Translational Systemics

for

Precision Medicine

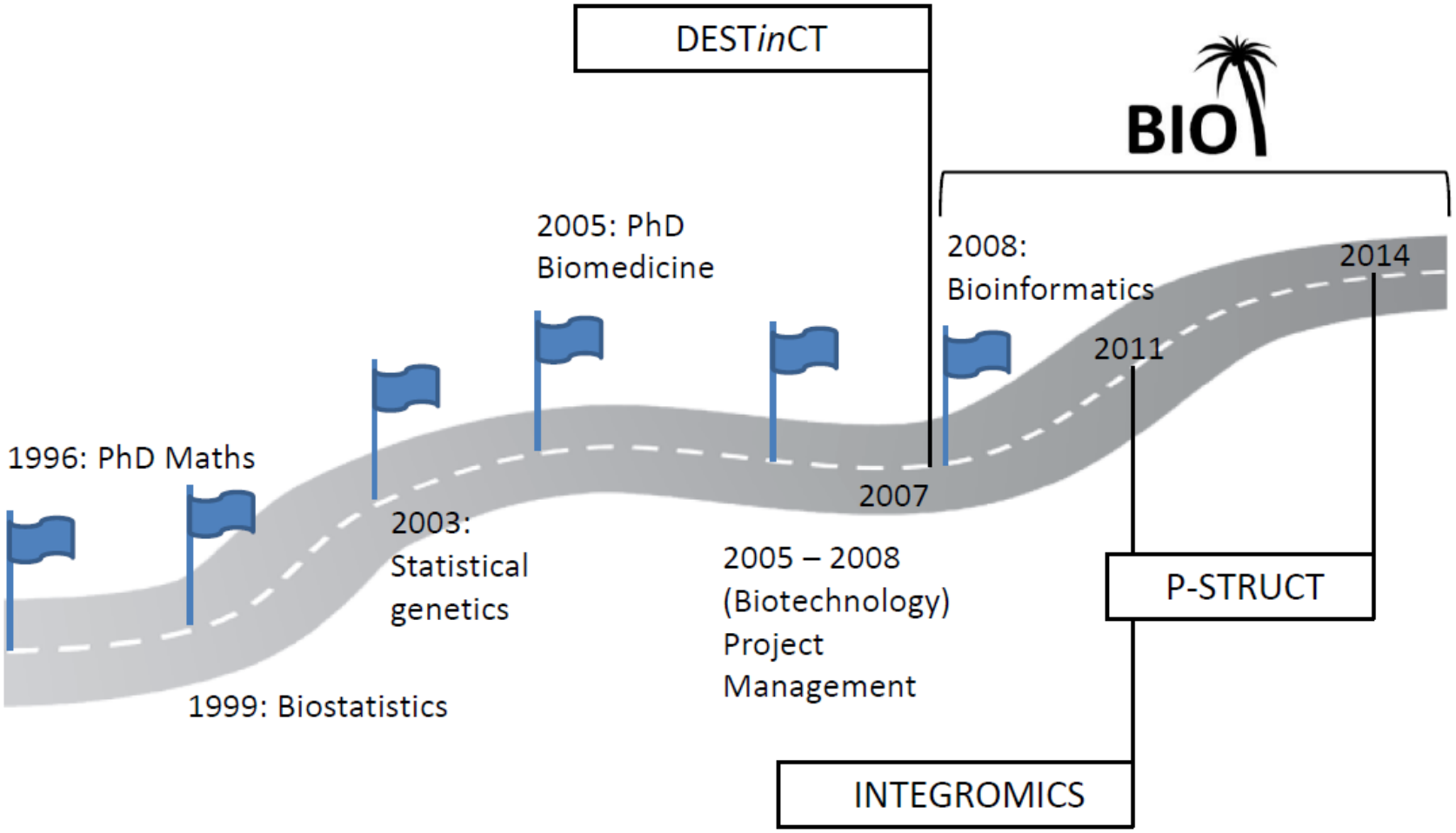
Kristel Van Steen, PhD[2] (*)

kristel.vansteen@ulg.ac.be

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

Systems Medicine Lab, KU Leuven, Belgium

BIO


DESTinCT
BIO
2005: PhD
Biomedicine
2008:
Bioinformatics
2014
2011
1996: PhD Maths
2003:
Statistical
genetics
2007
2005 – 2008
(Biotechnology)
Project
Management
1999: Biostatistics
P-STRUCT
INTEGROMICS

OUTLINE

- Precision medicine
- Practical implementations
- Analytical considerations
- Take-home messages

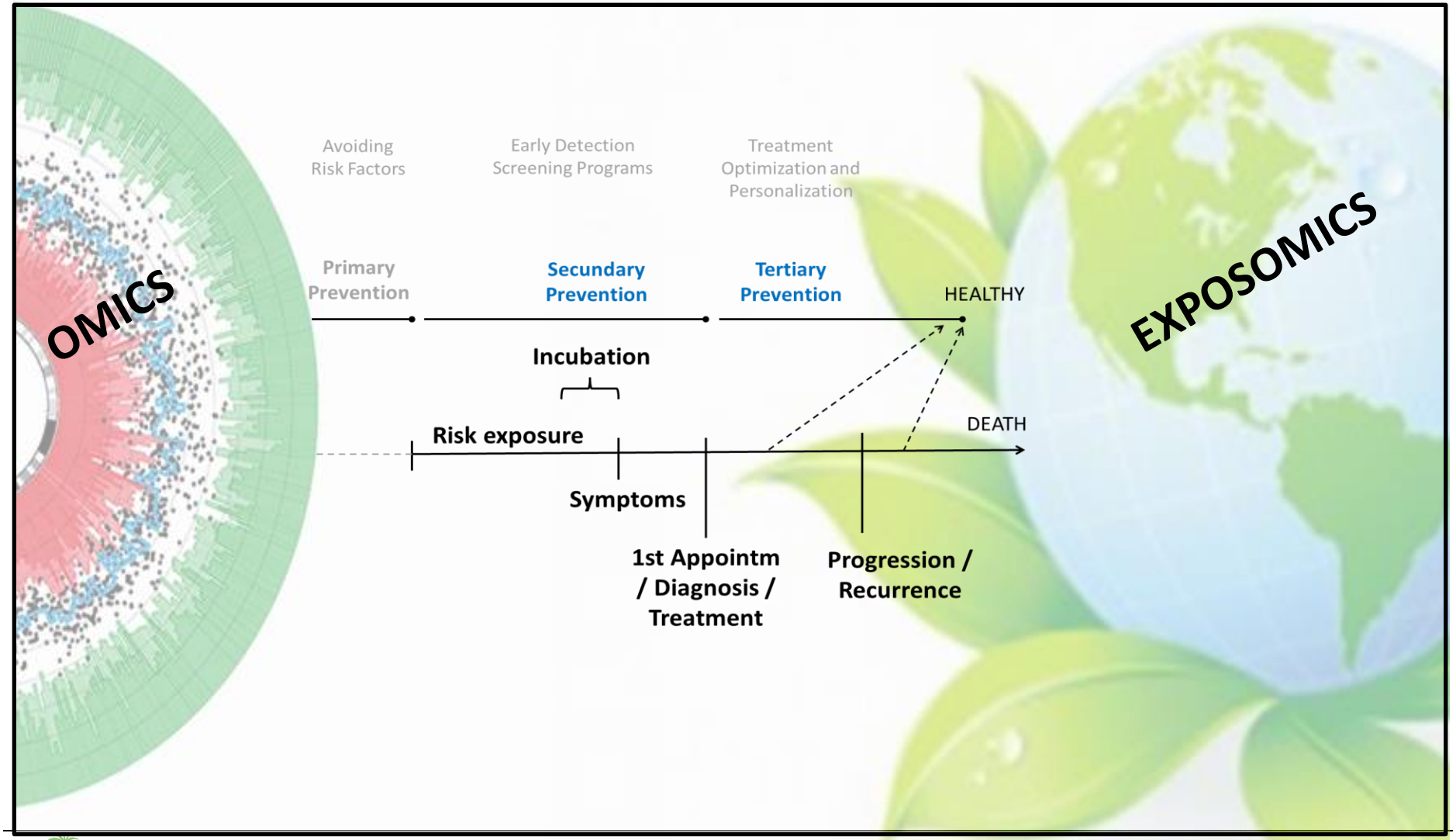
Precision Medicine

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."

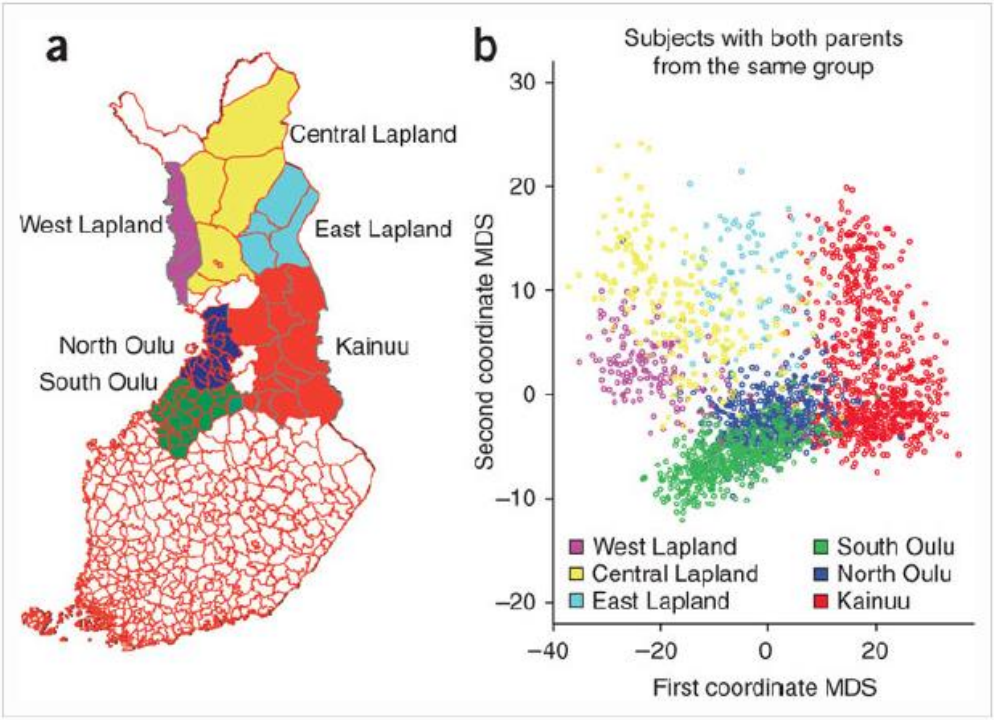
(HORIZON2020 Advisory Group)

An individual's ecosystem comes with a data deluge



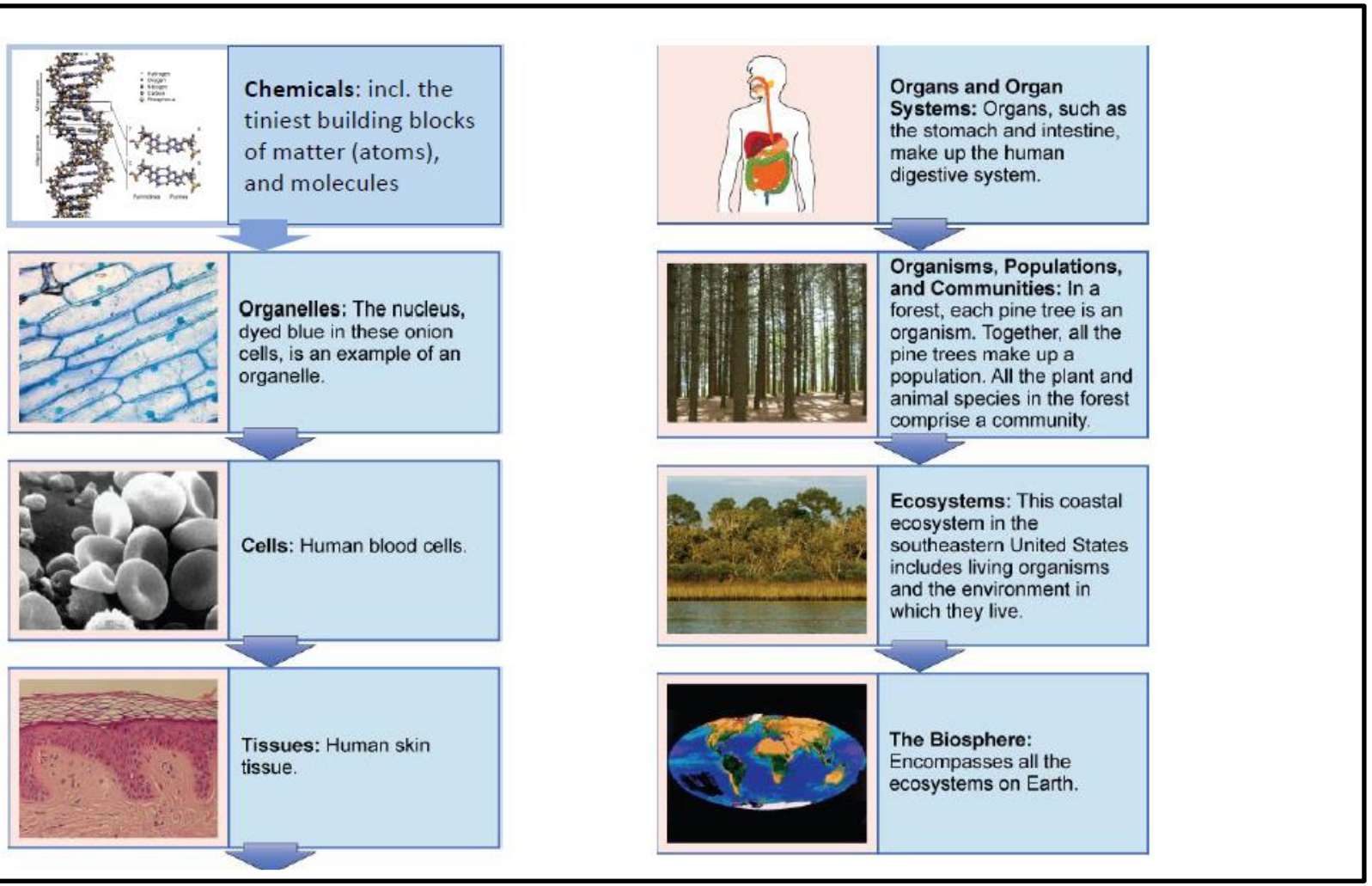
Population heterogeneity

- There can be population structure in all populations, even those that appear to be relatively “homogeneous”

