To Interact or not to Interact

a tale of two visions

Kristel Van Steen, PhD[2]

kristel.vansteen@ulg.ac.be

Systems and Modeling Unit, Montefiore Institute, University of Liège, Grande Traverse 10, 4000 Liège, Belgium

Bioinformatics and Modeling, GIGA-R, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium

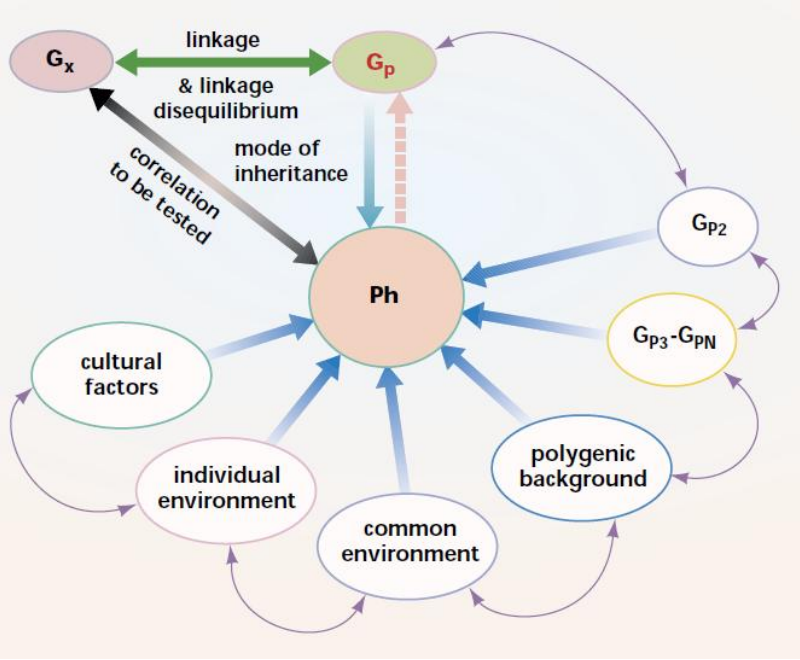
Université de Liège

Outline

- The origin of “interactions”
- Travelling the world of interactions
- How to best build our working space
- Model-Based Multifactor Dimensionality Reduction
- An example on Alzheimer’s disease
- Validation and replication: An impossible task?

The origin of interactions

The complexity of complex diseases

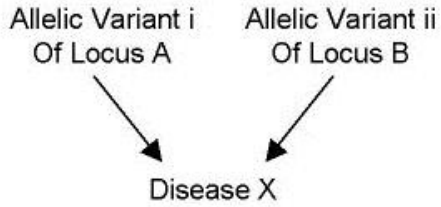
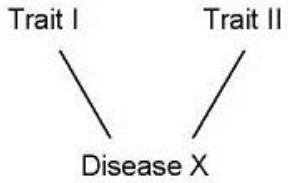
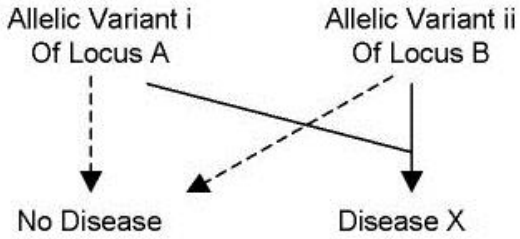


(Weiss and Terwilliger 2000)

There are likely to be *many* susceptibility genes each with combinations of *rare and common* alleles and genotypes that impact disease susceptibility primarily through *non-linear interactions* with *genetic and environmental* factors

(Moore 2008)

Factors complicating analysis of complex genetic disease

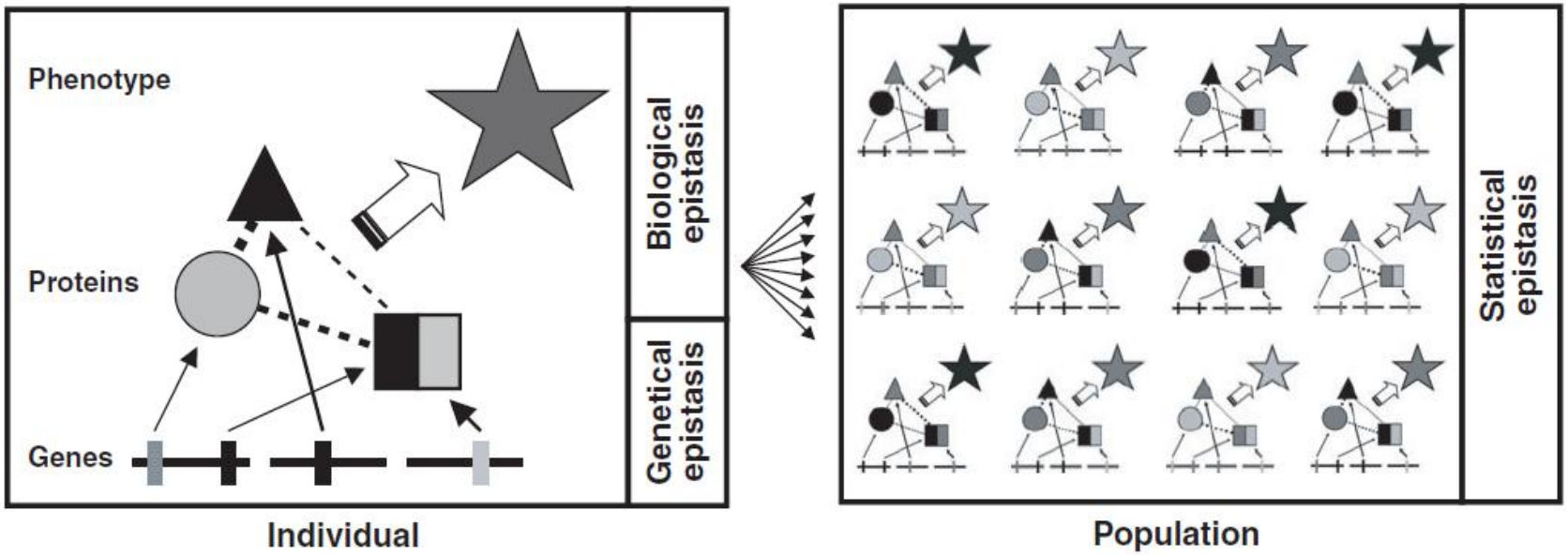
	Locus Heterogeneity	Trait Heterogeneity	Gene-Gene Interaction
Definition	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	when two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects
Diagram			
Example	**Retinitis Pigmentosa** (RP, OMIM# 268000) - genetic variations in at least fifteen genes have been associated with RP under an autosomal recessive model. Still more have been associated with RP under autosomal dominant and X-linked disease models[2] (http://www.sph.uth.tmc.edu/RetNet)	**Autosomal Dominant Cerebellar Ataxia** (ADCA, OMIM# 164500) - originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms,[6,7] and different genetic loci have been associated with the different subtypes[8]	**Hirschsprung Disease** (OMIM# 142623) - variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically such that they increase disease risk far beyond the combined risk of the independent variants[12]

(Thornton-Wells et al. 2006)

Factors complicating analysis of complex genetic disease

Gene-gene interactions

… when two or more DNA variations interact either directly to change transcription or translation levels, or indirectly by way of their protein product, to alter disease risk separate from their independent effects …



(Moore 2005)

The “observed” occurrences of epistasis – model organisms

- Carlborg and Haley (2004):
  - Epistatic QTLs without individual effects have been found in various organisms, such as birds[26,27], mammals[28–32], Drosophila melanogaster[33] and plants[18,34].
  - However, other similar studies have reported only low levels of epistasis or no epistasis at all, despite being thorough and involving large sample sizes[35–37].

This clearly indicates the complexity with which multifactorial traits are regulated; no single mode of inheritance can be expected to be the rule in all populations and traits.

