Graphical Models for Genomic Selection

Marco Scutari[1], Phil Howell[2]

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

[2] phil.howell@niab.com
NIAB

June 12, 2013

Background

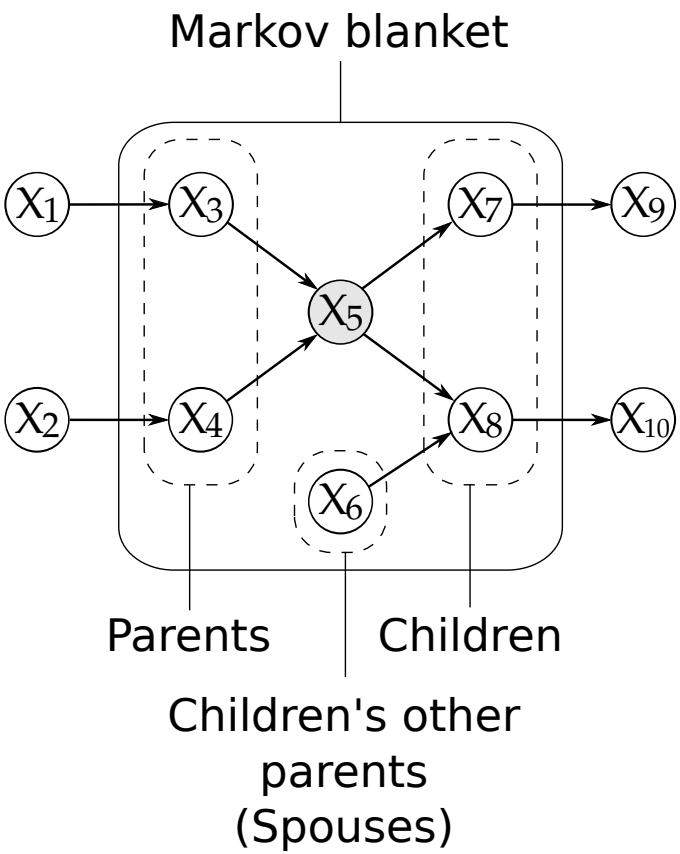
Bayesian networks: an overview

A Bayesian network (BN) [6, 7] is a combination of:

- directed graph $\mathcal{G} = (\mathbf{V}, E)$, 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, $\mathbf{X} = \{X_i\}$, which can be split into simpler local probability distributions according to the arcs $a_{ij} \in E$ 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 $\mathcal{G}$.

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 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 [11]. 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

Learning the Bayesian network

1. Feature Selection.
   1.1 For each trait, use the SI-HITON-PC algorithm [1, 10] 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 [5] 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 [9]. This follows the spirit of other hybrid approaches [3, 12], that have shown to be well-performing in literature.
3. Parameter Learning. Learn the parameters of the BN as a Gaussian BN [6]: each local distribution in a linear regression and the global distribution is a hierarchical linear model.

The Parameters of the Bayesian Network

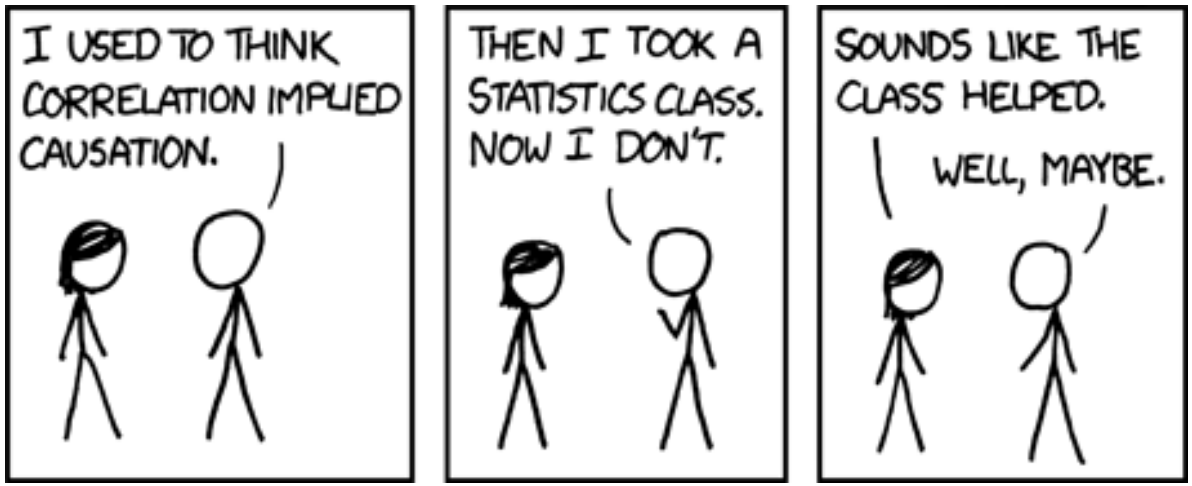
The local distribution of each trait X_i is a linear model

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

which can be estimated any frequentist or Bayesian approach in which the nodes in Π_{X_i} are treated as fixed effects (e.g. ridge regression [4], elastic net [13], 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.

A caveat about causal interpretations



http://xkcd.com/552/

Even though "good" BNs have a structure that mirrors cause-effect relationships [8], and even though there is ample literature on how to learn causal BNs from observational data, inferring causal effects from a BN requires great care even with completely independent data (i.e. with no family structure).