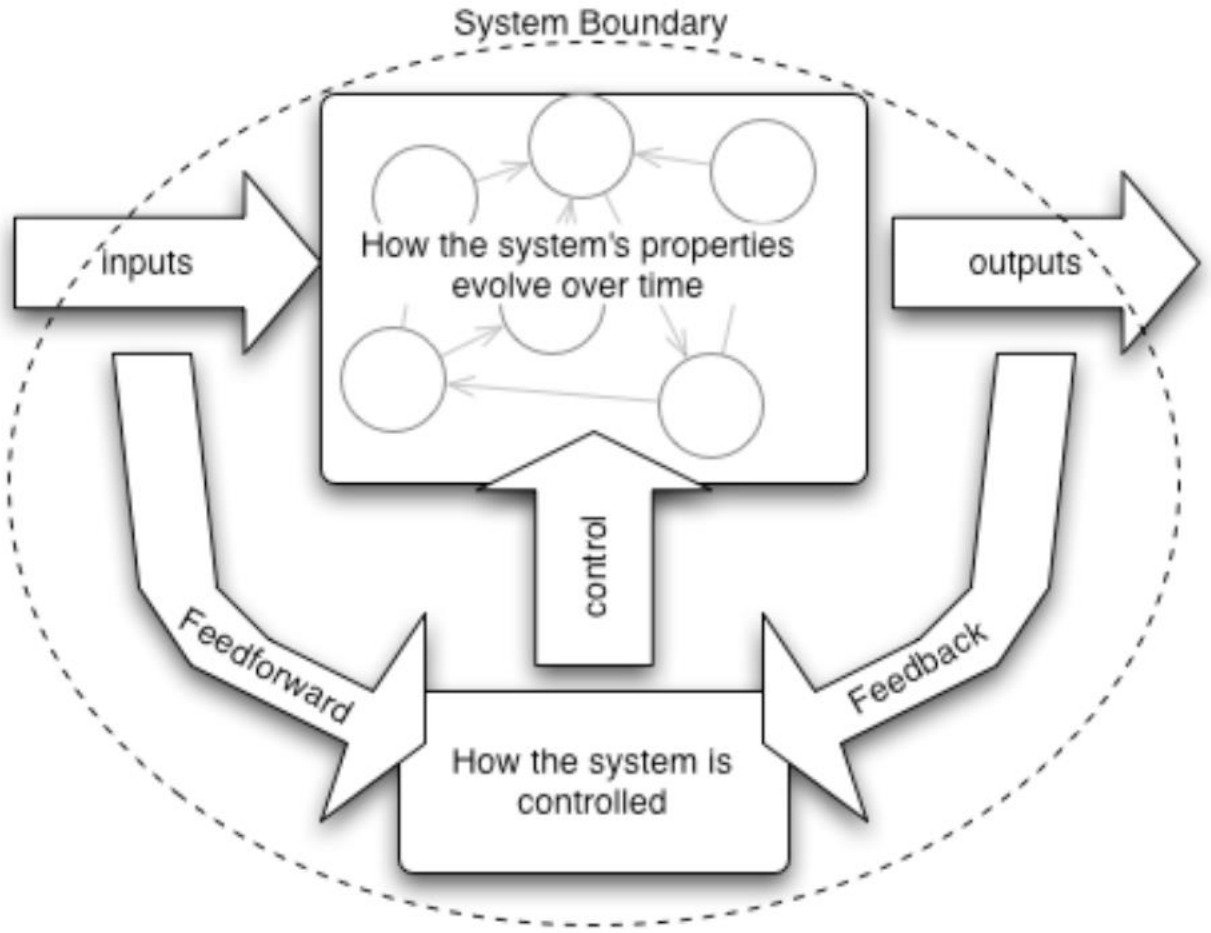


(Sabatti et al. 2009)

Patient heterogeneity

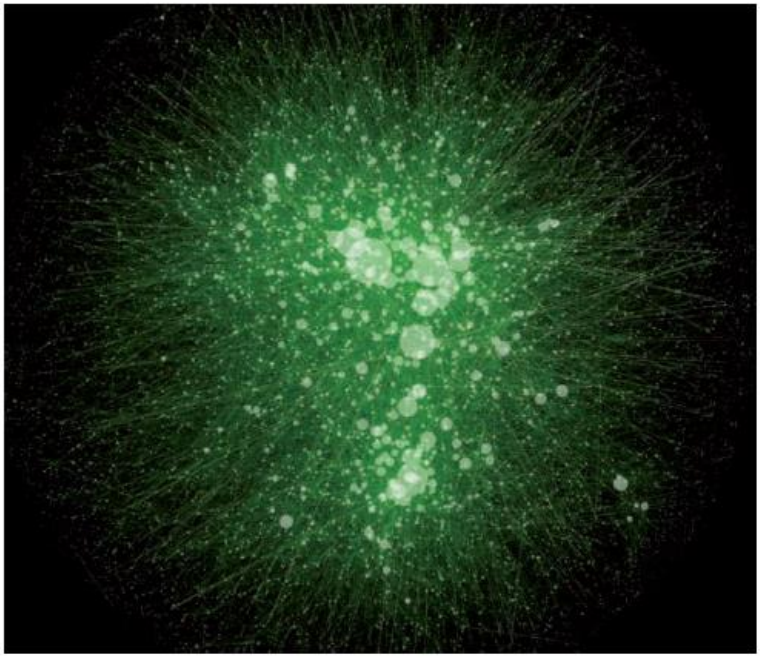


Interconnectivity: Systems and their eco-system



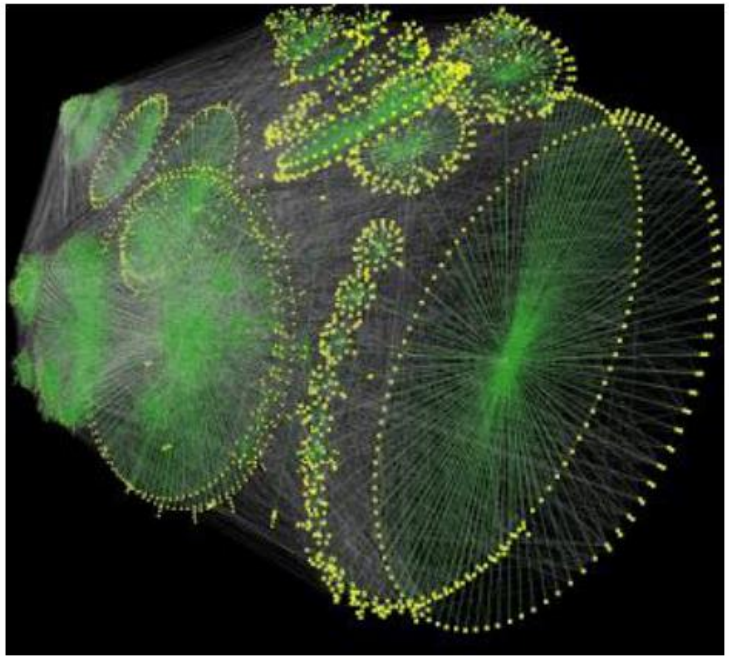
(@2004-5 Steve Easterbrook)

Interactions



Human interactome (PPI)

(Bonetta 2010)



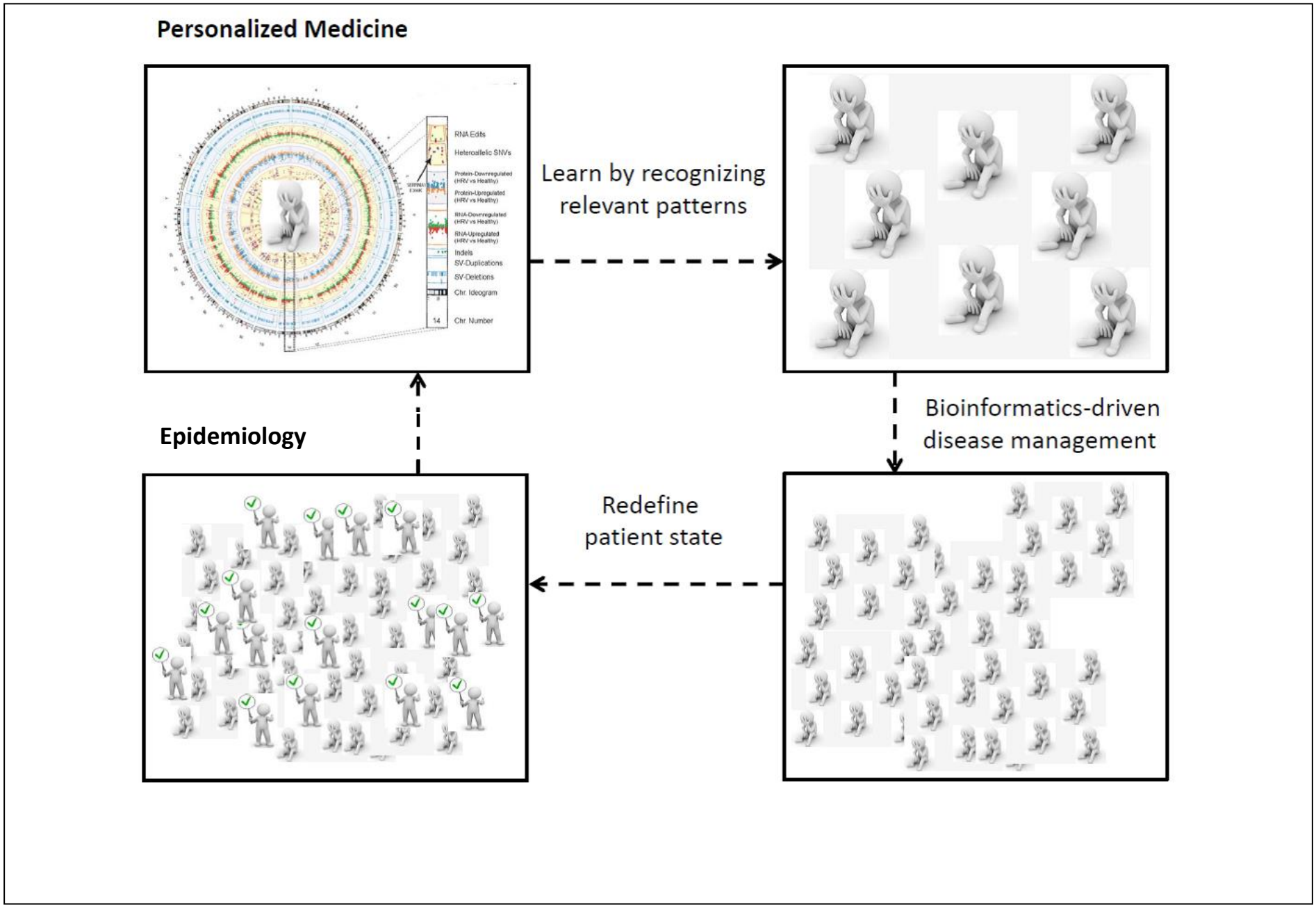
Fruit fly interactome

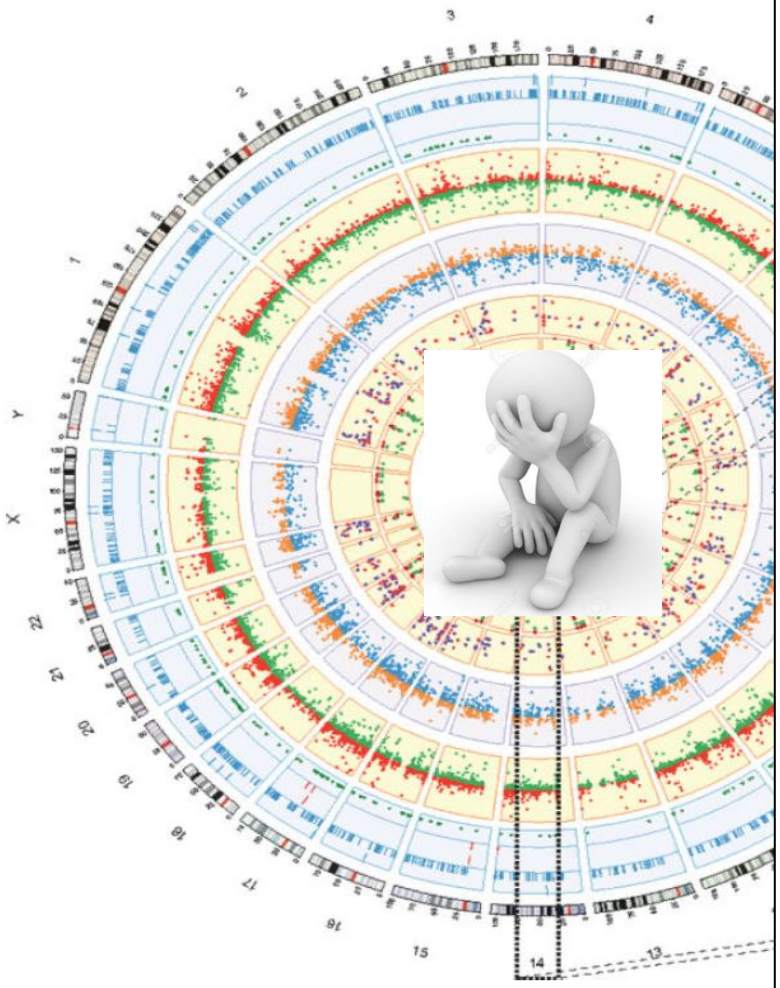
(www.molgen.mpg.de)

Reminder: “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***.

Precision medicine: practical implementation


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?

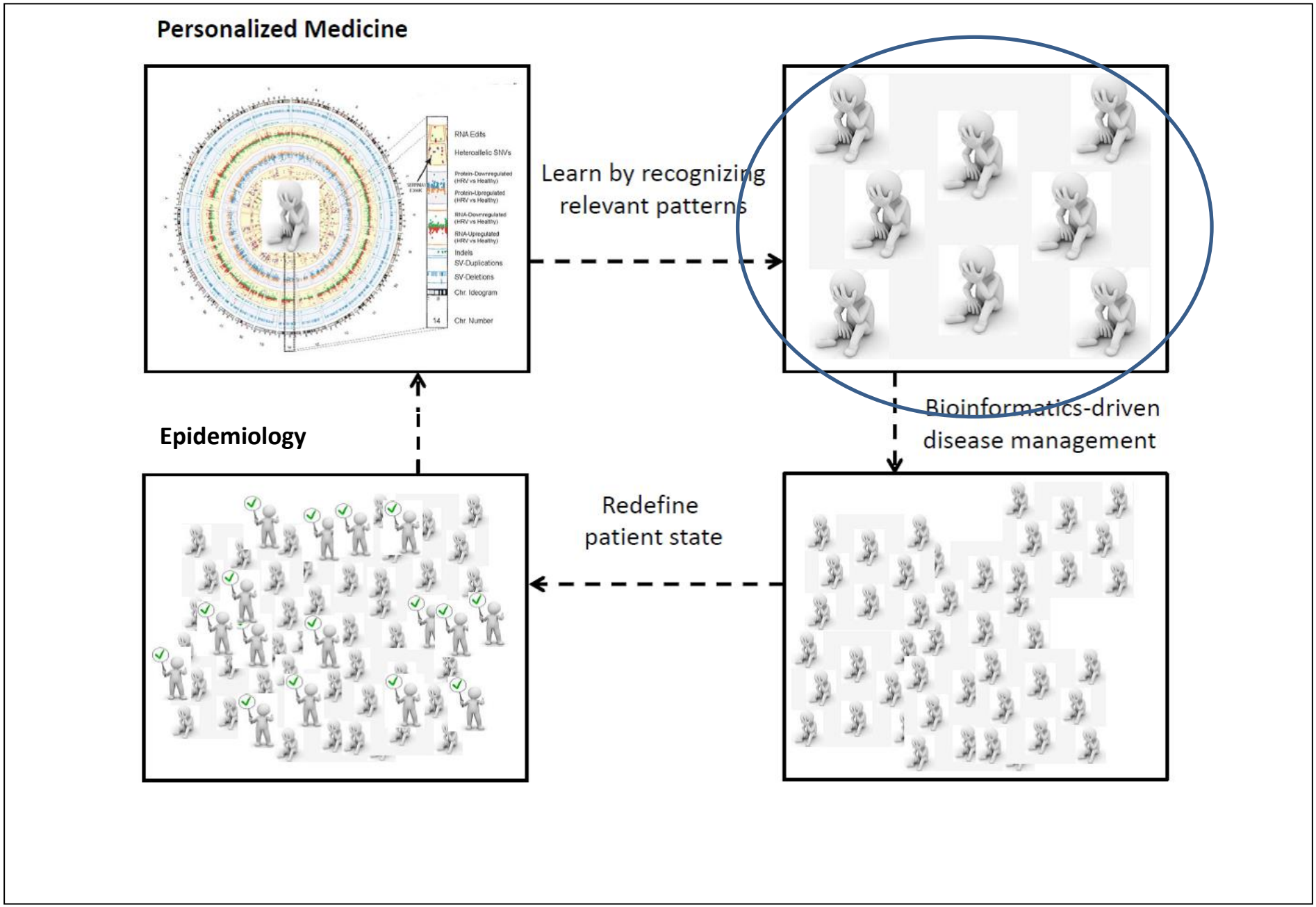
"Not in the form we are doing it. At the moment we have a very incomplete picture of what's going on, whereas if we were able to make thousands of measurements we would have a much better feeling. We just don't know, for the clinical tests, which thousand measurements are going to be most useful. We'll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