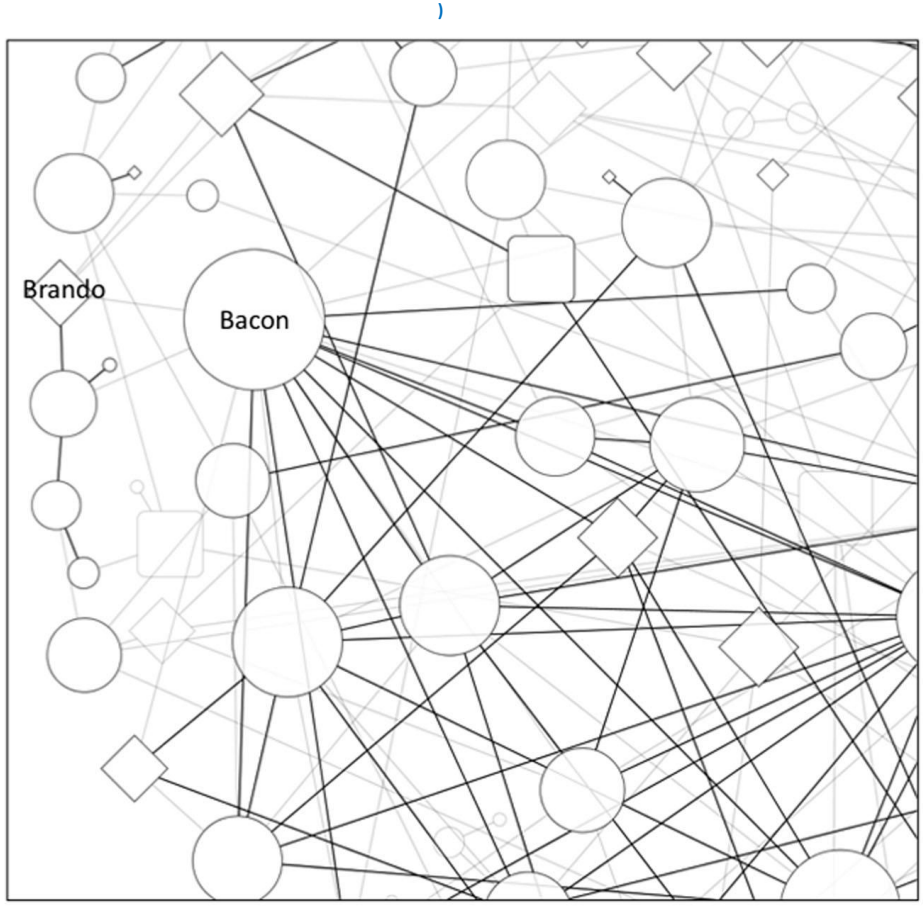
Great expectations

- From an evolutionary biology perspective, for a phenotype to be buffered against the effects of mutations, it must have an underlying genetic architecture that is comprised of networks of genes that are redundant and robust.
- The existence of these networks creates dependencies among the genes in the network and is realized as gene-gene interactions or (*trans*-) epistasis.
- This suggests that epistasis is not only important in determining variation in natural and human populations, but should also be more widespread than initially thought (rather than being a limited phenomenon).

Great expectations - empowering personal genomics

- Considering the epic complexity of the transcriptions process, the genetics of gene expression seems just as likely to harbor epistasis as biological pathways.
- When examining HapMap genotypes and gene expression levels from corresponding cell lines to look for cis-epistasis, over 75 genes pop up where SNP pairs in the gene's regulatory region can interact to influence the gene's expression.
- What is perhaps most interesting is that there are often large distances between the two interacting SNPs (with minimal LD between them), meaning that most haplotype and sliding window approaches would miss these effects. (Turner and Bush 2011)

Complementing insights from GWA studies



Edges **represent small gene–gene interactions between SNPs.** Gray nodes and edges **have weaker interactions.** Circle nodes **represent SNPs that do not have a significant main effect.** The diamond nodes **represent significant main effect association.** The size of the node **is proportional to the number of connections.**

(McKinney et al 2012)

Epistasis and phantom heritability

(Maher 2008)

Epistasis and phantom heritability

- Human genetics has been haunted by the mystery of “missing heritability” of common traits.
- Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability.
- The proportion of heritability explained by a set of variants is the ratio of (i) the heritability due to these variants (numerator), estimated directly from their observed effects, to (ii) the total heritability (denominator), inferred indirectly from population data.
- The prevailing view has been that the explanation for missing heritability lies in the numerator – variants still to identify

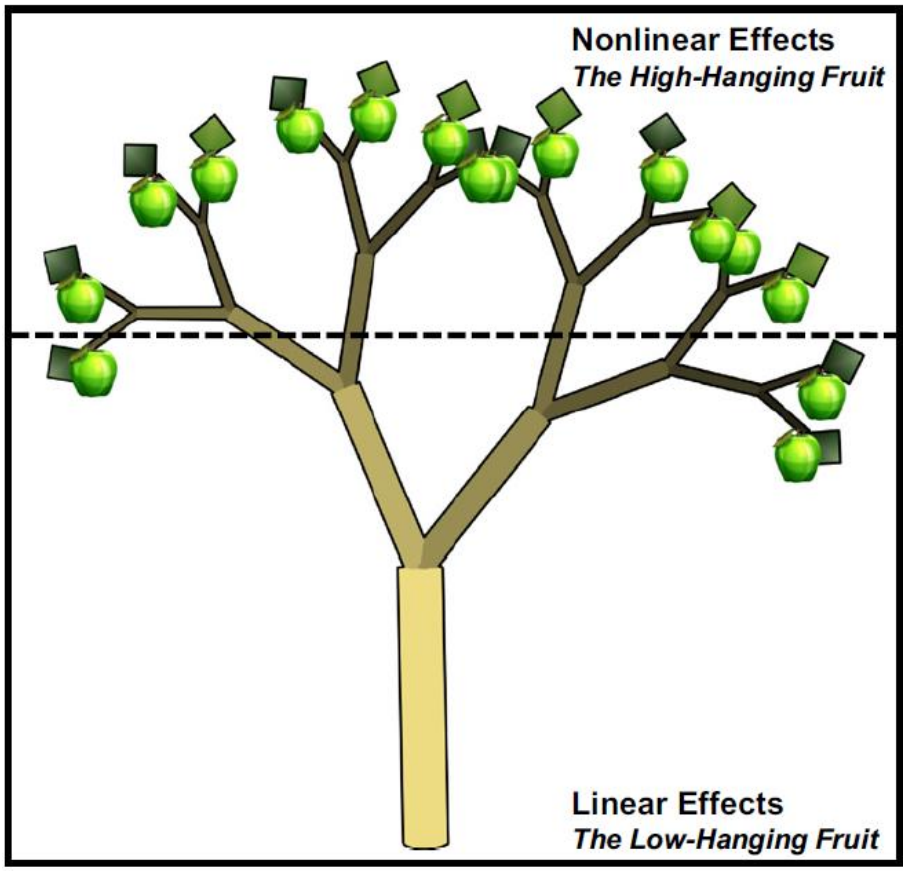
Epistasis and phantom heritability

- Overestimation of the total heritability can create “phantom heritability.”
  - estimates of total heritability implicitly assume the trait involves no genetic interactions (epistasis) among loci
  - this assumption is not justified
  - under such models, the total heritability may be much smaller and thus the proportion of heritability explained much larger.
- For example, 80% of the currently missing heritability for Crohn's disease could be due to genetic interactions, if the disease involves interaction among three pathways. (Zuk et al 2012)

Traveling the world of interactions

Université de Liège

Université de Liège



- Few SNPs with moderate to large independent and additive main effects

- Most SNPs of interest will only be found by embracing the complexity of the genotype-to-phenotype mapping relationship that is likely to be characterized by nonlinear gene-gene interactions, gene-environment interaction and locus heterogeneity.

