Towards Molecular Reclassification of Disease

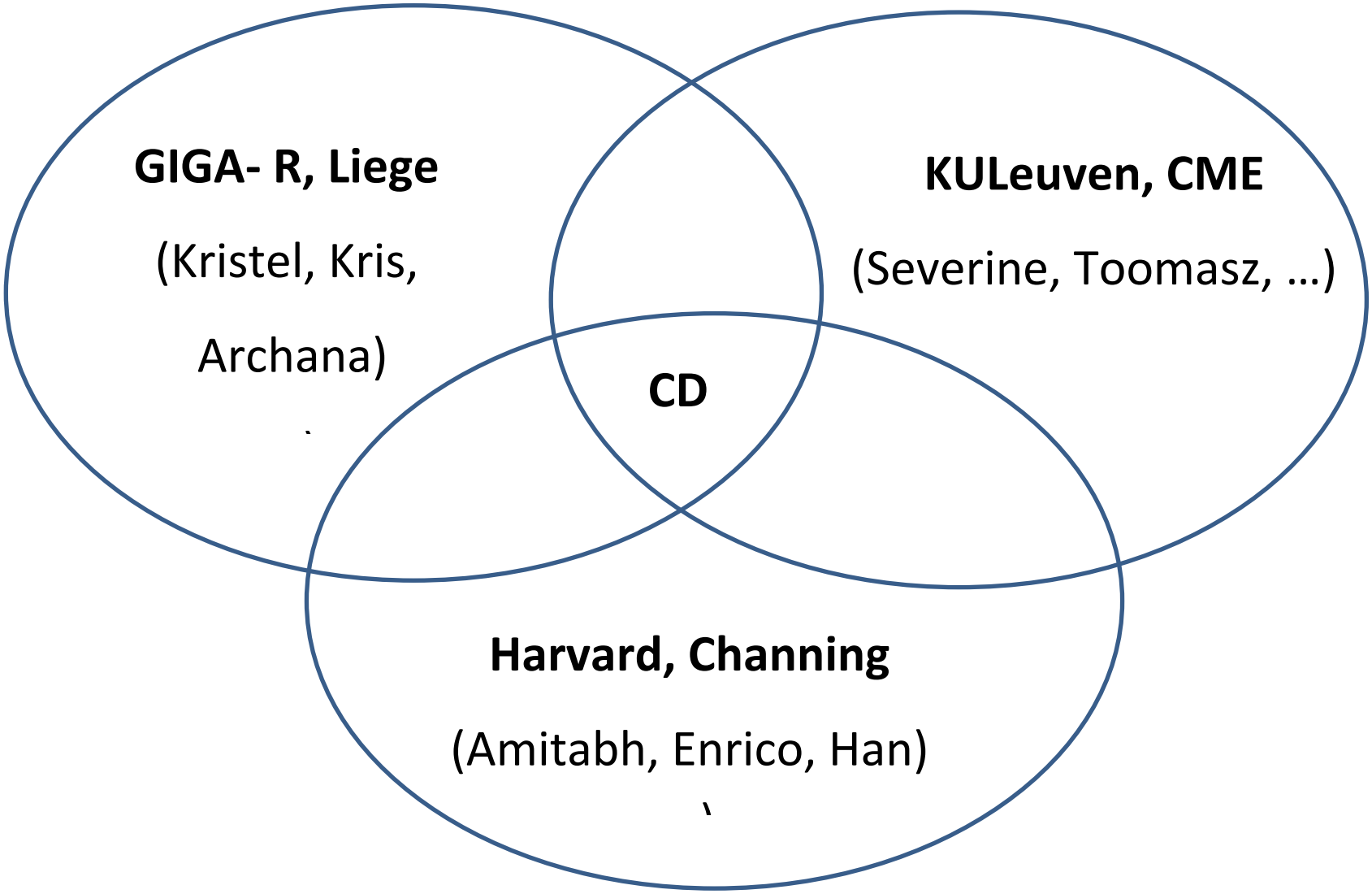
Kristel Van Steen, PhD[2] (*)

kristel.vansteen@uliege.be

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

Systems Medicine Lab, KU Leuven, Belgium

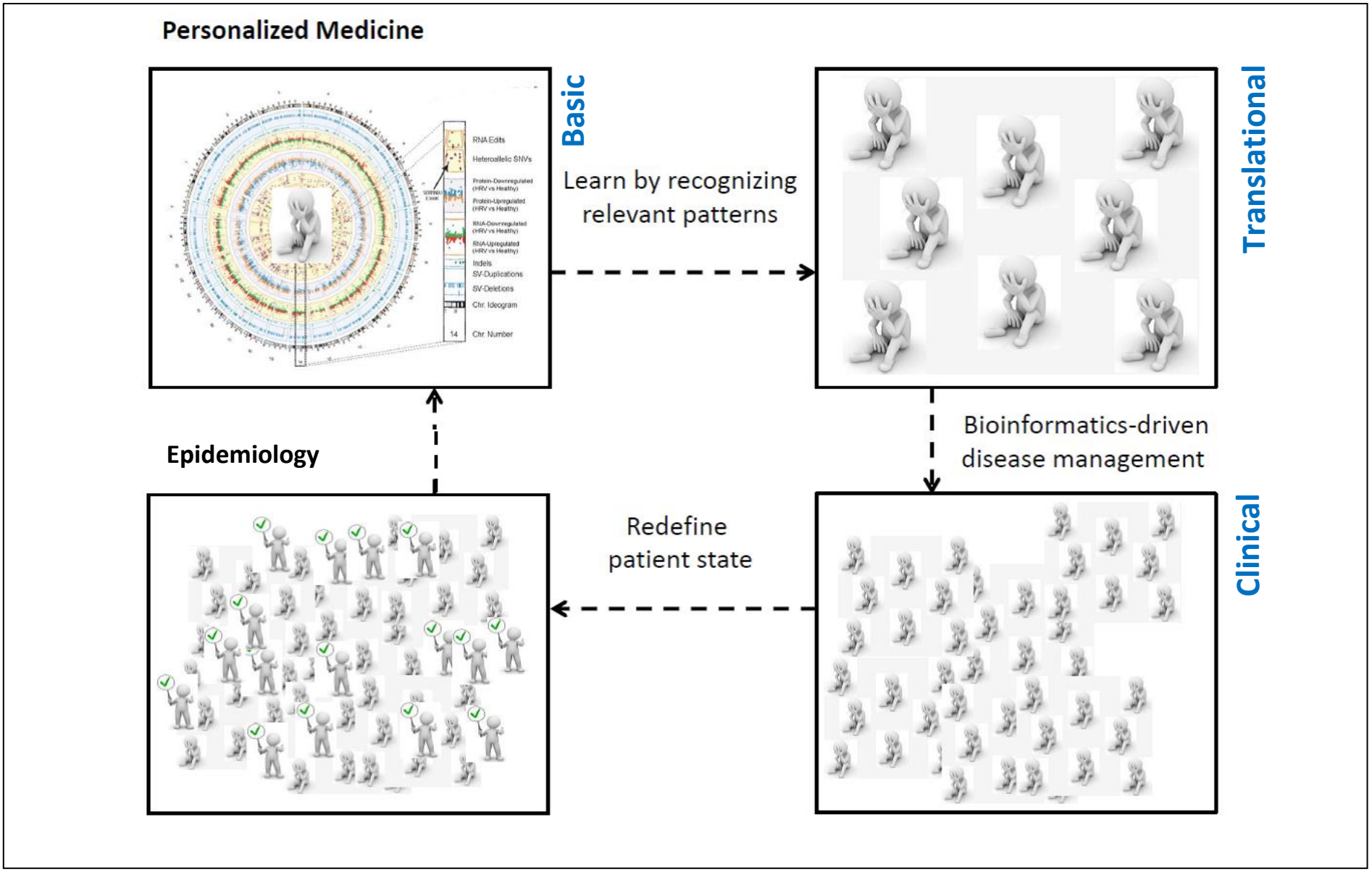
COLLABORATIVE WORK

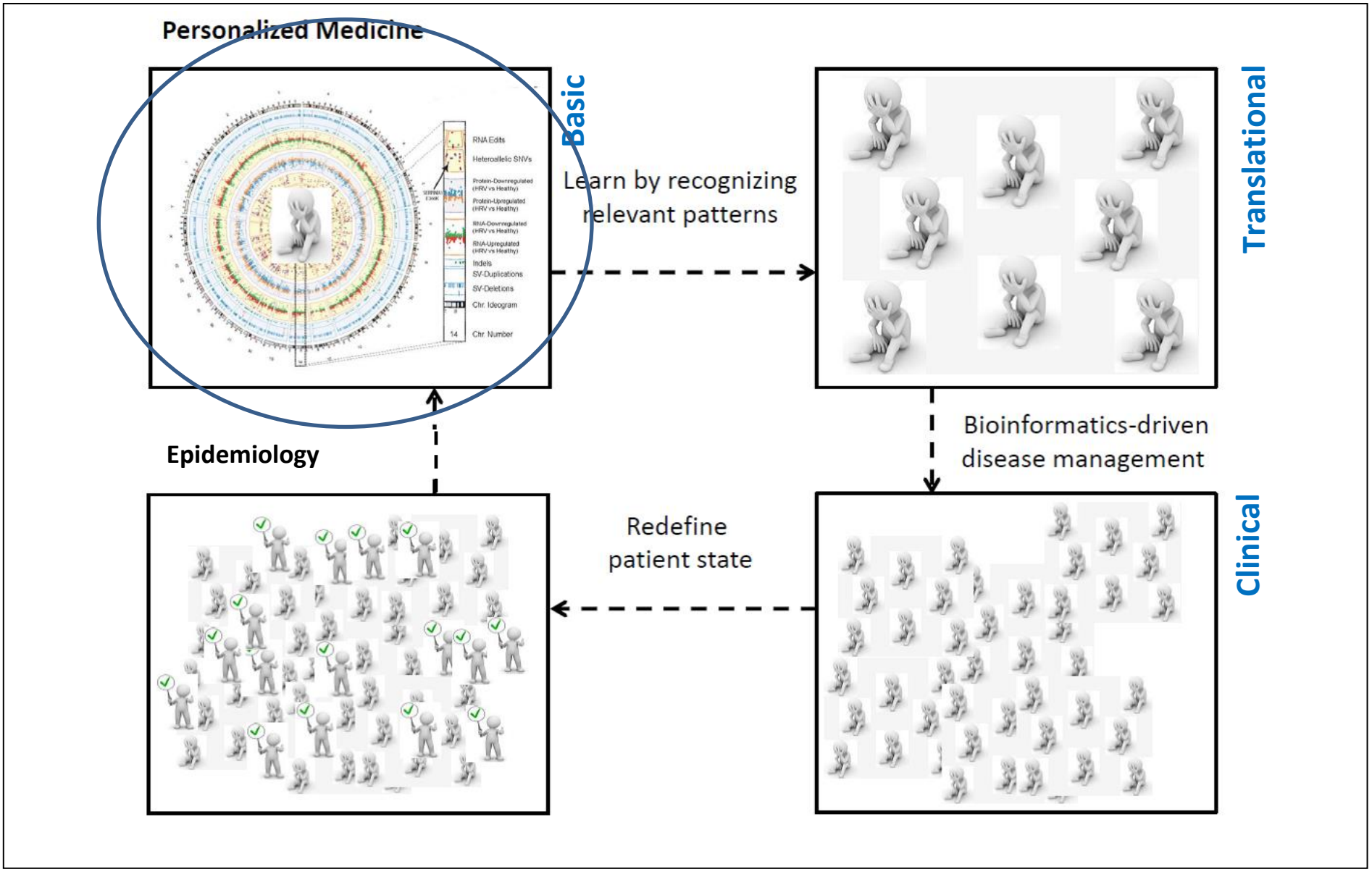


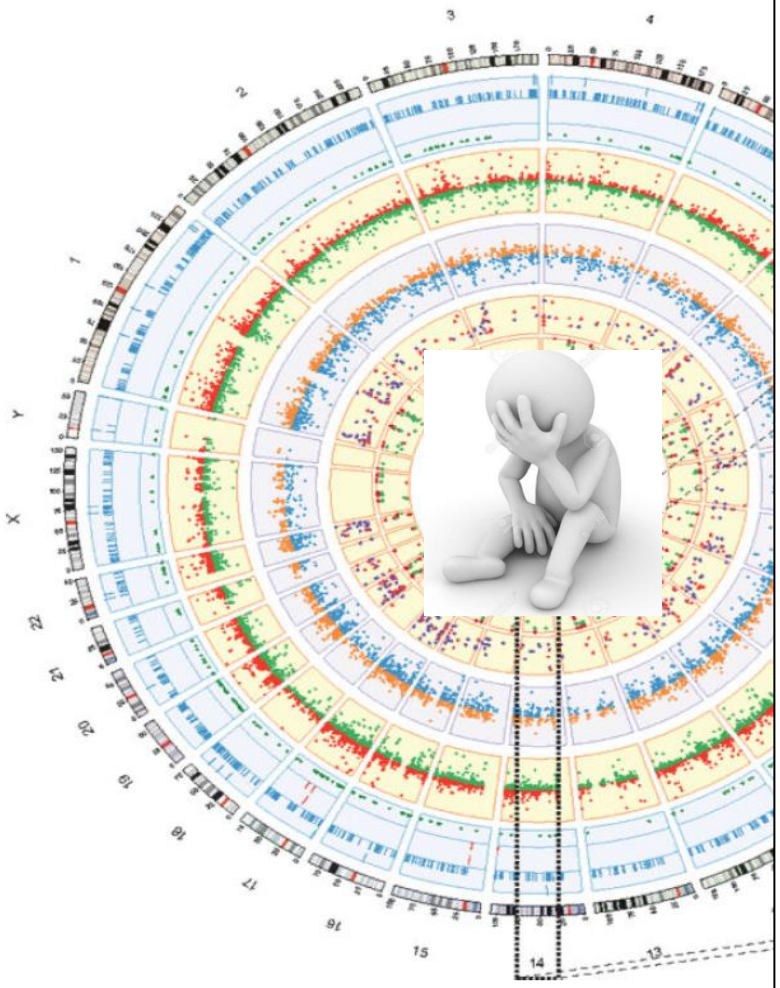
OUTLINE

- **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
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
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology


