Practical Considerations

in

Genome-Wide Association Interaction Studies

kristel.vansteen@ulg.ac.be

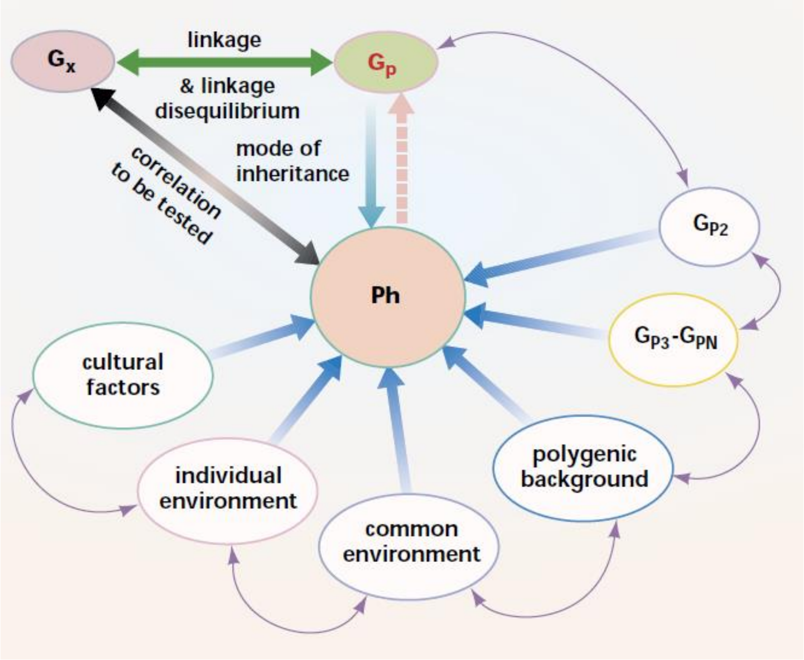
maggiew@cuhk.edu.hk

Outline

- Motive
- Opportunity
  - Data context
  - Disease context
- Means
  - GWAI protocol
  - W-test and MB-MDR
- Take-home messages

MOTIVE

The complexity of complex diseases



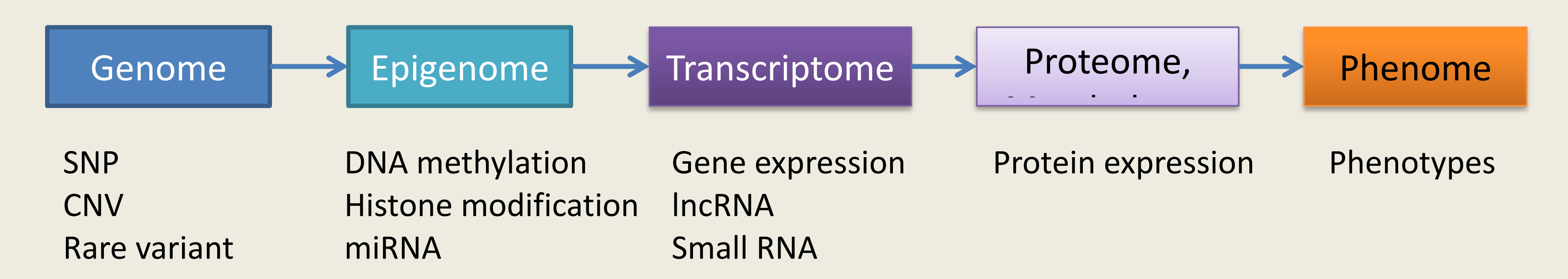
(Weiss and Terwilliger 2000)

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)

“Interactions”: A natural phenomenon

- There is a biological information flow from the genome to the ultimate cellular / disease phenotype
- The processes that are responsible for generating and modifying cellular components are generally dictated by molecular interactions:
  - protein–DNA “interactions” in the case of transcription, and
  - protein–protein “interactions”,
  - DNA-DNA “interactions”



Genome → Epigenome → Transcriptome → Proteome, → Phenome

SNP
CNV
Rare variant

DNA methylation
Histone modification
miRNA

Gene expression
lncRNA
Small RNA

Protein expression

Phenotypes

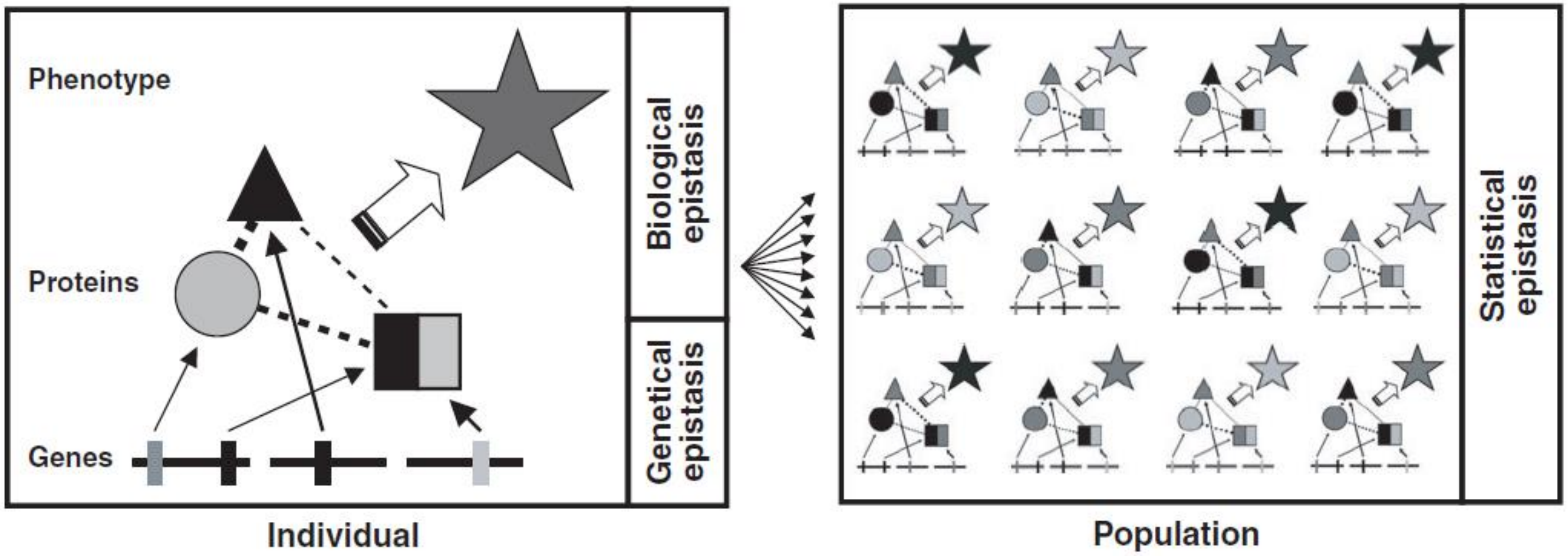
“Interactions” in humans

- 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).

(Moore et al. 2005)

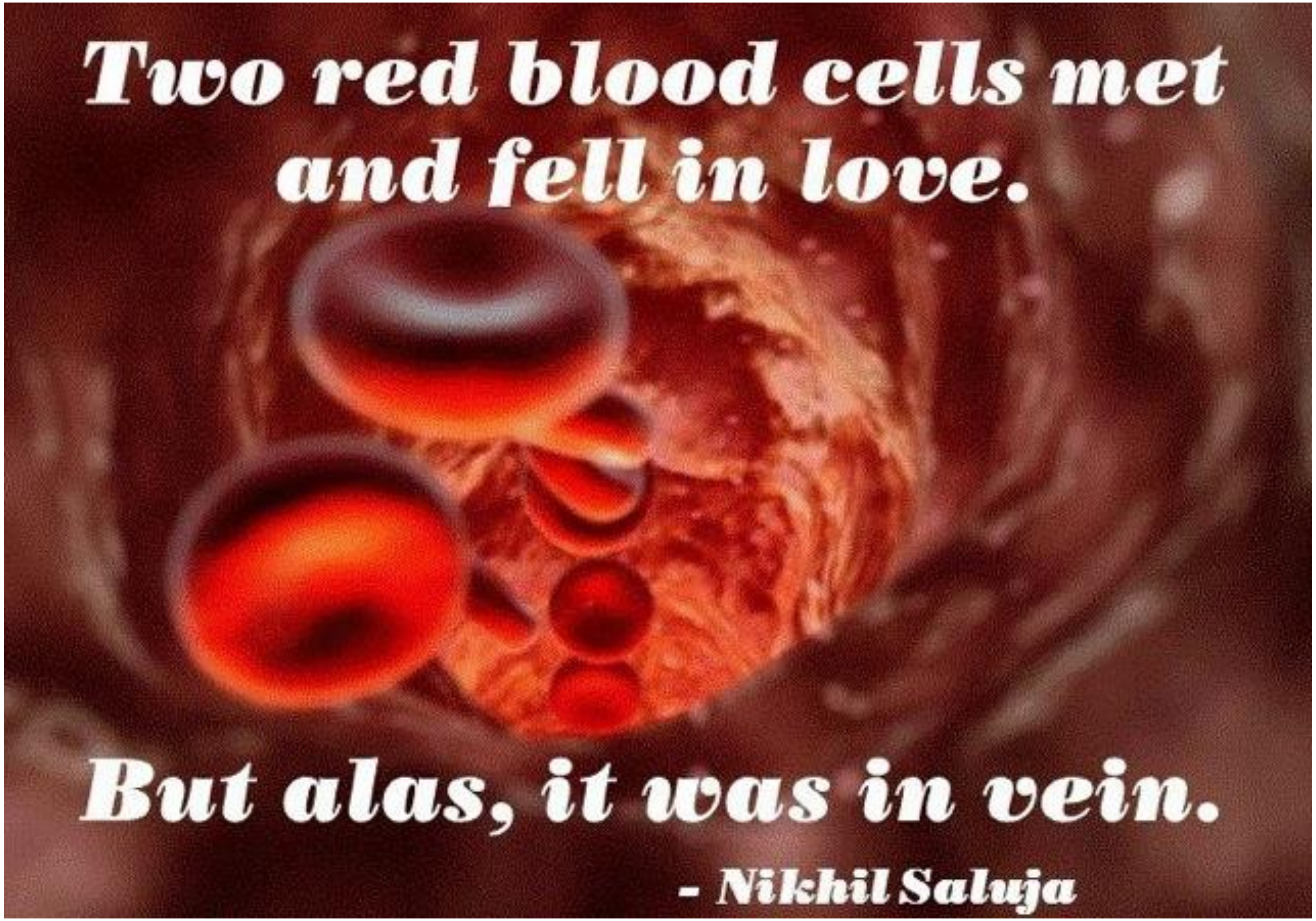
DNA x DNA “interactions”

- Two or more DNA variations may “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)

What’s in a name?

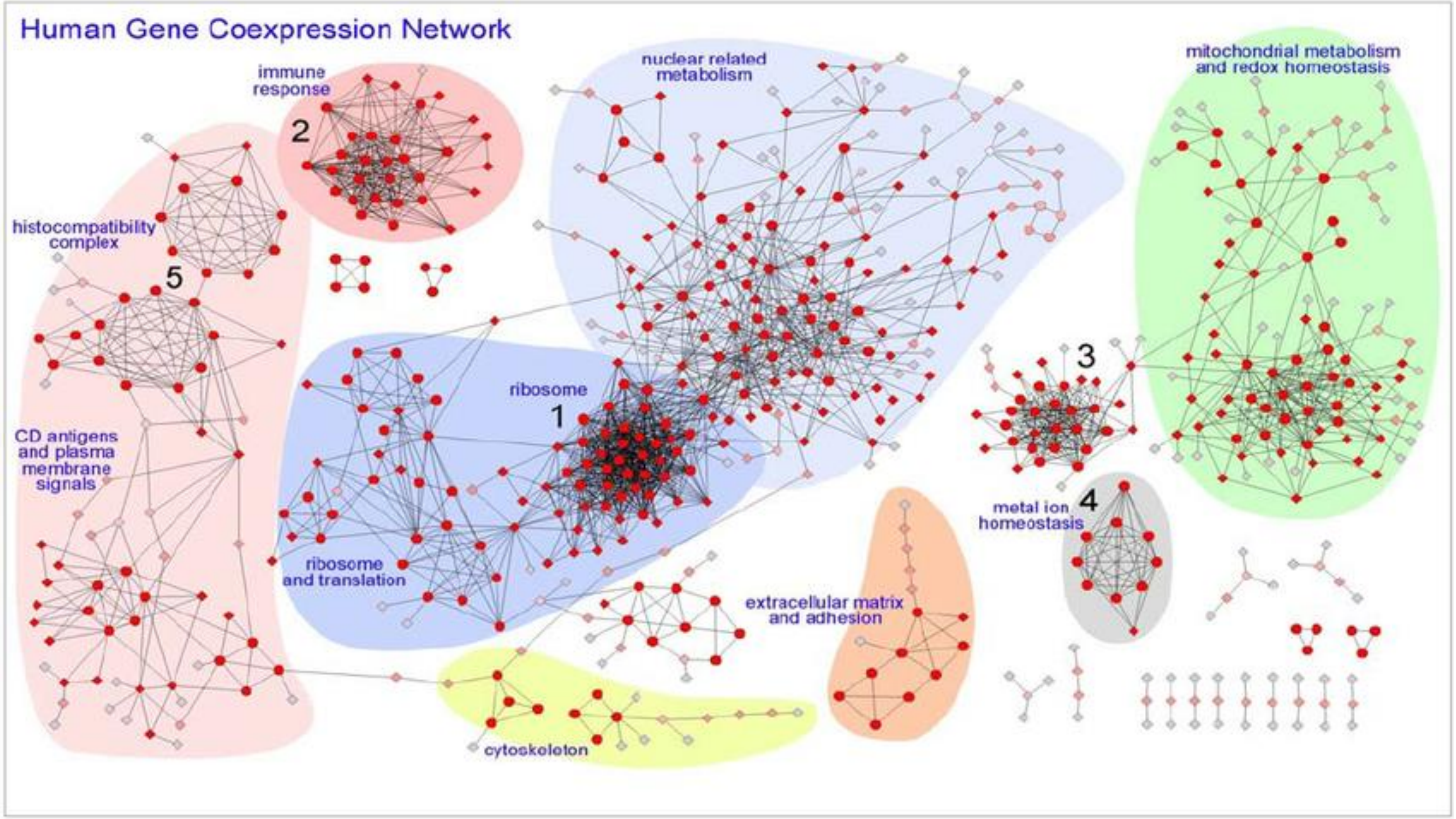


Biological interactions

- Biological interactions are the effects that the organisms in a community have on one another. In the natural world no organism exists in absolute isolation, and thus every organism must interact with the environment and other organisms.
- An organism's interactions with its environment are fundamental to the survival of that organism and the functioning of the ecosystem as a whole

(Elton 1968; Wikipedia)

Biological interactions – inference of gene-gene interactions using microarray data


Human Gene Coexpression Network
immune response
2
histocompatibility complex
5
CD antigens and plasma membrane signals
nuclear related metabolism
ribosome
1
ribosome and translation
cytoskeleton
extracellular matrix and adhesion
metal ion homeostasis
4
3
mitochondrial metabolism and redox homeostasis

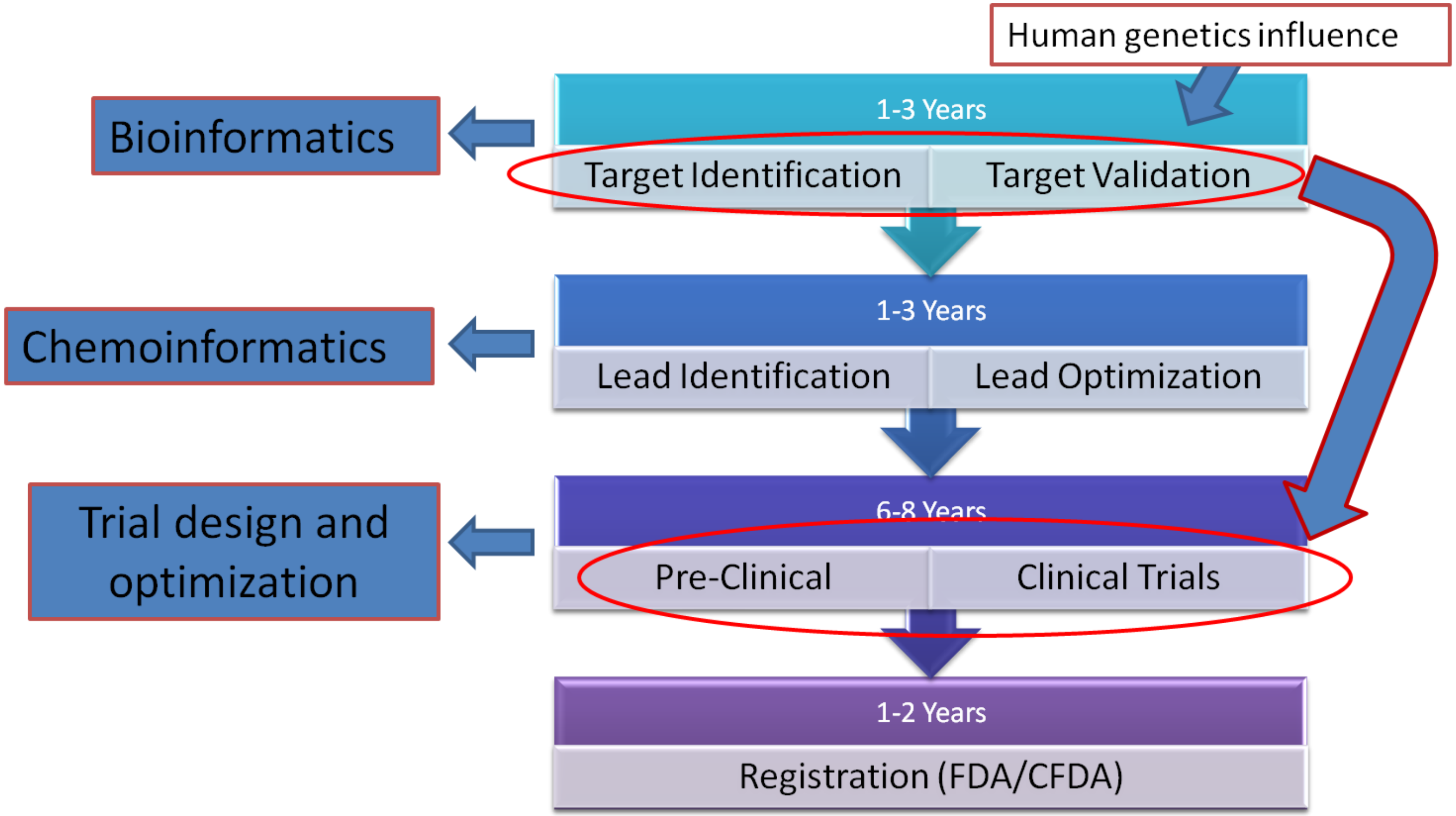
(Prieto et al. 2008)

Disease associated “target” identification

“We consider a target to be a molecular structure (chemically definable by at least a molecular mass) that will undergo a specific interaction with chemicals that we call drugs because they are administered to treat or diagnose a disease”

(Imming et al. 2006)

Drug discovery process



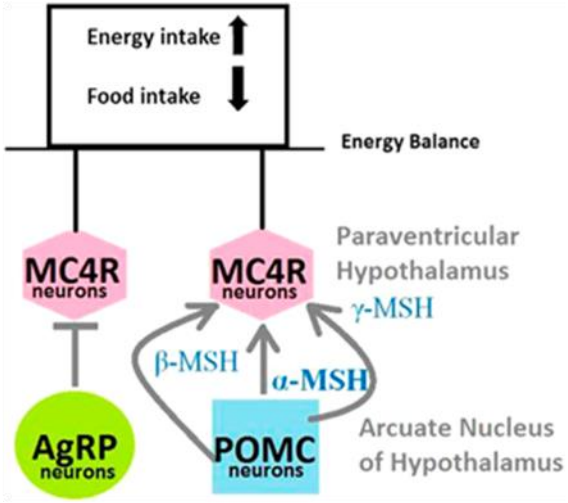
Disease associated “target” identification