(Moore and Williams 2009)

Université de Liège

From GWA to GWAI studies …

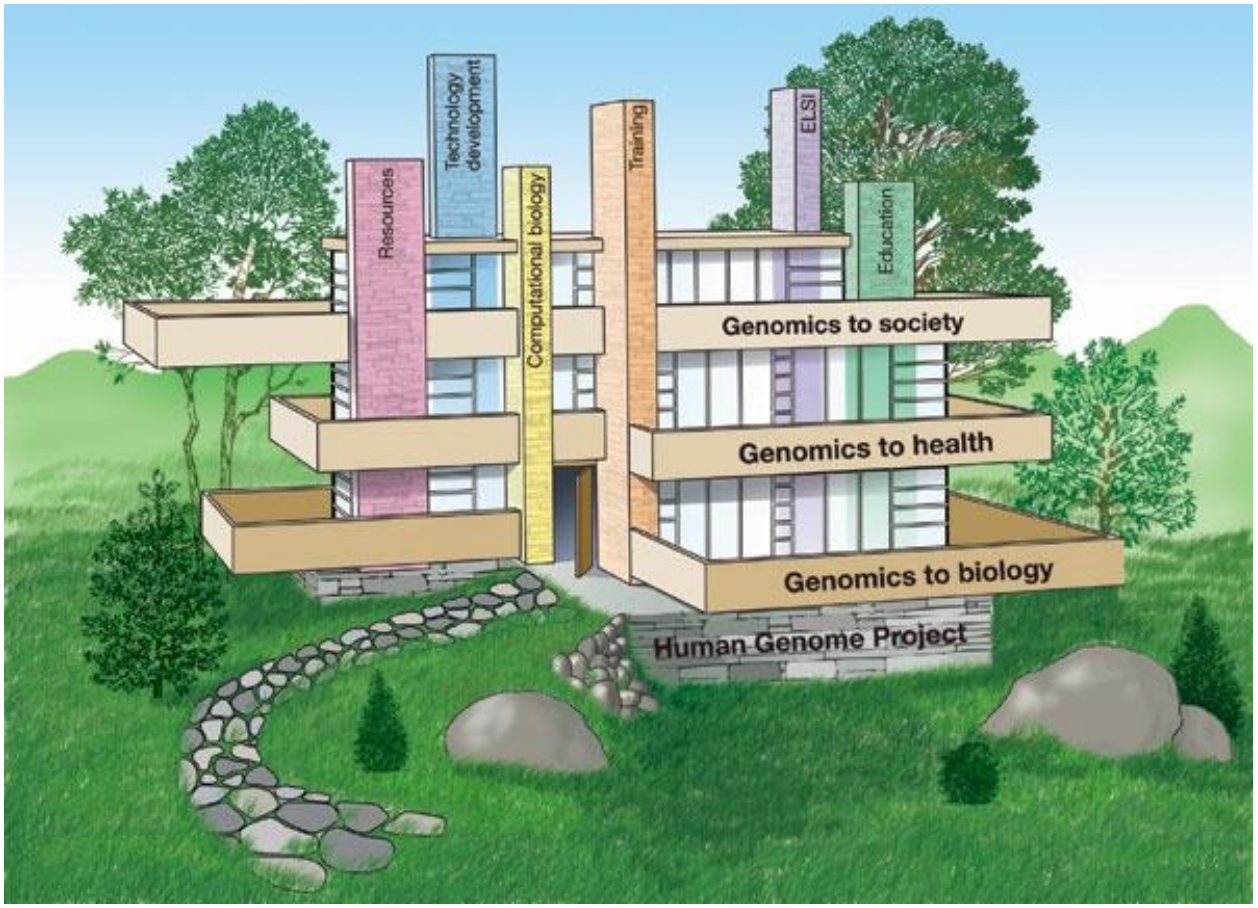
- Genome-Wide Association Interaction (GWAI) studies have not been as successful as GWA studies:
  - Possible negligible role of epistatic variance in a population? (Davierwala et al 2005)
  - Consequence of not yet available powerful epistasis detection methods or approaches?

    " Gene-gene interactions are commonly found when properly investigated " (Templeton 2000)

How to best build our working space

Université de Liège

Creating an atmosphere of "interdisciplinarity"



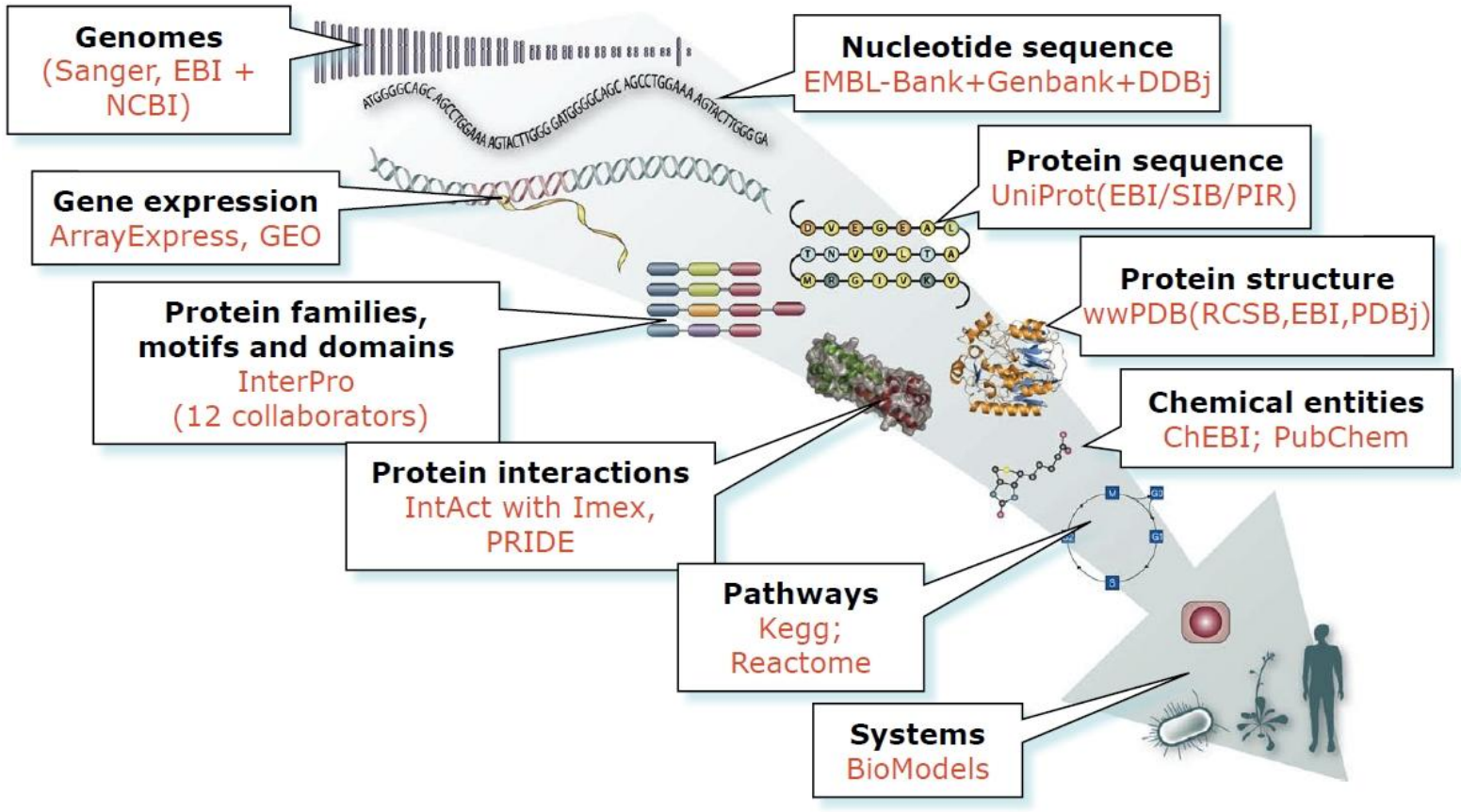
(http://www.genome.gov: the future of human genomics) + harmonization of biobanks

Université de Liège

Creating an atmosphere of "integration"