Personalized Medicine
Basic
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
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology



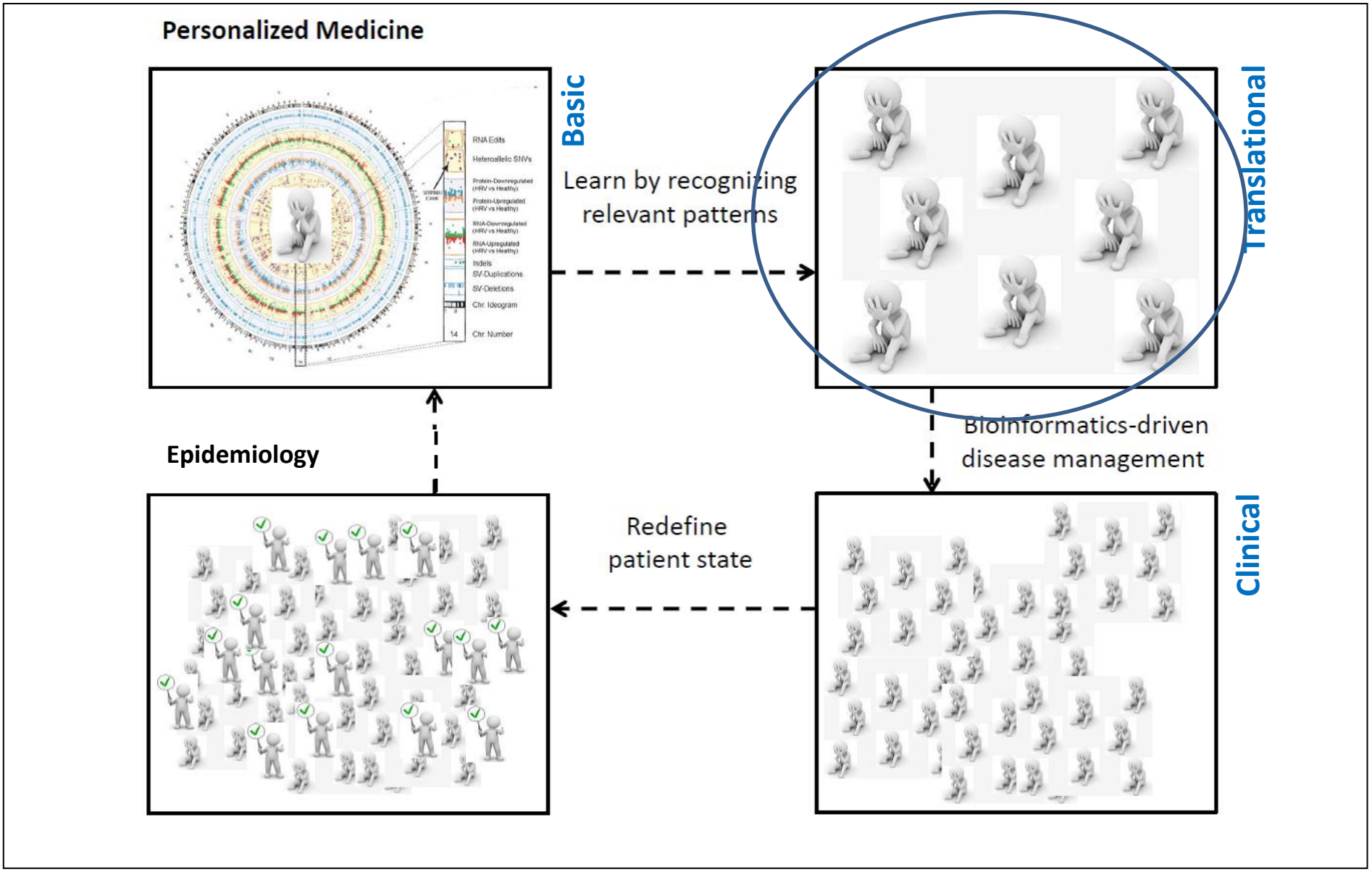
Do you think that omics profiling will be routinely used in the clinic in future?

"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."

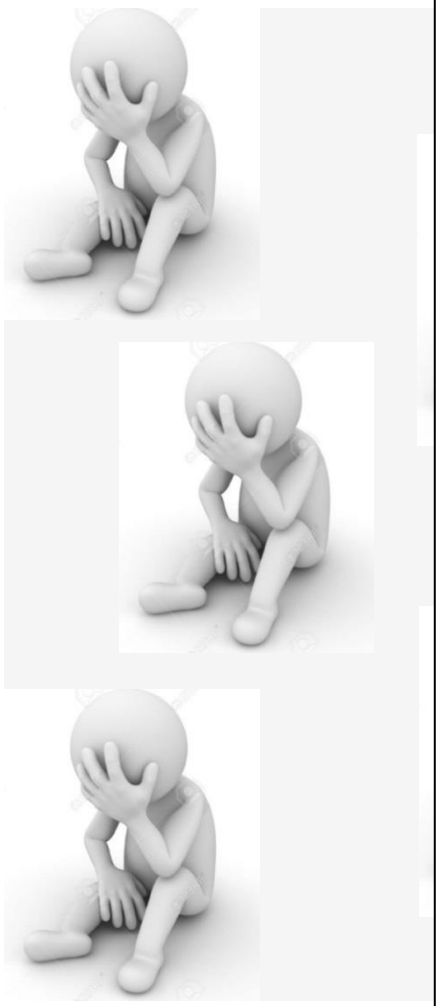
(Snyder 2014)

Redundancy – Informativity

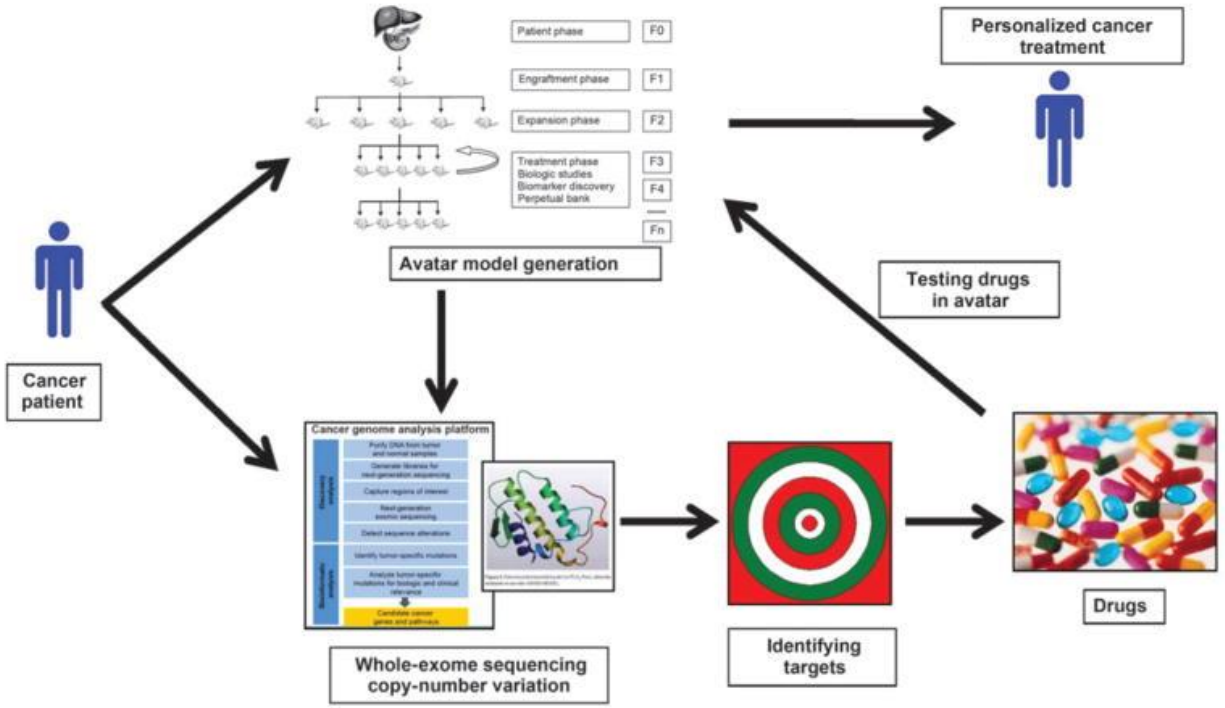
Missingness


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

Bionformatics-driven treatment assignment tool

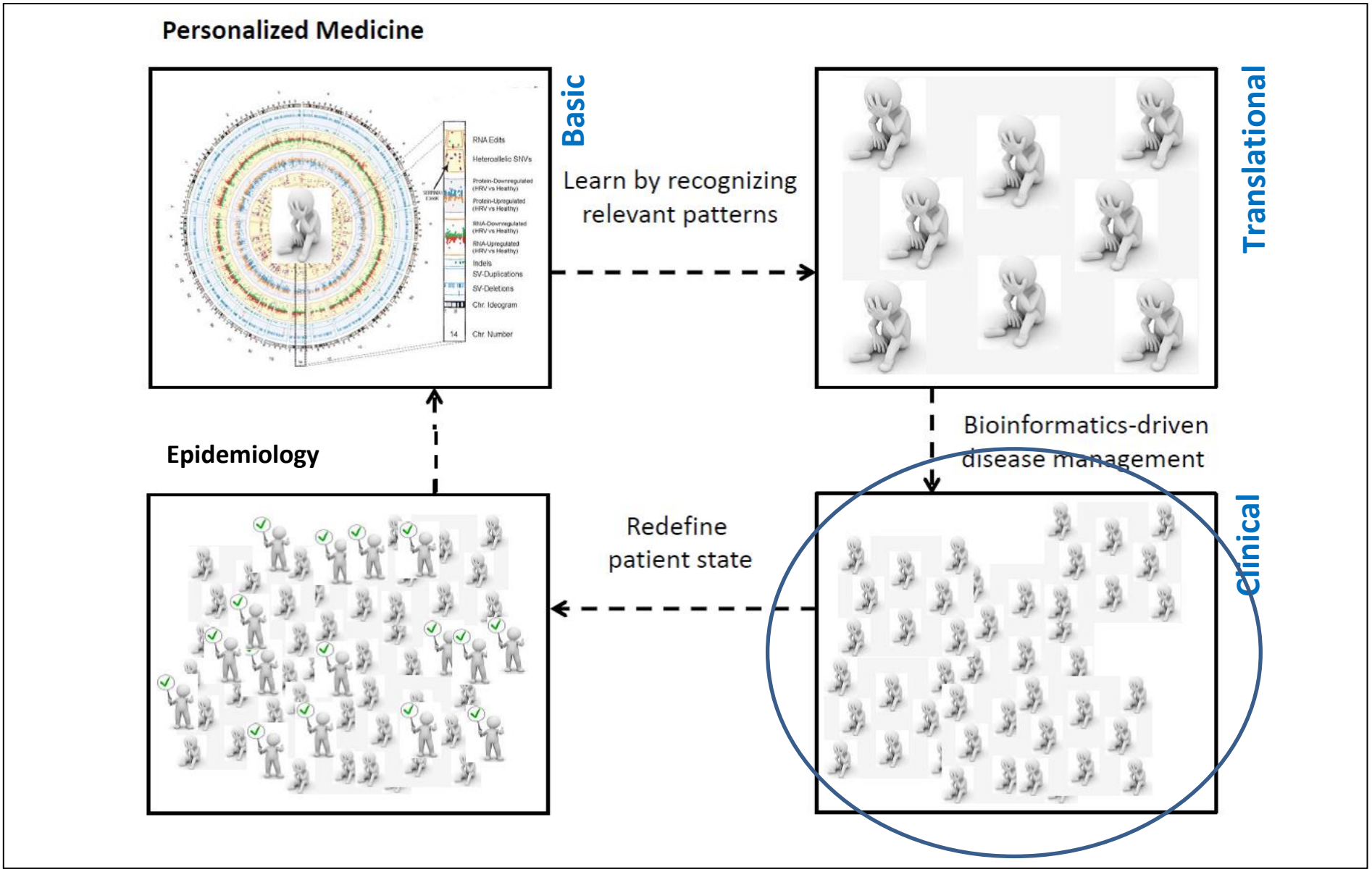


Integrating sequencing and avatar mouse models



Problems …

(Garralda et al. 2014)


Personalized Medicine
Basic
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
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology

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]

