Bayesian Networks, MAGIC Populations and Multiple Trait Prediction

Marco Scutari

scutari@stats.ox.ac.uk
Department of Statistics
University of Oxford

August 21, 2016

Bayesian Networks and Genomic Prediction

Multi-Trait Genomic Predictions

The availability of dense genome markers and of simultaneous measurements of multiple traits makes it possible to:

- better elucidate genetic architectures in genome-wide association studies; and to
- predict complex traits with low heritabilities in genomic selection programs.

For this purpose, we need statistical models that provide:

- an intuitive representation of the relationships linking both traits and markers;
- competitive predictive accuracy;
- enough flexibility to accommodate heterogeneous variables such as environmental effects;
- easy inference on both markers and traits.

Many Approaches Based on Random Effects Models...

GENOMIC SELECTION

Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy

Yi Jia* and Jean-Luc Jannink*,†

doi:10.1534/genetics.112.144246

Calus and Veerkamp Genetics Selection Evolution 2011, 43:26
http://www.gsejournal.org/content/43/1/26

GSE Genetics Selection Evolution

RESEARCH Open Access

Accuracy of multi-trait genomic selection using different methods

Mario PL Calus* and Roel F Veerkamp

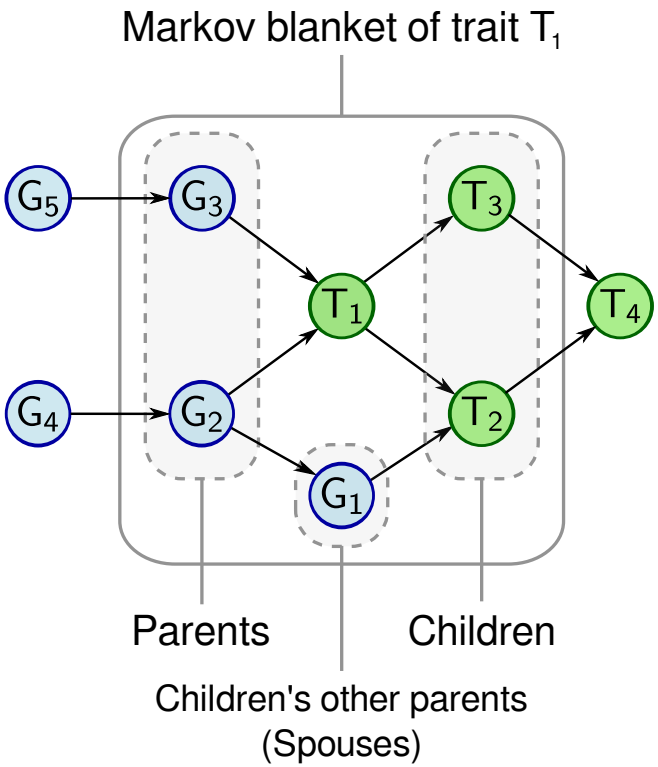
MULTIPARENTAL POPULATIONS

Whole-Genome Analysis of Multienvironment or Multitrait QTL in MAGIC

Arunas P. Verbyla,*,†,1 Colin R. Cavanagh,‡ and Klara L. Verbyla§

doi:10.1534/g3.114.012971

... and My Take using Bayesian Networks



Bayesian networks [4, 10] encode dependencies using a directed acyclic graph, which dictates how the joint distribution of traits and markers factorizes into local distributions: each one 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 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 [8]. Using common sense, we can make some additional assumptions:

- traits can depend on markers, but not vice versa;
- dependencies between traits should follow the order of the respective measurements (e.g. longitudinal traits, traits measured before and after harvest, etc.);
- dependencies in multiple omics data (e.g. SNP + gene expression or SNPs + methylation) should follow the central dogma of molecular biology.

Most markers can be 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 at that point.

Assumptions on the direction of the dependencies allow to reduce Markov blankets learning to learning the parents and the children of each trait, which is a much simpler task.

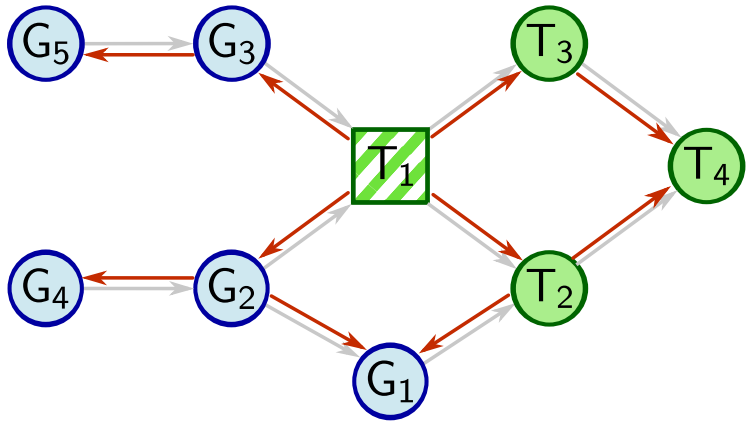
Inference on Traits and Markers

Posterior and causal inference have been studied in depth for Bayesian networks – see [6] from Judea Pearl. Both boil down to computing the probability of some event of interest under different conditions after modifying the corresponding local distributions:

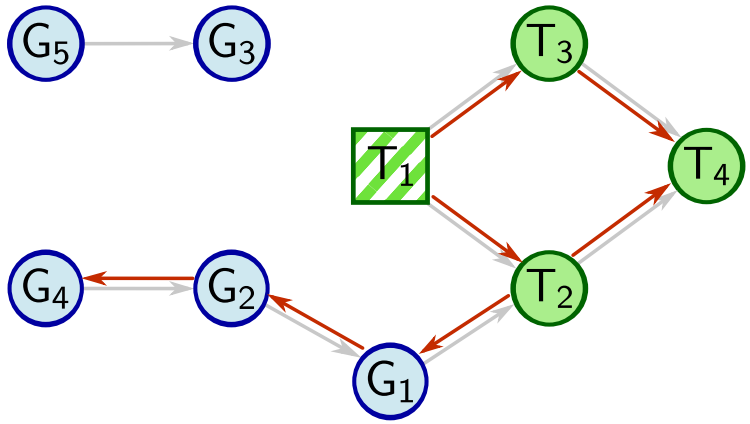
- with exact algorithms such as belief propagation or
- with approximate algorithms such as likelihood weighting.

As a result the question of how to learn the network structure remains the most interesting point from an applied research perspective.

