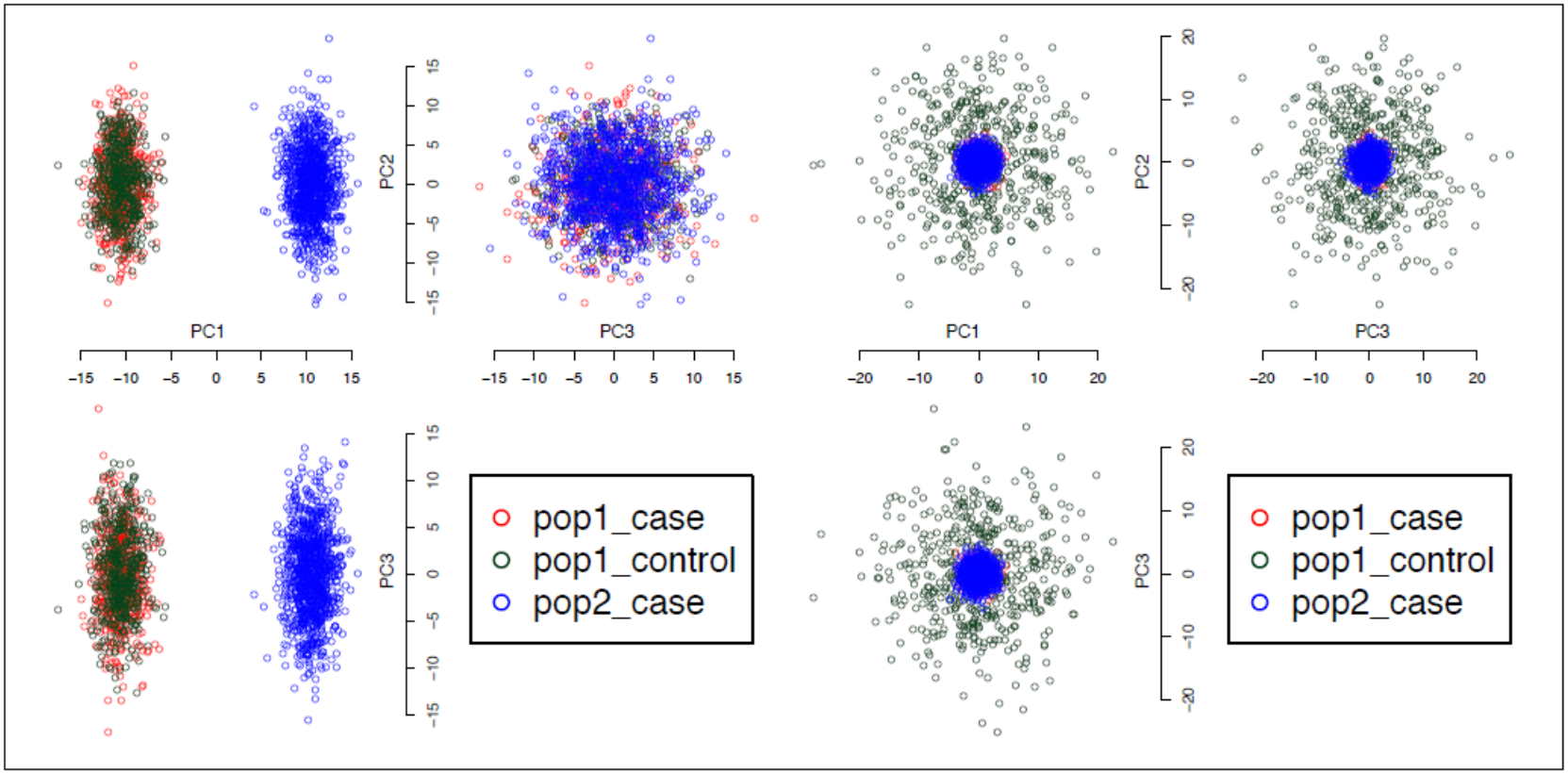
Bärbel Maus[1,2*], Camille Jung[3,4,5], Jestinah M. Mahachie John[1,2], Jean-Pierre Hugot[3,4,6], Emmanuelle Génin[7,8], Kristel Van Steen[1,2]

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

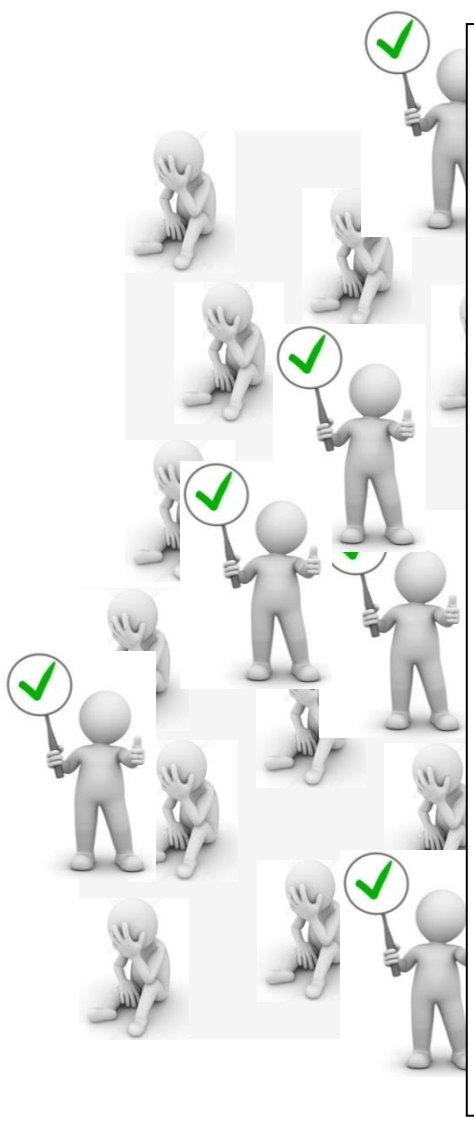
(Maus et al. 2013)

Heterogeneity as a target and a nuisance

Linear population structure correction (Chaichoompu 2017+)



Pooled case/control PCs (left) vs Case-Projected PCs (right)



What does it mean to be „Diseased“?

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MACHINE LEARNING
POPULATION GENETICS

Received
30 September 2014

Accepted
8 January 2015

Highlighting nonlinear patterns in population genetics datasets

Gregorio Alanis-Lobato[1,2]*, Carlo Vittorio Cannistraci[3]*, Anders Eriksson[1,4], Andrea Manica[4] & Timothy Ravasi[1,2]

