Evaluation of methods in gene association studies: yet another case for Bayesian networks

Gábor Hullám, Péter Antal,
András Falus and Csaba Szalai

MŰEGYETEM 1782

Department of Measurement
and Information Systems
Budapest University of Technology and Economics

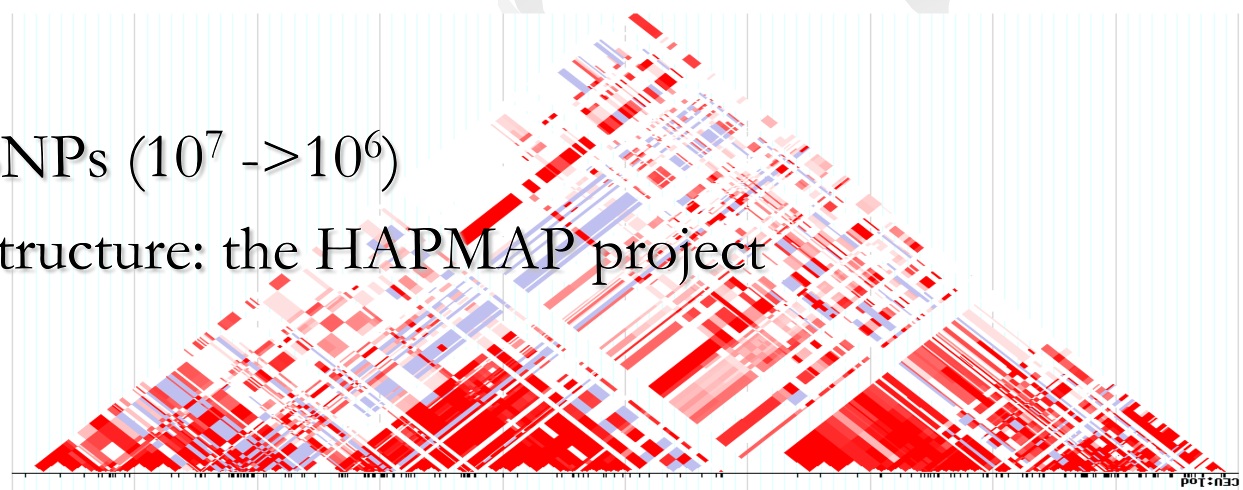
Department of Genetics
Cell and Immunobiology
Semmelweis University

- Genetic association studies (GAS)
- A Bayesian approach to GAS
- Bayesian networks in GAS
- Evaluation of methods

Motivation: Exploring the variome

- Variome
  - Single-Nucleotide Polymorphisms (SNPs)
  - Copy-Number Variations (CNVs).
  - Genome rearrangements
  - Methylome

- SNPs
  - Number of SNPs (10^7 -> 10^6)
  - Correlation structure: the HAPMAP project



GAS phases

PGAS Data analysis → confirmations refutations

384x48 PGAS s_{11}	384x48 PGAS s_{12}	384x48 PGAS s_{13}

Study design → SNP datasets

GWAS Data analysis → candidate regions. genes

Genome wide association study (GWAS)

GAS Facts

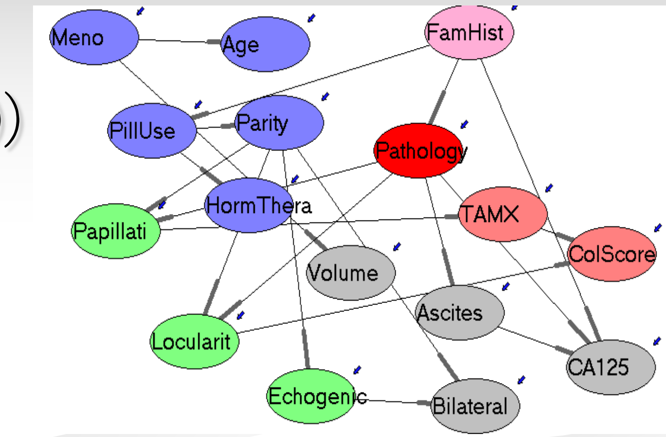
- Publications: ~40K
- SNPs on plate: 100K-2M
- Sample size: 30K
- Number of gene association studies
  - GWAS: ~100
  - PGAS-: ~10K
- Confirmed associations:
  - <1000
  - Small attributable risk
- Why?
  - Common disease – common variance hypothesis - **multifactorial diseases**, many weak interactions
  - Rare haplotype hypothesis (Minor allele freq. <1%)