- Past target finding process:
  - natural observations, existing knowledge, start from guessing, very low success rate to surpass Phase I and II clinical trials
- Human genetics improve the quality of drug targets sent to trials
  - Enabled by the advances in genetic sequencing technology
- Important to consider genetic interaction, even cross platforms in drug discovery process
- Statistical bioinformatics and machine learning methods that handles this “big data” are needed

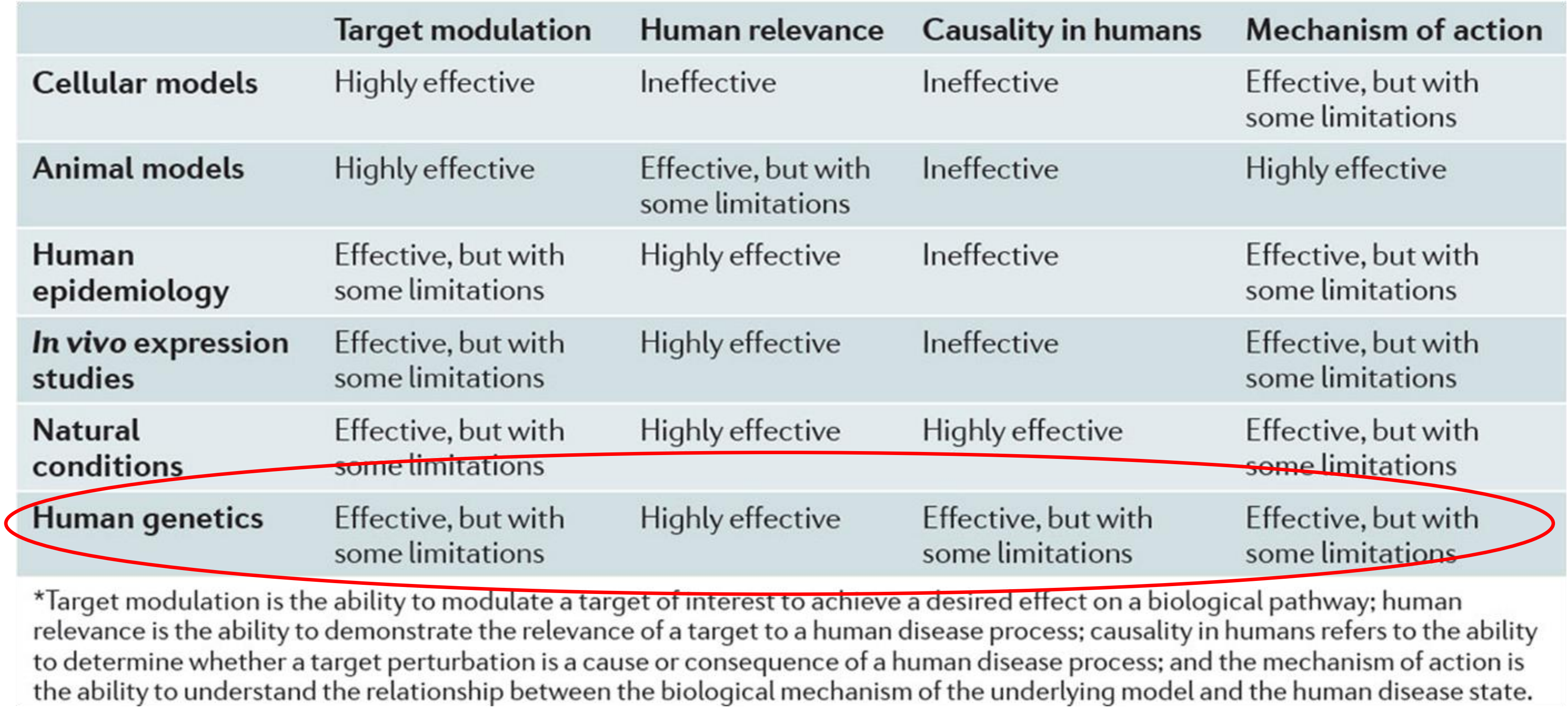
From genomics to drugs

- Example: melanocortin MC4 receptor – a star potential drug target for controlling obesity
- Identified in 1998 in French family genetic linkage study
- A G-protein coupled receptor responsive to melanocyte stimulating hormornes.
- 3-6% of extreme obesity is linked with it
- Major pharmaceutical companies have ongoing intense attempts to develop low molecular weight agonists for this receptor.

Jarl E. S. Wikberg and Felikss Mutulis (2008) Nature Review Drug Discovery



Characteristics of preclinical models for target validation



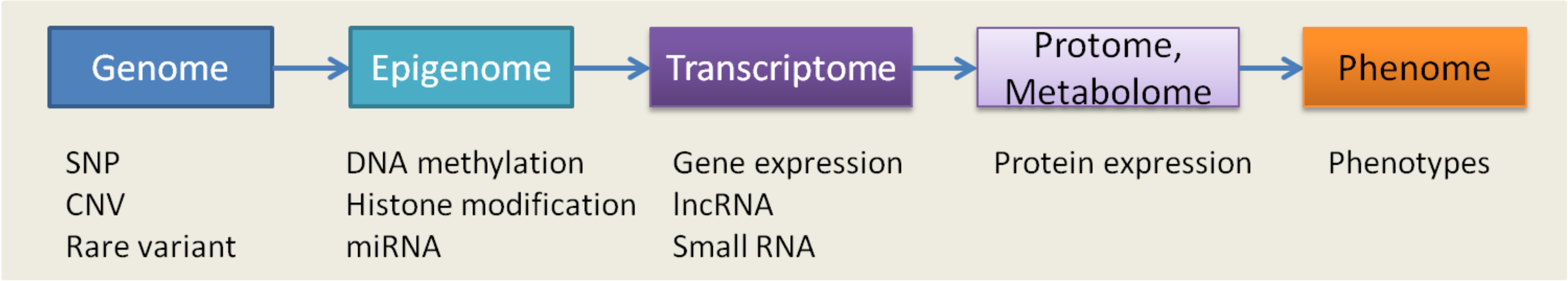
	Target modulation	Human relevance	Causality in humans	Mechanism of action
Cellular models	Highly effective	Ineffective	Ineffective	Effective, but with some limitations
Animal models	Highly effective	Effective, but with some limitations	Ineffective	Highly effective
Human epidemiology	Effective, but with some limitations	Highly effective	Ineffective	Effective, but with some limitations
***In vivo* expression studies**	Effective, but with some limitations	Highly effective	Ineffective	Effective, but with some limitations
Natural conditions	Effective, but with some limitations	Highly effective	Highly effective	Effective, but with some limitations
Human genetics	Effective, but with some limitations	Highly effective	Effective, but with some limitations	Effective, but with some limitations

*Target modulation is the ability to modulate a target of interest to achieve a desired effect on a biological pathway; human relevance is the ability to demonstrate the relevance of a target to a human disease process; causality in humans refers to the ability to determine whether a target perturbation is a cause or consequence of a human disease process; and the mechanism of action is the ability to understand the relationship between the biological mechanism of the underlying model and the human disease state.

Robert M. Plenge, Edward M. Scolnick & David Altshuler (2013) Nature Review Drug Discovery. “Validating therapeutic targets through human genetics”

OPPORTUNITY

Data context: Bioinformatics data availability

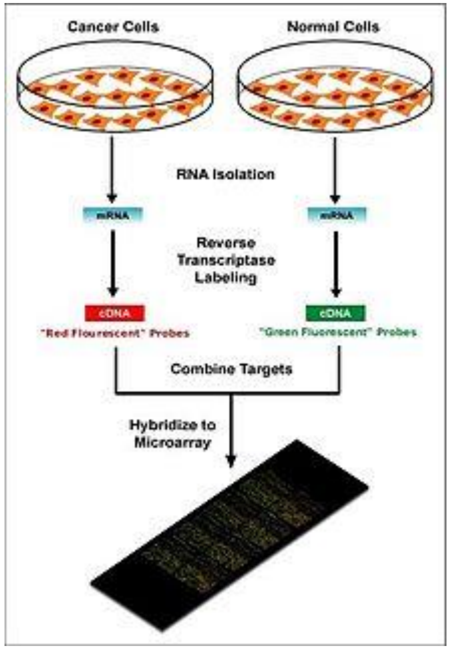


Genome	Epigenome	Transcriptome	Protome, Metabolome	Phenome
SNP	DNA methylation	Gene expression	Protein expression	Phenotypes
CNV	Histone modification	lncRNA		
Rare variant	miRNA	Small RNA		

- Problem complexity increases
- Rare variant data
- Interaction effects – high order
- Population stratification
- Confounding variables

Microarray data

- Motivation: much of the important information about a gene is not evident directly from its sequence
- For example, we want to know exactly when, where and how much the gene is expressed.
- It can be achieved by measuring the relative mRNA levels in cells, an indirect measure of protein levels.
- Microarray measures thousands of genes simultaneously; A sample can either be a single cell or a population of cells.



https://upload.wikimedia.org/wikipedia/en/thumb/c/c8/Microarray-schema.jpg/220px-Microarray-schema.jpg

GWAs data

- The value in this types of data is genotype: GG, GC, CC or AA, AT, TT
- We can analyze the data under recessive, dominant or co-dominant genetic model assumptions.
- Example: a SNP's genotype for 10 subjects:

  (AA, AA, AT, AT, AA, AA, TT, AA, AA, AA)
- There are 10 allele pairs, in which T appears least frequent, and is regarded as the minor allele.
- So the SNP can be coded as counts of the minor allele.

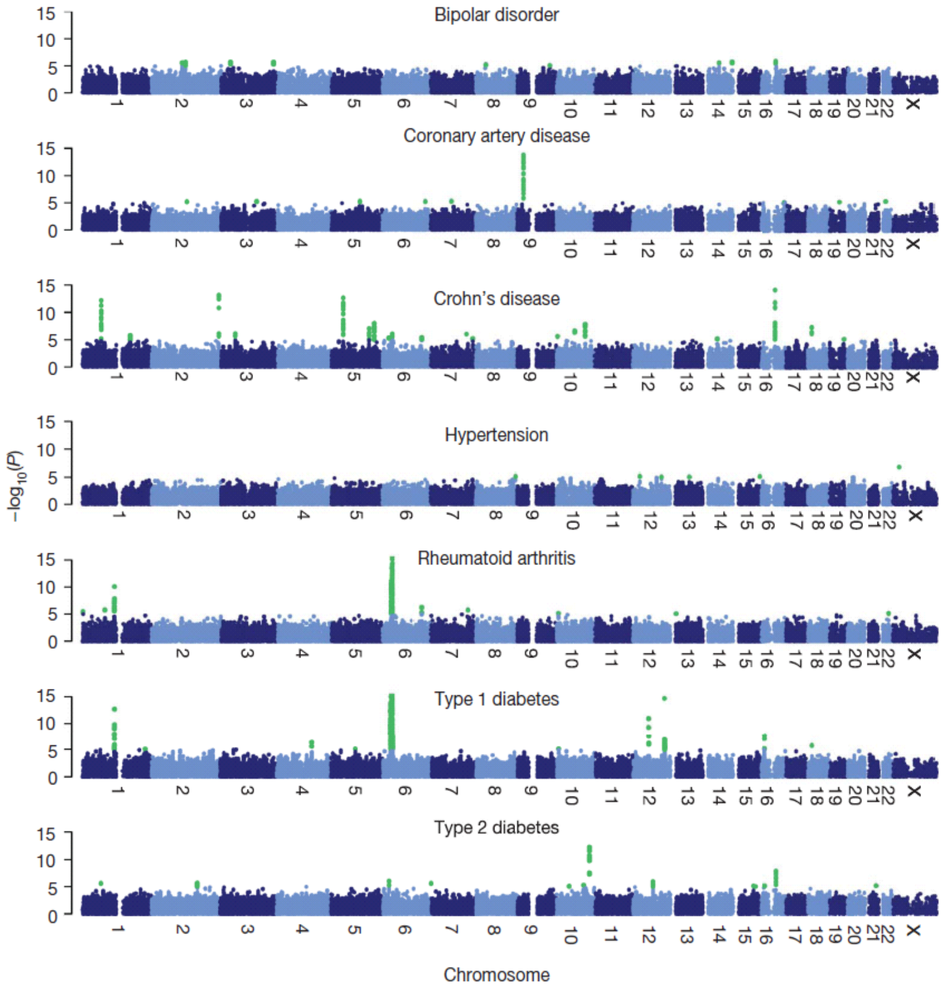
  (0, 0, 1, 1, 0, 0, 2, 0, 0, 0)
- Minor allele frequency (MAF) = sum of minor alleles/ (2*subjects) = 4/(2x10) = 0.2.

GWAs data

- GWAS targets to sequence the SNPs with MAF ranges >5%.
- The majority of SNPs in GWAS have MAF 10-40%.
- Next Generation Sequencing (NGS) targets to include SNPs with MAF <1%, called the rare variant, which composed of over 90% of the genome.

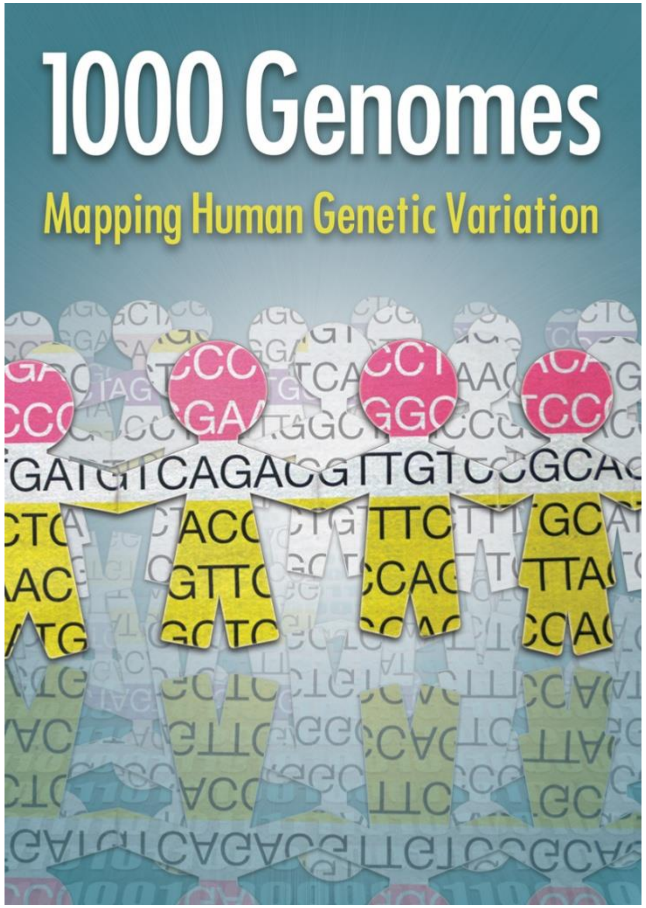
GWAs data

- In 2007, Welcome Trust Consortium sequenced 14,000 cases and 3,000 shared controls of 7 human diseases.
- The GWAS genome-wide significance level is around 5x10^{-7}.
- Green dot represent markers passed genome-wide significance level



(WTCCC 2007, Nature)

Next Generation Sequencing (NGS) data

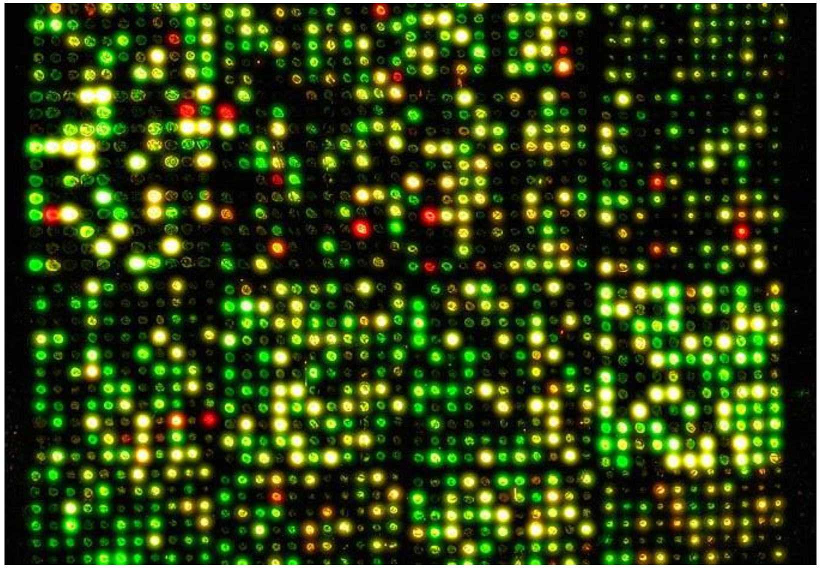


- 1000 Genome Project (EMBL-EBI):
  - 84.7 million SNPs
  - 3.6 million Short Indels and 60,000 structural variants
  - Phased genotypes of 2504 individuals up to 2015
  - From 26 populations
- Whole genome sequencing (WGS) for lower than 1000USD per genome
- Great opportunity to understand human common diseases, but also exerts great challenge for analytics.

http://www.1000genomes.org/, “A global reference for human genetic variation” Nature 526 68-74 2015

Scale matters

- Microarray data:
  - mRNA: > 20,000 probes
  - lncRNA: > 20,000 probes
  - miRNA: > 700 probe sets
- GWAS data
  - > 500,000 SNPs
- NGS data:
  - > 10 million SNPs
- When interaction effect is considered, the dimension goes exponentially up.
- For 20000 genes, the number of pairwise combinations is $2\text{x}10^8$!



http://www.dailymail.co.uk/sciencetech/article-2070248/The-human-genome-unravelled--literally-3D--ball-string-hold-key-work.html

Upscaling data and downscaling - the multiple testing problem?

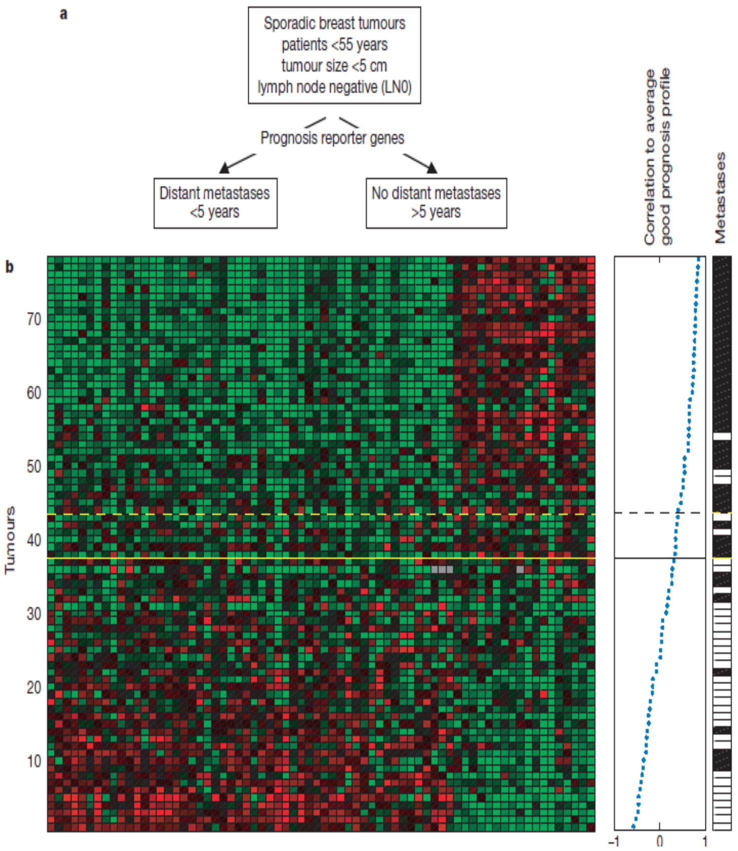
- Suppose we are comparing 100 genes simultaneously, and their values follow normal distribution. Under this null hypothesis, even in a sample without signal, if we set the significance level at 0.05, 5 genes will be declared significant.
- The more number of genes, the more serious is the issue.
- In real data where there are 25,000 genes, significance level at 0.05 will produce 1,250 genes with P-value < 0.05. If we expect the number of true positive is around 100, then 1150 markers will be false positives. The false positive ratio is 92%!