Posterior probability given T_1



Causal intervention on T_1



Learning the Model

Parametric Assumptions

In the spirit of classic additive models, we use a Gaussian Bayesian network and assume the joint distribution of traits and markers is multivariate normal. Then the local distribution of each trait T_i is a linear regression model

$$\begin{aligned} T_i &= \boldsymbol{\mu}_{T_i} + \Pi_{T_i}\boldsymbol{\beta}_{T_i} + \boldsymbol{\varepsilon}_{T_i} \\ &= \boldsymbol{\mu}_{T_i} + \underbrace{T_j\beta_{T_j} + \ldots + T_k\beta_{T_k}}_{\text{traits}} + \underbrace{G_l\beta_{G_l} + \ldots + G_m\beta_{G_m}}_{\text{markers}} + \boldsymbol{\varepsilon}_{T_i}, \quad \boldsymbol{\varepsilon}_{T_i} \sim N(0, \sigma^2_{T_i}\mathbf{I}) \end{aligned}$$

and the local distribution of each marker is likewise

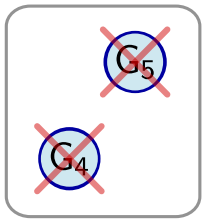
$$\begin{aligned} G_i &= \boldsymbol{\mu}_{G_i} + \Pi_{G_i}\boldsymbol{\beta}_{G_i} + \boldsymbol{\varepsilon}_{G_i} = \\ &= \boldsymbol{\mu}_{G_i} + \underbrace{G_l\beta_{G_l} + \ldots + G_m\beta_{G_m}}_{\text{markers}} + \boldsymbol{\varepsilon}_{G_i}, \qquad \boldsymbol{\varepsilon}_{G_i} \sim N(0, \sigma^2_{G_i}\mathbf{I}). \end{aligned}$$

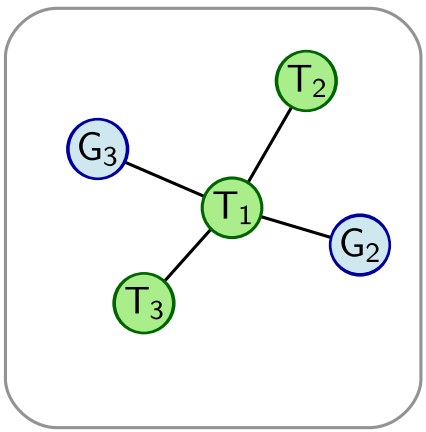
in which the regressors (Π_{T_i} or Π_{G_i}) are treated as fixed effects. Each Π_{G_i} contains the markers in LD with G_i since $\mathrm{COR}(G_i, G_l|\Pi_{G_i}) \neq 0 \Leftrightarrow \beta_{G_l} \neq 0$. This is also intuitively true for markers that are children of G_i, as LD is symmetric.

Learning the Bayesian Network

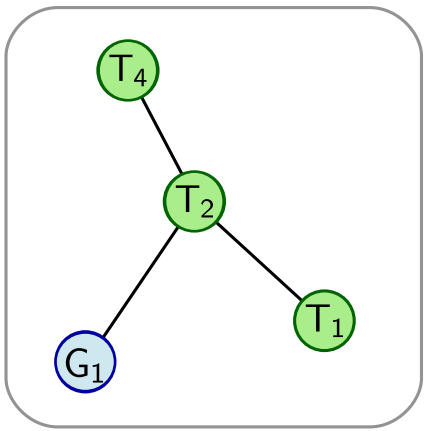
1. Feature Selection.
   - 1.1 Independently learn the parents and the children of each trait with the SI-HITON-PC algorithm [1]; children can only be other traits, parents are mostly markers, spouses can be either. Both are selected using the exact Student's t test for partial correlations.
   - 1.2 Drop all the markers that are not parents of any trait.

Redundant markers that are not in the Markov blanket of any trait

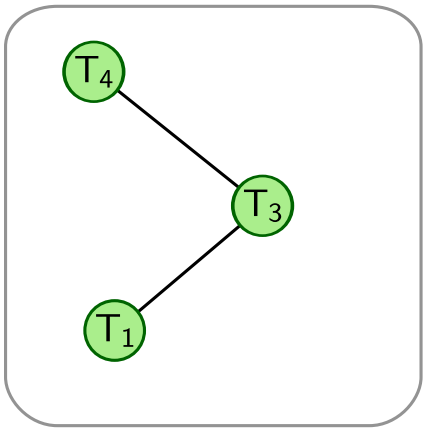




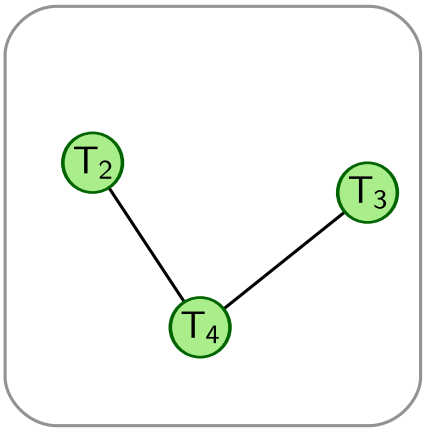
Parents and children of T_1



Parents and children of T_2



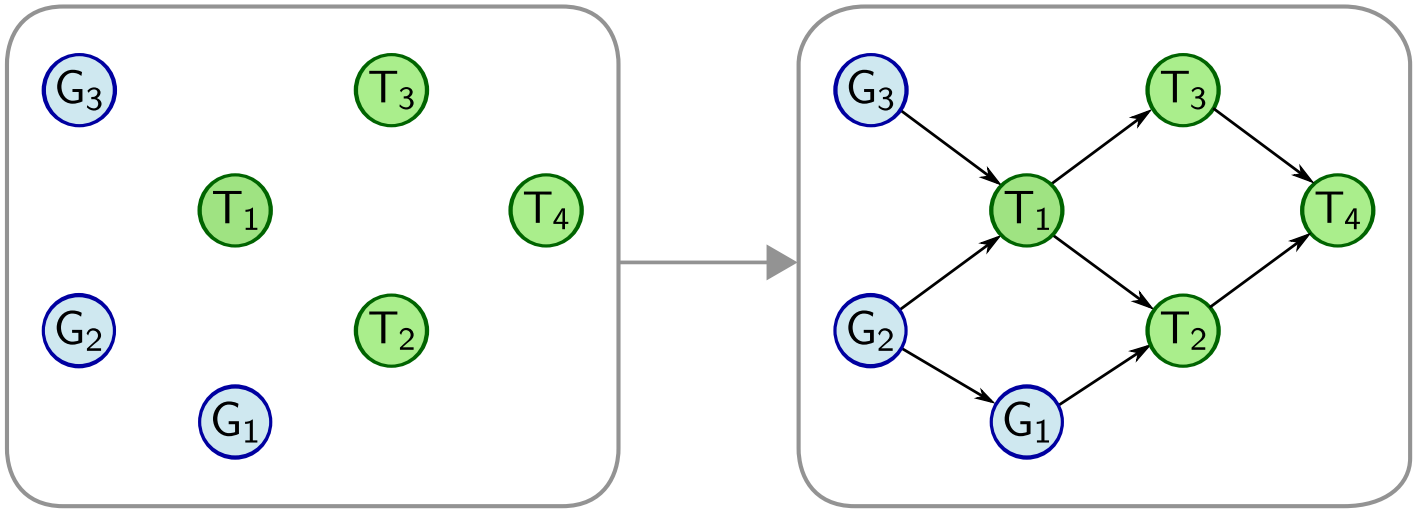
Parents and children of T_3



Parents and children of T_4

Learning the Bayesian Network

2. Structure Learning. Learn the structure of the network from the nodes selected in the previous step, setting the directions of the arcs according to the assumptions above. The optimal structure can be identified with a suitable goodness-of-fit criterion such as BIC. This follows the spirit of other hybrid approaches [3, 12], that have shown to be well-performing in literature.

