Bayesian Network Modelling

in Genetics and Systems Biology

Marco Scutari

m.scutari@ucl.ac.uk
Genetics Institute
University College London

October 15, 2013

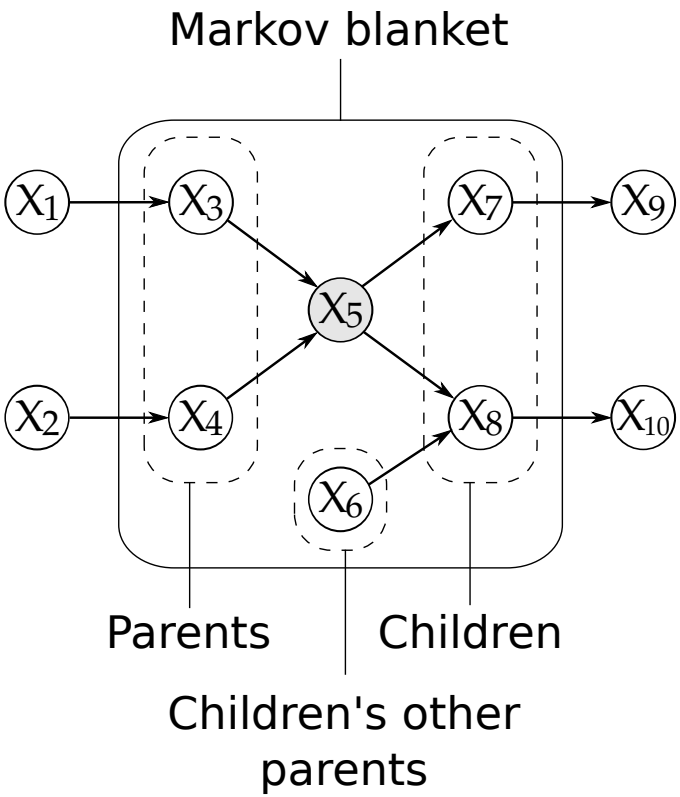
Bayesian Networks: an Overview

A Bayesian network (BN) [14, 19] is a combination of:

- a directed graph (DAG) $\mathcal{G} = (\mathbf{V}, A)$, in which each node $v_i \in \mathbf{V}$ corresponds to a random variable X_i (a gene, a trait, an environmental factor, etc.);
- a global probability distribution over $\mathbf{X} = \{X_i\}$, which can be split into simpler local probability distributions according to the arcs $a_{ij} \in A$ present in the graph.

This combination allows a compact representation of the joint distribution of high-dimensional problems, and simplifies inference using the graphical properties of $\mathcal{G}$. Under some additional assumptions arcs may represent causal relationships [20].

The Two Main Properties of Bayesian Networks



The defining characteristic of BNs is that graphical separation implies (conditional) probabilistic independence. As a result, the global distribution factorises into local distributions: each is associated with a node X_i and depends only on its parents Π_{X_i},

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathrm{P}(X_i \mid \Pi_{X_i}).$$

In addition, we can visually identify the Markov blanket of each node X_i (the set of nodes that completely separates X_i from the rest of the graph, and thus includes all the knowledge needed to do inference on X_i).

Bayesian Networks in Genetics & Systems Biology

Bayesian networks are versatile and have several potential applications because:

- dynamic Bayesian networks can model dynamic data [8, 13, 15];
- learning and inference are (partly) decoupled from the nature of the data, many algorithms can be reused changing tests/scores [18];
- genetic, experimental and environmental effects can be accommodated in a single encompassing model [22];
- interactions can be learned from the data [16], specified from prior knowledge or anything in between [17, 2];
- efficient inference techniques for prediction and significance testing are mostly codified.

Data: SNPs [16, 9], expression data [2, 22], proteomics [22], metabolomics [7], and more...

Markov Blankets for Feature Selection

Markov Blankets can Preserve Prediction Power

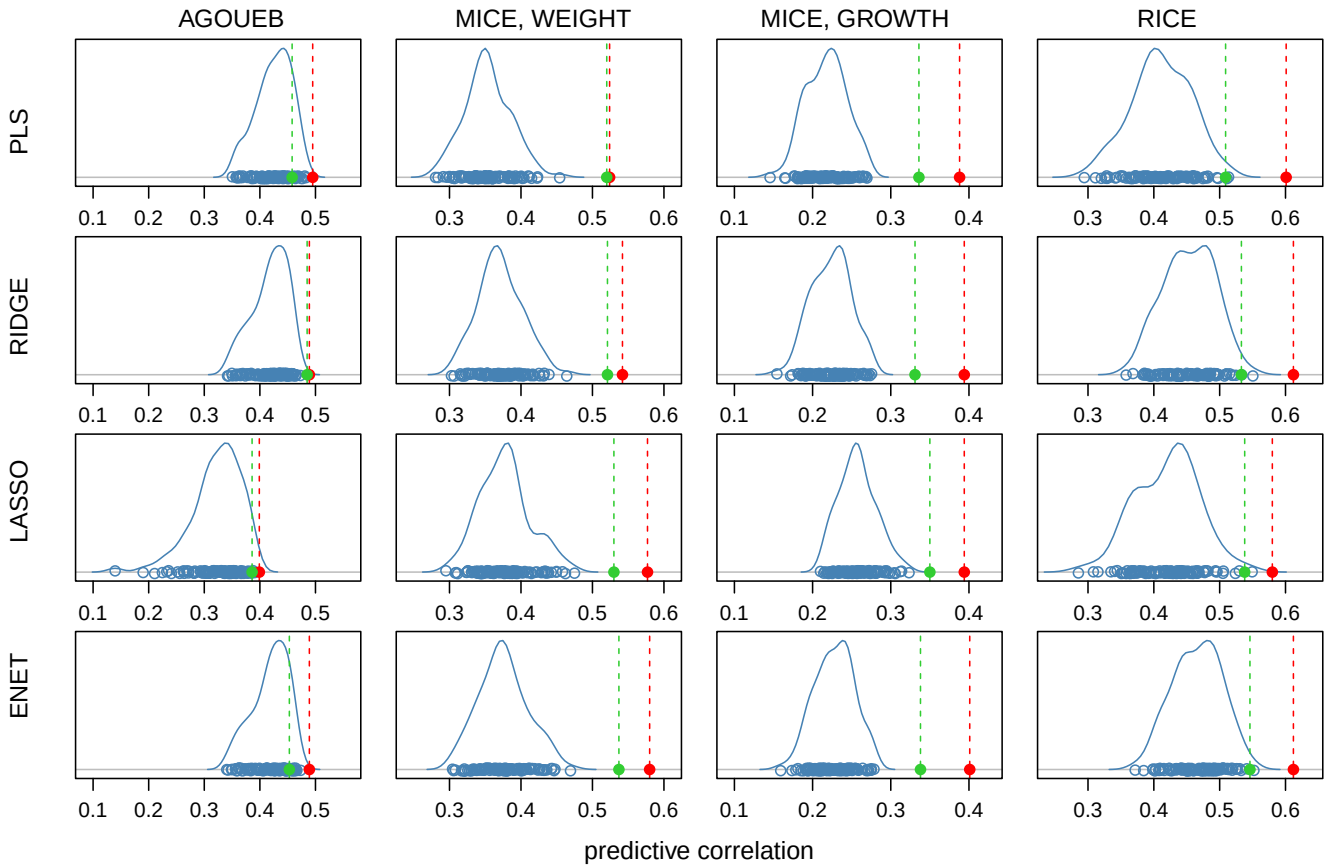
Model	ρ_{CV}	$\rho_{CV,MB}$	Δ
AGOUEB, YIELD (185/810 SNPs, 23%)			
PLS	0.495	0.495	+0.000
Ridge	0.501	0.489	−0.012
LASSO	0.400	0.399	−0.001
Elastic Net	0.500	0.489	−0.011
MICE, GROWTH RATE (543/12.5K SNPs, 4%)			
PLS	0.344	0.388	+0.044
Ridge	0.366	0.394	+0.028
LASSO	0.390	0.394	+0.004
Elastic Net	0.403	0.401	−0.001
MICE, WEIGHT (525/12.5K SNPs, 4%)			
PLS	0.502	0.524	+0.022
Ridge	0.526	0.542	+0.016
LASSO	0.579	0.577	−0.001
Elastic Net	0.580	0.580	+0.000
RICE, SEEDS PER PANICLE (293/74K SNPs, 0.4%)			
PLS	0.583	0.601	+0.018
Ridge	0.601	0.612	+0.011
LASSO	0.516	0.580	+0.064
Elastic Net	0.602	0.612	+0.010

Predictions based Markov blankets may have the same precision as genome-wide predictions for large $\alpha(\simeq 0.15)$ [25]. The data:

- AGOUEB (227 obs.): winter barley, yield [30, 3, 21];
- MICE (1940 obs.): WTCCC heterogeneous mouse populations, more than 100 traits [27, 29];
- RICE (413 obs.): *Oryza sativa* rice, 34 recorded traits [31].