with HTP omics data

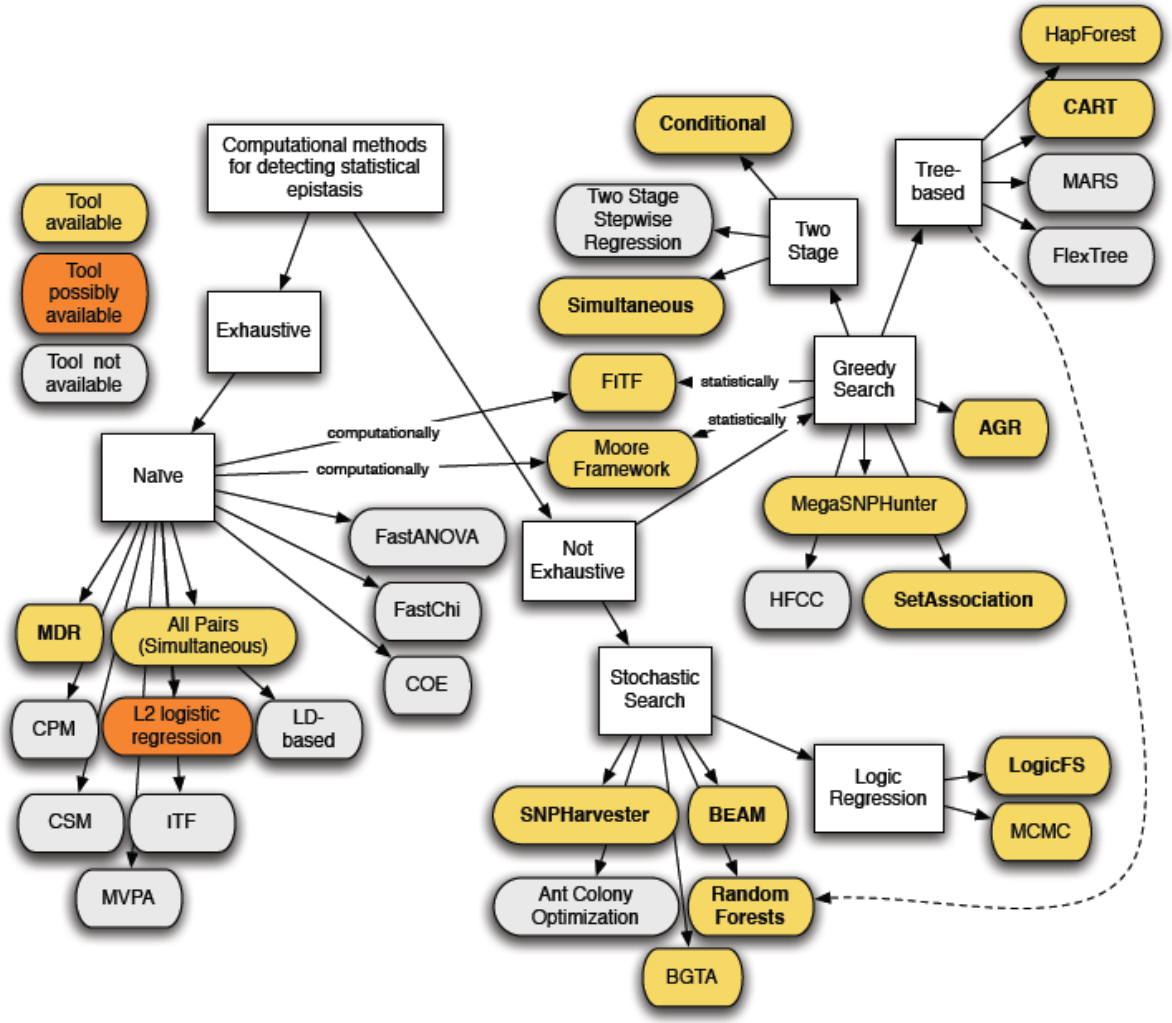
(J Thornton, EBI)



Université de Liège

Extending the toolbox

(Kilpatrick 2009)



Université de Liège

Extending the toolbox

- Comes with a caveat: need for thorough comparison studies using reference data sets!
- Several criteria exist to classify epistasis detection methods:
  - Exploratory versus non-exploratory
  - Testing versus Modeling
  - Direct versus Indirect testing
  - Parametric versus non-parametric
  - Exhaustive versus non-exhaustive search algorithms
  - ... (Van Steen et al 2011)

The "observed" occurrences of epistasis – humans

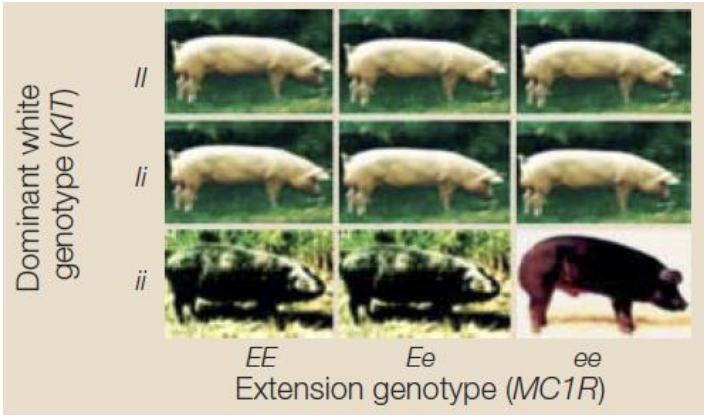
- Phillips et al (2008):
  - There are several cases of epistasis appearing as a statistical feature of association studies of human disease.
  - A few recent examples include coronary artery disease[63], diabetes[64], bipolar effective disorder[65], and autism[66].
  - So far, only for some of the reported findings additional support could be provided by functional analysis, as was the case for multiple sclerosis (Gregersen et al 2006).
- More recent examples: e.g., breast cancer (Ashworth et al. 2011), Alzheimer's (Combarros et al 2009),

Taking it a few steps back … What's in a name?

- Our ability to detect epistasis depends on what we mean by epistasis

"compositional epistasis"

- The original definition (**driven by biology**) refers to distortions of Mendelian segregation ratios due to one gene masking the effects of another; a variant or allele at one locus prevents the variant at another locus from manifesting its effect (William Bateson 1861-1926).



(Carlborg and Haley 2004)

Université de Liège

Taking it a few steps back … What's in a name?

"statistical epistasis"

- A later definition of epistasis (**driven by statistics**) is expressed in terms of deviations from a model of additive multiple effects.
- This might be on either a linear or logarithmic scale, which implies different definitions (Ronald Fisher 1890-1962).

- It seems that the interpretation of GWAIs is hampered by undetected false positives

Model-Based Multifactor Dimensionality Reduction

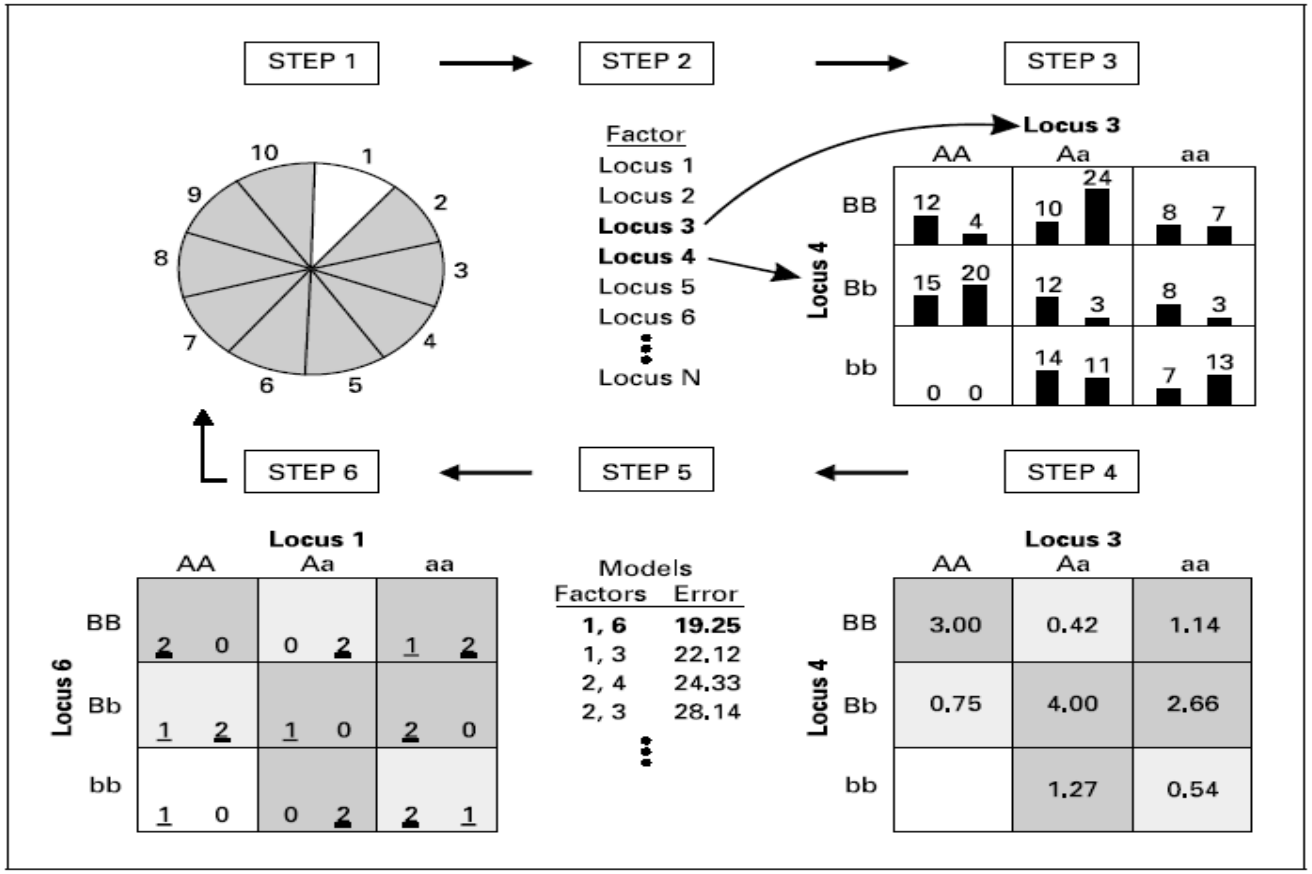
Université de Liège

Historical notes about MB-MDR

- Knowledge:
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small n big p problems may give rise to curse of dimensionality problems (Bellman 1961)
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies
- Alerts:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration

Historical notes about MB-MDR

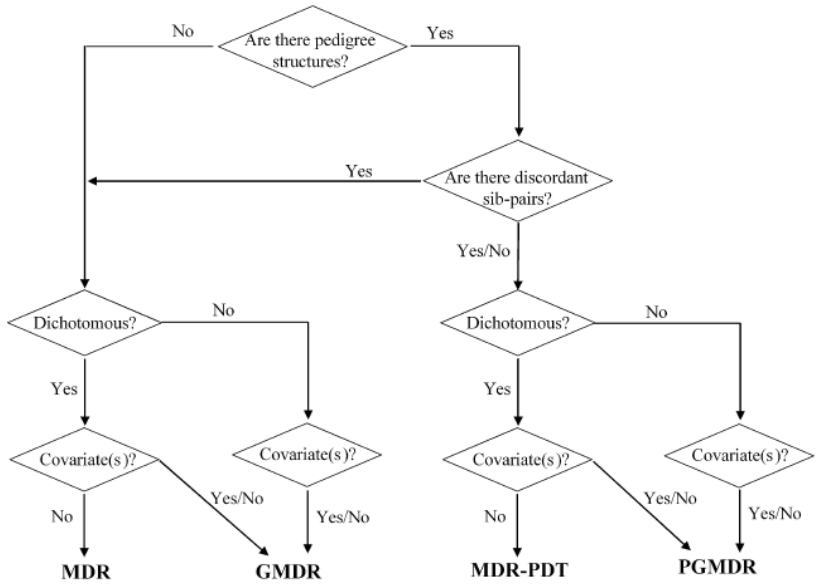
- Multifactor Dimensionality Reduction by MD Ritchie et al (2001)



Historical notes about MB-MDR

- Model-Based MDR by Calle et al (2007)

  - Unlike other MDR-like methods, MB-MDR breaks with the tradition of cross-validation to select optimal multilocus models with significant accuracy estimates



- Model-Based MDR by Calle et al (2007)

  - Rather, computation time is invested in optimal association tests to prioritize multilocus genotype combinations and statistically valid permutation-based methods to assess joint statistical significance
  - Results of association tests are used to “label” multilocus genotype cells (for instance: increased / reduced risk, based on sign of “effect”) and to “quantify” the multilocus signal wrt the trait of interest, “above and beyond lower order signals”

Historical notes about MB-MDR

- Model-Based MDR by Calle et al (2007, 2008)

Table 3. MB-MDR first step analysis for interaction between SNP 40 and SNP 252 in the bladder cancer study

SNP 40 x SNP 252 genotypes	Cases	Controls	OR	p-value	Category
c1 = (0,0)	88	77	1.01	0.9303	0
c2 = (0,1)	102	114	0.73	0.0562	L
c3 = (0,2)	38	34	0.98	1.0000	0
c4 = (1,0)	50	59	0.76	0.1229	0
c5 = (1,1)	96	37	2.68	0.0000	H
c6 = (1,2)	18	28	0.55	0.0675	L
c7 = (2,0)	12	6	1.99	0.3399	0
c8 = (2,1)	14	18	0.67	0.3668	0
c9 = (2,2)	6	6	0.84	1.0000	0

H: High risk; L: Low risk; 0: No evidence

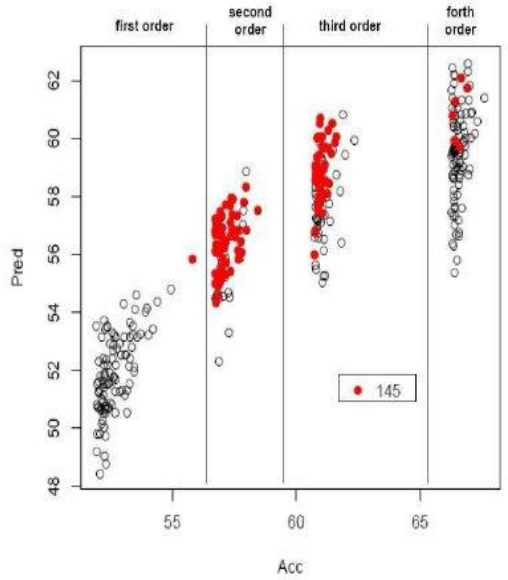
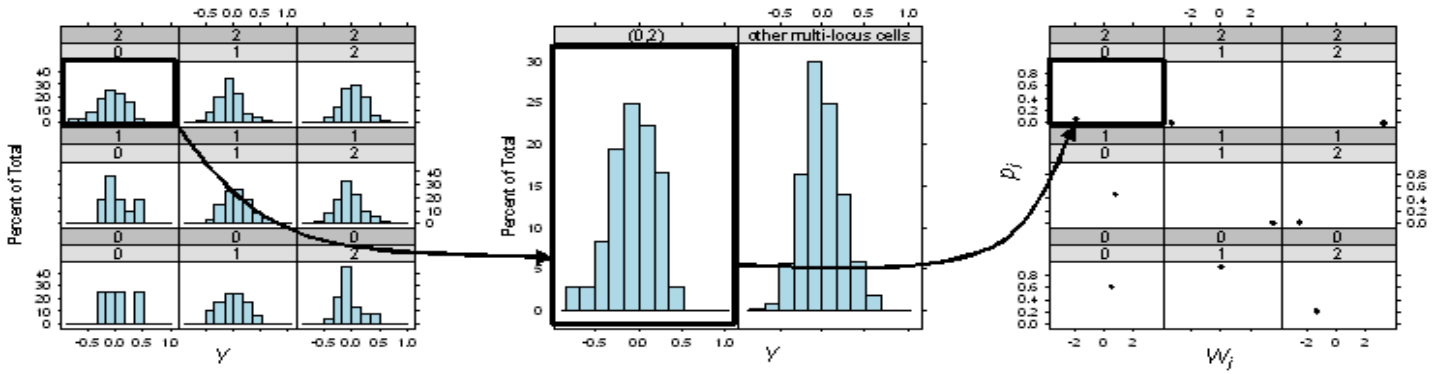


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

Historical notes about MB-MDR

- Model-Based MDR by Cattaert et al (2010)



  - Pooling "alike" (for instance, all low-risk and all high-risk) multilocus genotypes leads to statistic distribution that is different from the theoretical distribution (data snooping)
  - Score tests, one multilocus p-value and permutation-based strategy (Cattaert et al 2010), rather than Wald tests, and relying on MAF dependent reference distributions (Calle et al 2008)

Historical notes about MB-MDR

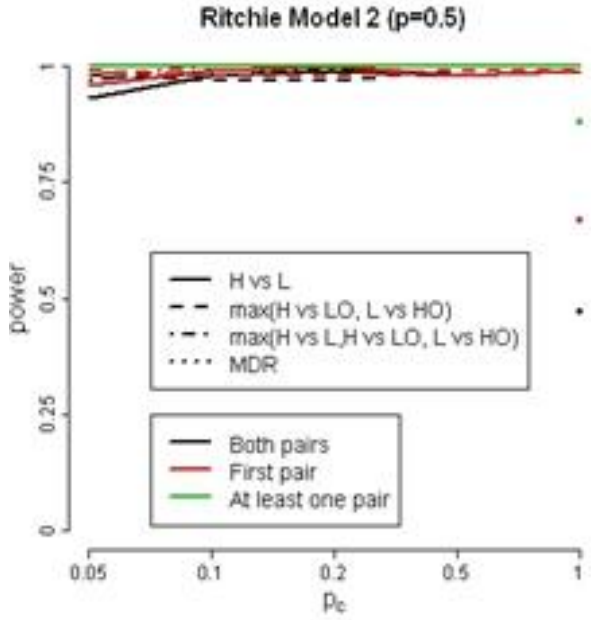
- Model-Based MDR by Cattaert et al (2010)

