Genotype-Environment Effects Analysis Using Bayesian Networks

Marco Scutari[1], Alison Bentley[2] and Ian Mackay[2]

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

[2] National Institute for Agricultural Botany (NIAB)
Cambridge, UK

December 7, 2014

Integrative Analyses in Statistical Genetics

Bayesian networks (BNs) represent a flexible tool for quantitative [6], qualitative and causal [9] reasoning, and are one of the building blocks used to specify complex models and Monte Carlo inference techniques in machine learning [8].

As such, they are well suited to integrative analyses in genetics and systems biology, that is, jointly modelling data from different sources:

- various forms of sequence data (e.g. SNPs, full sequence data);
- various qualitative and quantitative traits (e.g. disease scores, morphological characteristics);
- epigenetic data (e.g. methylation);
- products of gene transcriptions (e.g. RNA, proteins).

Depending on the data at hand, such analyses are called GWAS, GS, eQTL, GxE GWAS, mQTL, etc. and make up the vast majority of literature in the field.

Integrating Two Types of Data: GWAS and GS

The baseline model for genome-wide association studies (GWAS) and genomic selection (GS) is the linear mixed model [3], rebranded as GBLUP (Genetic BLUP, [7]). It is typically fitted on a single trait X_t at a time using a large number S of SNPs $\mathbf{X_S}$ in the form of 0/1/2 allele counts from a genome-wide profile:

$$X_t = \boldsymbol{\mu} + Z_{\mathbf{S}}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \mathbf{u} \sim N(\mathbf{0}, \mathbf{K}\sigma^2_{\mathbf{u}})$$

where $\boldsymbol{\mu}$ is the population mean, $Z_{\mathbf{S}}$ is the design matrix for the markers, $\mathbf{u}$ are random effects, $\boldsymbol{\varepsilon}$ is the error term and $\mathbf{K}$ is the kinship matrix encoding the relatedness between the individuals. When $\mathbf{K}$ can be expressed in the form $\mathbf{X_S}\mathbf{X_S}^{\mathrm{T}}$, GBLUP can be shown to be equivalent to the Bayesian linear regression

$$X_t = \boldsymbol{\mu} + \sum_{i=1}^{S} X^*_{s_i}\beta_i + \boldsymbol{\varepsilon} \qquad \text{with SNP effect prior} \qquad \boldsymbol{\beta} \sim N\left(\mathbf{0}, \frac{\sigma^2_g}{S}\mathbf{I}\right),$$

for some transformation of the X_{s_i} [10, 11].

Gaussian Bayesian Networks (GBNs)

GBNs use a DAG G to represent the dependence structure of the multivariate distribution of $\mathbf{X} = \{X_1, \ldots X_p\}$ under the following assumptions [6]:

1. $\mathbf{X}$ has a multivariate normal distribution; and
2. dependencies between the X_is are linear.

Under these assumptions $\mathrm{COV}(\mathbf{X}) = \Sigma$ is a sufficient statistic for the GBN and:

1. if X_i and X_j are graphically separated in G (d-separation, [6]), then $\Omega_{ij} = (\Sigma^{-1})_{ij} = 0$; and
2. the local distribution associated with each X_i is a linear regression on the parents Π_{X_i} of X_i, i.e.:

$$X_i = \mu_{X_i} + X_j\beta_j + \ldots + X_k\beta_k + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2).$$

Note that $\beta_j = -\Omega_{ij}/\Omega_{ii}$ in the above [2].

Assumptions for Genetic Data

In the spirit of commonly used additive genetic models [5, 7], we make some further assumptions on the GBN to obtain a sensible causal model:

1. traits can depend on SNPs (i.e. $X_{s_i} \to X_{t_j}$) but not vice versa (i.e. not $X_{t_j} \to X_{s_i}$), and they can depend on other traits (i.e. $X_{t_i} \to X_{t_j}, i \neq j$);
2. SNPs can depend on other SNPs (i.e. $X_{s_i} \to X_{s_j}, i \neq j$); and
3. dependencies between traits follow the temporal order in which they are measured.