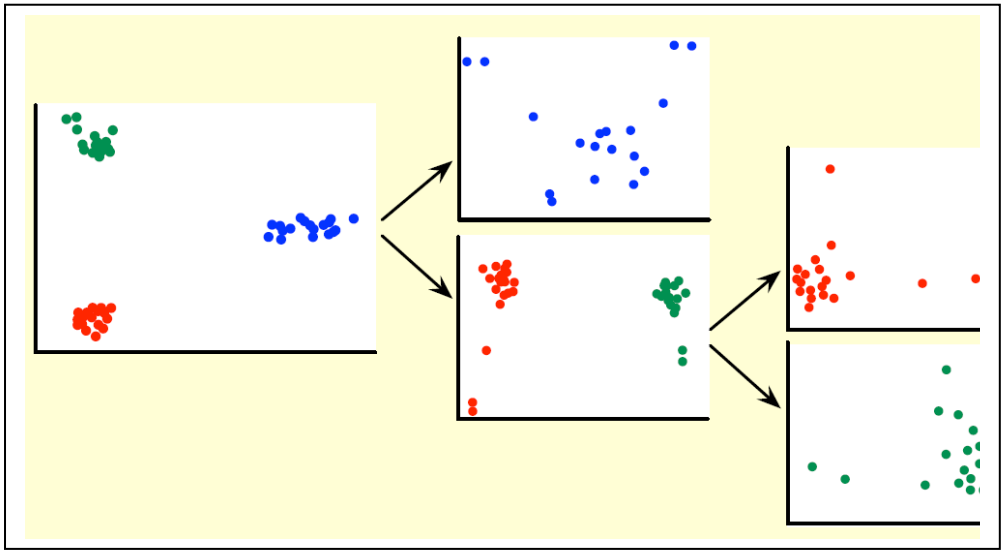
1 UMR843, INSERM, Paris, France, **2** Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, **3** UMR843, Institut National de la Sante et de la recherche Medicale, Paris, France, **4** Service de Gastroentérologie Pédiatrique, Hôpital Robert Debré, APHP, Paris, France, **5** CRC-CRB,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

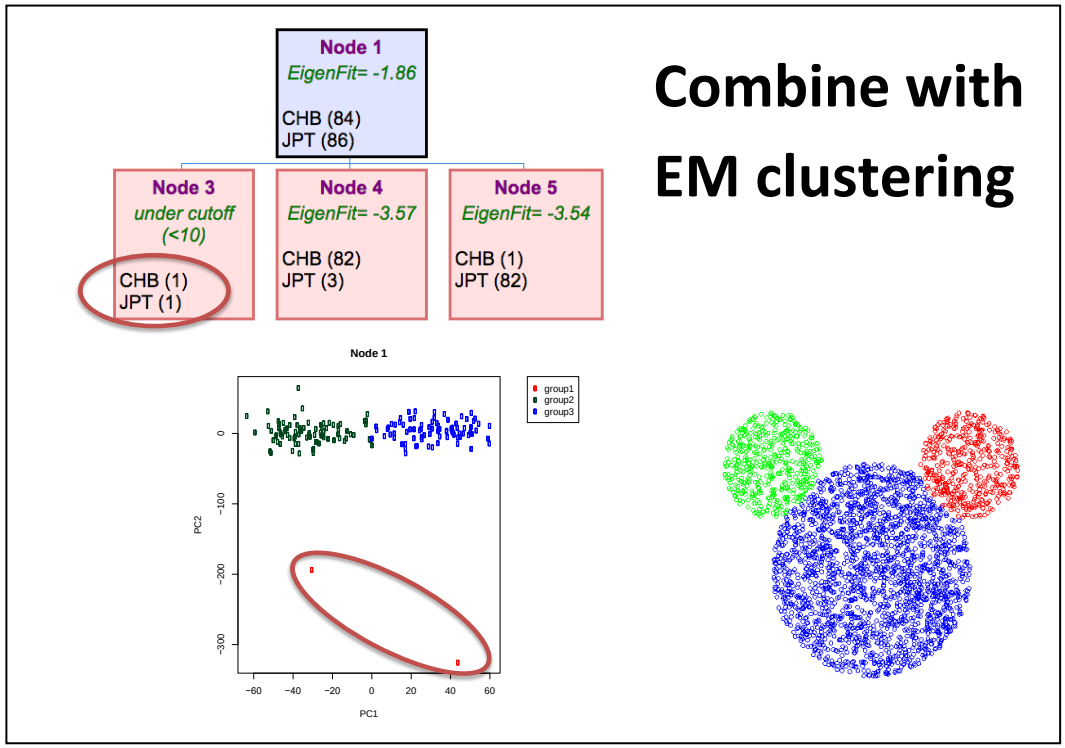
- Template: ipPCA (Intarapanich et al. 2009)
  - Performs PCA with genotype data (similar to EIGENSTRAT)
  - If substructure exists in PC space individuals are assigned to one of two clusters (2-means algorithm / fuzzy c-means)
  - Iteratively performs test for substructure and clustering on nested datasets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



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

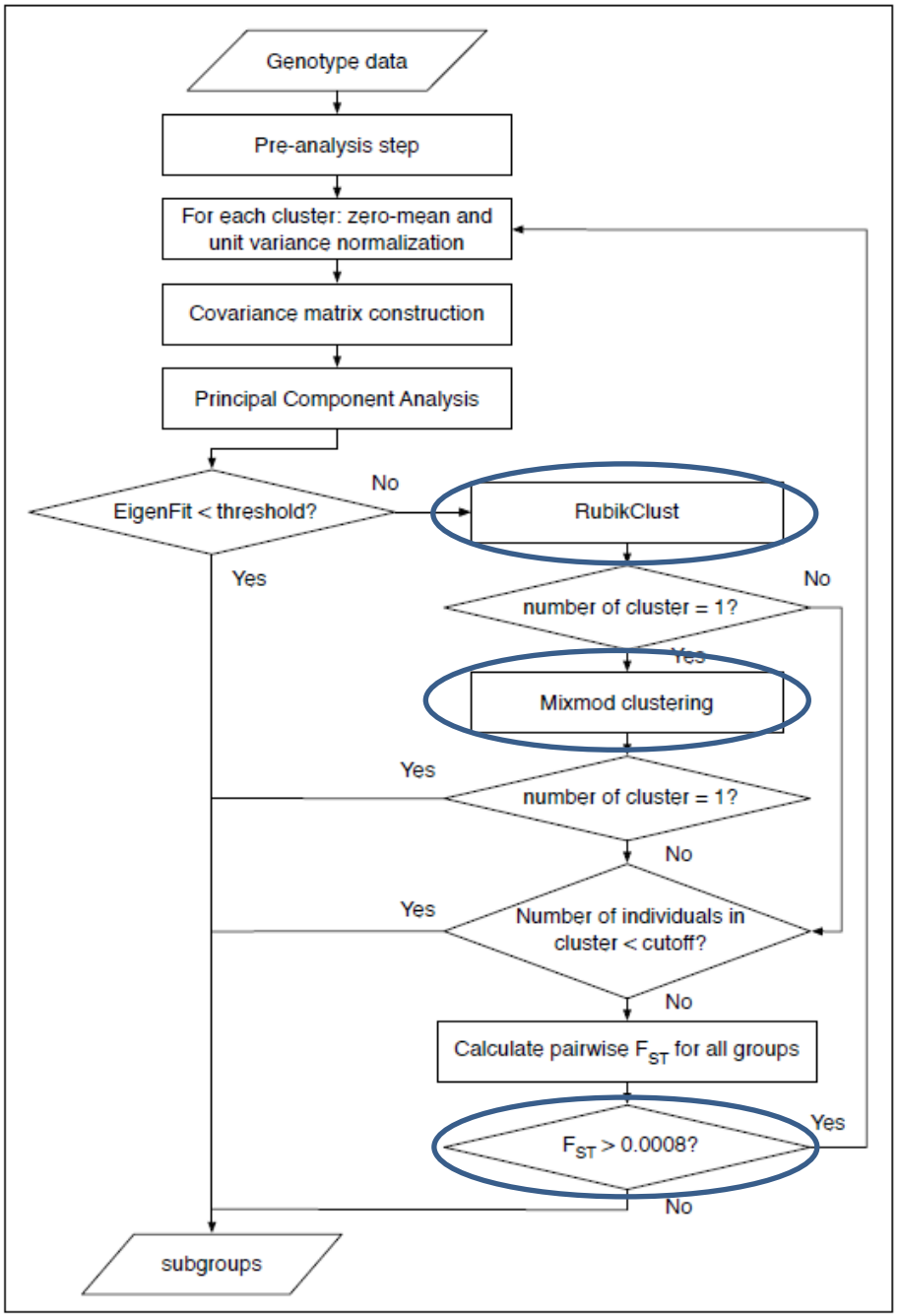
- ipPCA
  - Pros: outperformed others (STRUCTURE – 2000) in achieving higher accuracy for highly structured populations
  - Cons: binary splitting; outlier sensitive; difficult to integrate mixed data types
- Competitors:
  - SHIPS (2012) – divisive fine-scale structure detection; computational efficiency; 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

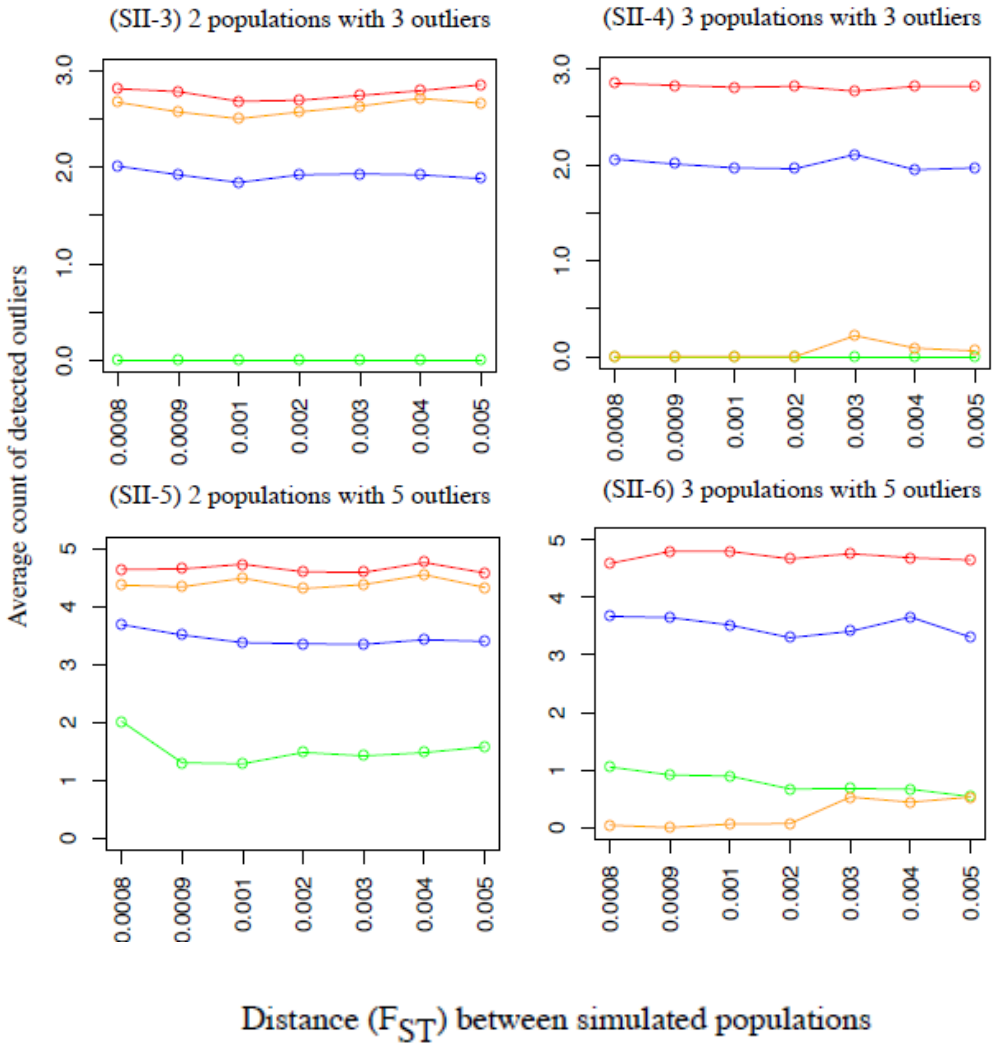


Combine with EM clustering

(Chaichoompu – thesis defense Oct 2017)



