Towards Molecular Reclassification of Disease

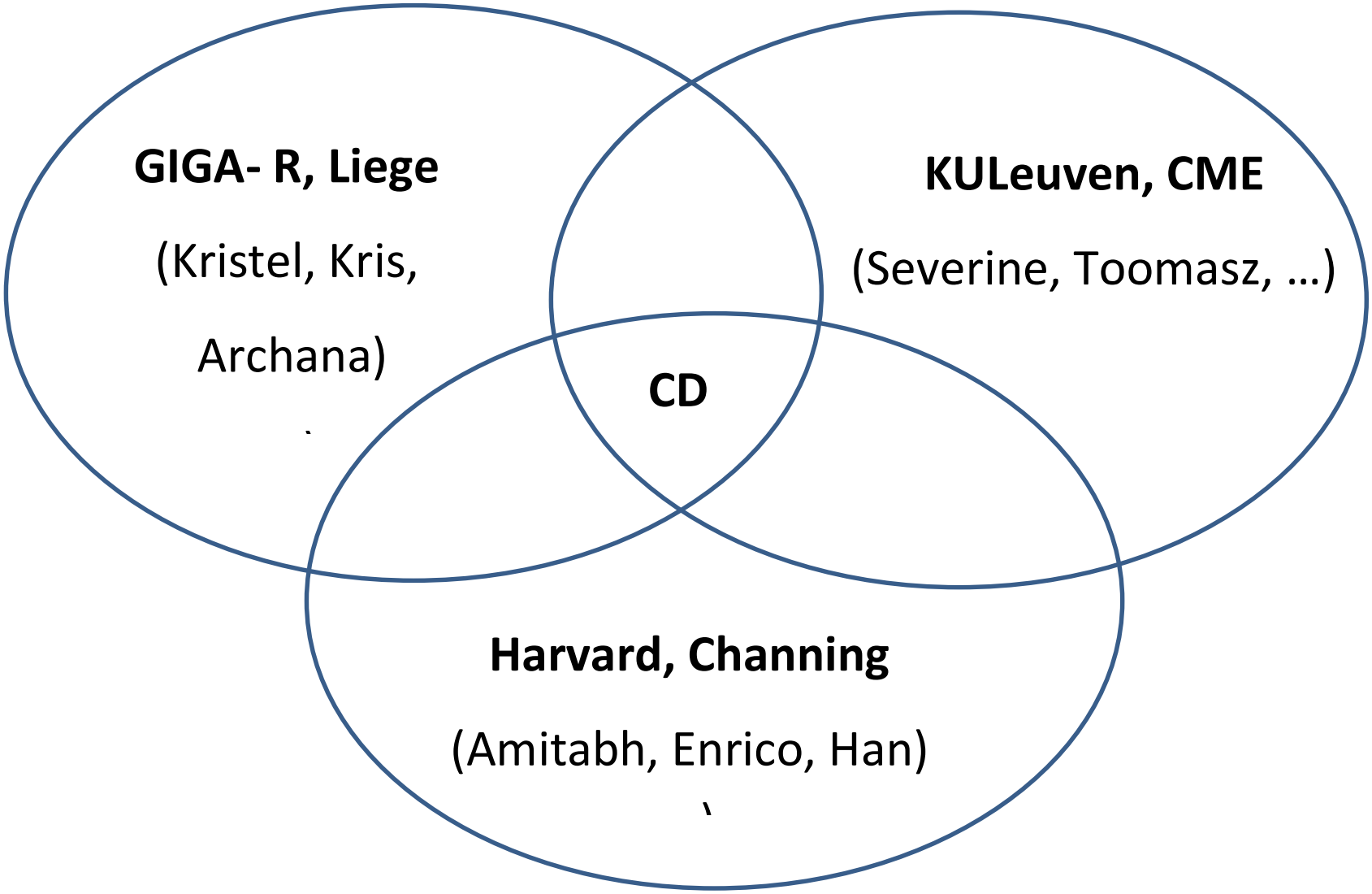
Kristel Van Steen, PhD[2] (*)