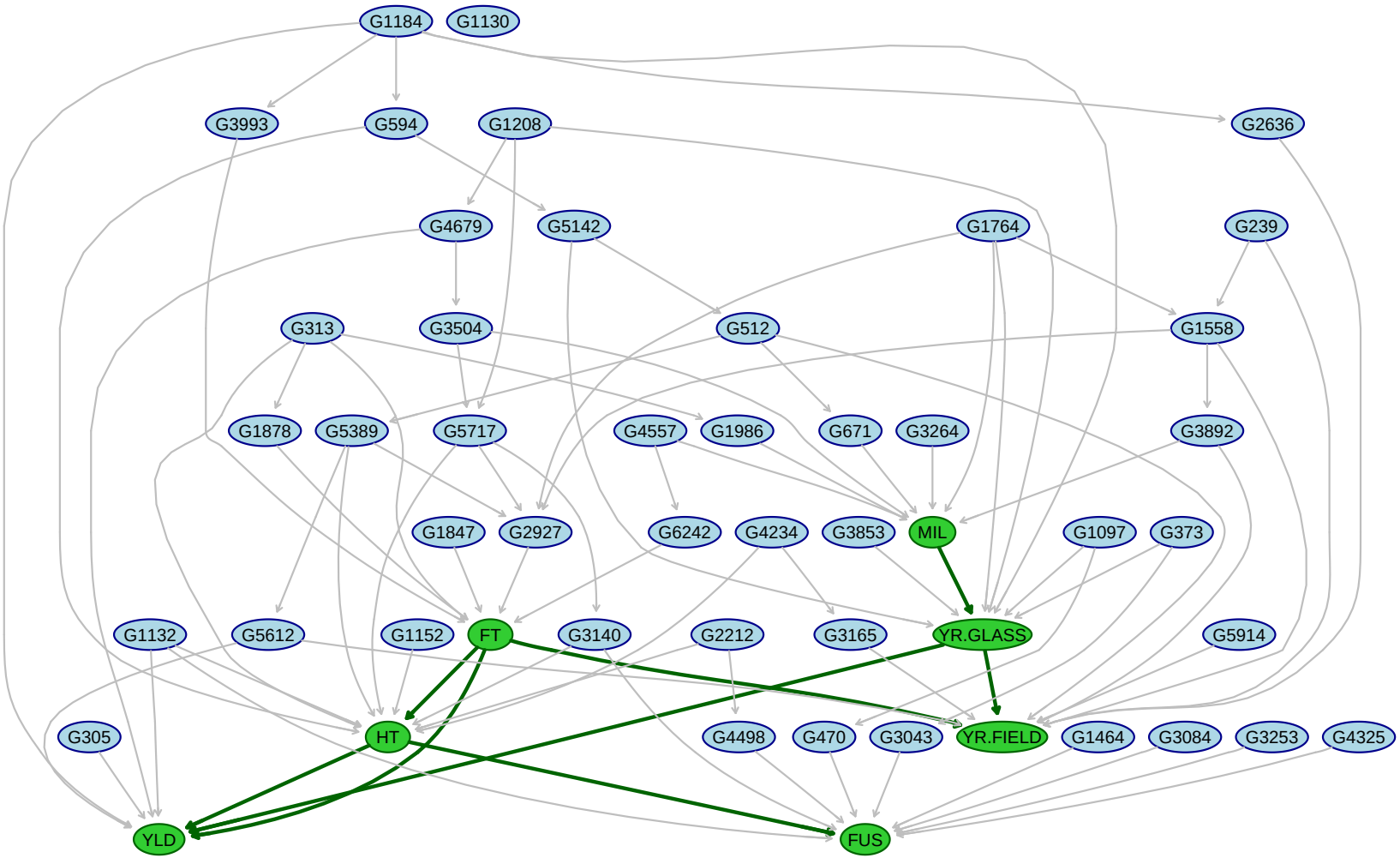
The MAGIC data

The MAGIC data (the same as in Ian's talk) include 721 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).