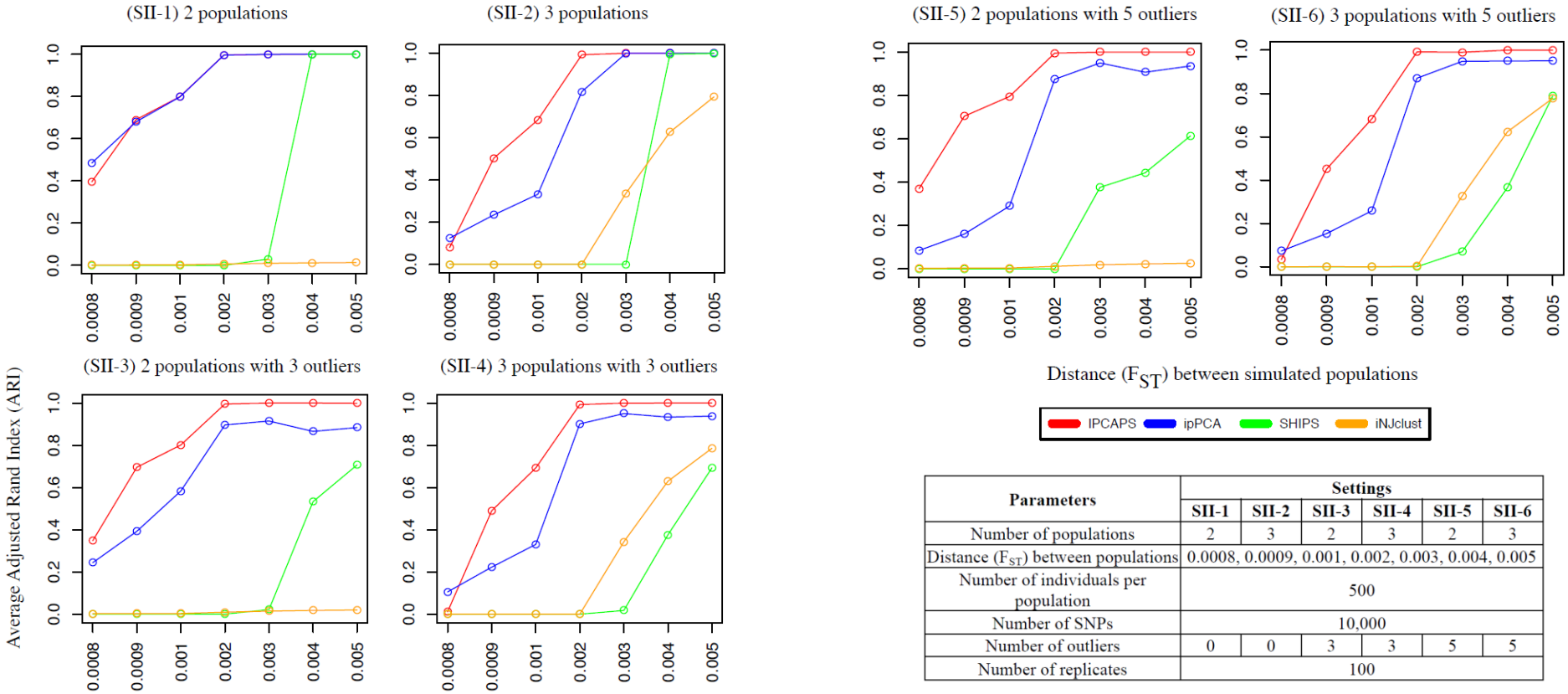
Performance of IPCAPS as outlier detection tool



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					

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					

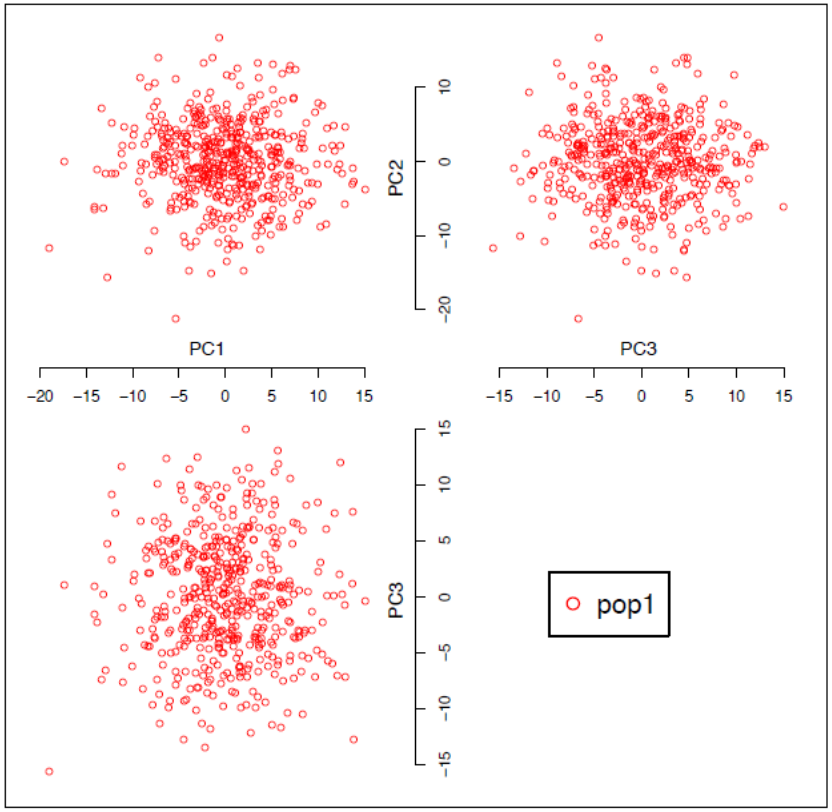
(Chaichoompu – thesis defense Oct 2017)

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)

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*]

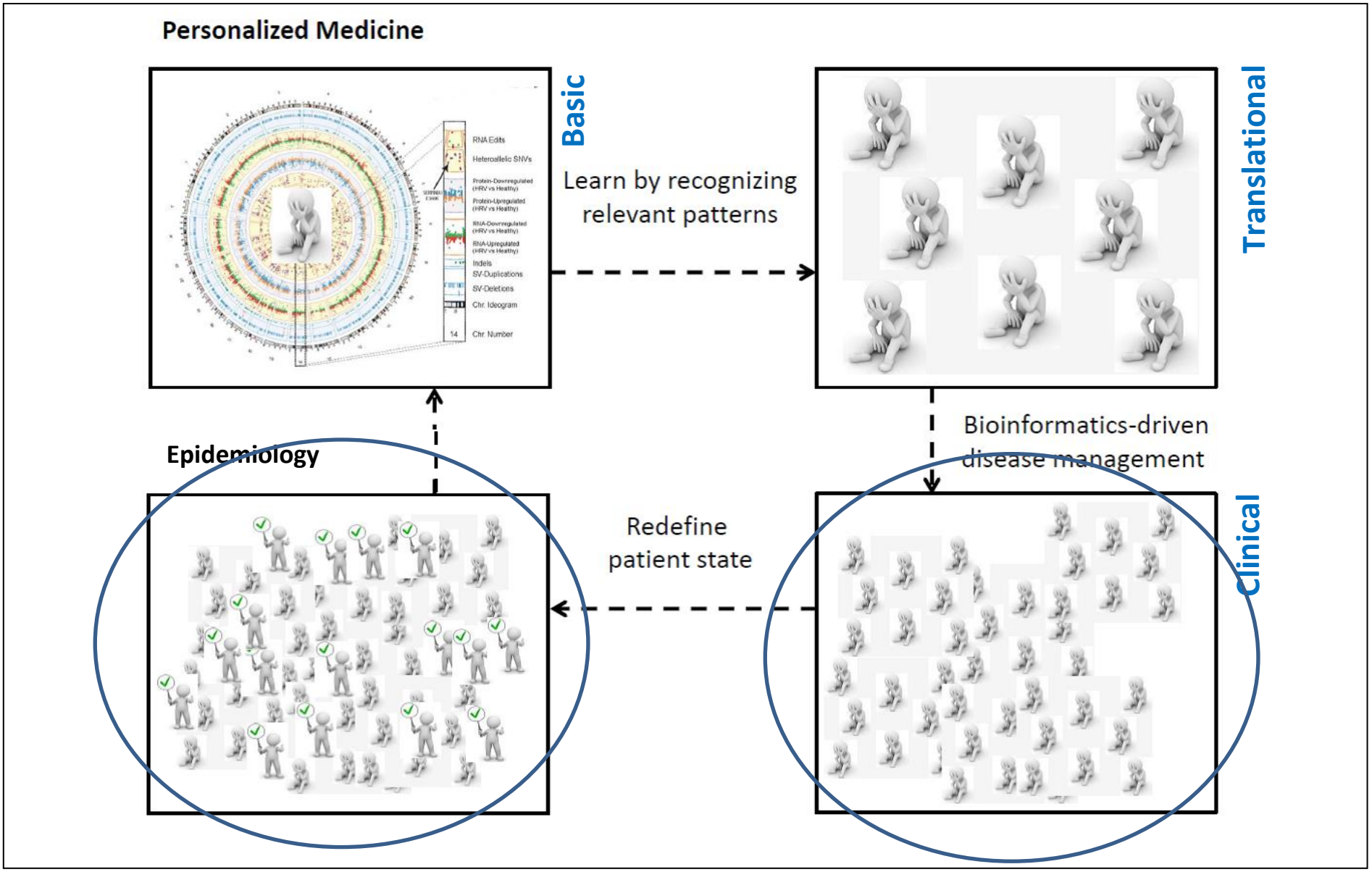
Abstract

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.

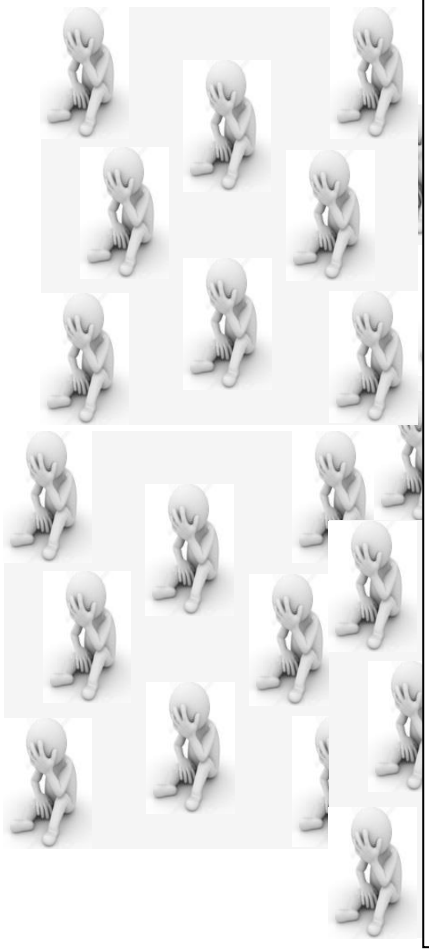
Results: This work presents an R package called IPCAPS, which uses SNP information for resolving possibly fine-scale population structure. The IPCAPS routines are built on the iterative pruning Principal Component Analysis (ipPCA) framework that systematically assigns individuals to genetically similar 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

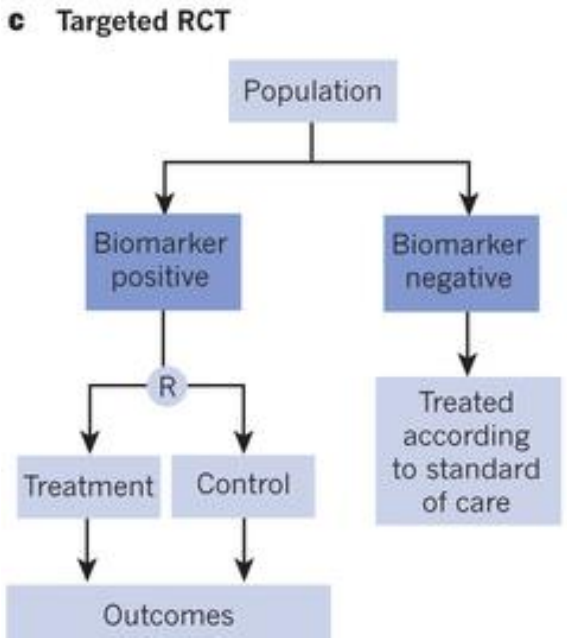
Keywords: Fine-scale structure, Iterative pruning, Population clustering, Population genetics, Outlier detection

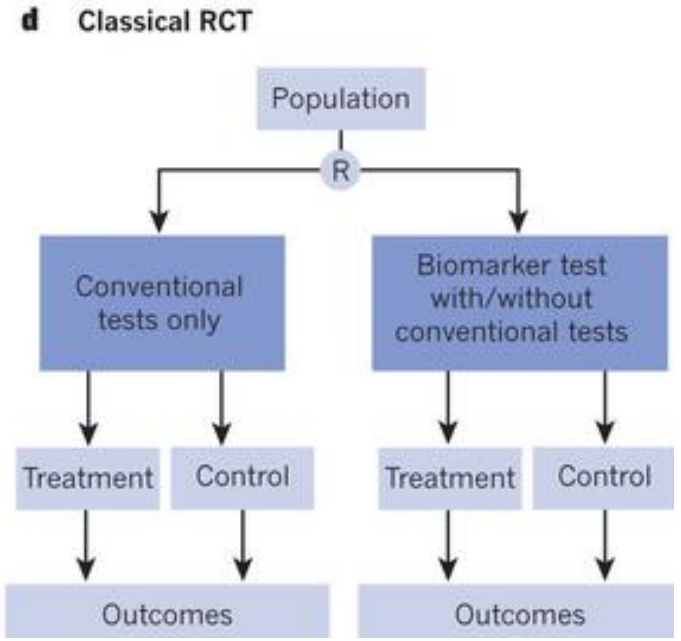

Personalized Medicine
Basic
RNA Edits
Heteroallelic SNVs
Protein-Downregulated (HRV vs Healthy)
Protein-Upregulated (HRV vs Healthy)
RNA-Downregulated (HRV vs Healthy)
RNA-Upregulated (HRV vs Healthy)
Indels
SV-Duplications
SV-Deletions
Chr. Ideogram
Chr. Number
Learn by recognizing relevant patterns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

Optimal study design including bioinformatics-driven PM?