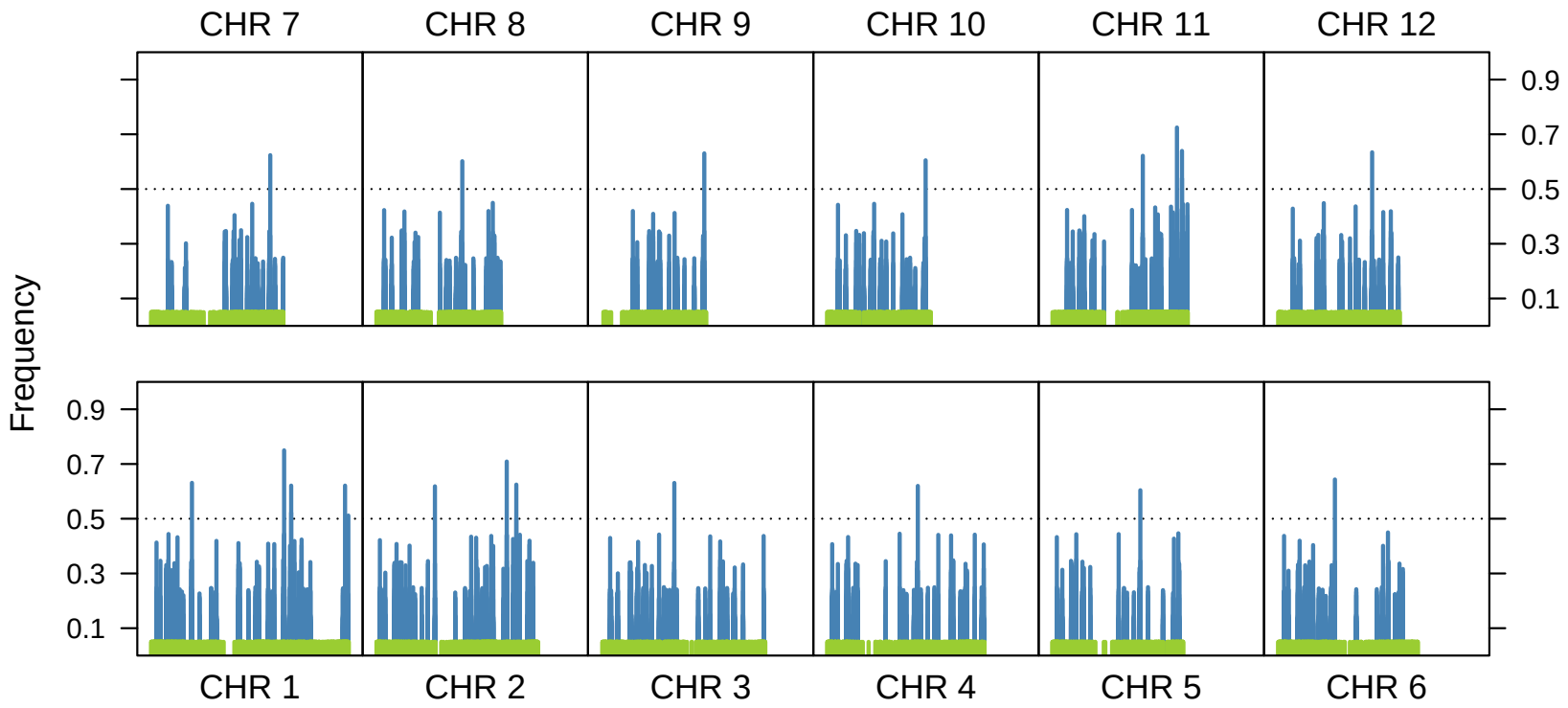
We observe no loss in predictive power after the Markov blanket feature selection. In fact, the reduced number of SNPs increases numerical stability and slightly improves the predictive power of the models.

More Informative with the Same Number of SNPs



Blue dots are random subsets, red dots are Markov blankets, green dots are single-SNP analyses, all with the same number of SNPs.

Markov Blankets and Mapping Information



Green ticks indicate the positions of all mapped SNPs for the RICE data; blue bars indicate the frequency of the SNPs included in the Markov blankets estimated from the rice data using cross-validation.

Causal Protein-Signalling Network from Sachs et al.

Source and Overview of the Data

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data
Karen Sachs, *et al.*
Science **308**, 523 (2005);
DOI: 10.1126/science.1105809

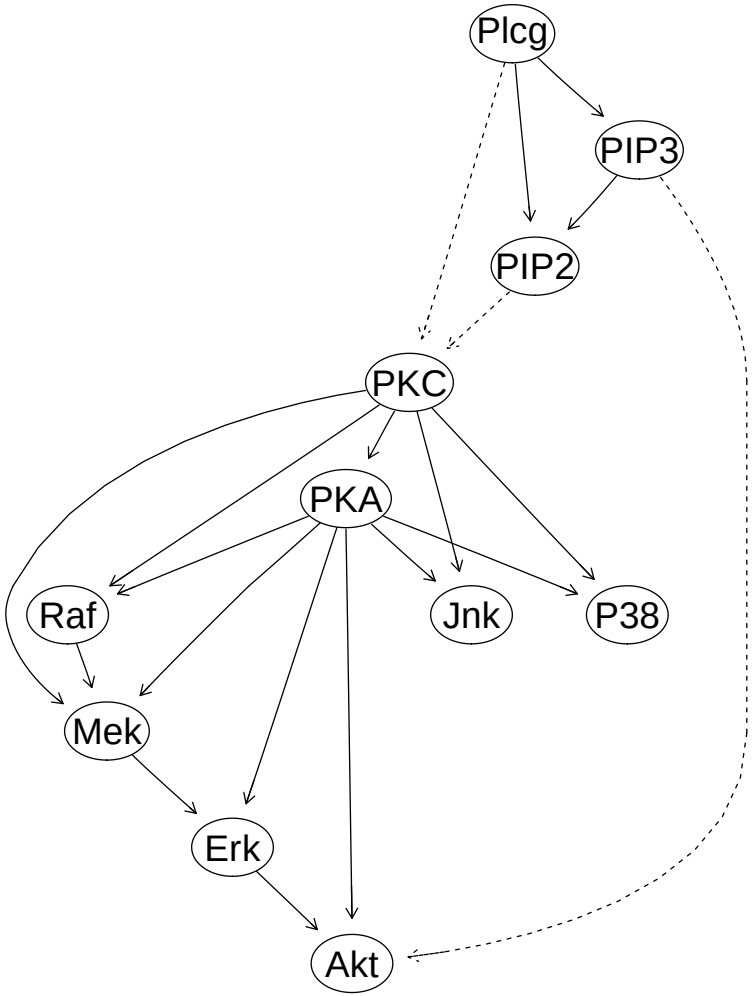
That's a landmark paper in applying Bayesian Networks because:

- it highlights the use of observational vs interventional data;
- results are validated using existing literature.