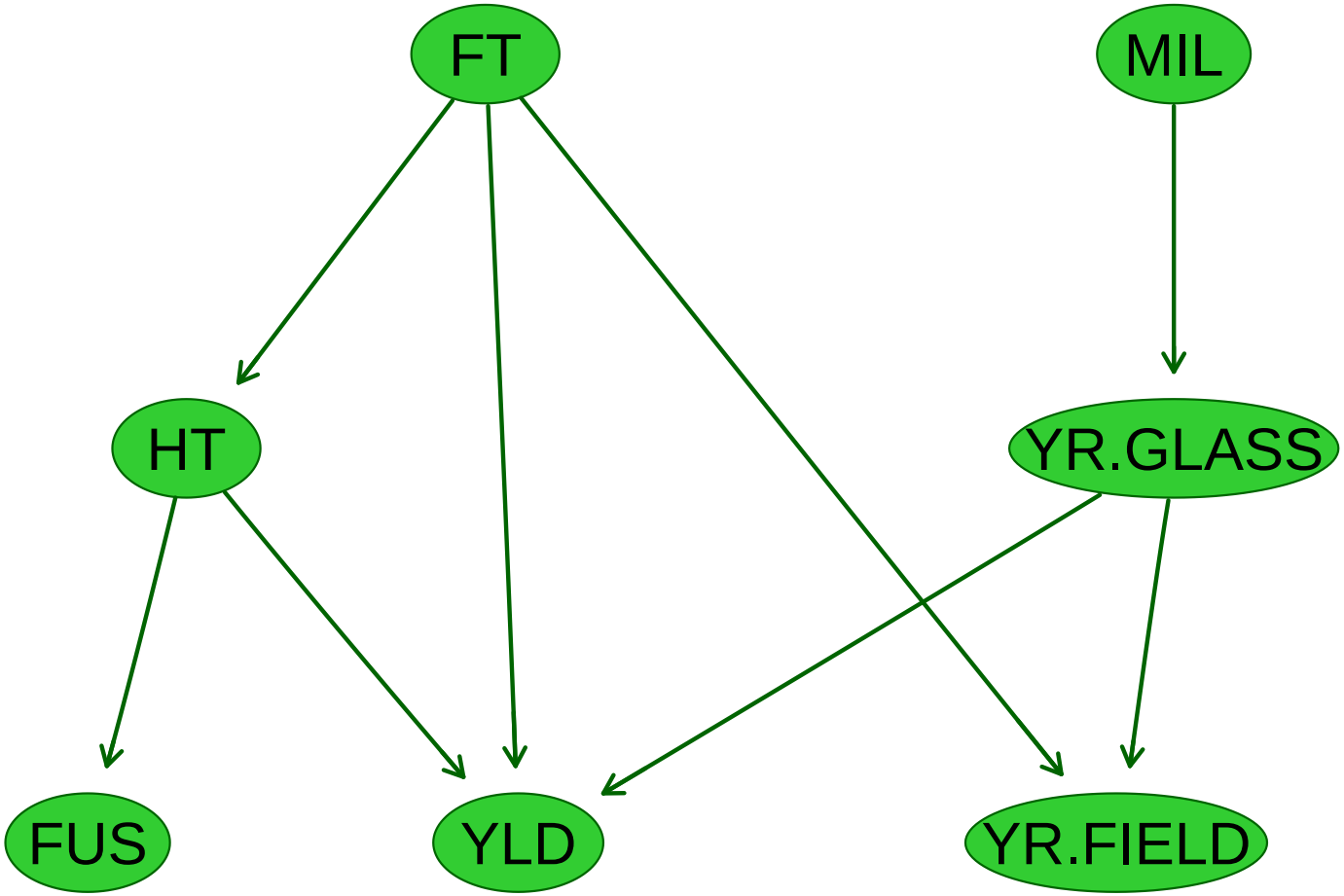
Varieties with missing phenotypes or family information and markers with $> 20\%$ missing data were dropped. The phenotypes were adjusted for family structure via BLUP and the markers screened for MAF > 0.01 and COR < 0.99.

Bayesian network learned from MAGIC



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

Phenotypic traits in MAGIC



Assessing arc strength with boostrap resampling

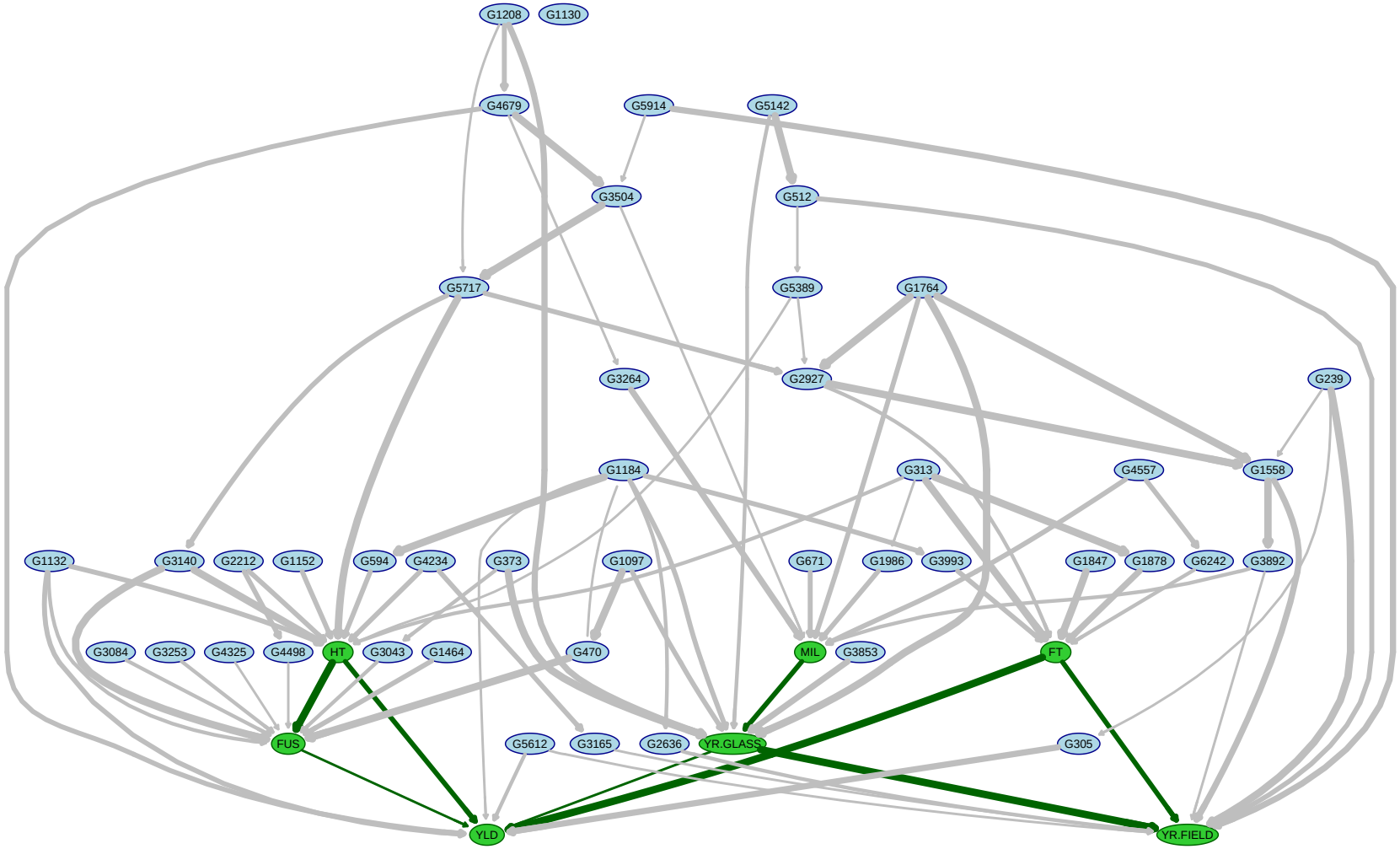
Friedman et al. [2] proposed an approach to assess the strength of each arc based on bootstrap resampling and model averaging:

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 edge e_i is present in the true network structure $\mathcal{G}_0 = (\mathbf{V}, E_0)$ as

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

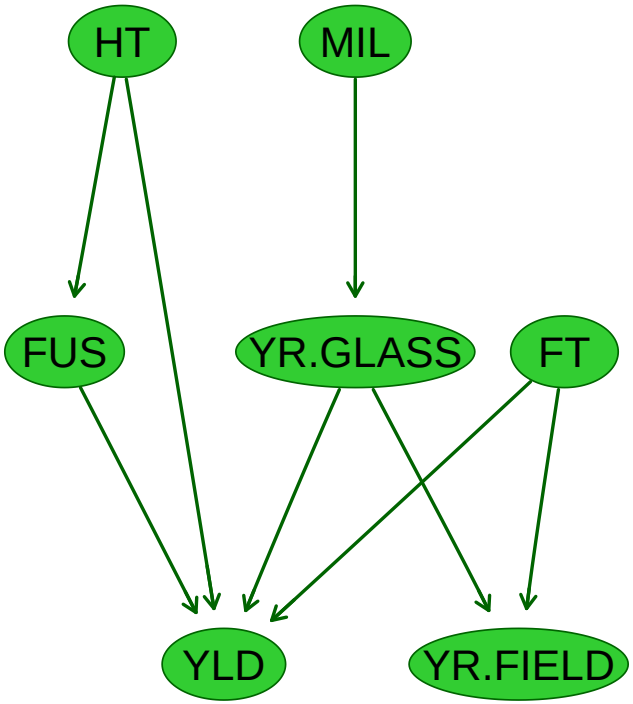
where $\mathbb{1}_{\{e_i \in E_b\}}$ is equal to 1 if $e_i \in E_b$ and 0 otherwise.

Averaged Bayesian network from MAGIC



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

Phenotypic traits in MAGIC



from	to	strength	direction
YR.GLASS	YLD	0.636	1.000
YR.GLASS	HT	0.074	0.648
YR.GLASS	YR.FIELD	1.000	0.724
YR.GLASS	FT	0.020	0.800
HT	YLD	0.722	1.000
HT	YR.FIELD	0.342	0.742
HT	FUS	0.980	0.885
HT	MIL	0.012	0.666
YR.FIELD	YLD	0.050	1.000
YR.FIELD	FUS	0.238	0.764
YR.FIELD	MIL	0.402	0.661
FUS	YR.GLASS	0.030	0.666
FUS	YLD	0.546	1.000
FUS	MIL	0.058	0.758
MIL	YR.GLASS	0.824	0.567
MIL	YLD	0.176	1.000
FT	YLD	1.000	1.000
FT	HT	0.420	0.809
FT	YR.FIELD	0.932	0.841
FT	FUS	0.436	0.692
FT	MIL	0.080	0.825

Arcs in the BN are highlighted in red in the table.