Testing precision-medicine strategies





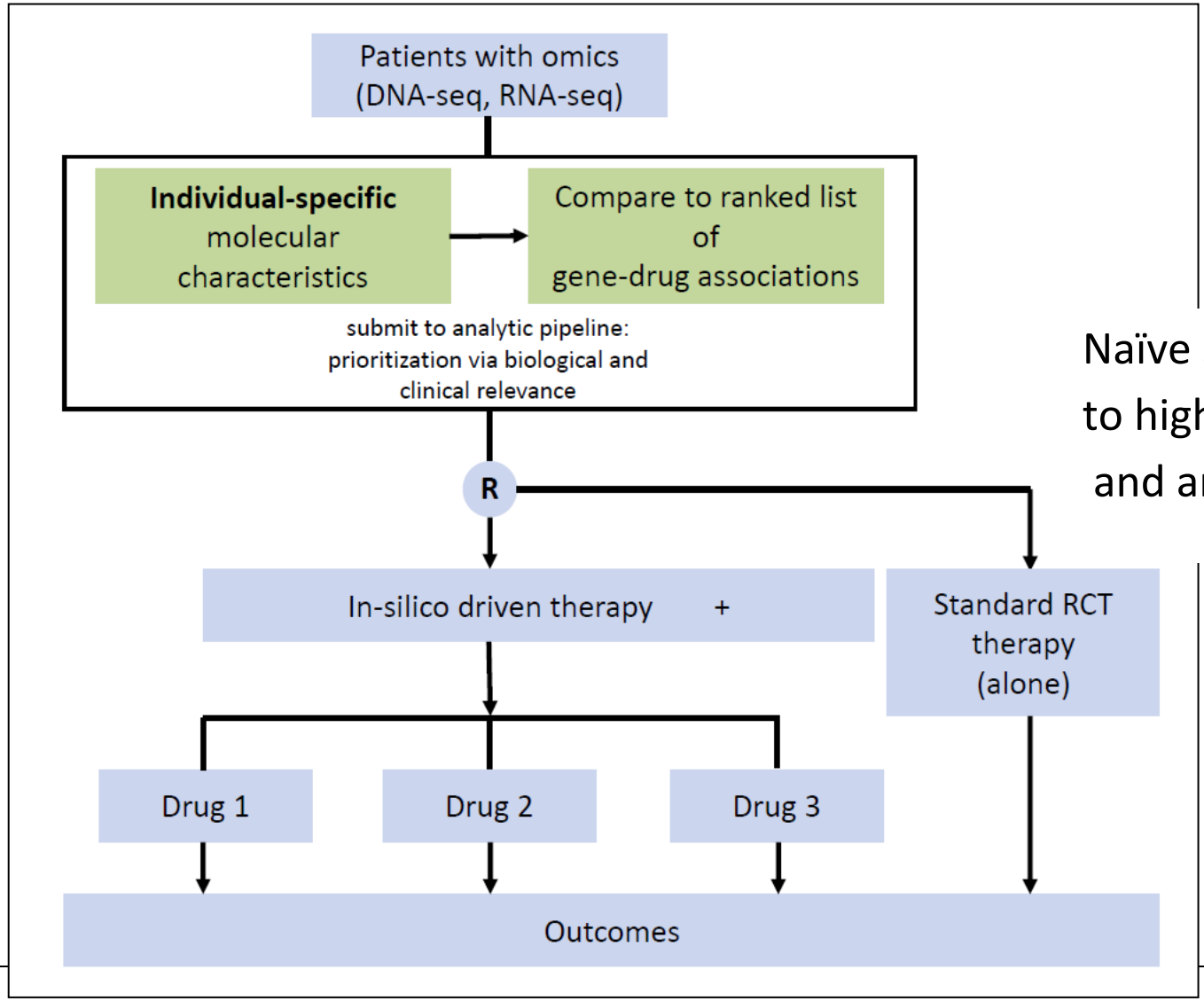
(Biankin et al. 2015)

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?

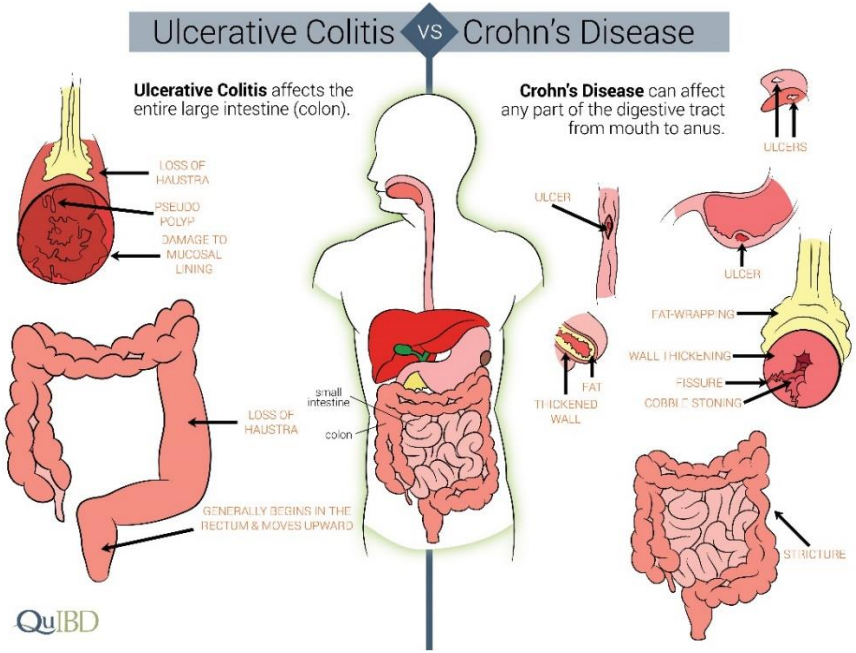


Naïve scheme
to highlight design
and analysis issues

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

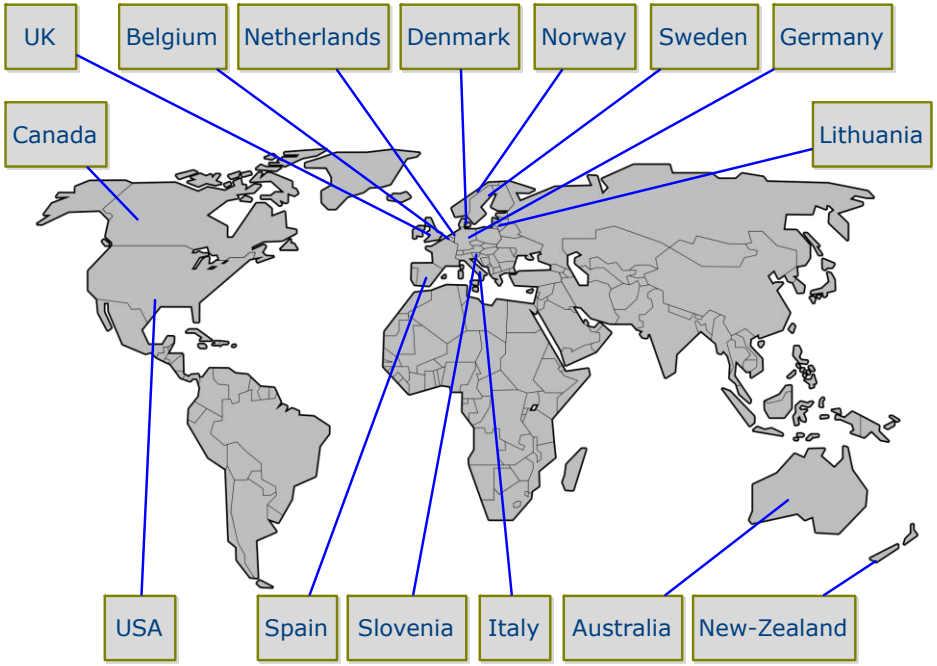
A list of authors and their affiliations appears at the end of the paper.

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

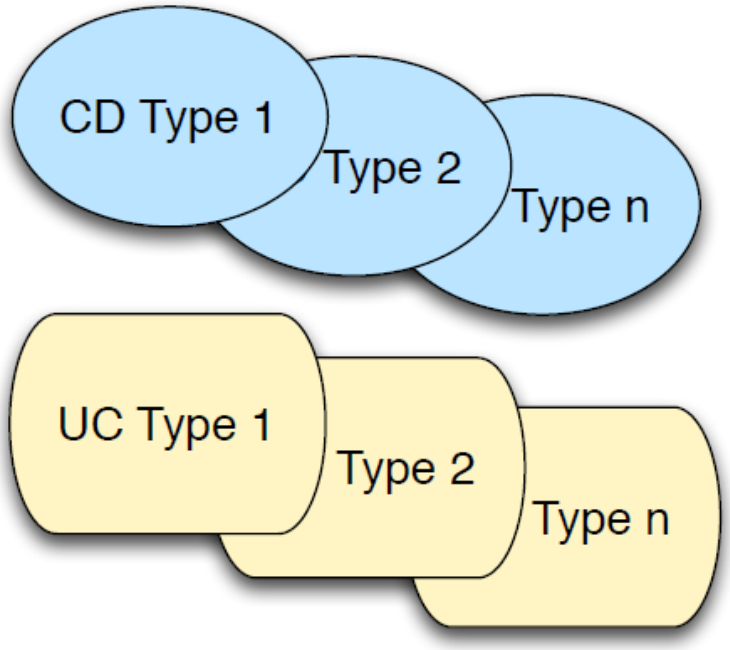
Geographic distribution of samples



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

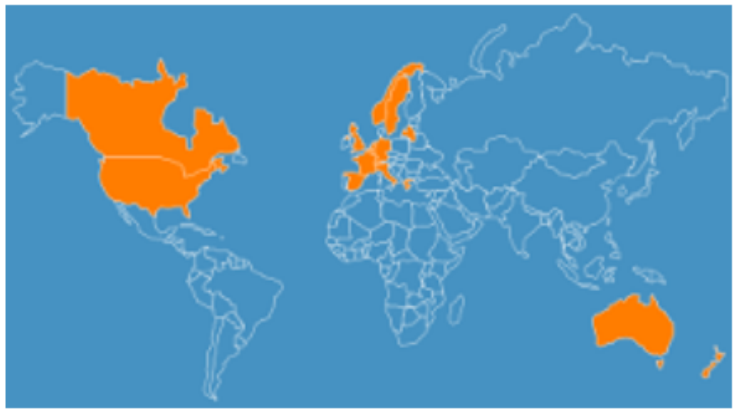
Research question

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)

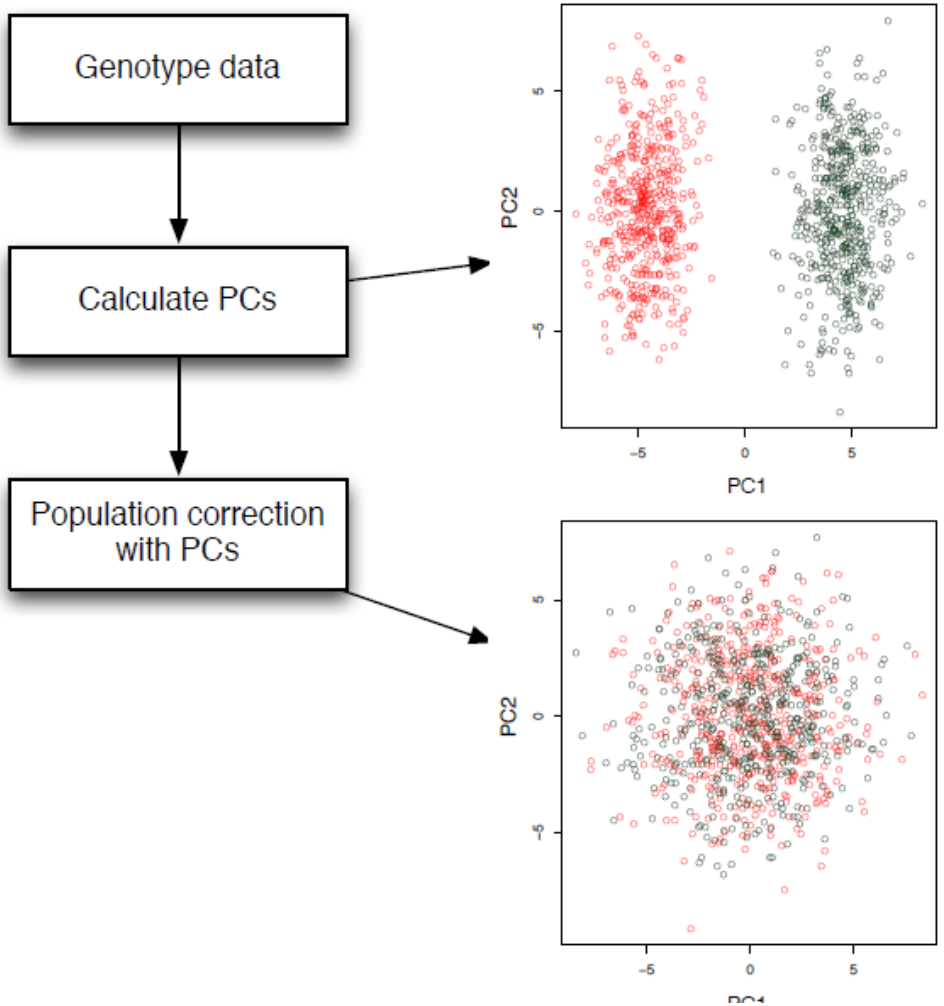
- **Step 1**: Split the patient data into discovery and replication sets; use controls to validate your PS adjustment strategy
- **Step 2**: Perform IPCAPS clustering on discovery and replication data, adjusted for confounding by population structure to determine the number of clusters or the settings 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

Step 1: How to capture population structure

Dataset	**Uncorrected** SNPs (I)		**Corrected with PCs from our curated SNPs** (pruned at r^2 of 0.2) (II)				**Corrected with PCs from the IIBDGC SNPs** (III)			
	Dis.	Rep.	5PCs		10PCs		5PCs		10PCs	
			Dis	Rep.	Dis.	Rep.	Dis.	Rep.	Dis.	Rep.
CON	5	4	3	7	1	1	3	9	3	7

Set	Uncorrected CON		CON			
	Dis.	Rep.	Dis.	Rep		
1	5	4	1	1		
2	3	5	1	1		
3	5	5	1	1		
4	5	5	1	1		
5	5	5	1	1		
6	5	4	1	1		
7	6	5	1	1		
8	6	4	1	1		
9	4	4	1	1		
10	4	5	1	1		
Aver.	4.8	4.6	1.0	1.0		

Input features adjusted for general population structure



PCs were calculated from all available data (2 populations), referred to as “Pooled PCs”

Population Correction: PCs regressed out from original SNPs.
PCs were calculated from adjusted SNPs.