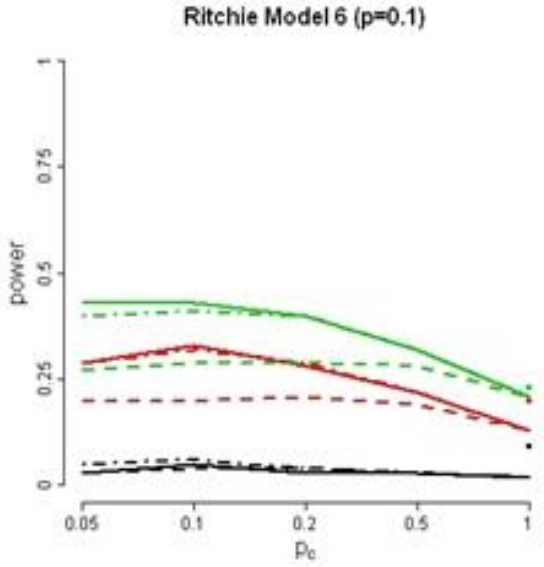
Model 2, $p = 0.5$

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

Model 6, $p = 0.1$

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

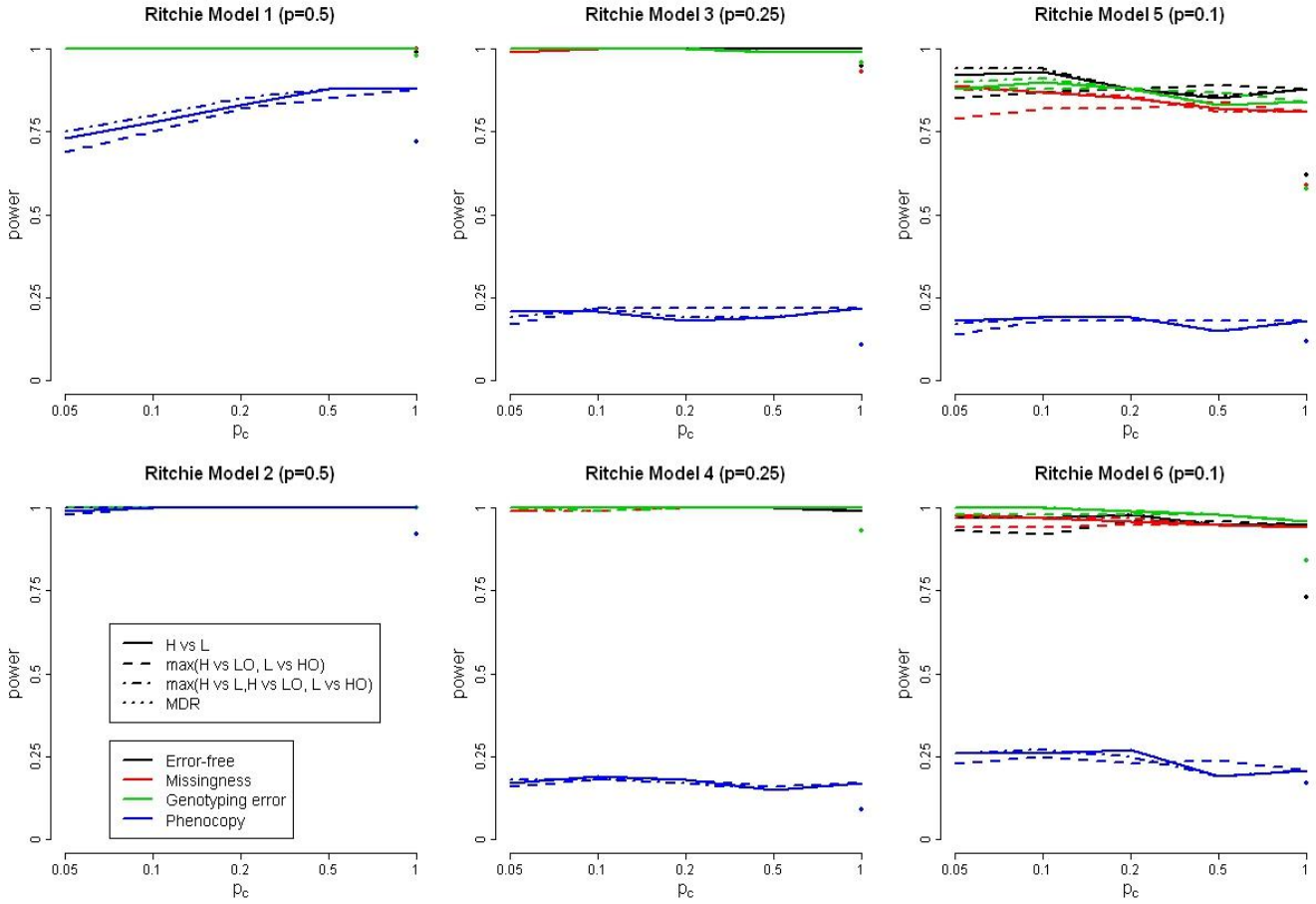




Université de Liège

Historical notes about MB-MDR

- Model-Based MDR by Cattaert et al (2010) – maximizing power



Université de Liège

Historical notes about MB-MDR

- Model-Based MDR by Van Lishout et al (2012 – under review) – speed
  - MaxT algorithm √
  - Association test statistics (parametric and non-parametric) √ +

SNPs	*MBMDR-3.0.2* sequential execution Binary trait	*MBMDR-3.0.2* sequential execution Continuous trait	*MBMDR-3.0.2* parallel workflow Binary trait	*MBMDR-3.0.2* parallel workflow Continuous trait
100	45 sec	1 min 35 sec	<1sec	<1sec
1,000	1 hour 16 minutes	2 hours 39 minutes	38 sec	1 min 17 sec
10,000	5 days 13 hours	11 days 19 hours	1 hour 3 min	2 hours 14 min
100,000	≈ 1.5 year	≈ 3 years	4 days 9 hours	≈ 9 days

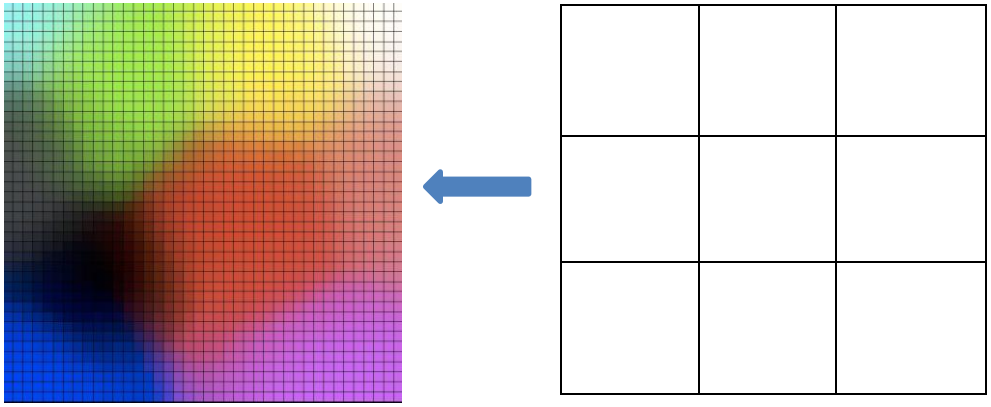
The parallel workflow was tested on a cluster composed of 10 blades, containing each four Quad-Core AMD Opteron(tm) Processor 2352 2.1 GHz.

The sequential executions were performed on a single core of this cluster.

The results prefixed by the symbol "≈" are extrapolated.