Current challenge: the discovery of epistasis

- Statistical epistasis: non-linear interaction of genes
- The goal is the exploration of…
  - explanatory variables of the target variable(s)
  - the interaction of explanatory variables
- Genetic association concepts can be formalized (partially) as machine learning concepts and as Bayesian network concepts

The model class: Bayesian networks

- directed acyclic graph (DAG)
  - nodes – domain entities
  - edges – direct probabilistic relations
- conditional probability models P(X|Pa(X))
- interpretations:

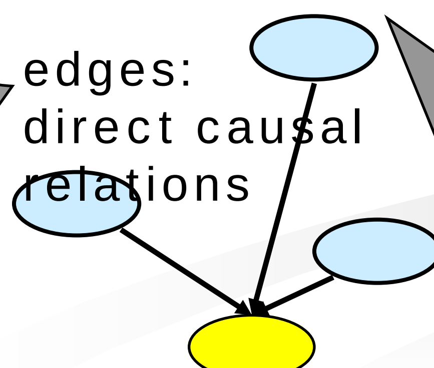


effective representation of the distribution

$$P(X_1,..X_N) = \prod_{i=1}^{N} P(X_i \mid Pa(X_i))$$

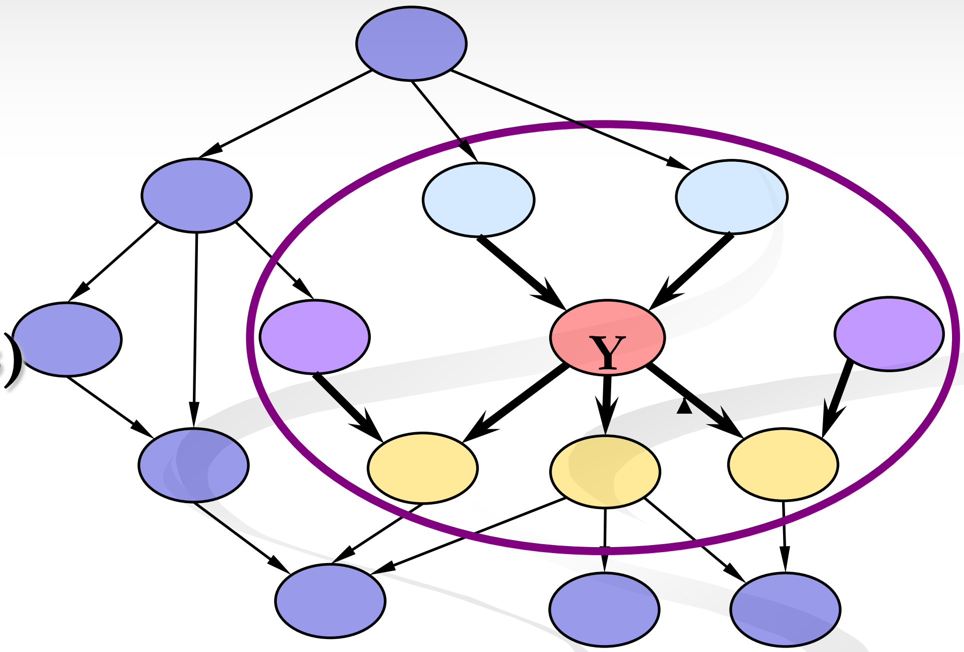
DAG structure: dependency map (d-separation)

edges: direct causal relations



Bayesian network features representing relevance

- **Markov Blanket (sub)Graphs (MBGs)**
  - (1) parents of the node
  - (2) its children
  - (3) parents of the children



- **Markov Blanket Sets (MBSs)**
  - the set of nodes which probabilistically isolate the target from the rest of the model
- **Markov Blanket Membership (MBM)**
  - pairwise relationship