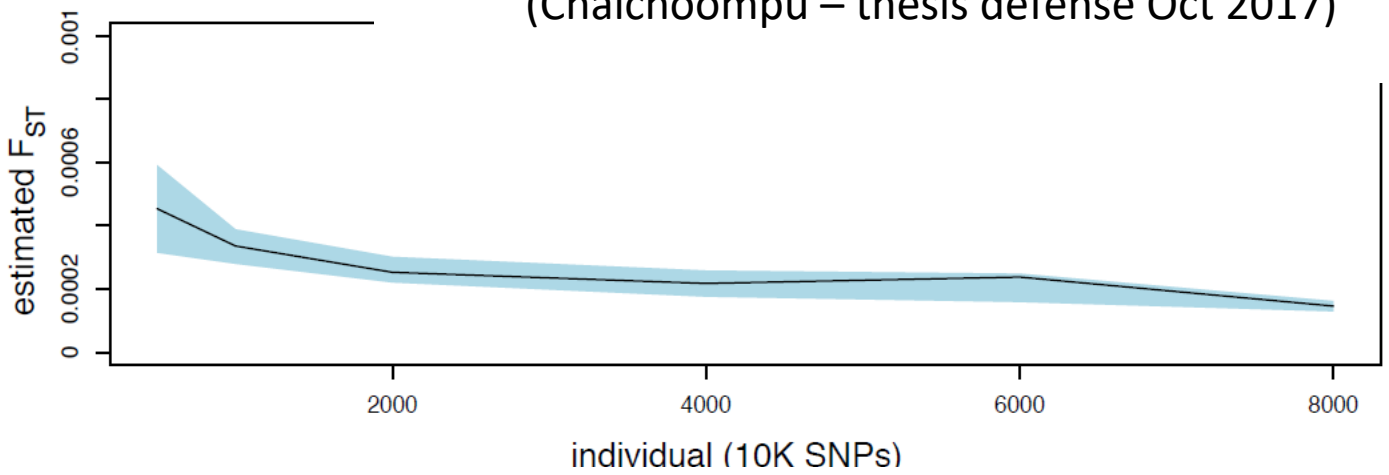
[1]Integrative Systems Biology Laboratory, Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Ibn Al Haytham Bldg. 2, Level 4, Thuwal 23955-6900, Kingdom of Saudi Arabia, [2]Division of Medical Genetics, Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA, [3]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany, [4]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, England.

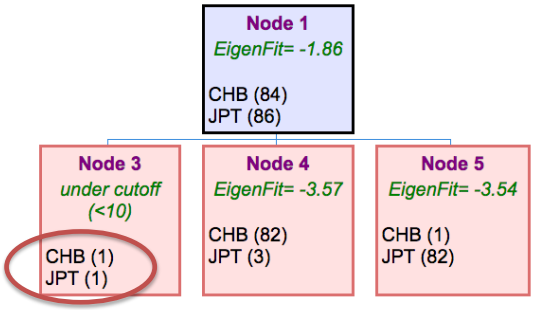
Non-linearity

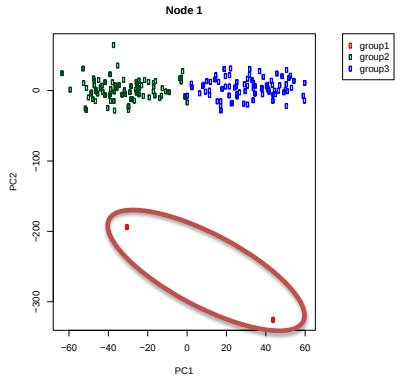
(Alanis-Lobato et al. 2015)

Fine-scale structure

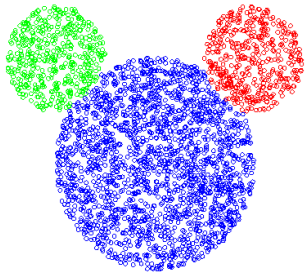
(Chaichoompu – thesis defense Oct 2017)