Multiple testing – controlling FWER

- Family-wise error rate (FWER): the probability of having one (or more) false positives in the global predicted set.
- Bonferroni Correction:
  - Set FWER < α, The total number of multiple tests is m, then individual test should declare significant using P-value < α/m.
  - For example, for 25000 genes, at α=0.05, the individual test significance level should be 0.05/25,000 = 2×10^{-6}.
  - Other correction methods, s.t. Sidak Procedure.

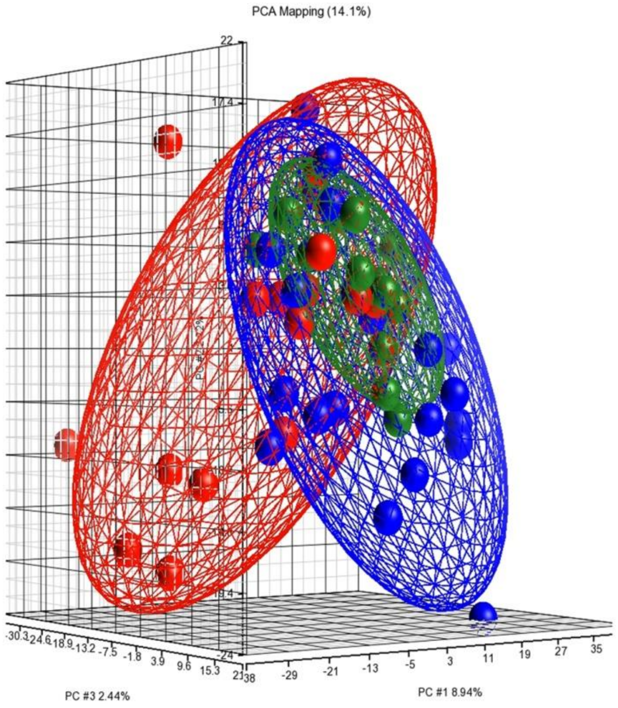
Methods for microarray data

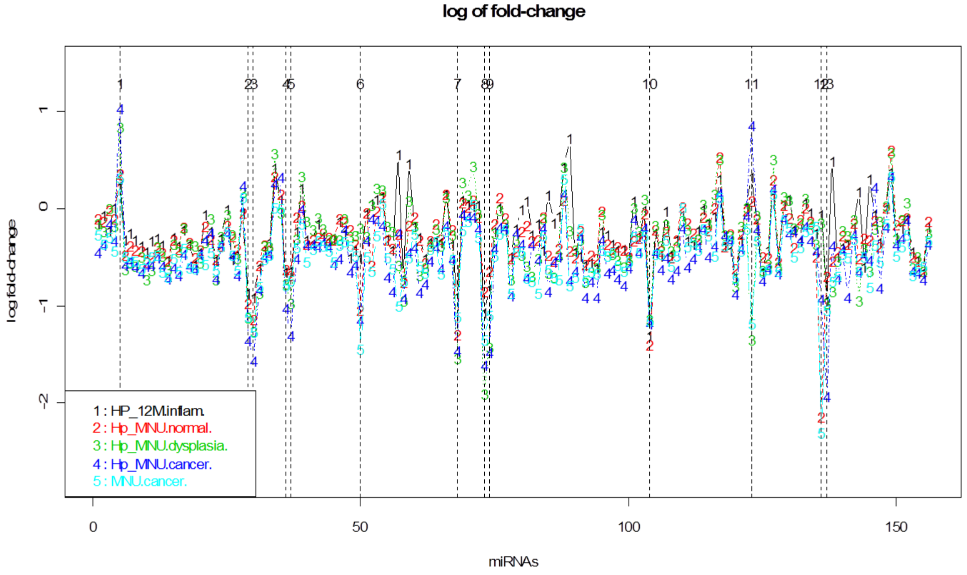
- *For disease classification*:
  - ○ Hierarchical Clustering
  - ○ Example: (van’t Veer et al 2002) A 70-gene profile selected from 25,000 genes, can classify breast cancer metastasis at 83% accuracy.
  - ○ Using gene expression profiling can reduce the unnecessary chemo- and hormornal therapies
  - ○ HC Measures the dissimilarties between two subjects, i and i', for the jth attributes: $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$
  - ○ The distance can be the correlation between subjects



Methods for microarray data

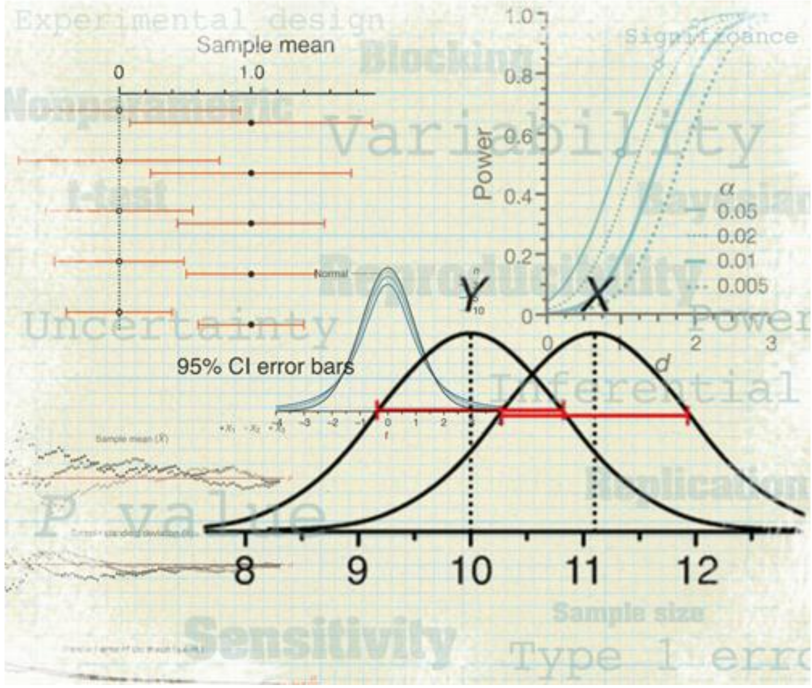
- *For visualizing data internal structure*
  - Principle component analysis (PCA)
- *For variable selections* to understand biological mechanism:
  - Fold Change
  - Linear regressions
  - Penalized generalized models





Methods for GWAs data

- Discrete Phenotypes
  - Chi-squared test
  - Logistic regression (PLINK incorporated)
  - Linkage Disequilibrium (LD) based tests
  - Multifactor Dimensionality Reduction (MDR)
- Continuous Phenotypes
  - Linear regression
  - F-test

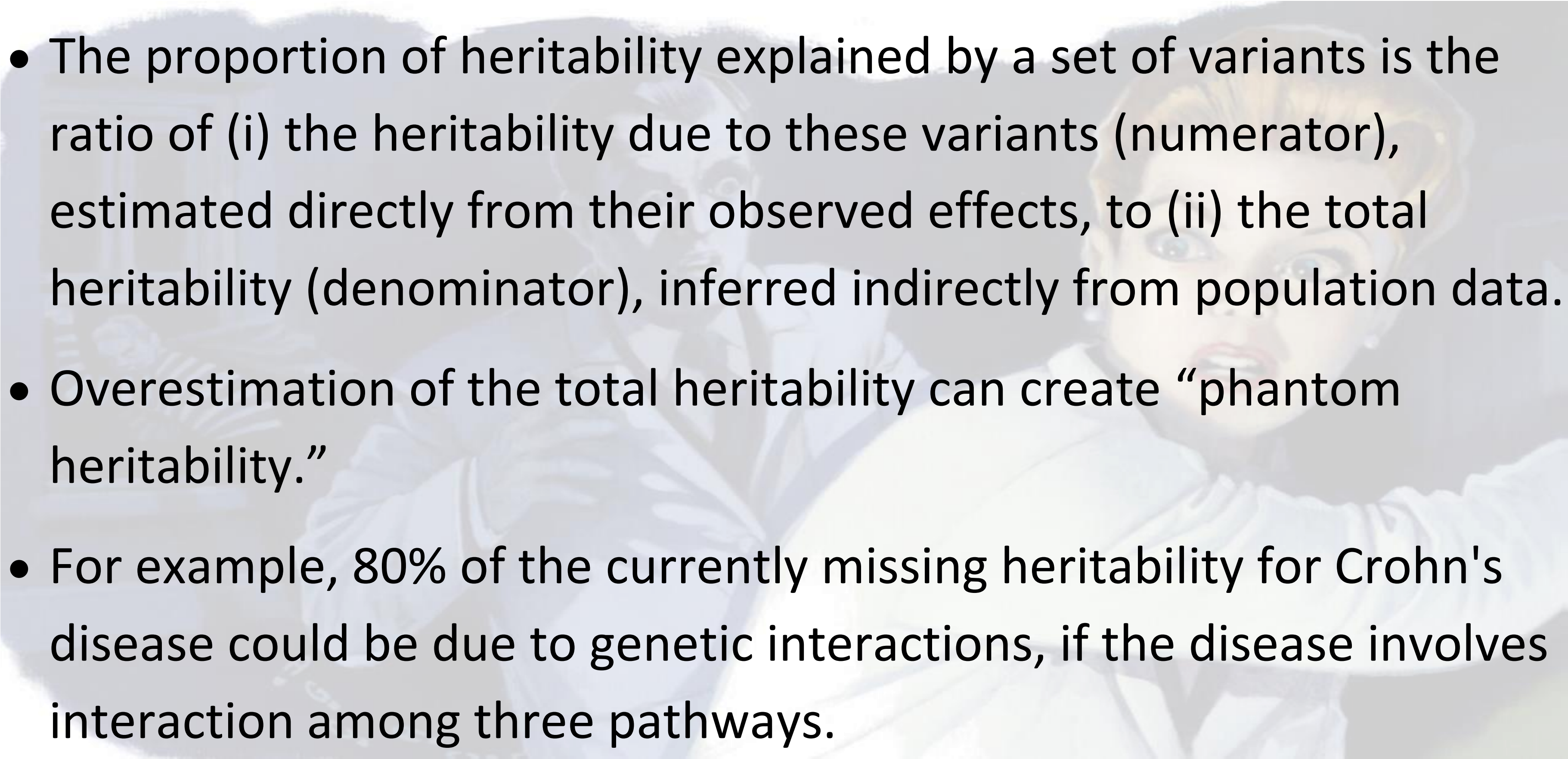


Methods for NGS data (Moutsianas and Agarwala et al 2015 PloS Genetics)

- Unidirectional gene-based tests:
    - Weighted-Sum Statistics (WSS) (Madsen & Browning 2009)
    - Combined Multivariate and Collapsing (CMC) (Liu and Leal 2008)
    - Variable threshold (VT) (Price et al. 2010)
    - Kernel Based Adaptive Cluster (KBAC) (Liu & Leal 2010)
- Bidirectional variance-component gene-based methods:
    - Sequential Kernel Association Test (SKAT) (Wu et al 2011)
- Combined tests
    - SKAT-O (Lee et al 2012)
    - Mixed Effects Score Test (MiST) (Sun et al 2013)

"... however, the power to detect even loci of relatively large effect is very low ..."

Unexplained 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.
- Overestimation of the total heritability can create "phantom heritability."
- 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.

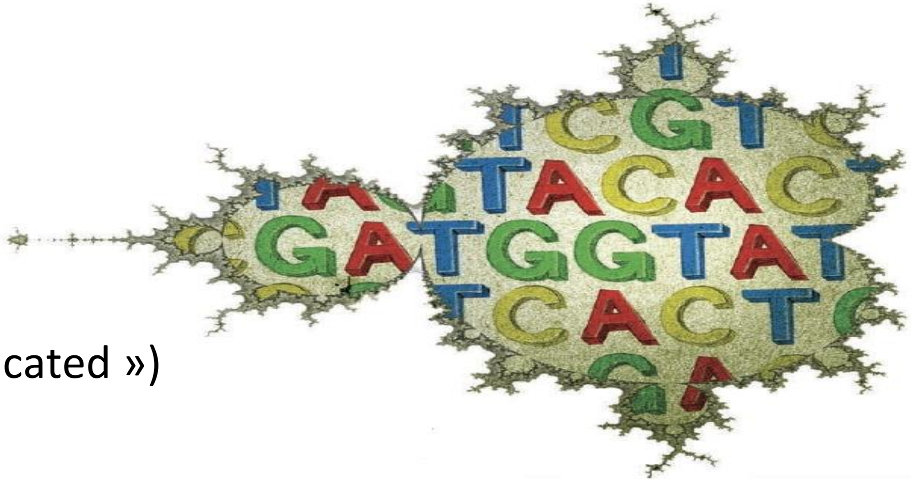
(Maher 2008, Zuk et al. 2012)

Unexplained heritability

Explanation	Rationale	Comments
Overestimated heritability estimates	These estimates are typically performed in the absence of gene-gene or gene-environment interactions (Young et al. 2014)	Limiting pathway modeling suggests that epistasis could account for missing heritability in complex diseases (Zuk et al. 2012)
Common genetic variants	More common variants are likely to be found in GWAs with larger sample sizes (drawback: more is less?)	Effect sizes of known GWAs loci may be underestimated since functional variants have often not yet been found
Rare genetic variants	Resequencing studies (e.g., WES) could identify rare genetic determinants of large effect size (Zuk et al. 2014)	Limited evidence for rare variants of major effect in complex diseases accounting for large amount of genetic variation – most rare variants analysis methods currently suffer from increased type I errors (Derkach et al. 2014)

Explanation	Rationale	Comments
Phenotypic and genetic heterogeneity	Most complex diseases are like syndromes with multiple potentially overlapping disease subtypes	Improvements in phenotyping of complex diseases will be required to understand genetic architecture.
Interactions	Gene-gene and gene-environment interactions are likely to be important for complex diseases (Moore et al 2005)	Limited evidence for statistical interactions in complex diseases; network-based approaches may be helpful (Hu et al. 2011)

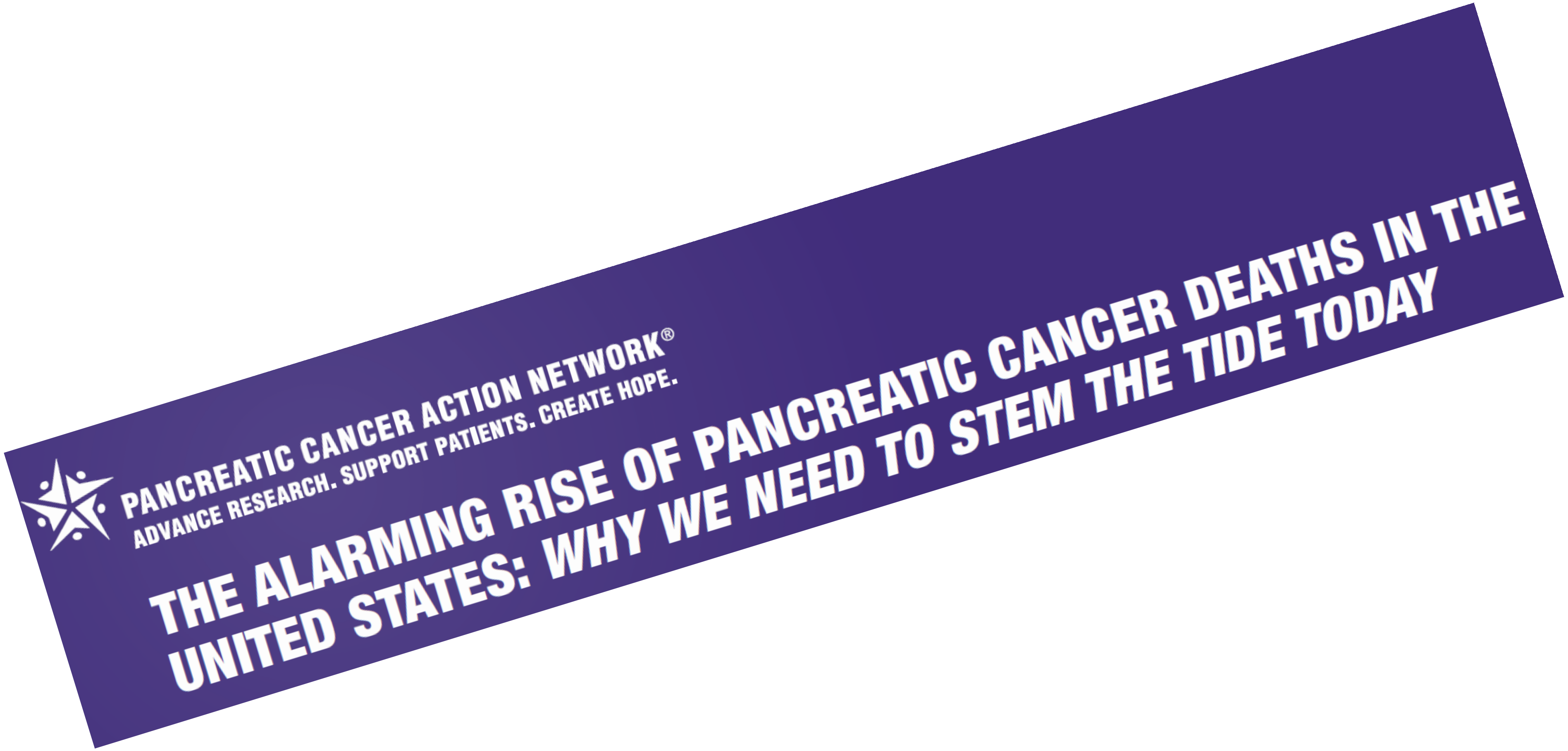
(adapted from Silverman et al. 2012)



(Hayden 2010

« Life is Complicated »)

Disease context: complex “complex diseases”

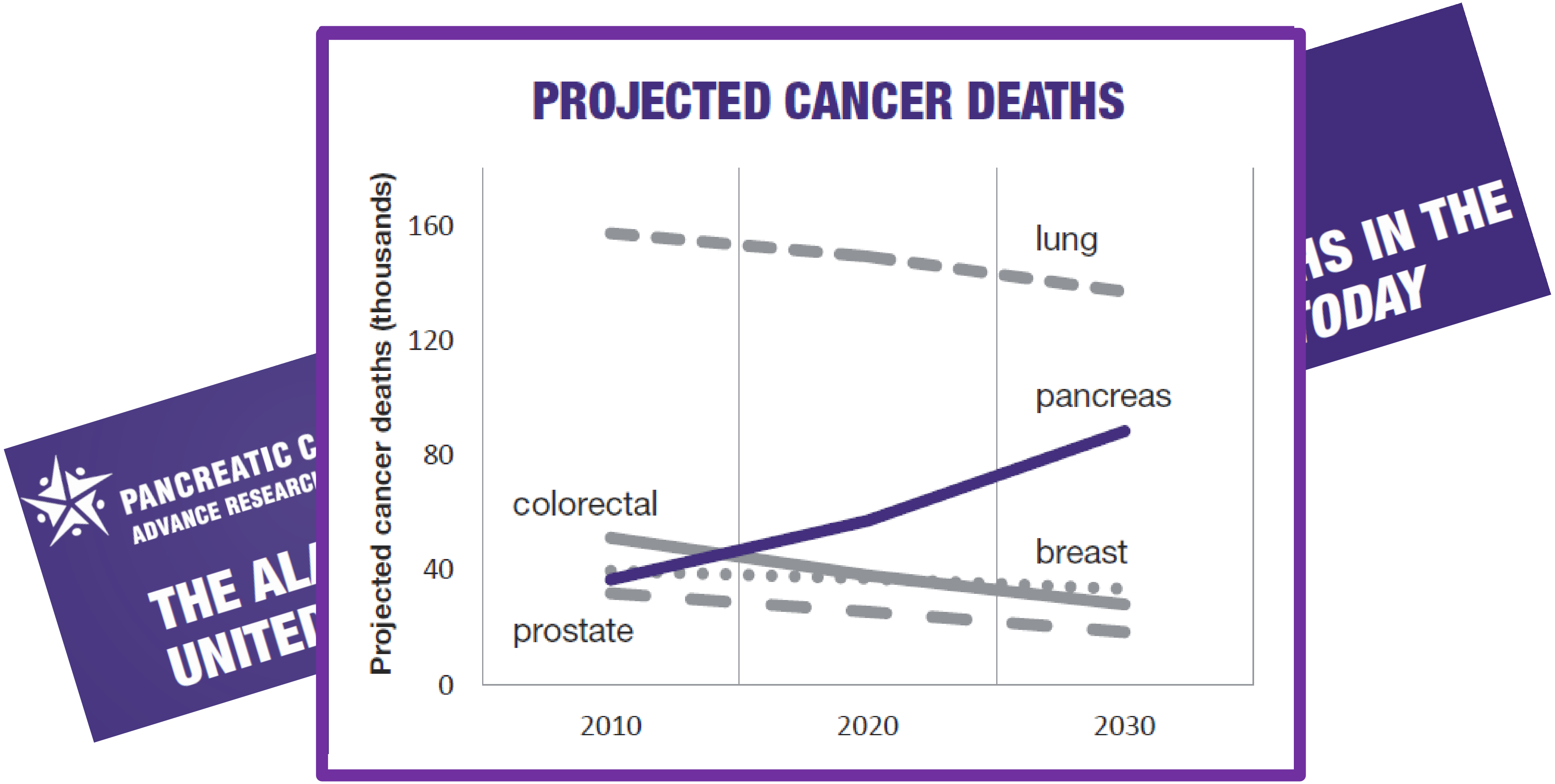


PANCREATIC CANCER ACTION NETWORK®

ADVANCE RESEARCH. SUPPORT PATIENTS. CREATE HOPE.

