Systems Health 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

OUTLINE

- Systems health
- Precision medicine: practical implementation
- Precision medicine: analytical considerations

  IPCAPS / gene-centric approaches
- Take-home message

Systems health

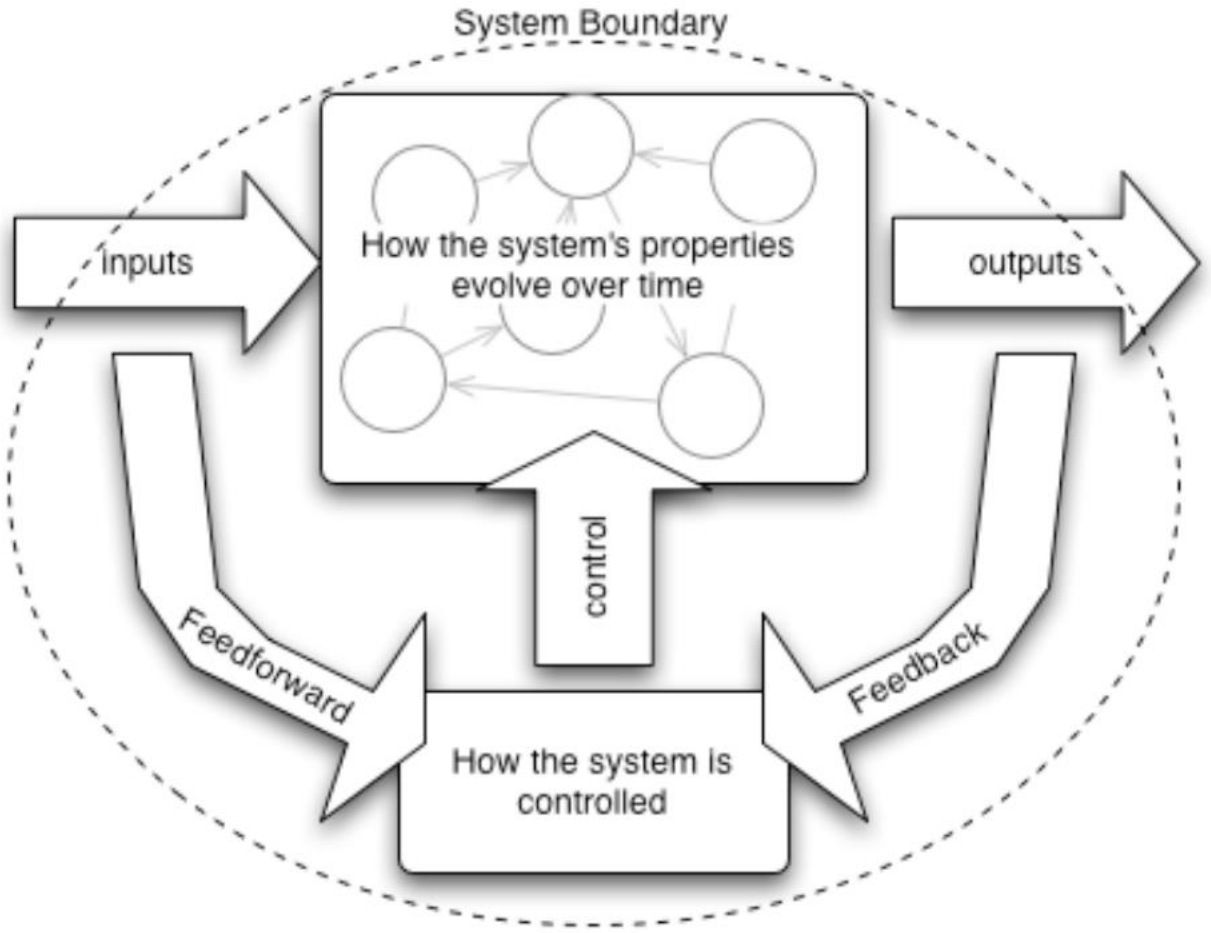
Systems

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.

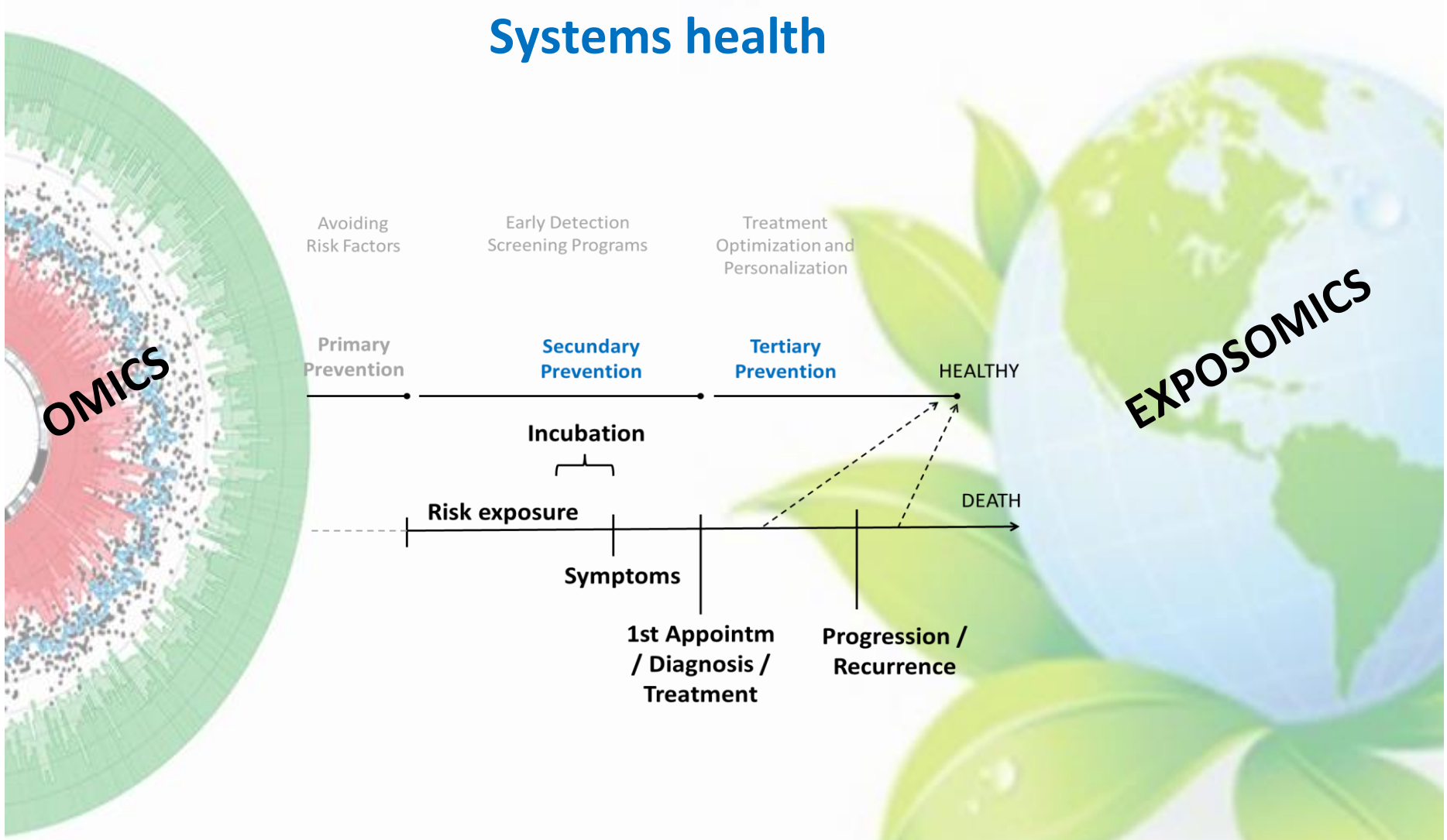
(Ackoff, 1970)

A System's Eco-system



(@2004-5 Steve Easterbrook)

Systems health



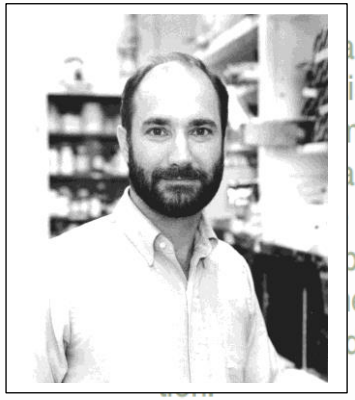
Individual-level characterizations

Cell PRESS

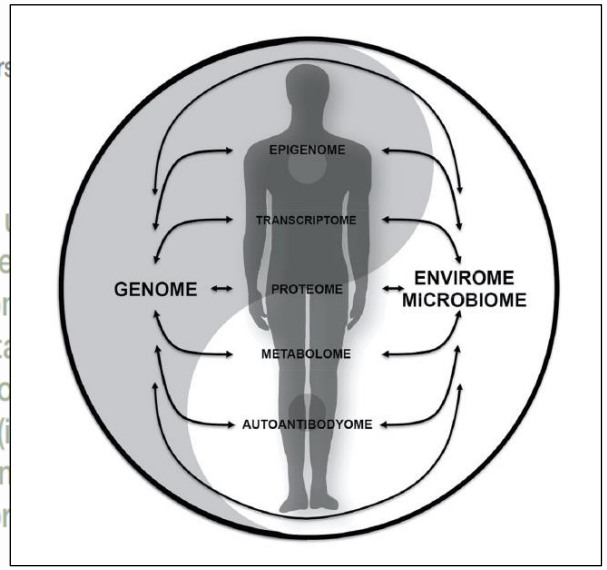
Chemistry & Biology
Review

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



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."

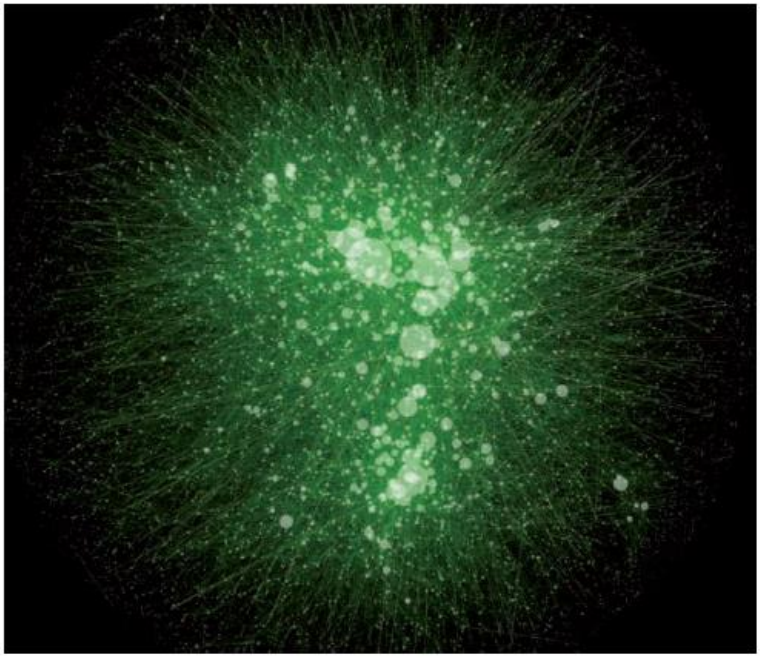
(HORIZON2020 Advisory Group)

Precision medicine: disease heterogeneity

	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]

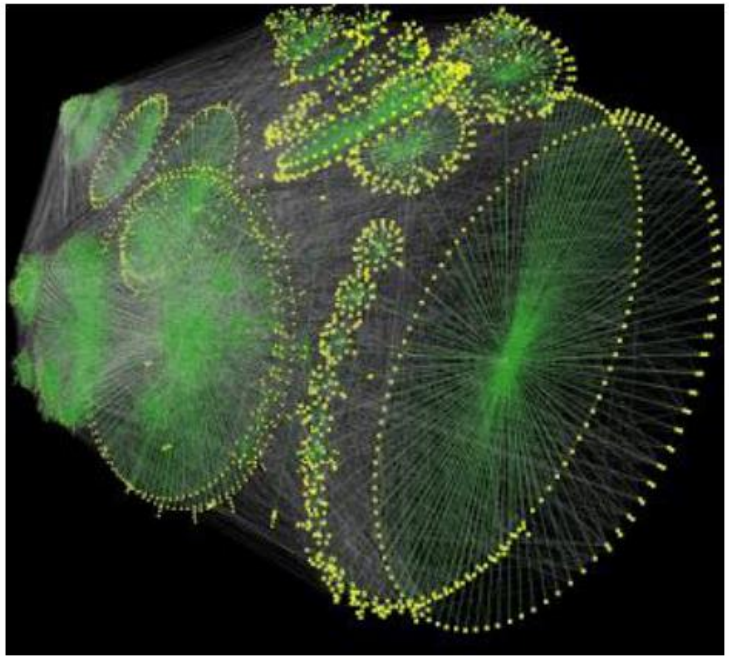
(Biancheri et al. 2013)

Systems health: interactions; Interactions and differentiation



Human interactome (PPI)

(Bonetta 2010)



Fruit fly interactome

(www.molgen.mpg.de)

Reminder: “the” interactome

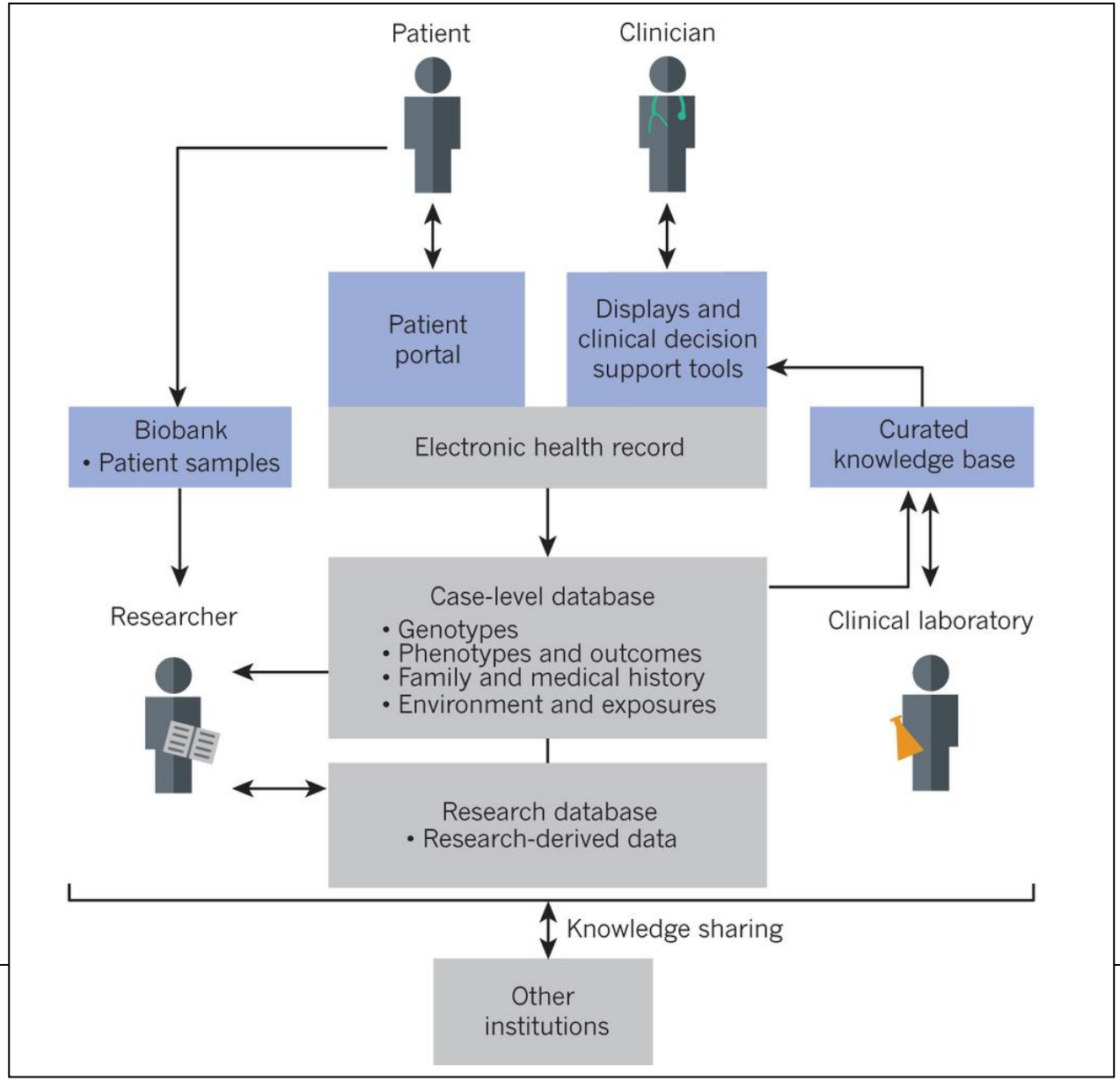
The **interactome** refers to the entire complement of interactions between DNA, RNA, proteins and metabolites within a cell. These interactions are influenced by genetic alterations and environmental stimuli. As a consequence, the interactome should be examined or considered in ***particular contexts***.