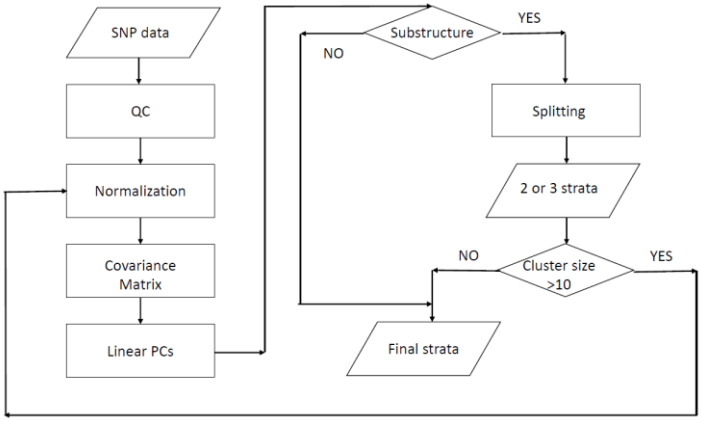




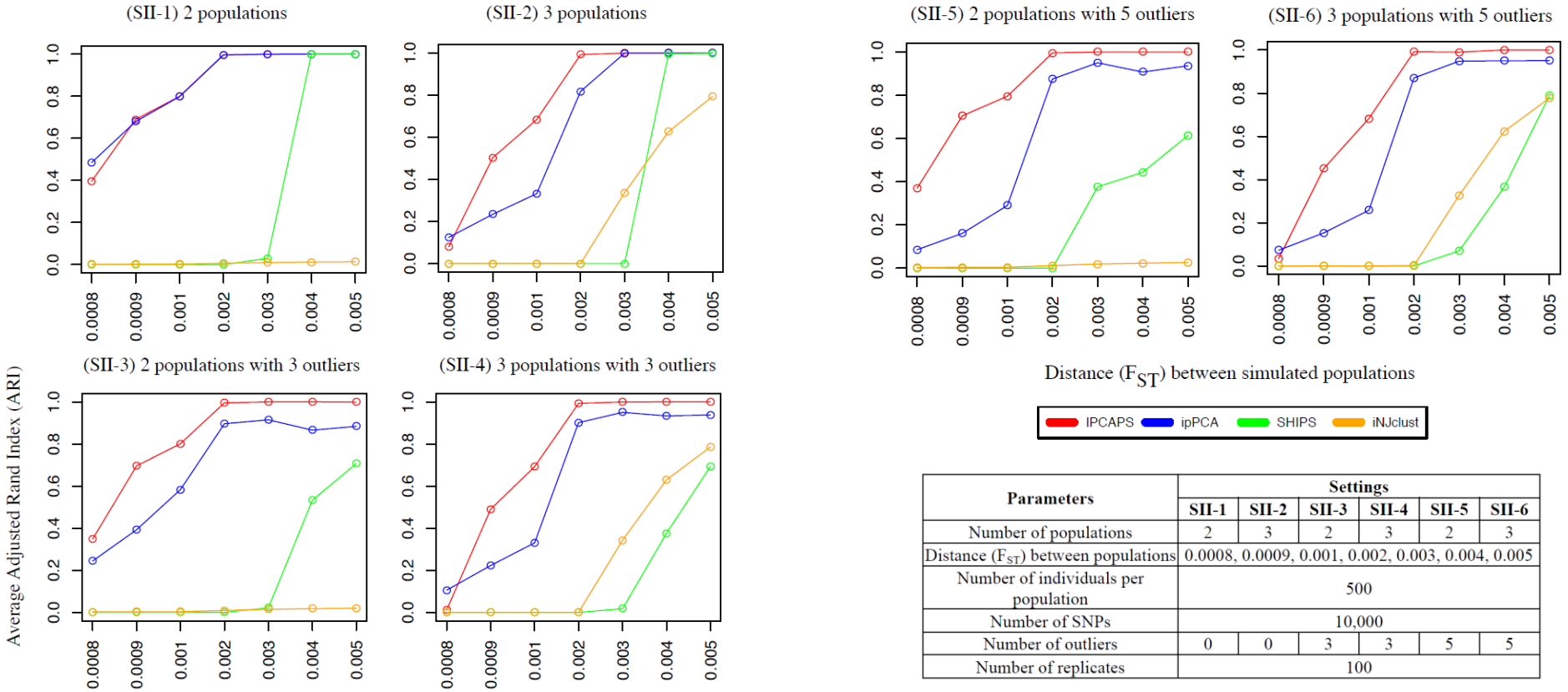
Combine with EM clustering



IPCAPS



Accuracy of IPCAPS

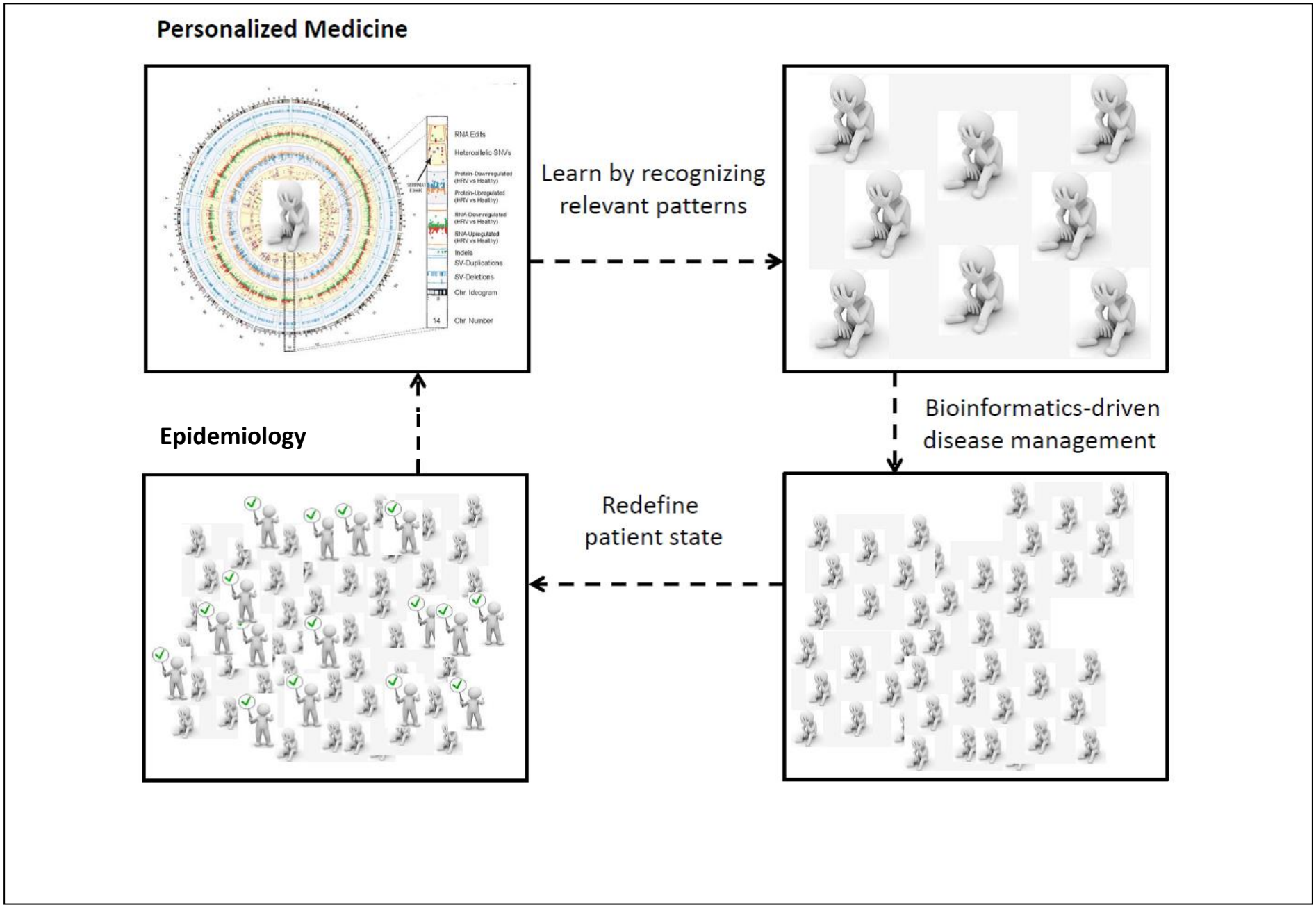


Parameters	Settings					
	SII-1	SII-2	SII-3	SII-4	SII-5	SII-6
Number of populations	2	3	2	3	2	3
Distance (F_{ST}) between populations	0.0008, 0.0009, 0.001, 0.002, 0.003, 0.004, 0.005					
Number of individuals per population	500					
Number of SNPs	10,000					
Number of outliers	0	0	3	3	5	5
Number of replicates	100					

F_{ST} among populations – examples

	Sp	*Fr*	*Be*	*UK*	*Sw*	*No*	*Ge*	*Ro*	*Cz*	*Sl*	*Hu*	*Po*	*Ru*	*CEU*	*CHB*	*JPT*
Fr	0.0008															
Be	0.0015	0.0002														
UK	0.0024	0.0006	0.0005													
Sw	0.0047	0.0023	0.0018	0.0013												
No	0.0047	0.0024	0.0019	0.0014	0.0010											
Ge	0.0025	0.0008	0.0005	0.0006	0.0011	0.0016										
Ro	0.0023	0.0017	0.0018	0.0028	0.0041	0.0044	0.0016									
Cz	0.0033	0.0016	0.0013	0.0014	0.0016	0.0024	0.0003	0.0016								
Sl	0.0034	0.0017	0.0015	0.0017	0.0019	0.0026	0.0005	0.0014	0.0001							
Hu	0.0030	0.0015	0.0013	0.0016	0.0020	0.0026	0.0004	0.0011	0.0001	0.0001						
Po	0.0053	0.0032	0.0028	0.0027	0.0023	0.0034	0.0012	0.0028	0.0004	0.0004	0.0006					
Ru	0.0059	0.0037	0.0034	0.0032	0.0025	0.0036	0.0016	0.0030	0.0008	0.0007	0.0009	0.0003				
CEU	0.0026	0.0008	0.0005	0.0002	0.0011	0.0012	0.0006	0.0028	0.0014	0.0016	0.0016	0.0026	0.0031			
CHB	0.1096	0.1094	0.1093	0.1096	0.1073	0.1081	0.1085	0.1047	0.1080	0.1069	0.1058	0.1086	0.1036	0.1095		
JPT	0.1118	0.1116	0.1114	0.1117	0.1095	0.1103	0.1107	0.1068	0.1102	0.1091	0.1079	0.1108	0.1057	0.1117	0.0069	
YRI	0.1460	0.1493	0.1496	0.1513	0.1524	0.1531	0.1502	0.1463	0.1503	0.1498	0.1490	0.1520	0.1504	0.1510	0.1901	0.1918

(Health et al. 2008)


Personalized Medicine
RNA Edits
Heteroallelic SNVs
Protein-Downregulated (HRV vs Healthy)
Protein-Upregulated (HRV vs Healthy)
RNA-Downregulated (HRV vs Healthy)
RNA-Upregulated (HRV vs Healthy)
Indels
SV-Duplications
SV-Deletions
Chr. Ideogram
14
Chr. Number
Learn by recognizing relevant patterns
Bioinformatics-driven disease management
Redefine patient state
Epidemiology