Step 2: Assessing the number of patient clusters

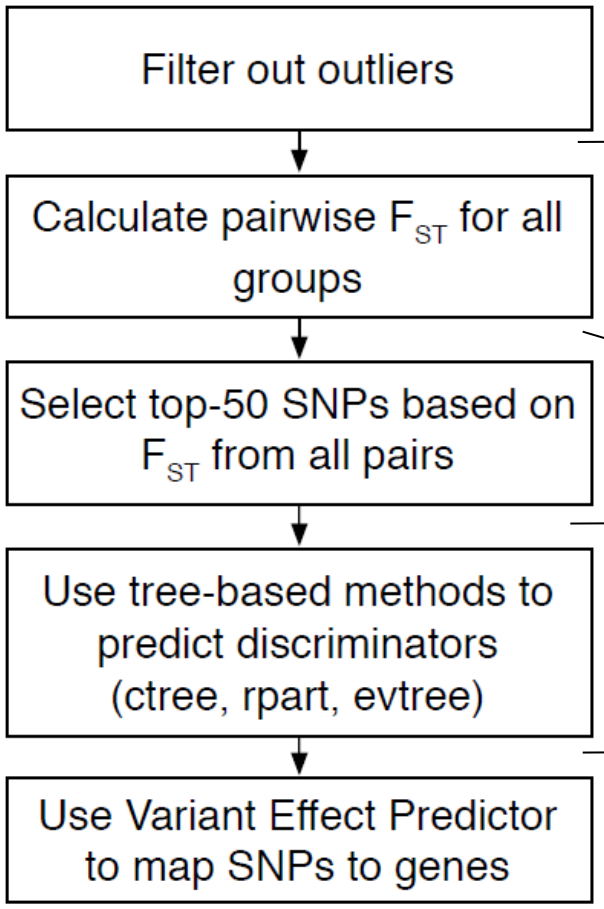
- Country stratified sampling; case/control proportional sampling

t	Uncorrected CON		CON		CD	
	Dis.	Rep.	Dis.	Rep	Dis.	Rep
1	5	4	1	1	3	8
2	3	5	1	1	3	5
3	5	5	1	1	3	3
4	5	5	1	1	3	3
5	5	5	1	1	3	5
6	5	4	1	1	3	3
7	6	5	1	1	3	3
8	6	4	1	1	6	3
9	4	4	1	1	3	8
10	4	5	1	1	6	5
Aver.	4.8	4.6	1.0	1.0	3.6	4.6

- Note that here it does not make sense to assess the stability of the clusterings in terms of individual assignments
- Number of clusters as extra stopping criterium in IPCAPS for pooled data:

  round below min (av. # Dis., av # Rep.)

First indication about "interesting" SNPs (disc. / repl.)



At least 20 individuals in a cluster, otherwise not retained (proportion)

Hudson's method with formulae of Bathia et al. (2013): stabler than Nei, and Weir and Cockerham; no over-estimation for unbalanced samples

Cluster membership as outcome; not to be recommended with large nr of clusters

OK at this stage: final assessments based on SNPnexus (Ullah et al. 2018)

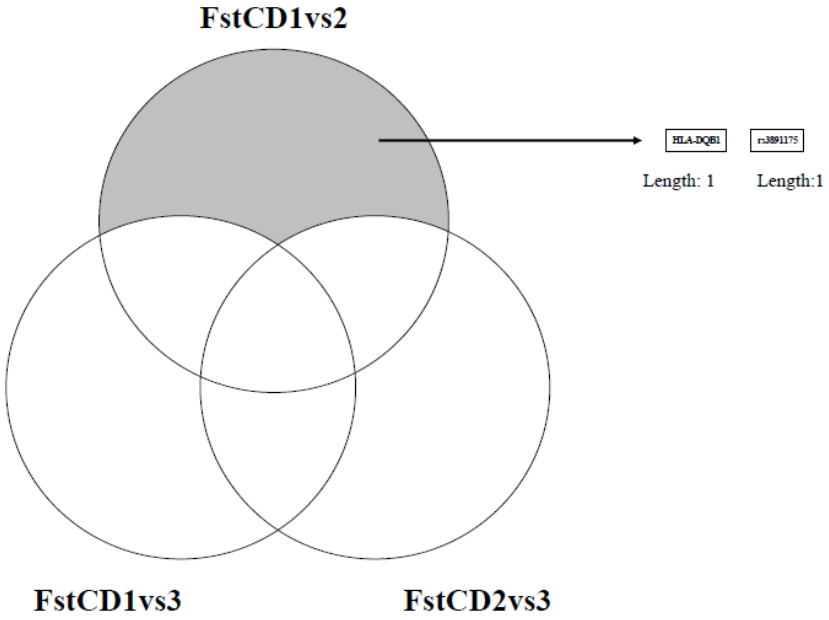
First indication about "interesting" SNPs ("dtree" package in R - 2018)

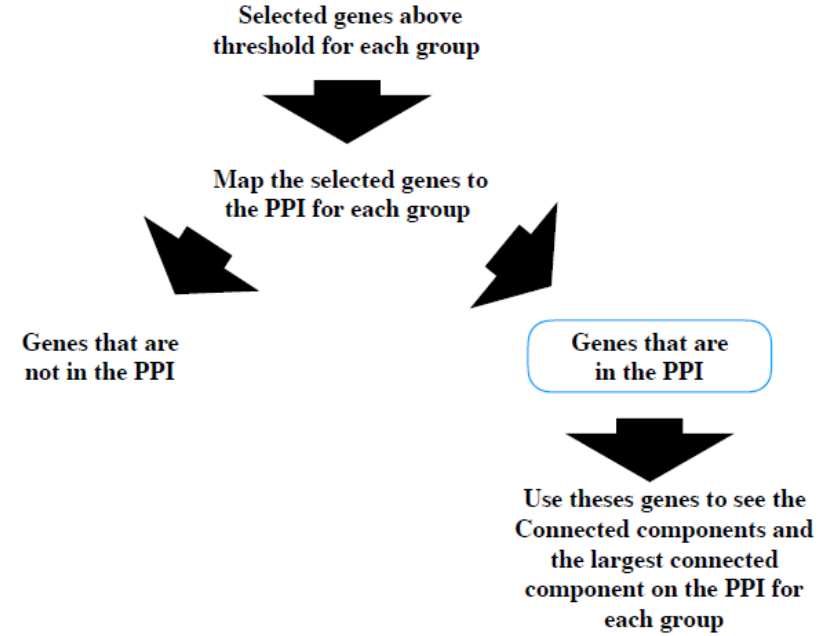
SNPs	Chr	Positions	Associated genes	Additional information	Runs
rs80261410	2	136049426	-	intergenic	9
rs11681014	2	134377531	MGAT5	intron	7
rs200930008	11	18246053	SAA2	splice region, intron	7
rs3749946	6	31481085	-	intergenic	6
rs4833103	4	38813881	-	intergenic	4
rs10280281	7	16365684	ISPD	intron	4
rs6922431	6	31497253	MICB	upstream gene	4
rs12210050	6	475489	-	intergenic	3
rs17796359	18	30401672	-	intergenic	2
rs2621377	6	32795333	-	intergenic	2
rs1982850	9	114238521	COL27A1	intron	2
rs3131063	6	30795979	HCG20	downstream gene	2
rs2284178	6	31464348	HCP5	non coding transcript exon	2
rs2516436	6	31452100	HCP5	intron, non coding transcript	2
rs2516464	6	31448379	HCP5	intron, non coding transcript	2
rs2524076	6	31276053	HLA-C	upstream gene	2
rs68600	6	32935947	HLA-DMB	intron	2
rs443198	6	32222629	NOTCH4	synonymous	2
rs10769905	11	8420169	STK33	intron	2
rs464367	6	33276864	WDR46	downstream gene	2
rs9366778	6	31301396	XXbac-BPG248L24.13	intron, non coding transcript	2
rs2844513	6	31420437	HCP5	intron, non coding transcript	1
rs434841	6	32223264	NOTCH4	intron	1
rs9267845	6	32225921	NOTCH4	upstream gene	1

- SNPs listed in this table appear at least 2 times out of 20 runs (10 discovery and 10 replication sets; non-independent data; different nr of clusters)
- 24 driver SNPs linked to 13 genes
- Only 2 IBD genes: HLA-C (Kulkarni et al. 2013) and MICB (Jostins et al. 2012)

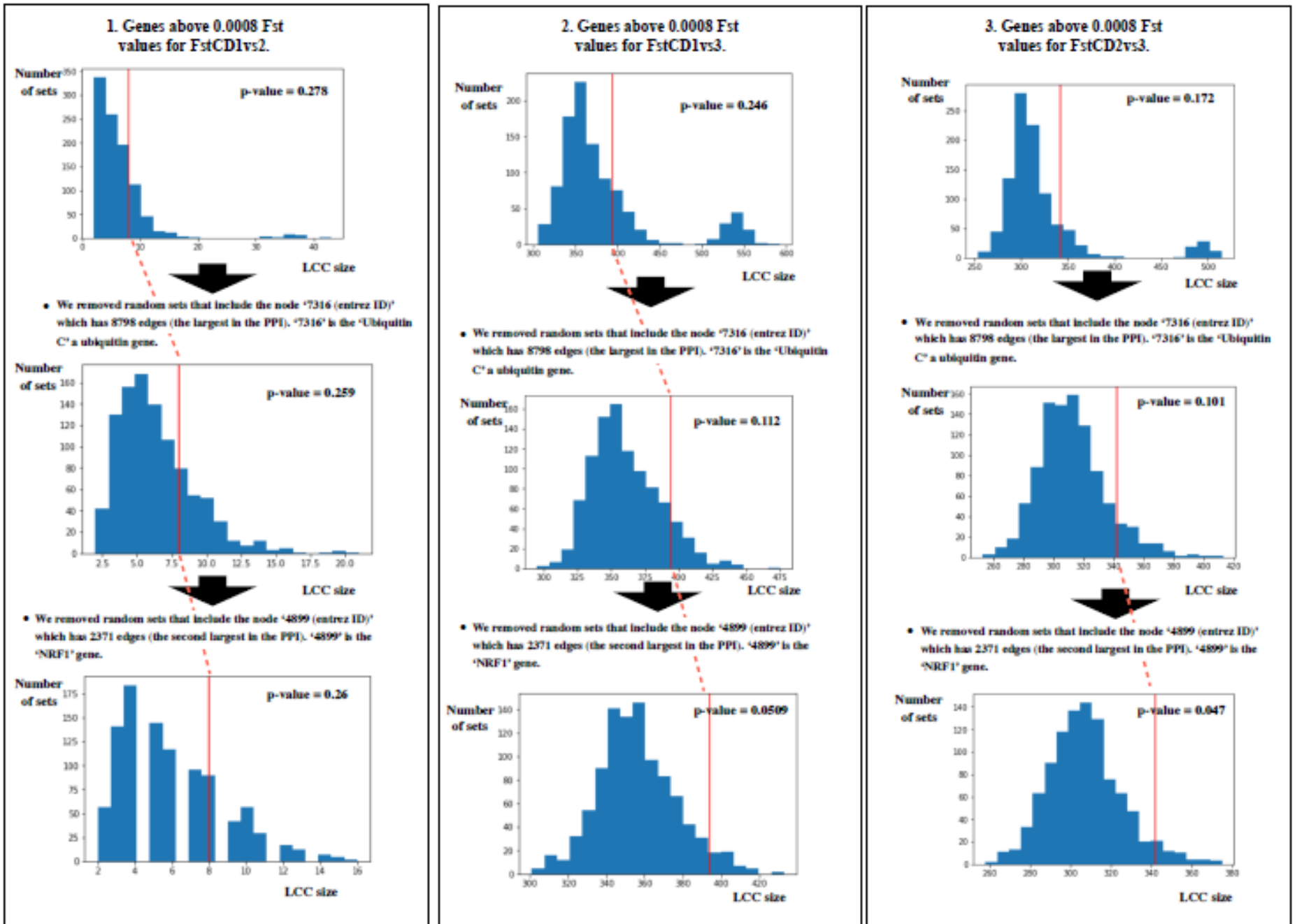
Step 3 + 4: Find discriminators between the 3 IPCAPS clusters

- Focus on pairs of clusters
  - Take SNPs with $F_{ST} \geq 0.008$ (note: not discriminative in the statistical sense)



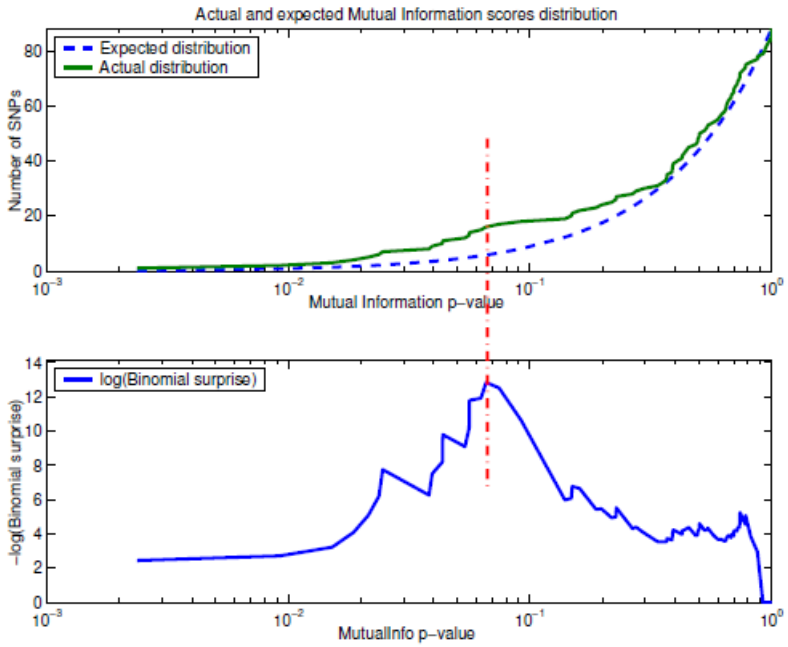


Remove more hubs ? Distance between LCCs ?