Inference

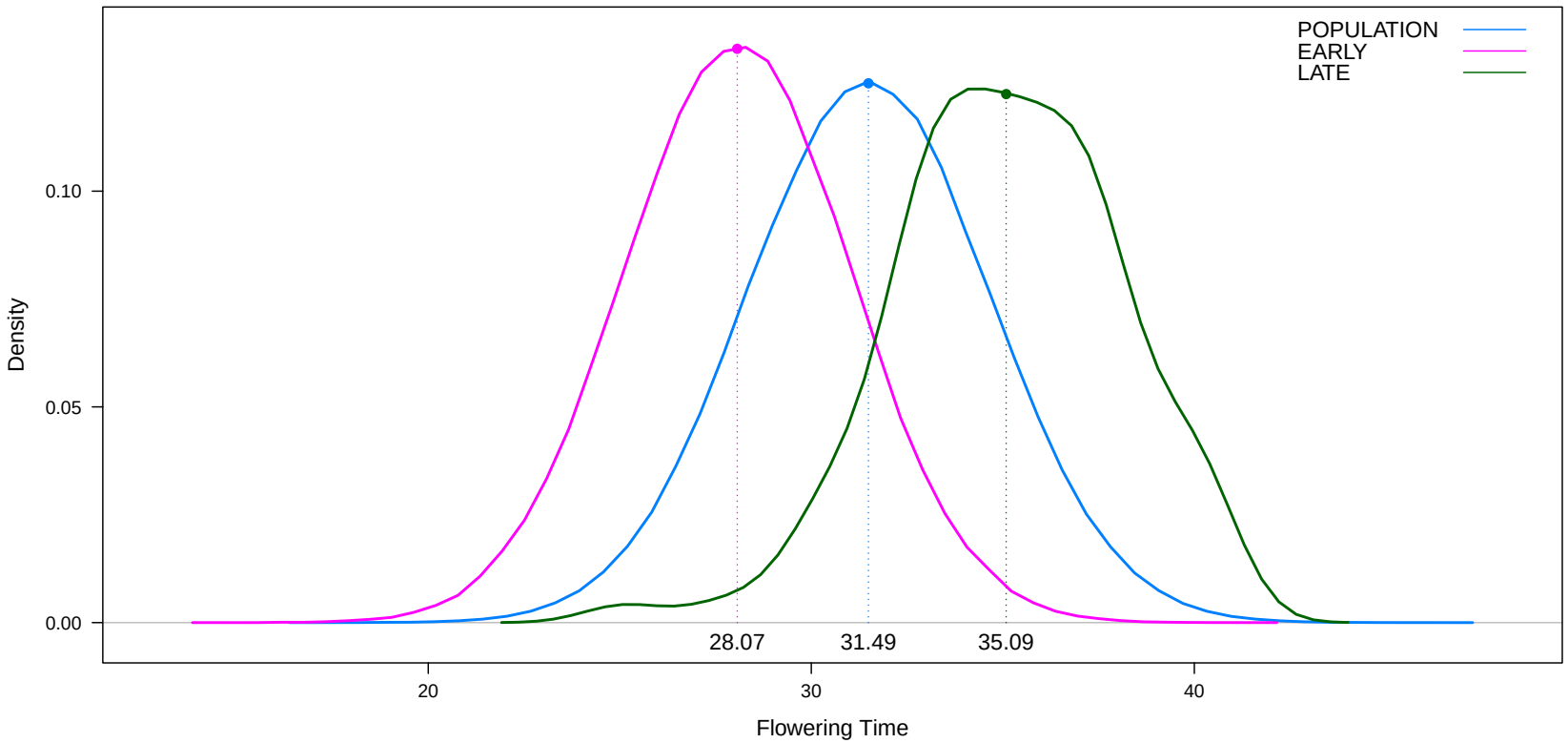
Inference in Bayesian networks

Inference for BNs usually takes two forms:

- conditional probability queries, in which the distribution of one or more nodes of interest is investigated conditional on a second set of nodes (which are either completely or partially fixed);
- maximum a posteriori queries, in which the most likely outcome of a certain event (involving one or more nodes) conditional on evidence on a set of nodes (which are often completely fixed for computational reasons).

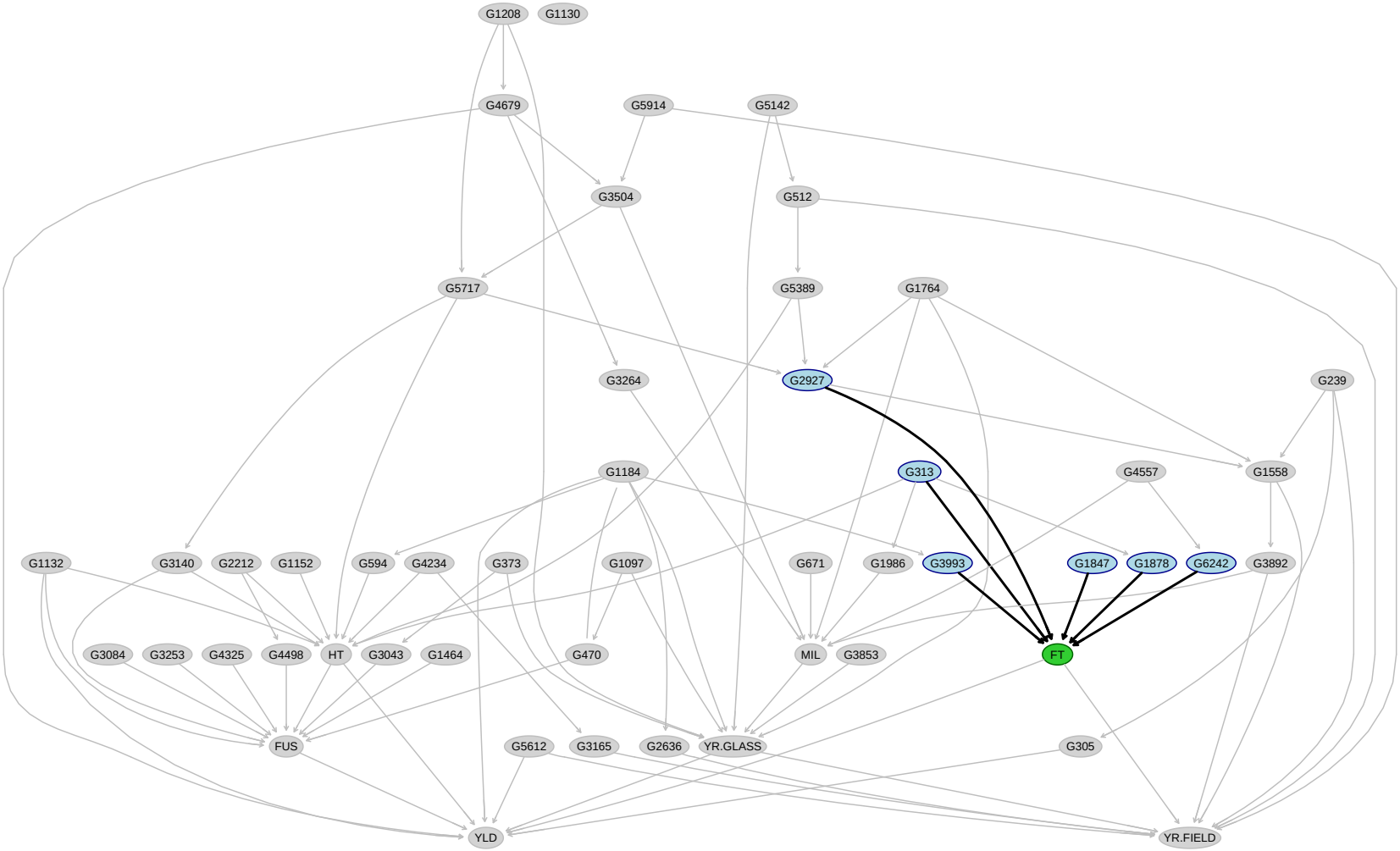
In practice this amounts to answering "what if?" questions (hence the name queries) about what could happen in observed or unobserved scenarios using posterior probabilities or density functions.