THE ALARMING RISE OF PANCREATIC CANCER DEATHS IN THE UNITED STATES: WHY WE NEED TO STEM THE TIDE TODAY

Disease context: complex “complex diseases”



Disease context: complex "complex diseases"

Did you know that **Pancreatic Cancer** has the lowest survival rate of all cancers?

WG2 : « integration of omics data » (work group leader: K Van Steen)

http://eupancreas.com

Addressing complexity in "complex diseases" - pancreatic cancer

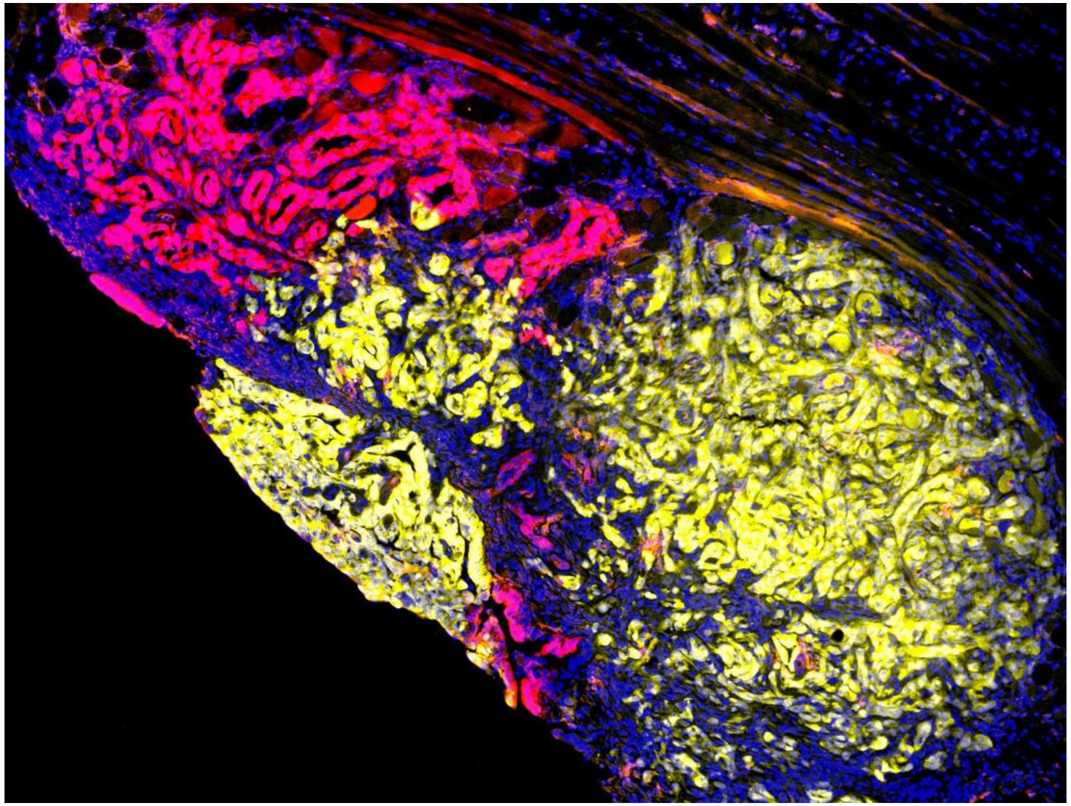
- 5-year survival rate below 5%
- Most cases are diagnosed at an advanced stage, making patients poor candidates for surgical treatment
- Major reasons for the dismal prognosis: lack of early appreciable symptoms, tendency of rapid local or distant metastasis, and intrinsic resistance to conventional chemotherapeutics

"Because effective systemic therapy capable of controlling the aggressive pancreatic cancer biology is currently lacking, the need for a better understanding of detailed mechanisms underlying pancreatic cancer development and progression is **URGENT**"

(Xie and Xie 2015)

Examples of interactions in pancreatic cancer

Cell-cell interactions



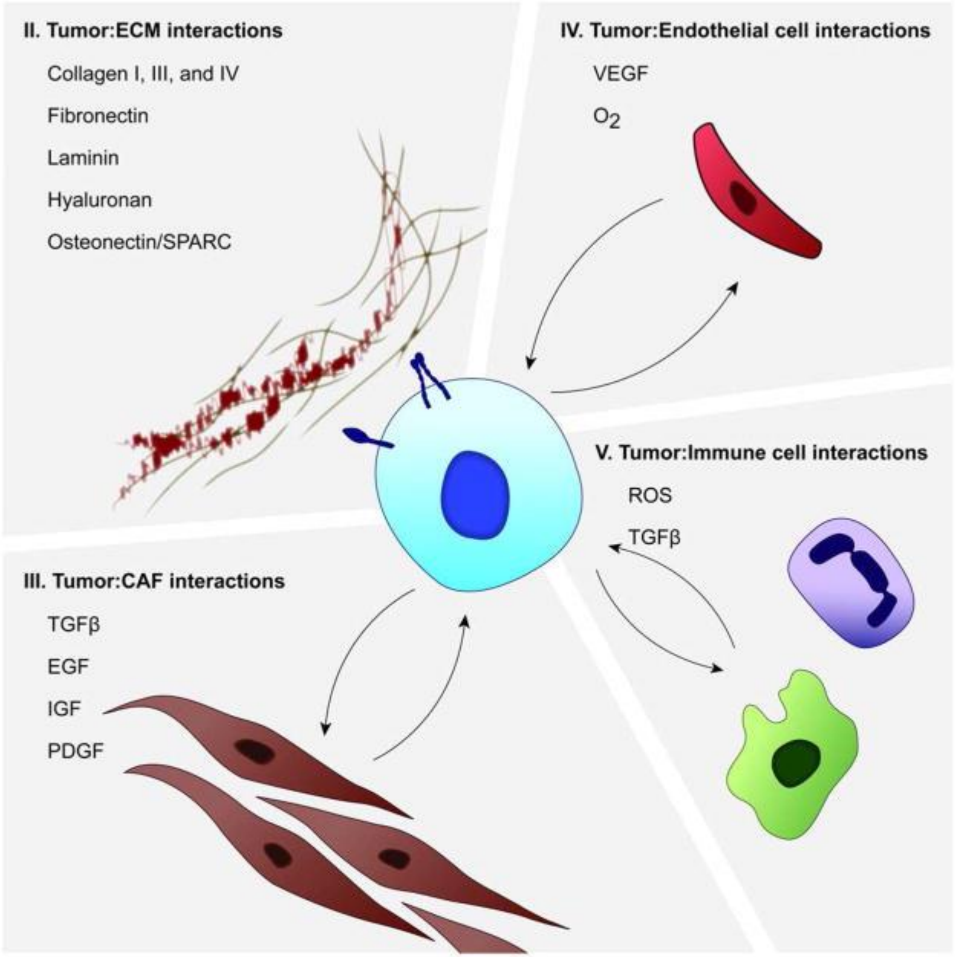
http://www.uphs.upenn.edu/news/News_Releases/2015/08

Multi-colored metastasis in the peritoneal lining of the abdomen comprised of red and yellow fluorescent cells demonstrating that pancreatic cancer spreads through interactions between different groups of cells.

Tumor-stromal interactions

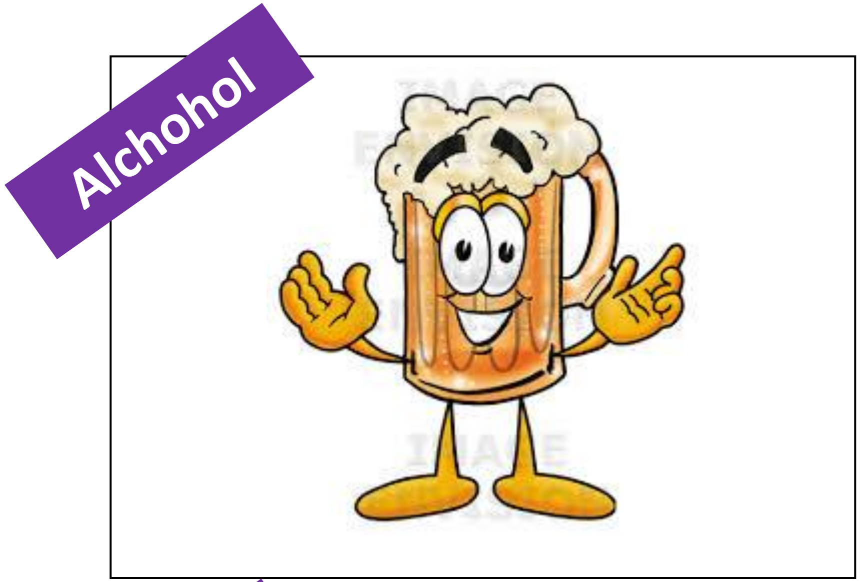
- Treatments focusing on pancreatic cancer cells alone have failed to significantly improve patient outcome over many decades
- Research efforts have now moved to understanding the pathophysiology of the stromal reaction and its role in cancer progression

(Whatcott et al. 2014)


II. Tumor:ECM interactions
Collagen I, III, and IV
Fibronectin
Laminin
Hyaluronan
Osteonectin/SPARC
IV. Tumor:Endothelial cell interactions
VEGF
O_2
V. Tumor:Immune cell interactions
ROS
TGFβ
III. Tumor:CAF interactions
TGFβ
EGF
IGF
PDGF

Gene-environment interactions

(Jansen et al. 2015)

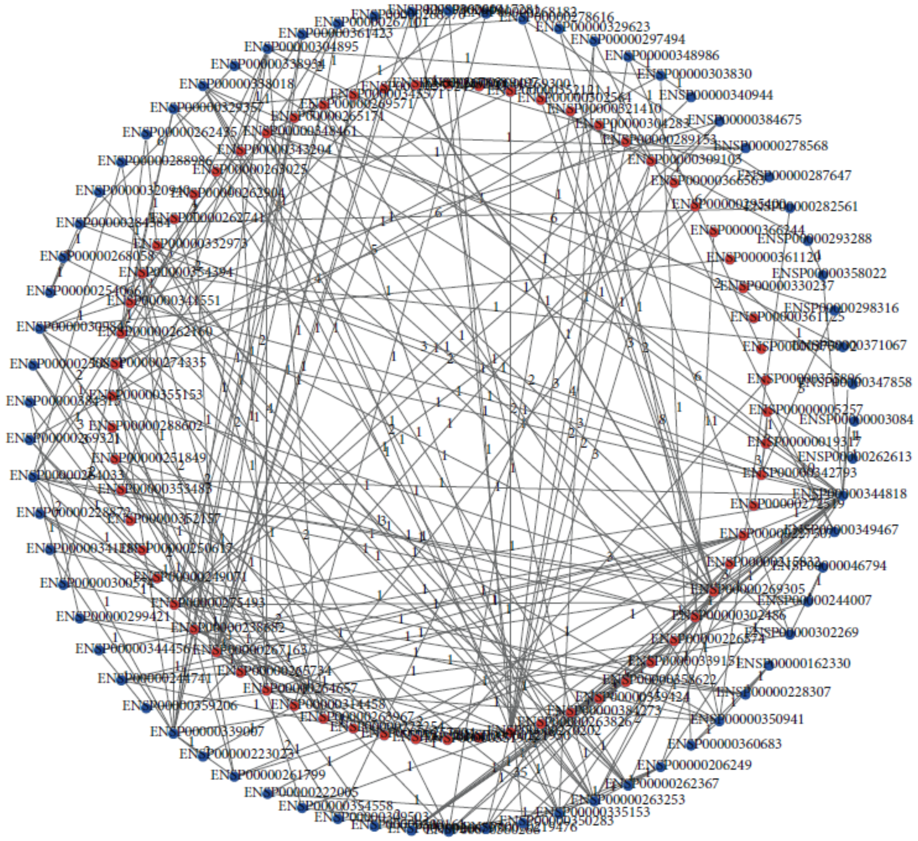


Protein-protein interactions

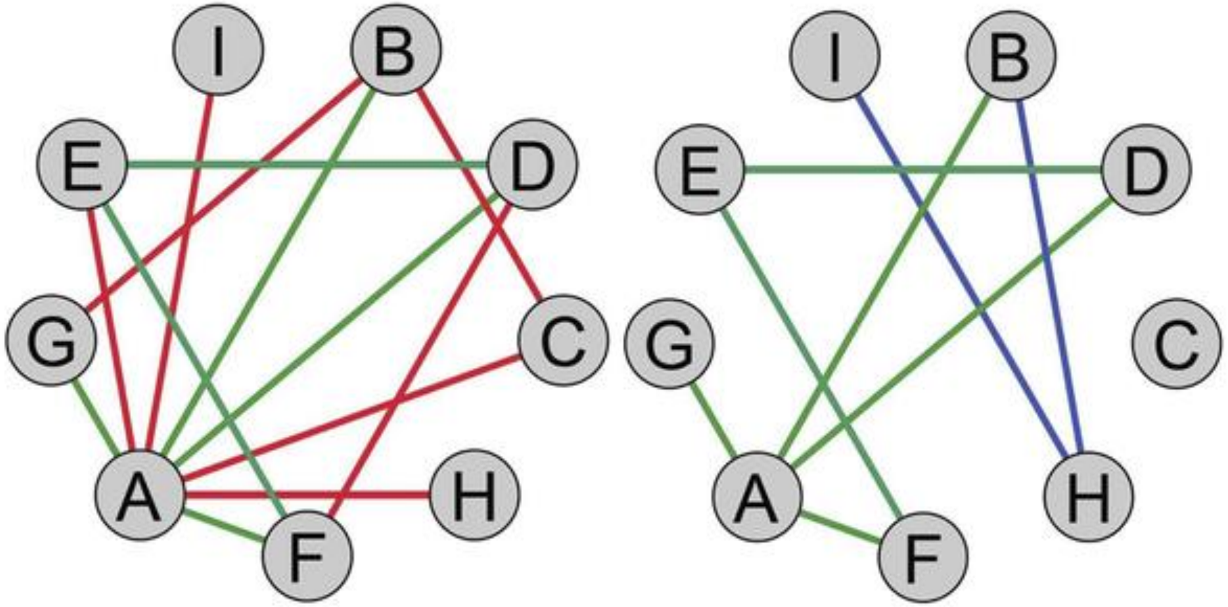
(Yuan et al. 2015)

A graph consisting of 2,080 shortest paths:

- The nodes on the inner circle (red nodes) represent 65 PC-related genes.
- The nodes on the outer circle (blue nodes) represent 69 shortest path genes.
- The numbers on the edges represent the weights of the edges.



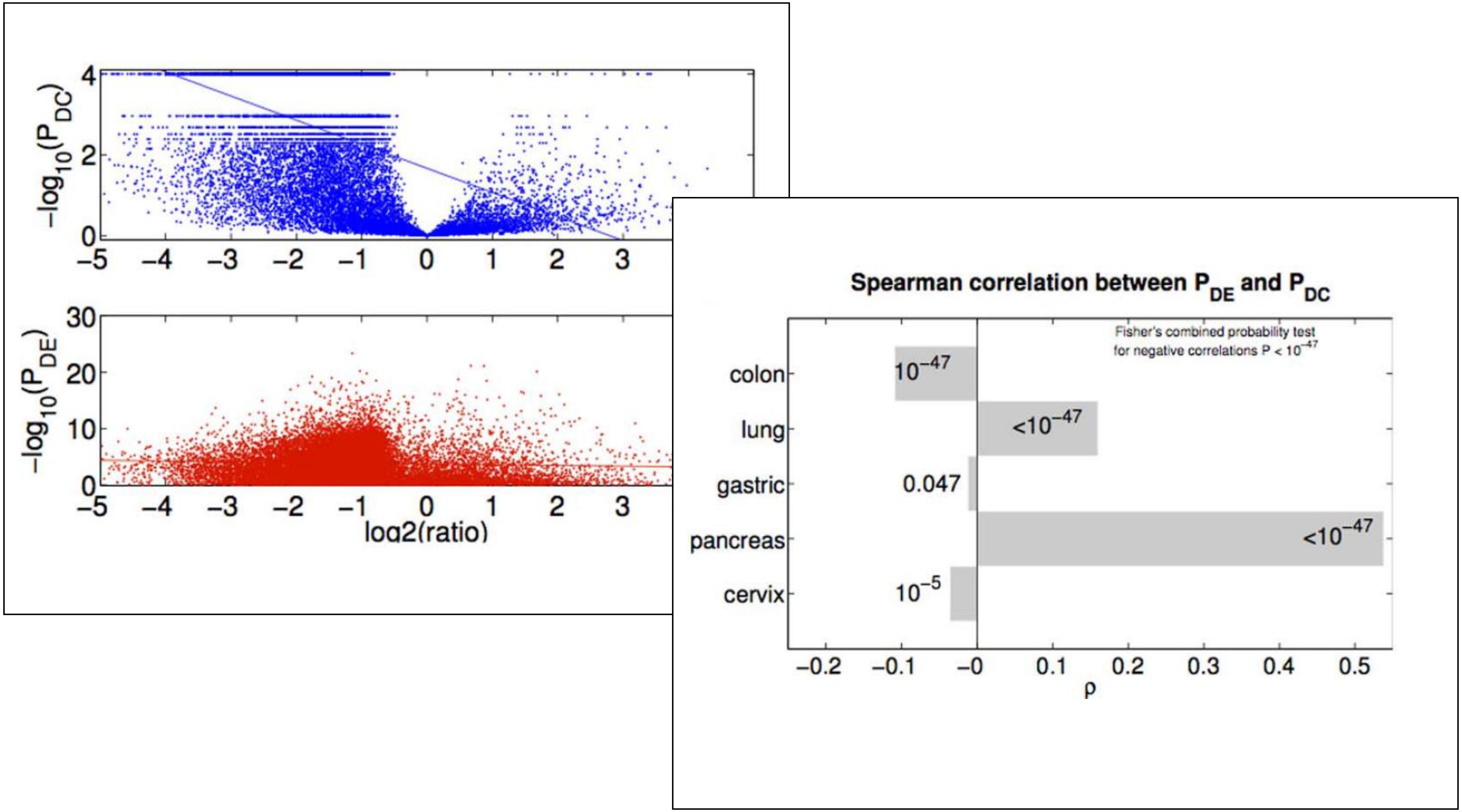
Gene-coexpression networks((Anglani et al. 2014)