kristel.vansteen@uliege.be

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

Systems Medicine Lab, KU Leuven, Belgium

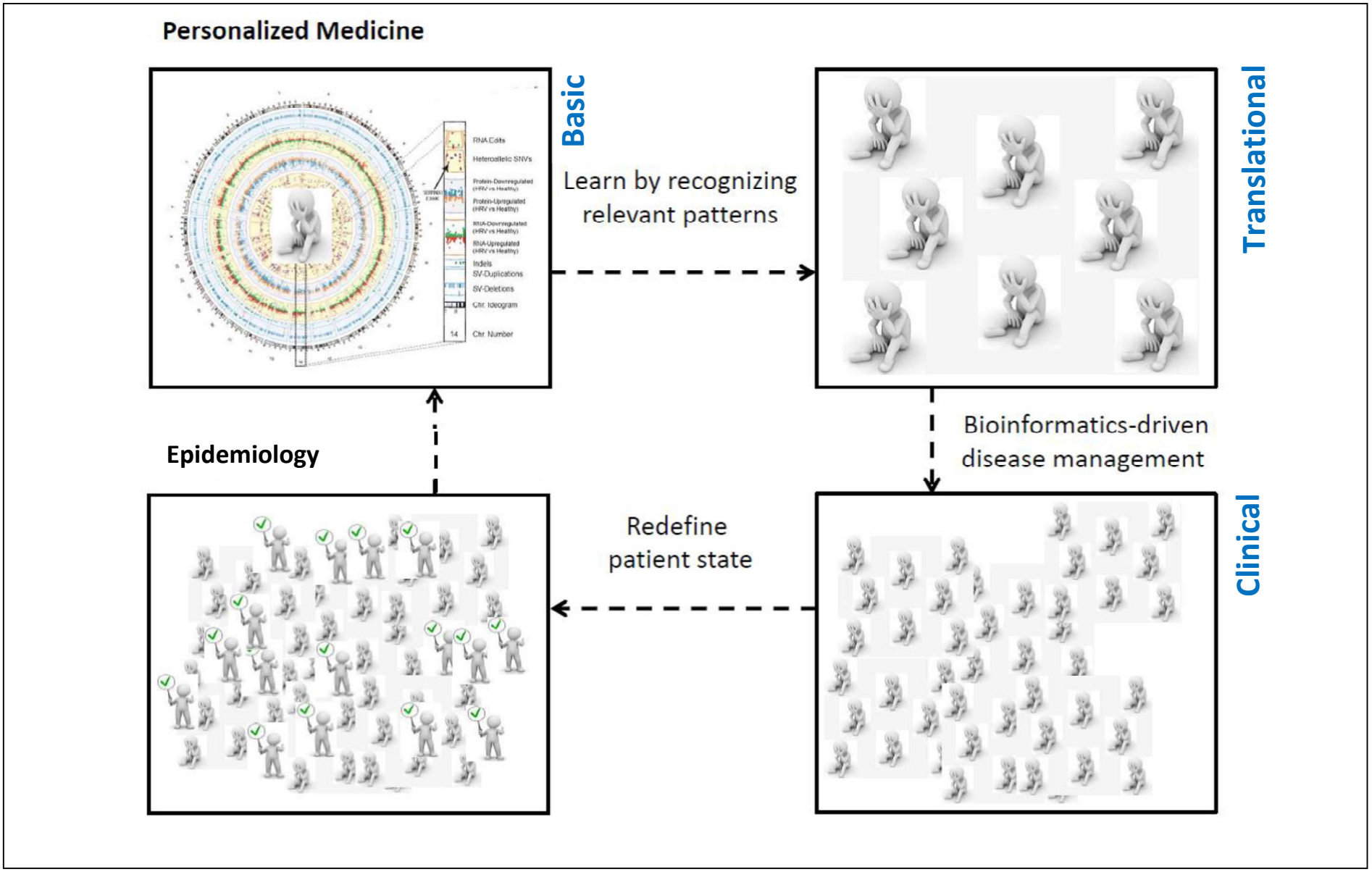
COLLABORATIVE WORK

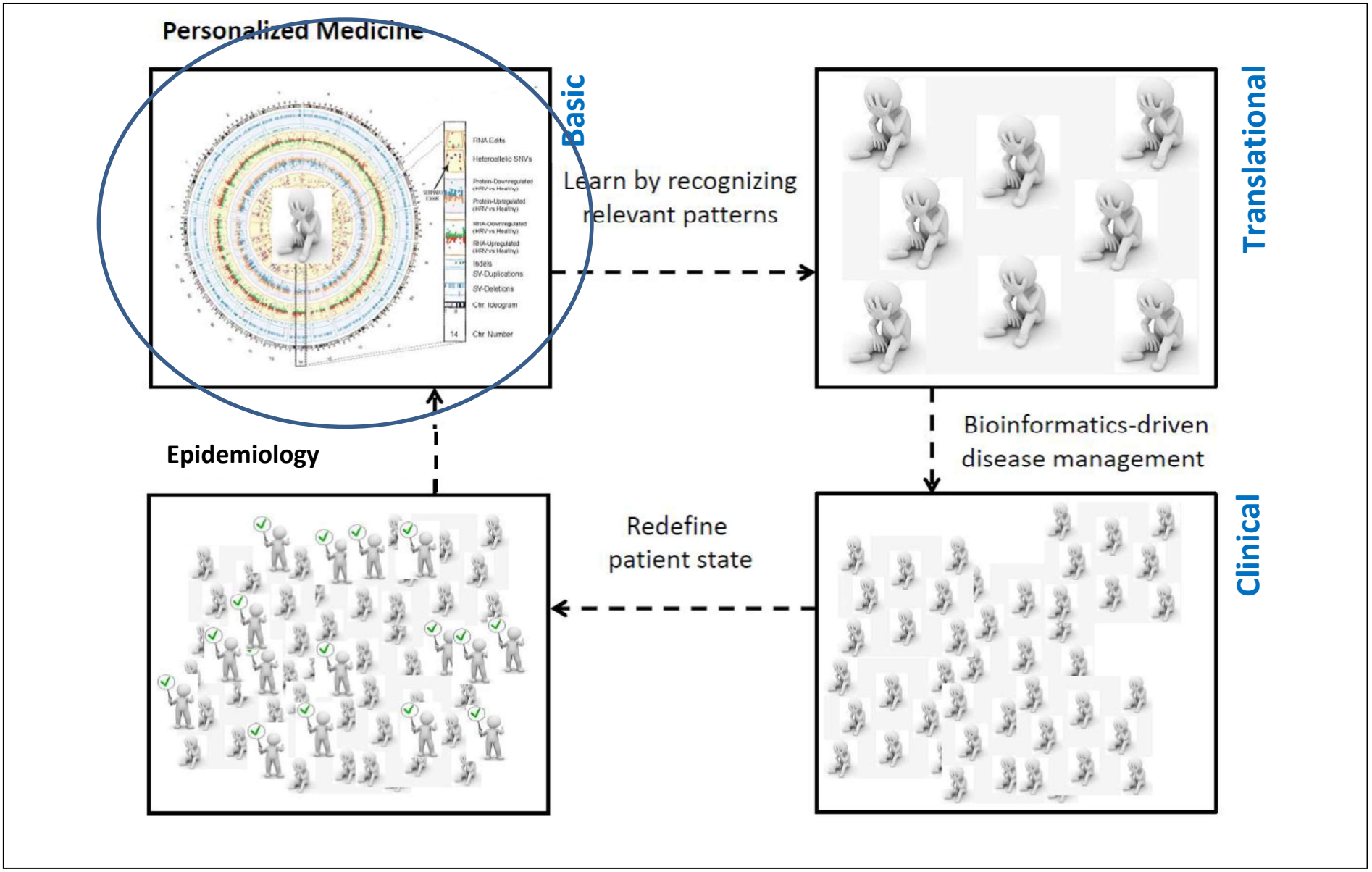


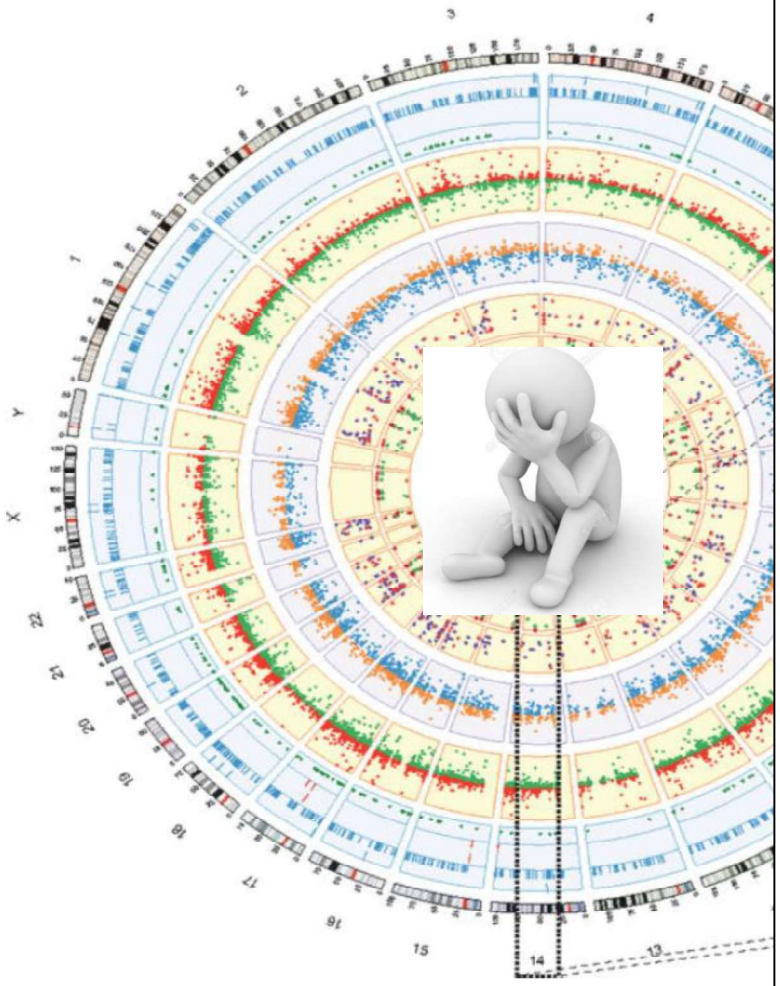
OUTLINE

- **General context: biomarker discovery for precision medicine**
  - Basic Science – How do things work?
  - Translational Science – Turning knowledge into sth useful?
  - Clinical Science – Is it really useful?
- **Application of IPCAPS to CD**
- **Take-home messages**

Biomarker discovery for Precision Medicine


Personalized Medicine
Basic
Learn by recognizing relevant patterns
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology
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


Personalized Medicine
Basic
Learn by recognizing relevant patterns
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology
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



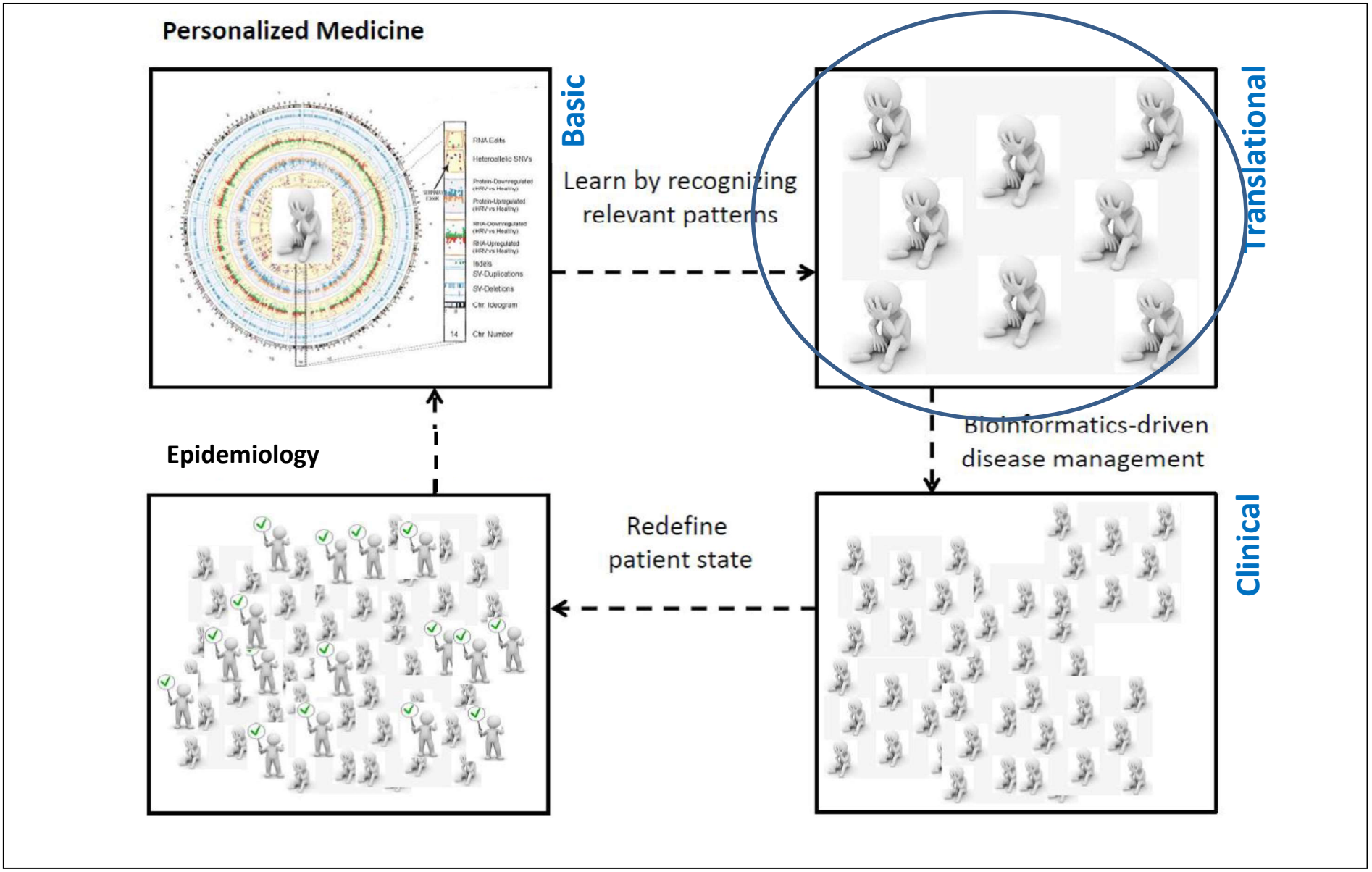
Do you think that omics profiling will be routinely used in the clinic in future?