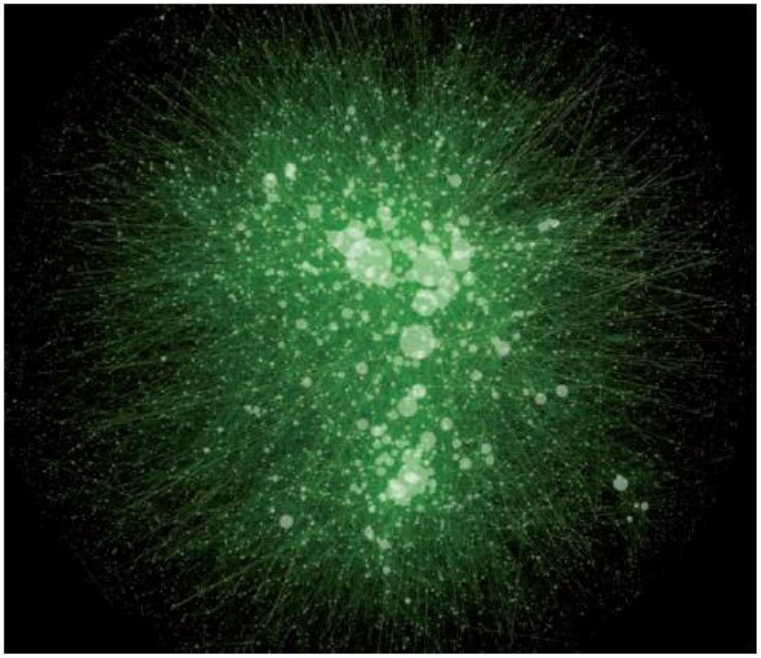
Tier 2: Systems thinking in analytics

Challenges and opportunities

5 Challenges and Opportunities

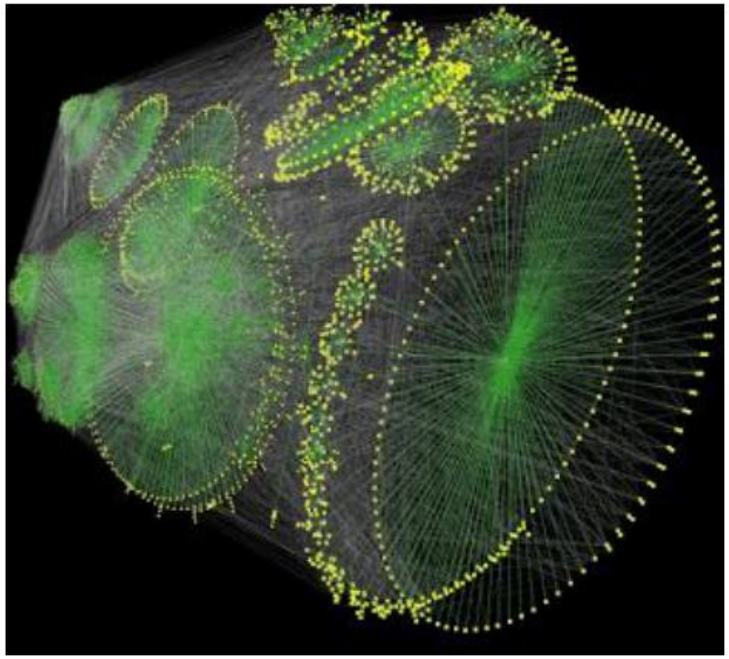
- Joint locus effect or pure interaction (lower-order effects adjustments)
- Redundant epistasis (redundancy due to associated features)
- Non-linear confounders (single or multi-omics characteristics)
- IT infrastructure and resources (data storage, backups, computations)
- Replication and validation – translation to the clinic (finding “similar” independent data and validity of model systems)

Interactions contribute to distinction



Human interactome (PPI)

(Bonetta 2010)



Fruit fly interactome

(www.molgen.mpg.de)

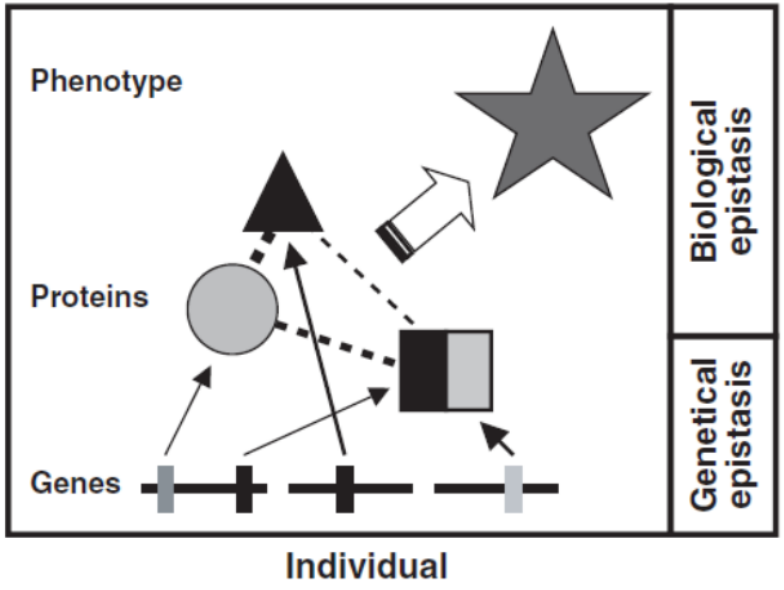
The “interactome”

The **interactome** refers to the entire complement of interactions between DNA, RNA, proteins and metabolites within a cell. These interactions are influenced by genetic alterations and environmental stimuli. As a consequence, the interactome should be examined or considered in ***particular contexts***.

Focus on DNA-DNA interactions

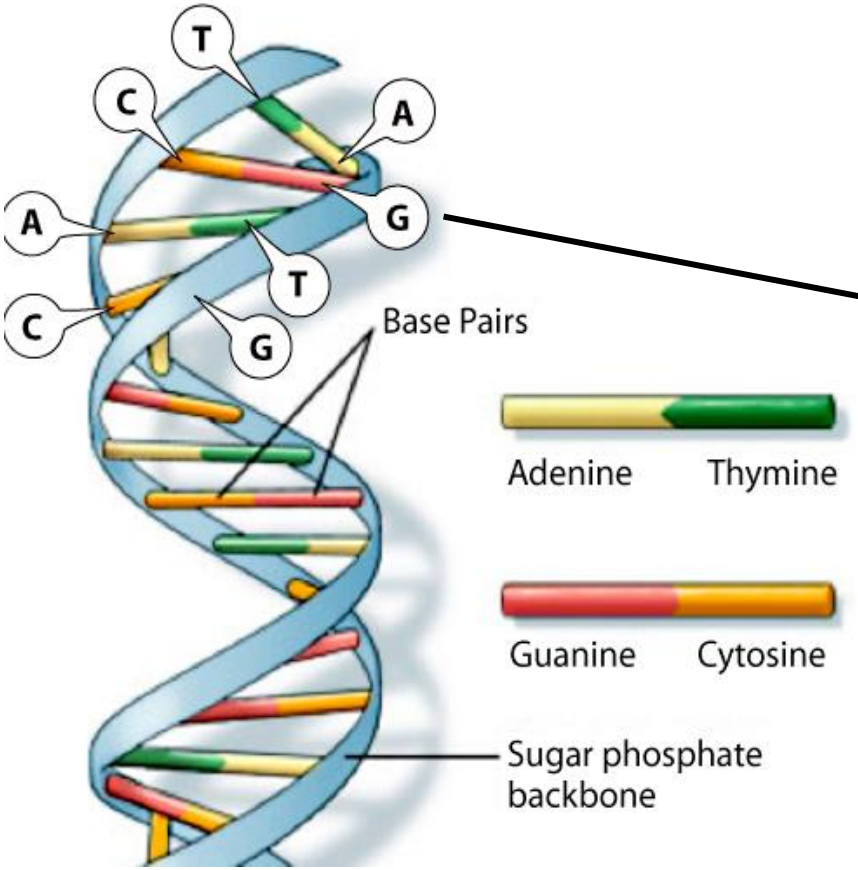
Biological viewpoint

- 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)