Under these assumptions, the local distribution of each trait is

$$\begin{aligned} X_{t_i} &= \boldsymbol{\mu}_{t_i} + \Pi_{X_{t_i}} \boldsymbol{\beta}_{t_i} + \boldsymbol{\varepsilon}_{t_i} \\ &= \boldsymbol{\mu}_{t_i} + \underbrace{X_{t_j}\beta_{t_j} + \ldots + X_{t_k}\beta_{t_k}}_{\text{traits}} + \underbrace{X_{s_l}\beta_{s_l} + \ldots + X_{s_m}\beta_{s_m}}_{\text{SNPs}} + \boldsymbol{\varepsilon}_{t_i}, \quad \boldsymbol{\varepsilon}_{t_i} \sim N(0, \sigma^2_{t_i}\mathbf{I}) \end{aligned}$$

and the local distribution of each SNP is

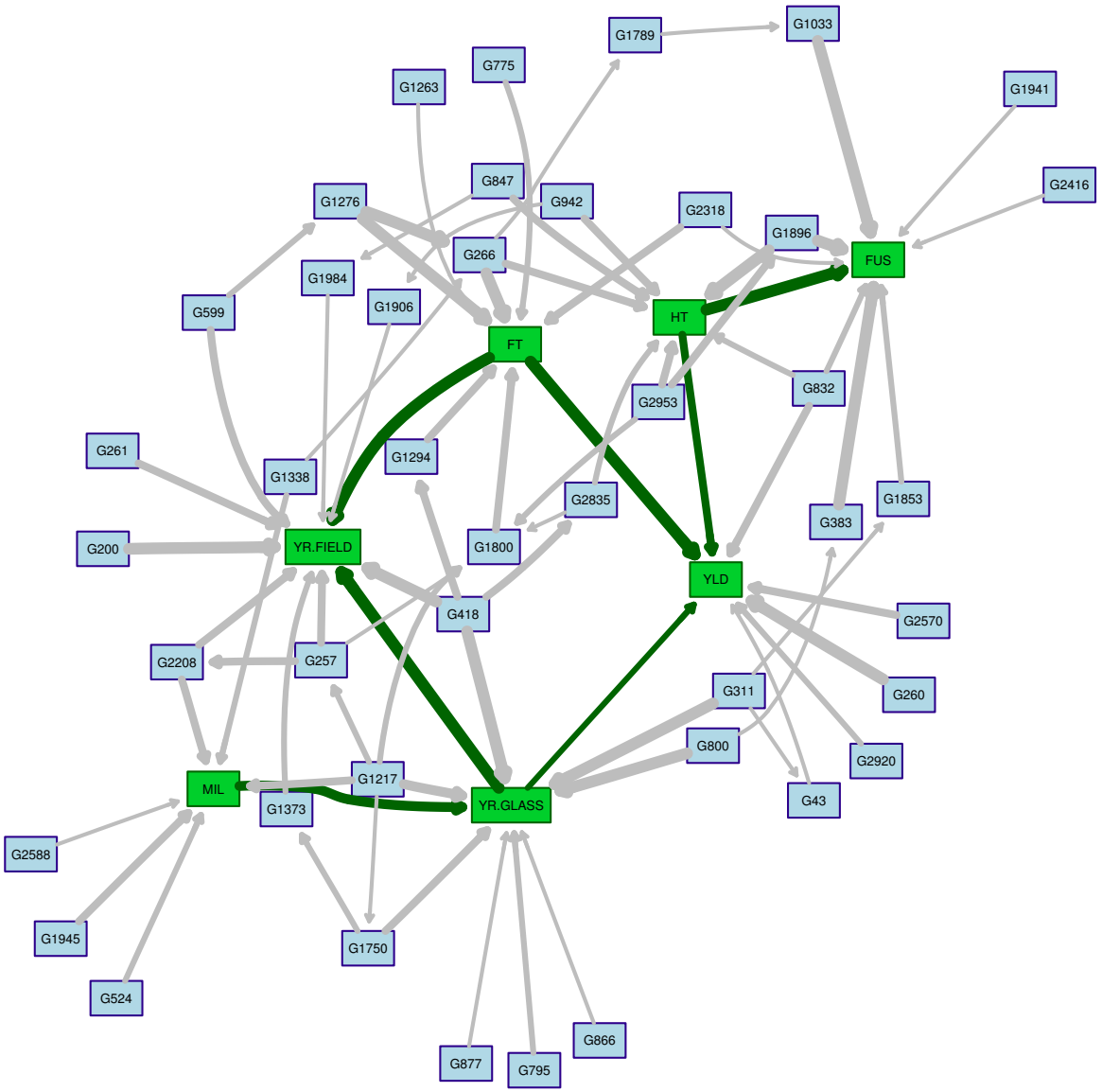
$$X_{s_i} = \boldsymbol{\mu}_{s_i} + \underbrace{X_{s_l}\beta_{s_l} + \ldots + X_{s_m}\beta_{s_m}}_{\text{SNPs}} + \boldsymbol{\varepsilon}_{s_i}, \qquad \boldsymbol{\varepsilon}_{s_i} \sim N(0, \sigma^2_{s_i}\mathbf{I}).$$

Learning GBNs from Genetic Data

We used the R packages bnlearn [12] and penalized [4] to implement the following hybrid approach to GBN learning [13].