"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."

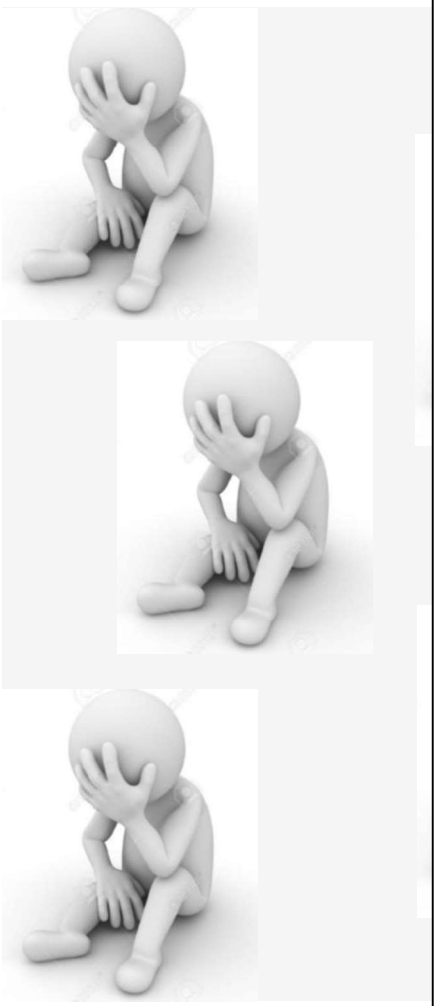
(Snyder 2014)

Redundancy – Informativity

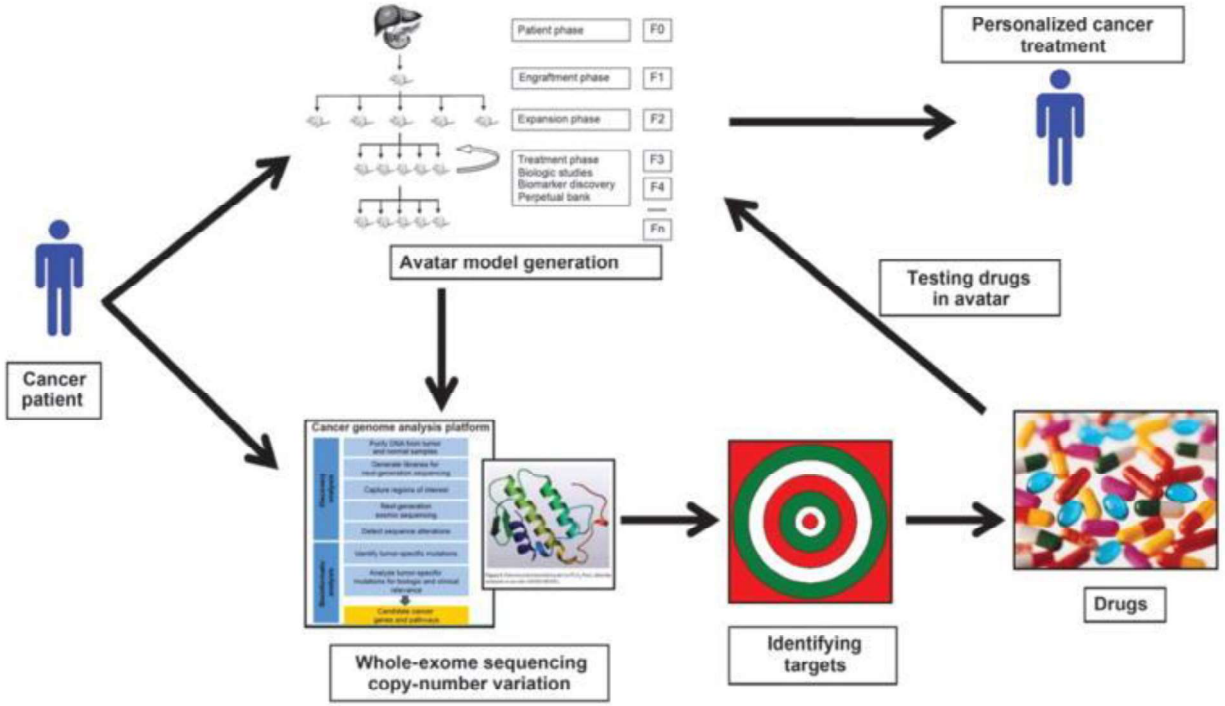
Missingness


Personalized Medicine
Basic
Learn by recognizing relevant patterns
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology
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

Bionformatics-driven treatment assignment tool

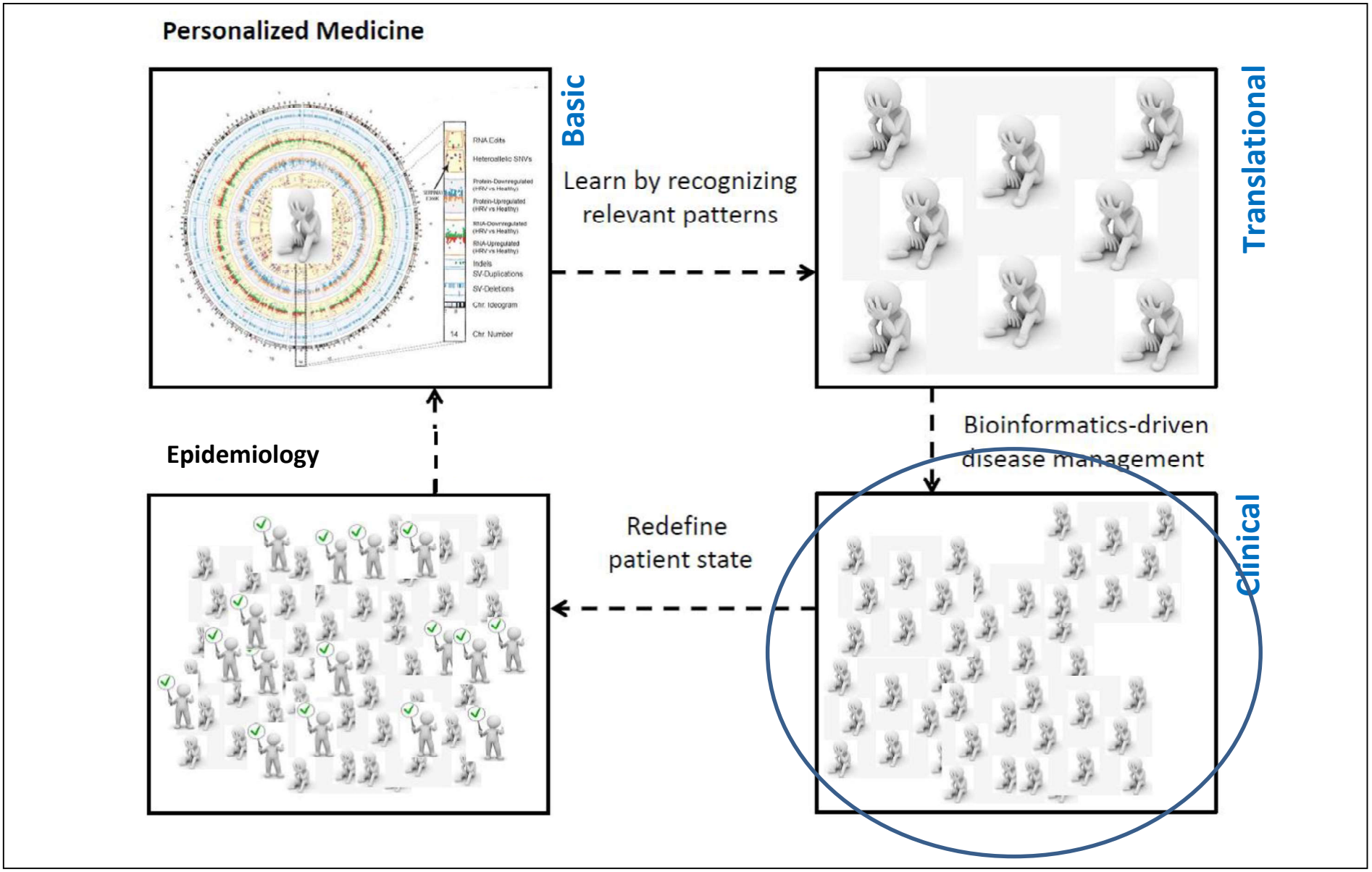


Integrating sequencing and avatar mouse models



Problems …

(Garralda et al. 2014)