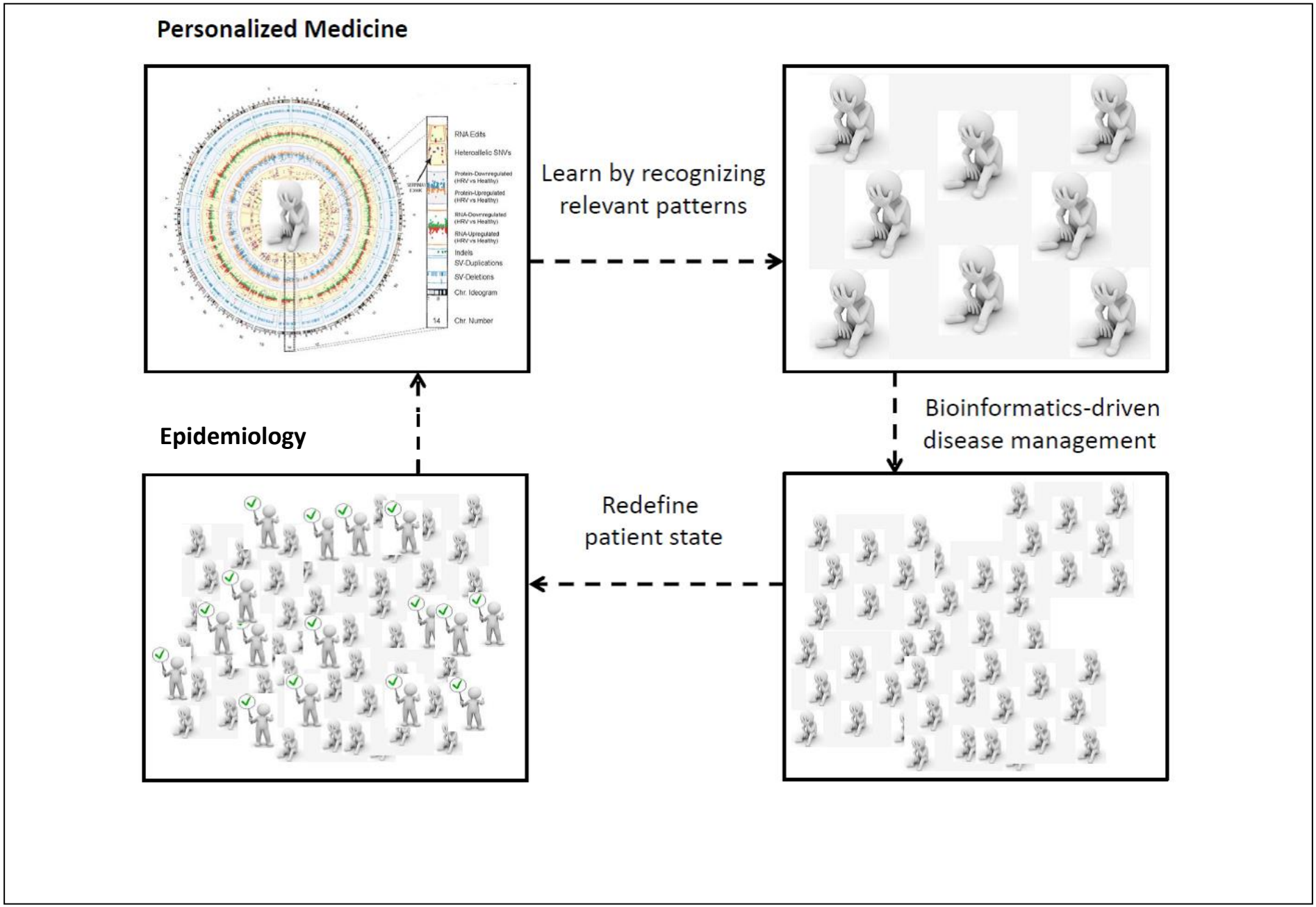
Precision medicine:

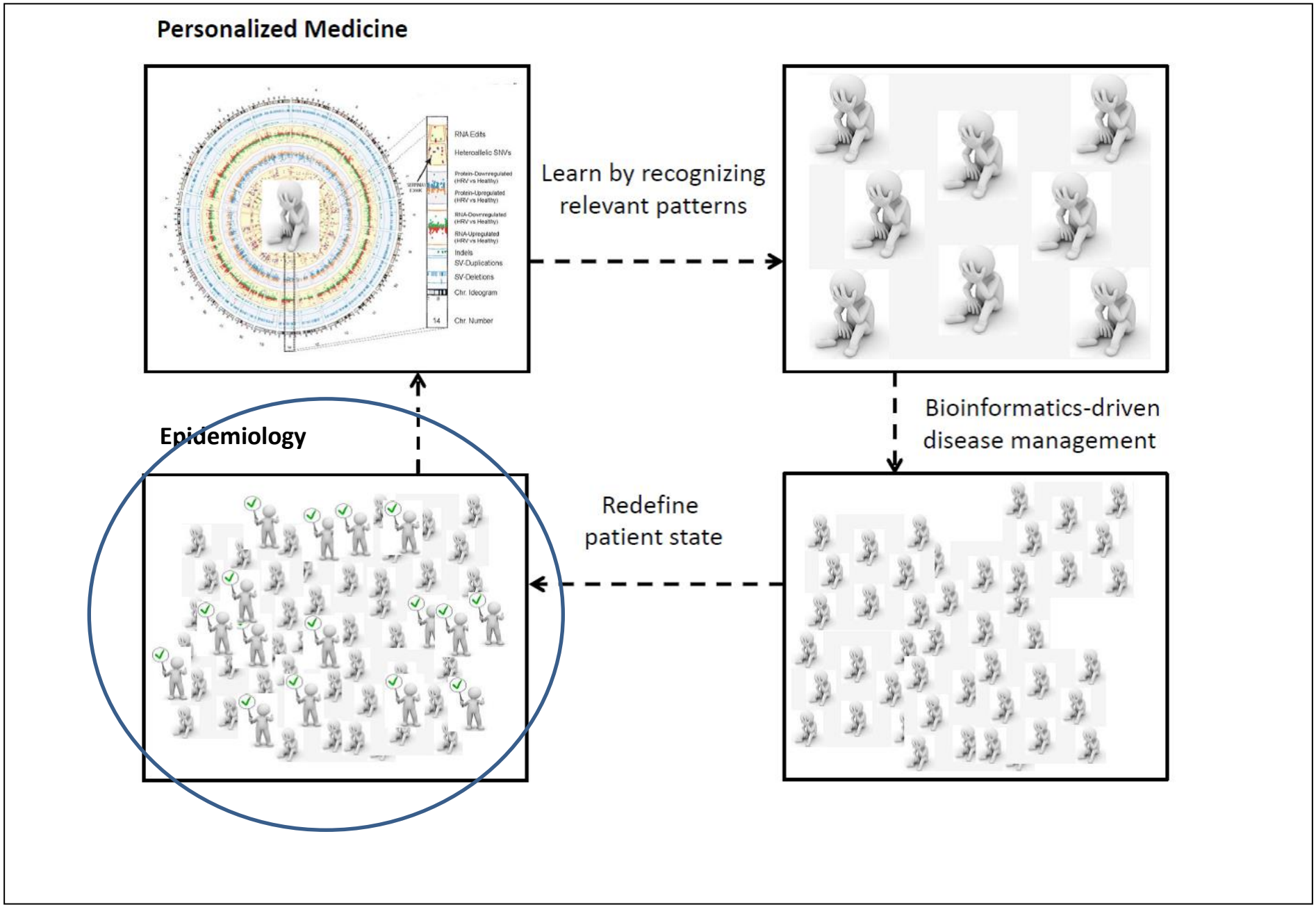
practical implementation

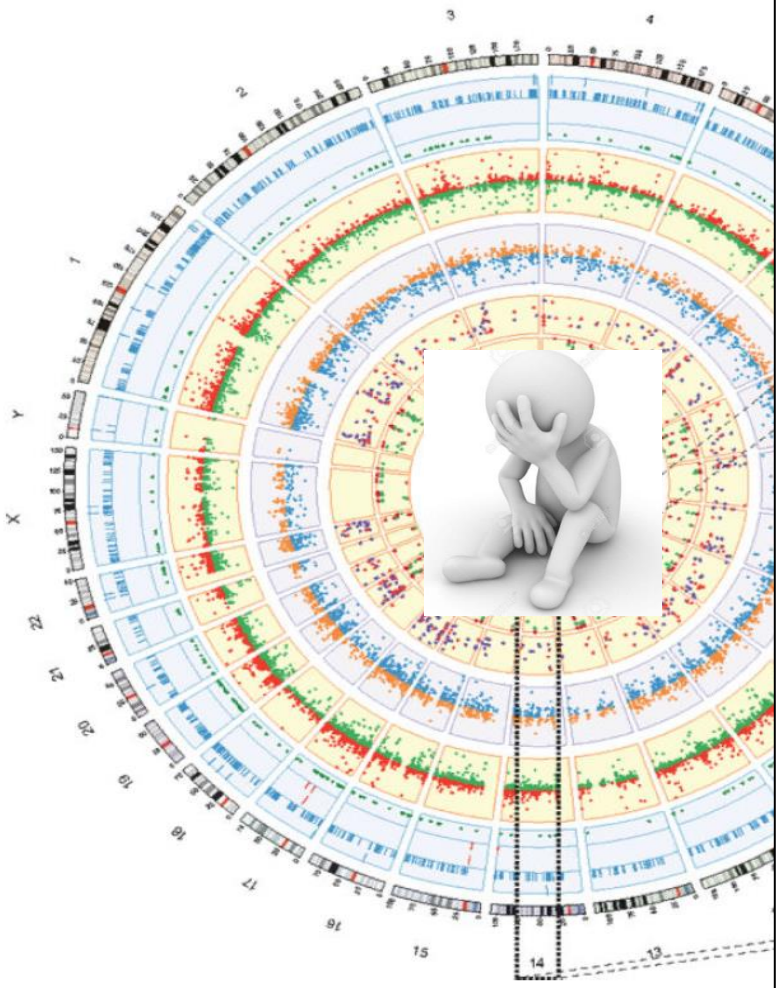
A Patient's Eco-System

(Aronson and Rehm 2015)




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


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

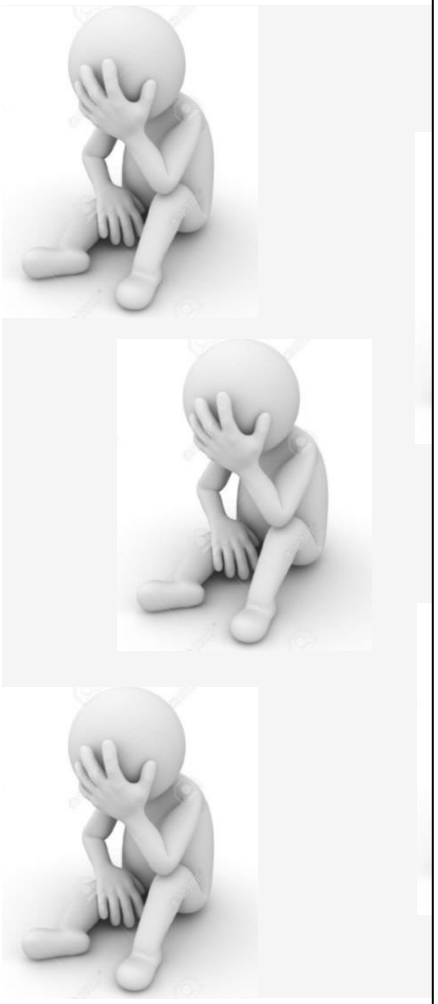


Do you think that omics profiling will be routinely used in the clinic in future?