1. Structure Learning.
   - 1.1 For each trait X_{t_i}, use the SI-HITON-PC algorithm [1] and the t-test for correlation to learn its parents and children; this is sufficient to identify the Markov blanket $\mathcal{B}(X_{t_i})$ because of the assumptions on the GBN. The choice of SI-HITON-PC is motivated by its similarity to single-SNP analysis.
   - 1.2 Drop all the markers which are not in any $\mathcal{B}(X_{t_i})$.
   - 1.3 Learn the structure of the GBN from the nodes selected in the previous step, setting the directions of the arcs as discussed above. We identify the optimal structure as that which maximises BIC.
2. Parameter Learning. Learn the parameters of the local distributions using ordinary least squares or ridge regression.

A GWAS Model from a Wheat Mapping Population



50 nodes (7 traits, 43 SNPs) from 600 obs. and 3.2K SNPs.

78 arcs, interpreted as putative causal effects.

Thickness represents arc strength, computed as the frequency of each arc in the 100 GBNs used in model averaging.

Scutari M, Howell P, Balding DJ, Mackay I (2014).
Multiple Quantitative Trait Analysis Using Bayesian Networks.
Genetics, 198(1), 129–137.

Adding Environmental Effects: GxE Interactions

The BN model in the previous slide has quite a few limitations, especially when interpreted as a causal model:

- It only uses SNPs to explain traits; there are multiple levels of unobserved biological processes in the middle acting as confounders.
- It assumes all observations are collected under the same conditions (environmental and/or exogenous), which is rarely the case for large experiments, and are homogeneous overall (e.g. no stratification or individuals from different ethnicities/subspecies).
- It assumes all variables are continuous, so that they can be meaningfully modelled with linear regression on their natural scale.

A step forward in addressing these concerns is moving from GBNs to conditional Linear Gaussian Bayesian networks (CLGBNs) to include environmental effects as discrete variables and model genotype-by-environment interactions (GxE) and those with the traits.

Conditional Linear Gaussian Bayesian networks (CLGBNs)

CLGBNs extend traditional GBNs using mixture of Gaussians under the following assumptions [6, 8]:

1. discrete variables can only have discrete parents;
2. the local distribution for a discrete variable is a conditional probability table (CPT); and
3. the local distribution for a continuous variable is a set of linear regressions, one for each configuration δ of the discrete parents Δ_{X_i} (if any), with the continuous parents Γ_{X_i} as explanatory variables:

$$X_{i\delta} = \mu_{i\delta} + X_{j\delta}\beta_{j\delta} + \ldots + X_{k\delta}\beta_{k\delta} + \varepsilon_{i\delta}, \quad \varepsilon_{i\delta} \sim N(0, \sigma^2_{i\delta}).$$

Note that, unlike most literature on mixture models, the δ does not arise from a latent variable but from an observed one.

Learning CLGBNs from Genetic Data

In addition to the assumptions used to learn GBNs, now we also assume that:

- traits and genes can depend environmental effects and experimental variables but not vice versa.

And the hybrid learning approach from [13] is modified as follows.

1. Structure Learning.
   - 1.1 For each trait X_{t_i}, use the SI-HITON-PC algorithm [1] and the t-test for correlation to learn its parents and children among the genes; then do a second pass also considering the environmental effects using SI-HITON-PC and a log-likelihood ratio test.
   - 1.2 Drop all the markers which are not in any $\mathcal{B}(X_{t_i})$.
   - 1.3 Learn the structure of the CLGBN from the nodes selected in the previous step, setting the directions of the arcs as discussed above. We identify the optimal structure as that which maximises BIC.
2. Parameter Learning. Learn the parameters of the local distributions using empirical frequencies for the discrete variables and ordinary least squares or ridge regression for the continuous variables.