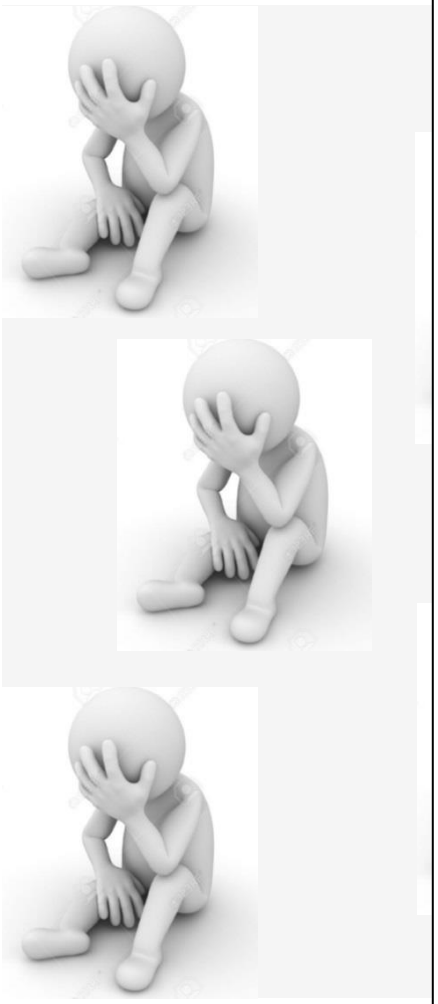
(Snyder 2014)

Redundancy – Informativity ?

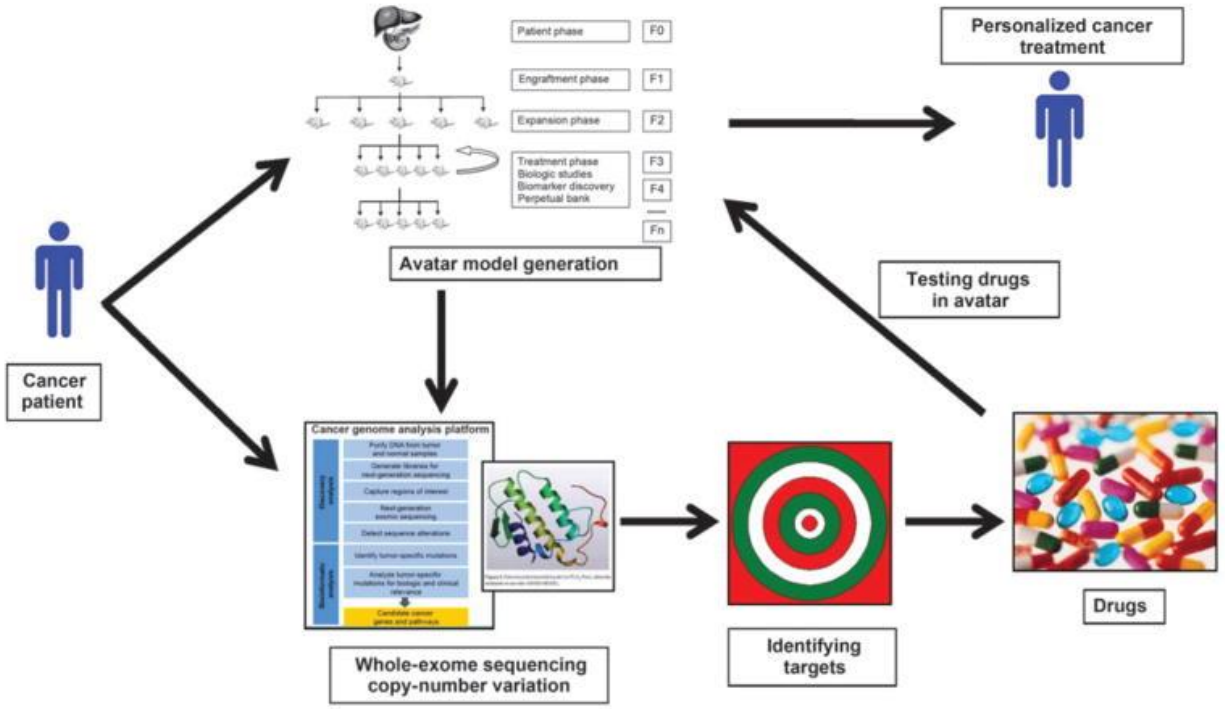
Feature selection

- In machine learning contexts, our goal may be to reduce the number of dimensions without losing “predictive” power:
  - exhaustive search
  - random feature selection
  - minimum redundancy maximum relevance feature selection
  - simultaneous learning feature representation and cluster assignment using deep learning network (Xie et al. 2016)
- Machine learning *can* benefit from redundancy
  - leading to increased performance by adding robustness (Lorenzen 1999)


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
Chr. Number
Learn by recognizing relevant patterns
Bioinformatics-driven disease management
Redefine patient state
Epidemiology



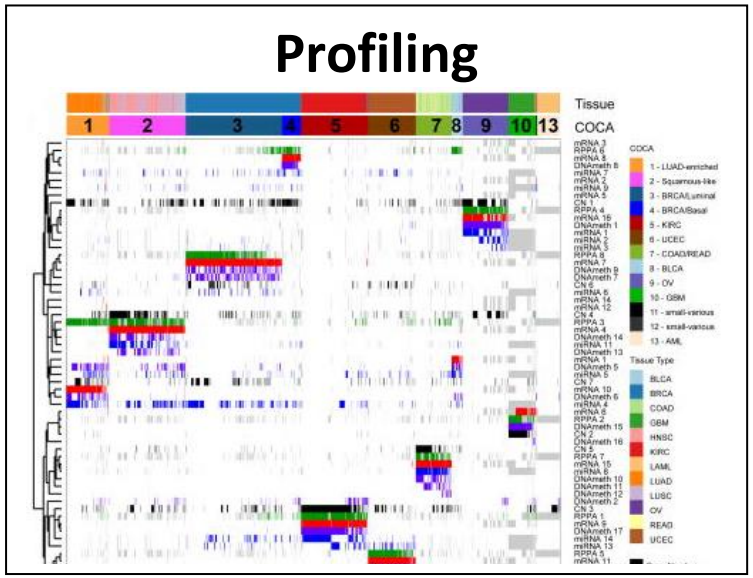
Integrating sequencing and avatar mouse models



Missingness ?

(Garralda et al. 2014)

More complex example: multiplatform profiling

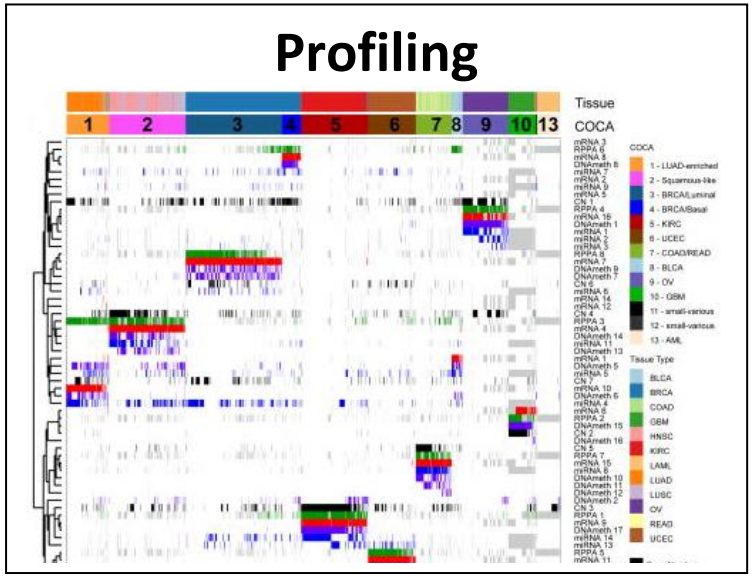


(Hoadley et al. 2014 ; Pan-Cancer-12)

- **Integration** is the process of connecting systems (which may have fusion in them) into a larger system (Oxley & Thorsen, 2004)
- A **trans-disciplinary approach** should provide generic frameworks and should provide organizing principles for the interaction of diff. types of analytics (Van Steen, Cluj, 2015)

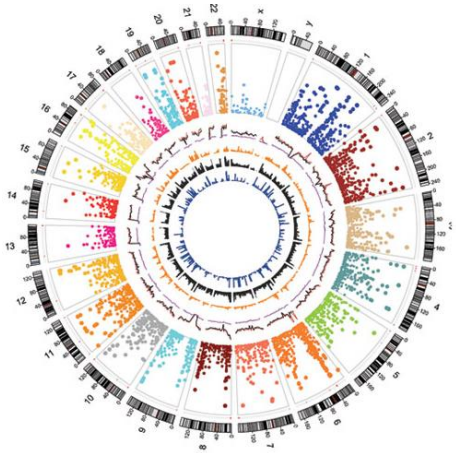
Different routes lead to … EORTC

Profiling



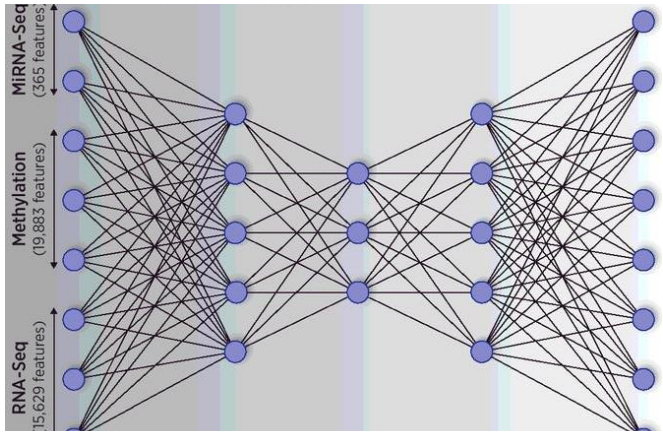
(Hoadley et al. 2014; Consensus Clust)

Comparing



(Jun Li et al. ’12; GWASrap)

Prediction



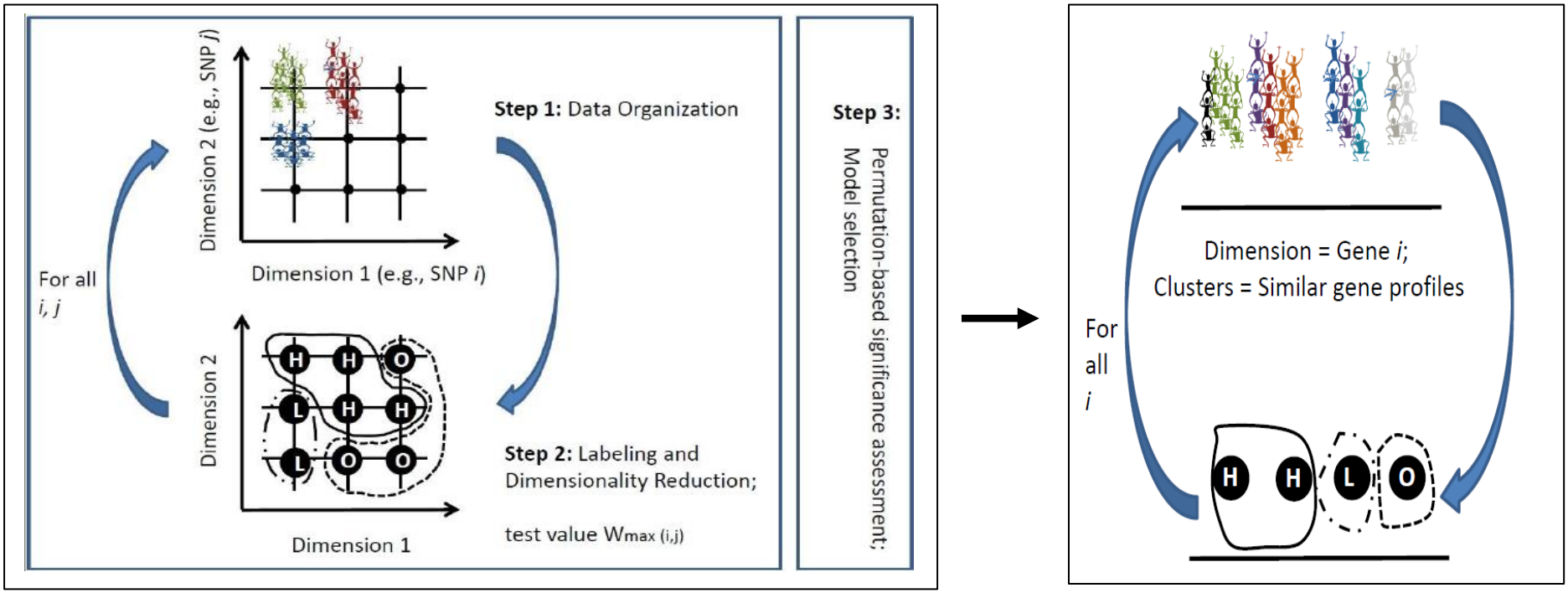
(Chaudhary et al. 2018; Deep Learning)

BIO

Different routes lead to … Singapore

- **Data integration** (heterogeneous data types) – WELL PROGRESSING

Ex: MB-MDR + diffusion kernels on graphs

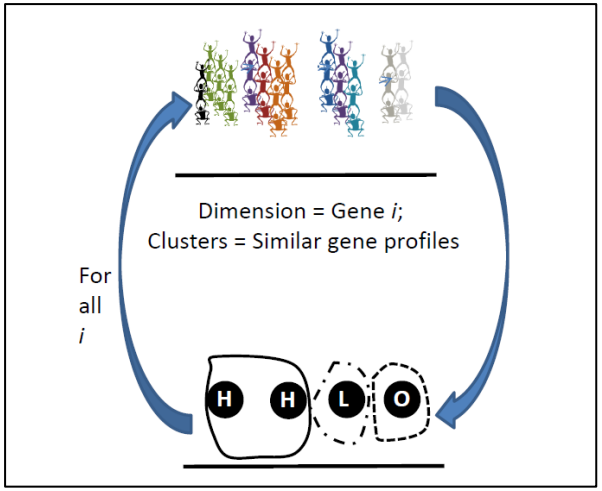


Different routes lead to … Singapore

- **Data integration** (heterogeneous data types) – WELL PROGRESSING

Ex: MB-MDR + diffusion kernels on graphs

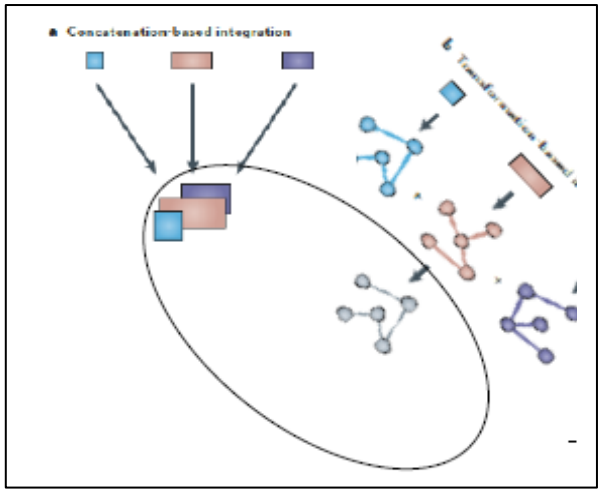
to perform omics-integrated gene-based sample clustering