Personalized Medicine
Basic
Learn by recognizing relevant patterns
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology
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

Homogeneity vs heterogeneity

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

Homogeneity vs heterogeneity

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]

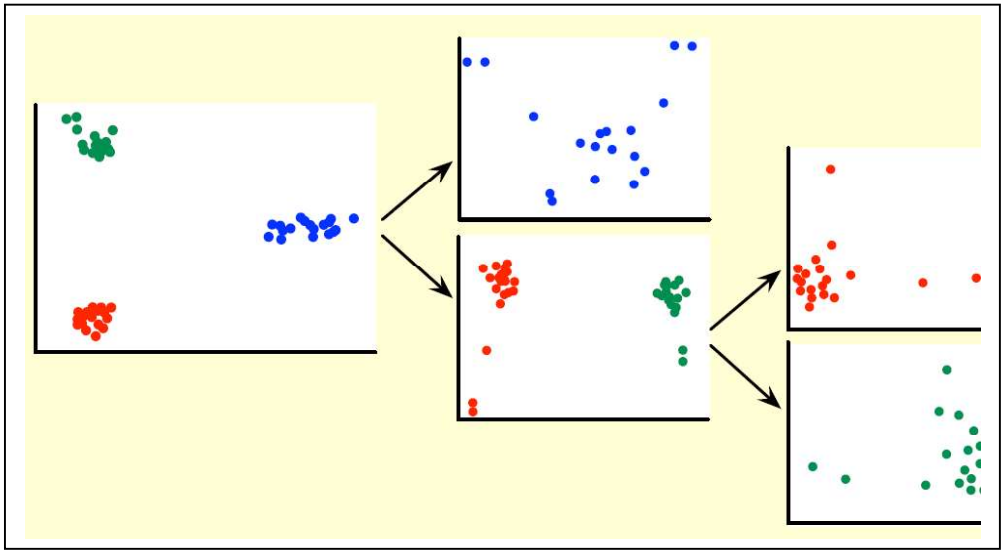
1 UMR843, INSERM, Paris, France, 2 Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, 3 UMR843, Institut National de la Sante et de la recherche Medicale, Paris, France, 4 Service de Gastroentérologie Pédiatrique, Hôpital Robert Debré, APHP, Paris, France, 5 CRC-CRB, CHI Creteil, 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 <u>as a target and a nuisance</u>

BIO3's approach: create a fine-scale structure detection tool

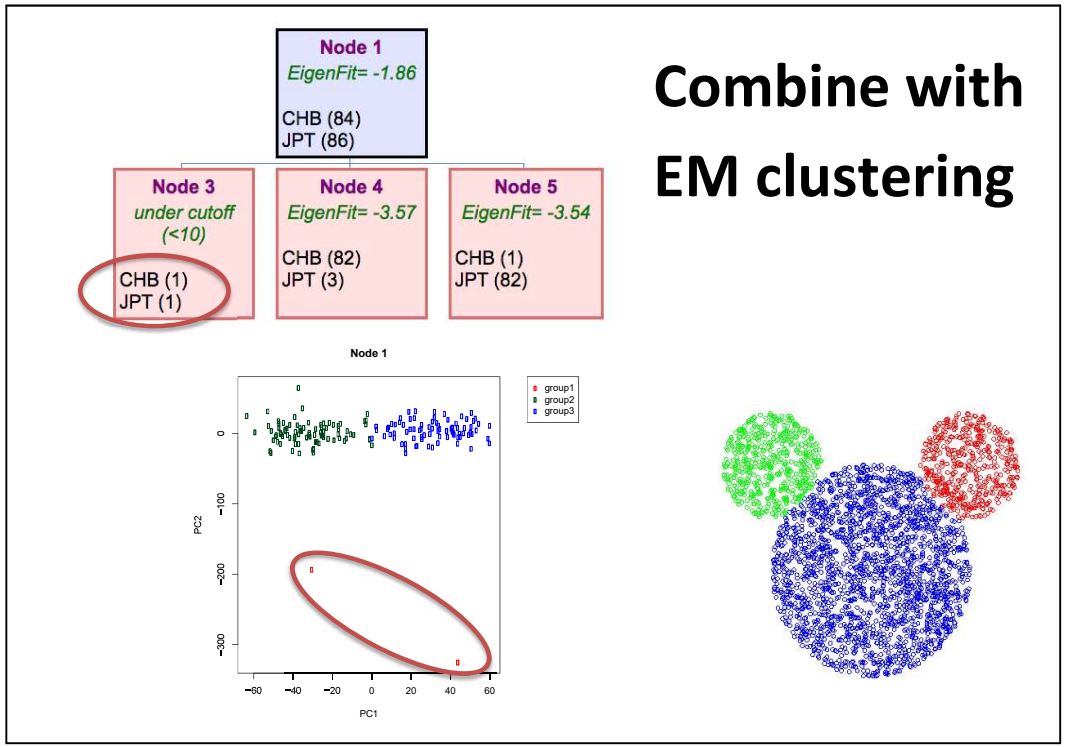
- Template: ipPCA (Intarapanich et al. 2009)
  - Performs PCA with genotype data (similar to EIGENSTRAT)
  - If substructure exists in PC space individuals are assigned to one of two clusters (2-means algorithm / fuzzy c-means)
  - Iteratively performs test for substructure and clustering on nested datasets until stopping criterium is satisfied (no substructure)



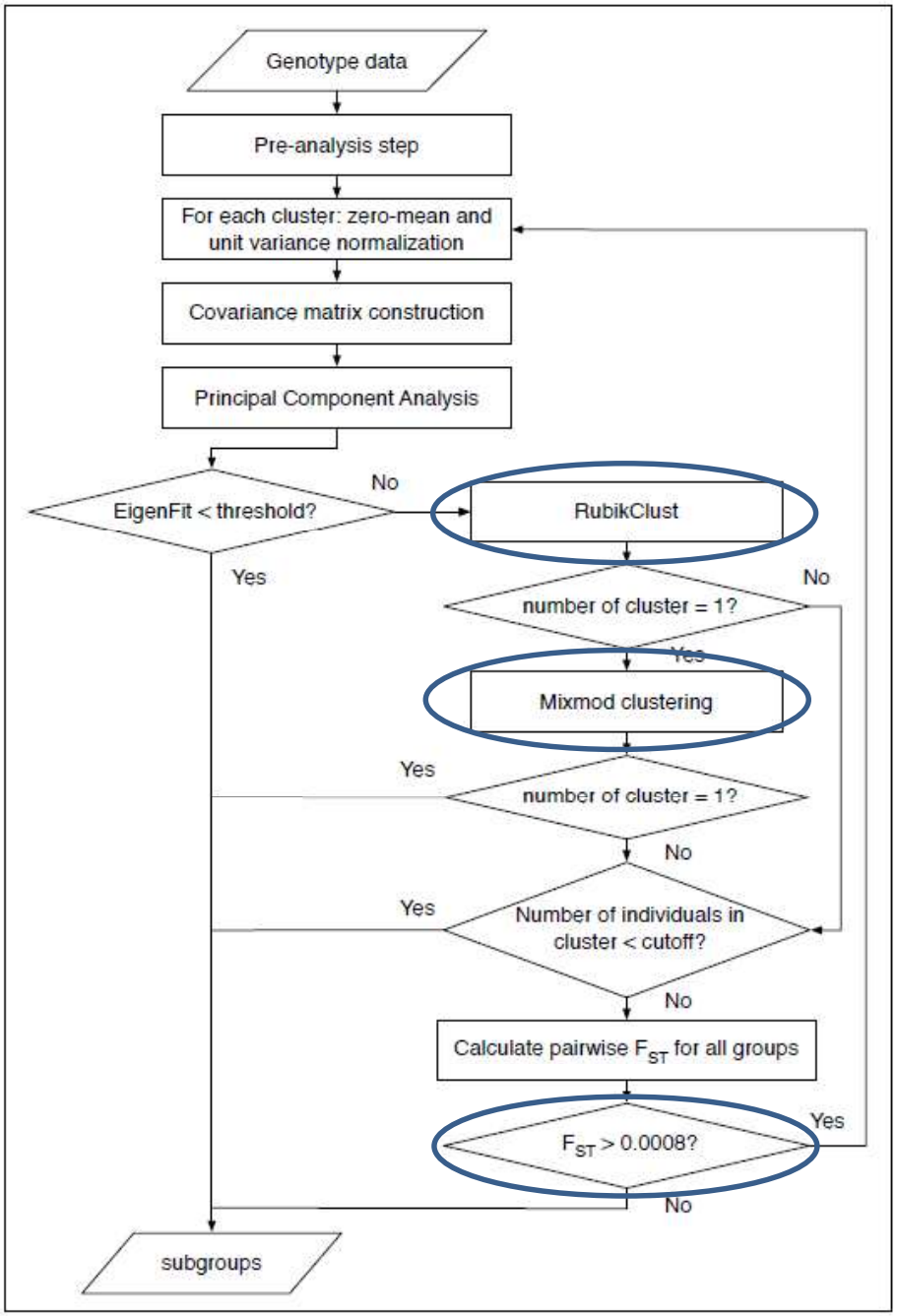
BIO3's approach: create a fine-scale clustering tool