The data consist in the 5400 simultaneous measurements of 11 phosphorylated proteins and phospholypids derived from thousands of individual primary immune system cells:

- 1800 data subject only to general stimolatory cues, so that the protein signalling paths are active;
- 600 data with with specific stimolatory/inhibitory cues for each of the following 4 proteins: Mek, PIP2, Akt, PKA;
- 1200 data with specific cues for PKA.

Analysis and Validated Network



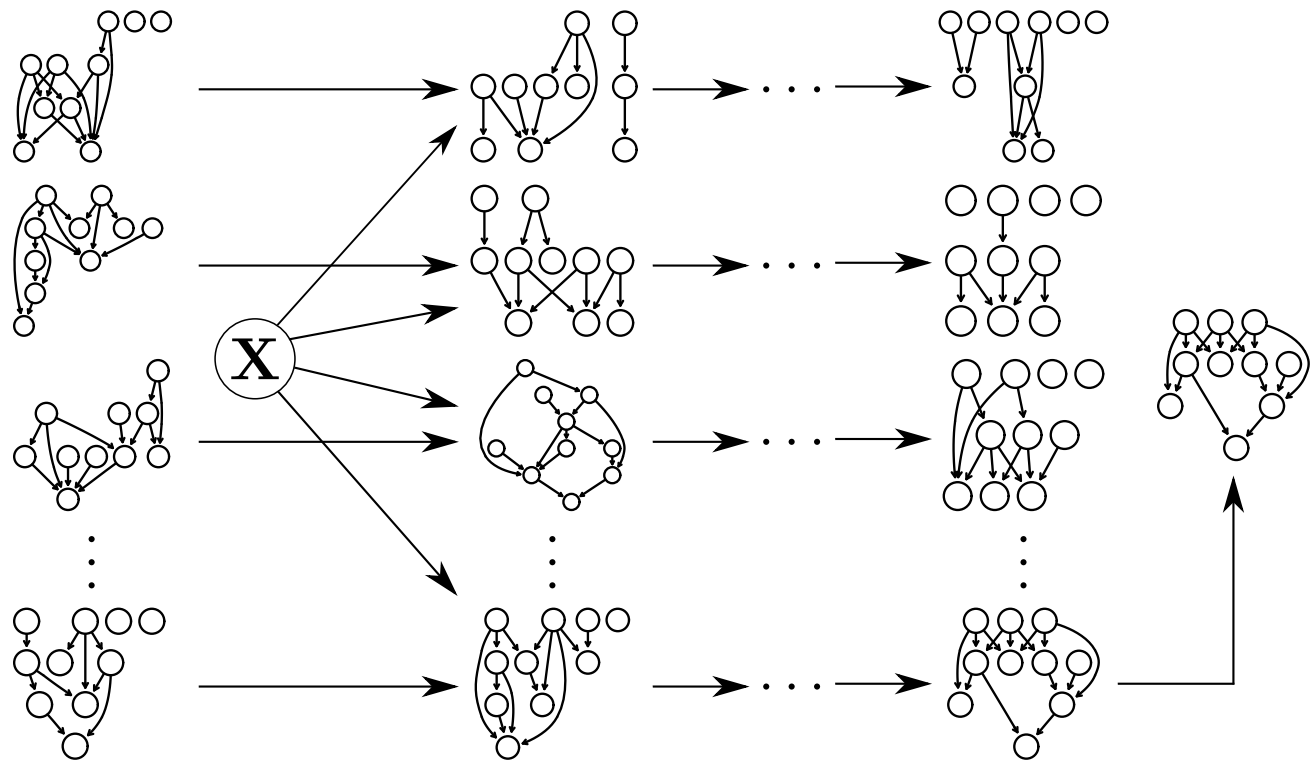
1. Outliers were removed and the data were discretised using the approach described in [10].
2. A large number of DAGs were learned and averaged to produce a more robust model. The averaged DAG was created using the arcs present in at least 85% of the DAGs.
3. The validity of the averaged BN was evaluated against established signalling pathways from literature.

Discretising Gene Expression Data

Hartemink's Information Preserving Discretisation [10]:

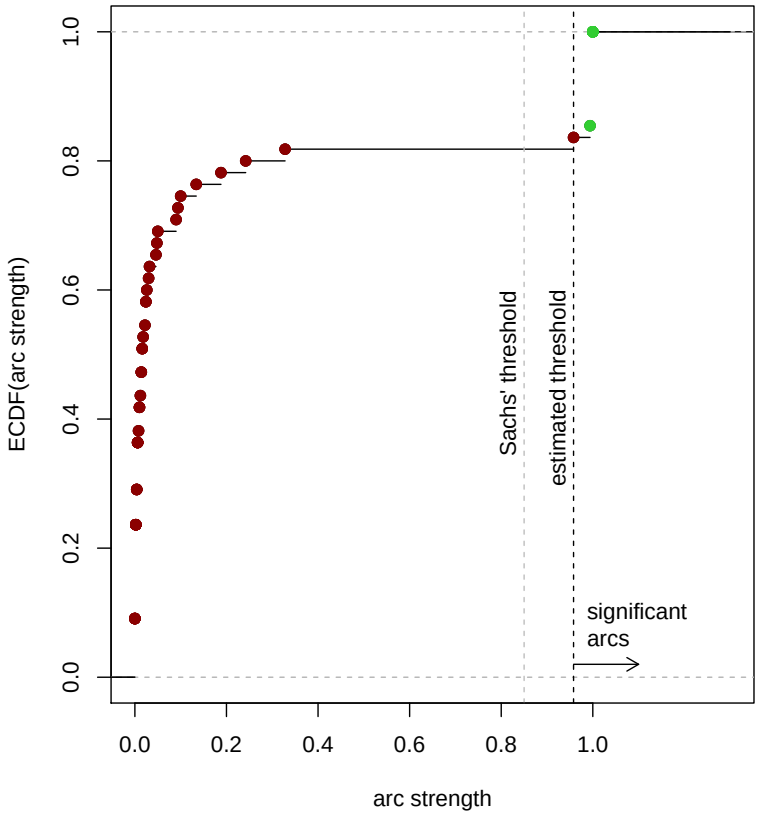
1. Discretise each variable independently using quantiles and a large number k_1 of intervals, *e.g.* $k_1 = 50$ or even $k_1 = 100$.
2. Repeat the following steps until each variable has $k_2 \ll k_1$ intervals, iterating over each variable X_i, $i = 1, \ldots, p$ in turn:
   - 2.1 compute pairwise mutual information coefficients
   $$\mathrm{M}_{X_i} = \sum_{j \neq i} \mathrm{MI}(X_i, X_j);$$
   - 2.2 collapse each pair l of adjacent intervals of X_i in a single interval, and from the resulting variable $X_i^*(l)$ compute
   $$\mathrm{M}_{X_i^*(l)} = \sum_{j \neq i} \mathrm{MI}(X_i^*(l), X_j);$$
   - 2.3 keep the best $X_i^*(l)$: $X_i = \mathrm{argmax}_{X_i(l)} \mathrm{M}_{X_i^*(l)}$.