GA-to-BN

- (model-based) pairwise association➔ Markov Blanket Memberhsips (MBM)
- Multivariate analysis ➔ Markov Blanket sets (MB)
- Multivariate analysis with interactions ➔ Markov Blanket Subgraphs (MBG)
- Causal relations/models ➔ Partially directed Bayesian network (PDAG)

- **Hierarchy**
  - DAG=>PDAG=>MBG=>MB=>MBM

Advantages of GA-to-BN - 1

- **Strong relevance - direct association:** Clear semantics and dedicated goal for the explicit. faithful representation of strongly relevant (e.g. non-transitive) relations
- **Graphical representation**: It offers better overview of the dependence-independence structure. e.g. about interactions and conditional relevance.
- **Multiple targets:** It inherently works for multiple targets.

Advantages of GA-to-BN – 2

- **Incomplete data:** It offers integrated management of incomplete data within Bayesian inference.
- **Causality:** Model-based causal interpretation of associations
- **Haplotype level:** Offers integrated approach to haplotype reconstruction and association analysis (assuming unphased genotype data)

Challenges of applying BNs in GAS

- High computational complexity
- High sample complexity ➔ Bayesian statistics

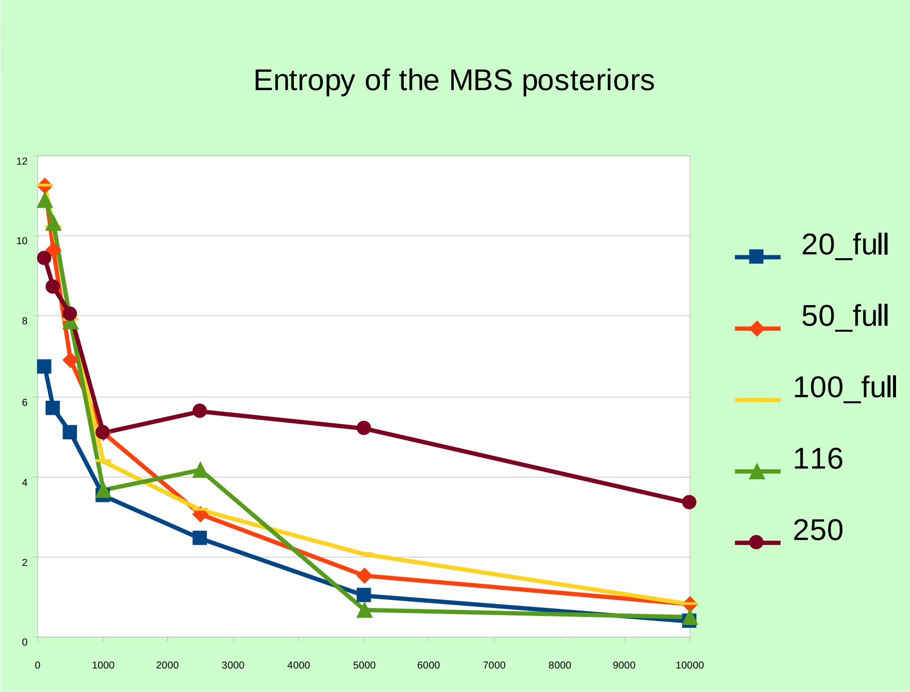
  ➔ Bayesian model averaging