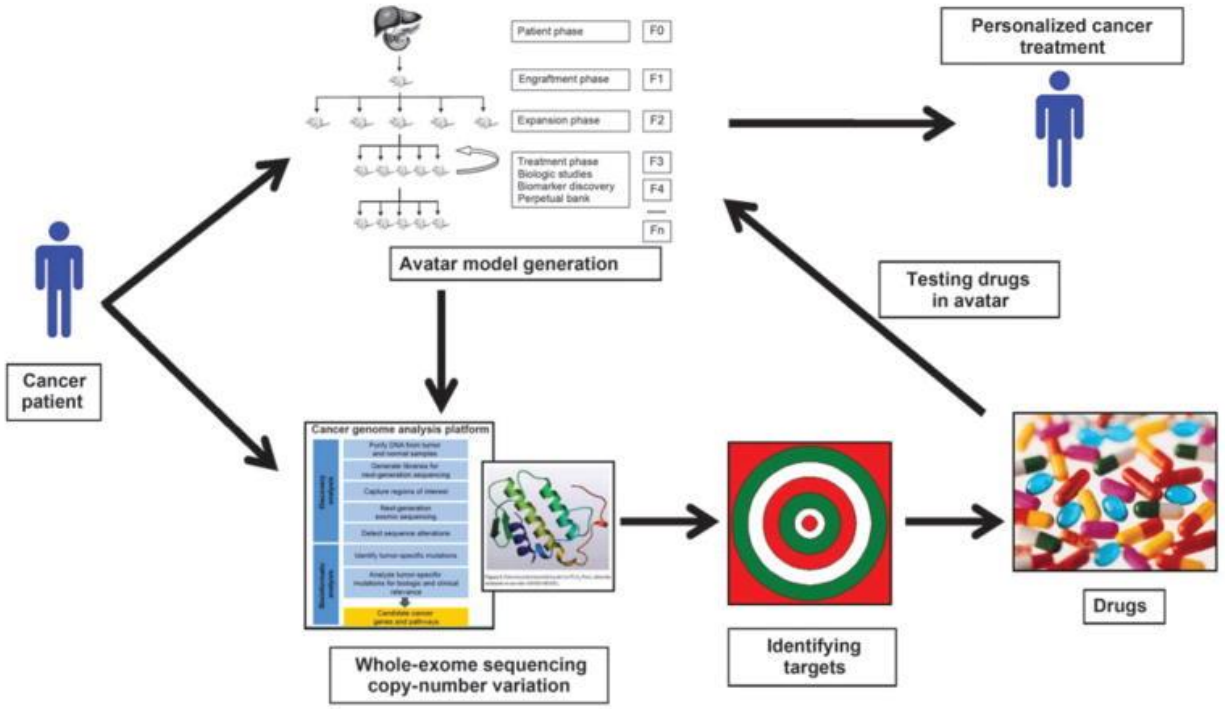
"Not in the form we are doing it. At the moment we have a very incomplete picture of what's going on, whereas if we were able to make thousands of measurements we would have a much better feeling. We just don't know, for the clinical tests, which thousand measurements are going to be most useful. We'll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

(Snyder 2014)

Redundancy - Informativity

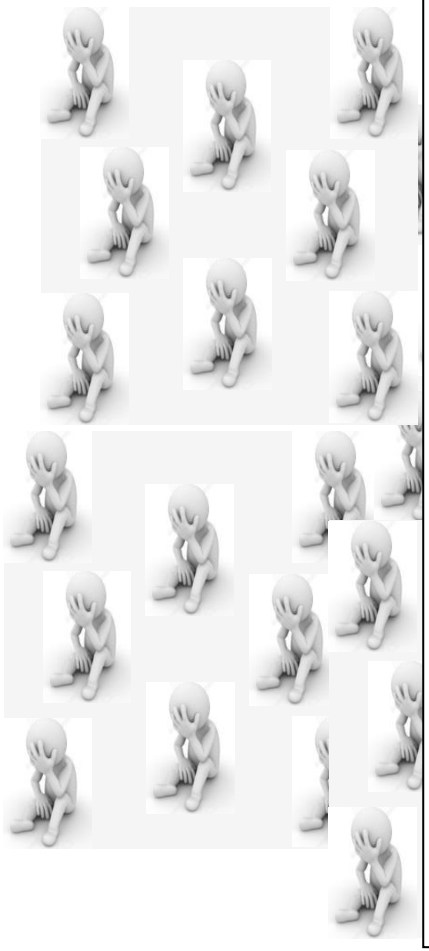


Integrating sequencing and avatar mouse models

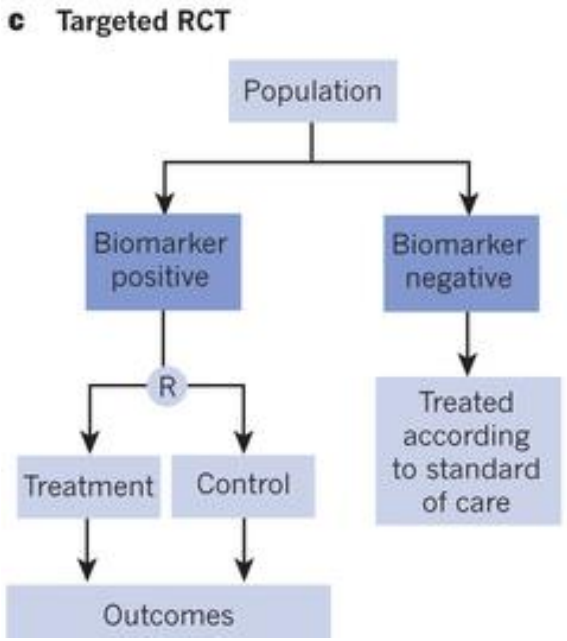


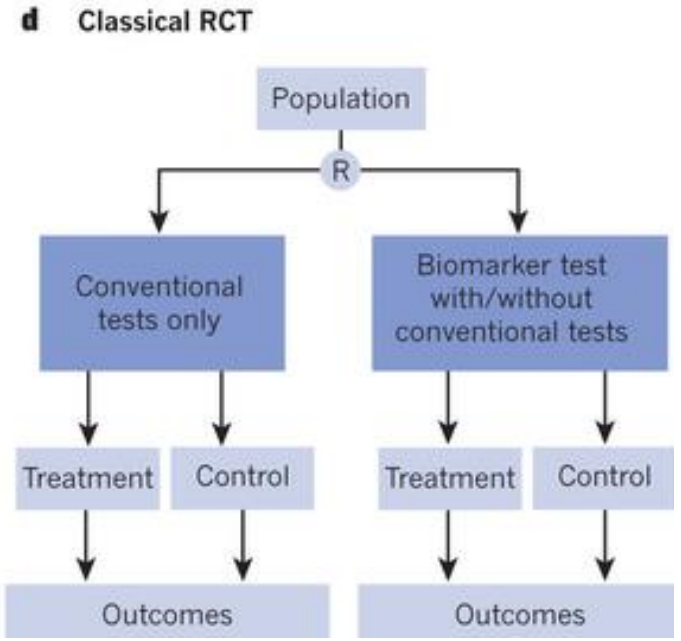
Missingness

(Garralda et al. 2014)



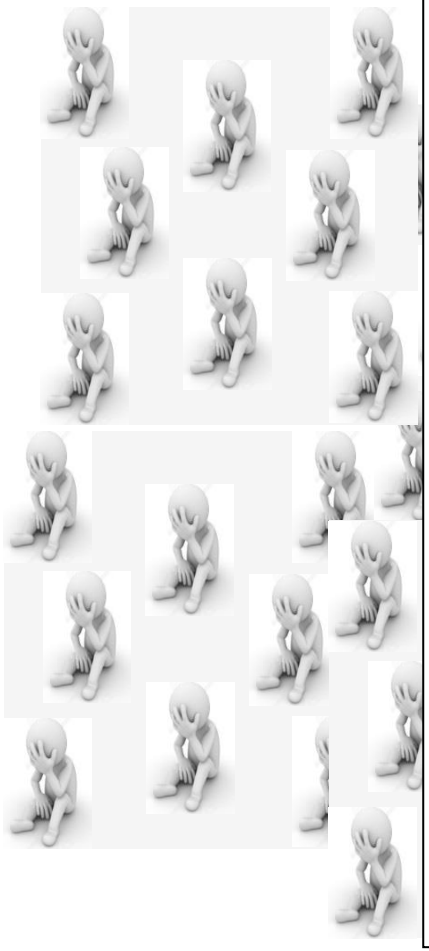
Testing precision-medicine strategies





Replication and validation

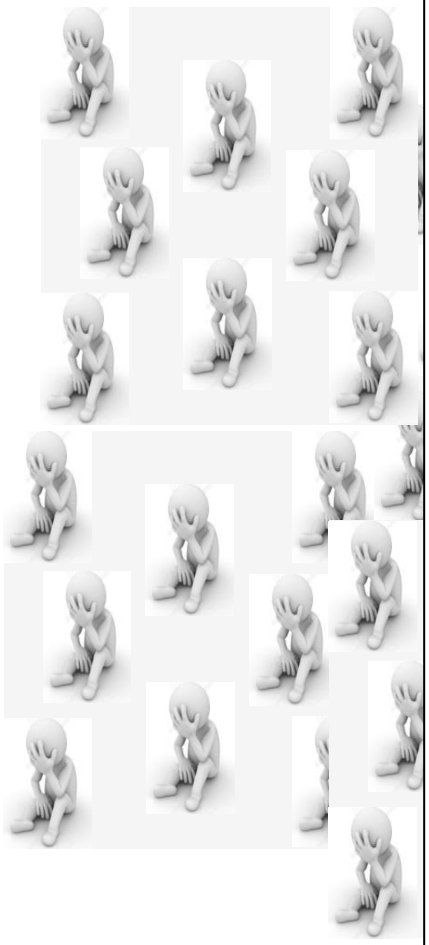
(Biankin et al. 2015)



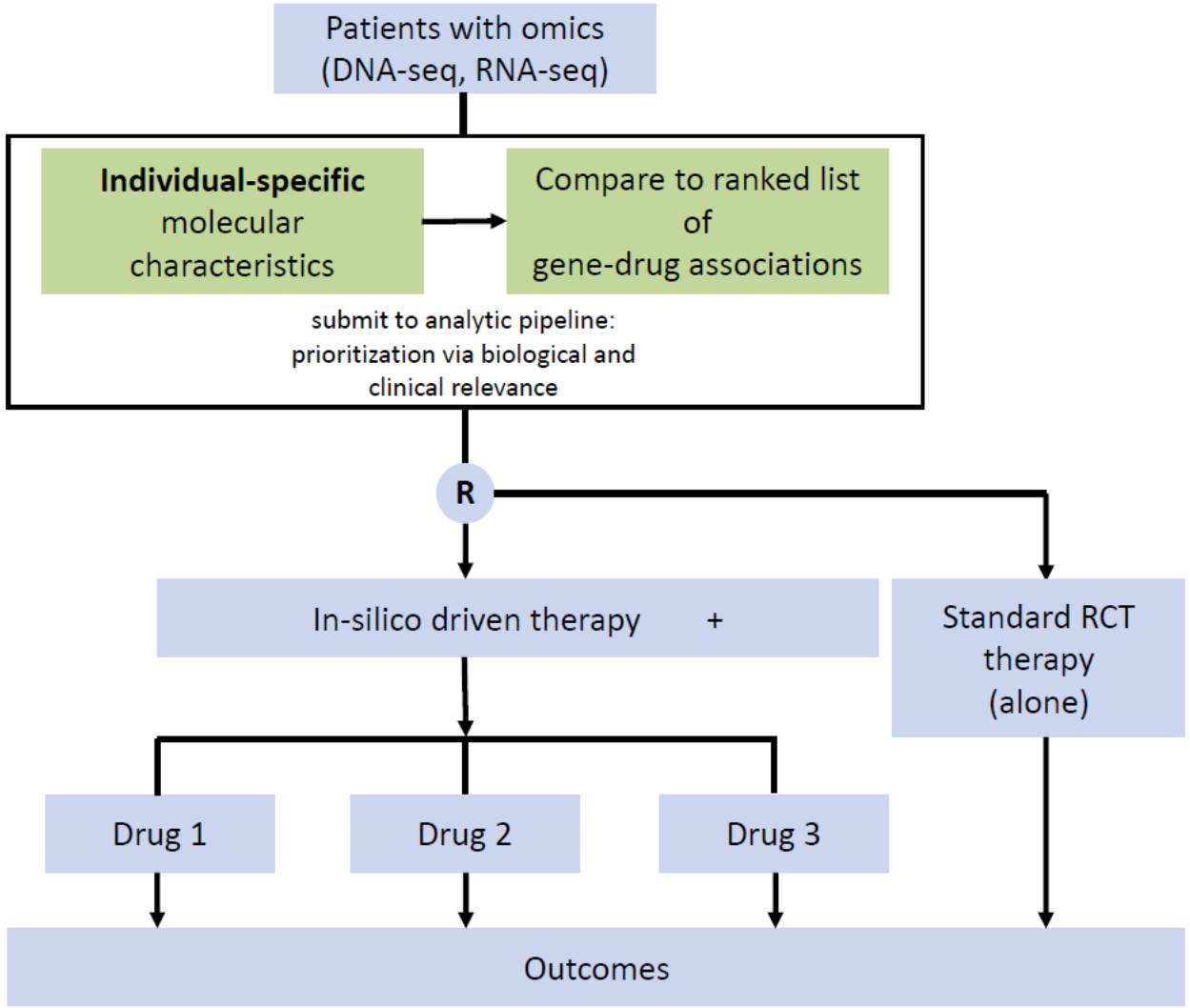
Testing precision-medicine strategies

- Umbrella CTs: 1 disease, different genetic mutations which define subcohorts, each receiving randomized treatment regimen
- Basket CTs: multiple diseases with the same genetic mutation, randomized treatment allocation (multi-dimensional mutational profile: assign treatment based on the mutation detected in the higher pct of cancer cells …)

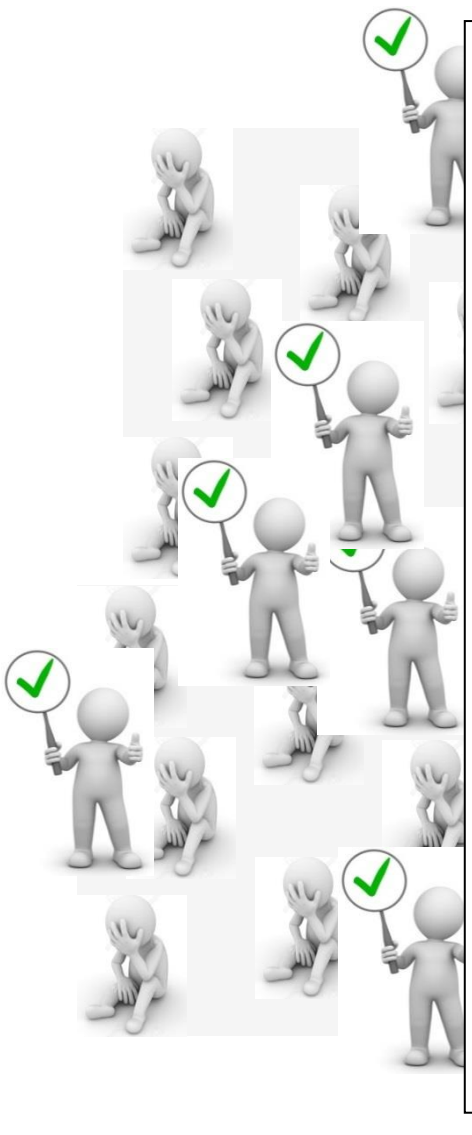
(Sumitrhra Mandrekar,
INSERM atelier 248, Bordeaux, 2017)