- ipPCA
  - Pros: outperformed others (STRUCTURE – 2000) in achieving higher accuracy for highly structured populations
  - Cons: binary splitting; outlier sensitive; difficult to integrate mixed data types
- Competitors:
  - SHIPS (2012) – divisive fine-scale structure detection; computational efficiency; together with STRUCTURE best accuracy (individual assignment and nr of clusters)
  - iNJClust (2014) – non-parametric; tree clustering (phylogenetic trees); fixation index F_{ST}

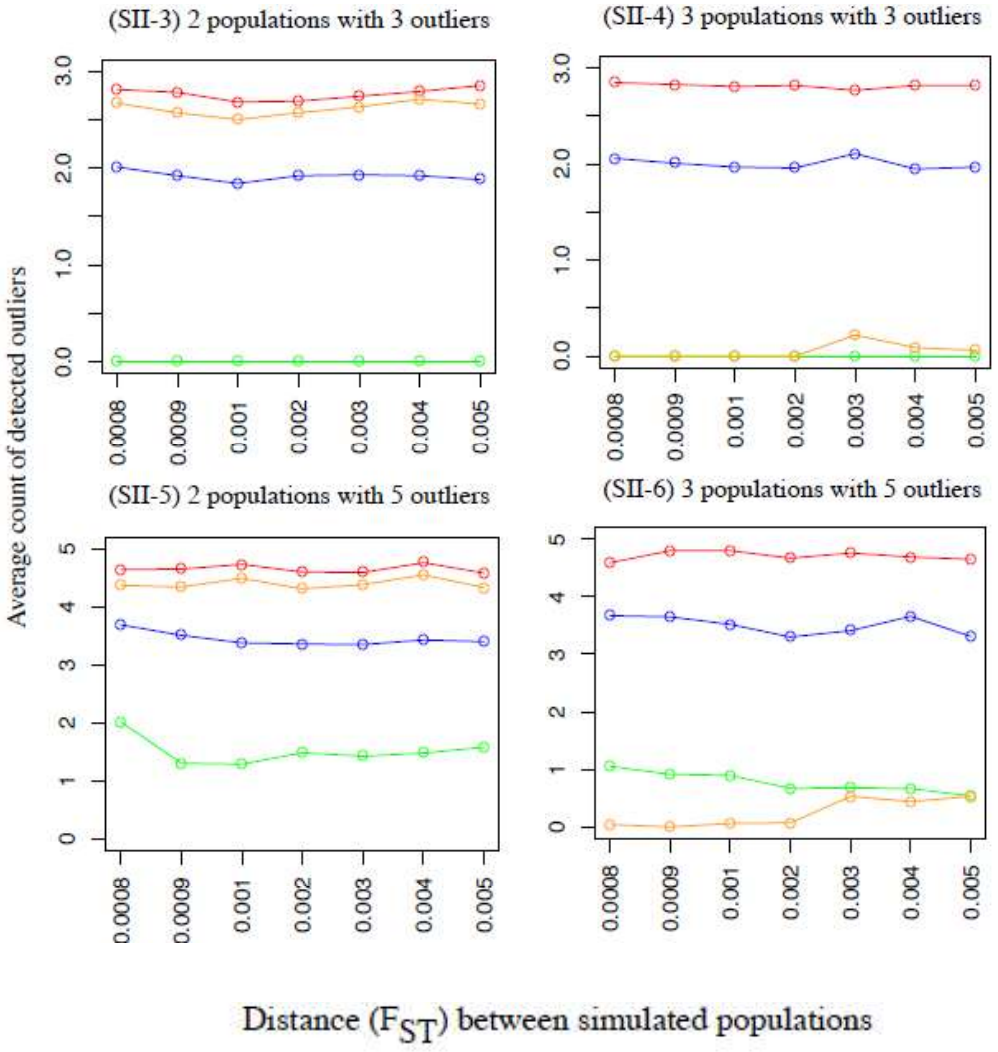
BIO3's approach: IPCAPS



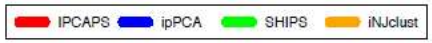
(Chaichoompu – thesis defense Oct 2017)



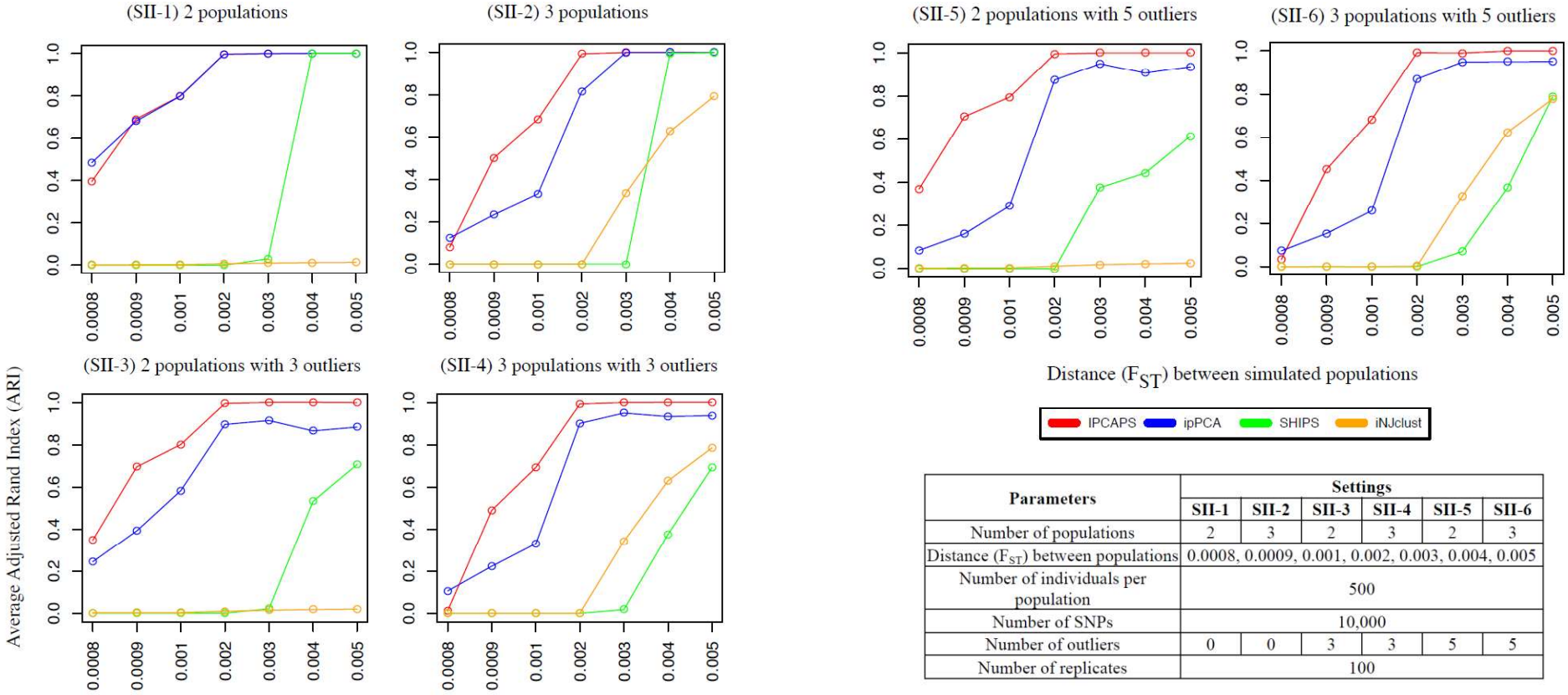
Performance of IPCAPS as outlier detection tool