Another Wheat Data Set, From Multiple Countries

We prototyped this approach on the wheat population described in:

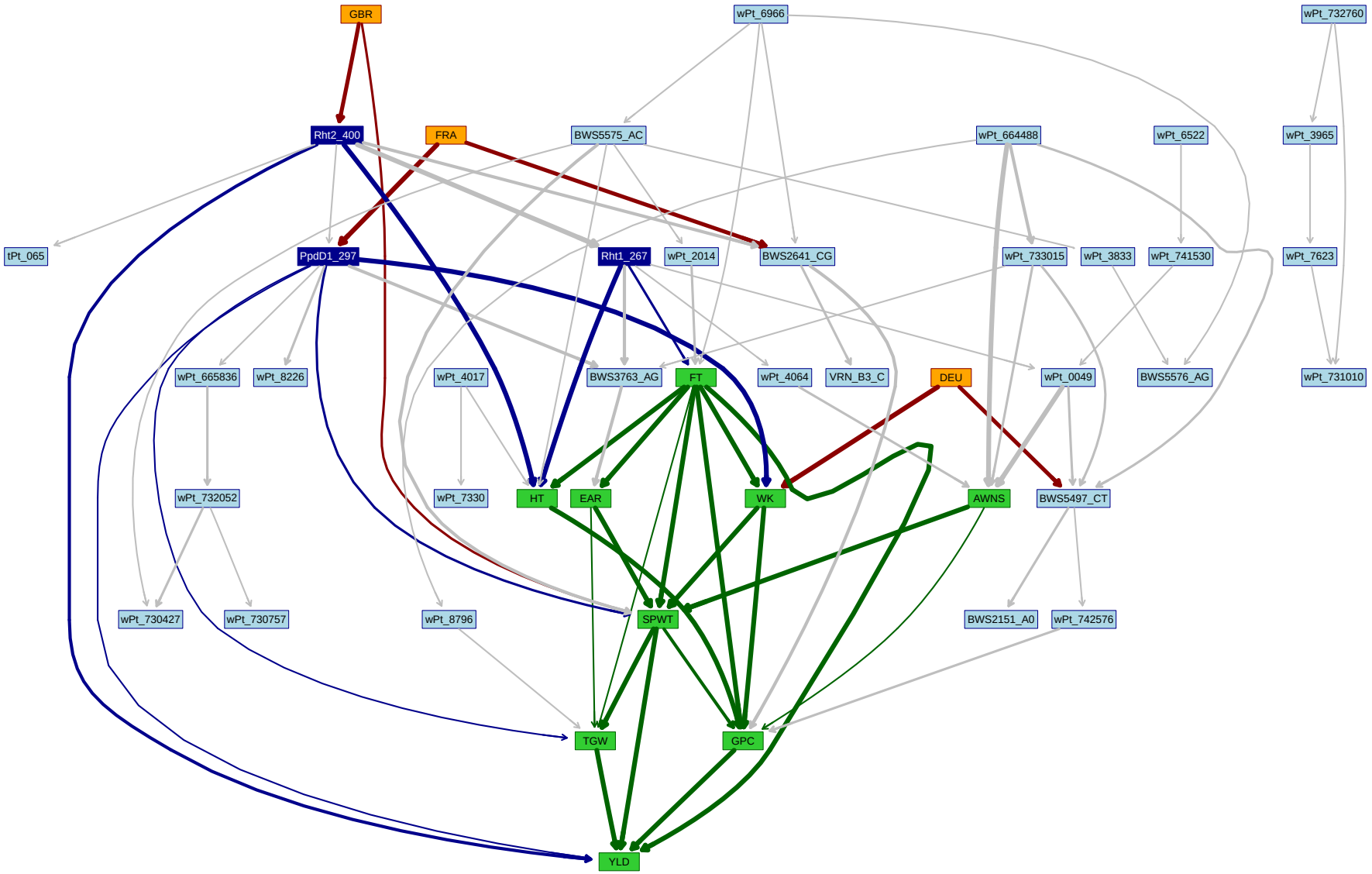
Bentley AR, Scutari M, Gosman N *et al.* (2014). Applying Association Mapping and Genomic Selection to the Dissection of Key Traits in Elite European Wheat.

Theoretical and Applied Genetics, 127(12), 2619–2633.