Flowering time: what if we fix directly related alleles?

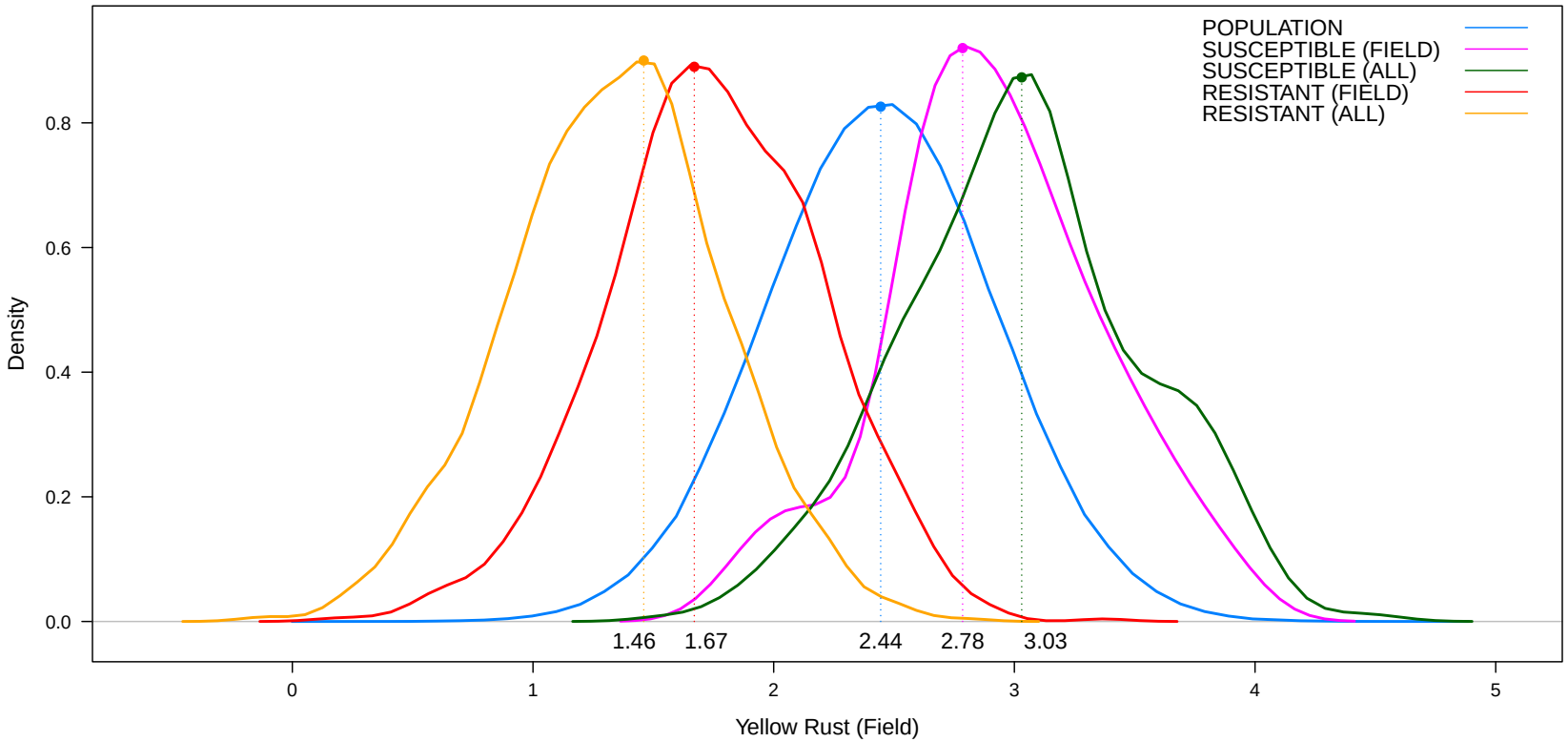


Fixing 6 genes that are parents of FT in the BN not to be homozygotes for late flowering (EARLY) or for early flowering (LATE). Heterozygotes are allowed in both cases.

Flowering time: which nodes we used...