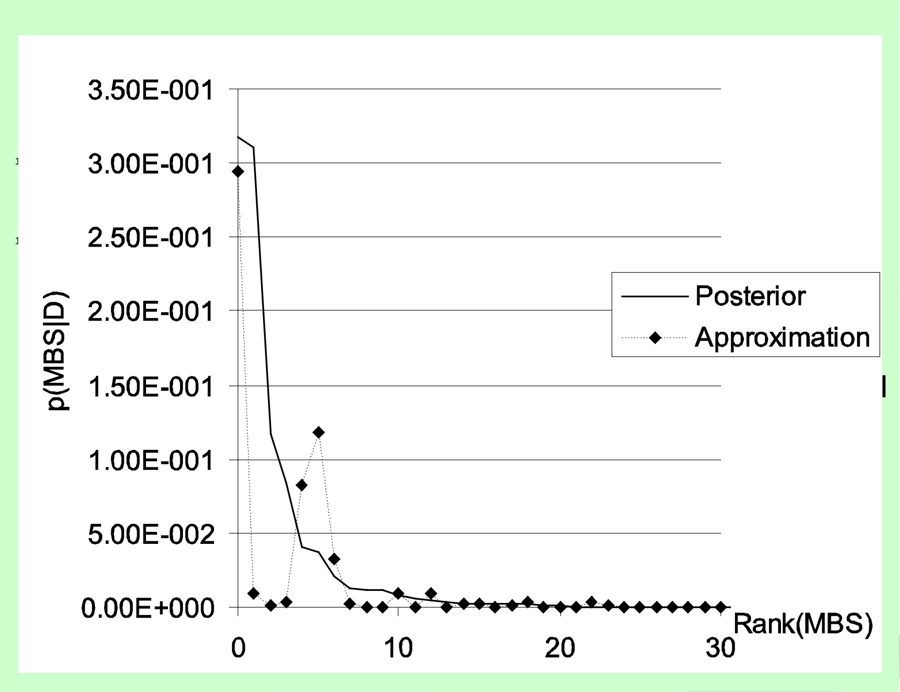
  ➔ Feature posterior

$$P(F = f) = \sum_{G:F^G = f} P(G)$$

- **Goal**: approximate the full-scale summation (integral)
- **A solution**: Metropolis coupled Markov chain Monte Carlo (MCMCMC)

Uncertainty in multivariate analysis





Advantages of the Bayesian framework

- **Automated correction for "multiple testing"**
  - The measure of uncertainty at a given level automatically indicates its applicability
- **Prior incorporation**: better prior incorporation both at parameter and structural levels.
- **Post fusion:** better semantics for the construction of meta probabilistic knowledge bases
- **Normative uncertainty for model properties** (cf. bootstrap)

The basis for comparison

Our approach is a model based exploration of the underlying structure

(note: multiple targets, causal and direct aspects)

$\neq$

Prediction of class labels

Comparison of GAS tools

Dedicated GAS tools

- BEAM
- BIMBAM
- SNPAssoc
- SNPMstat
- Powermarker

General purpose FSS tools

- MDR
- Causal Explorer

…

Application domain: The genomic background of asthna

- moderate number of clinical variables (in the range of 50)
- hundreds of genotypic SNP variables for each patient
- thousands of gene expression measurements