Testing precision-medicine strategies



BIO



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

OPEN ACCESS Freely available online

PLOS one

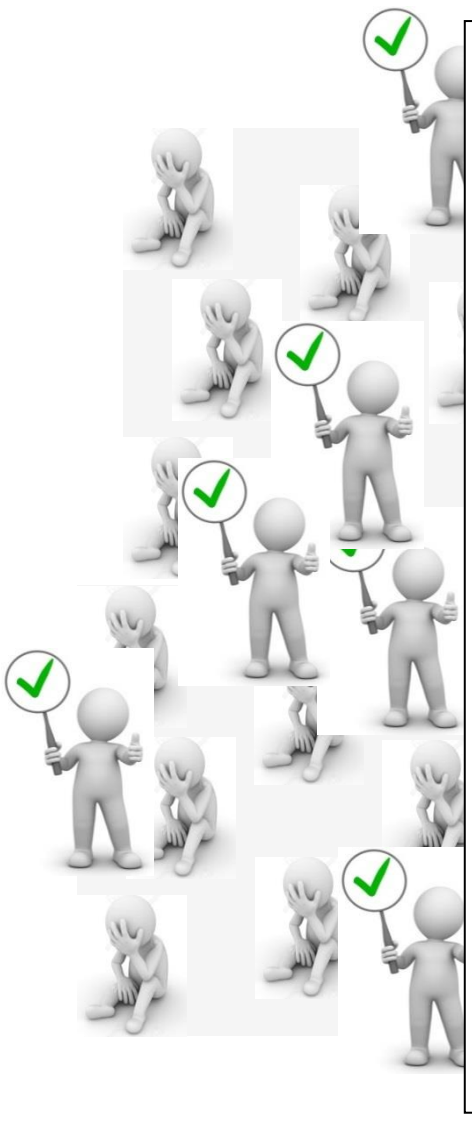
Molecular Reclassification of Crohn's Disease by Cluster Analysis of Genetic Variants

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



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

OPEN ACCESS Freely available online

PLOS ONE

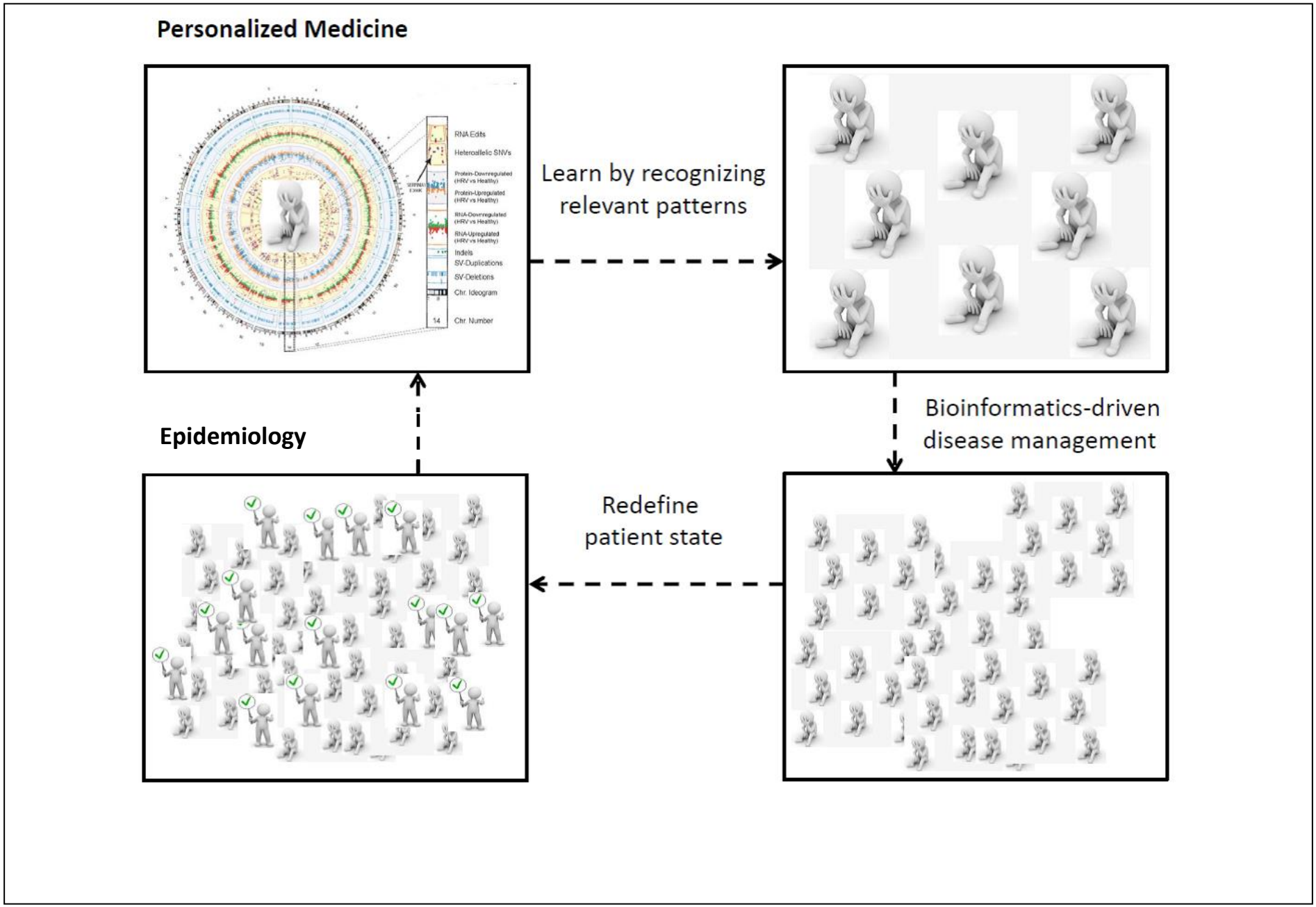
Molecular Reclassification of Crohn’s Disease: A Cautionary Note on Population Stratification

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

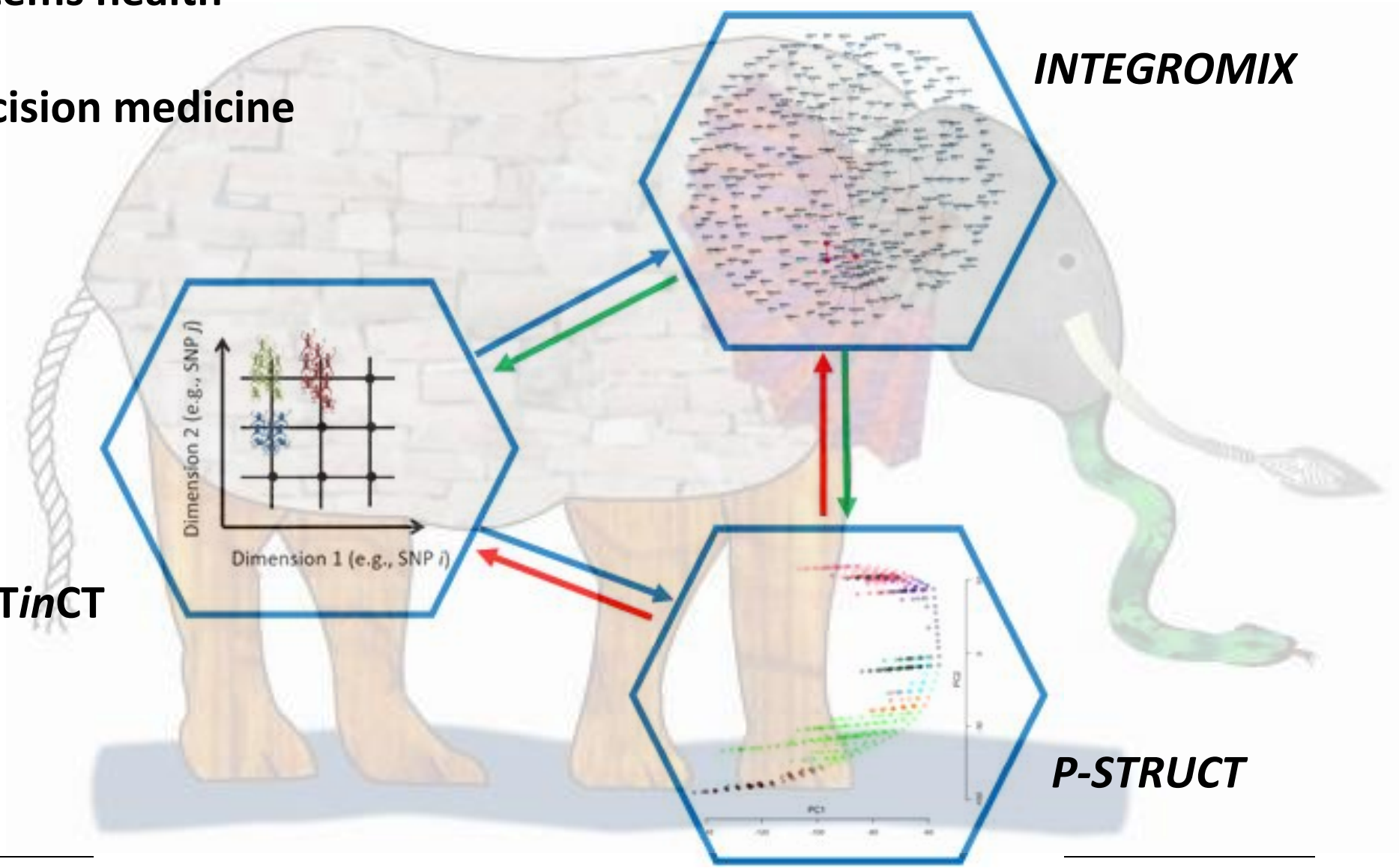
(Maus et al. 2013)

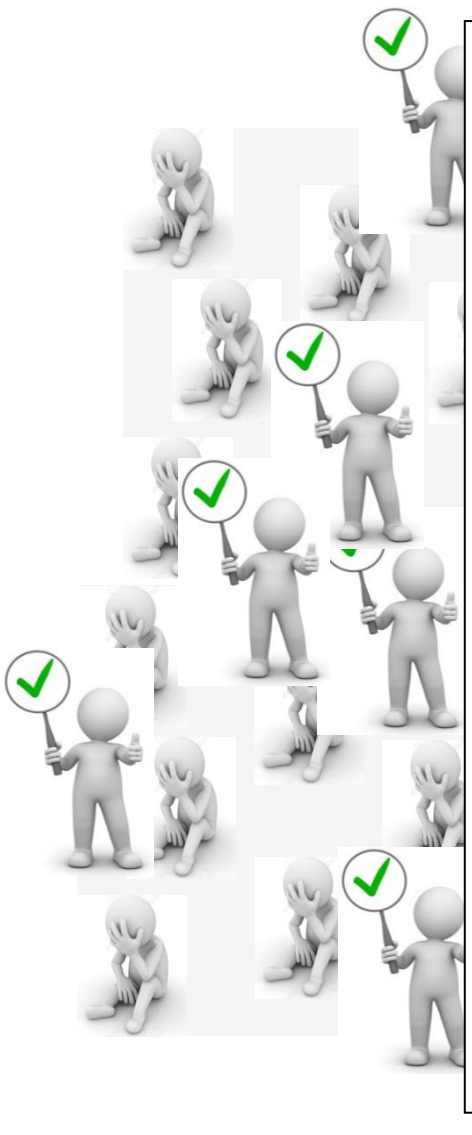
Heterogeneity as a target and a nuisance


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

Systems health for precision medicine





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]

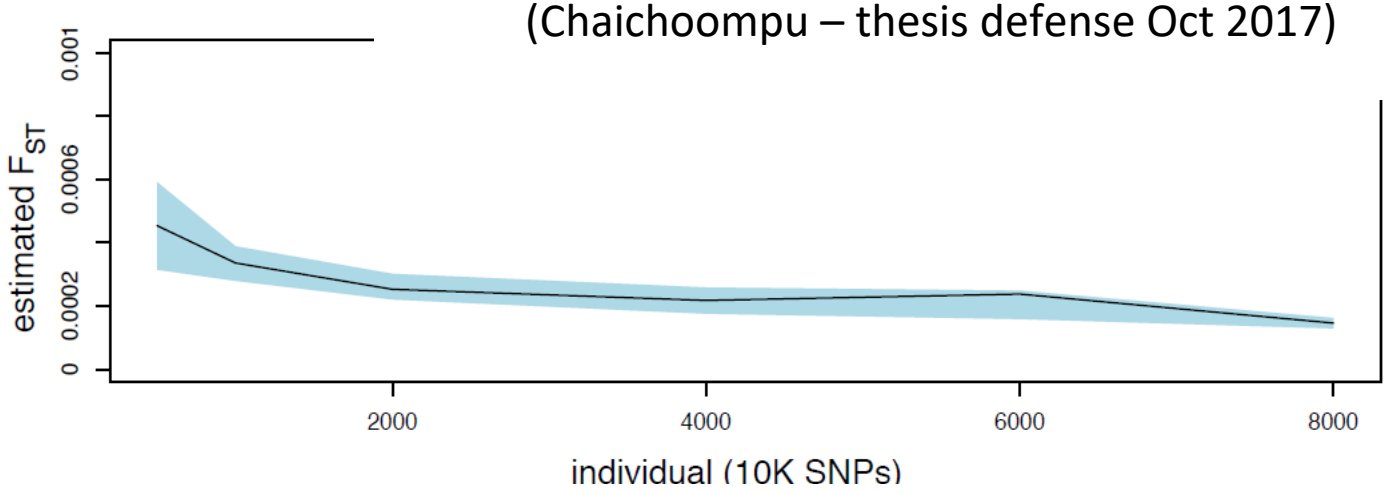
1 UMR843, INSERM, Paris, France, **2** Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, **3** UMR843, Institut National de la Sante et de la recherche Medicale, Paris, France, **4** Service de Gastroentérologie Pédiatrique, Hôpital Robert Debré, APHP, Paris, France, **5** CRC-CRB,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

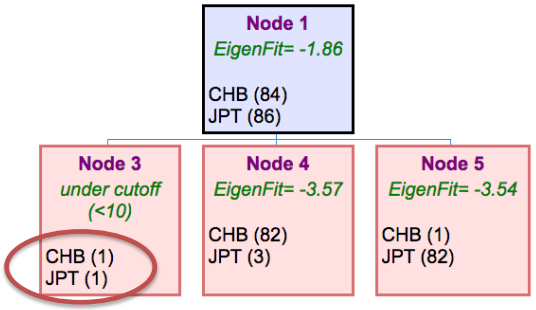
(Maus et al. 2013)

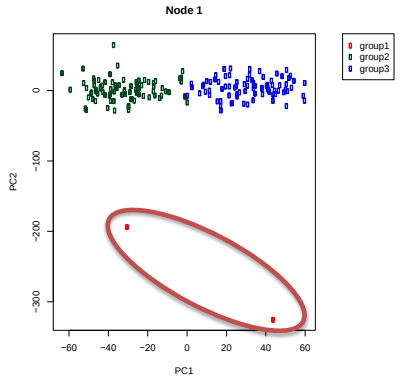
Heterogeneity as a target and a nuisance