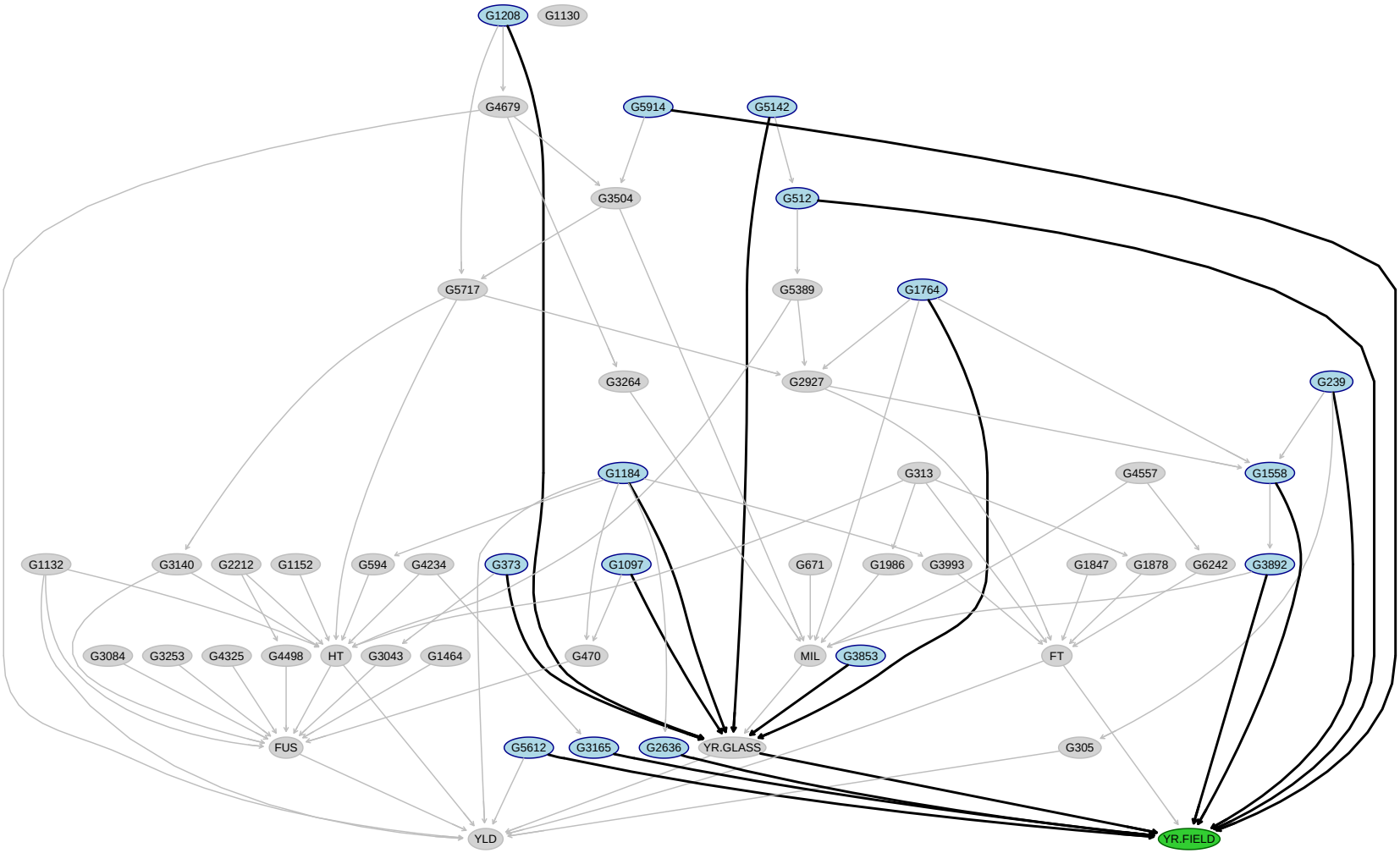


Yellow rust: what if we fix (in)directly related alleles?