(DEST*in*CT : MB-MDR)

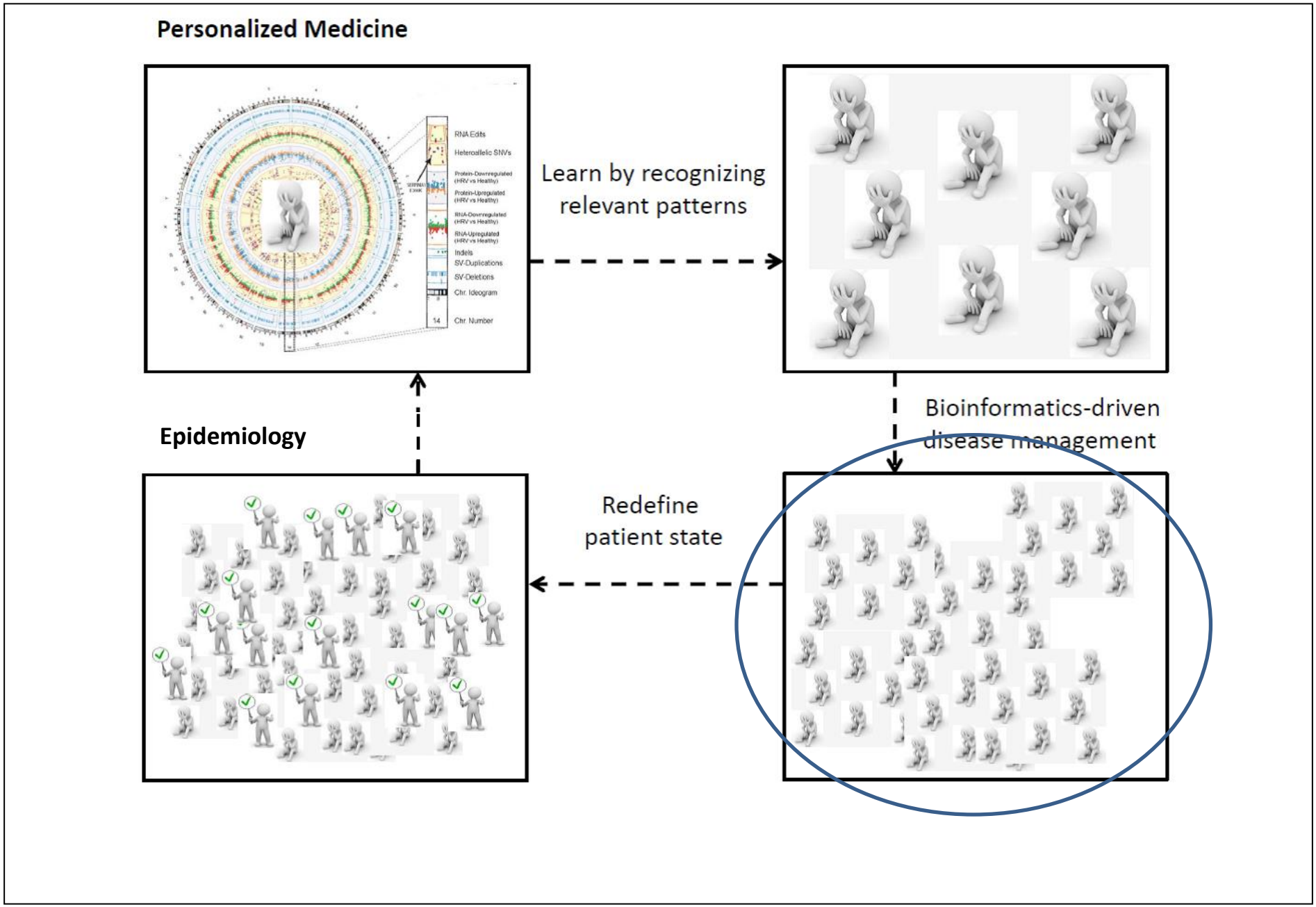
- Component-based
- Kernel-based
- Network-based

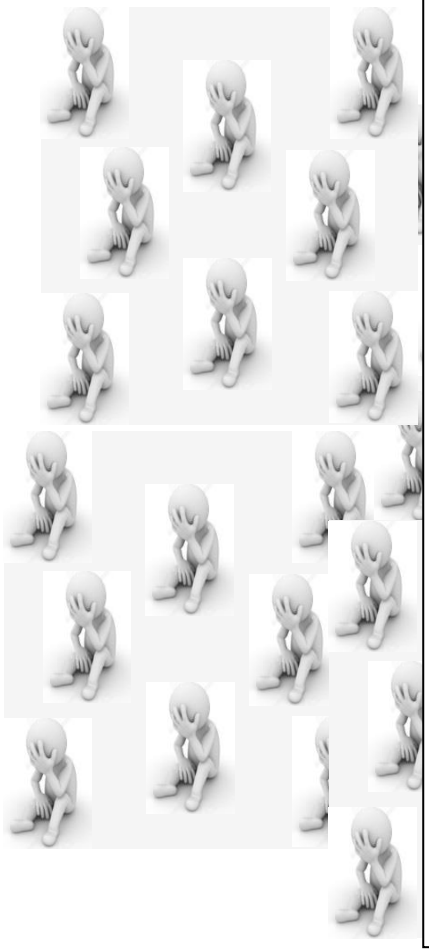
(Fouladi et al. 2015-2018)



(Ritchie et al. 2015)

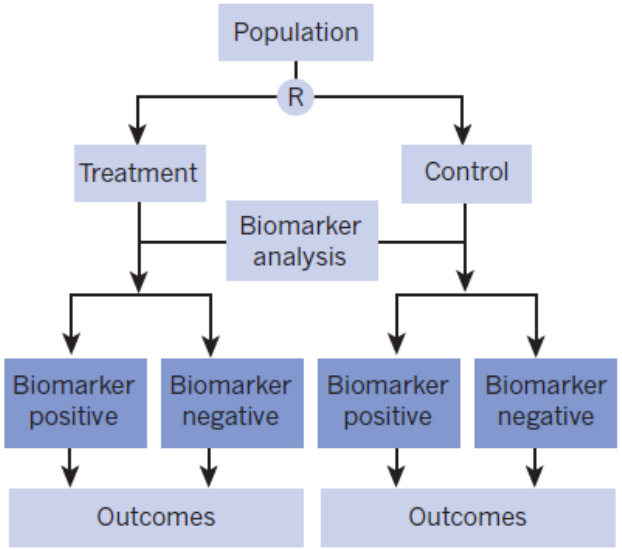
- **Analytic integration** (modelling paradigms) – INFANCY


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

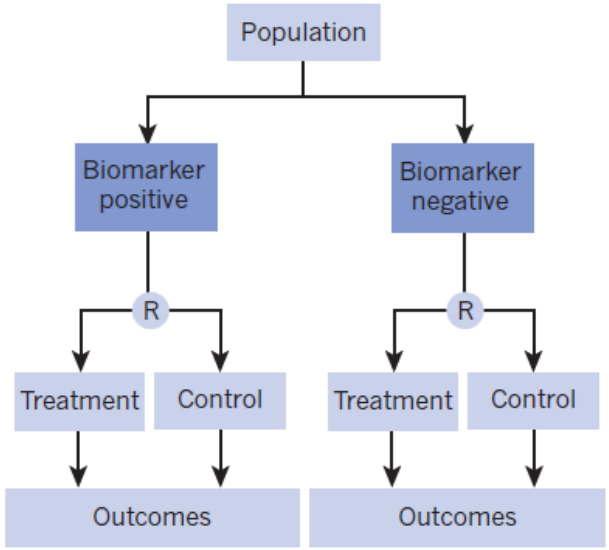


Testing precision-medicine strategies

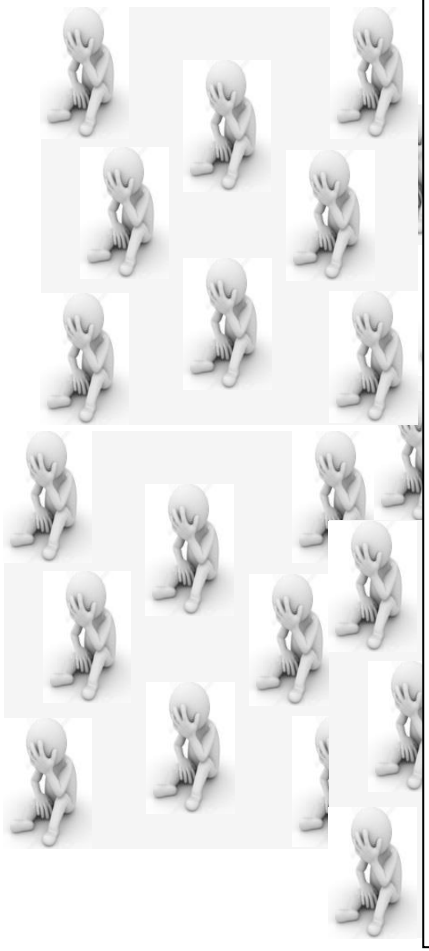
a Biomarker analysis within existing RCT



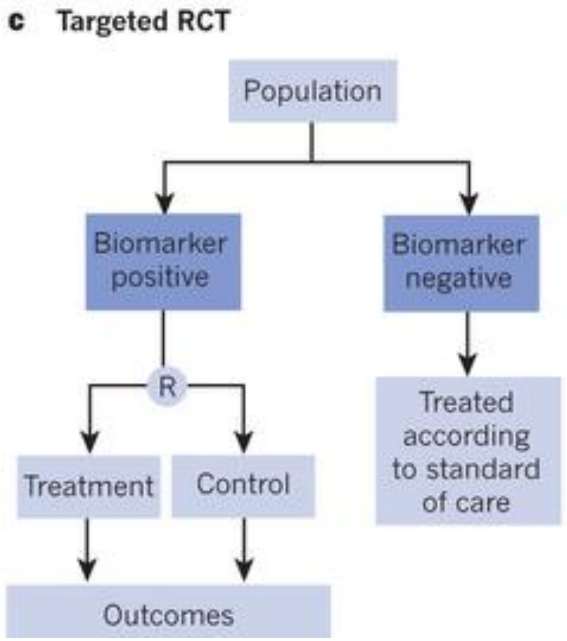
b Non-targeted RCT (stratified by biomarker)

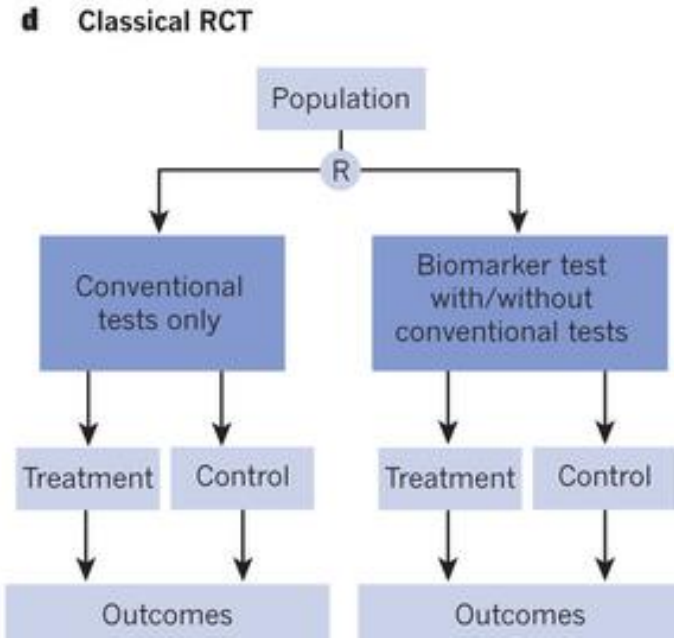


(Biankin et al. 2015)