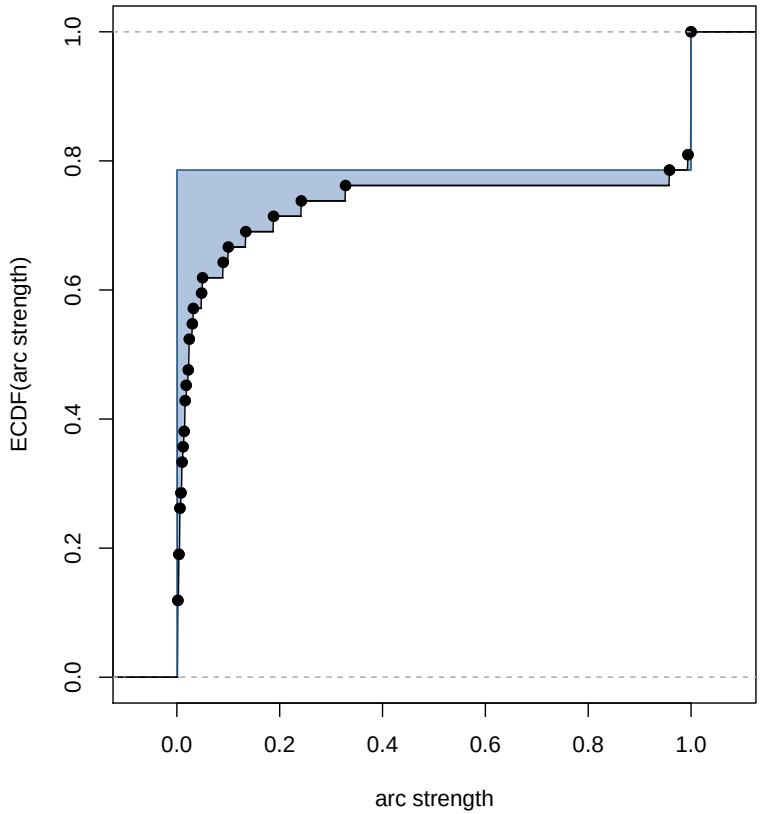
Learning Multiple DAGs from the Data



Searching for high-scoring models from different starting points models increases our coverage of the space of the possible DAGs; the frequency with which an arc appears is a measure of the strength of the dependence.

Model Averaging for DAGs

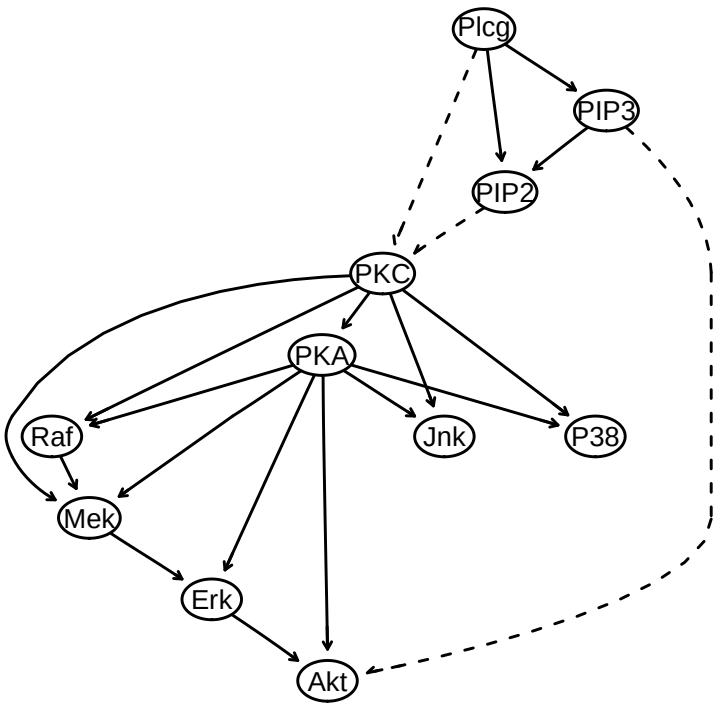




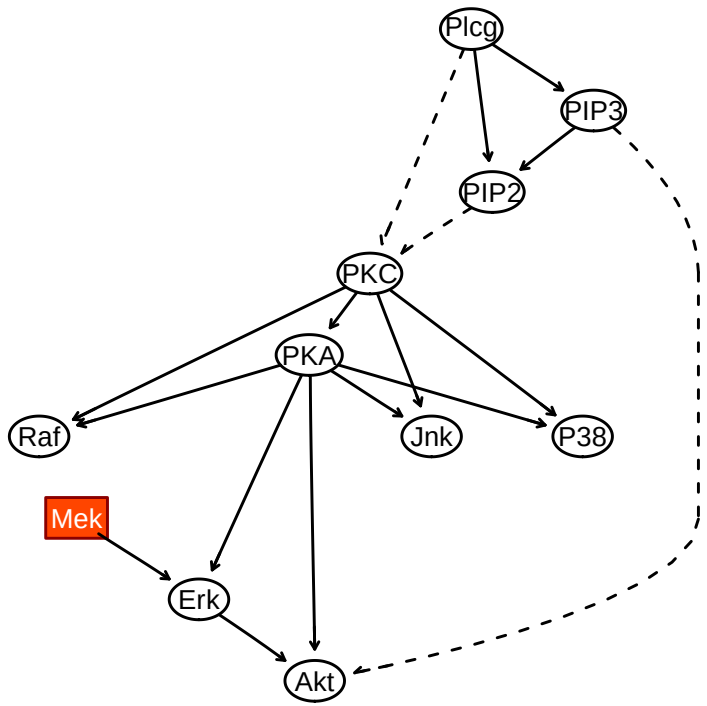
Arcs with significant strength can be identified using a threshold [26] estimated from the data by minimising the distance from the observed ECDF and the ideal, asymptotic one (the blue area in the right panel).

Combining Observational and Interventional Data

model without interventions



model with interventions



Observations must be scored taking into account the effects of the interventions, which break biological pathways; the overall network score is a mixture of scores adjusted for each experiment [4].

Genomic Selection and Genome-Wide Association Studies

Bayesian Networks for GS and GWAS

From the definition, if we have a set of traits and markers for each variety, all we need for GS and GWAS are the Markov blankets of the traits [25]. Using common sense, we can make some additional assumptions:

- traits can depend on markers, but not vice versa;
- traits that are measured after the variety is harvested can depend on traits that are measured while the variety is still in the field (and obviously on the markers as well), but not vice versa.

Most markers are discarded when the Markov blankets are learned. Only those that are parents of one or more traits are retained; all other markers' effects are indirect and redundant once the Markov blankets have been learned. Assumptions on the direction of the dependencies allow to reduce Markov blankets learning to learning the parents of each trait, which is a much simpler task.

Learning the Bayesian network

1. Feature Selection.
   1.1 For each trait, use the SI-HITON-PC algorithm [1, 24] to learn the parents and the children of the trait; children can only be other traits, parents are mostly markers, spouses can be either. Dependencies are assessed with Student's t-test for Pearson's correlation [12] and $\alpha = 0.01$.
   1.2 Drop all the markers which are not parents of any trait.
2. Structure Learning. Learn the structure of the BN from the nodes selected in the previous step, setting the directions of the arcs according to the assumptions in the previous slide. The optimal structure can be identified with a suitable goodness-of-fit criterion such as BIC [23]. This follows the spirit of other hybrid approaches [6, 28], that have shown to be well-performing in literature.
3. Parameter Learning. Learn the parameters of the BN as a Gaussian BN [14]: each local distribution in a linear regression and the global distribution is a hierarchical linear model.

The Parameters of the Bayesian Network

The local distribution of each trait X_i is a linear model

$$\begin{aligned} X_i &= \mu + \Pi_{X_i}\boldsymbol{\beta} + \varepsilon \\ &= \mu + \underbrace{X_j\beta_j + \ldots + X_k\beta_k}_{traits} + \underbrace{X_l\beta_l + \ldots + X_m\beta_m}_{markers} + \varepsilon \end{aligned}$$

which can be estimated any frequentist or Bayesian approach in which the nodes in X_i are treated as fixed effects (e.g. ridge regression [11], elastic net [32], etc.).

For each marker X_i, the nodes in Π_{X_i} are other markers in LD with X_i since $\text{COR}(X_i, X_j|\Pi_{X_i}) \neq 0 \Leftrightarrow \beta_j \neq 0$. This is also intuitively true for markers that are children of X_i, as LD is symmetric.