- Choices
  - Focus on top (significant) F_{ST} / Association? (cfr recent work Moore)
    - discriminators that are common to all 3 pairwise comparisons? Or pool the discriminators found for each pairwise comparison (see also **Step 2**)
  - What about statistically benchmarking sets of SNPs that jointly associate with a property of interest?
    (Tsalenko et al. 2003)

    → expected to affect LCC significance



- Decision 1:
  - Keep / complete runs for F_{ST} bounded by 0.008
    - Use information on bump removal and the impact of taking another bound on LCC size significance as motivation to only use statistically significant SNPs (yet with a modest entry criterium)
    - Focus on 1vs2, 1vs3, 2vs3 "discriminating" SNPS (F_{ST} bounded by 0.008) to interpret (may go in supplement)
    - Do interpretation analysis on the intersection of 1vs2, 1vs3, 2vs3 SNPS (F_{ST} bounded by 0.008) in addition (i.e. Steps 4 and 5)

- Decision 2:
  - Rerun everything only maintaining SNPs that occur in 3 GWAS (all curated SNPs – clarify whether the QC was repeated on the pooled data or not)
    - First construct Venn-diagram showing three circles (1vs2, 1vs3, 2vs3) and listing the number of GWAS hits at 0.05, using a simple chi-square.
    - Then start looking into 1vs2, 1vs3, 2vs3 “discriminating” SNPs (GWAS hits at 0.05) to do LCC analyses (size; distance), no repetition of checking impact of threshold, enrichment analysis
    - As before, do interpretation analysis on the intersection of 1vs2, 1vs3, 2vs3 SNPS (GWAS hits at 0.05) in addition.

Step 4: Discrimination - in the pipeline

- **Enrichment or depletion of functional terms (GO, pathways)**

  Choices to be made

  - All differentiating SNPs vs key driver SNPs
  - SNPs or regions around them → genes
    - Based on 250kb left and right
    - Based on LD block (e.g. $r^2 > 0.50$)
    - Note: some regions will harbor multiple genes; adapted enrichment analysis (Jostins et al. 2012)

  Decisions made

  - See before; LD block accounting in this paper when determining "coding" or "non-coding" (see next)

Step 4: Discrimination - in the pipeline

- **Coding SNPs (**link to "edgetics"**)**
  - functionGVS to annotate as coding SNP if it was classified as "missense" or "nonsense" or if it was in LD (r2>0.8) with a SNP with that classification
  - SNPnexus (Ullah et al 2018): broader functional annotation of *noncoding* variants and expanding annotations to the most recent human genome assembly; going beyond f.i. SIFT and PolyPhen predictions for deleterious effect of coding variants on protein functions

Step 4: Discrimination - in the pipeline (depends on input markers…)

- **In silico eQTL analysis**
  - eQTL database is GTEx Portal
  - FUMA: expression quantitative trait loci (eQTL) mapping (Watanabe et al 2017); SNP2GENE and GENE2FUNC core processes
- **Interaction enrichment**
  - Each gene is measured for enrichment in either direct or indirect (via other proteins in PPI) interactions with genes in other loci
- **Edge enrichment**:
  - Do genes in a targeted pathway have the tendency to be connected in the identified LCC?

Step 4: Discrimination - in the pipeline (depends on input markers…)

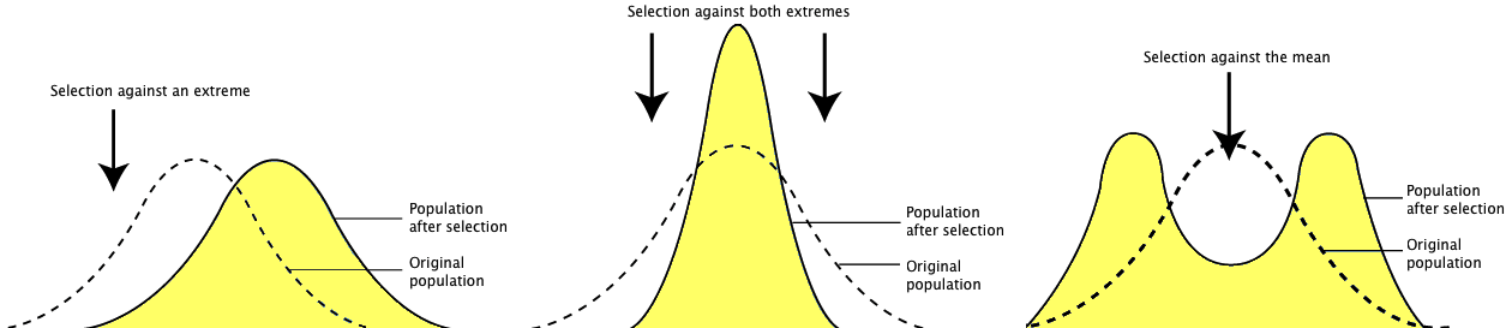
- **Redo enrichment analysis on "significant" group differentiators**
  (including those that are common for the three pairs of subtypes)
  - Ontology (GO) and KEGG pathway enrichment analysis using clusterprofiler (Yu et al., 2012) with Benjamini-Hochberg adjusted p-values.
  - Classify GO terms into three categories: biological process, molecular function and cellular component and performed enrichment for each category
  - Considering anticipated role of immune specific pathways, identify enriched immunogenic signatures via MsigDB c7 data

Step 4: Discrimination - in the pipeline (depends on input markers…)

- **Population genetics** [Toomas Kivisild; KULeuven]
  - Assess frequency of the highly differentiating SNPs between subgroups in ancient European populations to see whether there have been significant changes in those SNPs between the past and the present
  - to assess whether there is more long range (>2cM, >5cM) IBD sharing within as opposed to among the 3 CD subgroups and whether the long-range IBD (as a signal of selection) would be enriched around the highlighted SNPs

- **Population genetics** [Toomas Kivisild; KULeuven]
  - Stabilizing versus directional selection: Why might observed and expected phenotype frequencies differ?

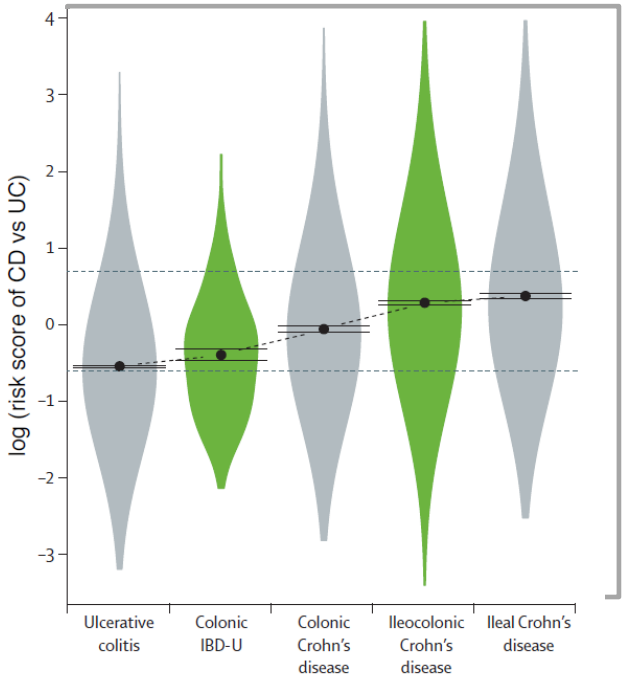
    Scenarios with natural selection at work and the phenotypic consequences over time:
    - Situation one favors only one tail of the distribution. Perhaps the tallest, perhaps the shortest, but not both. This is **directional selection.**
    - Both tails of the distribution are selected against, and only the middle is favored. This is called **stabilizing selection**.
    - The extremes on both ends are favored. This is called **disruptive selection**.



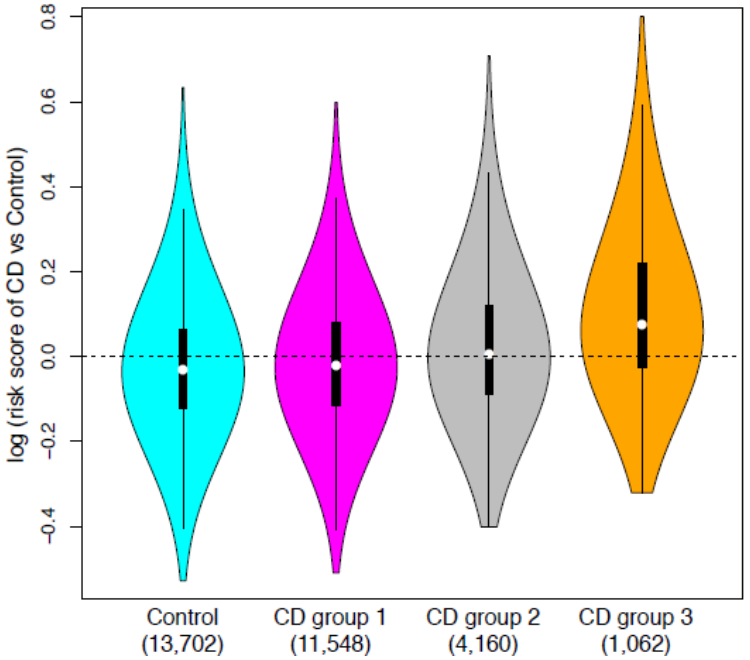
Step 5: Characterize the clusters – risk score prediction

Choices to be made

- CD vs healthy? Which variants? (~GWAS PRS)



(Cleynen et al 2016)



(based on 24 SNPs; Once set of SNPs is fixed, recompute p-values based on Mann Whitney U test as the overlap between the median regions does not hint towards significant differences)

Step 5: Characterize the clusters – in the pipeline

- **WGCN and module enrichment of differentiating/driver genes**

  In general, when control gene expression data are available, perform DE between each subtype and controls (not for this paper)

- **Observed minus expected genotype frequencies**

  observed in each subtype separately; expected via dbgap / reference allele frequencies) will highlight special genes (use SNP-to-GENE maps already used before)

  links to heterozygosity tests; heterozygosity has been positively associated to fitness and population survival; mean heterozygosity has been used to measure the degree of genetic variation

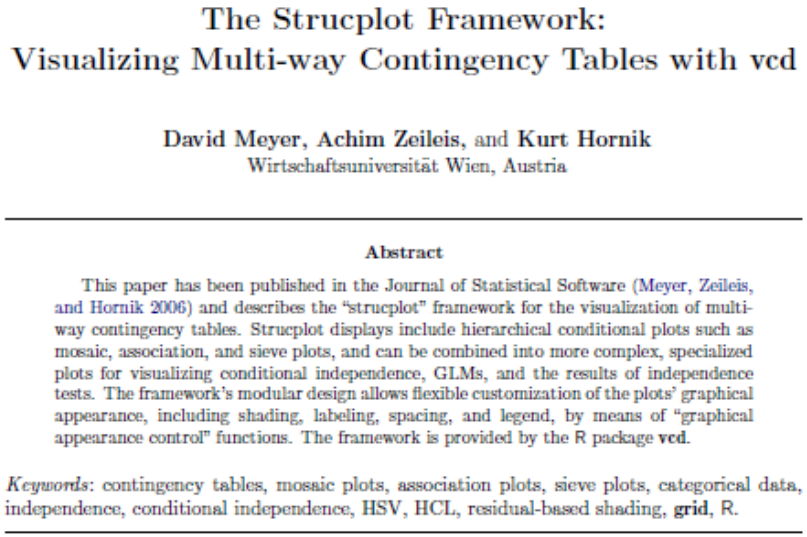
Step 5: Characterize the clusters – in the pipeline

- **Link to phenotypes**
  - How:
    - Multiple correspondence analysis with group membership
    - Missing data (e.g., PHENIX - phenotype imputation method for genetic studies, Dahl et al. 2012; Regularized Iterative Multiple Correspondence Analysis)
  - Which ones?
    - Sex
    - Smoking (smoker, ex-smoker, never-smoker)
    - Age of diagnosis
    - Affected body part
    - Disease behavior

- Plots include:

Multiple correspondence analysis plot

Table visualization including double decker plots for CD subgroups



The Strucplot Framework:
Visualizing Multi-way Contingency Tables with vcd

David Meyer, Achim Zeileis, and Kurt Hornik
Wirtschaftsuniversität Wien, Austria

Abstract

This paper has been published in the Journal of Statistical Software (Meyer, Zeileis, and Hornik 2006) and describes the "strucplot" framework for the visualization of multi-way contingency tables. Strucplot displays include hierarchical conditional plots such as mosaic, association, and sieve plots, and can be combined into more complex, specialized plots for visualizing conditional independence, GLMs, and the results of independence tests. The framework's modular design allows flexible customization of the plots' graphical appearance, including shading, labeling, spacing, and legend, by means of "graphical appearance control" functions. The framework is provided by the R package **vcd**.