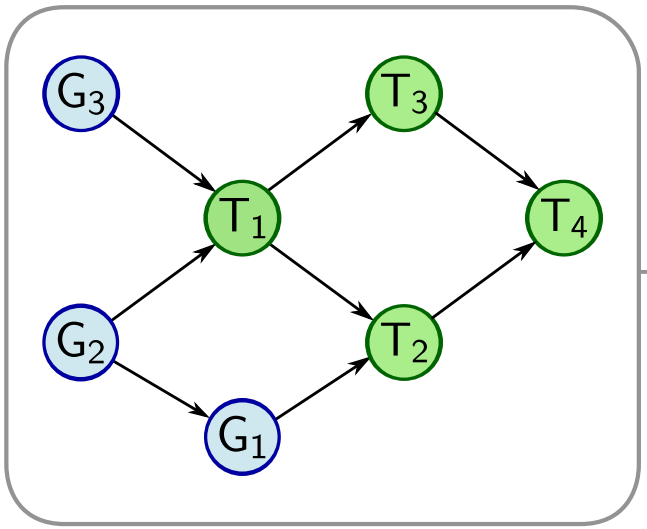


Empty network

Learned network

Learning the Bayesian Network

3. Parameter Learning. Learn the parameters: each local distribution in a linear regression and the global distribution is a hierarchical linear model. Typically least squares works well because SI-HITON-PC selects sets of weakly correlated parents; ridge regression can be used otherwise.



Learned network

$$
\begin{aligned}
G_1 &= \mu_{G_1} + G_2\beta_{G_2} + \varepsilon_{G_1} \\
G_2 &= \mu_{G_2} + \varepsilon_{G_2} \\
G_3 &= \mu_{G_3} + \varepsilon_{G_3} \\
T_1 &= \mu_{T_1} + G_2\beta_{G_2} + G_3\beta_{G_3} + \varepsilon_{T_1} \\
T_2 &= \mu_{T_2} + T_1\beta_{T_1} + G_1\beta_{G_1} + \varepsilon_{T_2} \\
T_3 &= \mu_{T_3} + T_1\beta_{T_1} + \varepsilon_{T_3} \\
T_4 &= \mu_{T_4} + T_2\beta_{T_2} + T_3\beta_{T_3} + \varepsilon_{T_4}
\end{aligned}
$$

Local distributions

The Semi-Interleaved HITON-PC Algorithm

Input: each trait T_i in turn, other traits (T_j) and all markers (G_l), a significance threshold α.
Output: the set $\mathbf{CPC}$ parents and children of T_i in the Bayesian network.

1. Perform a marginal independence test between T_i and each T_j ($T_i \perp\!\!\!\perp T_j$) and G_l ($T_i \perp\!\!\!\perp G_l$) in turn.
2. Discard all T_j and G_l whose p-values are greater than α.
3. Set $\mathbf{CPC} = \{\varnothing\}$.
4. For each the T_j and G_l in order of increasing p-value:
   - 4.1 Perform a conditional independence test between T_i and T_j/G_l conditional on all possible subsets $\mathbf{Z}$ of the current $\mathbf{CPC}$ ($T_i \perp\!\!\!\perp T_j \mid \mathbf{Z} \subseteq \mathbf{CPC}$ or $T_i \perp\!\!\!\perp G_l \mid \mathbf{Z} \subseteq \mathbf{CPC}$).
   - 4.2 If the p-value is smaller than α for all subsets then $\mathbf{CPC} = \mathbf{CPC} \cup \{T_j\}$ or $\mathbf{CPC} = \mathbf{CPC} \cup \{G_l\}$.

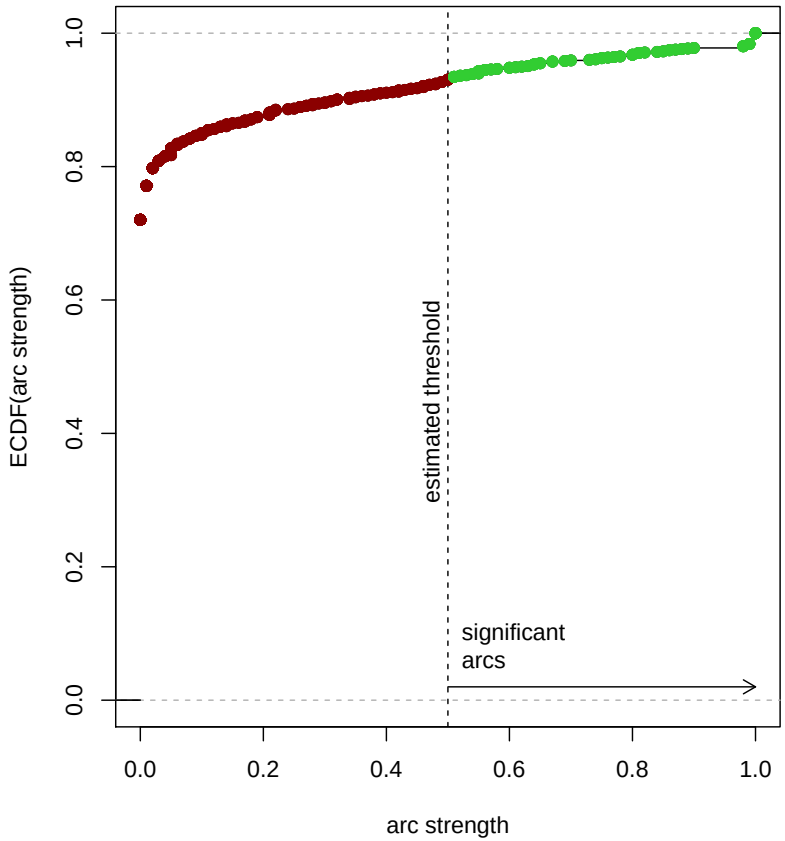
NOTE: the algorithm is defined for a generic independence test, you can plug in any test that is appropriate for the data.