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					

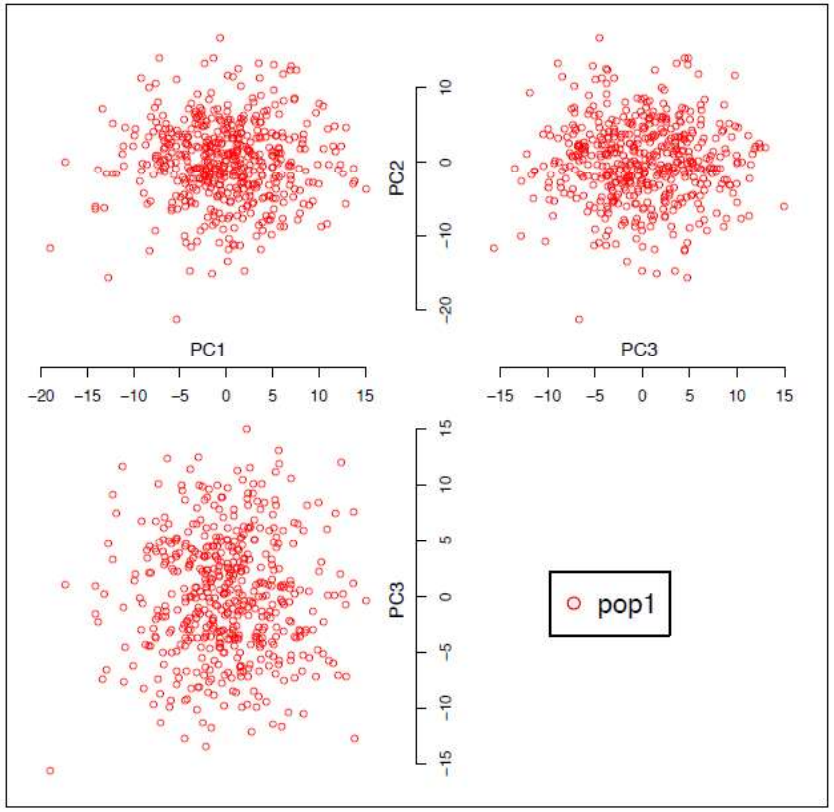
(Chaichoompu – thesis defense Oct 2017)

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; more on https://www.biorxiv.org/content/10.1101/234989v1.full)

Chaichoompu *et al. Source Code for Biology and Medicine* (2019) 14:2
https://doi.org/10.1186/s13029-019-0072-6

Source Code for Biology and Medicine

SOFTWARE Open Access

IPCAPS: an R package for iterative pruning to capture population structure

Kridsadakorn Chaichoompu[1*], Fentaw Abegaz[1], Sissades Tongsima[2], Philip James Shaw[3], Anavaj Sakuntabhai[4,5], Luísa Pereira[6,7] and Kristel Van Steen[1,8*]

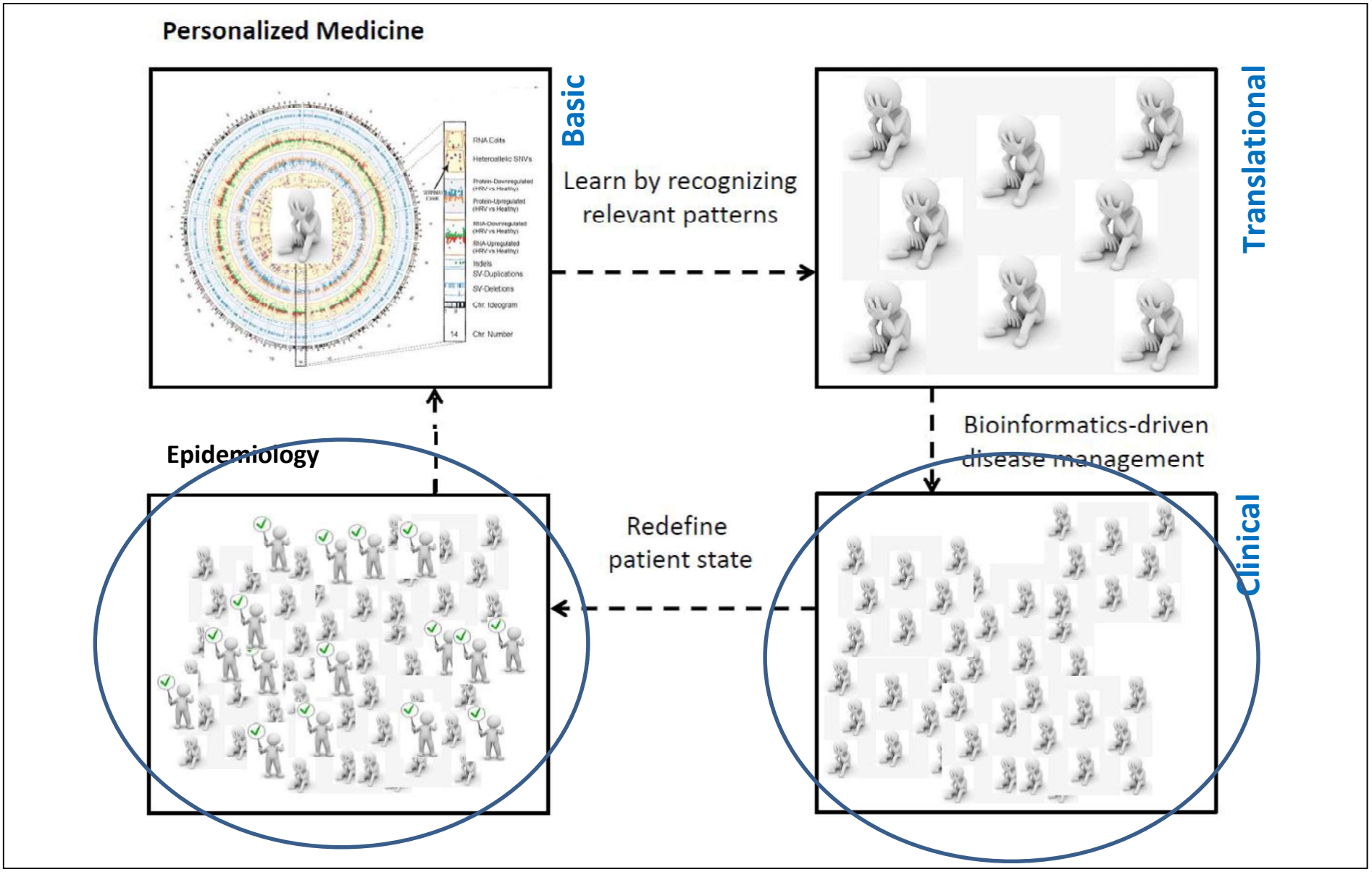
Abstract

Background: Resolving population genetic structure is challenging, especially when dealing with closely related or geographically confined populations. Although Principal Component Analysis (PCA)-based methods and genomic variation with single nucleotide polymorphisms (SNPs) are widely used to describe shared genetic ancestry, improvements can be made especially when fine-scale population structure is the target.

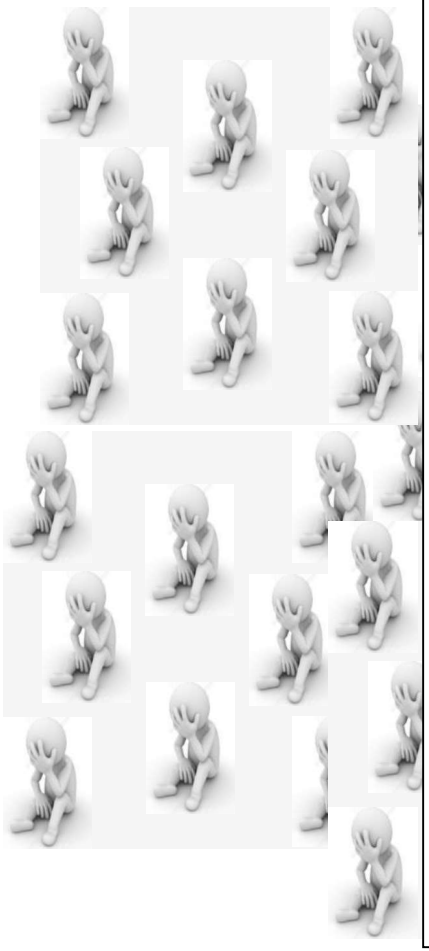
Results: This work presents an R package called IPCAPS, which uses SNP information for resolving possibly fine-scale population structure. The IPCAPS routines are built on the iterative pruning Principal Component Analysis (ipPCA) framework that systematically assigns individuals to genetically similar subgroups. In each iteration, our tool is able to detect and eliminate outliers, hereby avoiding severe misclassification errors.