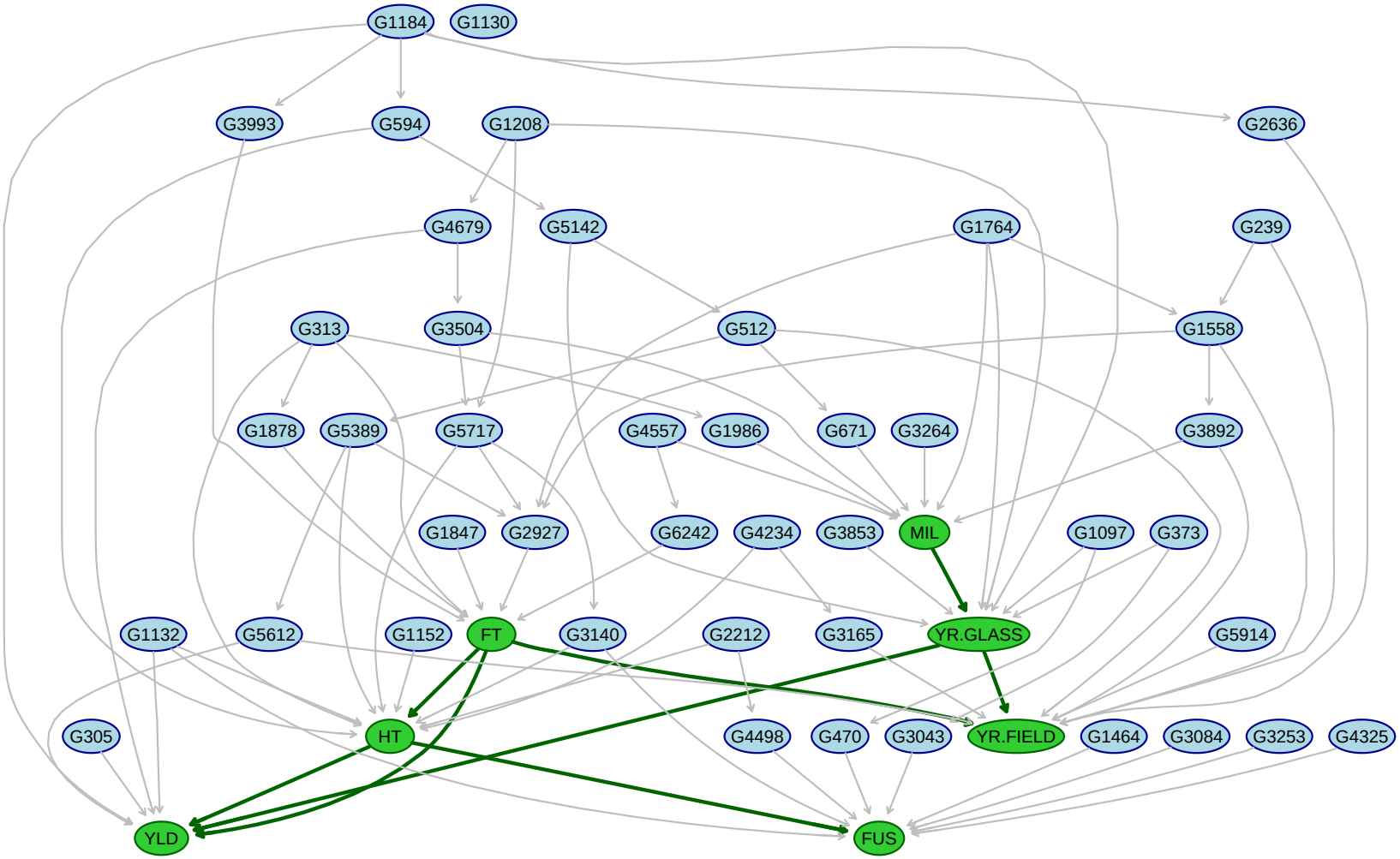
The MAGIC Wheat Data

The MAGIC data include 721 wheat varieties, 16K markers and the following phenotypes:

- flowering time (FT);
- height (HT);
- yield (YLD);
- yellow rust, as measured in the glasshouse (YR.GLASS);
- yellow rust, as measured in the field (YR.FIELD);
- mildew (MIL) and
- fusarium (FUS).

Varieties with missing phenotypes or family information and markers with $> 20\%$ missing data were dropped. The phenotypes were adjusted for family structure via BLUP and the markers screened for MAF > 0.01 and COR < 0.99.

Bayesian Network Learned from MAGIC



51 nodes (7 traits, 44 markers), 86 arcs, 137 parameters for 600 obs.

Assessing Arc Strength with Bootstrap Resampling

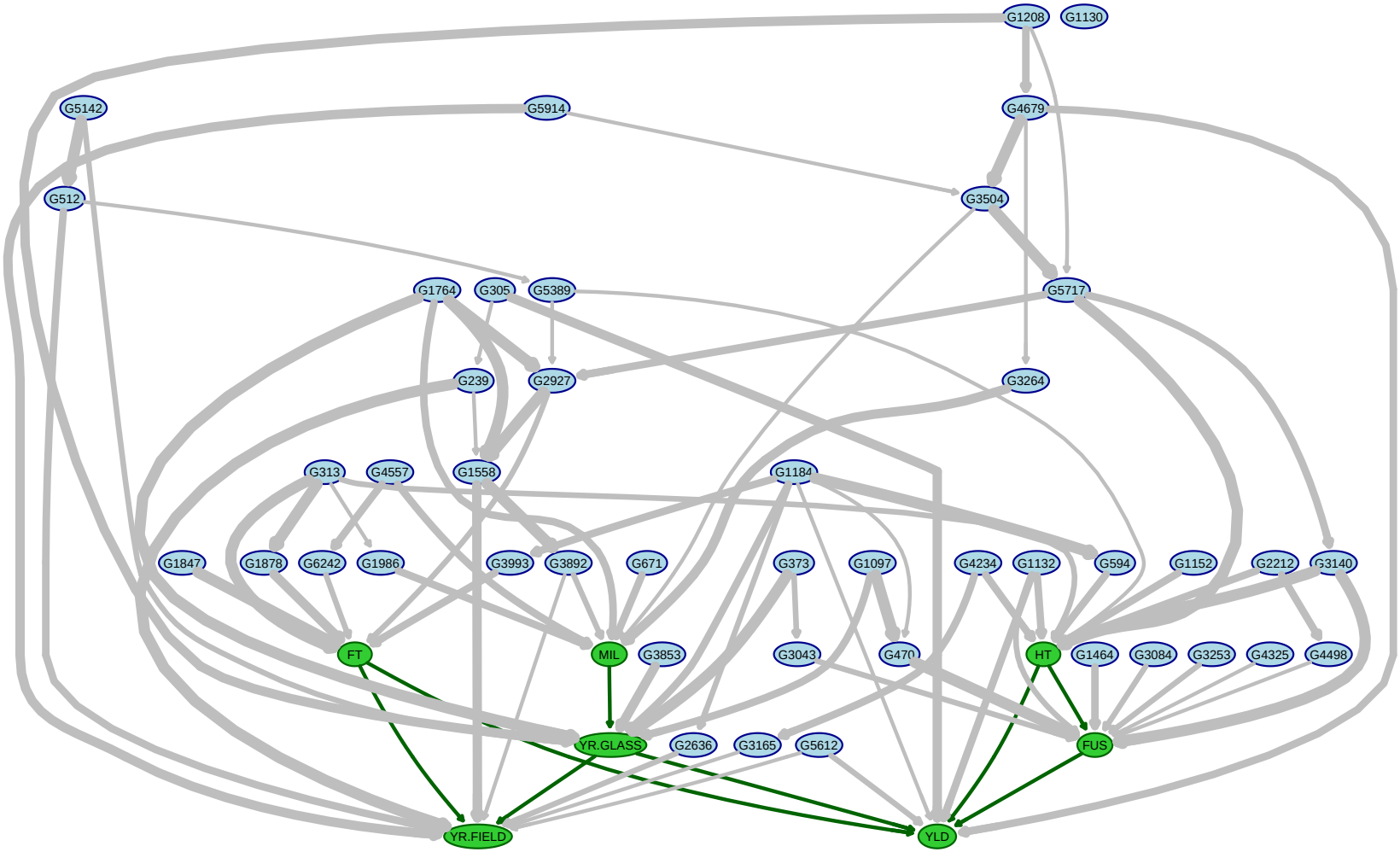
Friedman et al. [5] proposed an approach to assess the strength of each arc based on bootstrap resampling and model averaging:

1. For $b = 1, 2, \ldots, m$:
   - 1.1 sample a new data set $\mathbf{X}_b^*$ from the original data $\mathbf{X}$ using either parametric or nonparametric bootstrap;
   - 1.2 learn the structure of the graphical model $\mathcal{G}_b = (\mathbf{V}, E_b)$ from $\mathbf{X}_b^*$.
2. Estimate the confidence that each possible arc a_i is present in the true network structure $\mathcal{G}_0 = (\mathbf{V}, A_0)$ as

$$\hat{p}_i = \hat{\mathrm{P}}(a_i) = \frac{1}{m} \sum_{b=1}^{m} \mathbb{1}_{\{a_i \in A_b\}},$$

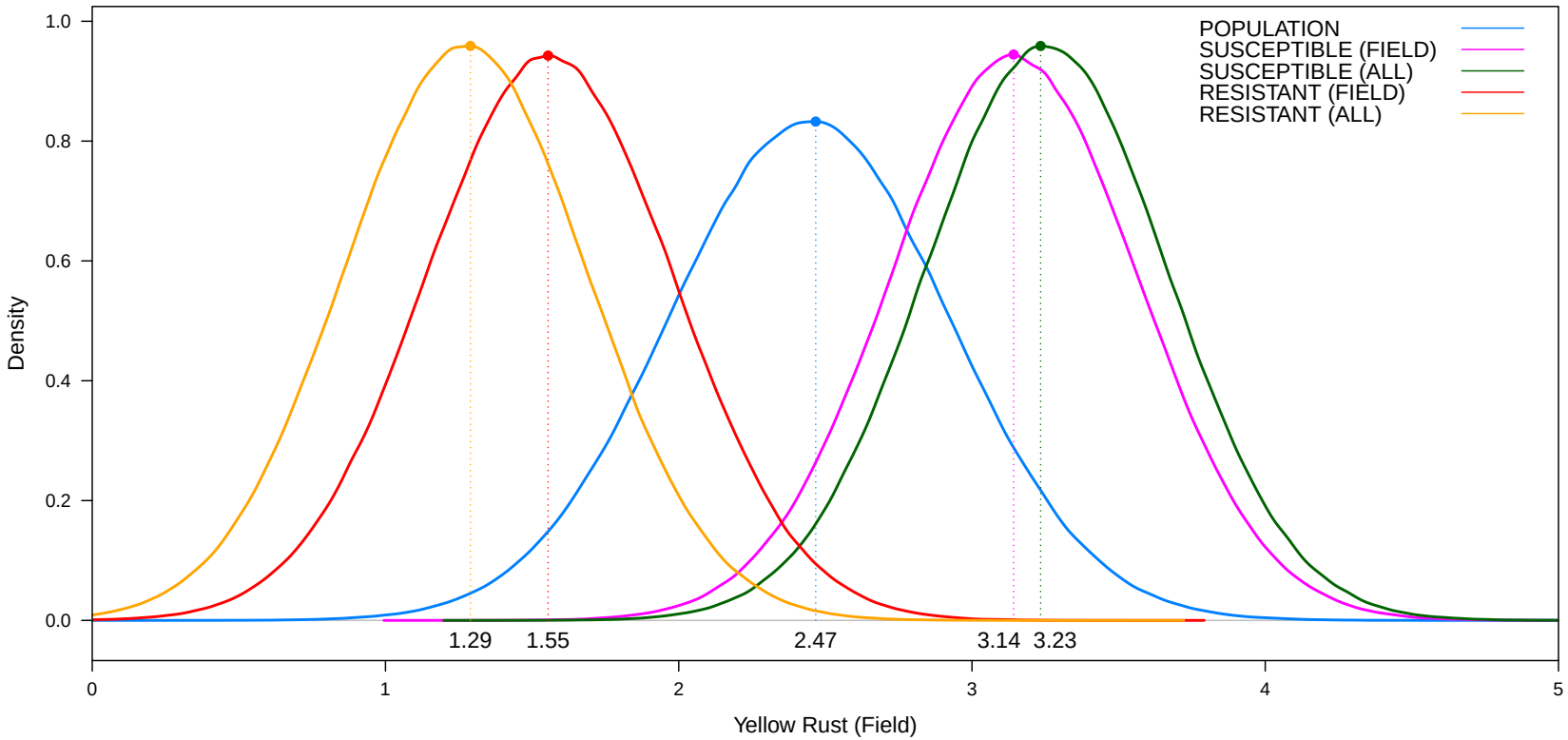
where $\mathbb{1}_{\{a_i \in A_b\}}$ is equal to 1 if $a_i \in A_b$ and 0 otherwise.

Averaged Bayesian Network from MAGIC



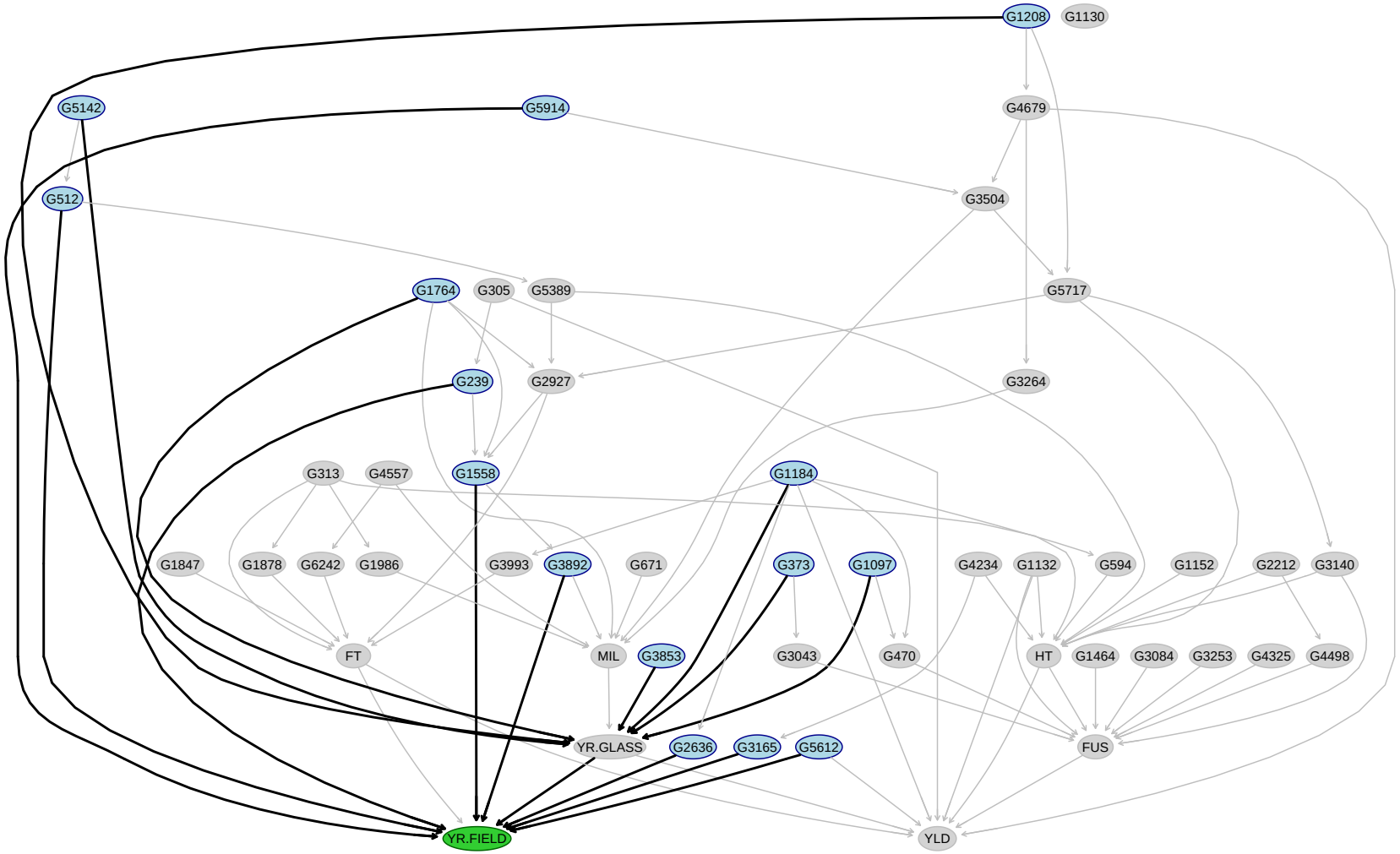
81 out of 86 arcs from the original BN are significant.