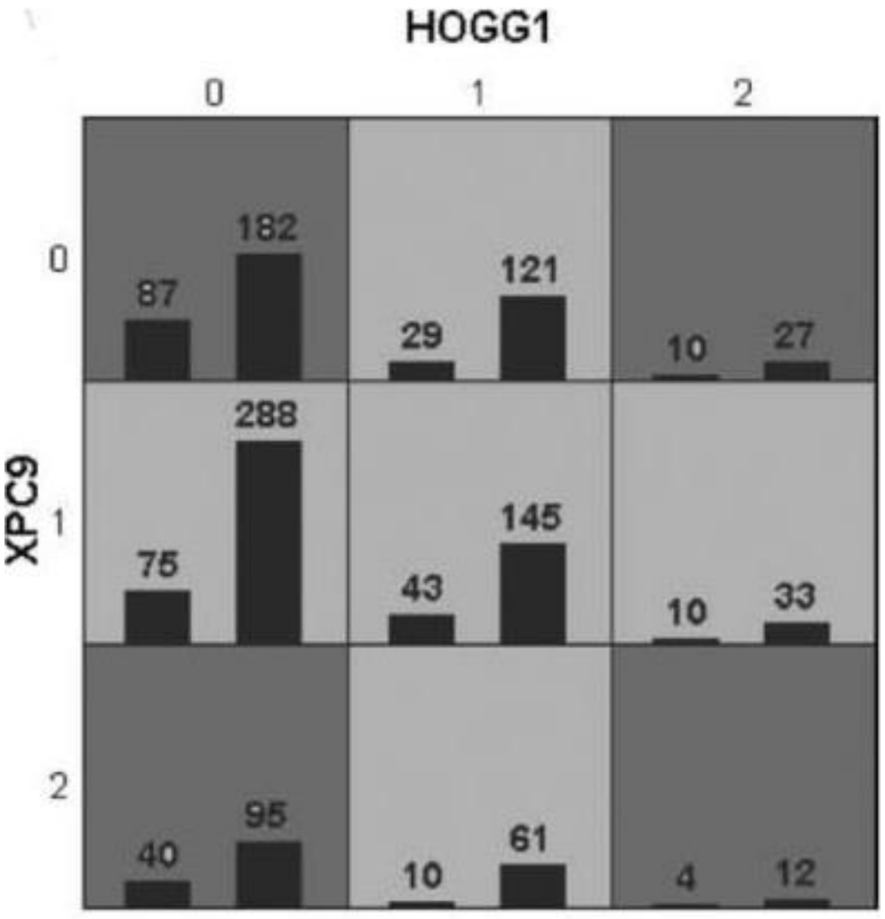
- Healthy condition on the left and disease-affected tissue on the right. Green links remain unchanged in the two phenotypes
- Red connections are loss from healthy to cancer network
- Blue edges are novel connections in the cancer tissue

Gene co-expression networks

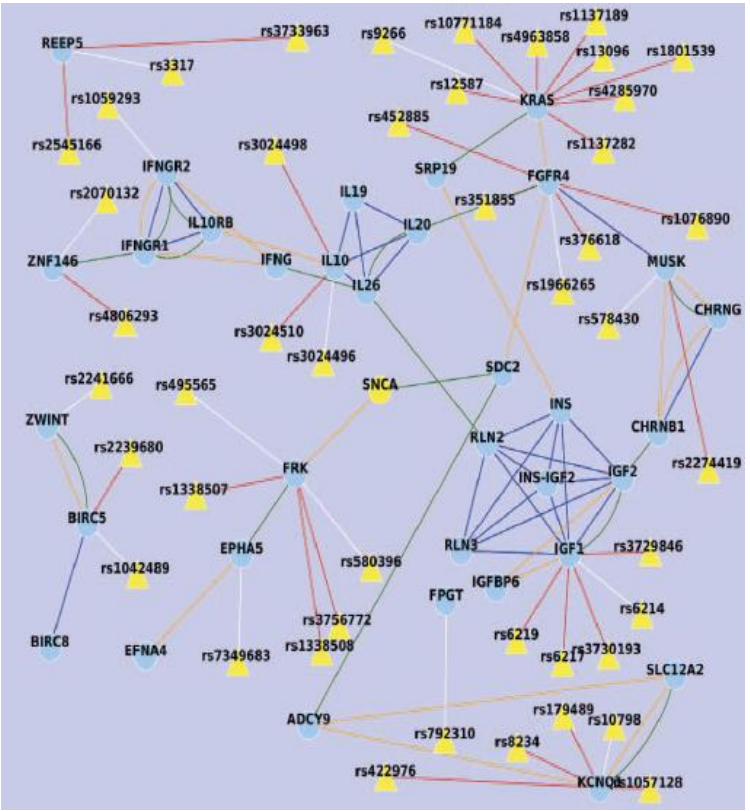
(Anglani et al. 2014)



Gene-gene interactions using SNPs



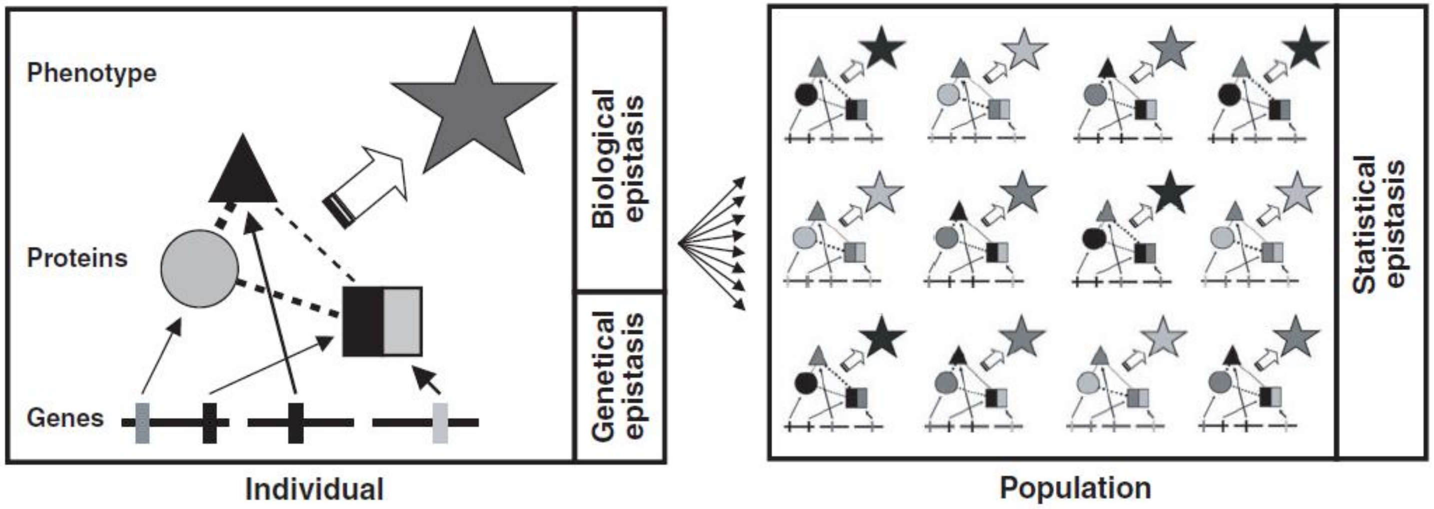
(Duell et al. 2008)



(Joseph et al. 2015)

Focus on Epistasis

- The original definition (**driven by biology**) refers to a variant or allele at one locus preventing the variant at another locus from manifesting its effect (William Bateson 1861-1926).
- A later definition of epistasis (**driven by statistics**) is expressed in terms of deviations from a model of additive multiple effects (Ronald Fisher 1890-1962).

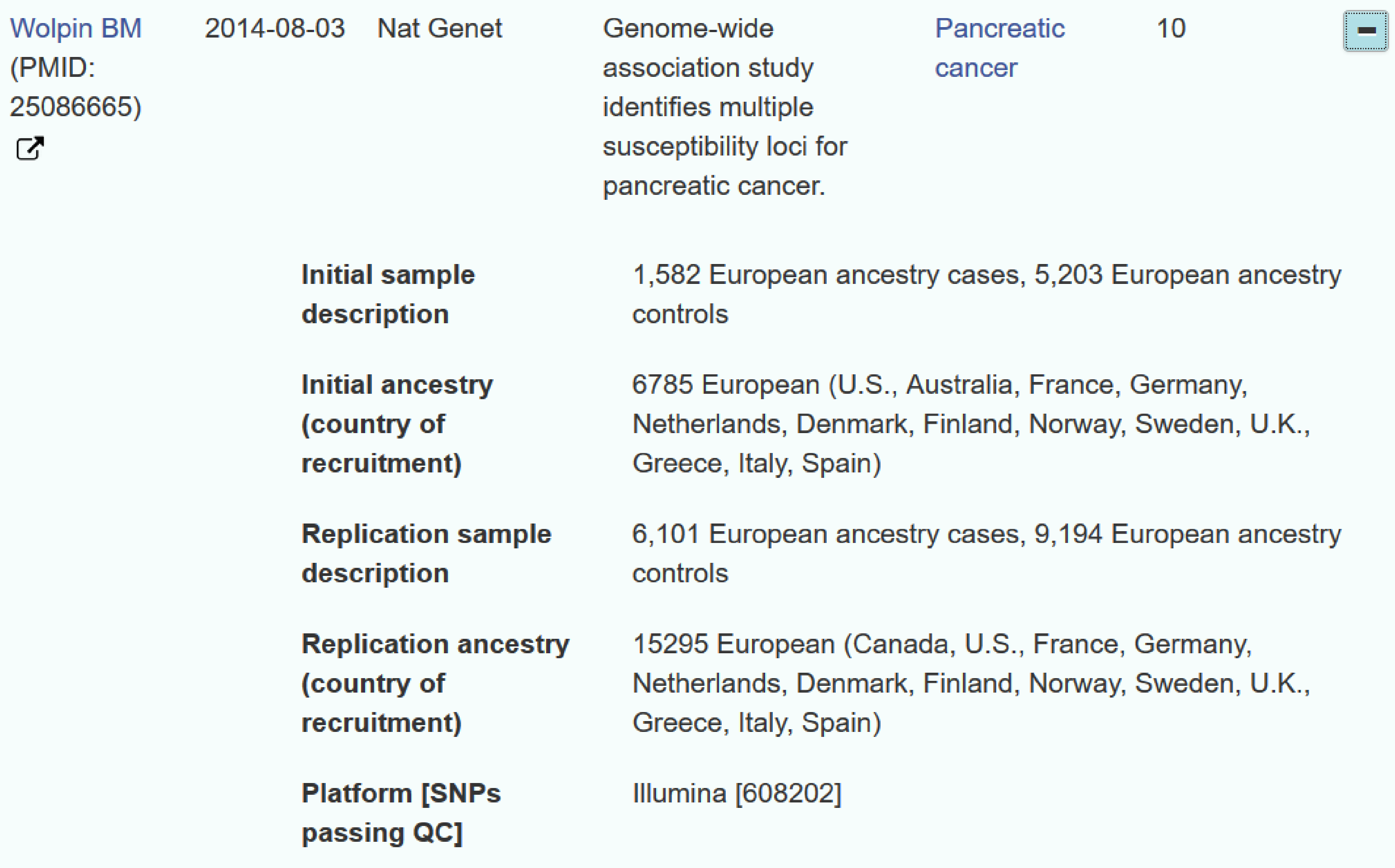


(Moore 2005)

Epistasis appearance versus detection

- Examples of DNA-DNA interactions from model organisms (Carlborg and Haley 2004):
  - Epistatic QTLs without individual effects have been found in various organisms, such as birds, mammals, Drosophila melanogaster and plants.
  - Other similar studies have reported only low levels of epistasis or no epistasis at all, despite being thorough and involving large sample sizes.
- 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…

GWAs Catalogue – “Pancreas Cancer”



Wolpin BM (PMID: 25086665)	2014-08-03	Nat Genet	Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer.	Pancreatic cancer	10

Initial sample description	1,582 European ancestry cases, 5,203 European ancestry controls
Initial ancestry (country of recruitment)	6785 European (U.S., Australia, France, Germany, Netherlands, Denmark, Finland, Norway, Sweden, U.K., Greece, Italy, Spain)
Replication sample description	6,101 European ancestry cases, 9,194 European ancestry controls
Replication ancestry (country of recruitment)	15295 European (Canada, U.S., France, Germany, Netherlands, Denmark, Finland, Norway, Sweden, U.K., Greece, Italy, Spain)
Platform [SNPs passing QC]	Illumina [608202]

(http://www.ebi.ac.uk/gwas/search?query=pancreas%20cancer#study)

MEANS

Although there is growing appreciation that attempting to map genetic interactions in humans may be a fruitful endeavor, there is no consensus as to the best strategy for their detection, particularly in the case of genome-wide association where the number of potential comparisons is enormous

(Evans et al. 2006)

One popular method singled out

- Interactions are commonly assessed by regressing on the product between both ‘exposures’ (genes / environment)

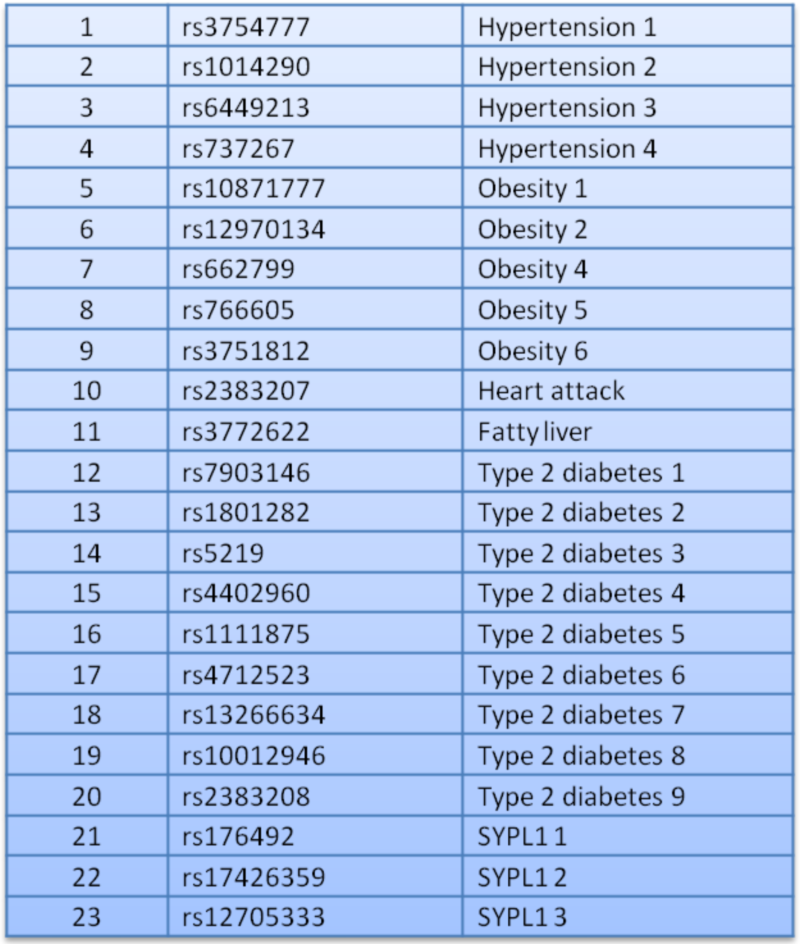
$$E[Y|G_1, G_2, X) = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_X X + \beta G_1 G_2$$

with X a possibly high-dimensional collection of confounders.

- PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/

Example – interaction effect using candidate SNPs and continuous phenotypes

- 23 candidate SNPs are selected from the literature
- Number of 2-way interaction terms: C(23, 2) = 253
- Total number of variables: 23 + 253 = 276.
- Total number of subjects: 115
- Number of subjects < number of variables



1	rs3754777	Hypertension 1
2	rs1014290	Hypertension 2
3	rs6449213	Hypertension 3
4	rs737267	Hypertension 4
5	rs10871777	Obesity 1
6	rs12970134	Obesity 2
7	rs662799	Obesity 4
8	rs766605	Obesity 5
9	rs3751812	Obesity 6
10	rs2383207	Heart attack
11	rs3772622	Fatty liver
12	rs7903146	Type 2 diabetes 1
13	rs1801282	Type 2 diabetes 2
14	rs5219	Type 2 diabetes 3
15	rs4402960	Type 2 diabetes 4
16	rs1111875	Type 2 diabetes 5
17	rs4712523	Type 2 diabetes 6
18	rs13266634	Type 2 diabetes 7
19	rs10012946	Type 2 diabetes 8
20	rs2383208	Type 2 diabetes 9
21	rs176492	SYPL1 1
22	rs17426359	SYPL1 2
23	rs12705333	SYPL1 3

The Lasso – Penalized regression (generalized linear models)

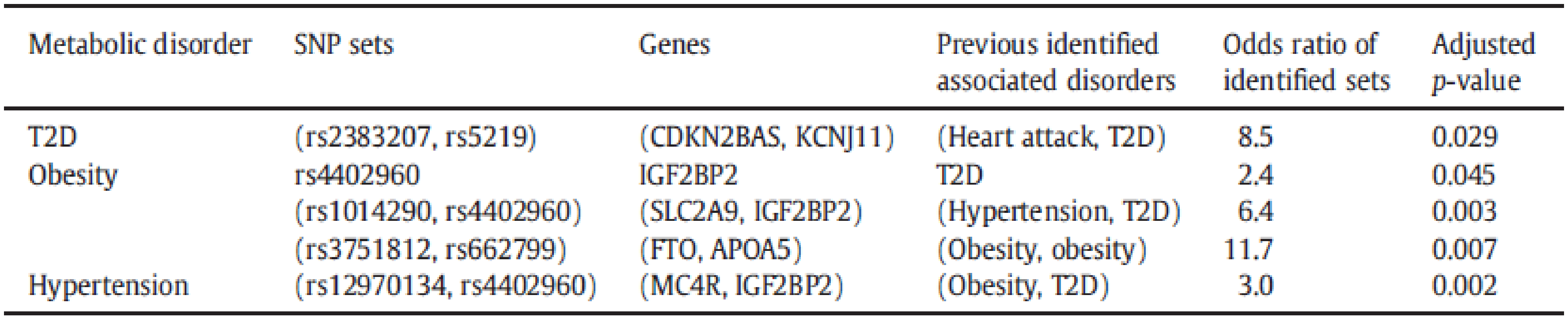
- Use the Lasso (Tibshirani 1995) to select variables:

$$\hat{\beta} = \arg\min \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Subject to $\sum_{j=1}^{p} |\beta_j| \leq t$

- When t is sufficiently small, some coefficients will be zero
- The lasso can be used to select variables and perform classification.