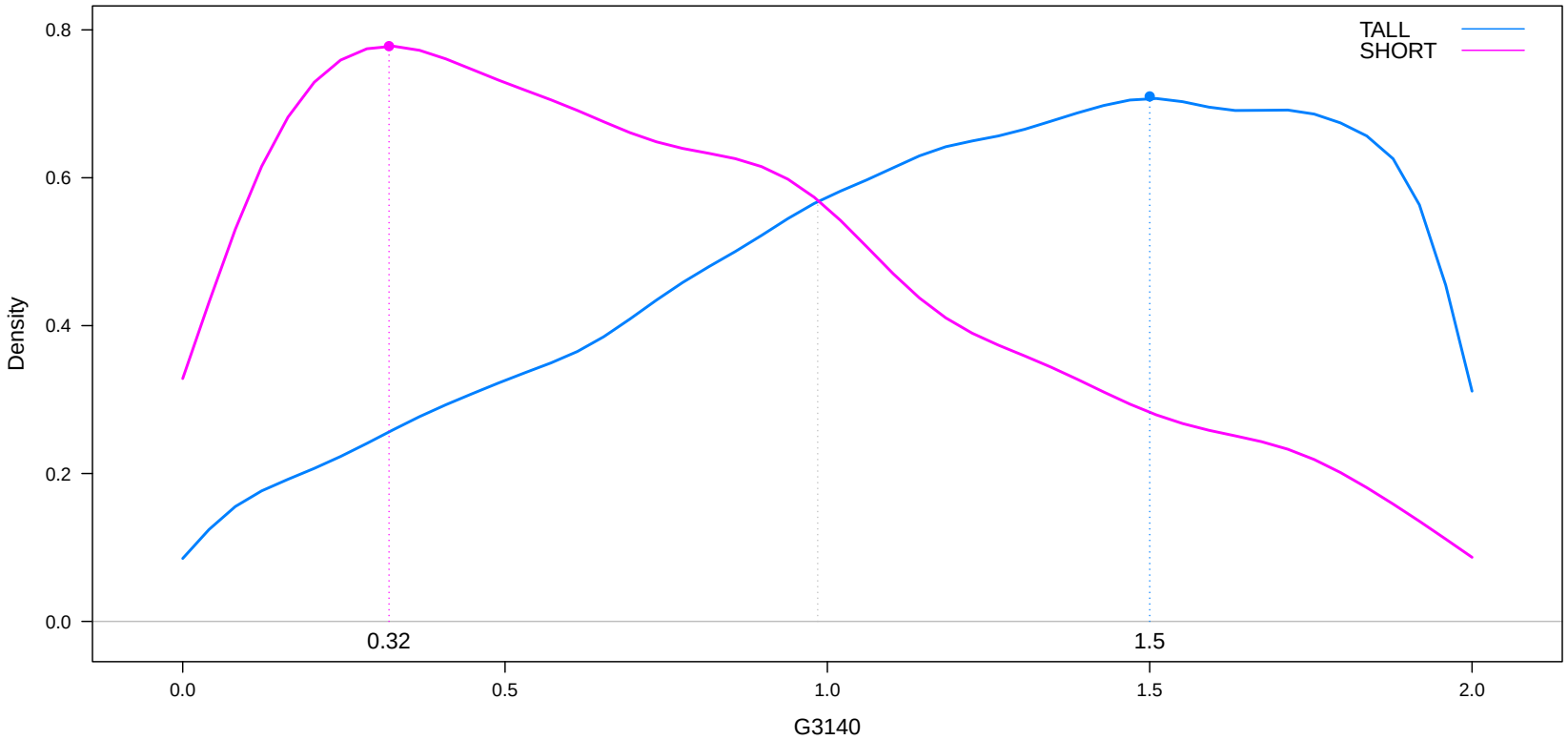


Fixing 8 genes that are parents of YR.FIELD, then another 7 that are parents of YR.GLASS, either not to be homozygotes for yellow rust susceptibility or for yellow rust resistance. Heterozygotes are allowed in both cases.

Yellow rust: nodes farther away can help...

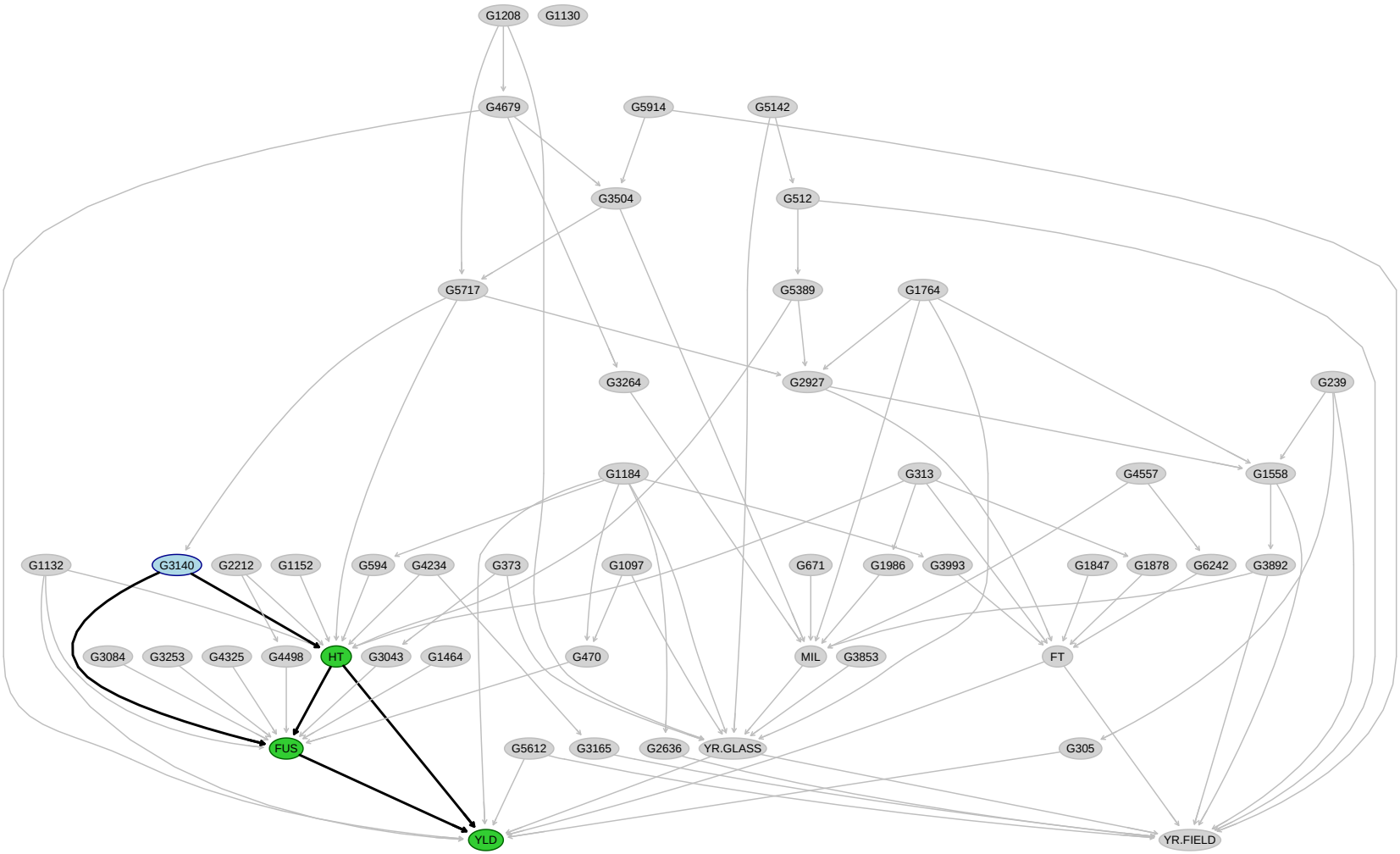


G3140: can we guess the allele?



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 markers, both within and between 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.

Thanks!

Acknowledgements

NIAB

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 programmes
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

UCL

David Balding	my Supervisor

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.

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.

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. E. Hoerl and R. W. Kennard.

Ridge Regression: Biased Estimation for Nonorthogonal Problems.

Technometrics, 12(1):55–67, 1970.

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. Koller and N. Friedman.

Probabilistic Graphical Models: Principles and Techniques.

MIT Press, 2009.

References II

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.

G. E. Schwarz.

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

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 (submitted).
Statistical Applications in Genetics and Molecular Biology, 2013.

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.

H. Zou and T. Hastie.

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