Fine-scale structure detection

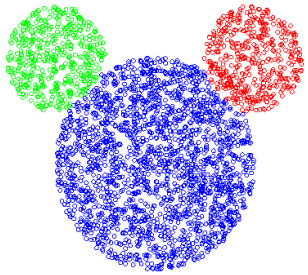
(Chaichoompu – thesis defense Oct 2017)



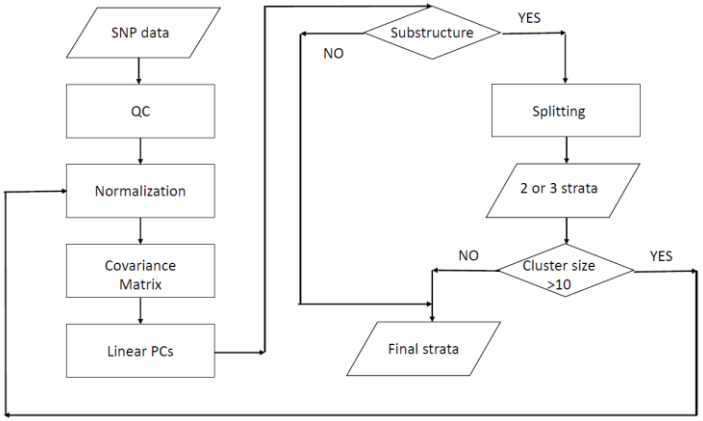




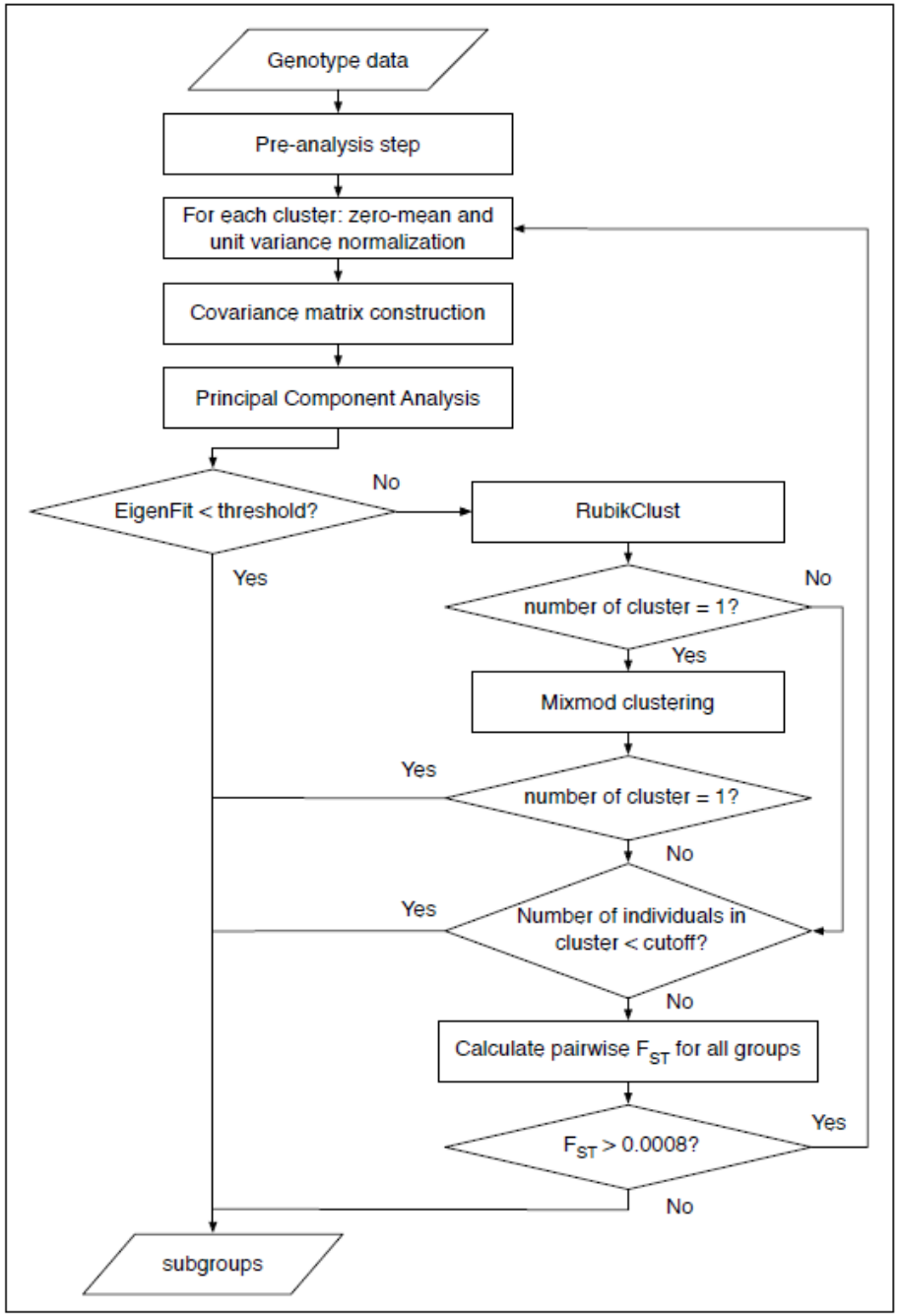
Combine with EM clustering



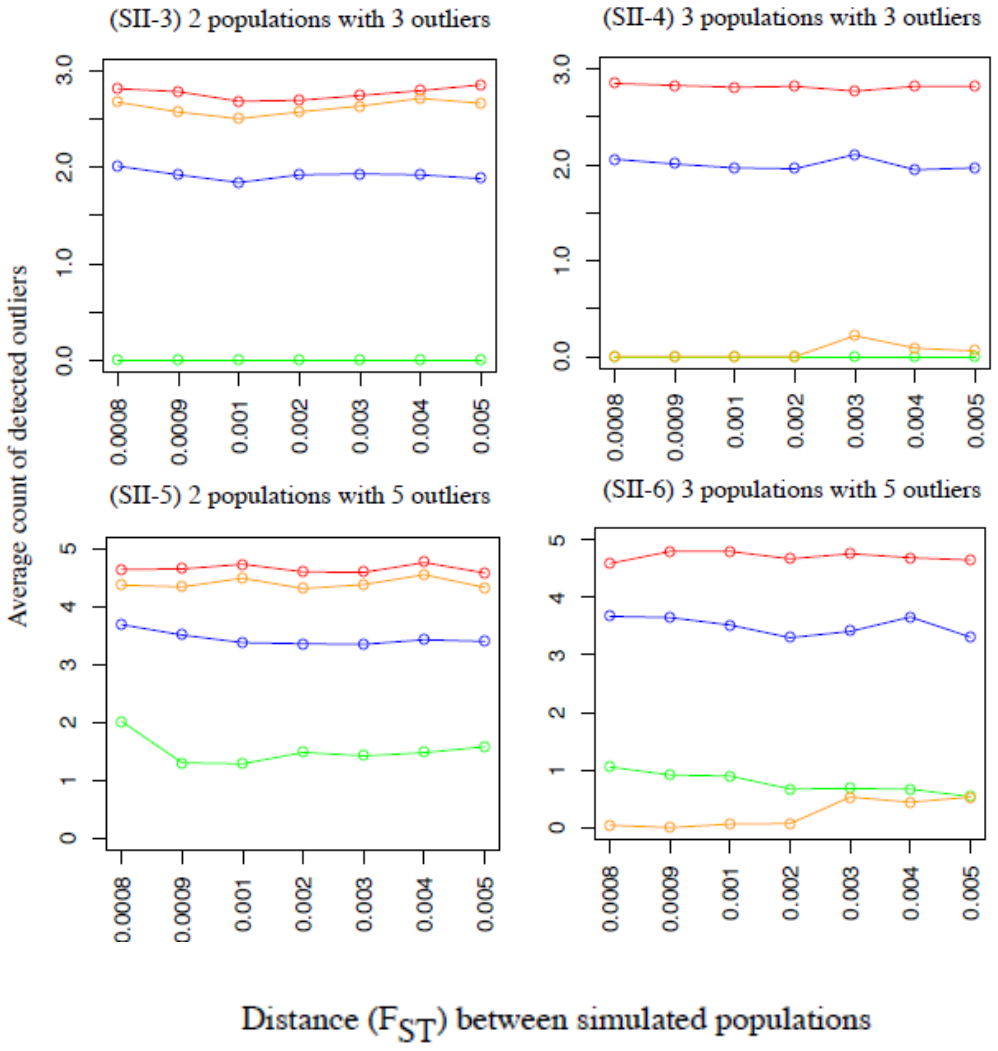
IPCAPS



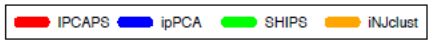
IPCAPS workflow



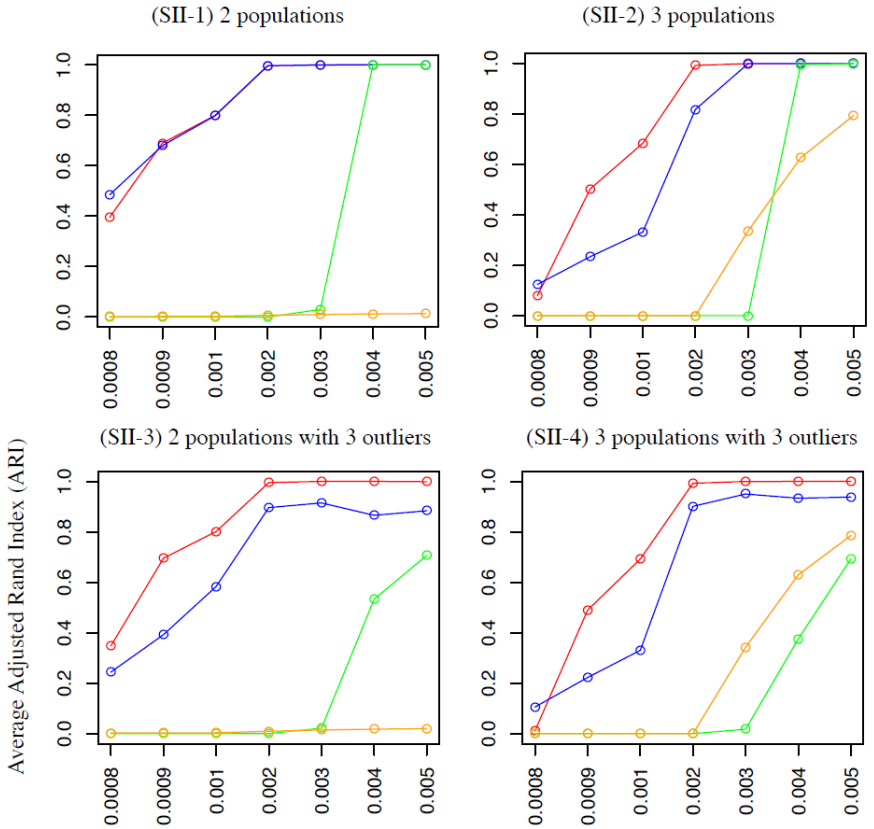
Performance of IPCAPS as outlier detection tool

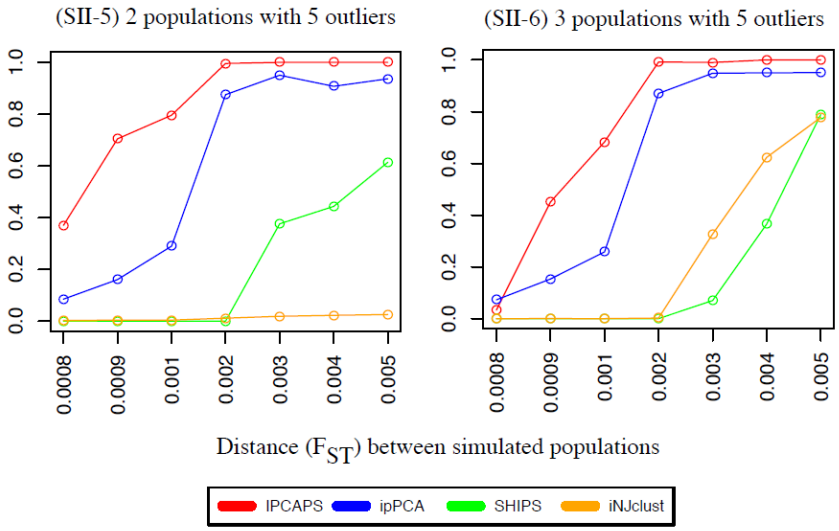


Parameters	Settings					
	SII-1	SII-2	SII-3	SII-4	SII-5	SII-6
Number of populations	2	3	2	3	2	3
Distance (F_{ST}) between populations	0.0008, 0.0009, 0.001, 0.002, 0.003, 0.004, 0.005					
Number of individuals per population	500					
Number of SNPs	10,000					
Number of outliers	0	0	3	3	5	5
Number of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



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					

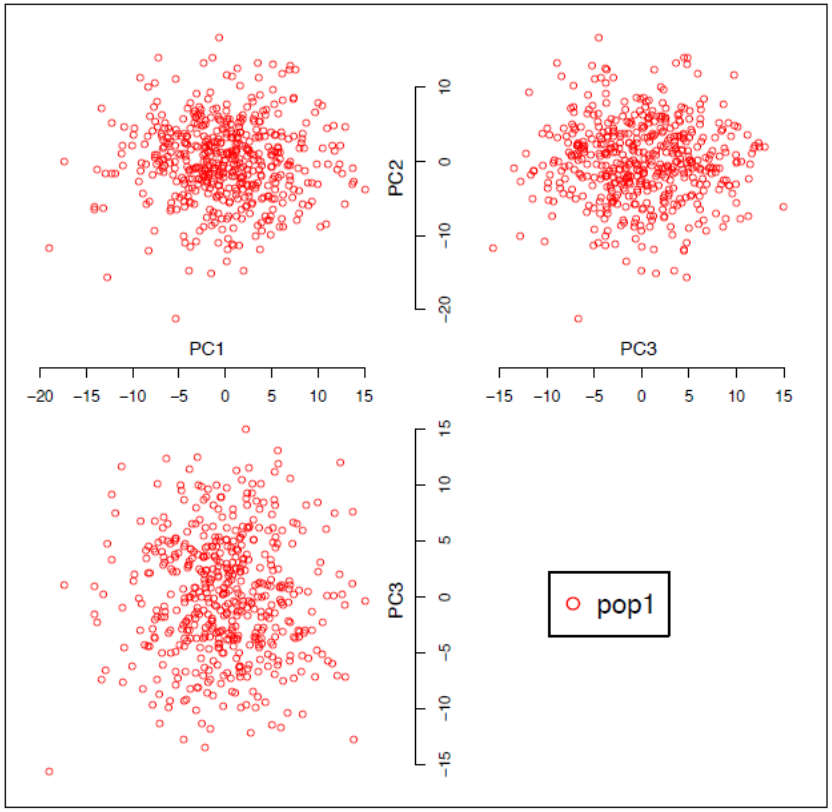
(Chaichoompu – thesis defense Oct 2017)