Keywords: contingency tables, mosaic plots, association plots, sieve plots, categorical data, independence, conditional independence, HSV, HCL, residual-based shading, **grid**, R.

Step 5: Characterize the clusters – in the pipeline

- **Link to gene-drug databases**
  - How:
    - Consider characterizing genes for each CD subtype (may be based on gene expression or deviation from heterozygosity profiling with each subtype)
    - Use SNP-to-Gene mapping tool used before and consult

DGIdb 3.0: a redesign and expansion of the drug–gene interaction database

Kelsy C. Cotto[1,†], Alex H. Wagner[1,*,†], Yang-Yang Feng[1], Susanna Kiwala[1], Adam C. Coffman[1], Gregory Spies[1], Alex Wollam[1], Nicholas C. Spies[1], Obi L. Griffith[1,2,3,4,*] and Malachi Griffith[1,2,3,4,*]

[1]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO 63108, USA, [2]Siteman Cancer Center, Washington University School of Medicine, St Louis, MO 63110, USA, [3]Department of Medicine, Washington University School of Medicine, St Louis, MO 63110, USA and [4]Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

Received September 16, 2017; Revised October 20, 2017; Editorial Decision October 21, 2017; Accepted November 06, 2017

	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]

Step X - Stability

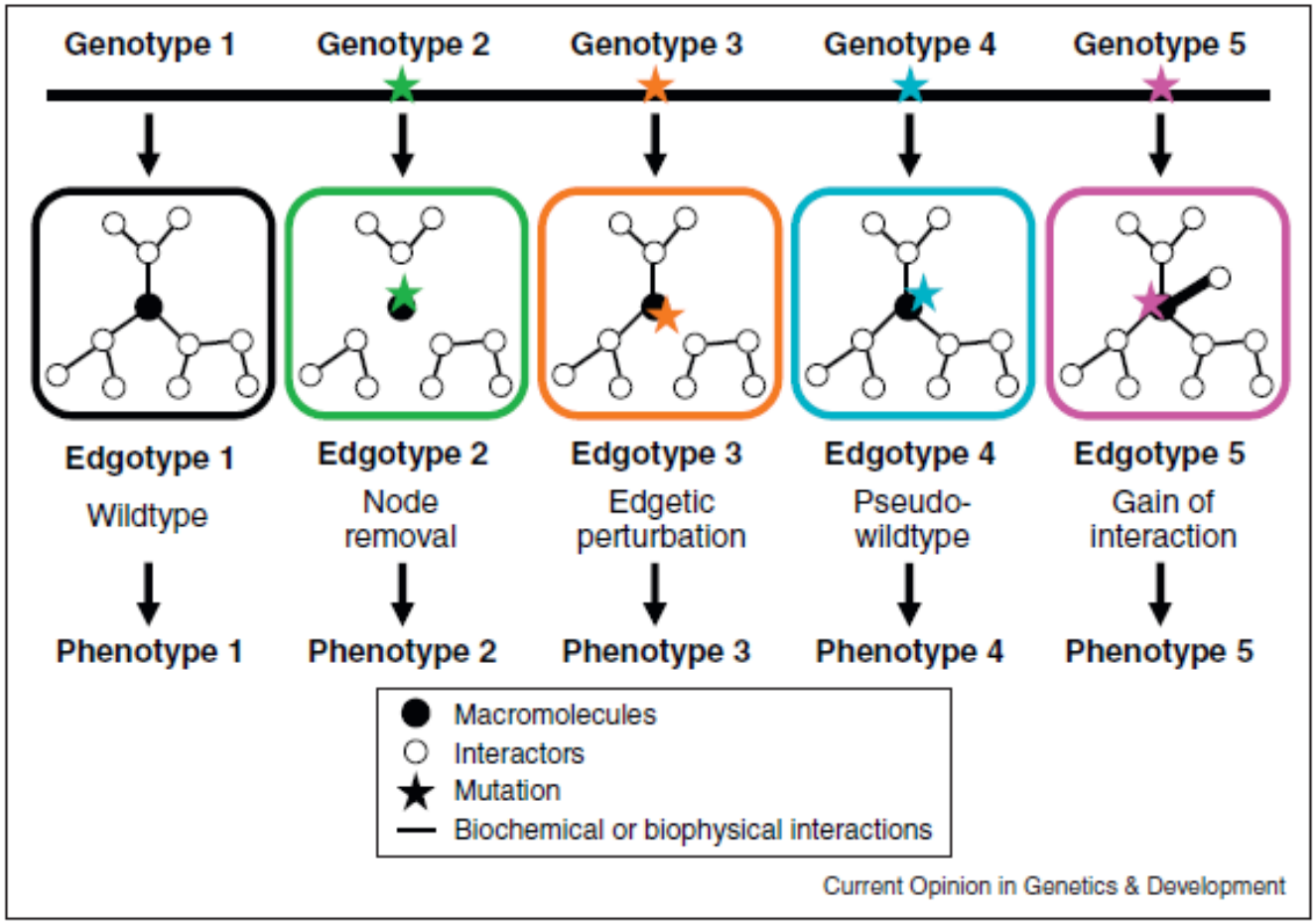
- Bootstrap sampling (X times) and distribution of Jaccard similarity between a cluster and its most similar bootstrap cluster (Hennig 2006) [cutoff of Jaccard: at least 0.75]
- Cluster-wise assessment of cluster stability (Ahlqvist et al. 2018) / Maximal compactness and maximal separation of clusters/ minimal hypervolume and maximal density of clusters (Du 2010) → special criteria for fuzzy clustering algorithms; these measures can also be used to find optimal parameter settings in the selected clustering algorithms
- Molecular stability: EBIC for optimized biclustering (Orzechowski et al. 2018) ; "runibic" as a Bioconductor package for parallel row-based biclustering of gene expression data (Orzechowski et al. 2019) → do you observe that the three CD groups cluster with gene expression modules?

Step X: Replication or how useful is it (translational science)

- Marker selection is related to replication success
  - Not everyone has been typed for … (~missingness)
  - Selection of “informative” markers (~informativity in the context of “prediction” ≠ informativity in the context of “differentiaton” → “characterizing” markers ~ group membership predictive markers?)
- Assign X new patients to the discovery clusters they are most similar to (which distance measure?)
  - Which characteristics can be attributed to replication clusters?
  - Compare replication to predicted clusters (choices!)

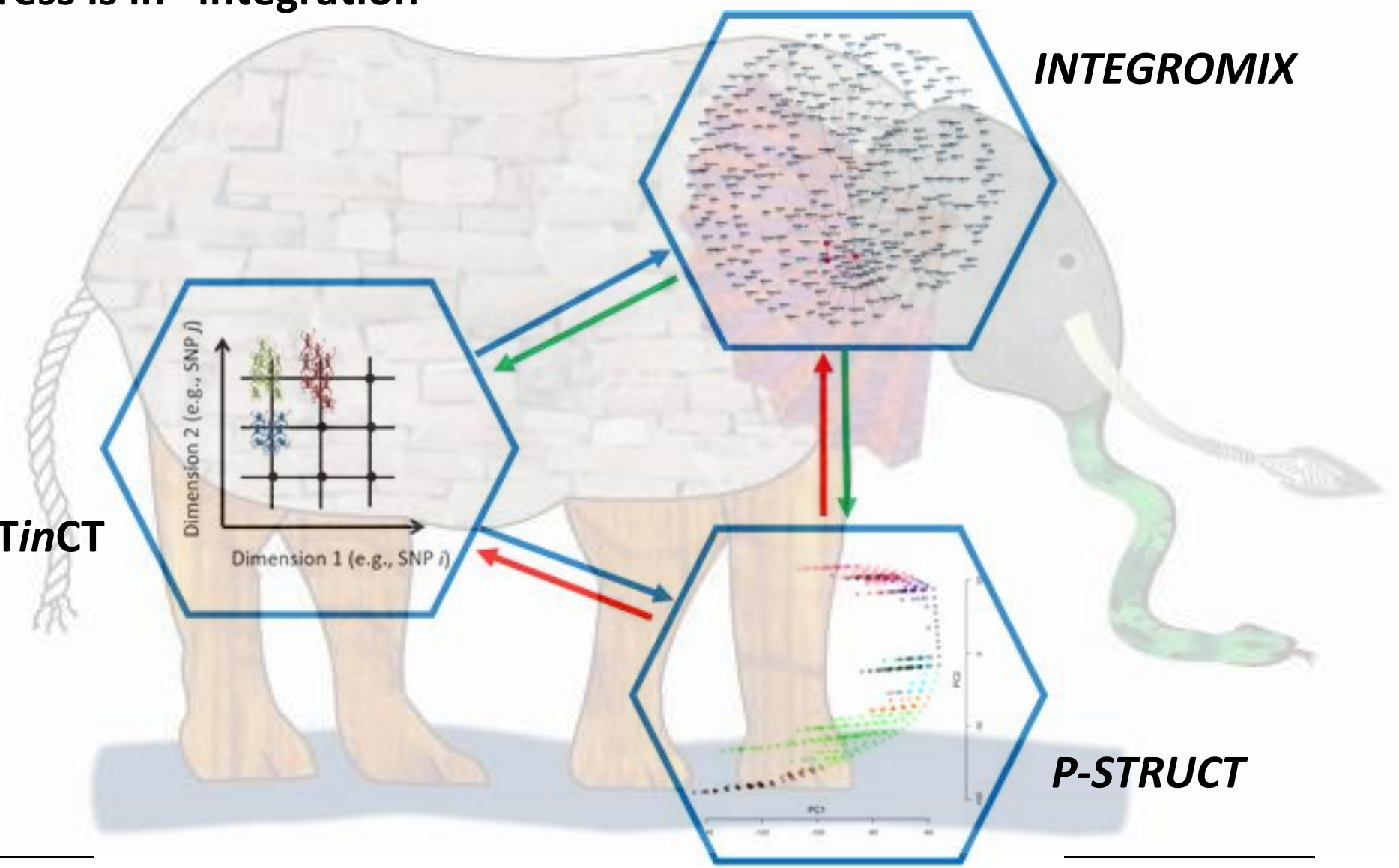
Take-home messages

Patient entry criteria - Information is in “the edges”



(moving towards individual networks)

Progress is in “integration”



Embrace learning representations for clustering

- Clustering techniques are abundant and have been studied extensively in terms of distance functions and grouping algorithms
- Much less studies in this area from the “machine learning” community:
  - Deep embedded clustering which **simultaneously learns feature representation and cluster assignments** using deep neural networks (Xie et al. 2016)
  - Regularized unsupervised **multiple kernel learning to integrate** data for clustering: linear weighted combinations of affinity matrices; weights optimized using multiple kernel learning (Speicher et al. 2015)

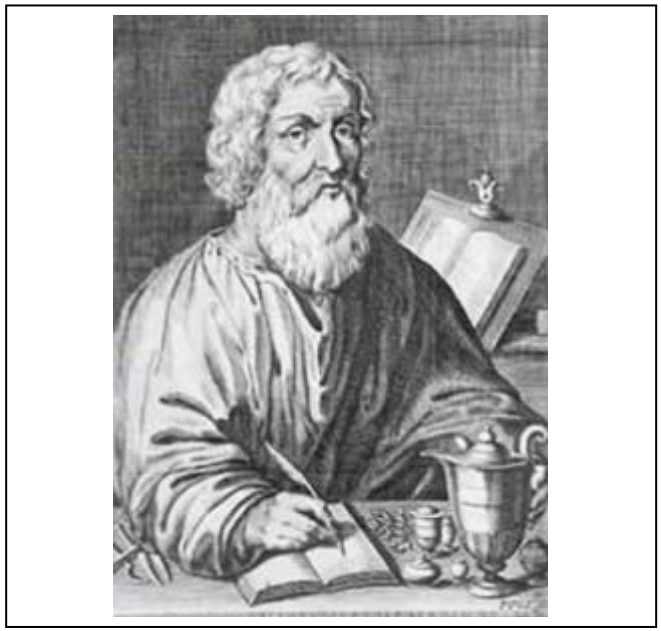
Don’t forget about presumably healthy populations

- To benchmark
- To target for interventions: risk prediction

- Again lessons can be learned from work on “interactions”
  - Collaborators extended **MB-MDR** to generate **prediction rules**
  - The new algorithm (available in R) can use information hidden in interactions more efficiently than two other state-of-the-art algorithms; it clearly **outperforms Random Forest and Elastic Net** if interactions are present.
  - The performance of these algorithms is comparable if no interactions are present

(Gola et al. 2019)

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



Hippocrates (460-370 BC)

Acknowledgements

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

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