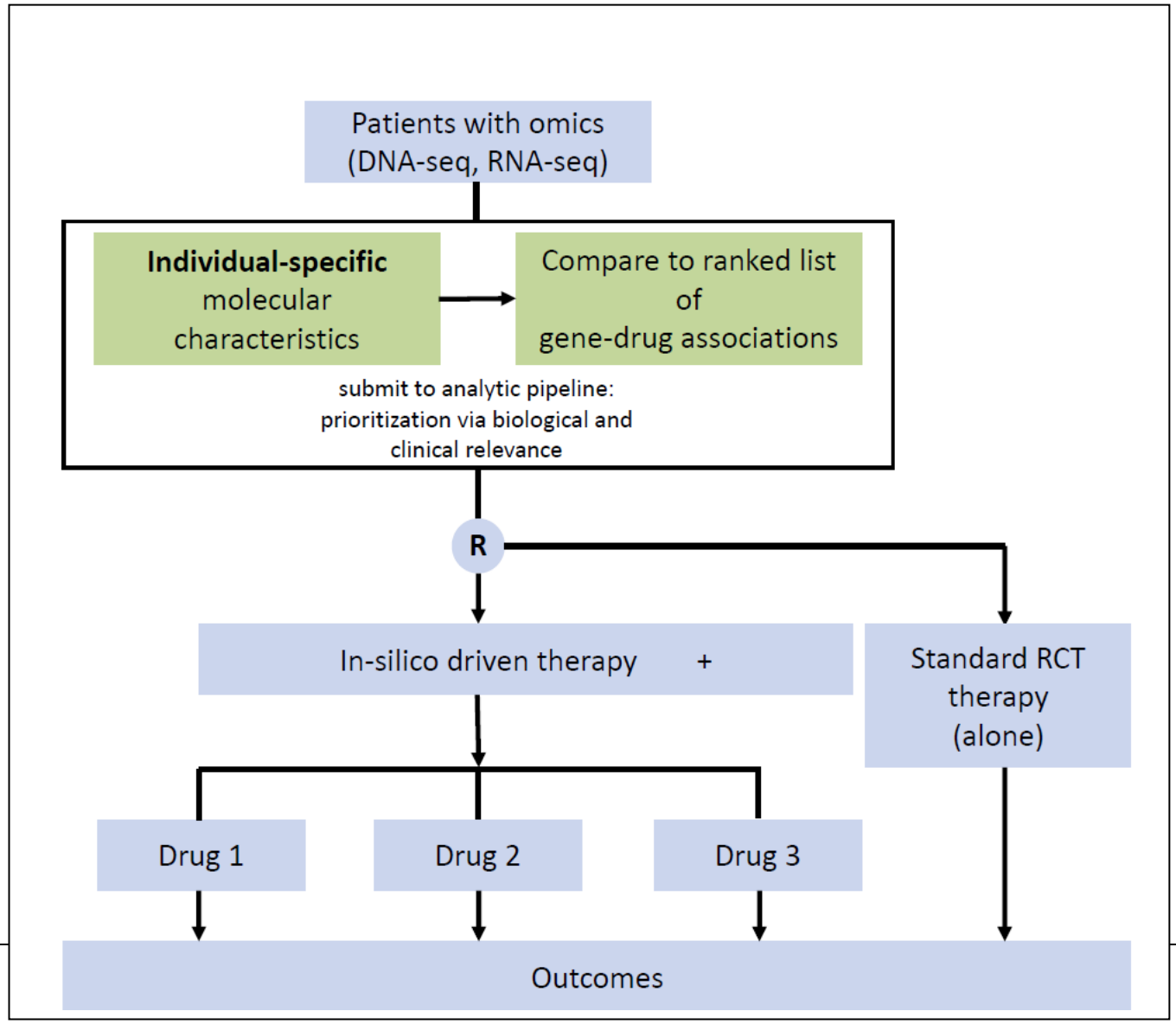
Testing precision-medicine strategies

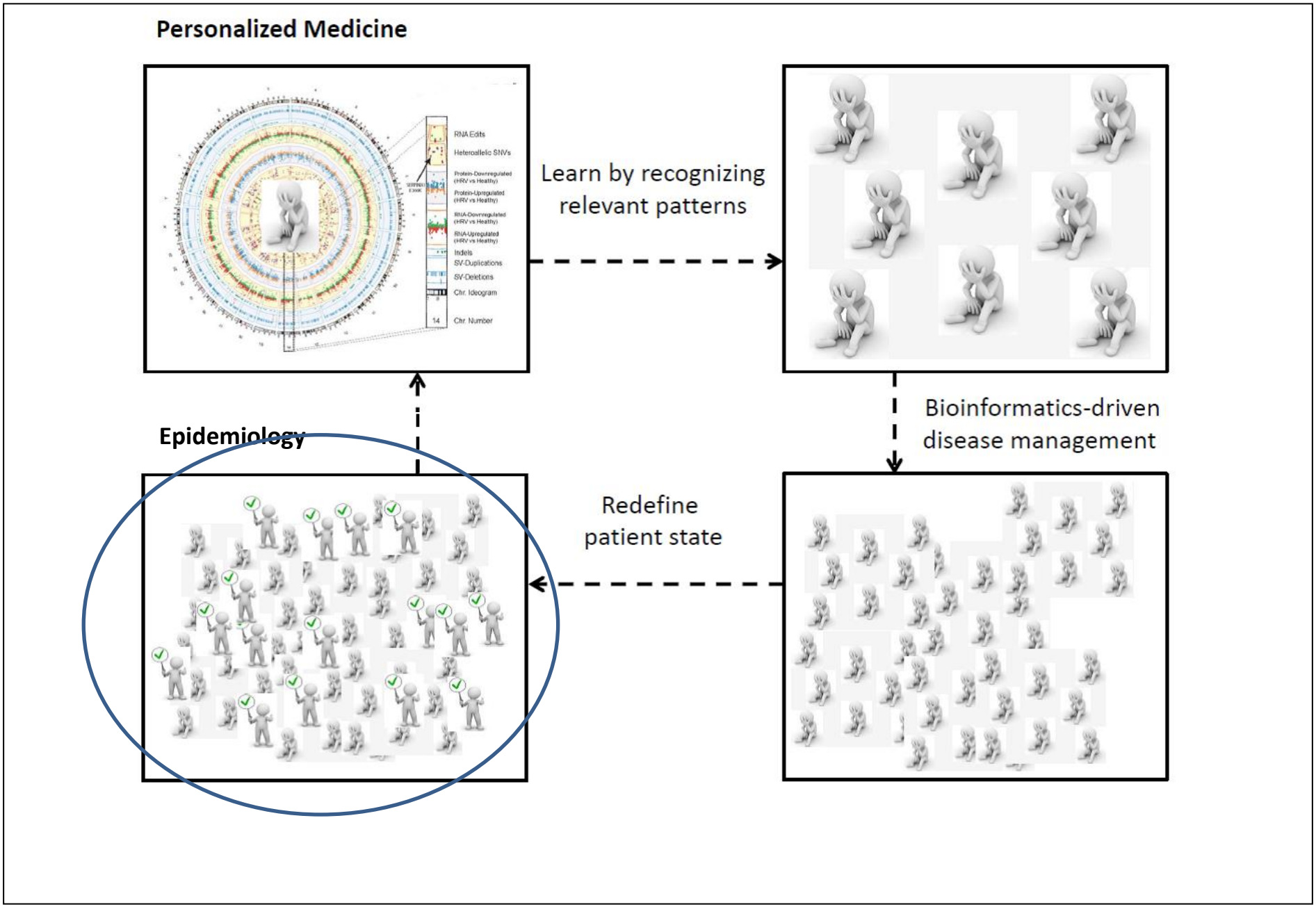


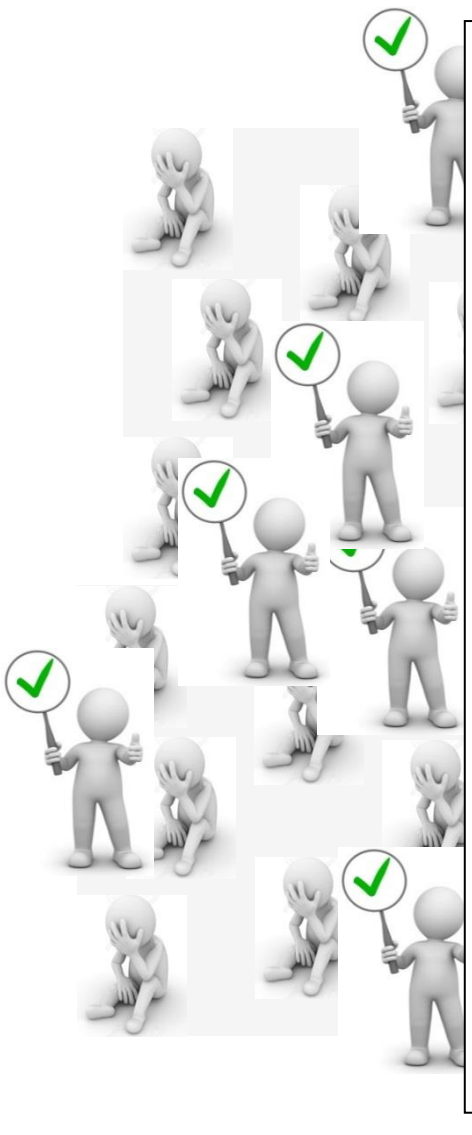


(Biankin et al. 2015)

Further specification of CTs in view of personalized medicine (c)




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
Epidemiology
Redefine patient state



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

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

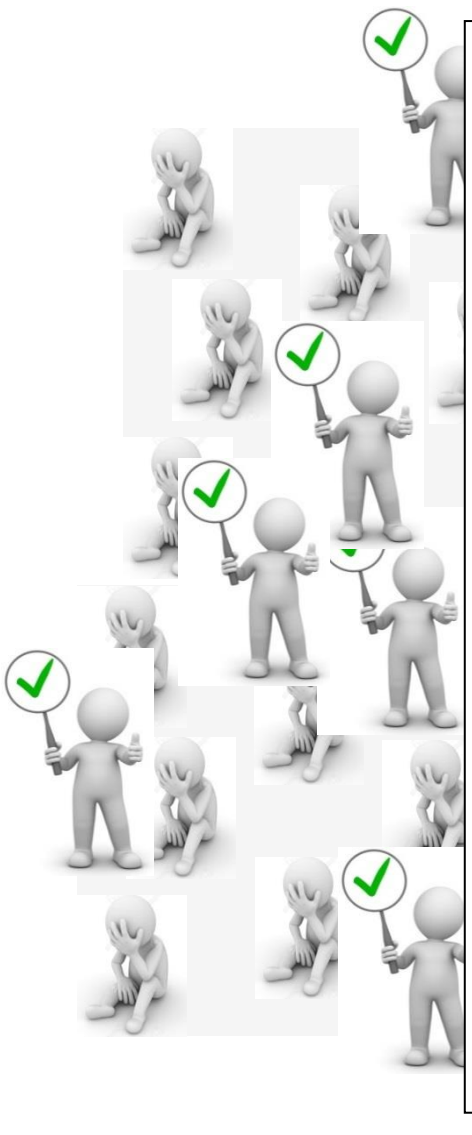
(Cleynen et al. 2012)

Heterogeneity as a target

Further specification of CTs in view of personalized medicine

- Basket CTs: multiple diseases with the same genetic mutation, randomized treatment allocation
- Umbrella CTs: 1 “disease”, different genetic mutations which define subcohorts, each receiving randomized treatment regimen
- Adding complexity:
  - cellular heterogeneity - assign based on the mutation detected in the higher percentage of cancer cells?
  - highly multi-dimensional profiles

(Sumitrhra Mandrekar,
INSERM atelier 248, Bordeaux, 2017)



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

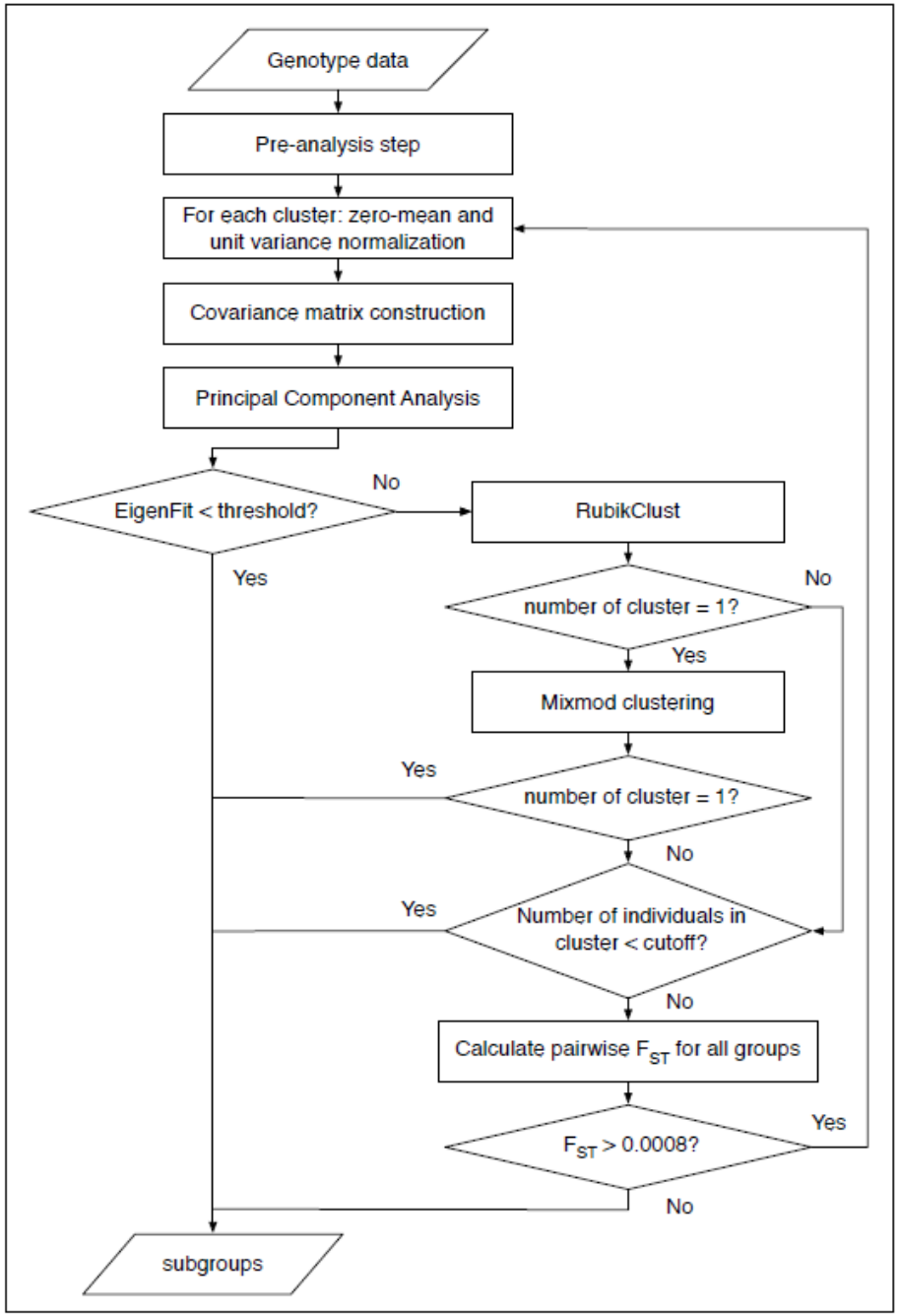
Bärbel Maus[1,2*], Camille Jung[3,4,5], Jestinah M. Mahachie John[1,2], Jean-Pierre Hugot[3,4,6], Emmanuelle Génin[7,8], Kristel Van Steen[1,2]

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

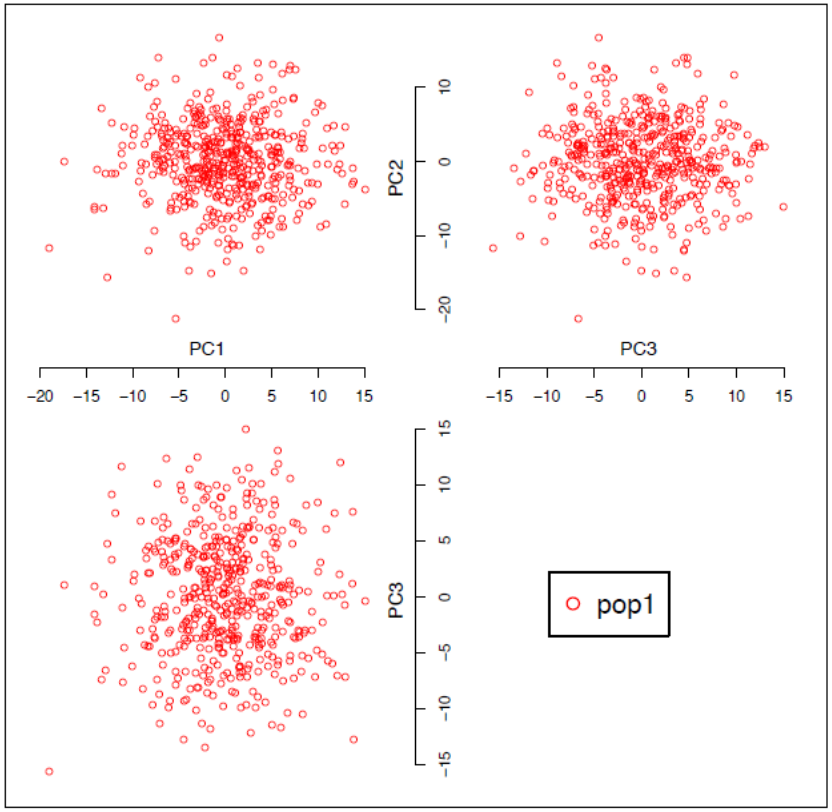
(Maus et al. 2013)