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

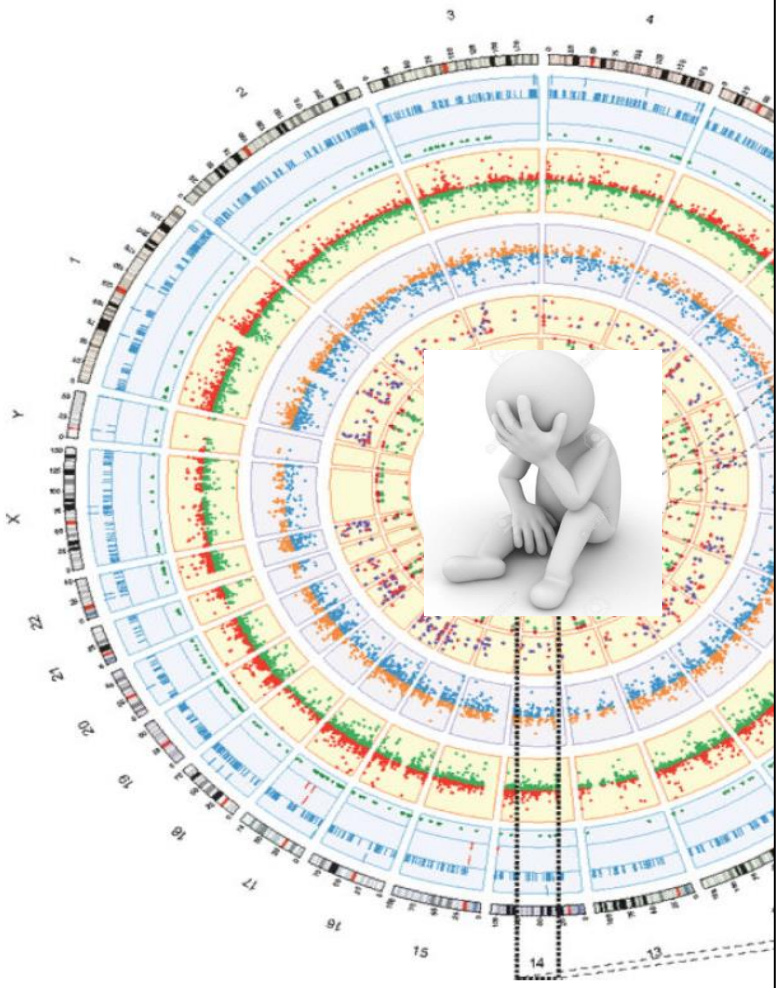
(Heath et al. 2008)

Type I error of IPCAPS



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

(Kridsadakorn Chaichoompu 2017,
PhD thesis – Chapter 2)



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

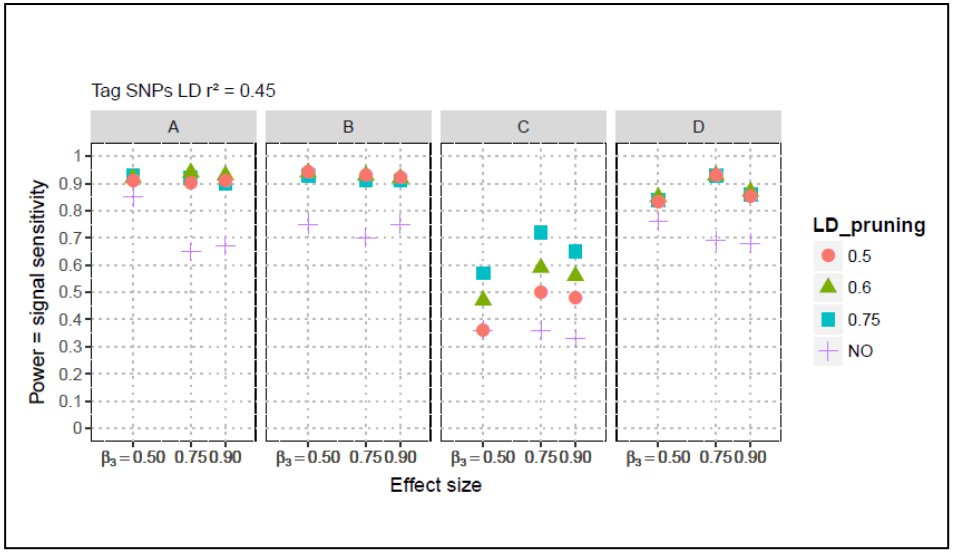
"Not in the form we are doing it. At the moment we have a very incomplete picture of what's going on, whereas if we were able to make thousands of measurements we would have a much better feeling. We just don't know, for the clinical tests, which thousand measurements are going to be most useful. We'll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

(Snyder 2014)

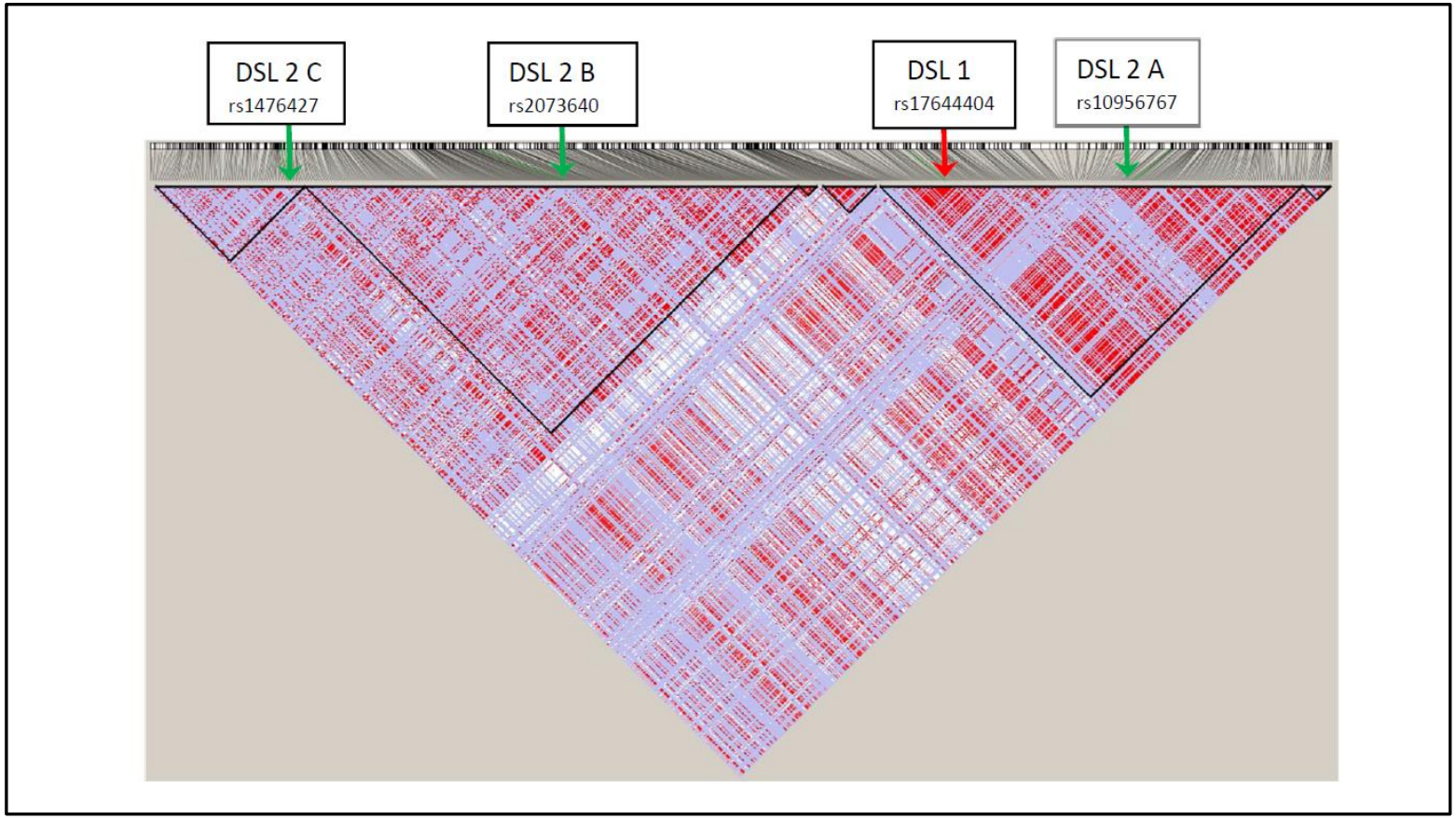
Redundancy - Informativity

Highly correlated features

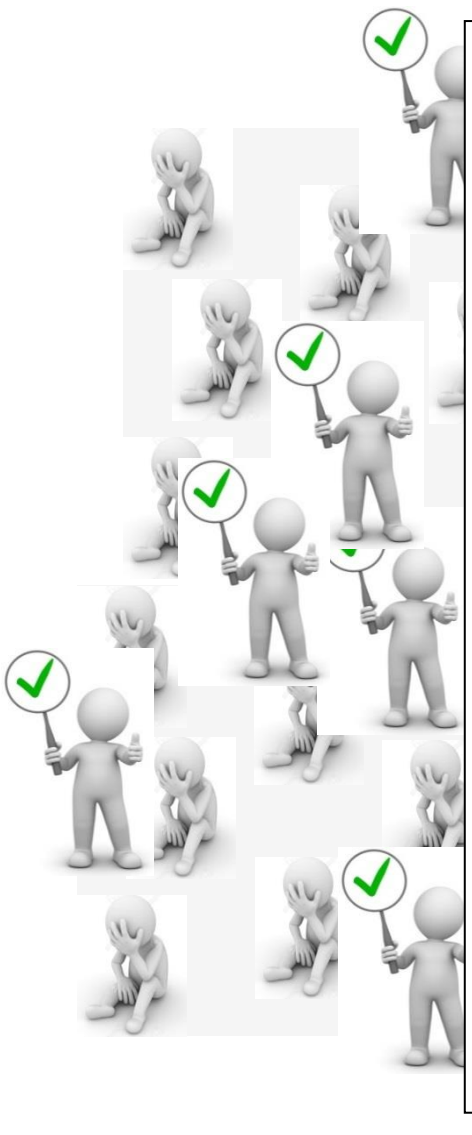
- Statistics and Linkage Disequilibrium (LD) pruning
- Results (Marc Joiret – 2017 intern BIO3):
  - Exact signal sensitivity may be low when actual actors were pruned out
  - No pruning gives the lowest signal sensitivity
  - Sufficient pruning gives acceptable signal sensitivity
  - Lowest power when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



Highly correlated features



(Marc Joiret – 2017 BIO3 intern)



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

OPEN ACCESS Freely available online

PLOS ONE

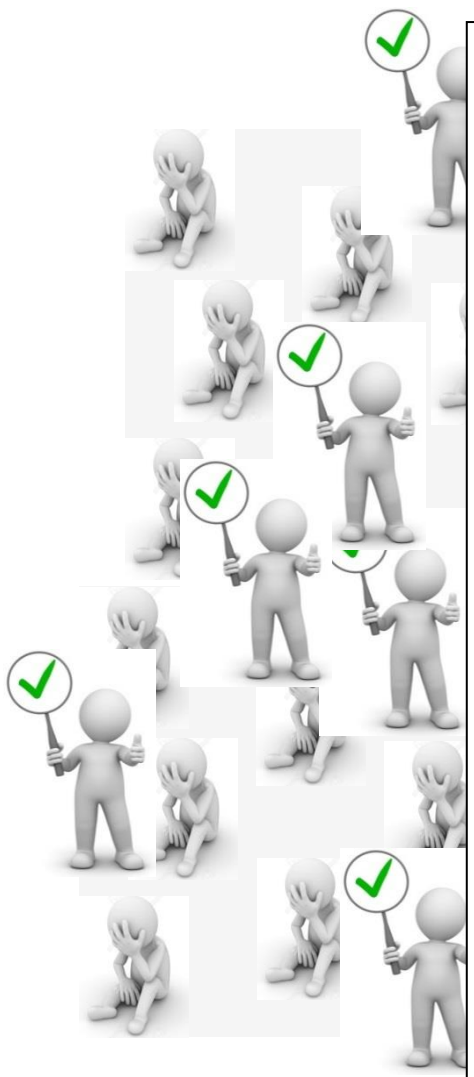
Molecular Reclassification of Crohn’s Disease: A Cautionary Note on Population Stratification

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



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]

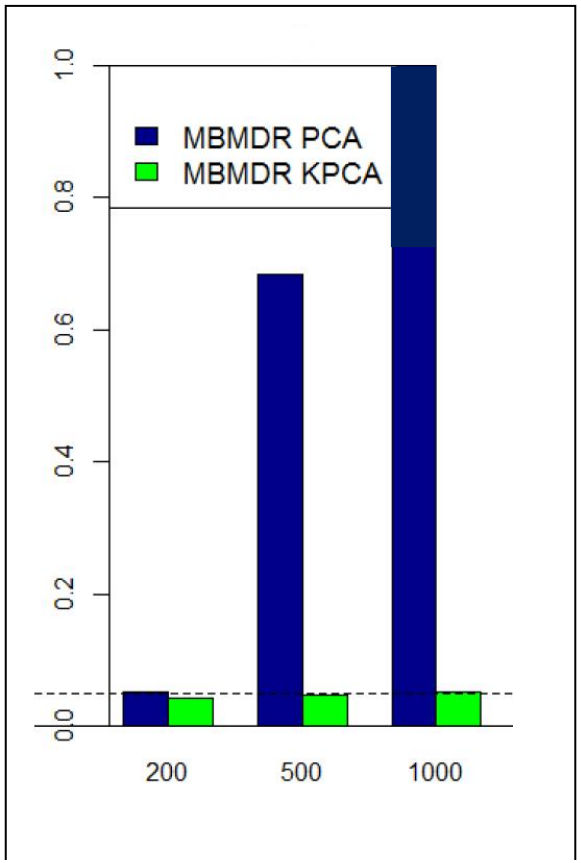
[1]Integrative Systems Biology Laboratory, Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Ibn Al Haytham Bldg. 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.