Common genetic variations



Single Nucleotide Polymorphisms (SNPs)	Frequency in general population
G	95%
A	5% > 1%

Non-biological viewpoints

- A post-Bateson definition of epistasis driven by **statistics** is expressed in terms of deviations from a model of additive multiple effects (Ronald Fisher 1890-1962).

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_G X_1 + \beta_H X_2 + \beta X_1 X_2$$

- **Neural networks**: In the expression below, interaction is between X_1 and X_2 since they are non-additive, and the additive effect of the single feature X_3 is categorized as a main effects

$$" Y = X_1 X_2 + X_3 "$$

Little work exists on interpreting statistical interactions captured by neural network methods.

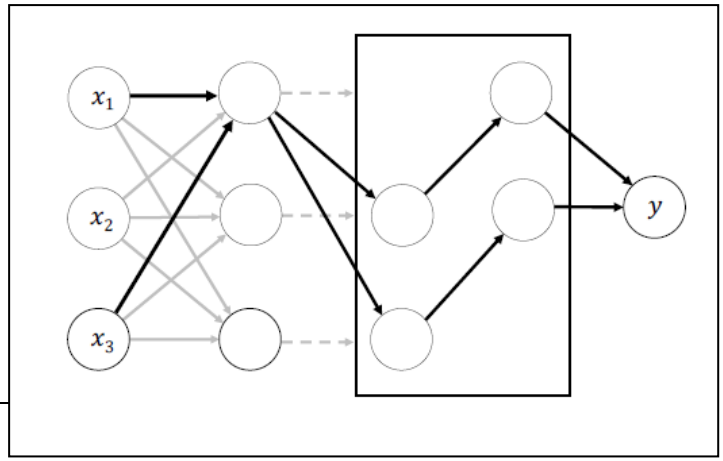
Machine learning and interactions

- **Deep neural networks** have been recognized as some of the best performing machine learning methods:

(Uppu et al. 2016)

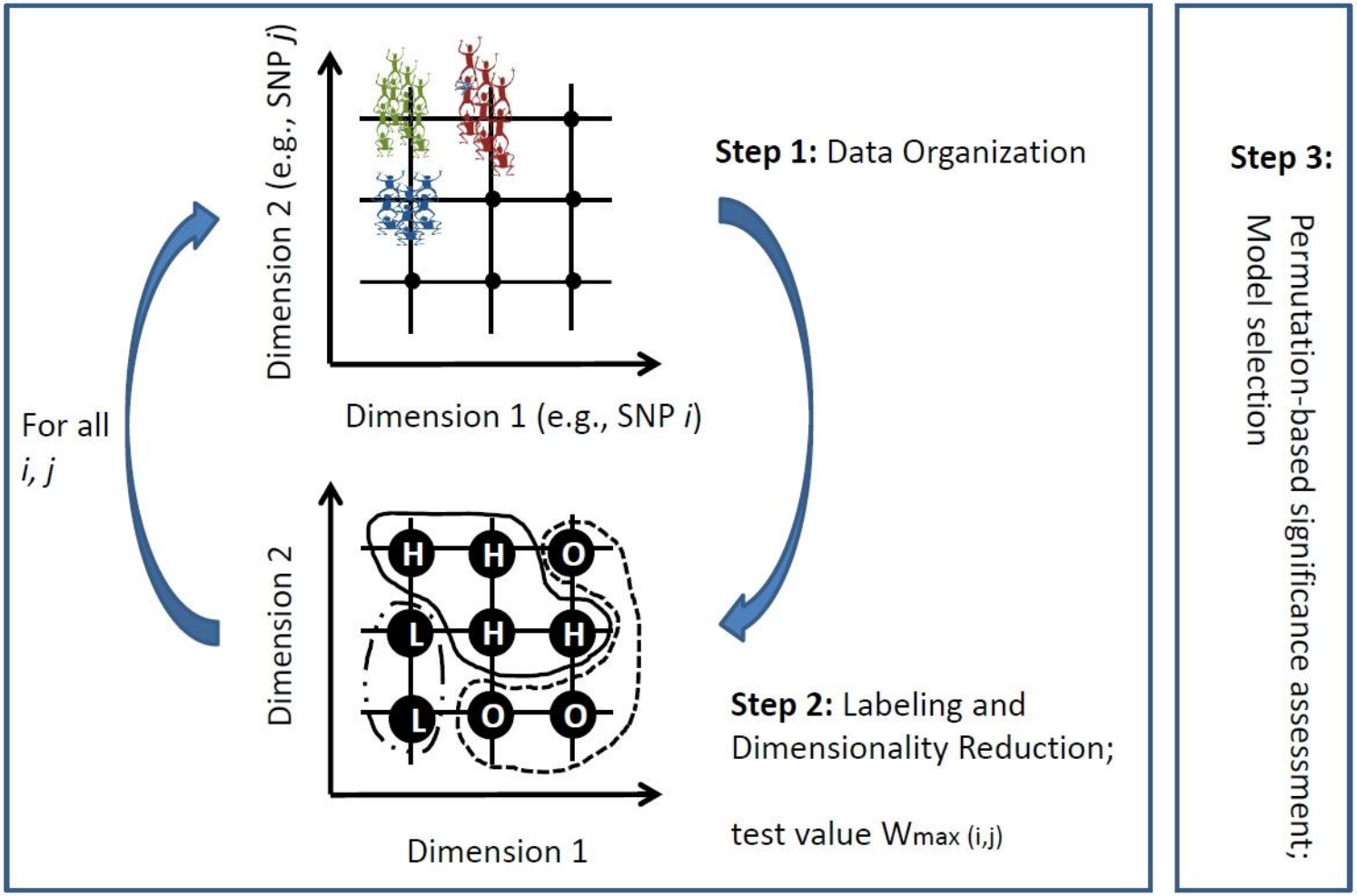
Methods	Accuracy
Deep learning	**68.78**
RF	55.85
LR	67.07
Naïve Bayes	62.68
GBM	65.85

- Interactions can be detected in neural networks by interpreting the network parameters, assuming that any input features interacting with each other must follow strongly weighted connections before the final output (Tsang et al. 2017).