Model Averaging and Assessing Predictive Accuracy

We perform all the above in 10 runs of 10-fold cross-validation to

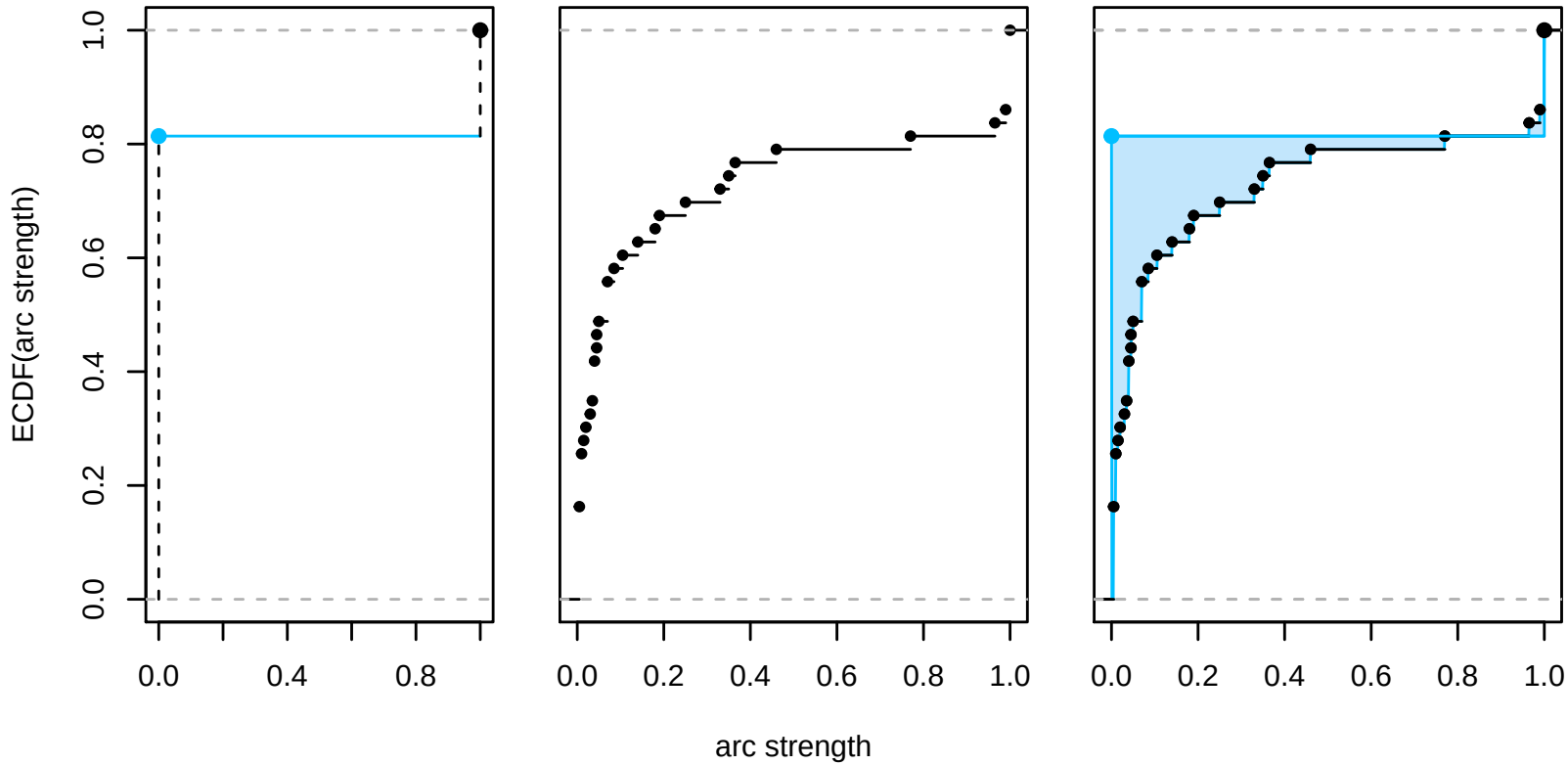
- assess predictive accuracy with e.g. predictive correlation;
- obtain a set of networks to produce an averaged, de-noised consensus network [9]. The threshold for arc strength is estimated from the data.

As a side effect we get a model-agnostic arc strength estimate: the frequency of each arc in the 10×10 Bayesian networks.



Computing a Threshold for Arc Strength

A simple way of computing such a threshold is by considering that Bayesian network learning is consistent, so the empirical distribution function of arc strengths converges to a single-step distribution function as $n \to \infty$. The height of that step is the threshold, which can be estimated by minimizing the area between the empirical and asymptotic distribution functions.