Asthma

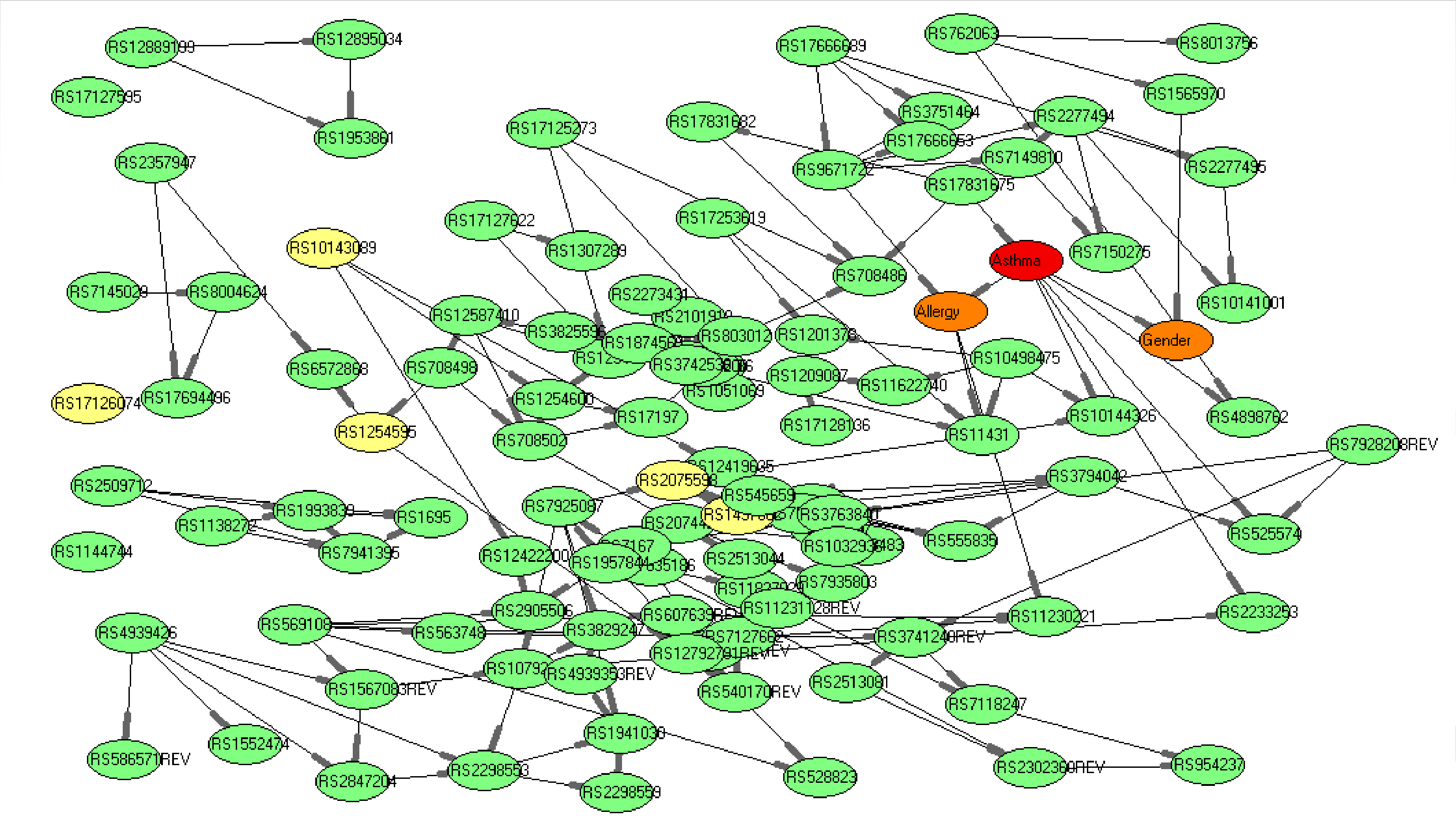
- Complex disease mechanism
- Half of the patients do not respond well to current treatments
- Unknown pathways in the asthmatic process