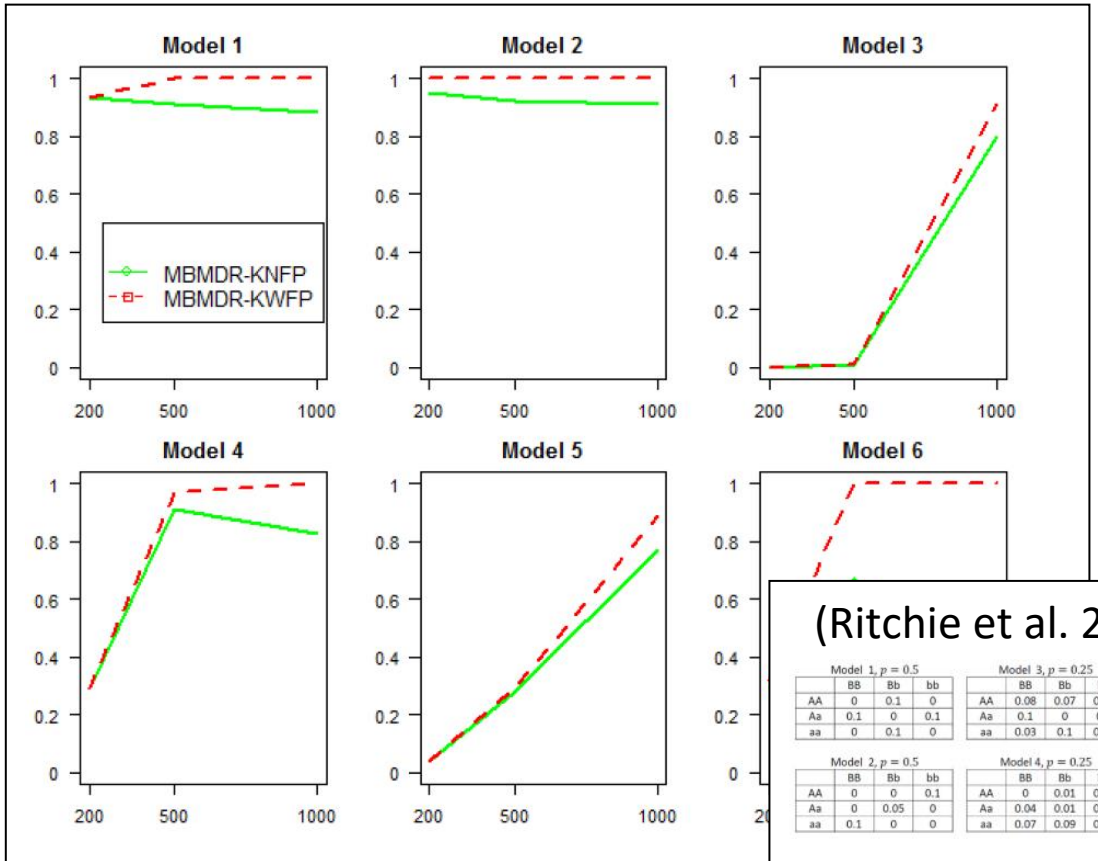
Non-linearity

(Alanis-Lobato et al. 2015)

(Non-linear) confounders

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





(Ritchie et al. 2003)

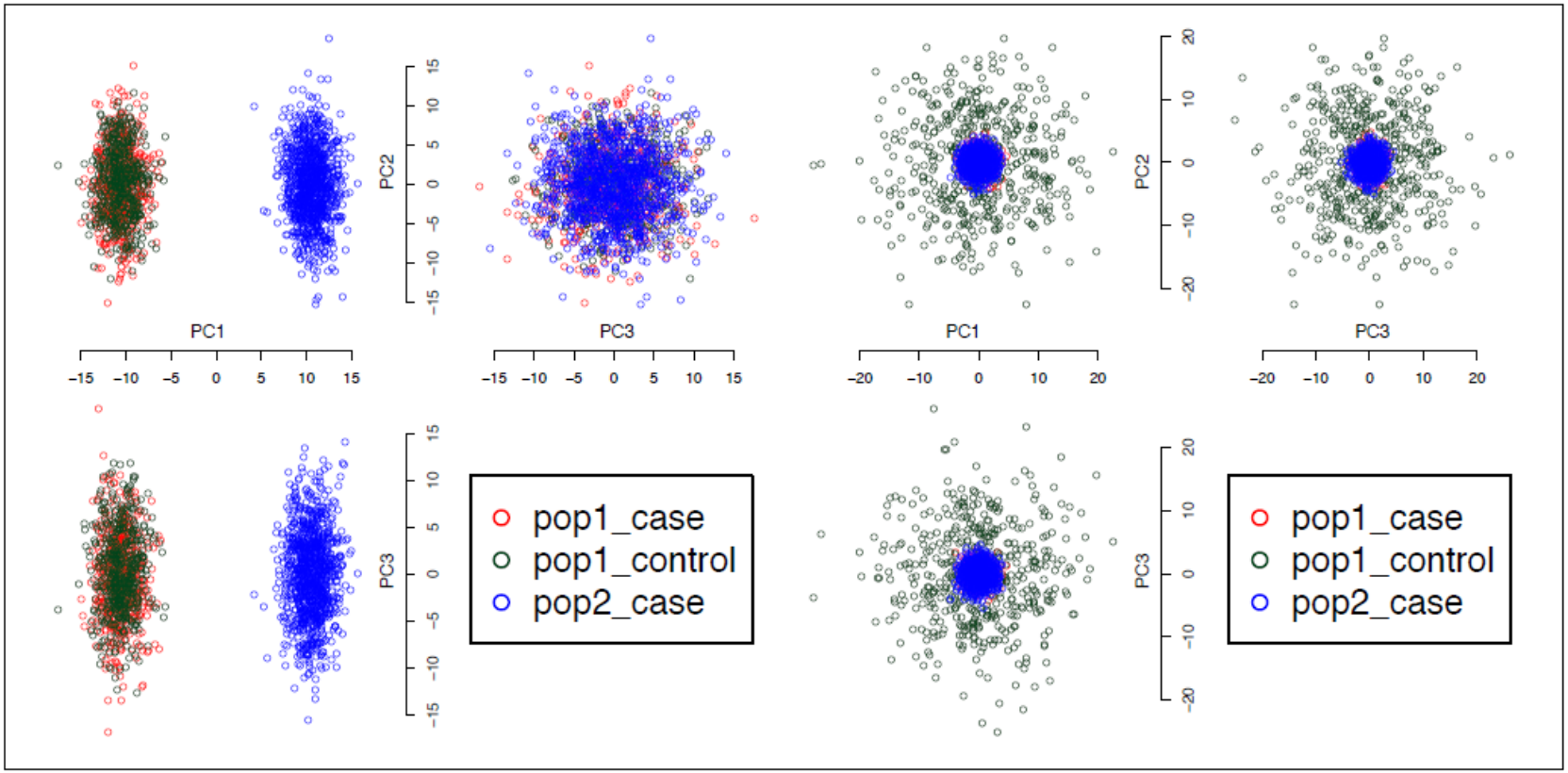
Above : 60/40 CC ratio, structural epistasis according to corresponding full penetrance Rtichie epistasis model ; Below : 50/50 (200+200)

	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

BIO

(Cattaert et al. 2011)

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

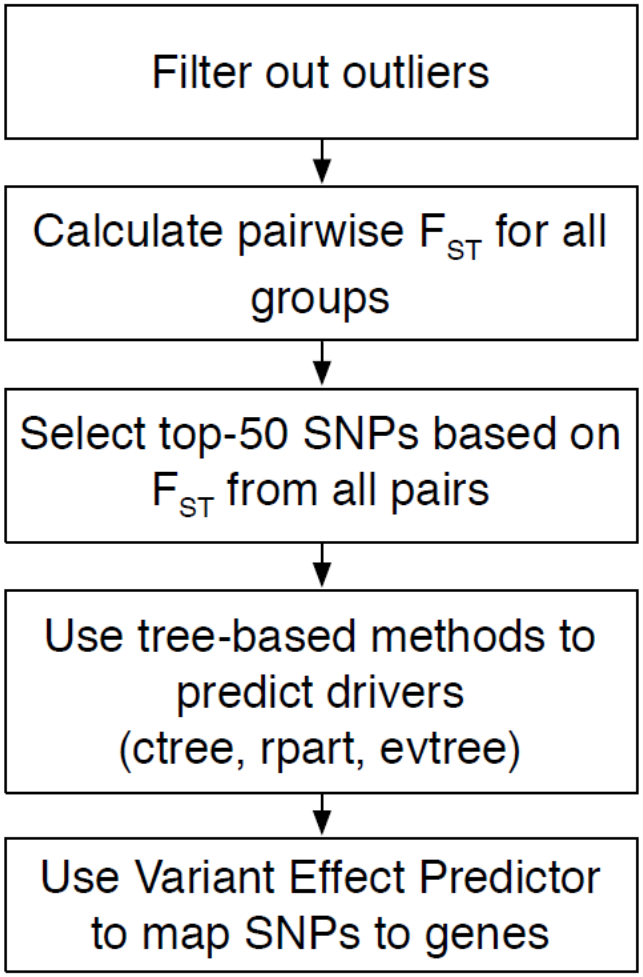
(cluster sizes less than 20 are considered to be outlying and are removed)

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

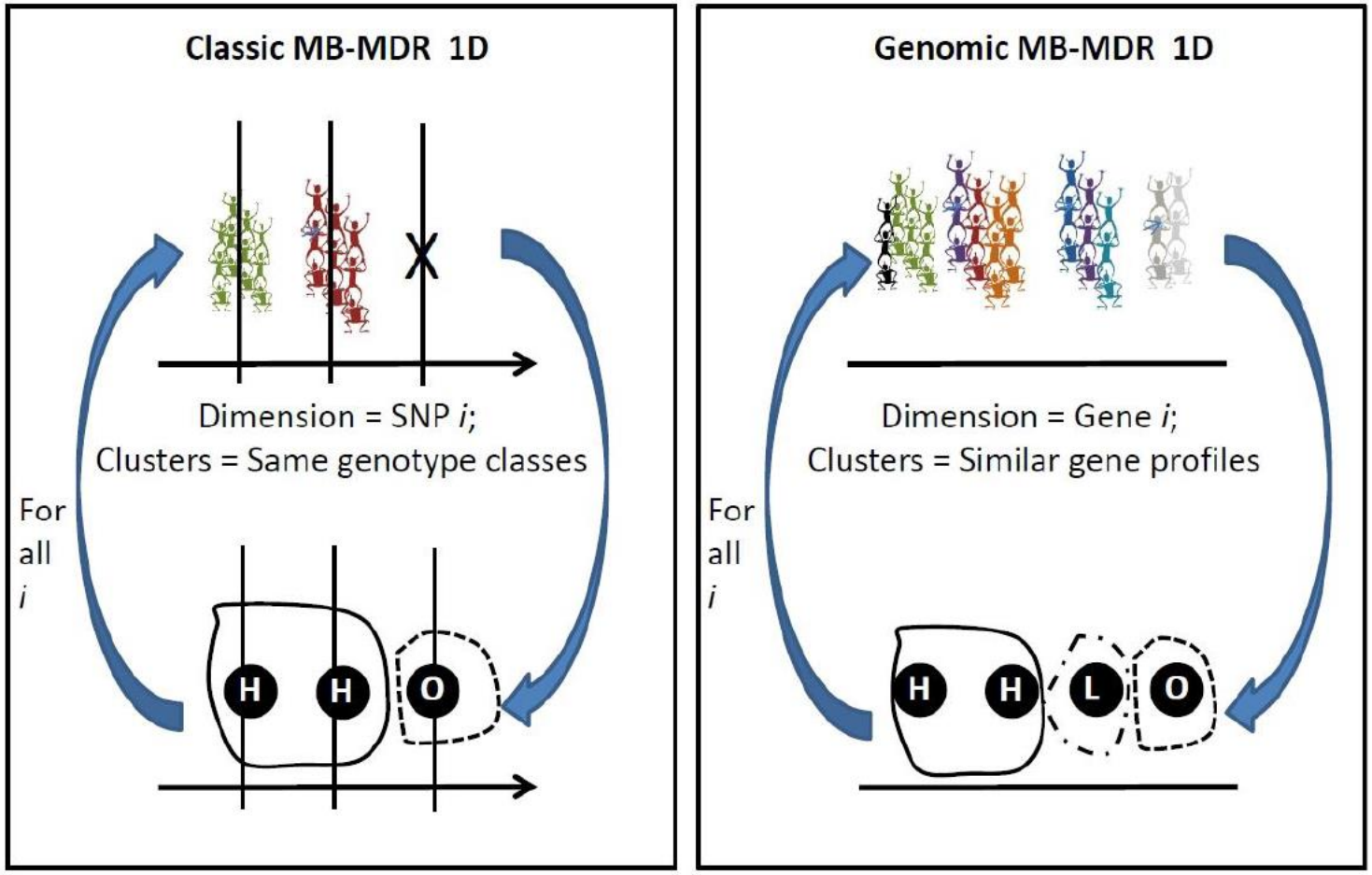


Go gene-centric - GWAS

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)

(Yuanlong Liu 2017, PhD thesis – chapter 1)

Go gene-centric - GWAIS



Original Paper

Human Heredity

Hum Hered 2015;79:157–167
DOI: 10.1159/000381286

Published online: July 28, 2015

Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis

Ramouna Fouladi Kyrylo Bessonov François Van Lishout Kristel Van Steen

Systems and Modeling Unit, Montefiore Institute, and Bioinformatics and Modeling, GIGA-R, Univer
Liège, Belgium

Gene-based representation via kernel principal components

Error control - conservativeness

Original Paper

Human Heredity

Hum Hered 2015;79:157–167
DOI: 10.1159/000381286

Published online: July 28, 2015

Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis

Ramouna Fouladi Kyrylo Bessonov François Van Lishout Kristel Van Steen

Systems and Modeling Unit, Montefiore Institute, and Bioinformatics and
Liège, Belgium