MAGIC Populations

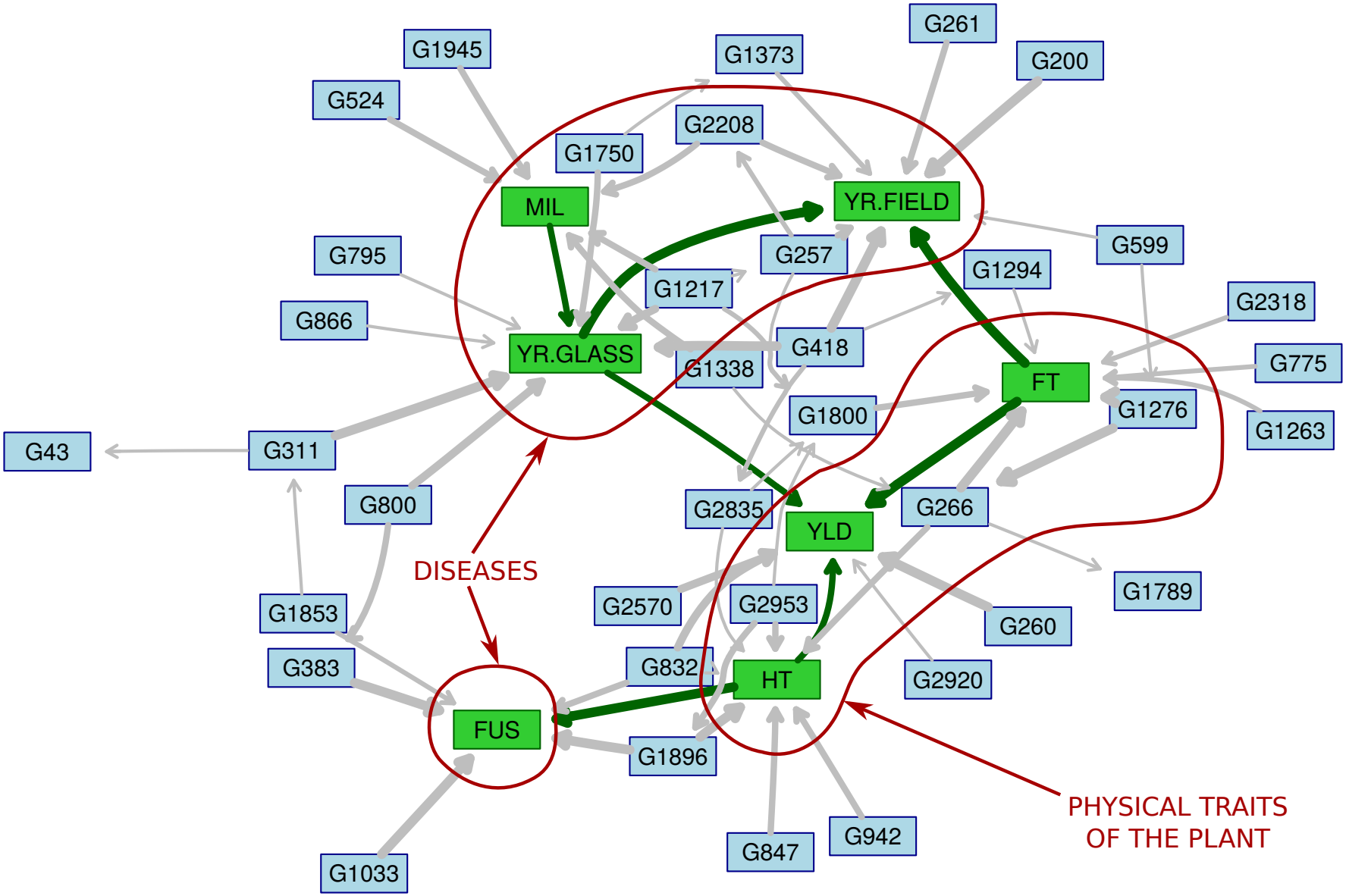
MAGIC Populations: Wheat and Rice

Multiparent advanced generation inter-cross (MAGIC) populations are ideal for learning complex models because of their high genetic recombination, diversity and large sample size. Here we consider two:

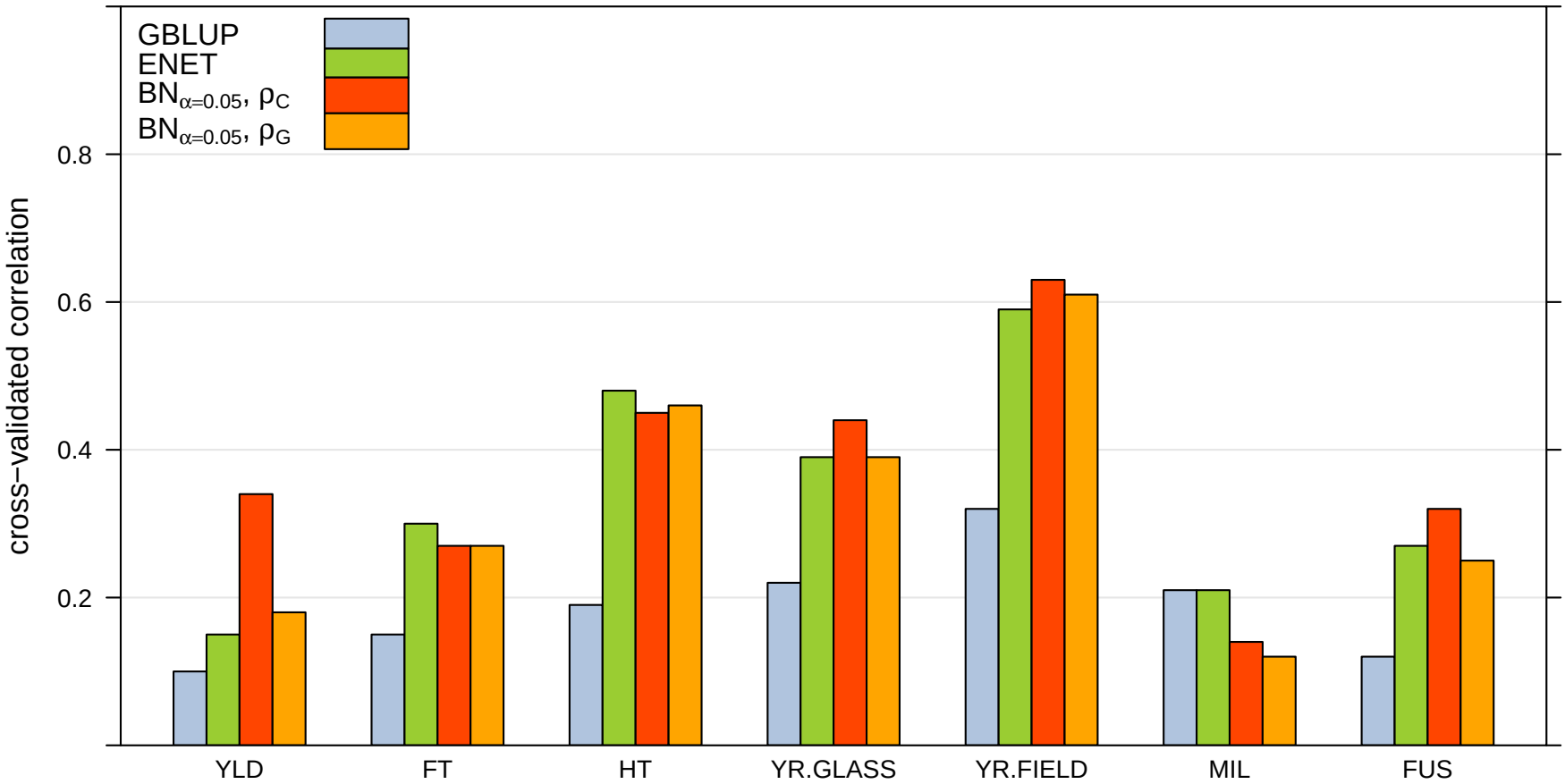
- A winter wheat population [5, 7] with 721 varieties and 16K markers with 7 phenotypes.
- An indica rice population [unpublished] with 1087 varieties and 4K markers with 10 phenotypes.

Phenotypic traits include flowering time (FT), height (HT), yield (YLD), a number of disease scores (YR, MIL, FUS; BROWN); and, in the case of rice, physical and quality traits for the grains (GL, GW, AMY, GT, CHALK).

WHEAT: a Bayesian Network (44 nodes, 66 arcs)