Evaluation on an artificial data set

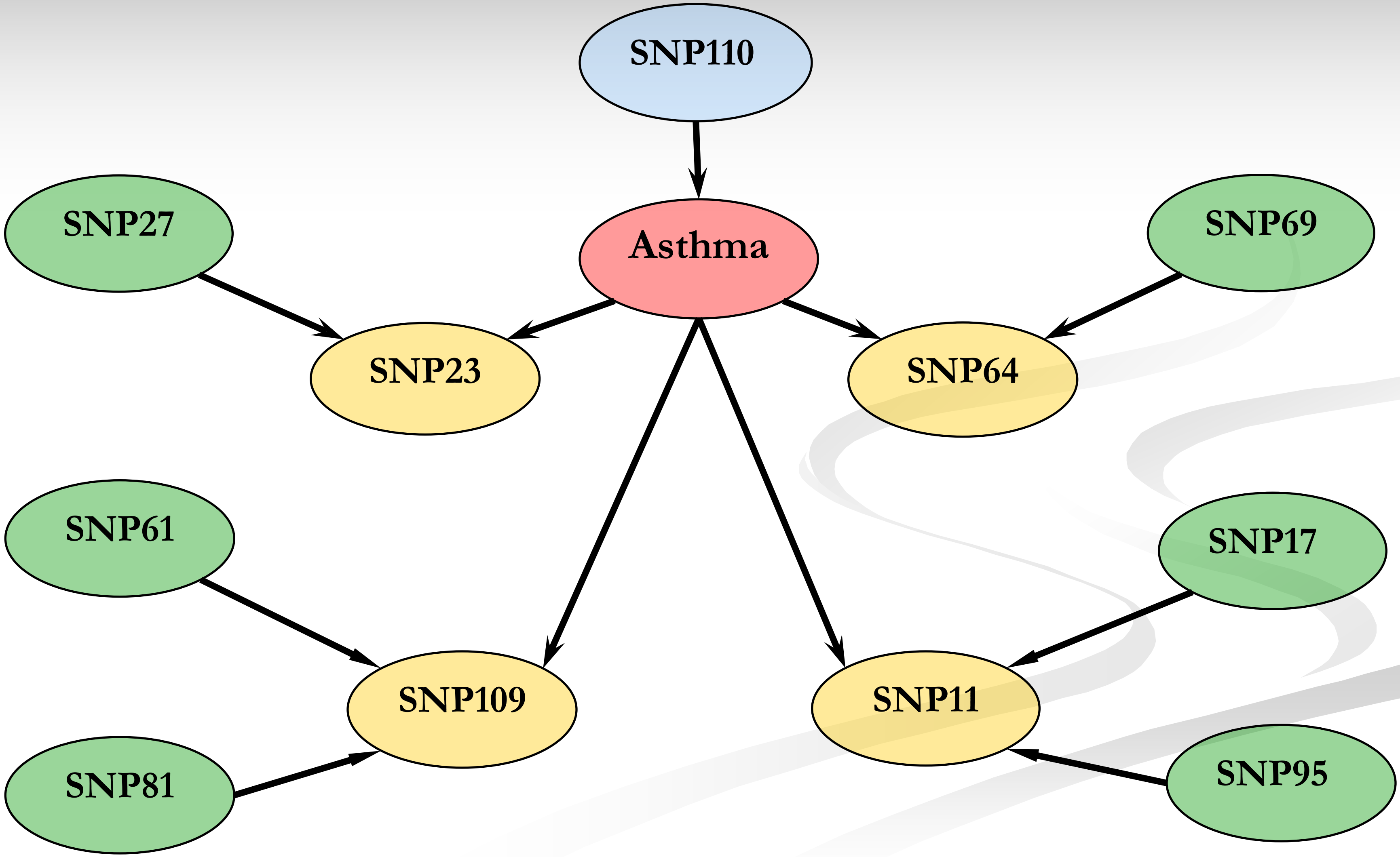
- Artificial model based on a real-world domain: the genomic background of asthma