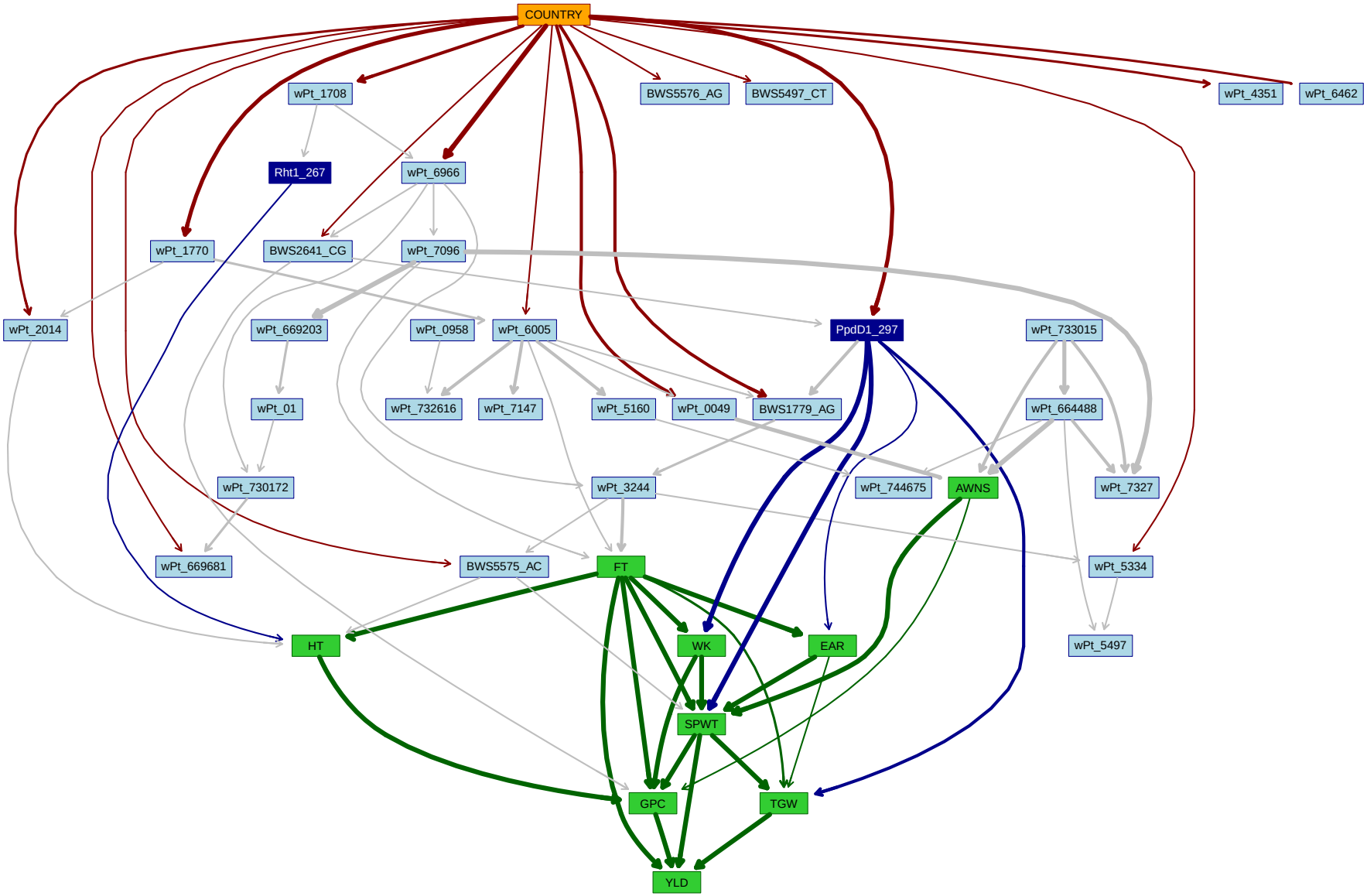
This data set contains 376 wheat varieties from different countries (210 FRA, 90 DEU, 75 GBR) trialled in the same set of fields in GBR, DEU and FRA to produce a variety of gene-environment interactions. After preprocessing marker profiles include 2.1K DaRTs and SNPs and 3 known genes: *PpdD1_297* (flowering time) and *Rht1_267*/*Rht2_400* (dwarfing genes). Traits include:

- Yield (YLD, t/ha)
- Flowering time (FT, days)
- Height (HT, cm)
- Winter Kill (WK, 1–9)
- Grain Protein Content (GPC, %)
- Thousand Grain Weight (TGW, weight/hl)
- Specific Weight (SPWT, weight/hl)
- Earing (EAR, ears/m^2)
- Awns (AWNS, 0–1)