(Tsang et al. 2017)

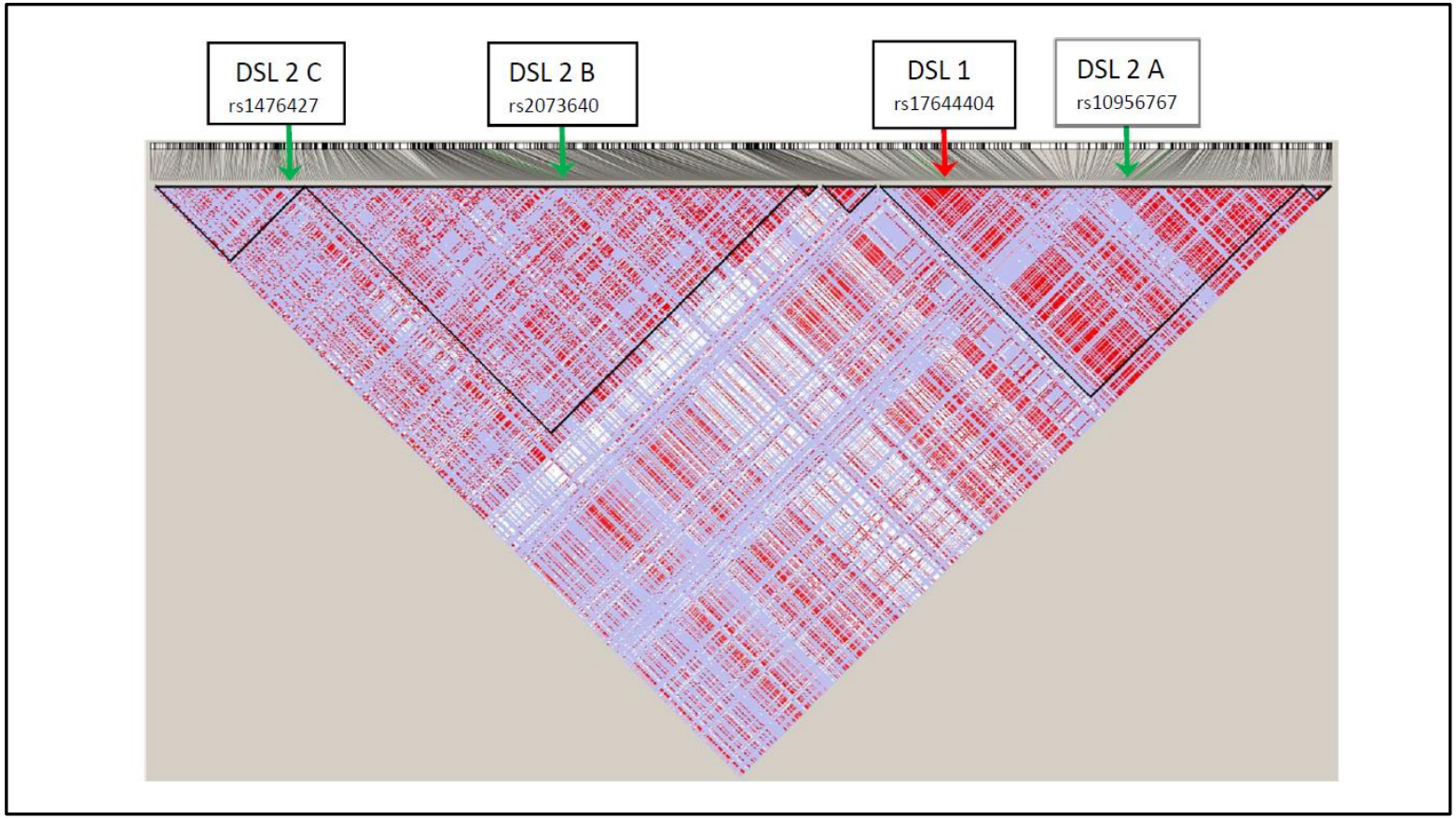
Model-Based MDR (BIO3 team – 2010+)



MB-MDR and MDR are conceptually different (BIO3 team – 2010+)

- Computation time is invested in
  - optimal **association tests** to label multi-locus genotype combinations and
  - in statistically valid permutation-based methods to assess **joint statistical significance** of multiple SNP pairs
- Labels are related to substantially improve/worsen trait values (H/L). In case there is **no** such **evidence**, the multi-locus label is not forced to be H or L (but will be O).
- In the **presence of main effects**, MB in MB-MDR ensures false positive control at 5%

Highly correlated features



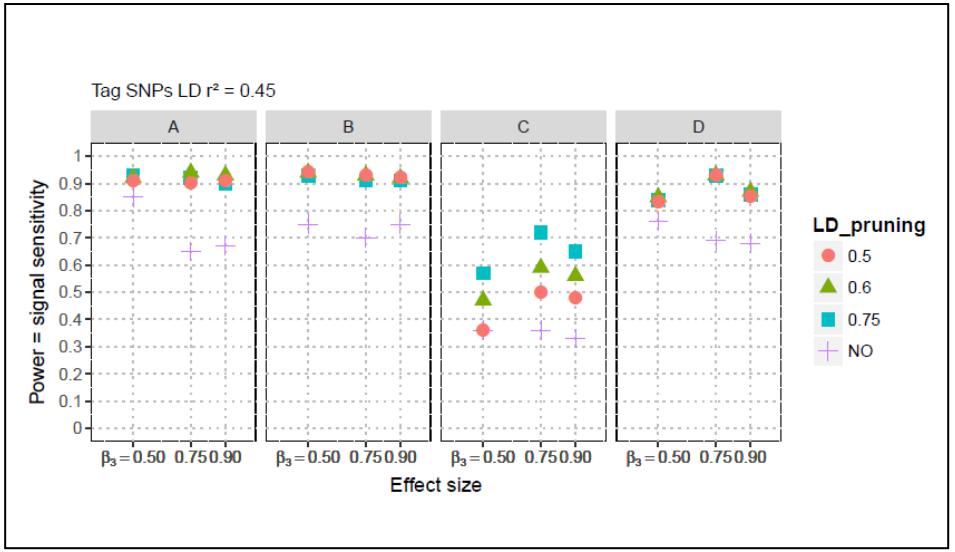
(Marc Joiret – 2017 BIO3 intern)

Highly correlated features

- Machine learning and Correlation based Feature Selection (CFS)
- Results (Hall 1999):
  - quickly identifies and screens irrelevant, redundant, and noisy features
  - classification accuracy using the reduced feature set equals or improves over using the complete feature set
  - degradation when important actors were not selected