Example – interaction effect using candidate SNPs and continuous phenotypes



Metabolic disorder	SNP sets	Genes	Previous identified associated disorders	Odds ratio of identified sets	Adjusted *p*-value
T2D	(rs2383207, rs5219)	(CDKN2BAS, KCNJ11)	(Heart attack, T2D)	8.5	0.029
Obesity	rs4402960	IGF2BP2	T2D	2.4	0.045
	(rs1014290, rs4402960)	(SLC2A9, IGF2BP2)	(Hypertension, T2D)	6.4	0.003
	(rs3751812, rs662799)	(FTO, APOA5)	(Obesity, obesity)	11.7	0.007
Hypertension	(rs12970134, rs4402960)	(MC4R, IGF2BP2)	(Obesity, T2D)	3.0	0.002

The selected interaction pairs can provide insights to gene regulation and disease pathology (Wang et al. 2014)

- Follows up by further experiments

Advantages of lasso

- Evaluate main effect, interaction effects and environmental factors simultaneously
- Handles all types of phenotype data and genetic data

Caution of Lasso

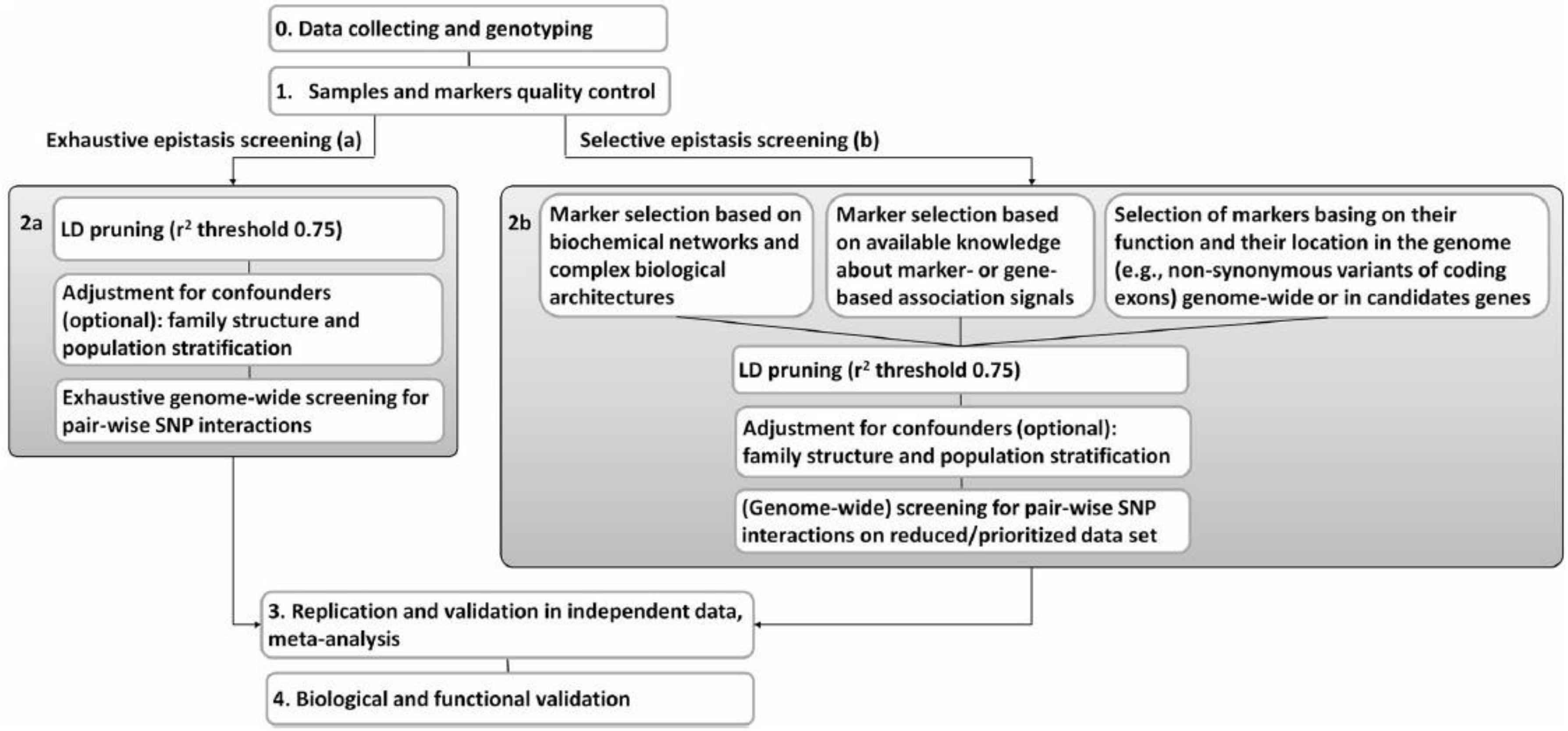
- If the shrinkage parameter is not properly tuned, might cause over-fitting problem
  - The remedial method is to select the tuning parameter through cross-validation
- Lasso does not provide p-value estimation for the coefficients
  - Needs other method to find out: linear regression, permutations

Caution of Lasso

- Even if lasso can handle n<p, the sample size to parameter ratio shouldn't go too small.
  - one will need to pre-select or pre-screen the input variables (Fan & Li 1999)

- In some instances, when including interaction parameters in a regression framework there is no direct correspondence between the interactive effects in the logistic regression models and the underlying penetrance based models displaying some kind of epistasis effect (North et al. 2005)

GWAI protocol


0. Data collecting and genotyping
1. Samples and markers quality control
Exhaustive epistasis screening (a)
Selective epistasis screening (b)
2a
LD pruning (r^2 threshold 0.75)
Adjustment for confounders (optional): family structure and population stratification
Exhaustive genome-wide screening for pair-wise SNP interactions
2b
Marker selection based on biochemical networks and complex biological architectures
Marker selection based on available knowledge about marker- or gene-based association signals
Selection of markers basing on their function and their location in the genome (e.g., non-synonymous variants of coding exons) genome-wide or in candidates genes
LD pruning (r^2 threshold 0.75)
Adjustment for confounders (optional): family structure and population stratification
(Genome-wide) screening for pair-wise SNP interactions on reduced/prioritized data set
3. Replication and validation in independent data, meta-analysis
4. Biological and functional validation

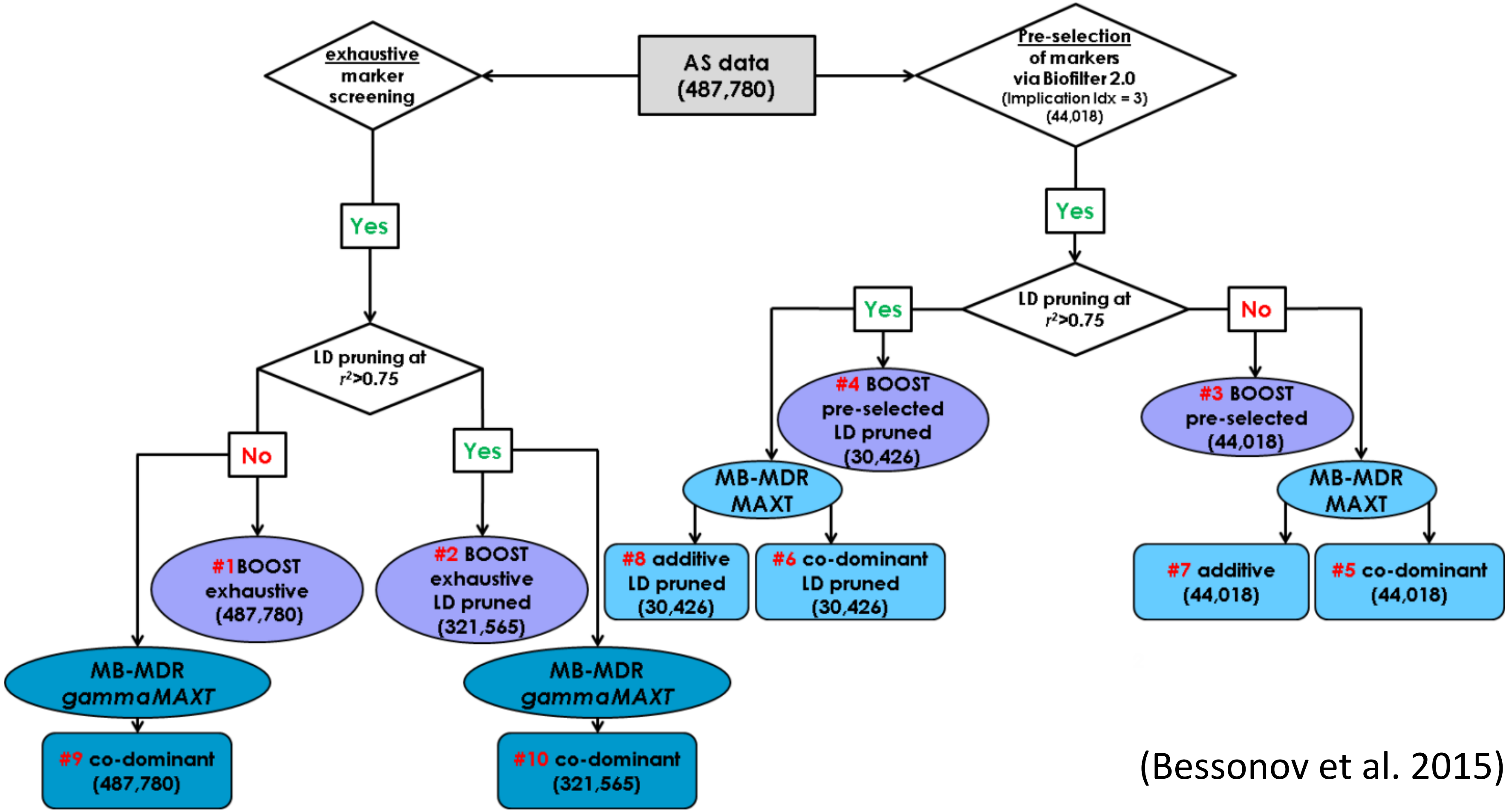
(Gusareva et al. 2014)

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

Prior biological information

- Some researchers incorporate prior biological “knowledge”:
  - Allow for uncertainty involved in the data source entries
  - Acknowledge the complementary characteristics of each of the available data sources
  - Think about the “significance” of evidence scores
- The advantage is reduced data dimension and potentially saving costs
- The draw-back is to be restricted by the biological assumption: hypothesis-driven versus hypothesis generating analysis

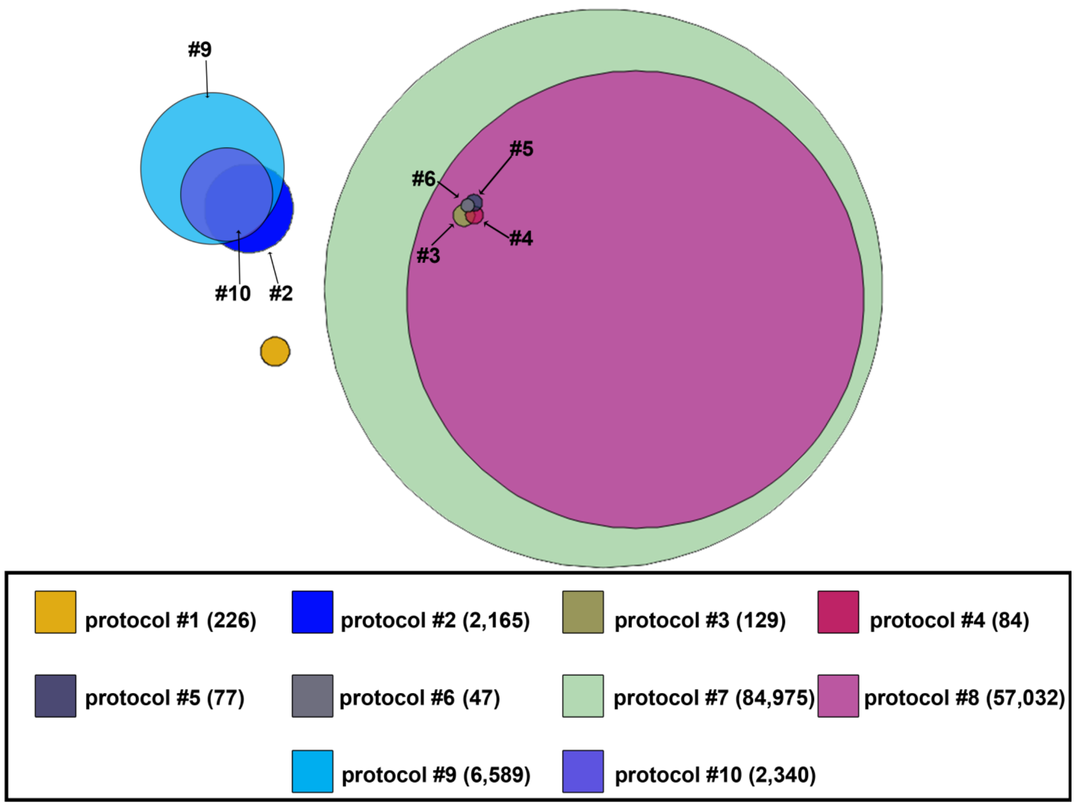
Slight protocol changes may lead to huge differences in results



AS data
(487,780)
exhaustive marker screening
Yes
LD pruning at r^2>0.75
No
Yes
#1 BOOST exhaustive (487,780)
#2 BOOST exhaustive LD pruned (321,565)
MB-MDR gammaMAXT
MB-MDR gammaMAXT
#9 co-dominant (487,780)
#10 co-dominant (321,565)
Pre-selection of markers via Biofilter 2.0 (Implication Idx = 3) (44,018)
Yes
LD pruning at r^2>0.75
Yes
No
#4 BOOST pre-selected LD pruned (30,426)
#3 BOOST pre-selected (44,018)
MB-MDR MAXT
MB-MDR MAXT
#8 additive LD pruned (30,426)
#6 co-dominant LD pruned (30,426)
#7 additive (44,018)
#5 co-dominant (44,018)

(Bessonov et al. 2015)

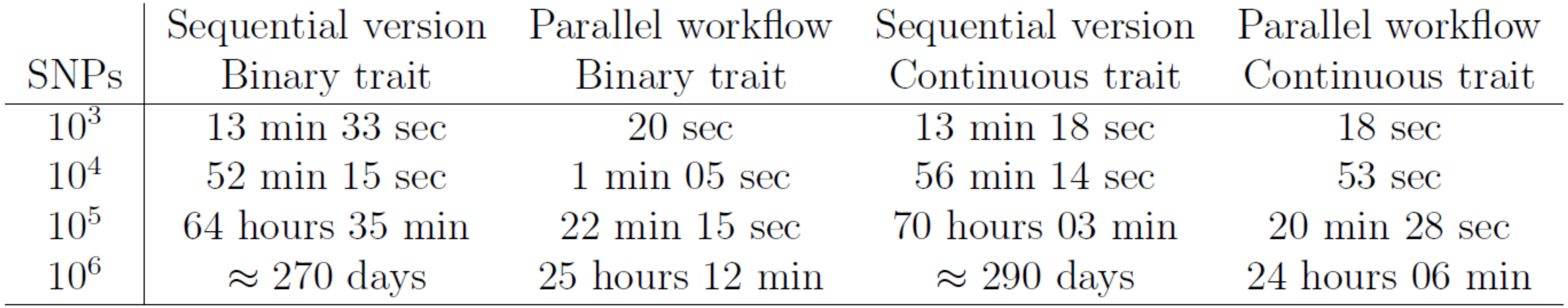
Slight protocol changes may lead to huge differences in results



- **Bessonov K, Gusareva ES, Van Steen K (2015)** A cautionary note on the impact of protocol changes for Genome-Wide Association SNP x SNP Interaction studies: an example on ankylosing spondylitis. Hum Genet - accepted **[non-robustness of GWAI analysis protocols]**

Speed

- From sequential to parallel workflows and a better usage of “null data”



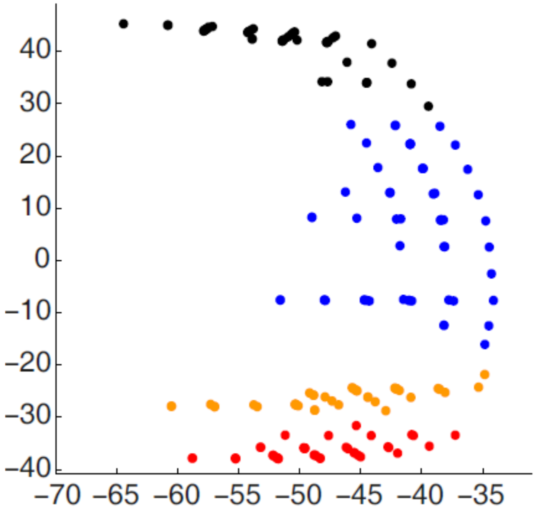
SNPs	Sequential version Binary trait	Parallel workflow Binary trait	Sequential version Continuous trait	Parallel workflow Continuous trait
10^3	13 min 33 sec	20 sec	13 min 18 sec	18 sec
10^4	52 min 15 sec	1 min 05 sec	56 min 14 sec	53 sec
10^5	64 hours 35 min	22 min 15 sec	70 hours 03 min	20 min 28 sec
10^6	$\approx$ 270 days	25 hours 12 min	$\approx$ 290 days	24 hours 06 min

(Example shown: mbmdr-4.2.2.out; results prefixed by “≈” are extrapolated; parallel workflow: 256-core computer cluster Intel L5420 2.5 GHz; Sequential: single core)

- **Van Lishout** F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, Theâtre E, Charloteaux B, Calle ML, Wehenkel L, Van Steen K (2012) An efficient algorithm to perform multiple testing in epistasis screening. BMC Bioinformatics. 2013 Apr 24;14:138 **[C++ MB-MDR made faster!]**
- **Van Lishout F,** Gadaleta F, Moore JH, Wehenkel L, Van Steen K (2015) gammaMAXT: a fast multiple-testing correction algorithm – submitted **[C++ MB-MDR made SUPER-fast]**

Confounding by shared genetic ancestry

(MB-MDR and pair-specific genomic control – Van Lishout)



AA, BB	AA, Bb	AA, bb
Aa, BB	Aa, Bb	Aa, bb
aa, BB	aa, Bb	Aa, bb

Multilocus Exposure

Disease Outcome

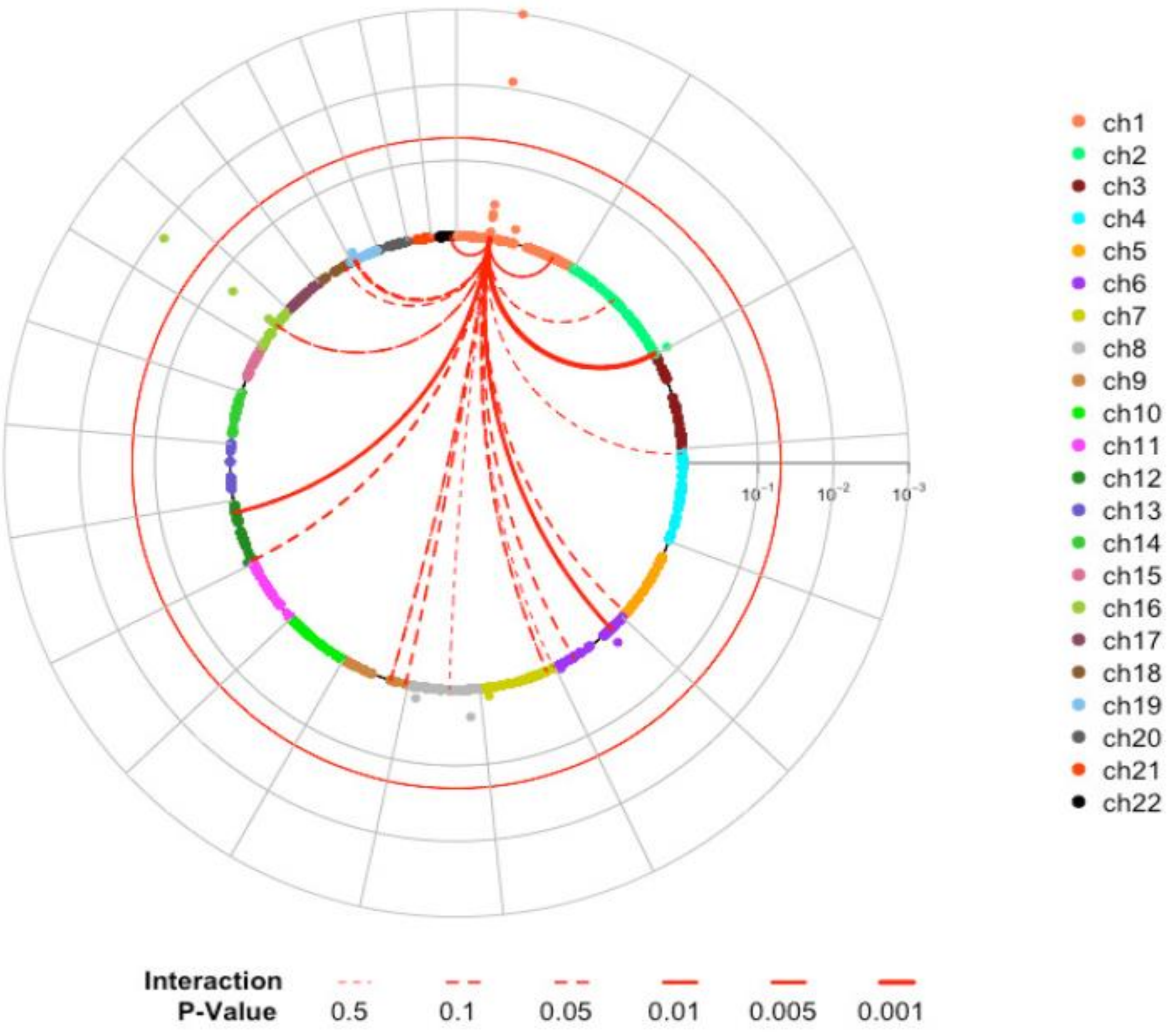
Replication

*“Leaving aside for the moment **what replication means** or should mean in the context of GWAIS, even for the currently so-called replicated genetic interactions it is unclear to what extent **a false positive has been replicated** due to the adopted methodological strategy itself or whether the replication of epistasis is not solely attributed to main effects (such as HLA effects) not properly accounted for.“*