Conclusions: IPCAPS supports different measurement scales for variables used to identify substructure. Hence, panels of gene expression and methylation data can be accommodated as well. The tool can also be applied in patient sub-phenotyping contexts. IPCAPS is developed in R and is freely available from http://bio3.giga.ulg.ac.be/ipcaps

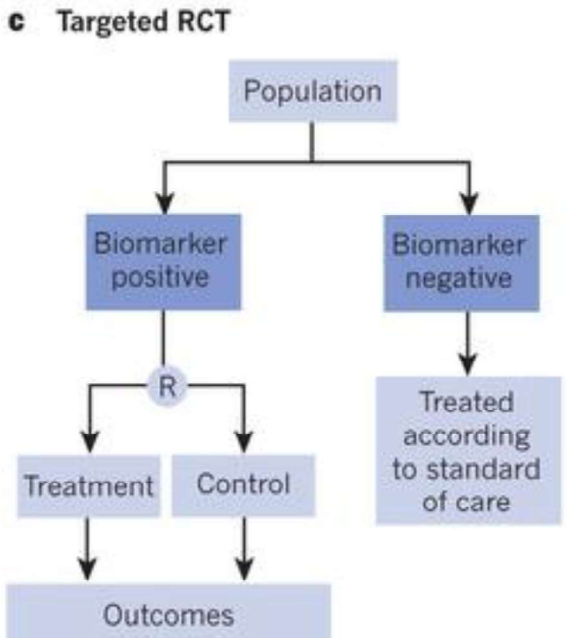
Keywords: Fine-scale structure, Iterative pruning, Population clustering, Population genetics, Outlier detection

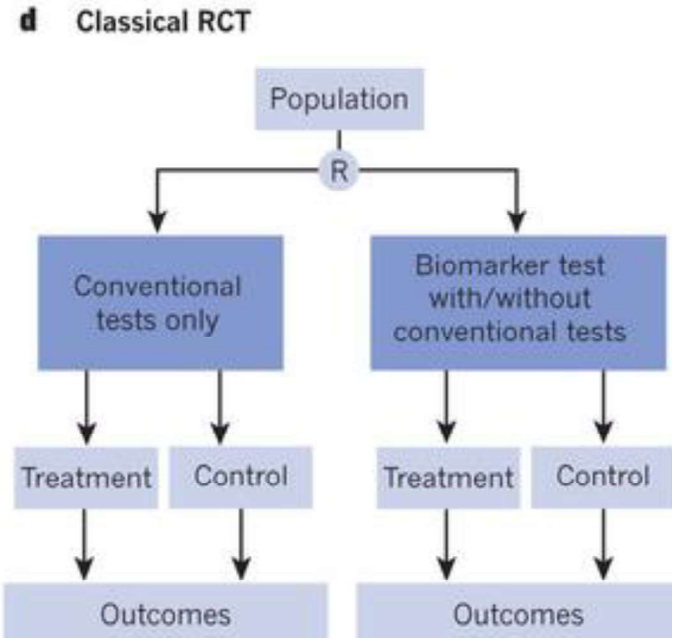

Personalized Medicine
Basic
Learn by recognizing relevant patterns
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology
RNA Edits
Heteroallelic SNVs
Protein-Downregulated (HRV vs Healthy)
Protein-Upregulated (HRV vs Healthy)
RNA-Downregulated (HRV vs Healthy)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

Optimal study design including bioinformatics-driven PM?



Testing precision-medicine strategies





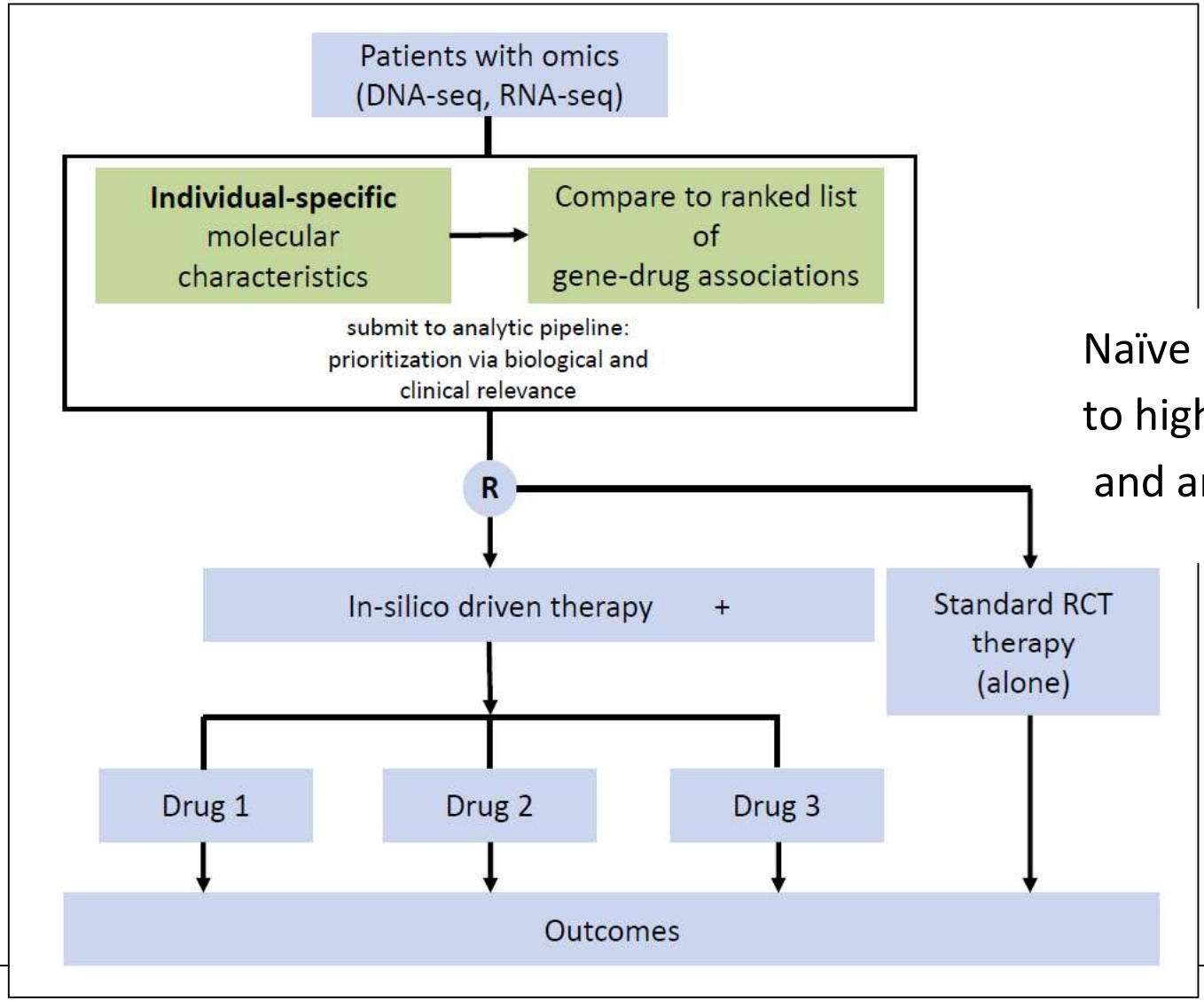
(Biankin et al. 2015)

CTs in view of personalized medicine – where are we?

- **Basket CTs**: multiple diseases with the same genetic mutation(s), randomized treatment allocation
- **Umbrella CTs**: 1 "disease", different genetic mutations which define sub-cohorts, each receiving randomized treatment regimen
- Added complexities:
  - highly multi-dimensional profiles are expected to lead to very small cohorts
  - cellular heterogeneity - assign based on the mutation detected in the higher percentage of cancer cells?

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

CTs in view of personalized medicine – where are we going?