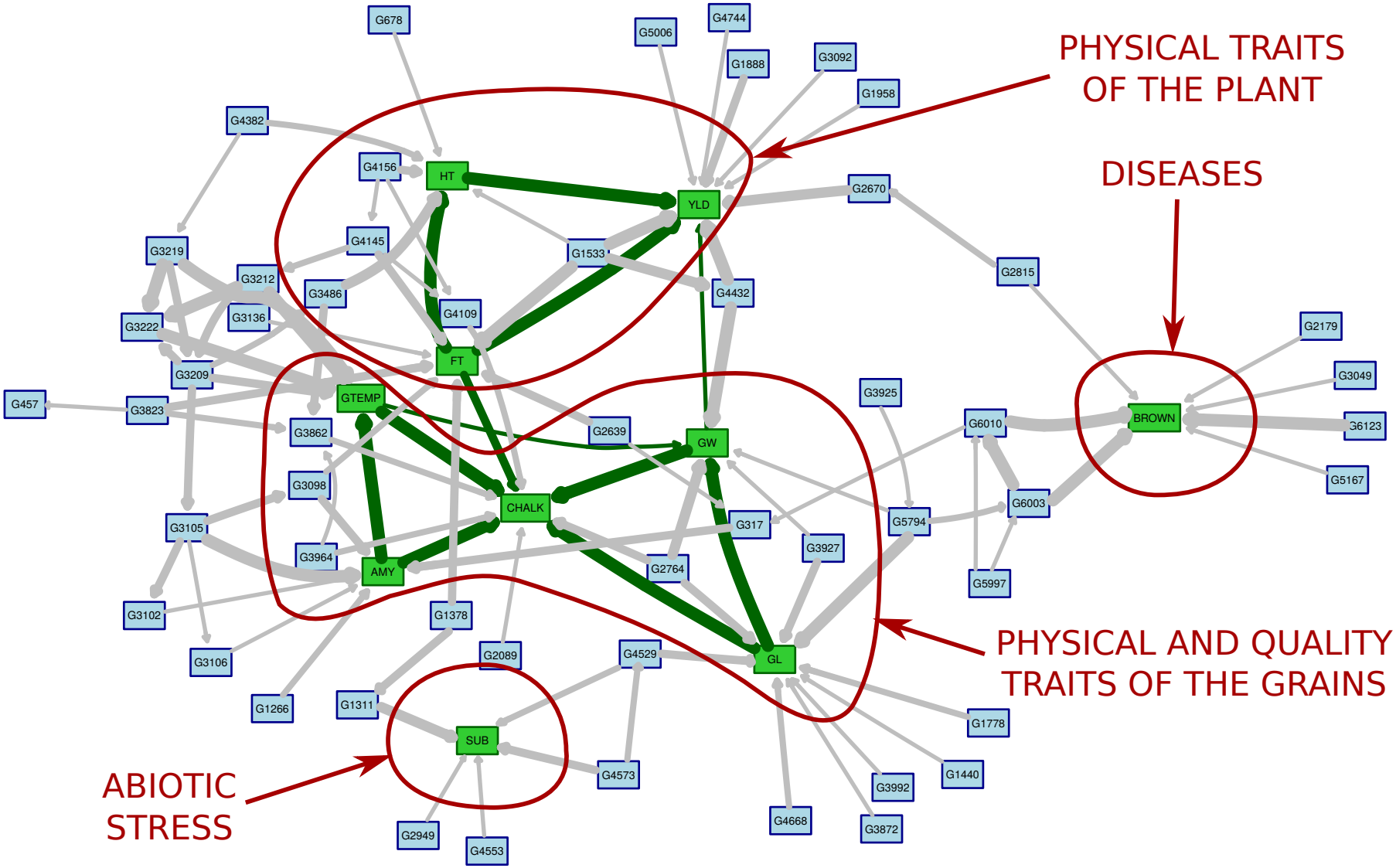


WHEAT: Predictive Performance from Cross-Validation

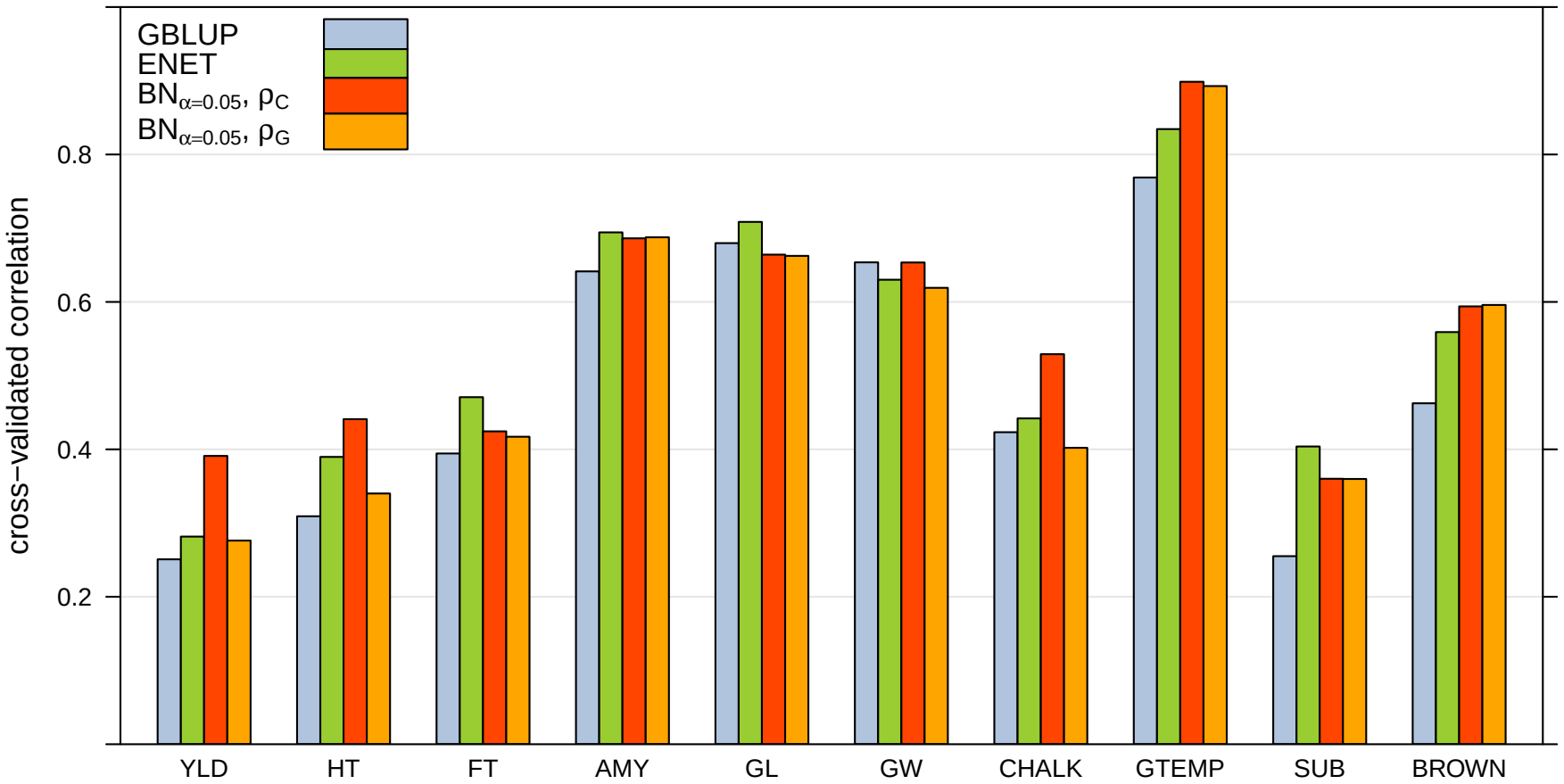


ρ_G = predictive correlation given all SNPs in the model.
ρ_C = predictive correlation given putative causal effects identified by the Bayesian network.
On average, GBLUP has $\rho = 0.18$, ENET has $\rho = 0.34$, and BN has $\rho_C = 0.37$, $\rho_G = 0.33$.

RICE: a Bayesian Network (64 nodes, 102 arcs)



RICE: Predictive Performance from Cross-Validation



On average, GBLUP has $\rho = 0.48$, ENET has $\rho = 0.54$, and BN has $\rho_C = 0.56$, $\rho_G = 0.53$.