*“Genome-wide SNP genotyping platforms consist predominantly of **tagSNPs** from across the genome. Most of these SNPs are not causal and have no functional consequences. **When two or more tagSNPs are combined in a genetic interaction** model, is it reasonable to assume that the same combination of tagSNPs interacts in an independent dataset?”*

(Ritchie and Van Steen 2015 – under review)

Visualization – the MogPlot (Van Lishout)



ANALTYIC MEANS

The W-test

W-test

- Handles case-control genotype data, measures main and interactions
- Null hypothesis: *If a SNP or SNP-pair has association effect to the phenotype, the probability distribution of the set is different in the case group from that in the control group.*
- Test form:

$$W = h\sum_{i=1}^{k}\left[\log \frac{\hat{p}_{1i}/(1-\hat{p}_{1i})}{\hat{p}_{0i}/(1-\hat{p}_{0i})} \Big/ SE_i\right]^2 \sim \chi_f^2$$

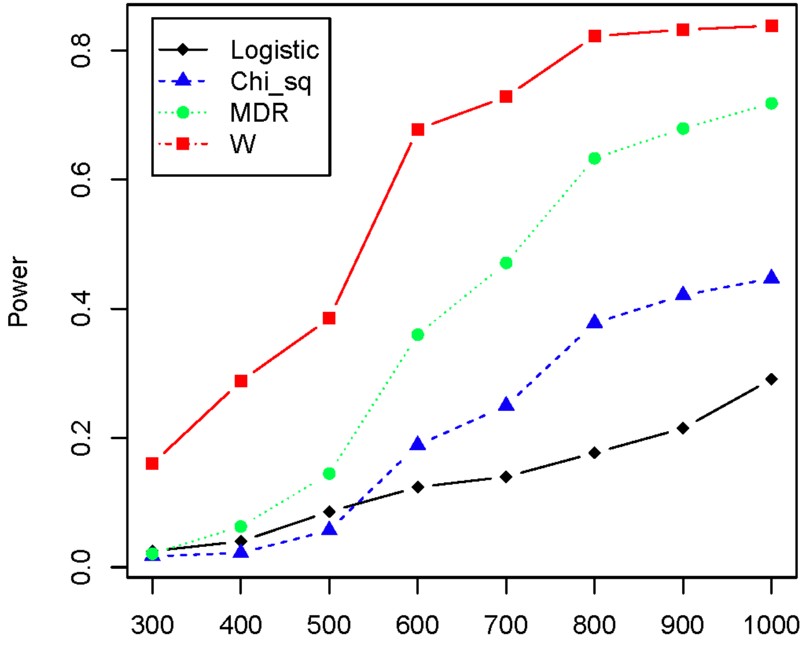
Where, $SE_i = \sqrt{\frac{1}{n_{0i}} + \frac{1}{n_{1i}} + \frac{1}{N_0 - n_{0i}} + \frac{1}{N_1 - n_{1i}}}$

- Probability distribution is data-set adaptive: h and f, obtained from bootstrapped samples of the working data set.
- h $\approx$ (k – 1)/k and f $\approx$ k – 1.

Properties of W-test

- Measures epistasis effect without making linear or non-linear assumption
- h, f: Let W-test robust in low frequency data environment or when sample size decrease
- Adjust for mild population structure
- Fast:
  - ○ Several hours for GWAS on cluster
- R package: wtest(), C software

Figure : Power in low frequency SNPs environment

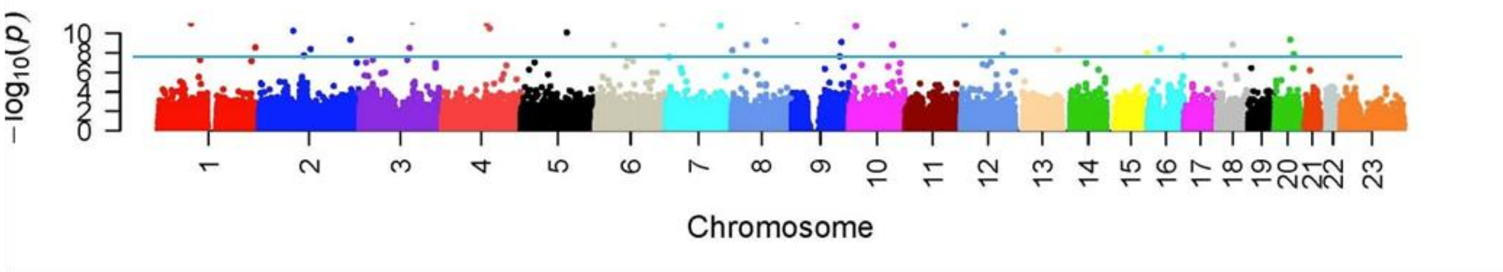


W-test: application on real GWAS data

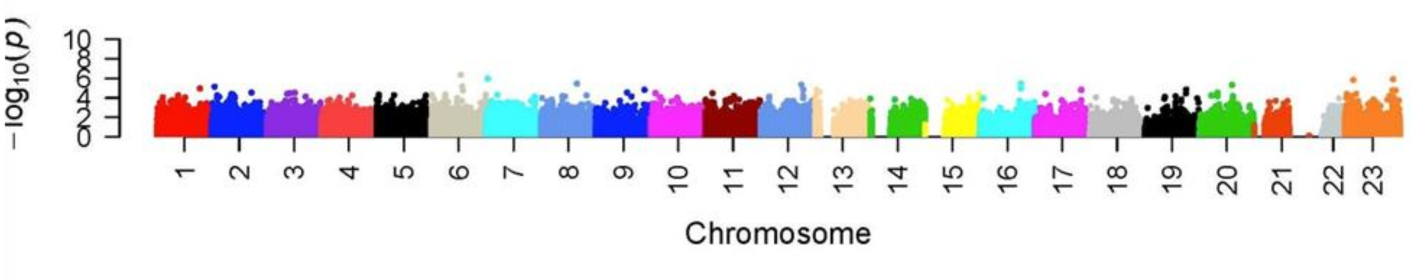
- Dataset 1: Welcome Trust Case-control Consortium (WTCCC) bipolar data set (Burton, Clayton et al. 2007).
    - 2,000 cases and 3,000 controls
    - 414,682 SNPs after quality control.
- Dataset 2: Genetic Association Information Network (GAIN) bipolar project in dbGaP database (McInnis, Dick et al. 2003)
    - 1,079 cases and 1,089 controls
    - 729,304 SNPs after quality control

Real data application – main effect

- WTCCC data



- GAIN data



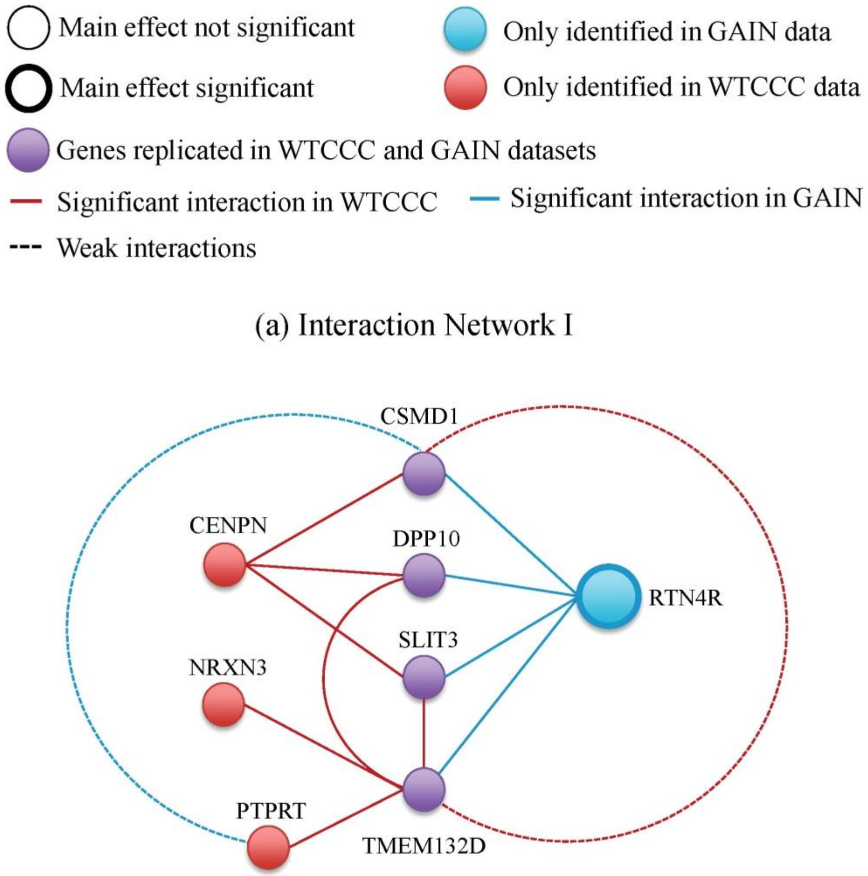
- 51 genome-wide significant markers in WTCCC, and 1 in GAIN data.
- 80% of the significant markers identified by W-test are in the low frequency MAF range

W-test : Interaction networks of two data sets

SLIT3: mammalian SLIT proteins may participate in the formation and maintenance of the nervous and endocrine systems by protein-protein interactions. (Itoh et al. 1998)

DPP10: Transcripts of 5.0 and 7.5 kb were also detected in **brain.** Analysis of mouse ESTs indicated that mouse Dpp10 was expressed in several **brain regions and retina**. DPP10 was recovered in the membrane fraction of transfected cells. (Qi et al. 2003)

PTPRT: a receptor-type protein tyrosine phosphatase for signal transduction and neurite extension, which promotes synapse formation and is reported to be highly expressed in the central nervous system.


Main effect not significant
Only identified in GAIN data
Main effect significant
Only identified in WTCCC data
Genes replicated in WTCCC and GAIN datasets
Significant interaction in WTCCC
Significant interaction in GAIN
Weak interactions
CSMD1
CENPN
DPP10
RTN4R
SLIT3
NRXN3
PTPRT
TMEM132D

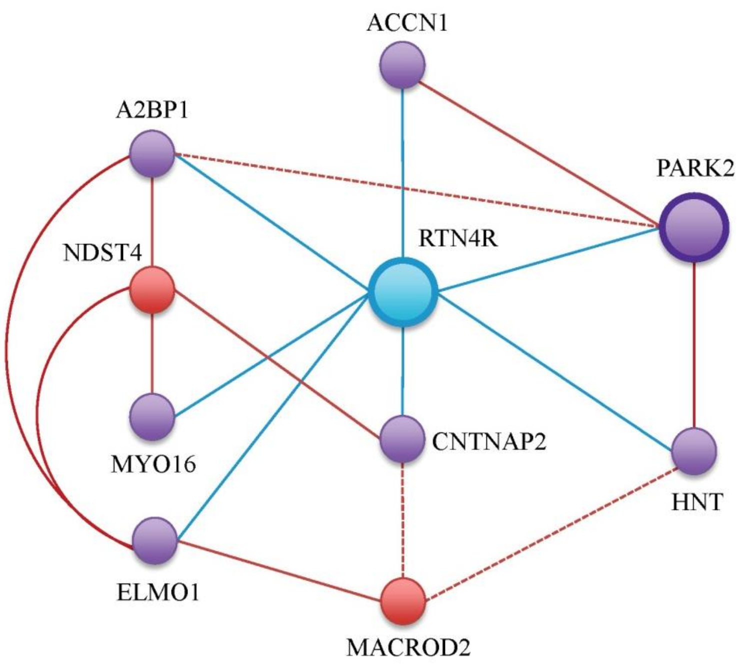
(a) Interaction Network I

W-test : Interaction networks of two data sets

ELMO1-A2BP1: has significant epistasis effect (p-value 3.86E-18), while the component SNPs are non-significant with p-values of 3.02×10^{-6} and 3.97×10^{-3}

A2BP1, or *RBFOX1*: RNA-binding protein that regulates alternative splicing in neurons and plays a key role in the development of human neurons reported in RNA-sequencing, cytogenetic, and molecular characterization studies

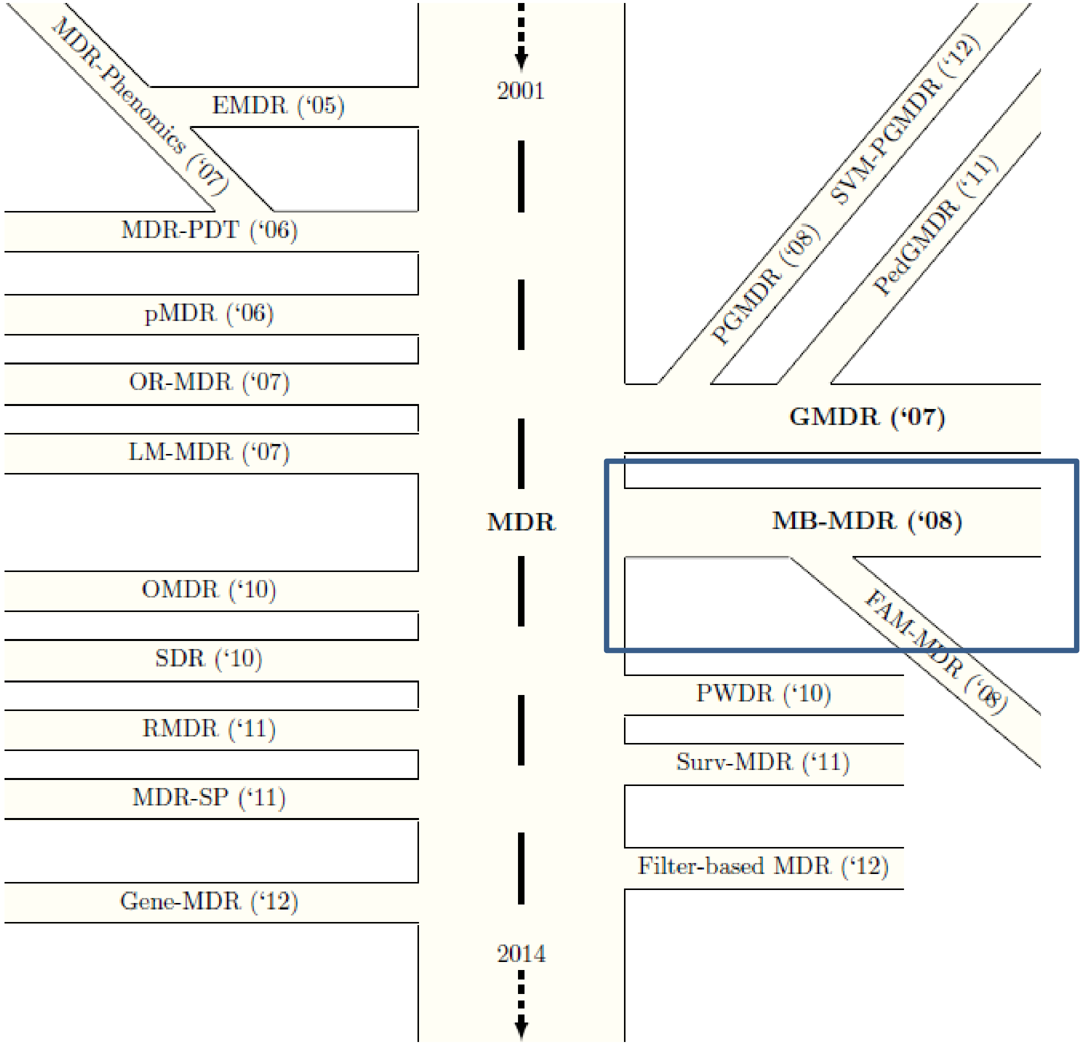
CNTNAP2: encodes a neuronal transmembrane protein member of the neurexin superfamily. broadly expressed in the developing rodent brain. Abrahams et al. (2007) noted that human CNTNAP2 expression was enriched in circuits involved in higher cortical functions, including language.


(b) Interaction Network II

ANALTYIC MEANS

MB-MDR

Several MDR roads lead to Rome



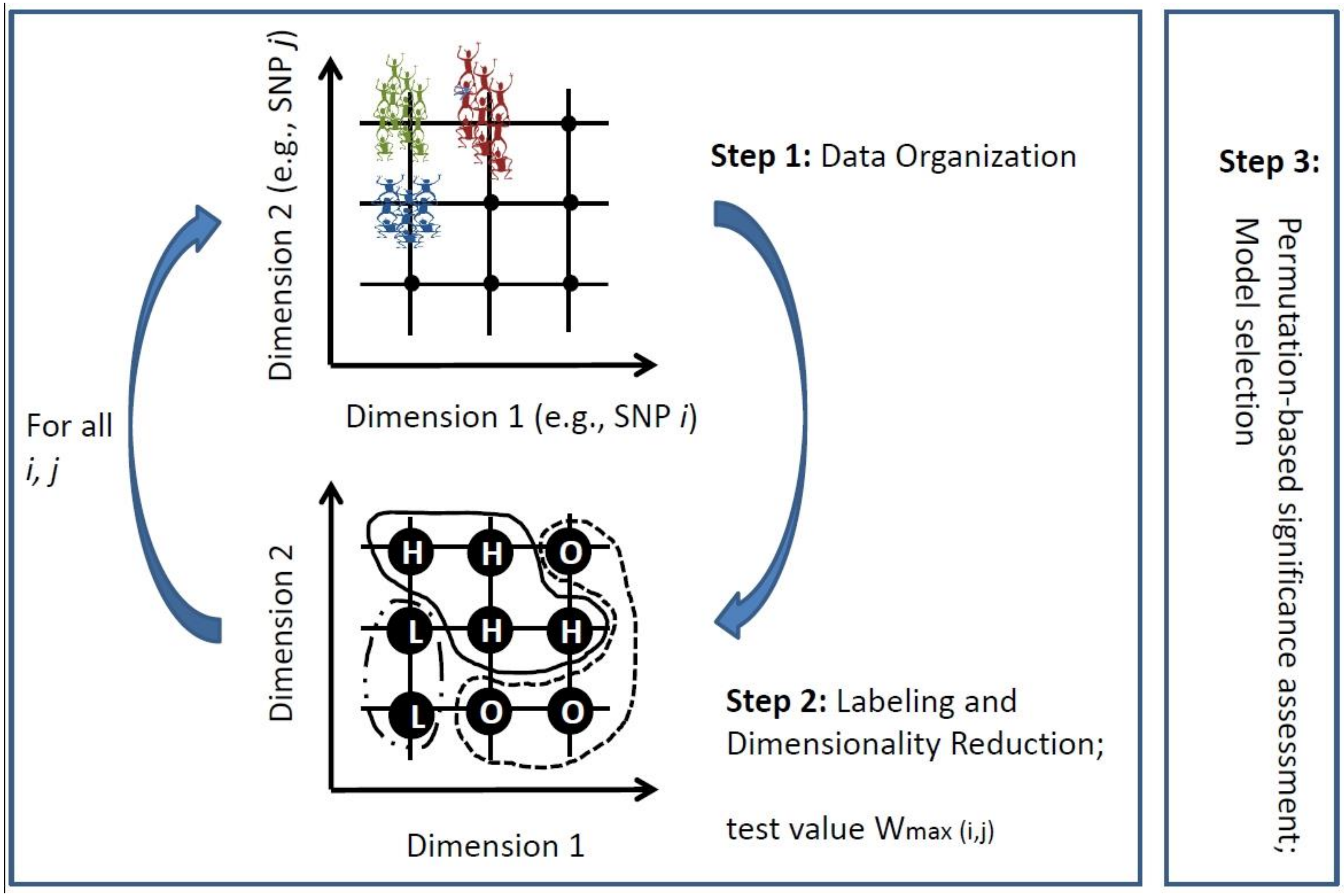
Which dimensions are reduced?