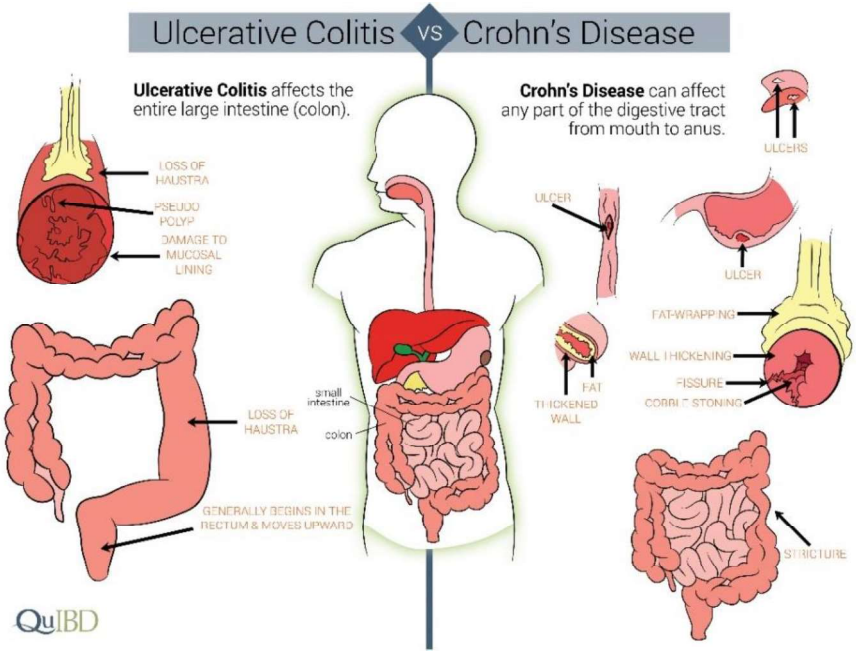


Naïve scheme
to highlight design
and analysis issues

Application of IPCAPS to CD

Inflammatory Bowel Disease (IBD)

- IBD involves chronic inflammation of all or part of the digestive tract.
- Commonly, gastroscopy and colonoscopy are used to diagnose IBD to check for inflammation.
- There are two main forms of inflammatory bowel disease: Crohn's Disease (CD) and ulcerative colitis (UC)
- IBD affects over 2.5 million people of European ancestry with rising prevalence in other populations



Immunochip

- Custom Illumina Infinium chip comprising 196,524 SNPs and small indels selected primarily based on GWAS analysis of 12 autoimmune and inflammatory diseases.
- In total, ~240,000 SNPs were selected for inclusion incl. finemapping and replication results + 25,000 null SNPs; e.g.
  - (0.2cM centered) around 289 established GWAS associations corresponding to 187 distinct loci plus suggestive associations
  - all SNPs and short indels in these regions from the 1000 Genomes Project (CEU samples)
  - variants discovered in resequencing experiments conducted by groups collaborating in the chip design/ replication study results

GWAS and Immunochip

- Meta-analysis of the Immunochip AND GWAS data identified 193 statistically independent signals of association at genome-wide significance ($p < 5\times10-8$) in at least one of CD, UC, IBD).
- Signals referring to the same functional unit were merged, leading to into 163 regions
- Strong evidence of association to the major histocompatibility complex (MHC). This region encodes a large number of immunological candidates, including the antigen-presenting classical HLA molecules → HLA heterogeneity between CD and UC

(Goyette et al. 2015)

LETTER

doi:10.1038/nature11582

Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease

163 loci in 2012

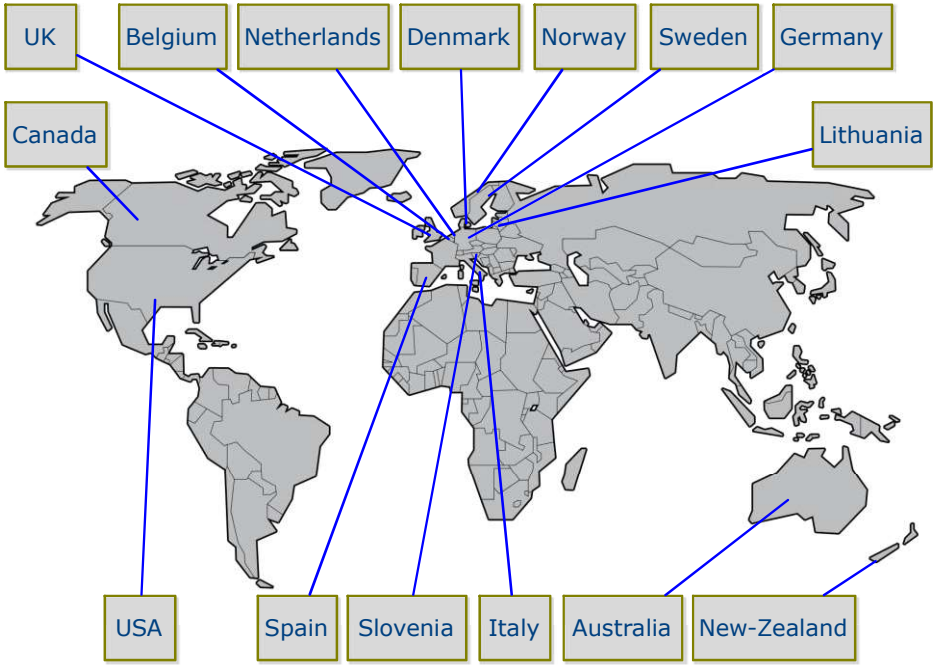
A list of authors and their affiliations appears at the end of the paper.

Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations

Jimmy Z Liu[1,25], Suzanne van Sommeren[2,3,25], Hailiang Huang[4], Siew C Ng[5], Rudi Alberts[2], Atsushi Takahashi[6], Stephan Ripke[4], James C Lee[7], Luke Jostins[8], Tejas Shah[1], Shifteh Abedian[9], Jae Hee Cheon[10], Judy Cho[11], Naser E Daryani[12], Lude Franke[3], Yuta Fuyuno[13], Ailsa Hart[14], Ramesh C Juyal[15], Garima Juyal[16], Won Ho Kim[10], Andrew P Morris[17], Hossein Poustchi[9], William G Newman[18], Vandana Midha[19], Timothy R Orchard[20], Homayon Vahedi[9], Ajit Sood[19], Joseph J Y Sung[5], Reza Malekzadeh[9], Harm-Jan Westra[3], Keiko Yamazaki[13], Suk-Kyun Yang[21], International Multiple Sclerosis Genetics Consortium[22], International IBD Genetics Consortium[22], Jeffrey C Barrett[1], Andre Franke[23], Behrooz Z Alizadeh[24], Miles Parkes[7], Thelma B K[16], Mark J Daly[4], Michiaki Kubo[13,26], Carl A Anderson[1,26] & Rinse K Weersma[2,26]

38 loci in 2015

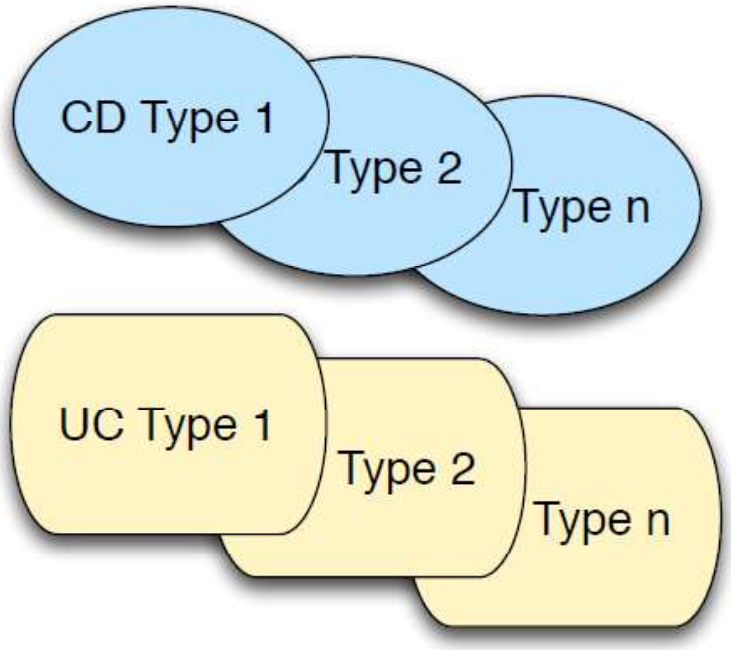
Geographic distribution of samples



Countries	CD	Control
UK	3,885	4,293
Belgium	2,545	1,614
USA	2,489	757
Germany	1,639	3,865
Italy	1,256	479
Netherlands	1,201	0
Australia	867	530
Canada	828	379
New-Zealand	698	477
Sweden	693	357
Spain	277	289
Slovenia	172	0
Norway	140	318
Lithuania	125	279
Denmark	67	90
Total	16,882	13,727

Research question

Aim



Patient sub-structure

Awareness

Underlying population structure

Disease heterogeneity – multi-source data

	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)

Basic analysis steps (~150,000 SNPs → 20,000 SNPs on ~7000 cases and ~7000 controls retained; QC step 1 on cases/controls separately; LD pruning at r^2=0.2 ~ PC computations)

- **Step 1**: Split the patient data into discovery and replication sets; use controls to validate your PS adjustment strategy
- **Step 2**: Perform IPCAPS clustering on discovery and replication data, adjusted for confounding by population structure to determine the number of clusters or the settings of your parameters (we did not do the latter because our settings were driven by simulations)
- **Step 3**: Perform IPCAPS on all available data (discovery and replication data pooled)
- **Step 4**: Determine and interpret cluster discriminants
- **Step 5**: Characterize your clusters