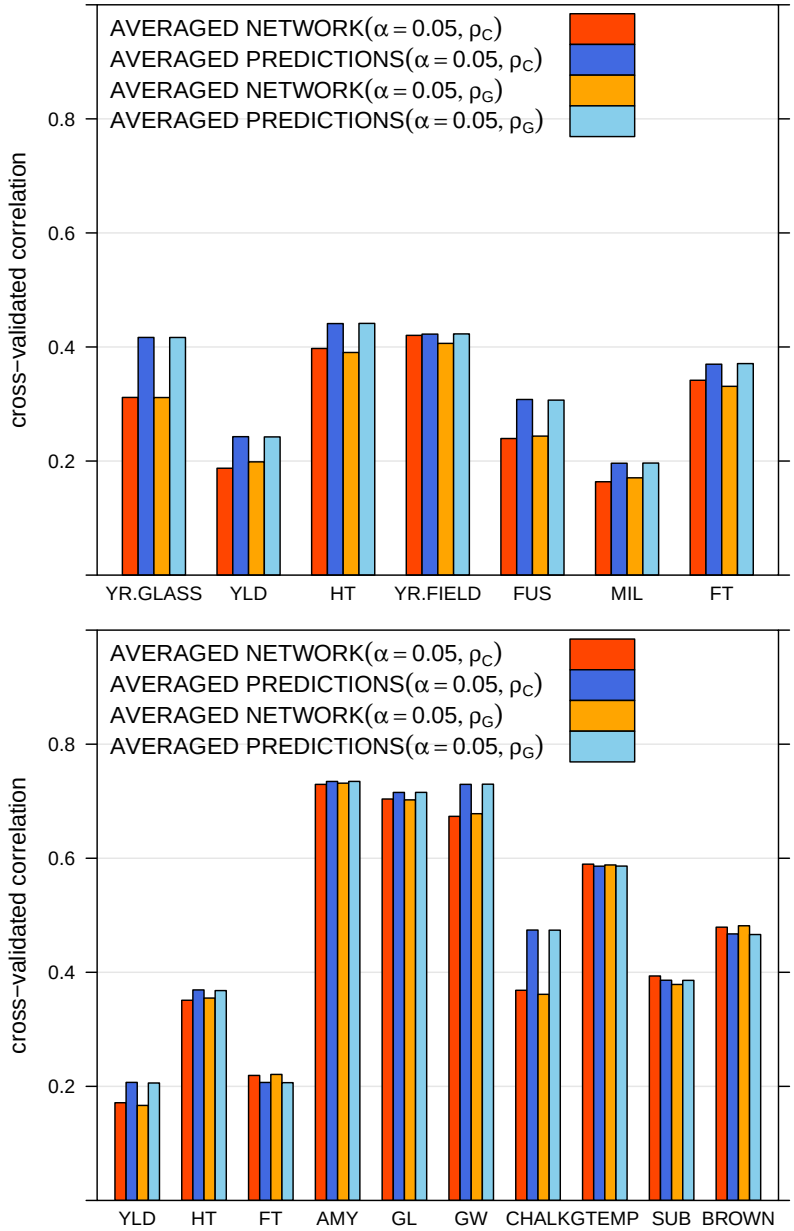
Predicting Traits for New Individuals

We can predict the traits:

1. from the averaged consensus network;
2. from each of the 10×10 networks we learn during cross-validation, and average the predictions for each new individual and trait.

Option 2. almost always provides better accuracy than option 1., especially for polygenic traits; 10×10 networks can cover the genome much better, and we have to learn them anyway.

So: averaged network for interpretation, individual networks for predictions.



A Few Notes on Learning Bayesian Networks

- Bayesian networks are largely self-correcting for multiplicity, leaving the type-I error α as the only tuning parameter.
- Bayesian networks try to determine the direction of relationships (unlike random effects models and undirected graphical models) and to address confounding.
- SNPs that are associated with more than one trait (pleiotropic effects) are included in the Bayesian network even when association with just a single trait is detected; at that point they can be linked to all the relevant traits. On the other hand, SNPs that are jointly associated but individually independent from a trait (epistatic effects) are not likely to be included.
- Performing feature selection impacts the ability of predicting traits influenced by many small genetic effects.

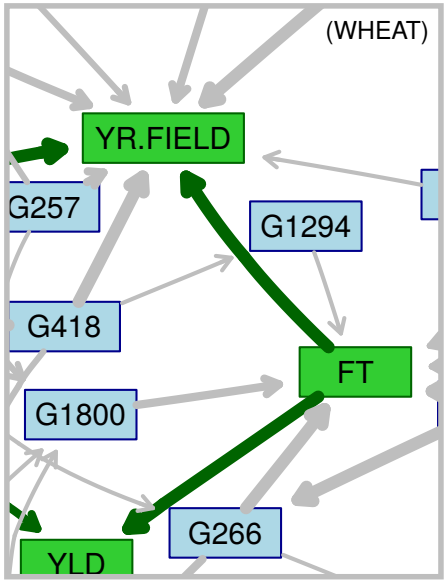
Inference and Interpretation

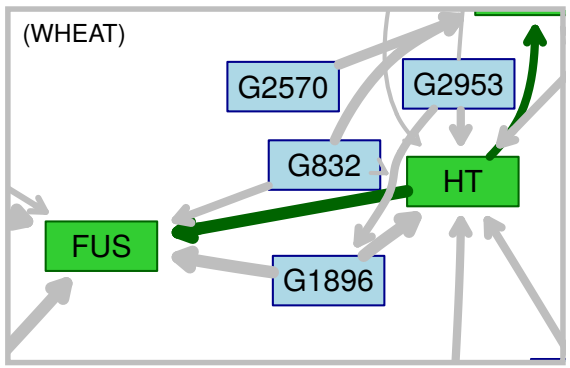
Causal Relationships Between Traits

One of the key properties of Bayesian networks is their ability of capturing the direction of the causal relationships among traits in the absence of latent confounders (the experimental design behind the data collection should take care of a number of them).

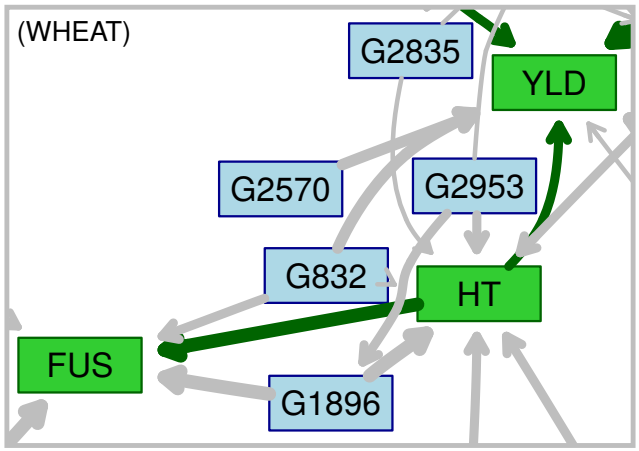
It works because each trait will have at least one incoming arc from the markers, say $G_l \rightarrow T_j$, and then $(G_l \rightarrow) T_j \leftarrow T_k$ and $(G_l \rightarrow) T_j \rightarrow T_k$ are not probabilistically equivalent [4]. So the network can

- suggest the direction of novel relationships;
- confirm the direction of known relationships, troubleshooting the experimental design and data collection.