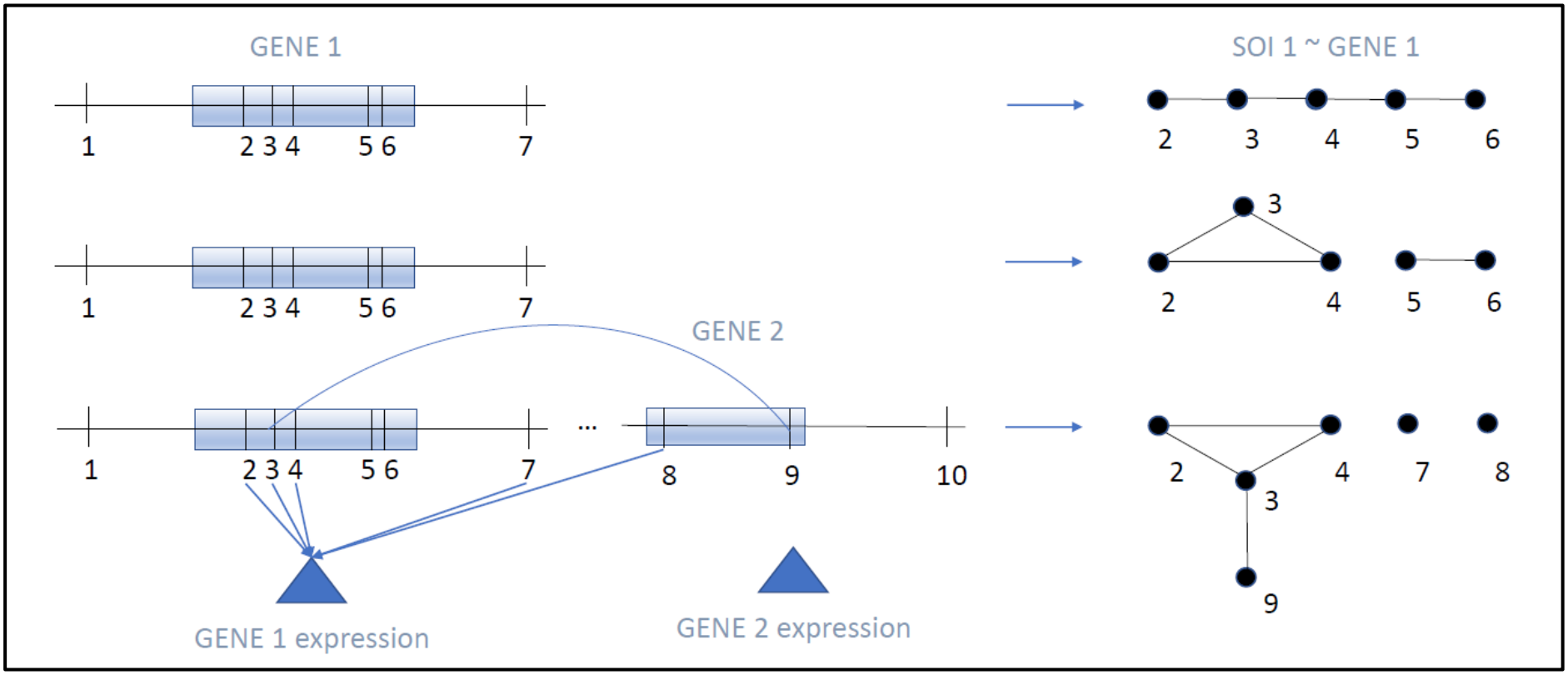
Gene-based representation via kernel principal components

and

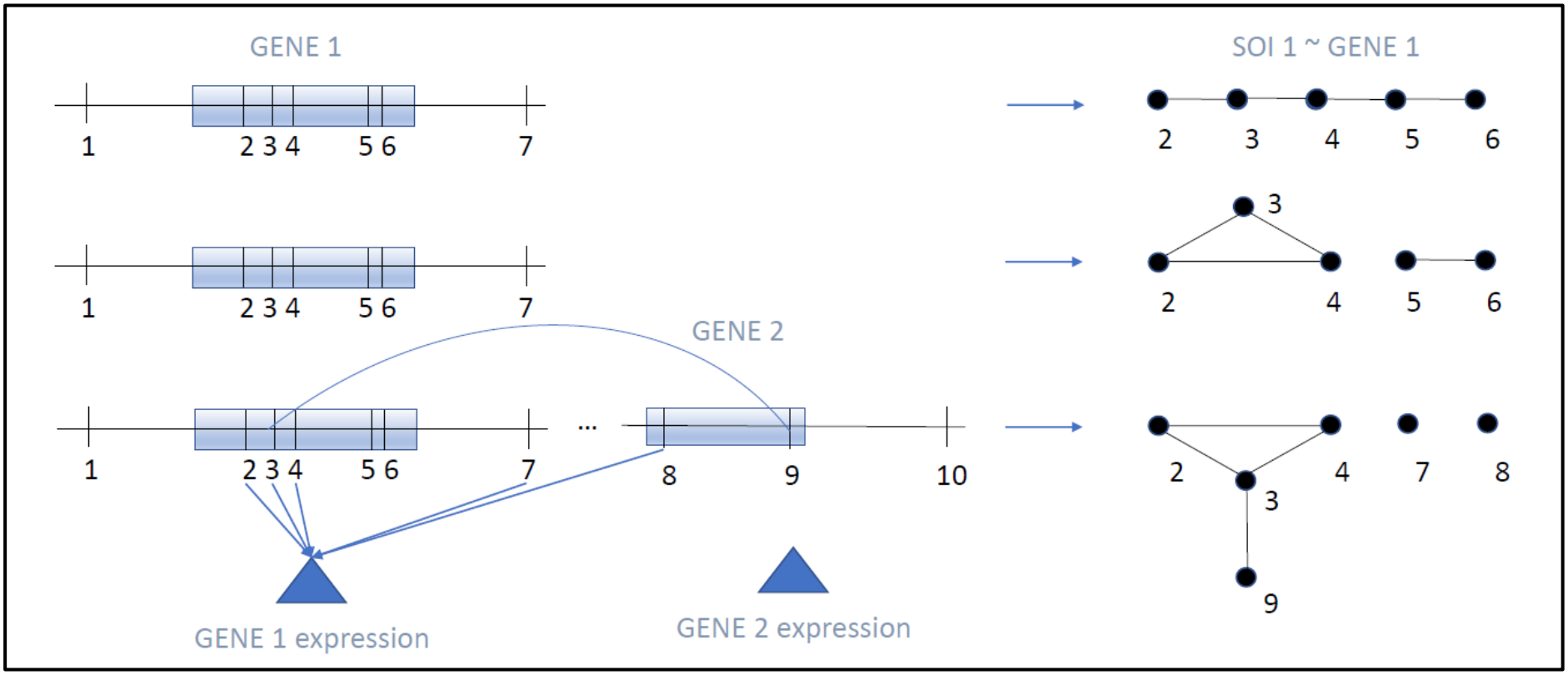
Diffusion kernels over graphs

BIO

Alternative gene representations – precision medicine



Alternative gene representations – precision medicine

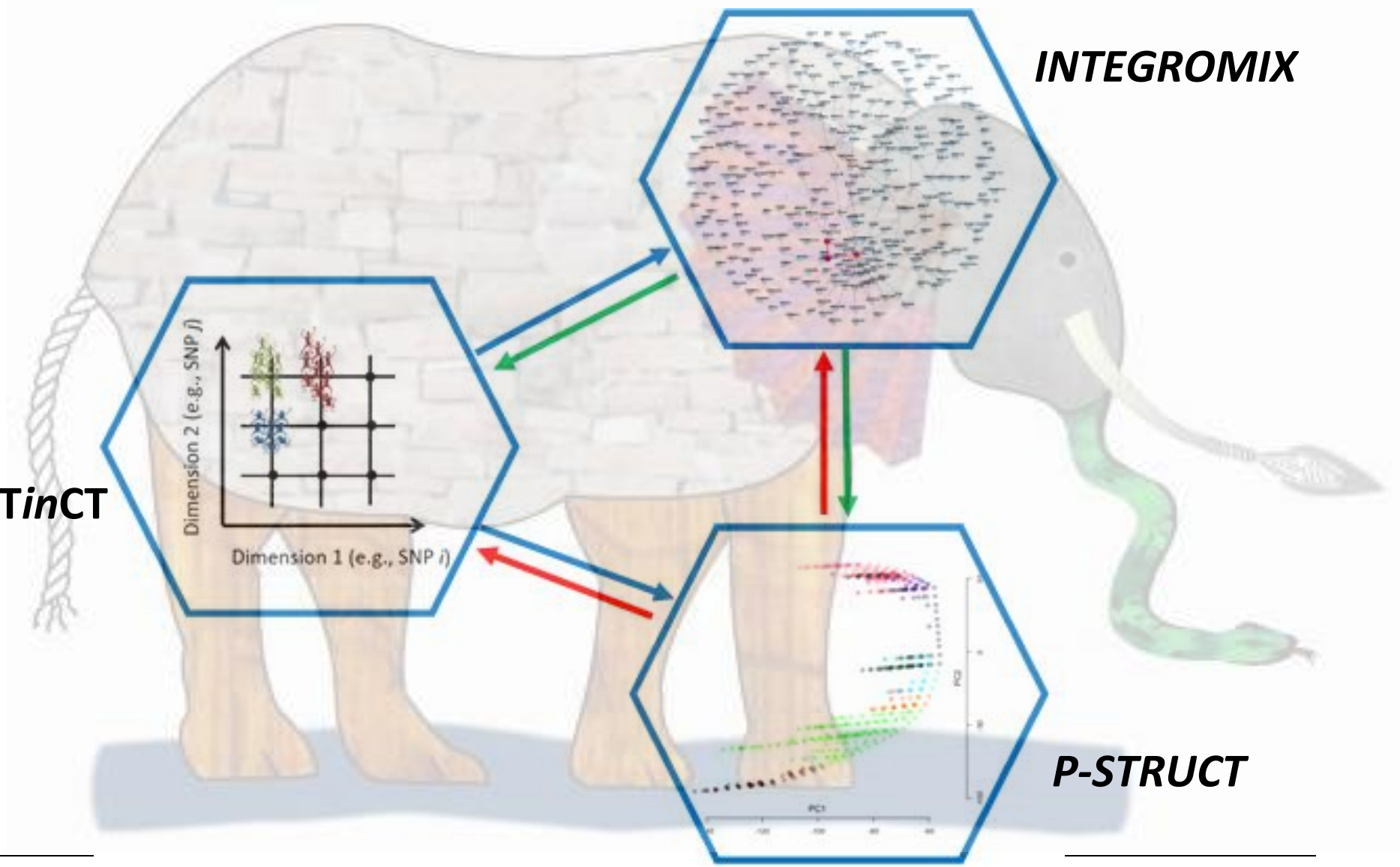


or having it trained from the data?

Take-home messages

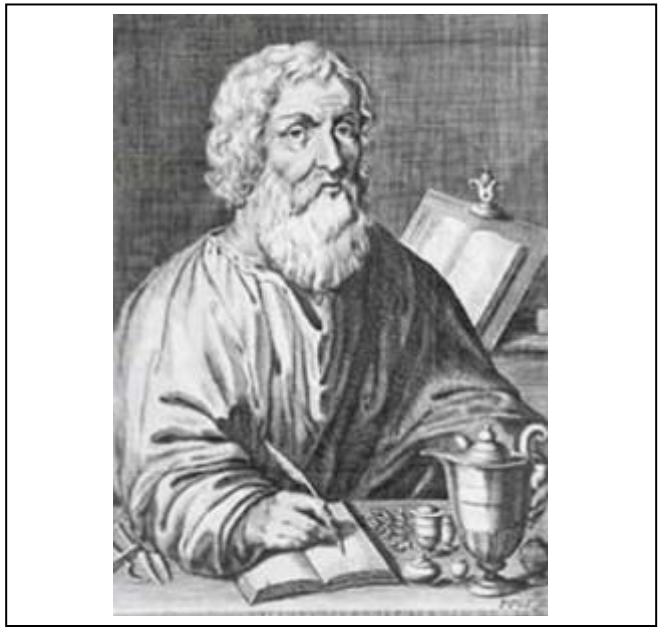
Challenges and opportunities

- 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)


INTEGROMIX
DESTinCT
Dimension 2 (e.g., SNP j)
Dimension 1 (e.g., SNP i)
P-STRUCT

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/

Supplements

Integration of inferred haplotypes in ipPCA

Information	SNPs (r^2<1) (I)	SNPs (r^2<0.8) (II)	SNPs (r^2<0.2) (III)	SNPs & LD blocks (IV)	Only LD blocks (V)
Number of SNPs (base pairs)	552K	359K	125K	97K	-
Number of LD-based haplotype blocks (blocks)	-	-	-	87K	87K
Number of clusters	4	4	4	4	4
Cluster overlap [total = 992] (individuals)	870	926	827	943	949

Reference clustering via ipPCA on Thai population
as reported by Wangkumhang et al. 2013

Performance in synthetic and real-life data

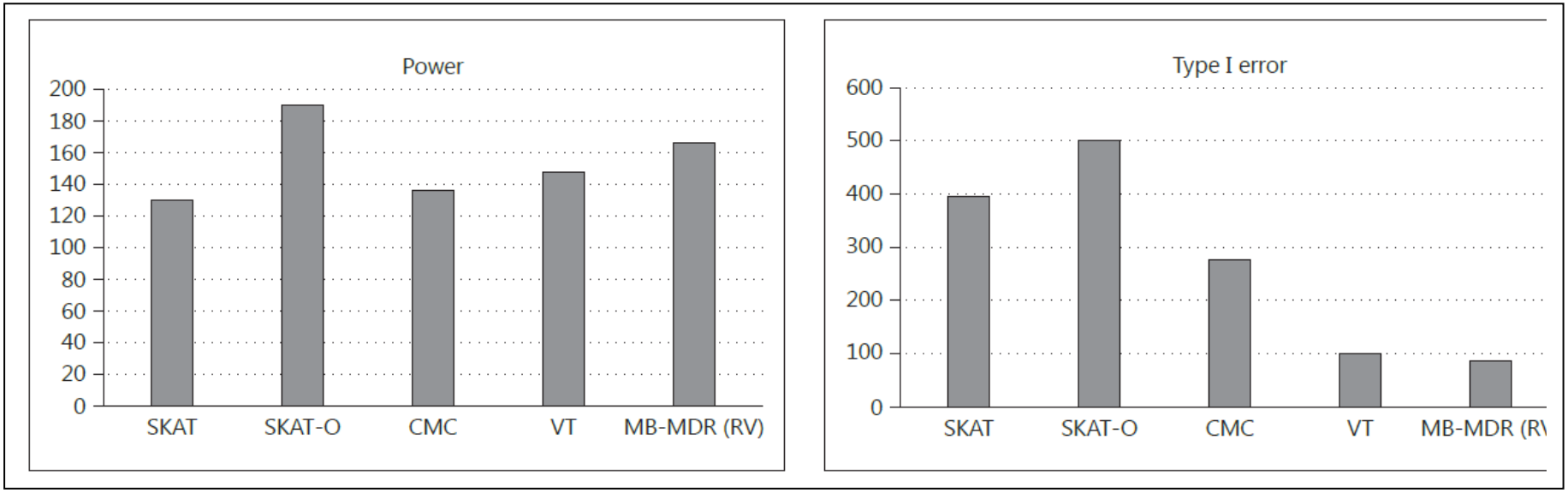
- Several methods applied to synthetic data are hampered by inflated type I error rates, including SKAT, SKAT-O and CMC (Derkach et al. 2014)
- VT seems to exhibit consistently controllable type I error rates in several scenarios (Dering et al. 2014)

(Dering et al. 2014)

	Type I	Av Power
aSum	0.12	(0.15)
C-α	0.06	0.10
CAST	0.06	0.09
CMAT	0.12	(0.16)
CMC	0.11	(0.20)
FPCA	0.05	0.07
KBAC	0.04	0.06
PWST	0.65	(0.85)
RC	0.08	(0.13)
RVT1	0.06	0.10
RVT2	0.05	0.11
SKAT	0.08	(0.11)
SKAT-O	0.10	(0.14)
VT	0.06	0.04
WSS	0.12	(0.17)

Performance in synthetic and real-life data

- Our results show similar trends, yet type I errors are smaller (restricting attention to a single chromosome -4- only)



Power : count/200 ; Type I error : count/(80*200)