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)
  - Lower order effects correction (omit at cell-labeling step) √ +
  - Two-locus effect modifiers √
  - Different faces of “dimensions” in dimensionality reduction



Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)

Human Heredity

Original Paper

Hum Hered 2004;58:82–92
DOI: 10.1159/000083029

Received: June 30, 2004
Accepted after revision: September 23, 2004

MDR and PRP: A Comparison of Methods for High-Order Genotype-Phenotype Associations

L. Bastone[a] M. Reilly[b] D.J. Rader[b] A.S. Foulkes[c]

[a]Division of Biostatistics, [b]Cardiovascular Division and Center for Experimental Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, Pa., and [c]Department of Biostatistics, School of Public Health and Health Sciences, University of Massachusetts, Amherst, Mass., USA

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)

Human Heredity

Original Paper

Hum Hered 2004;58:82–92
DOI: 10.1159/000083029

Received: June 30, 2004
Accepted after revision: September 23, 2004

MDR and PRP: A Comparison of Methods for High-Order Genotype-Phenotype Associatio

L. Bastone[a] M. Reilly[b] D.J. Rader[b] A.S. Foulkes[c]

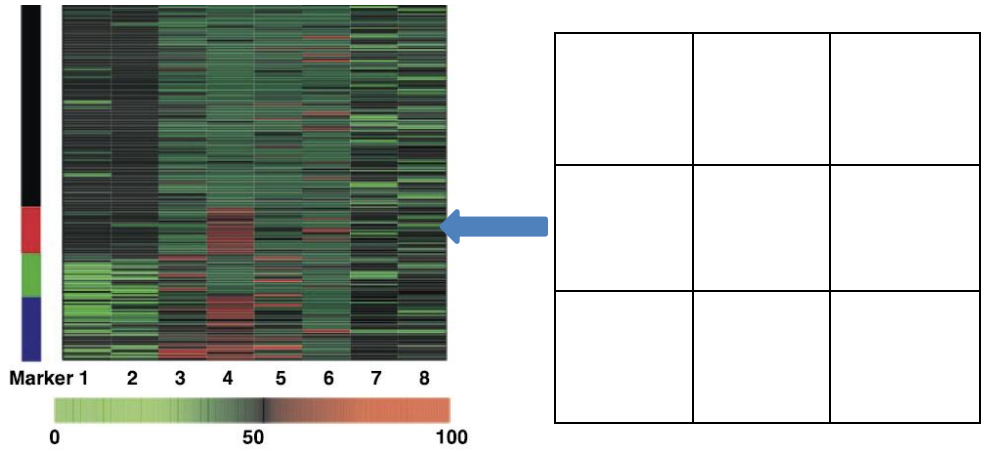
[a]Division of Biostatistics, [b]Cardiovascular Division and Center for Experimental Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, Pa., and [c]Department of Biostatis School of Public Health and Health Sciences, University of Massachusetts, Amherst, Mass., US

. Statistical methods suc as multifactor dimensionality reduction (MDR), the com binatorial partitioning method (CPM), recursive partitior ing (RP), and patterning and recursive partitioning (PRF are designed to uncover complex relationships withou relying on a specific model for the interaction, and ar therefore well-suited to this data setting. However, th theoretical overlap among these methods and their rela tive merits have not been well characterized. In thi paper we demonstrate mathematically that MDR is special case of RP

Université de Liège

Historical notes about MB-MDR

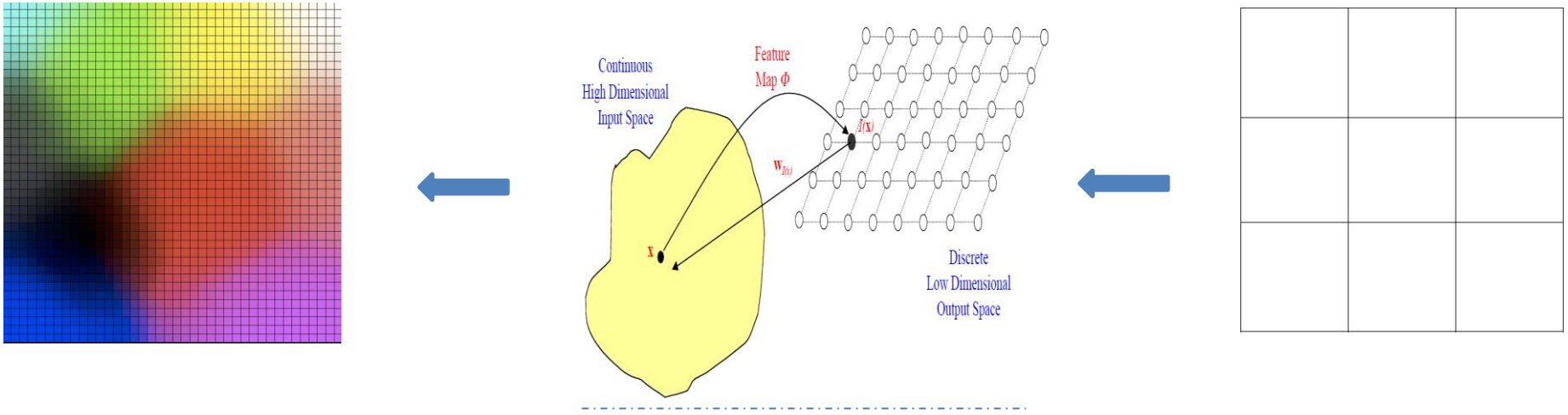
- Model-Based MDR by Van Steen lab (2012 and +)
  - Dimension (1,2) = (SNP1,SNP2) √
  - Dimension (1,2) = (SNP1, "categorized" continuous variable) √ +
  - Dimension (1,2) = (SNP1, genomic region with rare variants) +



(Shi et al 2006, unsupervised clustering with RFs)

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)
  - Dimension (1,2) = (SNP1, genomic region with rare variants) **+**
  - Dimension (1,2) = **+**



(SOMs: Bullinaria 2004)

An example on Alzheimer's disease

First hurdle: Selection of most appropriate method

- Honest methods comparisons should / can highlight the "core" (**the ABC**) of each method:

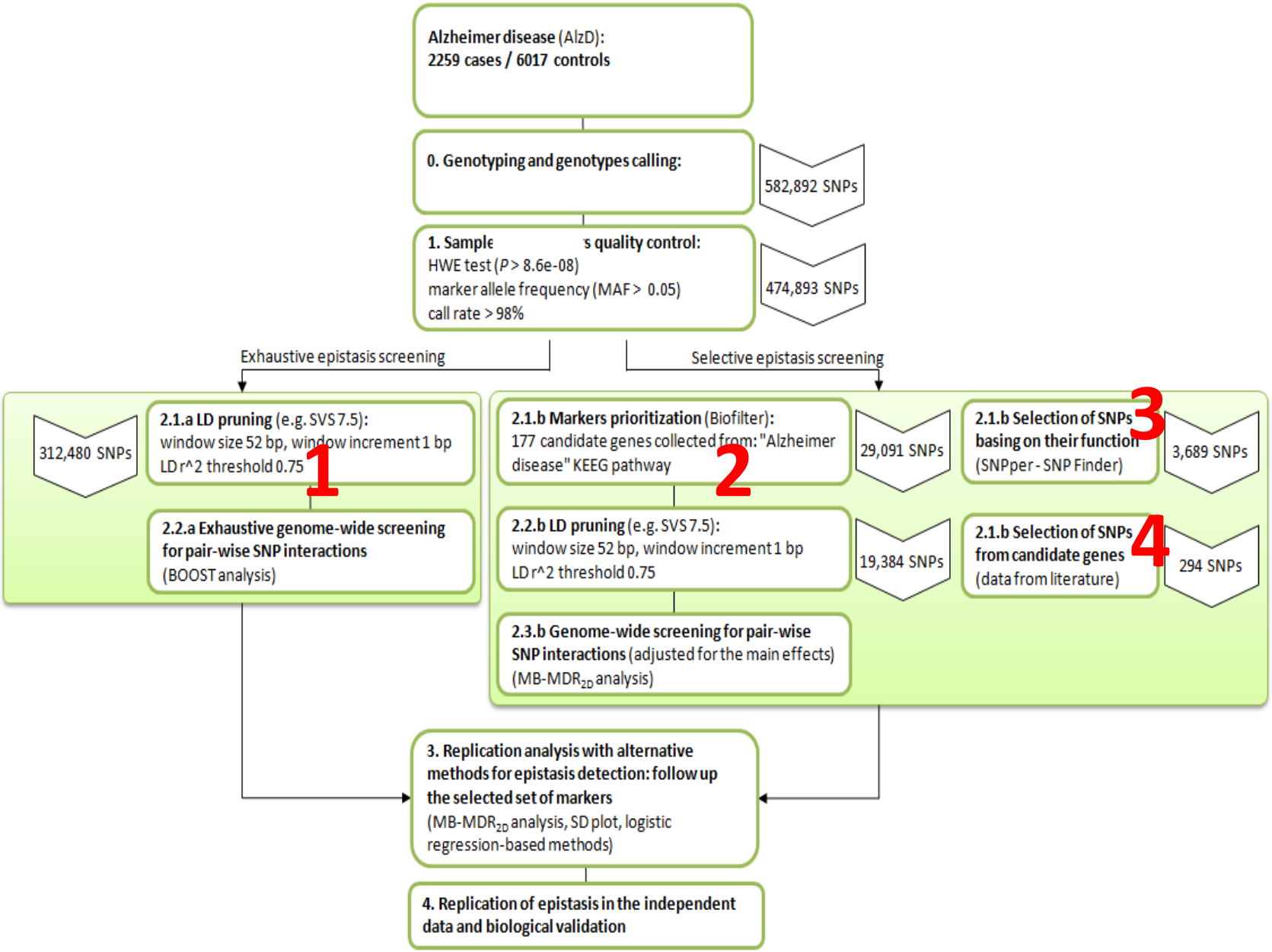
  A: Pre-processing (screening); **B**: core; **C**: multiple testing