Yellow Rust: What if We Fix (In)directly Related Alleles?

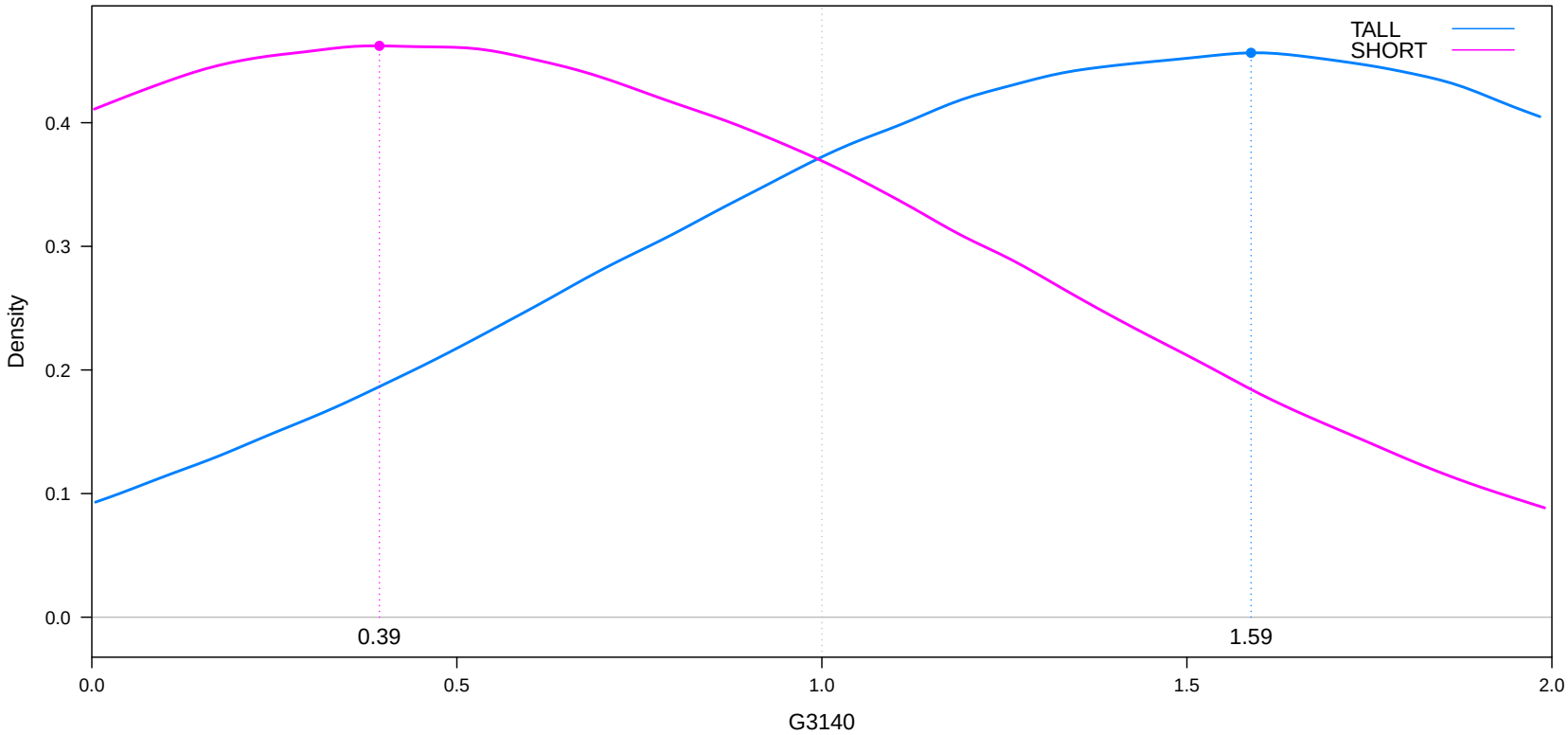


Fixing 8 genes that are parents of YR.FIELD, then another 7 that are parents of YR.GLASS, either to be homozygotes for yellow rust susceptibility or for yellow rust resistance.

Yellow Rust: Nodes Farther Away Can Help...

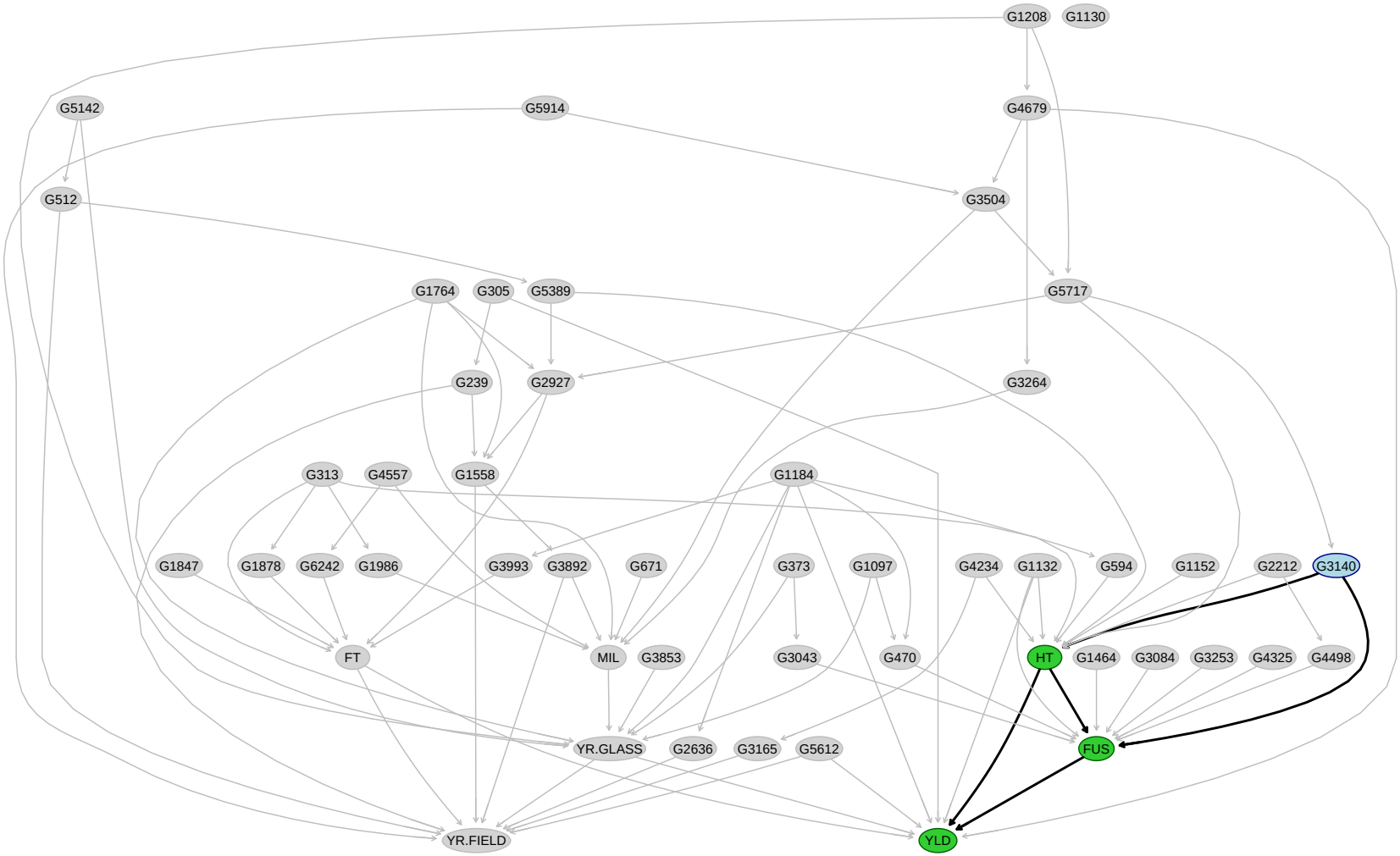


G3140: Can We Guess the Allele?