Heterogeneity as a nuisance

IPCAPS workflow

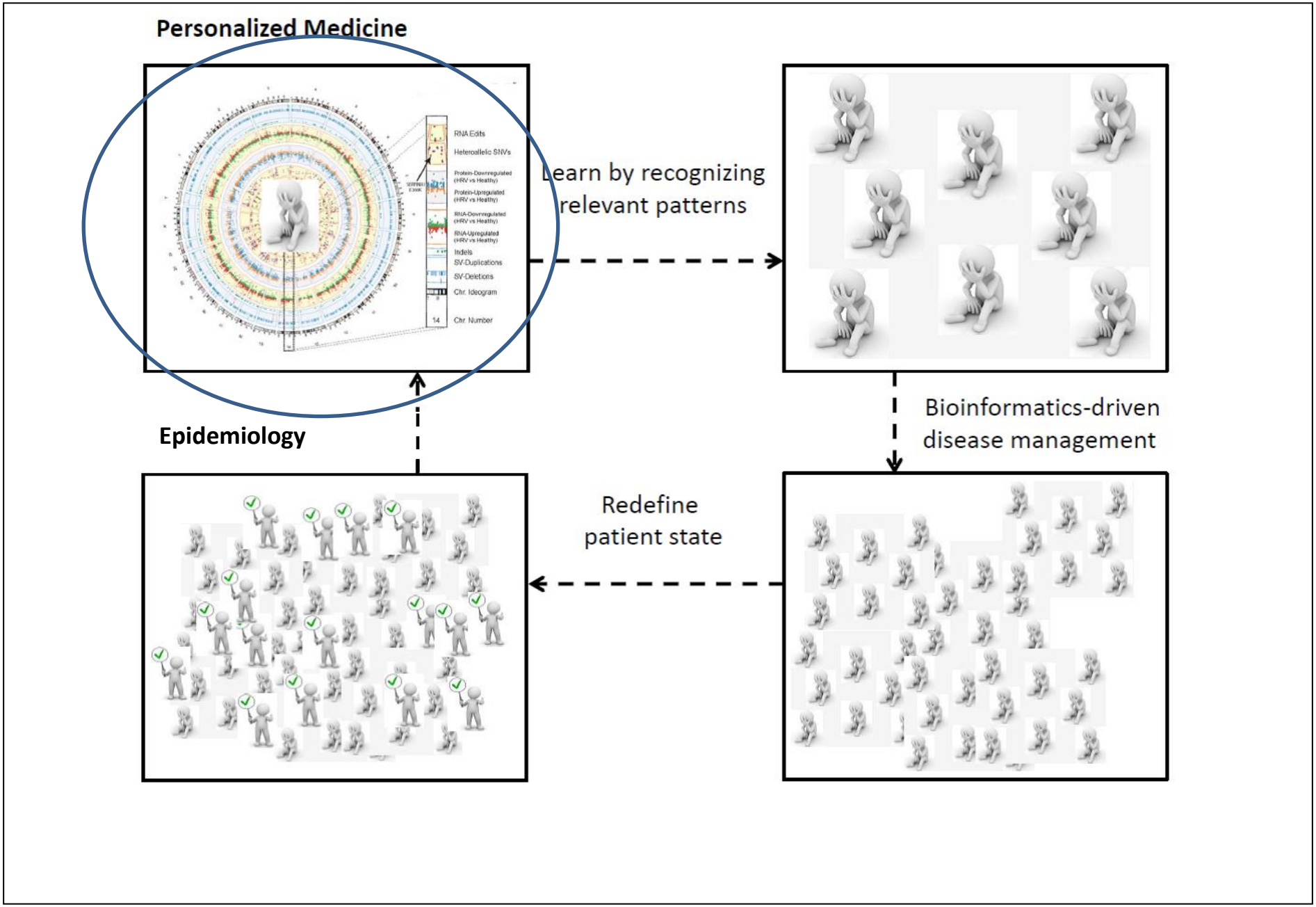


Type I error of IPCAPS



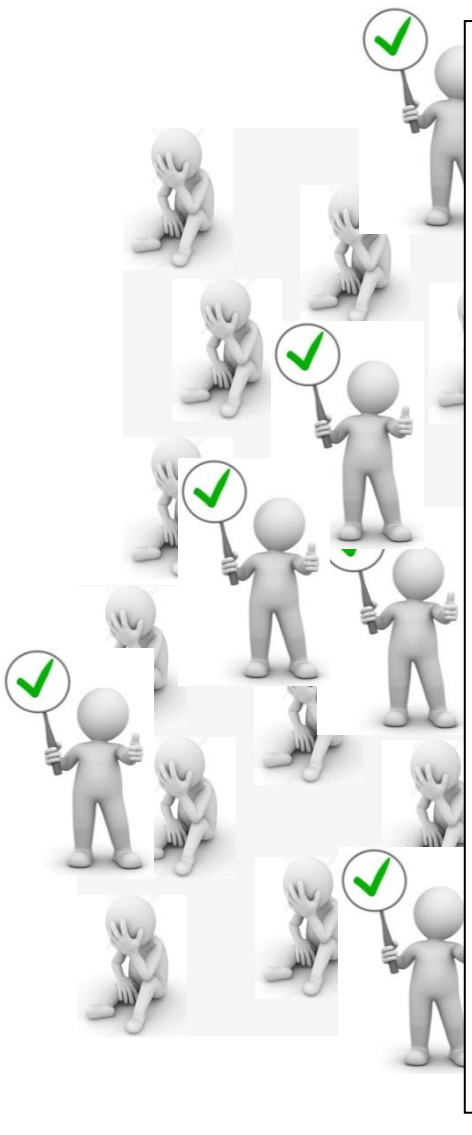
Method	Av. # clusters
IPCAPS	1
ipPCA	2
SHIPS	1
iNJclust	>150

(Kridsadakorn Chaichoompu 2017,
PhD thesis – Chapter 2)


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

Precision medicine:

analytical considerations – FINE-SCALE



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]

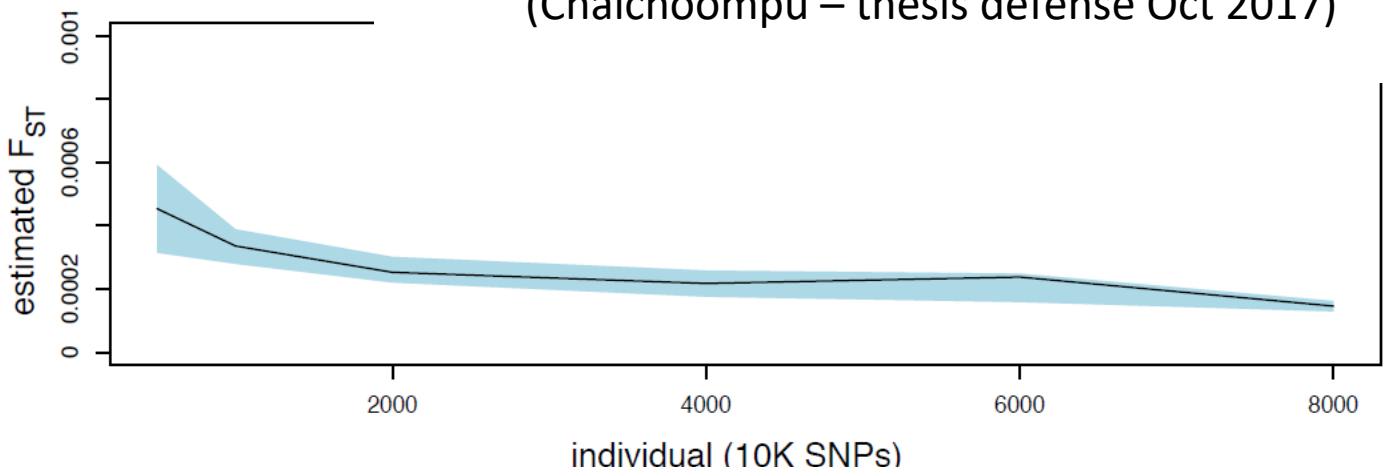
1 UMR843, INSERM, Paris, France, **2** Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, **3** UMR843, Institut National de la Sante et de la recherche Medicale, Paris, France, **4** Service de Gastroentérologie Pédiatrique, Hôpital Robert Debré, APHP, Paris, France, **5** CRC-CRB, 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

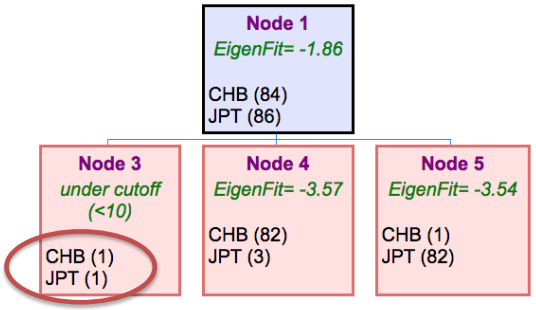
(Maus et al. 2013)

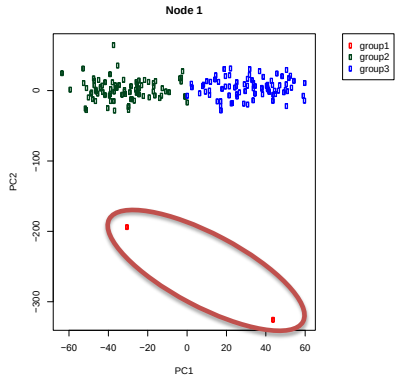
Heterogeneity as a target and a nuisance

Fine-scale structure detection

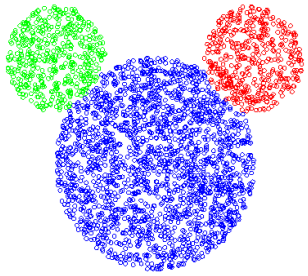
(Chaichoompu – thesis defense Oct 2017)



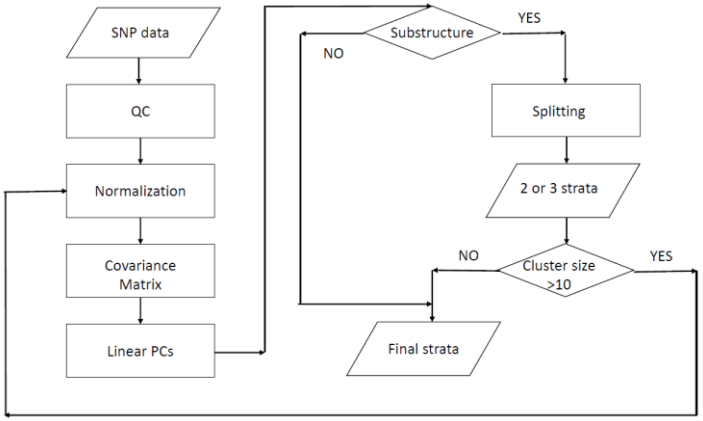




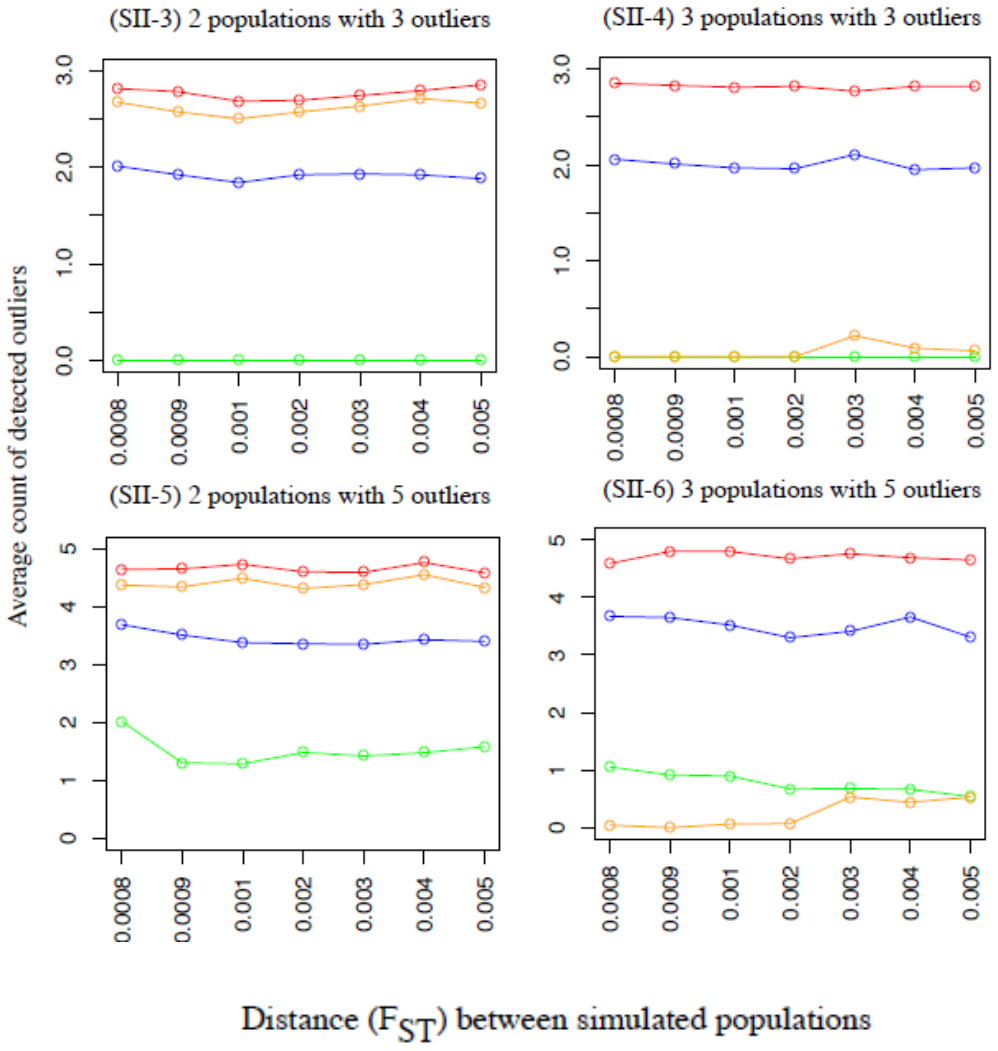
Combine with EM clustering



IPCAPS



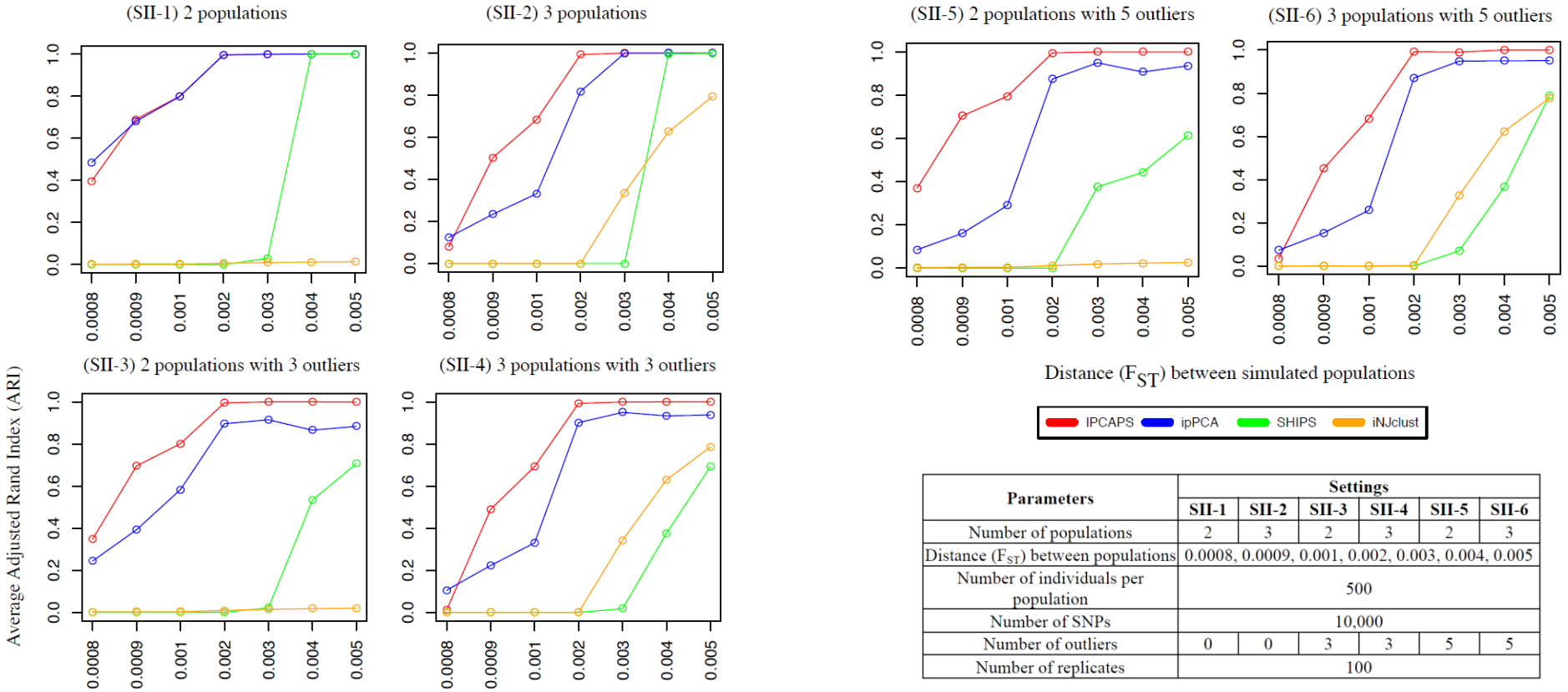
Performance of IPCAPS as outlier detection tool