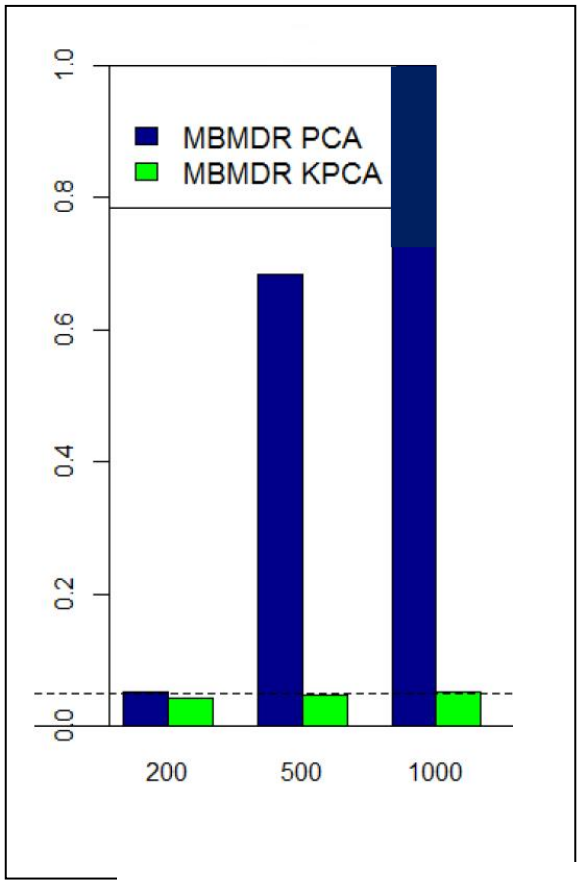
Highly correlated features

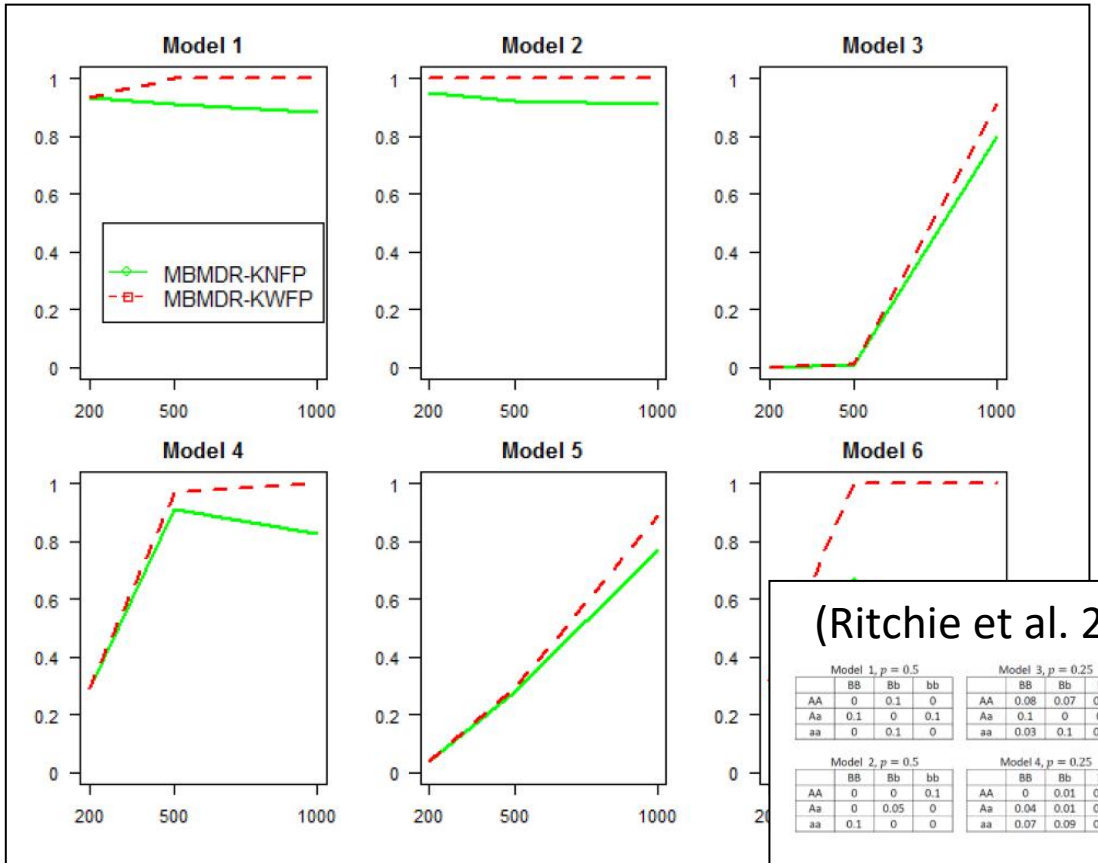
- Statistics and Linkage Disequilibrium (LD) pruning
- Results (Marc Joiret – 2017 intern BIO3):
  - Exact signal sensitivity may be low when actual actors were pruned out
  - No pruning gives the lowest signal sensitivity
  - Sufficient pruning gives acceptable signal sensitivity
  - Lowest power when DSLs reside at the boundaries of LD regions (scenario C)



(Non-linear) confounders

(Fouladi et al. 2016+ ; Abegaz et al. 2016+)





(Ritchie et al. 2003)

Model 1, $p = 0.5$

	BB	Bb	bb
AA	0	0.1	0
Aa	0.1	0	0.1
aa	0	0.1	0

Model 3, $p = 0.25$

	BB	Bb	bb
AA	0.08	0.07	0.05
Aa	0.1	0	0.1
aa	0.03	0.1	0.04

Model 5, $p = 0.1$

	BB	Bb	bb
AA	0.07	0.05	0.02
Aa	0.05	0.09	0.01
aa	0.02	0.01	0.03

Model 2, $p = 0.5$

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

Model 4, $p = 0.25$

	BB	Bb	bb
AA	0	0.01	0.09
Aa	0.04	0.01	0.08
aa	0.07	0.09	0.03

Model 6, $p = 0.1$

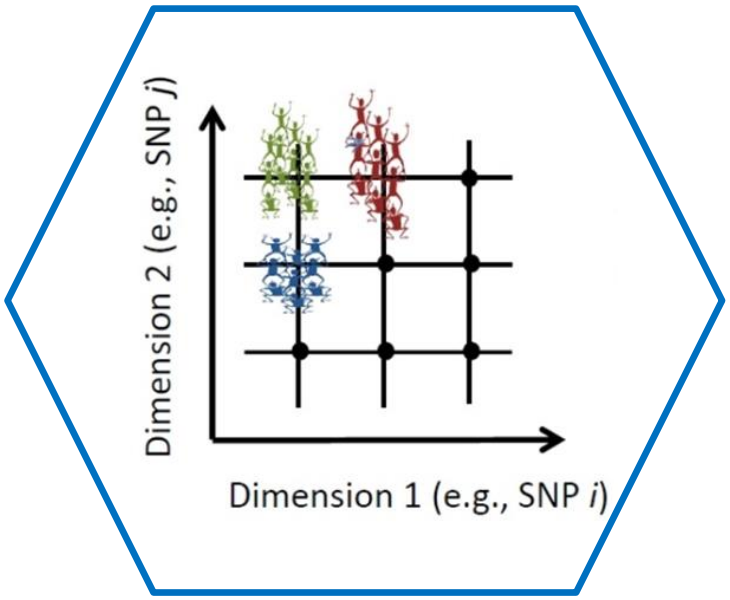
	BB	Bb	Bb
AA	0.09	0.001	0.02
Aa	0.08	0.07	0.005
aa	0.003	0.007	0.02

Above : 60/40 CC ratio, structural epistasis according to corresponding full penetrance Rtichie epistasis model ; Below : 50/50 (200+200)