- The real data set consists of:
  - 113 SNPs
  - 1117 samples

The reference model



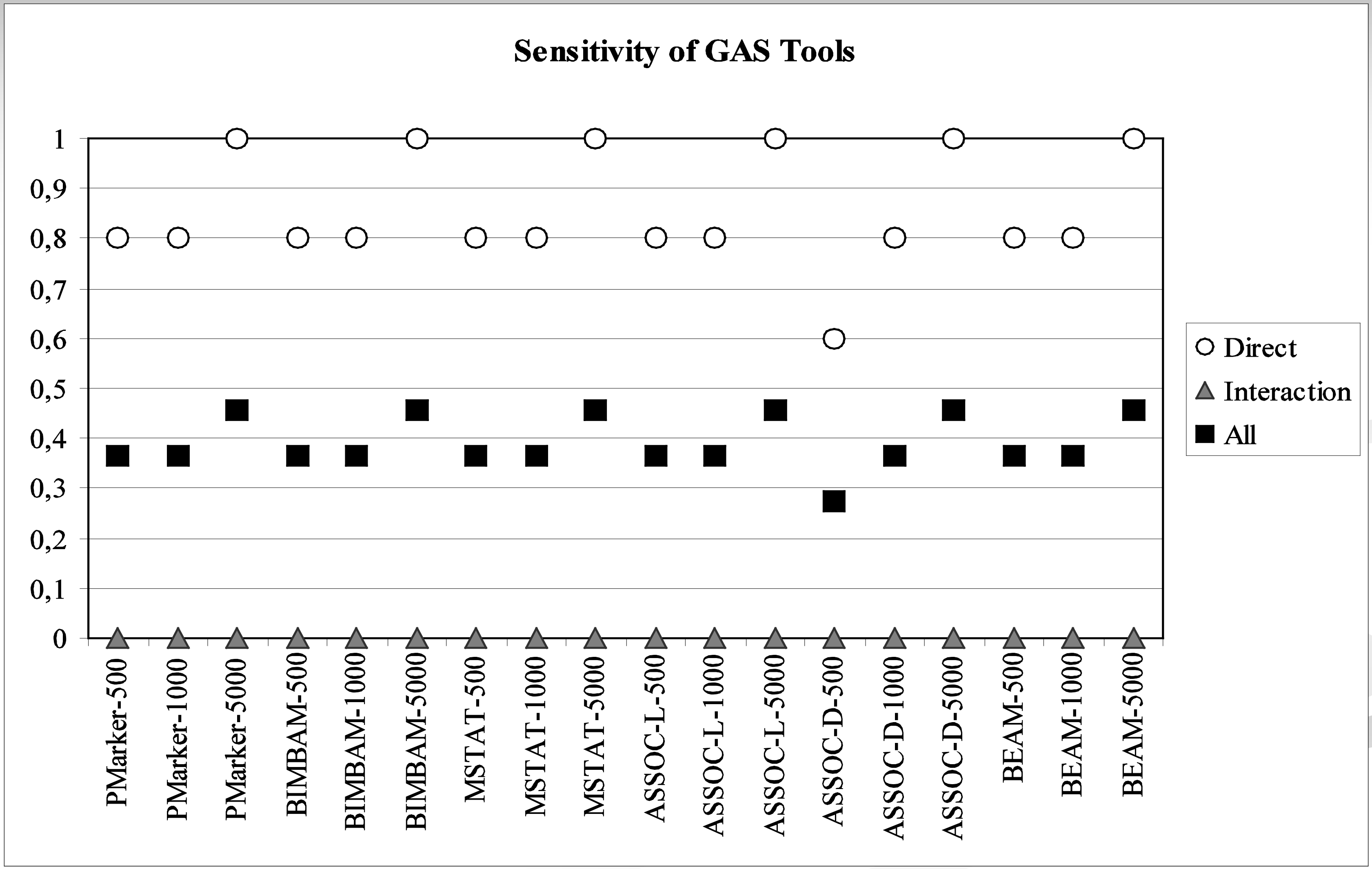
Reference MBG



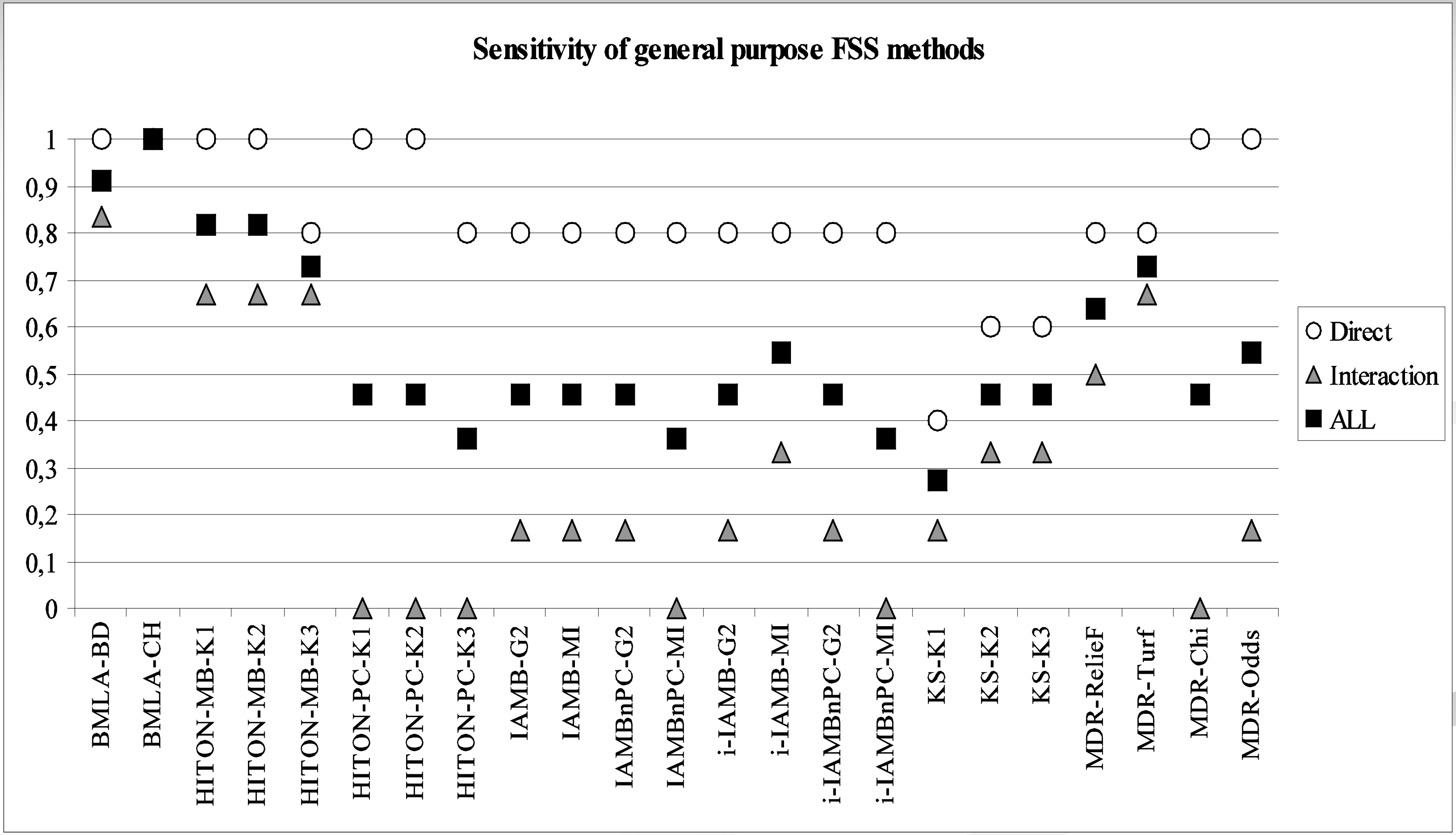
Results - 1

Software (Parameters)	Sensitivity	Specificity	Accuracy
BMLA (CH)	1	0.99	0.99115044
BMLA (BD)	0.92307692	1	0.99115044
HITON MB (k=1)	0.76923077	0.98	0.95575221
HITON MB (k=2)	0.76923077	0.99	0.96460177
HITON MB (k=3)	0.69230769	0.99	0.95575221
MDR – TurF	0.61538462	0.97	0.92920354
MDR – Relief	0.53846154	0.96	0.91150442
interIAMB (MI)	0.46153846	0.96	0.90265487

Results – 2.


Sensitivity of GAS Tools
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
PMarker-500
PMarker-1000
PMarker-5000
BIMBAM-500
BIMBAM-1000
BIMBAM-5000
MSTAT-500
MSTAT-1000
MSTAT-5000
ASSOC-L-500
ASSOC-L-1000
ASSOC-L-5000
ASSOC-D-500
ASSOC-D-1000
ASSOC-D-5000
BEAM-500
BEAM-1000
BEAM-5000
Direct
Interaction
All

Results – 3.



Sensitivity of general purpose FSS methods
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
BMLA-BD
BMLA-CH
HITON-MB-K1
HITON-MB-K2
HITON-MB-K3
HITON-PC-K1
HITON-PC-K2
HITON-PC-K3
IAMB-G2
IAMB-MI
IAMBnPC-G2
IAMBnPC-MI
i-IAMB-G2
i-IAMB-MI
i-IAMBnPC-G2
i-IAMBnPC-MI
KS-K1
KS-K2
KS-K3
MDR-RelieF
MDR-Turf
MDR-Chi
MDR-Odds
Direct
Interaction
ALL

Summary

- **General BN representation is feasible and gives superior performance for PGAS**
- **Bayesian statistics allows the quantification of applicability of BNs**
- Special extensions are necessary for
  - Multiple targets
  - Combined discovery of relevance and interactions (MBM, MBS, MBG)
  - Scalable multivariate analysis (k-MBS concept)
  - Feature aggregation

Antal et al.: A Bayesian View of Challenges in Feature Selection: Multilevel Analysis, Feature Aggregation, Multiple Targets, Redundancy and Interaction, JMLR Workshop and Conference Proceedings

Future work

- Specific local models (GA –specific local models)
- Integrated missing data management and GA analysis (cf. imputation)
- Noisy genotyping ➔ probabilistic data (see poster)
- Integrated haplotype reconstruction (see poster)
- Integrated study design and analysis (see poster)
- Scaling computation up to ~1000 variables