(Van Steen lab: in preparation)

		EpiCruncher																MB-MDR	PLINK	EPIBLASTER
		Bonferroni								Permutations										
		LR test				Score test				LR test				Score test						
		Test statistic		P-value		Test statistic		P-value		Test statistic		P-value		Test statistic		P-value				
		M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5			
rs17116117	rs2513574	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rs17116117	rs2519200	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rs17116117	rs4938056	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
rs17116117	rs1713671	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
rs13126272	rs11936062	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
rs17116117	rs7126080	x	x	x	x					x	x	x	x							
rs3770132	rs1933641					x		x						x		x				
rs12339163	rs1933641					x		x						x		x				
rs12853584	rs1217414										x				x		x	x		
rs17116117	rs1169722																			x
number significant		**6**	**6**	**6**	**6**	**7**	**5**	**7**	**5**	**6**	**7**	**6**	**6**	**7**	**6**	**7**	6	**6**	**3**	**3**

Université de Liège

Protocol for GWAI analysis



Available "knowledge" about epistasis: candidate genes

Gene	Gene name	Function	Location	Epistatic SNPs	Main effect for AlzD	Population (N cases/N controls)	Reference
INS	Insulin	Glucose metabolism	11p15.5	rs689	no	Germans (104/123)	Brune et al., 2003
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes	Northern Europeans (336/2426)	Kölsch et al., 2012
IL1A	Interleukin 1 alfa	Inflammatory cytokine	2q13	rs3783550	no	Northern Europeans (336/2426)	Heun et al., 2012
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes		
IL1B	Interleukin 1 beta	Inflammatory cytokine	2q13	rs16944	no	Northern Europeans (336/2426)	Heun et al., 2012
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes		
IL10	Interleukin 10	Inflammatory cytokine	1q32.1	rs1800896	yes	Northern Europeans (336/2426)	Heun et al., 2012
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs4253766	no		
IL1A	Interleukin 1 alfa	Inflammatory cytokine	2q13	rs1800587	no	Northern Europeans (336/2426)	Combarros et al., 2010
DBH	b-Hydroxylase	Onverts dopamine to norepinephrine in the synaptic vesicles of postganglionic sympathetic neurons	9q34.2	rs1611115	yes		
TF	Transferrin	Iron metabolism	3q22.1	rs1049296	no	UK (191/269)	Robson et al., 2004
HFE	Hemochromatosis		6p22.2	rs1800562	yes	Caucasians USA (1166/1404)	Kauwe et al., 2010
						North Europeans (336/2426)	Lehmann et al., 2012
TF	Transferrin	Iron metabolism	3q22.1	rs1130459	no	North Europeans (336/2426)	Lehmann et al., 2012
HFE	Hemochromatosis		6p22.2	rs1799945	yes		
MTHFR	Methylenetetrahydrofolate reductase	Homocysteine metabolism useful for normal brain functioning	1p36.22	rs1801131	yes	Indians (80/120)	Mansoori et al., 2012
IL6	Interleukin 6	Pro-inflammatory cytokine	7p15.3	rs1800795	no		
IL10	Interleukin 10	Limit inflammation in the brain	1q32.1	rs1800871	yes	North Spains (232/191) ,	Infante et al., 2004
IL6	Interleukin 6	Pro-inflammatory cytokine	7p15.3	rs2069837	yes	North Europeans (336/2426)	Combarros et al., 2009
ABCA1	ATP-binding cassette transporter A1	Intracellular cholesterol transport and maintance of cell cholesterol balance	9q31.1	rs2422493	no	Spanish (631/731)	Rodríguez-Rodríguez et al., 2010
NPC1	Niemann-Pick C1		18q11.2	rs18050810 rs4800488 rs2236707 rs2510344	no		

LRP1	low density lipoprotein receptor-related protein 1	Neuronal uptake of cholesterol	12q13.3	rs1799986	no	Spanish (246/237)	Vázquez-Higuera et al., 2009
MAPT	Microtubule-associated protein tau		17q21.33	rs2471738	no		
GSK3B	Glycogen synthase kinase-3 beta	Abnormal hyperphosphorylation of tau, neuronal uptake of cholesterol	3q13.33	rs334558	no	Spanish (246/237)	Vázquez-Higuera et al., 2009
CDK5R1	Cyclindependent kinase 5		17q11.2	rs735555			
NR1H2	Liver X receptor beta	Cholesterol metabolism	19q13.33	rs1052533 rs1405655	no	Spanish (414/442)	Infante et al., 2010
HMOX1	Heme oxygenase-1		22q12.3	rs2071746			

Different levels

- Genetic marker
- Locus
- Gene
- Window including either one of the previous
- Pathway

Revised analysis for candidate gene pairs

- MB-MDR analysis: 294 SNPs selected from France_AlzD panel of SNPs

MTHFR	IL10	IL1A	IL1B	TF	HFE	IL6	ABCA1	DBH	INS	LRP1	CDK5R1	MAPT	NPC1	NR1H2	HMOX1	PPARA	
	+	ns	+	+	+	+	+	+	+	+	ns	+	+	+	ns	+	MTHFR
		+	+	+	ns	ns	+	+	ns	+	ns	+	ns	ns	+	+	IL10
			ns	+	+	+	+	ns	+	ns	ns	+	ns	ns	ns	+	IL1A
				+	ns	ns	+	ns	ns	+	ns	+	+	ns	ns	ns	IL1B
					+	+	+	+	ns	+	ns	+	+	+	+	+	TF
						+	+	ns	+	+	ns	+	+	+	ns	+	HFE
							+	+	ns	ns	ns	+	+	+	+	+	IL6
								+	+	+	ns	+	+	+	+	+	ABCA1
									+	+	ns	+	+	ns	+	+	DBH
										ns	ns	+	ns	ns	+	+	INS
											ns	+	ns	ns	+	+	LRP1
												ns	ns	ns	ns	ns	CDK5R1
													+	ns	+	+	MAPT
														ns	ns	+	NPC1
															ns	ns	NR1H2
																+	HMOX1
																	PPARA

"+" - at least one SNP pair from the corresponding genes was associated with AlzD

(the marginal *p*-value < 0.05 for the MB-MDR$_{2D}$ analysis)

Replication is highlighted by green; no replication is highlighted by red.

Université de Liège

Replication and validation of GWAIs: An impossible task?

Université de Liège

(Mission Impossible @ google)

Acknowledgments (MB-MDR)

Université de Liège

Acknowledgments (Alzheimer's)

Kristel Van Steen (PI)
Montefiore Institute / GIGA-R,
University of Liege,
Belgium

Nilufer Taner
Mayo Clinic Rochester, Departments of Neurology and Neuroscience, Rochester, USA

Elena Gusareva
Montefiore Institute / GIGA-R,
University of Liege,
Belgium

Kristel Sleegers
Neurodegenerative Brain Diseases Group,
Department of Molecular Genetics, VIB,
University of Antwerp,
Belgium

Jean-Charles Lambert
INSERM U744, Lille, France
Institut Pasteur de Lille, Lille, France
Universite de Lille Nord de France, Lille, France

Université de Liège