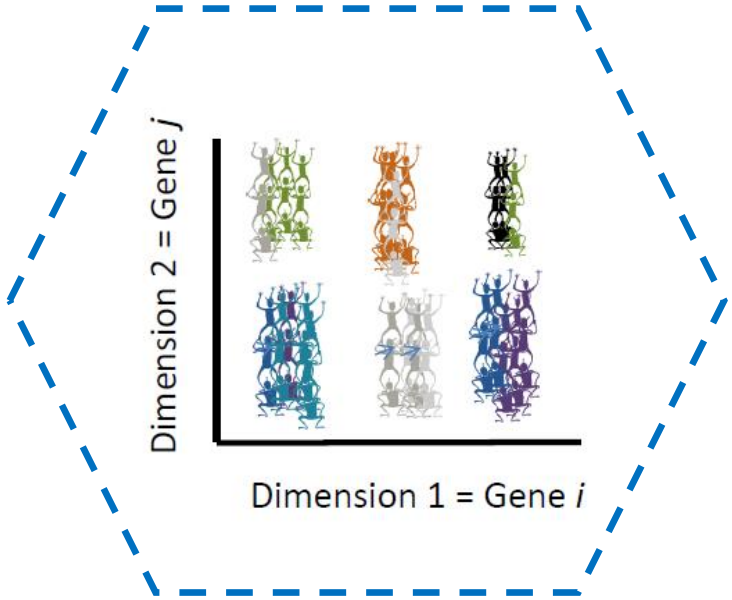
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Noise	MB	MDR	MB	MDR	MB	MDR	MB	MDR	MB	MDR	MB	MDR
None	100	99	100	100	100	95	100	93	93	62	97	73

(Cattaert et al. 2011)

Replication and validation



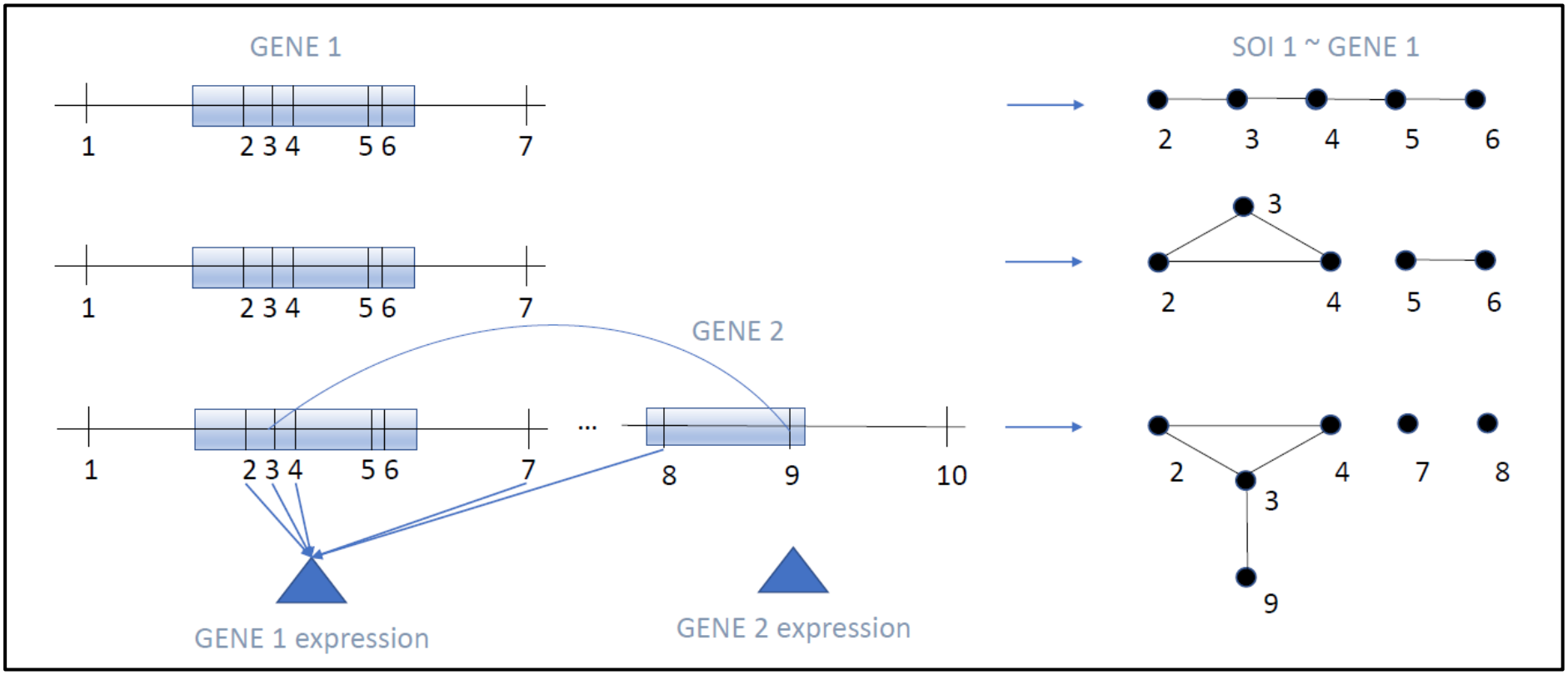
MB-MDR in integrative context



- Component-based
- Kernel-based
- Network-based

(Fouladi et al. 2015 ; 2016+)

Replication and validation



or having it trained from the data?

Computational feasibility

Multiple testing correction via “MAXT” in MBMDR-3.0.3:

SNPs	Sequential version Binary trait	Sequential version Continuous trait	Parallel workflow Binary trait	Parallel workflow Continuous trait
10^2	45 sec	1 min 35 sec	<1sec	<1sec
10^3	1 hour 16 min	2 hours 39 min	38 sec	1 min 17 sec
10^4	5 days 13 hours	11 days 19 hours	1 hour 3 min	2 hours 14 min
10^5	≈ 1.5 year	≈ 3 years	4 days 9 hours	≈ 9 days

The parallel workflow was tested on a cluster composed of 10 blades, containing each four Quad-Core AMD Opteron(tm) Processor 2352 2.1 GHz. The sequential executions were performed on a single core of this cluster. The results prefixed by the symbol “≈” are extrapolated.

(Van Lishout et al. 2013)

Computational feasibility: approximating vs exact

Multiple testing correction via “gammaMAXT” in MBMDR-4.2.2:

SNPs	Sequential version Binary trait	Parallel workflow Binary trait	Sequential version Continuous trait	Parallel workflow Continuous trait
10^3	13 min 33 sec	20 sec	13 min 18 sec	18 sec
10^4	52 min 15 sec	1 min 05 sec	56 min 14 sec	53 sec
10^5	64 hours 35 min	22 min 15 sec	70 hours 03 min	20 min 28 sec
10^6	$\approx$ 270 days	25 hours 12 min	$\approx$ 290 days	24 hours 06 min

The parallel workflow was tested on a 256-core computer cluster (Intel L5420 2.5 GHz 1333 MHz FSB). The sequential executions were performed on a single core of this cluster. The results prefixed by the symbol “$\approx$” are extrapolated.

(Van Lishout et al. 2015)

Conclusion

Imagine a world …

- in which machine learning taxonomy addresses an interdisciplinary community
- in which missing data handling strategies hold, despite sample heterogeneity
- in which multi-omics summaries can be learned from data
- in which confounding information is adequately described or accounted for
- in which disease prediction can be extended to accommodate a latent spectrum of diseases or a continuum of disease presentations
- in which neural network parameters aid in deriving meaningful/relevant relationships

Hippocrates (460-370 BC):

"It's far more important to know what person the disease has than what disease the person has."

Acknowledgements

GIGA-R, Medical Genomics Thematic Research Unit, Liège, Belgium

http://bio3.giga.ulg.ac.be/