Spotting Confounding Effects

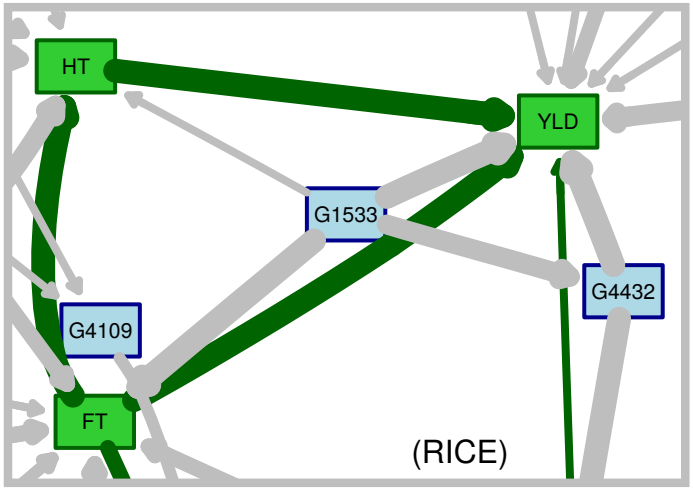


Traits can interact in complex ways that may not be obvious when they are studied individually, but that can be explained by considering neighboring variables in the network.
An example: in the WHEAT data, the difference in the mean YLD between the bottom and top quartiles of the FUS disease scores is +0.08.

So apparently FUS is associated with increased YLD! What we are actually measuring is the confounding effect of HT (FUS $\leftarrow$ HT $\rightarrow$ YLD); conditional on each quartile of HT, FUS has a negative effect on YLD ranging from −0.04 to −0.06. This is reassuring since it is known [11] that susceptibility to fusarium is positively related to HT, which in turn affects YLD [2].

Disentangling Pleiotropic Effects

When a marker is shown to be associated to multiple traits in a GWAS, we should separate its direct and indirect effects on each of the traits. (Especially if the traits themselves are linked!)
Take for example G1533 in the RICE data set: it is putative causal for YLD, HT and FT.



- The difference in mean between the two homozygotes is +4.5cm in HT, +2.28 weeks in FT and +0.28 t/ha in YLD.
- Controlling for YLD and FT, the difference for HT halves (+2.1cm);
- Controlling for YLD and HT, the difference for FT is about the same (+2.3 weeks);
- Controlling for HT and FT the difference for YLD halves (+0.16 t/ha).

So, the model suggests the marker is causal for FT and that the effect on the other traits is partly indirect. This agrees from the p-values from an independent GWAS study (FT: 5.87e-28 $<$ YLD: 4.18e-10, HT:1e-11).

Identifying Causal (Sets of) Markers

Compared to additive regression models, Bayesian networks make it trivial to compute:

- posterior probability of association for a marker and a trait after all the other markers and traits are taken into account to rule out linkage disequilibrium, confounding, pleiotropy, etc.;
- expected allele counts n_{LOW} and n_{HIGH} for a marker and low/high values of a set of traits ($n_{\text{LOW}} - n_{\text{HIGH}}$ should be large if the marker tags a causal variant and thus should show which allele is favorable).

	G1778	G3872	G4529	G1440	G5794
SMALL GRAINS	0.00	0.78	0.29	0.16	0.74
LARGE GRAINS	2.00	0.47	0.63	0.35	0.12
	G4668	G2764	G3927	G3992	G4432
SMALL GRAINS	0.24	0.29	0.18	0.09	0.00
LARGE GRAINS	0.17	0.00	0.62	0.29	0.82

SMALL GRAINS = bottom 10% GL, bottom 10% GW in the RICE data.
LARGE GRAINS = top 10% GL, top 10% GW.

Conclusions

Conclusions

- Bayesian networks provide a model for qualitative and quantitative reasoning that is easily interpretable thanks to its graphical nature.
- They also provide competitive predictive accuracy for multi-trait modeling, and can easily accommodate additional variables (experimental factors, different omics data, etc.).
- Causal inference has been thoroughly explored in the literature and can be used to address confounding and disambiguate the effects of pleiotropic markers.
- Posterior inference is easily performed by simulation for both traits and markers.

Acknowledgements

Ian Mackay	data preparation and general support
Phil Howell	has run the MAGIC programme and collected disease scores and yield data
Nick Gosman	involved in the running of the MAGIC programme
Rhian Howells	collected the flowering time data
Richard Hornsell	performed crossing to create the MAGIC population and preparation of DNA
Pauline Bancept	collected the glasshouse yellow rust data
Alison Bentley	performed crossing and subsequent analyses

IRRI

Chitra Raghavan	data co-ordination and mapping
RK Singh	population development and breeding applications
Mona Jubay	field operations, data collection and management
Hei Leung	principal investigator

Thank You!

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.

J. E. Flintham, A. Börner, A. J. Worland, and M. D. Gale.

Optimizing wheat grain yield: effects of Rht (Gibberellin-Insensitive) dwarfing genes.

J. Agr. Sci., 128(1):11–25, 1997.

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.

D. Koller and N. Friedman.

Probabilistic Graphical Models: Principles and Techniques.

MIT Press, 2009.

I. Mackay, P. Bansept, T. Barber, A. Bentley, J. Cockram, J. Elderfield, N. Gosman, A. Greenland, R. Horsnell, R. Howells, G. Rose, D. O'Sullivan, and P. Howell.

An Eight-Parent Multiparent Advanced Generation Inter-Cross Population for Winter-Sown Wheat: Creation, Properties, and Validation.

Genetics, 4(9):1603–1610, 2014.

J. Pearl.

Causality: Models, Reasoning and Inference.

Cambridge University Press, 2nd edition, 2009.

References II

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

Multiple Quantitative Trait Analysis Using Bayesian Networks.

Genetics, 198(1):129–137, 2014.

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

Improving the Efficiency of Genomic Selection.

Statistical Applications in Genetics and Molecular Biology, 12(4):517–527, 2013.

M. Scutari and R. Nagarajan.

On Identifying Significant Edges in Graphical Models of Molecular Networks.

Artificial Intelligence in Medicine, 57(3):207–217, 2013.

Christine Sinoquet and Raphaël Mourad.

Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics.

Oxford University Press, 2013.

Srinivasachary, N. Gosman, A. Steed, T. W. Hollins, R. Bayles, P. Jennings, and P. Nicholson.

Semi-dwarfing Rht-B1 and Rht-D1 loci of wheat differ significantly in their influence or resistance to fusarium head blight.

Theor. Appl. Genet., 118:695–702, 2009.

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.