- The estimated degrees of freedom for MDR and LR using K=1, 2 and 3 factors (standard errors in parentheses). LR exact refers to the asymptotic exact degrees of freedom

	Number of Factors K		
Method	**1**	**2**	**3**
MDR	1.9 (0.13)	5.6 (0.20)	17.4 (0.37)
LR	2.1 (0.4)	8.0 (0.26)	26.8 (0.53)
LR exact	2	8	26

(Park and Hastie 2007)

Model-Based Multifactor Dimensionality Reduction (MB-MDR)



Step 1: Data Organization

Step 2: Labeling and Dimensionality Reduction;

test value $W_{max\ (i,j)}$

Step 3:

Permutation-based significance assessment;
Model selection

MB-MDR

Summary of the steps involved in MB-MDR analysis:

- For every *k* variates (e.g., SNPs), Step 1 is a data organization step in which individuals are (naturally) allocated to *k*-dimensional (genotype) profiles.
- Step 2 labels individuals according to their profiles in multi-dimensional space and liberal association tests. Individuals with the same label are merged into a single group.
- Extreme groups are contrasted to each other via an association test, leading to a test value W_{max} for the selected k-tuple.
- The final k-models are selected in Step 3, using permutation-based significance assessments and adequate multiple testing control.

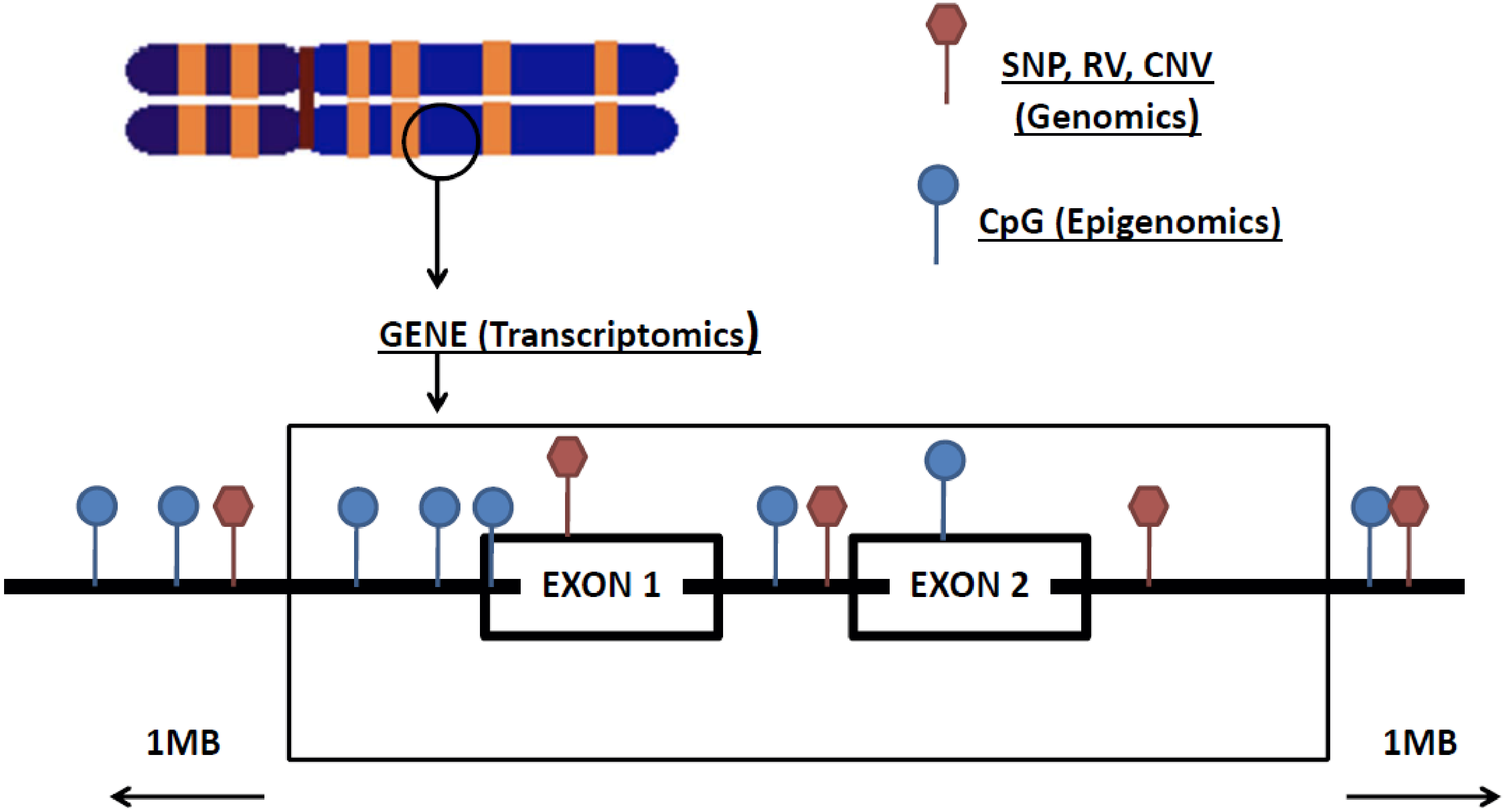
Shift from prediction accuracy (MDR) to association strength (MB-MDR)

- In MB-MDR, computation time is invested in optimal **association tests** to prioritize multi-locus genotype combinations
- Also in statistically valid permutation-based methods to assess **joint statistical significance**
- The labelling concept is extended beyond two “risk” groups and is based on the sign of “effects”
- Lower-order effects are potentially conditioned upon

Type I error **γ** Power **γ** Genetic heterogeneity **γ**

(see MB-MDR references)

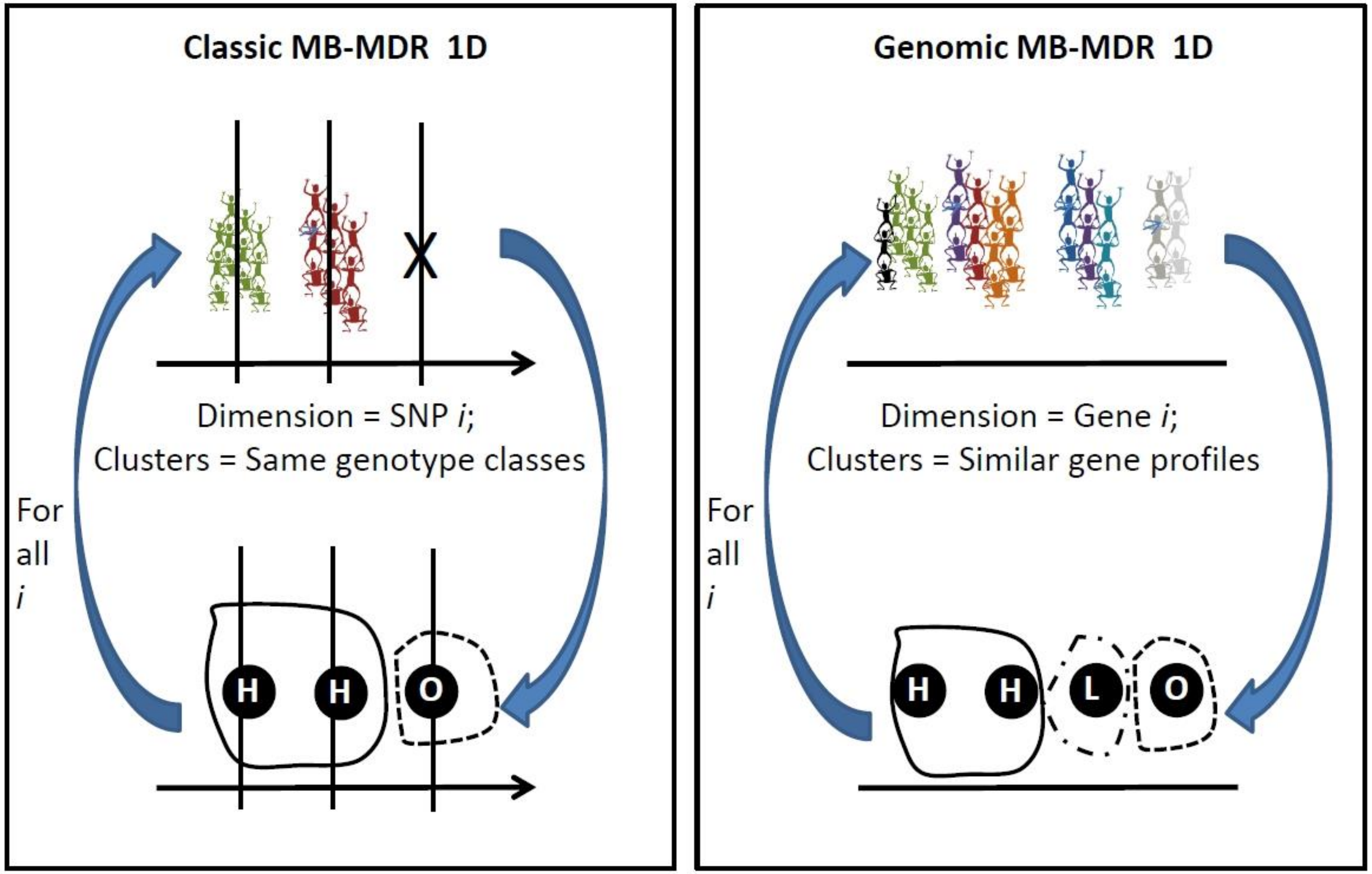
From SNPs to Sets-of-Interest



(Slide S Pineda 2014 – lab meeting)

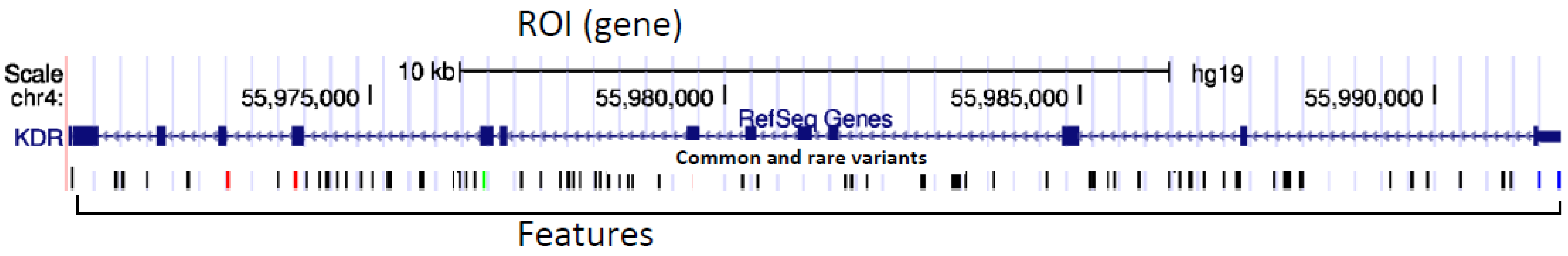
Gene-based MB-MDR

- **Fouladi R**, Bessonov K, Van Lishout F, Van Steen K (2015) Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis. Human Heredity – accepted **[aggregating based on similarity measures to deal with DNA-seq data]**

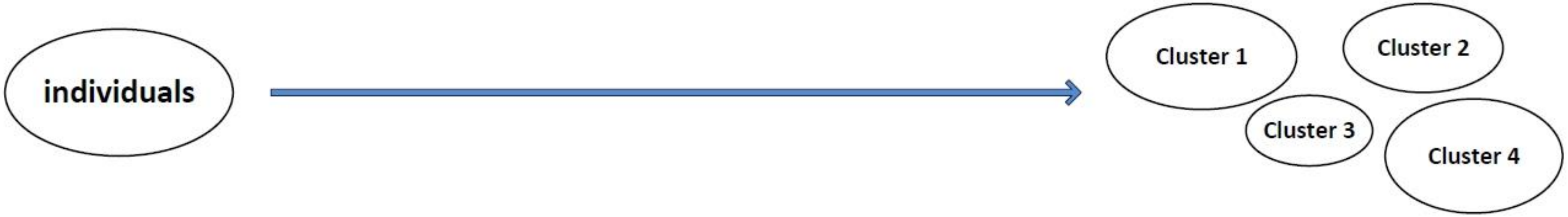


Gene-based MB-MDR (Fouladi et al. 2015 – DNA-seq)

- **Phase 1:** Select sets of interest (ROI) / Prepare the data

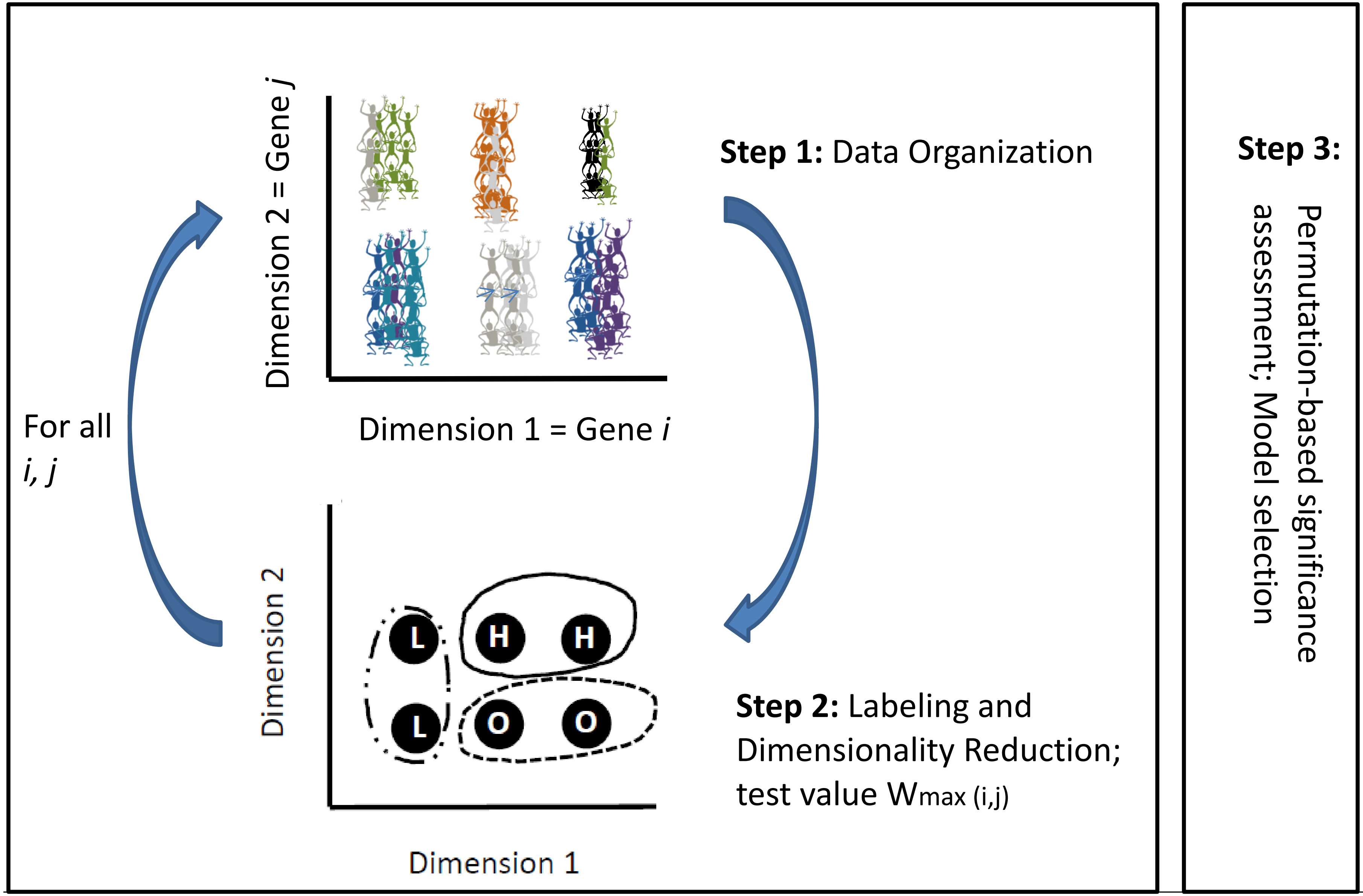


- **Phase 2:** Clustering individuals according to features (e.g., common and rare variants, epigenetic markers, … and kernel methodology)



- **Phase 3:** Application of classic MB-MDR on new constructs

Gene-based MB-MDR



For all *i, j*

Step 1: Data Organization

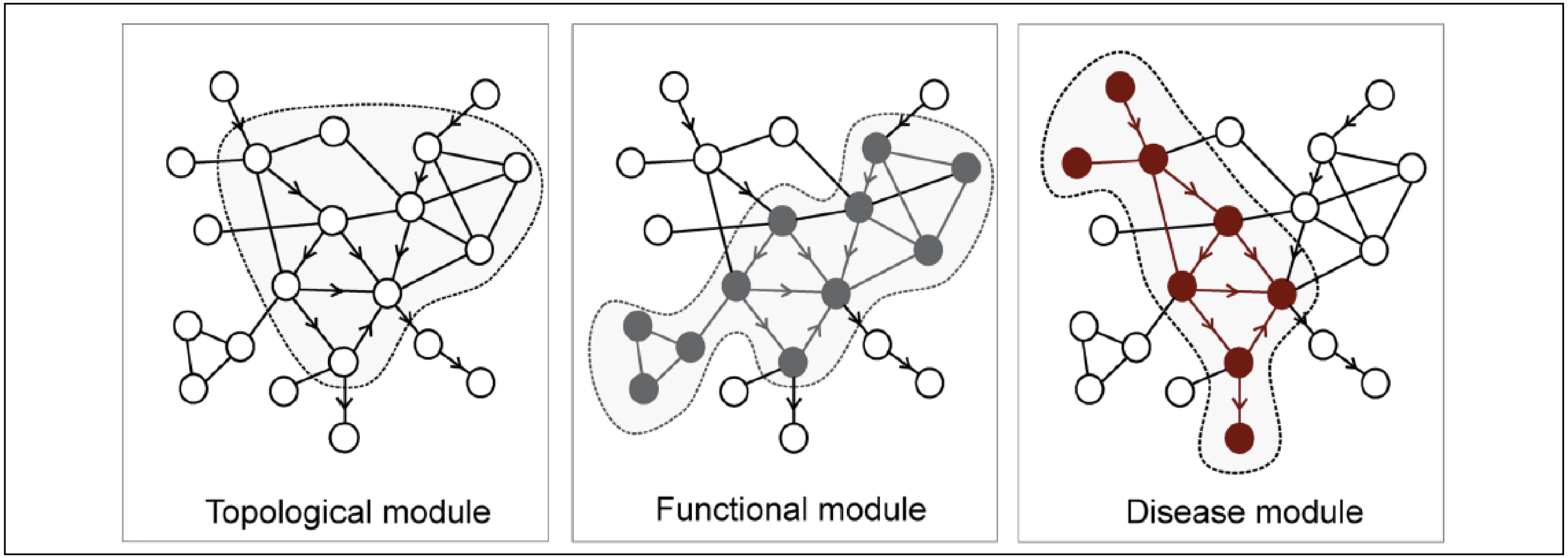
Step 2: Labeling and Dimensionality Reduction; test value W_{max} (i,j)

Step 3:

Permutation-based significance assessment; Model selection

Gene-based MB-MDR facilitates network medicine

- The underlying assumption is that the topological, functional, and disease modules overlap so that functional modules correspond to topological modules and a disease can be viewed as the breakdown of a functional module (Barabási et al. 2011)



TAKE HOME MESSAGES

(http://thebusyba.com)

Context matters

- Structured populations (genomic control)
- Meta-analysis (non-parametric)
- Replication and functional interpretation
- Set-based analyses

OUR AFFILIATIONS

KRISTEL VAN STEEN

GIGA-R Medical Genomics - BIO3 Unit

University of Liège

Belgium

http://www.statgen.ulg.ac.be/

http://www.montefiore.ulg.ac.be/~kvansteen/

Google Scholar, Research Gate

MAGGIE HAITIAN WANG

CU Medicine HONG KONG

香港中文大學醫學院

Faculty of Medicine

The Chinese University of Hong Kong

Centre for Clinical Research and Biostatistics (CCRB)
JC School of Public Health and Primary Care
The Chinese University of Hong Kong
Hong Kong S.A.R

www2.ccrb.cuhk.edu.hk/wtest/download.html

www.sphpc.cuhk.edu.hk/maggiew

Google Scholar, Research Gate