GBNs, Adding Countries as Standalone Dummy Variables



CLGBNs, Adding Countries as a Single Discrete Variable



Predictive Performance

GBN (69 nodes, 117 arcs, $p = 186$) vs CLGBN (227 nodes, 421 arcs, $p = 941$)					
	YLD	FT	HT	WK	GPC
ρ_C	0.94 vs 0.94	0.18 vs 0.21	0.86 vs 0.86	0.52 vs 0.46	0.94 vs 0.94
ρ_G	0.16 vs 0.17	0.18 vs 0.21	0.19 vs 0.21	0.25 vs 0.19	0.22 vs 0.24
ENET	0.17	0.27	0.20	0.18	0.26
GBLUP	0.13	0.15	0.14	0.11	0.14
	TGW	SPWT	EAR	AWNS	Avg.
ρ_C	0.89 vs 0.90	0.97 vs 0.97	0.83 vs 0.83	0.30 vs 0.28	0.71 vs 0.71
ρ_G	0.19 vs 0.21	0.23 vs 0.26	0.18 vs 0.22	0.30 vs 0.28	0.21 vs 0.22
ENET	0.21	0.31	0.20	0.27	0.23
GBLUP	0.13	0.15	0.14	0.09	0.13

ρ_G = predictive correlation given all SNPs and all environmental effects.
ρ_C = predictive correlation given putative causal effects identified by the BN.

Computed for $\alpha = 0.02$ averaging 10×10-fold cross-validations, $\sigma \leqslant 0.016$ for traits and $\sigma = 0.005$ for the average. ENET is a single-trait elastic net penalised regression [14]; GBLUP is a single-trait linear mixed model.

Pros & Cons of the Two Approaches

- GBNs use fewer nodes and parameters for the same α and predictive power, and thus produce models that are potentially more stable and possibly predict better at very low sample sizes. Even so, there is no evidence suggesting that CLGBNs are overfitting.
- However CLGBNs disentangle more GxE effects, because they allow different residual variances and regression coefficients for each environment (as opposed to different intercepts in GBNs).
- CLGBNs make it possible to compute posterior probabilities of the type $\mathrm{P}(\mathrm{COUNTRY} \mid \mathrm{SNPs}, \mathrm{TRAITS})$, which is not really possible in GBNs because each level of the environmental effects is a separate node in the model.
- Both GBNs and CLGBNs are competitive with the elastic net, which is a state-of-the-art approach to genomic prediction, and at the same time they provide and intuitive representation which is useful for quantitative and qualitative reasoning.

Thanks!

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.
J. Mach. Learn. Res., 11:171–234, 2010.

D. R. Cox and N. Wermuth.
Multivariate Dependencies: Models, Analysis and Interpretation.
Chapman & Hall, Boca Raton, 1996.

E. Demidenko.
Mixed Models: Theory and Applications with R.
Wiley, 2nd edition, 2009.

J. J. Goeman.
penalized R package, 2012.
R package version 0.9-41.

Y. Guan and M. Stephens.
Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems.
Annals of Applied Statistics, 5(3):1780–1815, 2011.

References II

D. Koller and N. Friedman.
Probabilistic Graphical Models: Principles and Techniques.
MIT Press, Cambridge, 2009.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard.
Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.
Genetics, 157:1819–1829, 2001.

K. P. Murphy.
Machine Learning: A Probabilistic Perspective.
MIT Press, 2012.

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

H.-P. Piepho.
Ridge Regression and Extensions for Genomewide Selection in Maize.
Crop Sci., 49(4):1165–1176, 2009.

References III

H.-P. Piepho, J. O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, and F. Technow.
Efficient Computation of Ridge-Regression Best Linear Unbiased Prediction in Genomic Selection in Plant Breeding.
Crop Sci., 52(3):1093–1104, 2012.

M. Scutari.
Learning Bayesian networks with the bnlearn R package.
J. Stat. Soft., 35(3):1–22, 2010.

M. Scutari, P. Howell, D. J. Balding, and I. Mackay.
Multiple Quantitative Trait Analysis Using Bayesian Networks.
Genetics, 198(1):129–137, 2014.

H. Zou and T. Hastie.
Regularization and variable selection via the elastic net.
J. Roy. Stat. Soc. B, 67(2):301–320, 2005.