If we have two varieties for which we scored low levels of fusarium (0 to 2), and are among the top 25% yielding, but one is tall (top 25%) and one is short (bottom 25%), which is the most probable allele for gene G3140?

G3140: Information Travels Backwards...



Conclusions

Conclusions

- Bayesian networks provide an intuitive representation of the relationships linking sets of phenotypes and genotypes, both between and within each other.
- Given a few reasonable assumptions, we can learn a Bayesian network for multiple trait GWAS and GS efficiently and reusing state-of-the-art general-purpose algorithms.
- Once learned, Bayesian networks provide a flexible tool for inference on both the markers and the phenotypes.
- Markov blankets are a valuable tool for feature selection, even when we are not learning a complete Bayesian network.

Thanks!

References

References I

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Xenofon.

Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation.

Journal of Machine Learning Research, 11:171–234, 2010.

K. C. Chipman and A. K. Singh.

Using Stochastic Causal Trees to Augment Bayesian Networks for Modeling eQTL Datasets.

BMC Bioinformatics, 12(7):1–17, 2011.

J. Cockram, J. White, D. L. Zuluaga, D. Smith, J. Comadran, M. Macaulay, Z. Luo, M. J. Kearsey, P. Werner, D. Harrap, C. Tapsell, H. Liu, P. E. Hedley, N. Stein, D. Schulte, B. Steuernagel, D. F. Marshall, W. T. Thomas, L. Ramsay, I. Mackay, D. J. Balding, The AGOUEB Consortium, R. Waugh, and D. M. O'Sullivan.

Genome-Wide Association Mapping to Candidate Polymorphism Resolution in the Unsequenced Barley Genome.

PNAS, 107(50):21611–21616, 2010.

G. F. Cooper and C. Yoo.

Causal Discovery from a Mixture of Experimental and Observational Data.

In *UAI '99: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, pages 116–125. Morgan Kaufmann, 1995.

N. Friedman, M. Goldszmidt, and A. Wyner.

Data Analysis with Bayesian Networks: A Bootstrap Approach.

In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 196 – 205. Morgan Kaufmann, 1999.

References II

N. Friedman, D. Pe'er, and I. Nachman.

Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm.
In *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 206–221. Morgan Kaufmann, 1999.

A. K. Gavai, Y. Tikunov, R. Ursem, A. Bovy, F. van Eeuwijk, H. Nijveen, P. J. F. Lucas, and J. A. M. Leunissen.

Constraint-Based Probabilistic Learning of Metabolic Pathways from Tomato Volatiles.
Metabolomics, 5(4):419–428, 2005.

M. Grzegorczyk and D. Husmeier.

Non-Stationary Continuous Dynamic Bayesian Networks.
Advances in Neural Information Processing Systems (NIPS), 22:682–690, 2009.

B. Han, X. Chen, Z. Talebizadeh, and H. Xu.

Genetic Studies of Complex Human Diseases: Characterizing SNP-Disease Associations Using Bayesian Networks.
BMC Systems Biology, 6(Suppl. 3):S14, 2012.

A. J. Hartemink.

Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks.
PhD thesis, School of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2001.

A. E. Hoerl and R. W. Kennard.

Ridge Regression: Biased Estimation for Nonorthogonal Problems.
Technometrics, 12(1):55–67, 1970.

References III

H. Hotelling.

New Light on the Correlation Coefficient and Its Transforms.
Journal of the Royal Statistical Society. Series B (Methodological), 15(2):193–232, 1953.

D. Husmeier.

Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks.
Bioinformatics, 19:2271–2282, 2003.

D. Koller and N. Friedman.

Probabilistic Graphical Models: Principles and Techniques.
MIT Press, 2009.

G. Lelandais and S. Lèbre.

Recovering Genetic Network from Continuous Data with Dynamic Bayesian Networks.
In D. J. Balding, M. Stumpf, and M. Girolami, editors, *Handbook of Statistical Systems Biology*. Wiley, 2011.

G. Morota, B. D. Valente, G. J. M. Rosa, K. A. Weigel, and D. Gianola.

An Assessment of Linkage Disequilibrium in Holstein Cattle Using a Bayesian Network.
Journal of Animal Breeding and Genetics, 129(6):474–487, 2012.

S. Mukherjee and T. P. Speed.

Network Inference using Informative Priors.
PNAS, 105:14313–14318, 2008.

References IV

R. Nagarajan, M. Scutari, and S. Lèbre.

Bayesian Networks in R with Applications in Systems Biology.
Use R! series. Springer, 2013.

J. Pearl.

Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann, 1988.

J. Pearl.

Causality: Models, Reasoning and Inference.
Cambridge University Press, 2nd edition, 2009.

N. Rostoks, L. Ramsay, K. MacKenzie, L. Cardle, P. R. Bhat, M. L. Roose, J. T. Svensson, N. Stein, R. K. Varshney, D. F. Marshall, A. Graner, T. J. Close, and R. Waugh.

Recent History of Artificial Outcrossing Facilitates Whole-Genome Association Mapping in Elite Inbred Crop Varieties.
PNAS, 106(49):18656–18661, 2006.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan.

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data.
Science, 308(5721):523–529, 2005.

G. E. Schwarz.

Estimating the Dimension of a Model.
Annals of Statistics, 6(2):461 – 464, 1978.

References V

M. Scutari.

bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference, 2013.
R package version 3.3.

M. Scutari, I. Mackay, and D. J. Balding.

Improving the Efficiency of Genomic Selection.
Statistical Applications in Genetics and Molecular Biology, 2013.
Submitted.

M. Scutari and R. Nagarajan.

On Identifying Significant Edges in Graphical Models of Molecular Networks.
Artificial Intelligence in Medicine, 57(3):207–217, 2013.
Special Issue containing the Proceedings of the Workshop "Probabilistic Problem Solving in Biomedicine" of the 13th Artificial Intelligence in Medicine (AIME) Conference, Bled (Slovenia), July 2, 2011.

L. C. Solberg, W. Valdar, D. Gauguier, G. Nunez, A. Taylor, S. Burnett, C. Arboledas-Hita, P. Hernandez-Pliego, S. Davidson, P. Burns, S. Bhattacharya, T. Hough, D. Higgs, P. Klenerman W. O. Cookson, Y. Zhang, R. M. Deacon, J. N. Rawlins, R. Mott, and J. Flint.

A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice.
Mamm. Genome, 17:129–146, 2006.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis.

The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.
Machine Learning, 65(1):31–78, 2006.

References VI

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. Rawlins, R. Mott, and J. Flint.
Genome-Wide Genetic Association of Complex Traits in Heterogeneous Stock Mice.
Nat. Genet., 8:879–887, 2006.

R. Waugh, D. Marshall, B. Thomas, J. Comadran, J. Russell, T. Close, N. Stein, P. Hayes, G. Muehlbauer, J. Cockram, D. O'Sullivan, I. Mackay, A. Flavell, AGOUEB, BarleyCAP, and L. Ramsay.
Whole-Genome Association Mapping in Elite Inbred Crop Varieties.
Genome, 53(11):967–972, 2010.

K. Zhao, C. Tung, G. C. Eizenga, M. H. Wright, M. L. Ali, A. H. Price, G. J. Norton, M. R. Islam, A. Reynolds, J. Mezey, A. M. McClung, C. D. Bustamante, and S. R. McCouch.
Genome-Wide Association Mapping Reveals a Rich Genetic Architecture of Complex Traits in Oryza Sativa.
Nat. Commun., 2:467, 2011.

H. Zou and T. Hastie.
Regularization and Variable Selection via the Elastic Net.
J. Roy. Stat. Soc. B, 67(2):301–320, 2005.