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					

IPCAPS ipPCA SHIPS iNJclust

Accuracy of IPCAPS as a clustering technique



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					

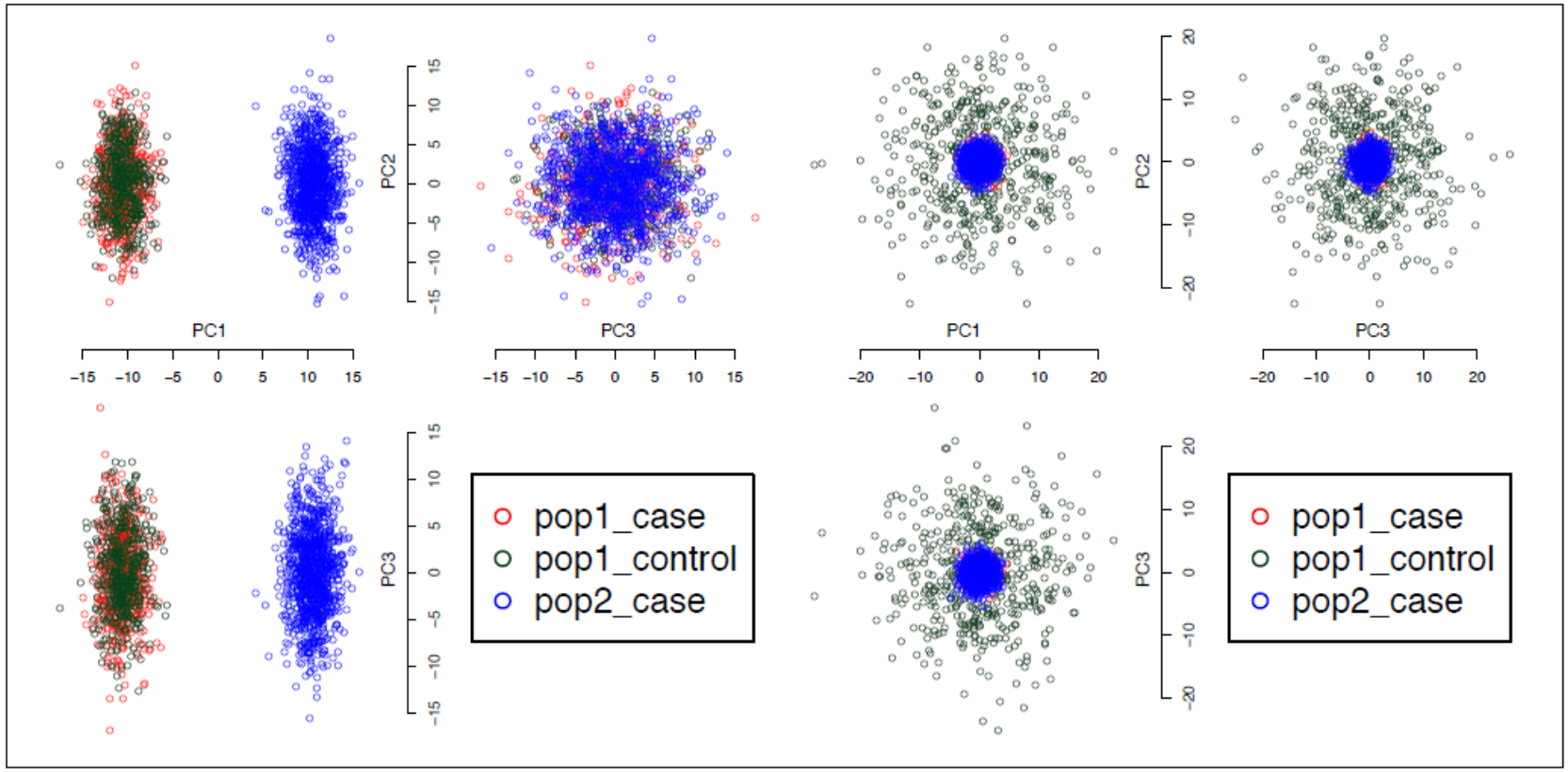
(Chaichoompu – thesis defense Oct 2017)

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

(Heath et al. 2008)

Linear population structure correction (Chaichoompu 2017+)



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

Pooled PCs but on which SNPs? (Chaichoompu 2017+)

Dataset	Uncorrected SNPs (I)		Corrected with PCs from our curated SNPs (II)				Corrected with PCs from the IIBDGC SNPs (III)				Corrected with clusters obtained by IPCAPS (IV)	
	Dis.	Rep.	5PCs		10PCs		5PCs		10PCs		Dis.	Rep.
			Dis.	Rep.	Dis.	Rep.	Dis.	Rep.	Dis.	Rep.		
CON	5	4	3	7	1	1	3	9	3	7	4	8
CD	8	4	5	8	3	8	6	3	8	3		
UC	6	7	7	7	3	3	1	5	1	5		
IBD	5	6	1	4	1	1	1	7	1	1		

Pooled PCs but on which SNPs? (Chaichoompu 2017+)

Set	Uncorrected CON		CON		CD		UC	
	Dis.	Rep.	Dis.	Rep.	Dis.	Rep.	Dis.	Rep.
1	5	4	1	1	3	8	3	3
2	3	5	1	1	3	5	3	3
3	5	5	1	1	3	3	3	5
4	5	5	1	1	3	3	3	3
5	5	5	1	1	3	5	3	3
6	5	4	1	1	3	3	3	3
7	6	5	1	1	3	3	3	3
8	6	4	1	1	6	3	3	3
9	4	4	1	1	3	8	3	5
10	4	5	1	1	6	5	3	3
Average	4.8	4.6	1.0	1.0	3.6	4.6	3.0	3.4

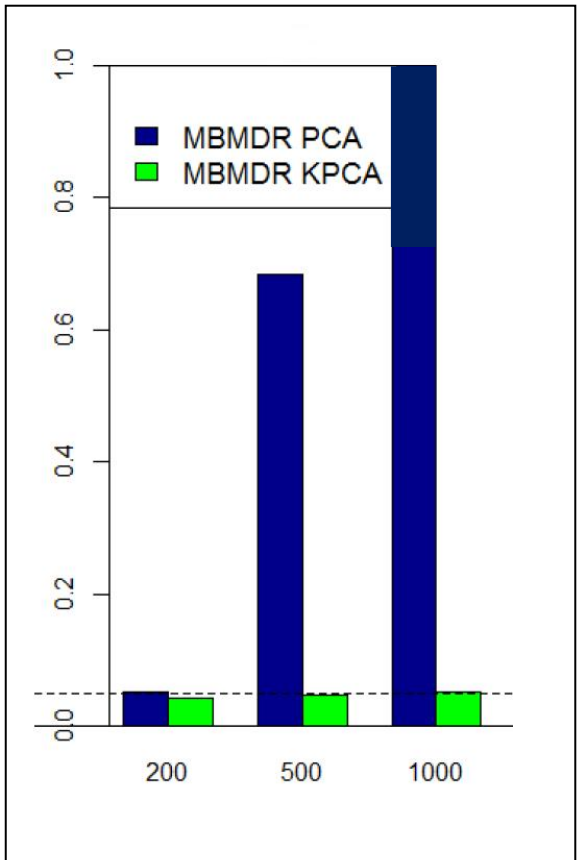
(cluster sizes less than 20 are considered to be outlying and are removed)

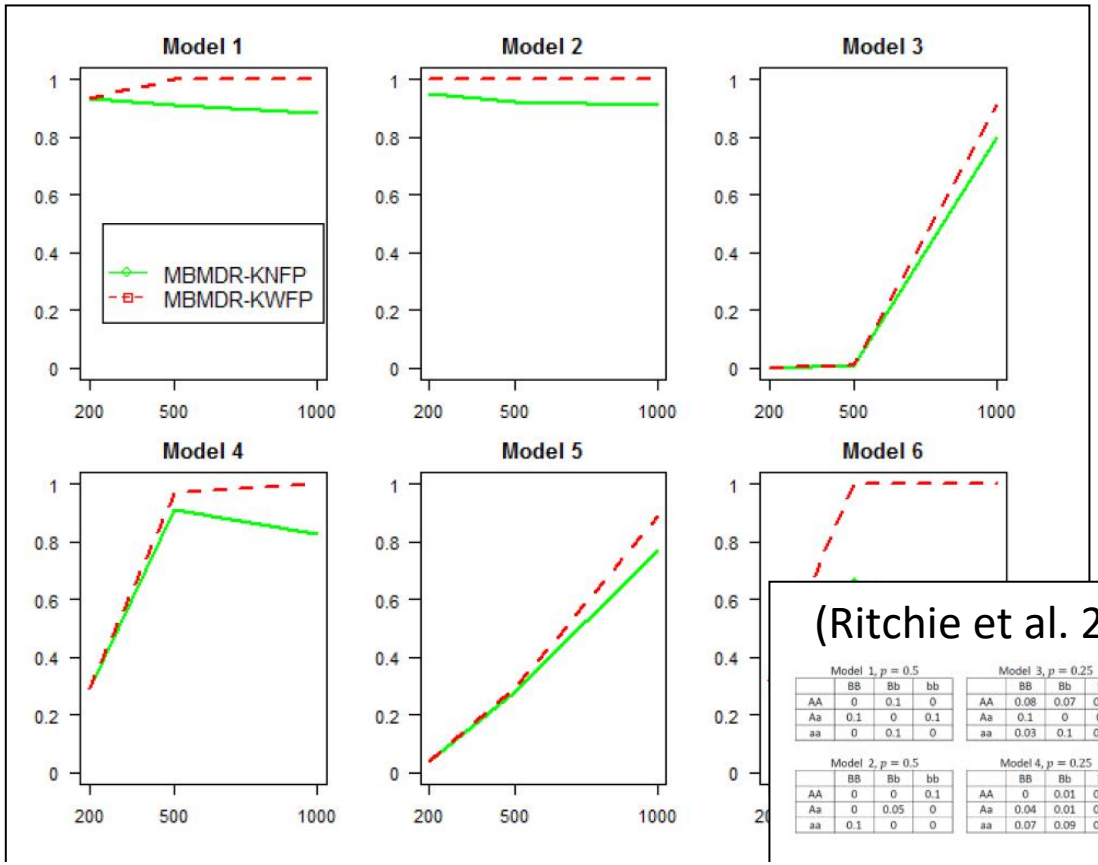
Precision medicine:

analytical considerations – NON-LINEAR

(Non-linear) confounders

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





(Ritchie et al. 2003)

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

BIOT

(Cattaert et al. 2011)

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)

Precision medicine:

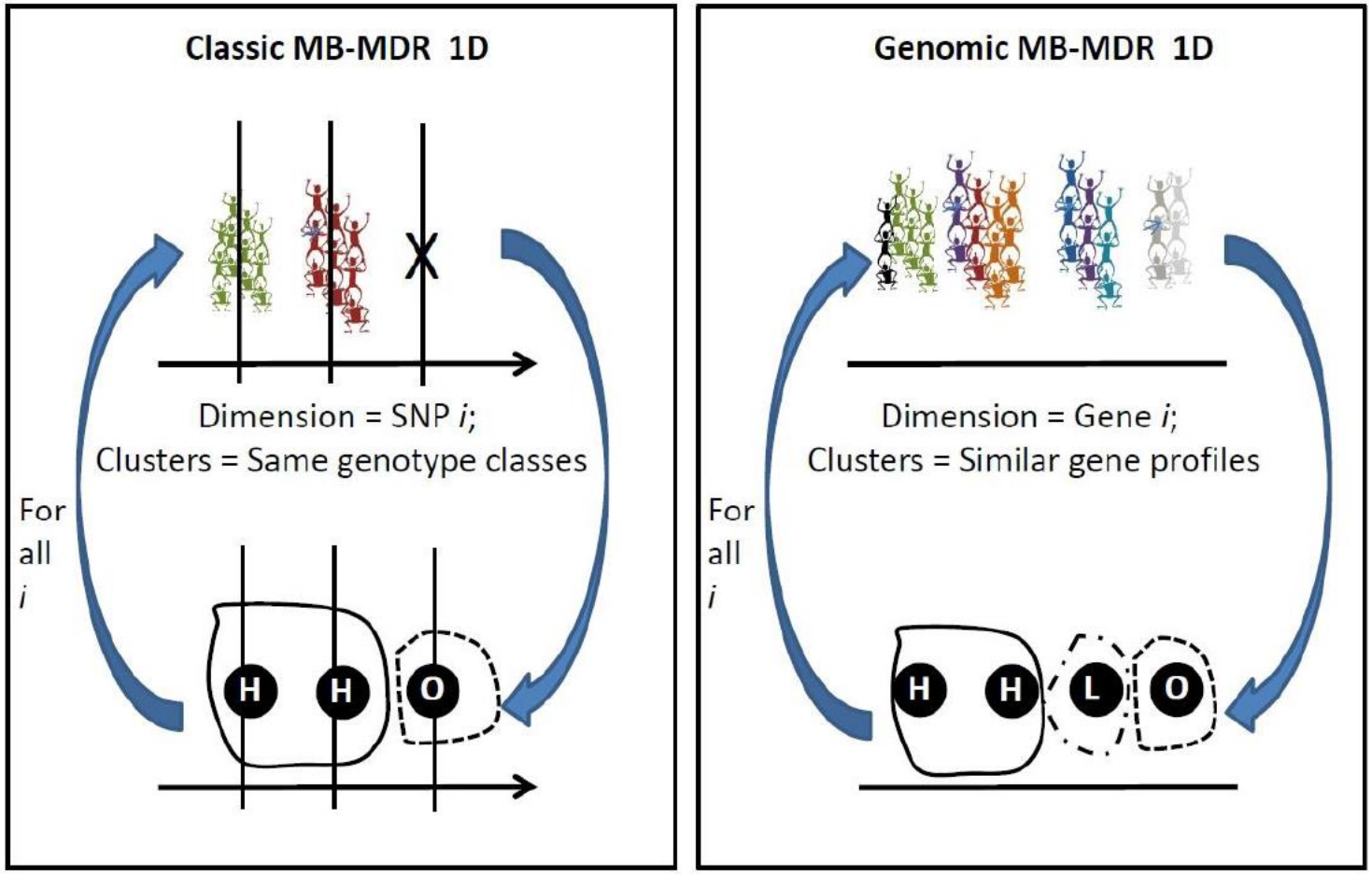
analytical considerations – REPLICATION

Replication and validation - GWAS

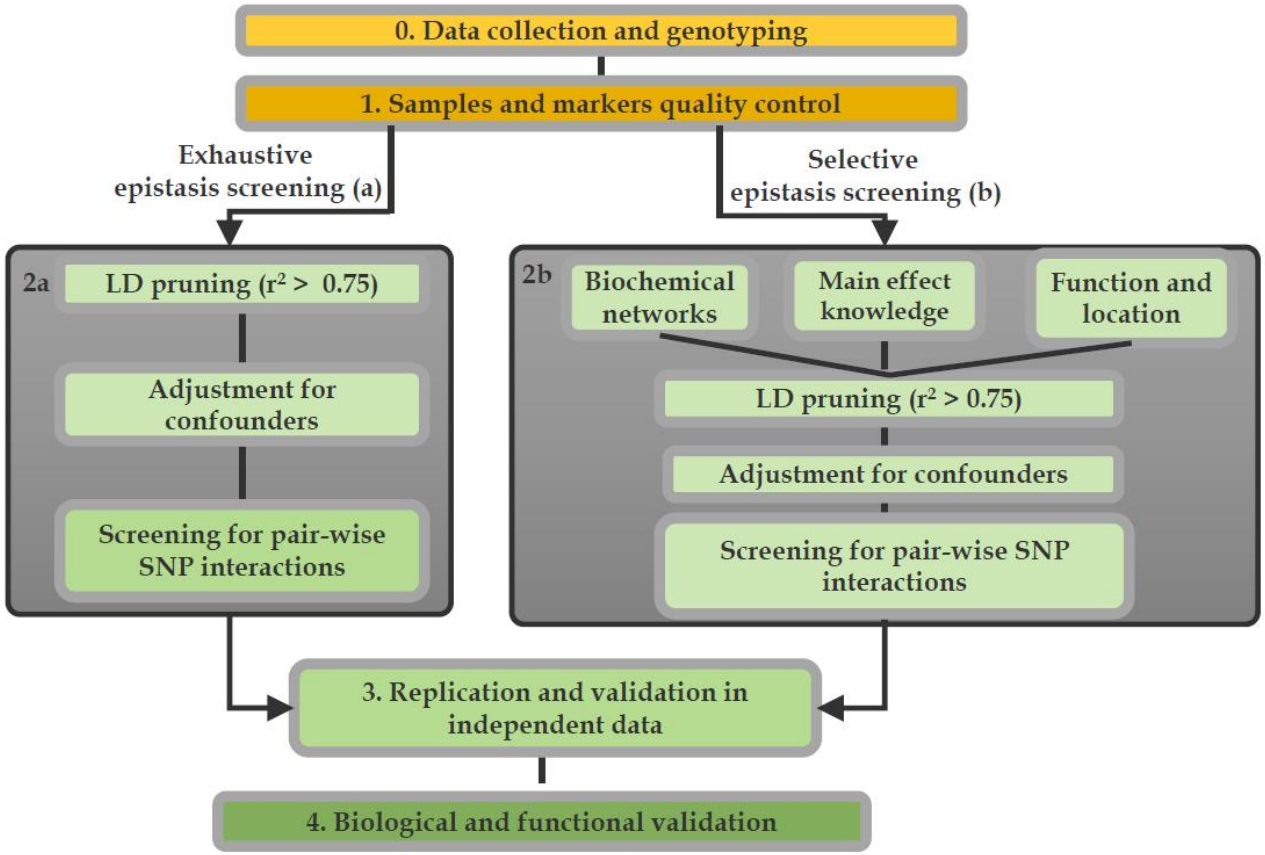
Gene representative statistics	**Related method**
$T = -2\sum_{i=1}^{m} \ln P_i$	COMBASSOC (Curtis et al., 2008)
$T = -2\sum_{i=1}^{m} \ln(1-P_i)$	Pearson's method (Pearson, 1938)
$T = \sum_{i=1}^{m} X_i$; where $X_i = Q_{\chi_1^2}(P_i)$ is the upper quintile of the χ_1^2 distribution evaluated at P_i	VEGAS (Liu et al., 2010), VEGAS2 (Mishra et al., 2015), PASCAL (Lamparter et al., 2016), fastBAT (Bakshi et al., 2016), MAGMA (Leeuw et al., 2015)
$T = \max_{i \le m} X_i$, or equivalently, $T = \min_{i \le m} P_i$	VEGAS, VEGAS2, PASCAL, MAGMA
$T = \max_{i \le m} Z_i$; where $Z_i = Q_{N(0,1)}(P_i)$ is the upper quintile of the standard normal distribution evaluated at P_i	MAGENTA
$T = -2 \times Q_1(\ln P_1, \ln P_2, \cdots, \ln P_m)$; Q_1: the first quartile	TopQ (Lehne et al., 2011)
$T(k) = \prod_{i=1}^{k} P_{(i)}$; $1 \le k \le N$ is a truncation point chosen a priori by user	Rank Truncated Product (Dudbridge et al., 2003)
$T = \prod_{i=1}^{N} P_i^{I(P_i \le \tau)}$; τ is a truncating parameter, typically set as τ=0.05	Truncated Product (Zaykin et al., 2002)

(taken from Yuanlong Liu et al., 2017)

Replication and validation – GWAIS



Interpretation - GWAIS



(Gusareva et al. 2014)

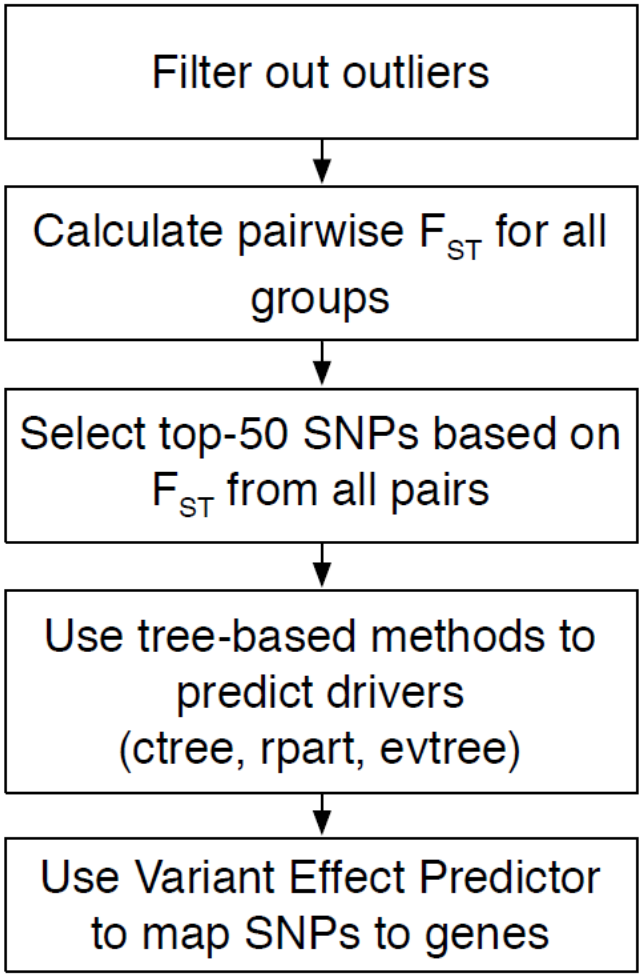
These critical steps are paramount to the *success* of GWAI studies

Interpretation - cluster discriminators

SNPs	Chr	Positions	Associated genes	Additional information
rs80261410	2	136049426	-	intergenic
rs11681014	2	134377531	MGAT5	intron
rs200930008	11	18246053	SAA2	splice region, intron
rs3749946	6	31481085	-	intergenic
rs4833103	4	38813881	-	intergenic
rs10280281	7	16365684	ISPD	intron
rs6922431	6	31497253	MICB [19]	upstream gene

...

Interpretation - cluster determinants

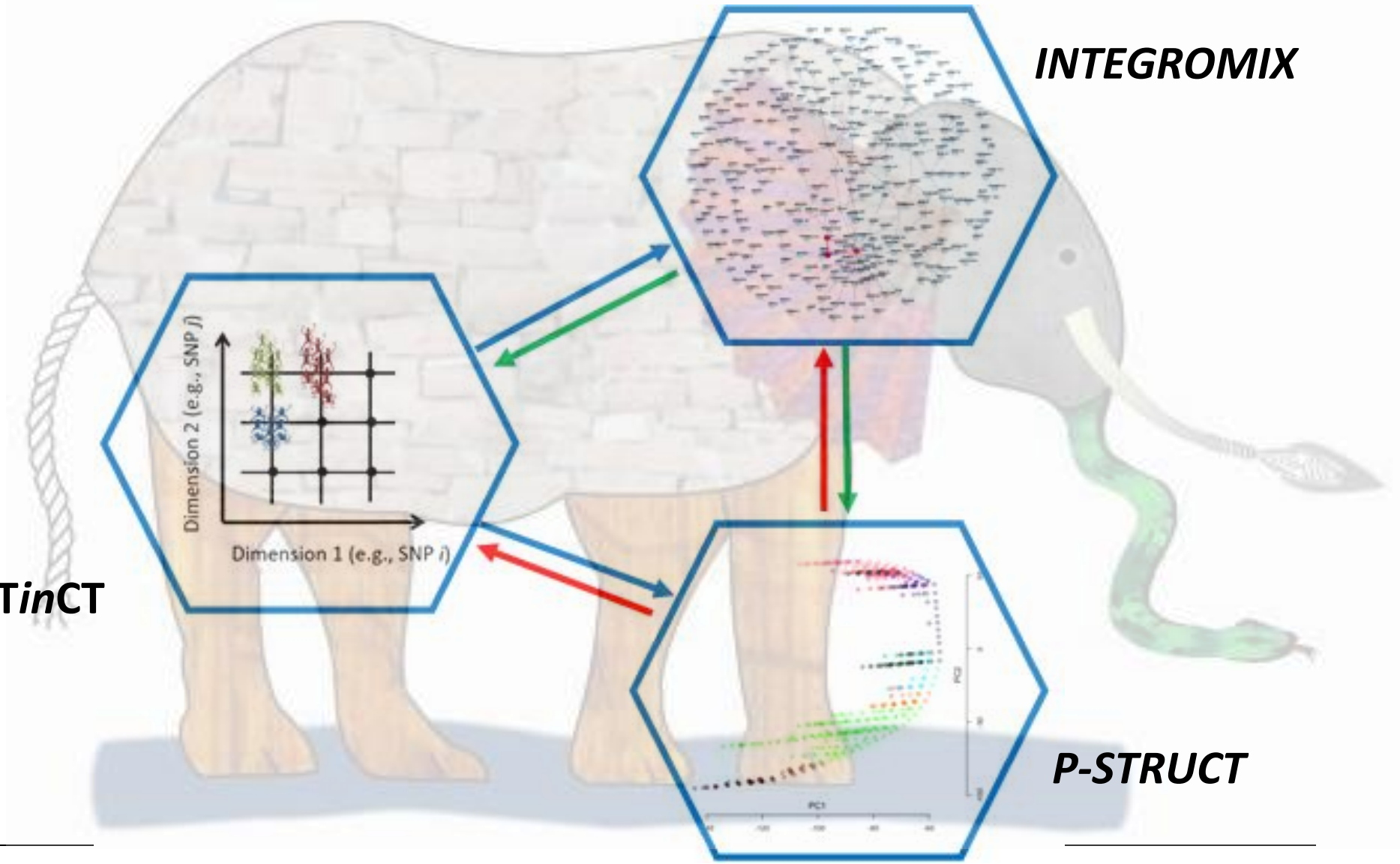


Take-home messages

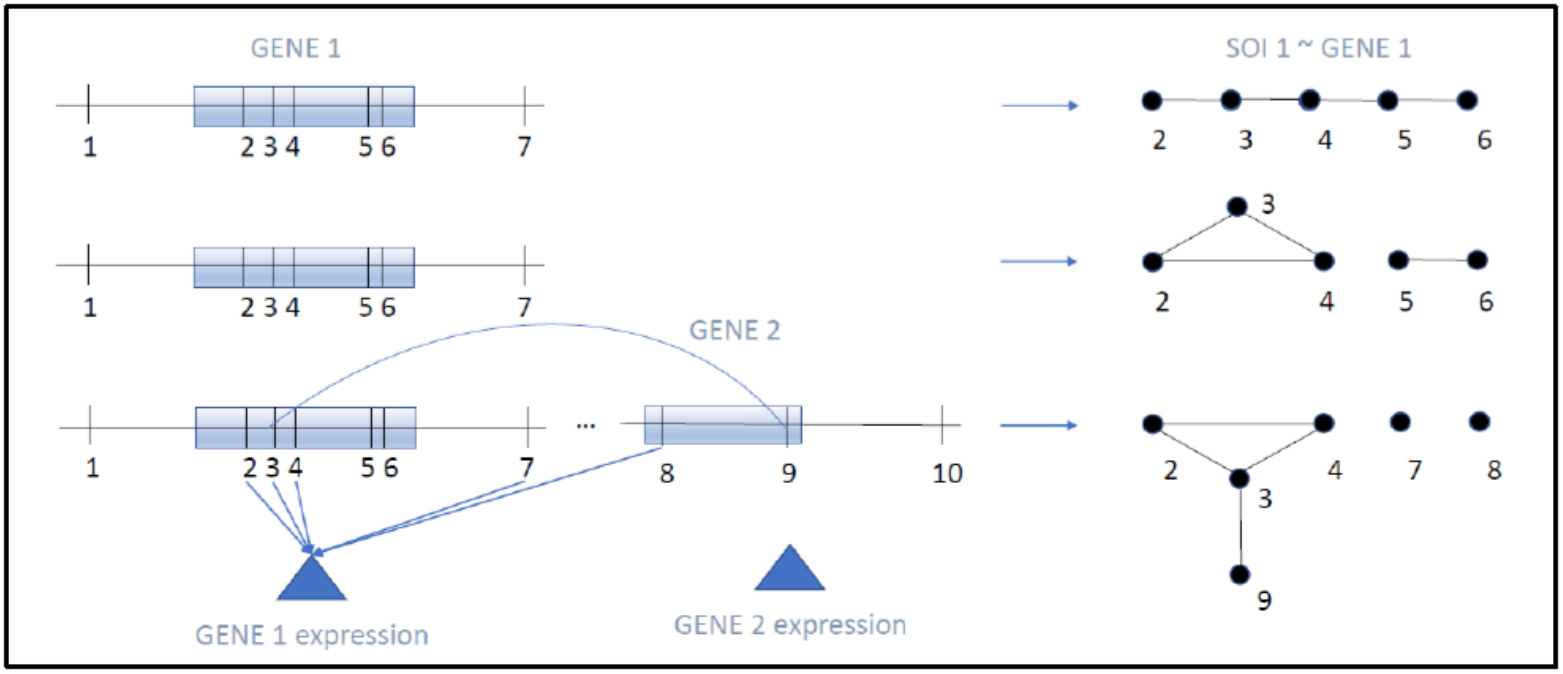
Imagine a world …

- in which missing data handling strategies hold, despite data heterogeneity
- in which multi-omics summaries can be (deep-) learned from data
- in which machine learning taxonomy addresses an interdisciplinary community
- 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

The road less traveled



Advanced integration in reduced systems



OR … have structure (deep-) learned from the data

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/