**University of São Paulo**
**"Luiz de Queiroz" College of Agriculture**

# Novel Bayesian networks for genomic prediction of developmental traits in biomass sorghum

**Jhonathan Pedroso Rigal dos Santos**

Thesis presented to obtain the degree of Doctor in Science.
Area: Genetics and Plant Breeding

**Piracicaba**
**2019**

**Jhonathan Pedroso Rigal dos Santos**
**Agronomist**

# Novel Bayesian networks for genomic prediction of developmental traits in biomass sorghum

Advisor:

Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Thesis presented to obtain the degree of Doctor in Science.

Area: Genetics and Plant Breeding

**Piracicaba**
**2019**

**Dados Internacionais de Catalogação na Publicação**
**DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Santos , Jhonathan Pedroso Rigal dos
Novel Bayesian networks for genomic prediction of developmental traits in biomass sorghum / Jhonathan Pedroso Rigal dos Santos . – – Piracicaba, 2019 .
52 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Bioenergia 2. Sorgo 3. Predição genômica 4. Redes Bayesianas I. Título.

## DEDICATORY

To my wife, Evelyn.

## ACKNOWLEDGMENTS


- To University of São Paulo, campus "Luiz de Queiroz" College of Agriculture (ESALQ), for all outstanding courses, seminars, physical and computational infrastructures that allowed the execution of the Doctorate's degree and project.
- To Cornell University for all outstanding seminars, physical and computational infrastructures that allowed the execution of this project.
- To my previous universities Universidade Estadual de Maringá (UEM) and Universidade Federal de Lavras (UFLA) for all background that allowed this doctorate.
- All financial resources supported by FAPESP grants 2017/03625-2 and 2017/25674-5 / CAPES (São Paulo Research Foundation) Finance Code 001 / Conselho Nacional de Desenvolvimento Cientıfico e Tecnológico (CNPq). The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000598. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
- To my national advisor Antonio Augusto Franco Garcia for his inspiring orientation, friendship, integrity and leadership.
- To my international advisor Michael A. Gore for his inspiring orientation, friendship, integrity and leadership.
- To my previous advisors Maria C. G. Vidigal, Pedro S. Vidigal Filho, Adriana Gonela, Carlos A. Scapim, Gilberto Pozar, Renzo G. Von Pinho, and Marcio Balestre for all background that allowed this doctorate.
- To my collaborators Edward S. Buckler, Roberto Lozano, Samuel B. Fernandes, Patrick J. Brown, Letícia Lara, Matheus Dalsente Krause, Deniz Akdemir and Davies Adeloye for all knowledge learned and friendship.
- To my friend, lab mate, and best man Kaio Olimpio for all knowledge, talks, friendship, and brotherhood.
- To Statistical Genetics lab (ESALQ/USP), The Gore lab (Cornell University), Buckler Lab, and TERRA-MEPP team for all knowledge and friendship.
- To my parents Maria Elenir Rigal dos Santos and Alvaro Cesar dos Santos (*In Memoriam*), my little brother Gustavo Eugenio Rigal dos Santos, family and friends for all love, education, friendship and inspiration.
- To the love of my life and wife Evelyn Graziane Rodrigues dos Santos.

# SUMMARY

# RESUMO

## Novas redes Bayesianas para predição genômica de caracteres de desenvolvimento em sorgo biomassa

O sorgo (*Sorghum bicolor* L. Moench spp.) é uma cultura bioenergética com várias características atrativas para serem exploradas no melhoramento de plantas para aumentar a eficiência de produção de bioenergia. A possibilidade de conectar informações genômicas em caracteres quantitativos ao longo do tempo, e entre caracteres, destacam as Redes Bayesianas como uma ferramenta probabilística poderosa para delinear novos modelos de predição genômica. Neste estudo, um painel diverso de 869 linhagens de sorgo foi fenotipado em quatro ambientes diferentes (2 locais em 2 anos) com medidas a cada duas semanas de 30 a 120 dias após o plantio (DAP), para altura de plantas e biomassa seca no fim da safra. Um procedimento de Genotipagem por sequenciamento foi executado, resultando na chamada de 100.435 marcadores baseados em Polimorfismos de Nucleotídeos Únicos (SNPs) bialélicos. Neste estudo foram desenvolvidos e avaliados os modelos de predição genômica: Rede Bayesiana (BN), Rede Bayesiana Pleiotrópica (PBN), e Rede Bayesiana Dinâmica (DBN). Os pressupostos para BN, PBN, e DBN foram independência, dependência entre caracteres, e dependência entre pontos no tempo, respectivamente. Para fins comparativos, formulações de modelos multivariados GBLUP foram utilizados considerando dependência entre pontos de tempo para altura de plantas (MTi-GBLUP), e ambos os pontos de tempo para a altura de plantas e biomassa seca (MTr-GBLUP), modelando matriz de variância-covariância não estruturada para efeitos genéticos e residuais. Índices de coincidência (IC) foram calculados para entender o sucesso na seleção indireta de biomassa seca usando medidas de altura de plantas, bem como um índice de coincidência baseado em linhagens (CIL), usando as amostras das posteriores das redes Bayesianas para entender a plasticidade genética ao longo do tempo. No esquema de validação cruzada 5-fold, as acurácias das predições variaram de 0,48 (PBN) a 0,51 (MTr-GBLUP) para biomassa seca e de 0,47 (DBN-DAP120) a 0,74 (MTi-GBLUP-DAP60) para altura de plantas. A validação cruzada forward-chaining mostrou um incremento substancial nas acurácias das predições ao usar o modelo DBN, com r = 0,6 (treinando no intervalo 30:45 para prever 120 DAP) até 0,94 (treinando no intervalo 30:90 para prever 105 DAP) em comparação com o BN e PBN, e semelhante aos modelos multivariados GBLUP. Os índices CI e CIL mostraram que o ranking de linhagens promissoras mudou minimamente após 45 DAP para altura de plantas. Estes resultados sugerem que 45 DAP é um estágio de desenvolvimento ideal para impor a estrutura de seleção indireta em dois níveis, onde a seleção indireta para a altura da planta no final da estação (caractere alvo de primeiro nível) pode ser feita com base na sua classificação com 45 DAP (caractere secundário), bem como para a biomassa seca (caractere alvo de segundo nível). Com o avanço das tecnologias robóticas para a fenotipagem baseada em campo, o desenvolvimento de novas abordagens, como a estrutura de seleção indireta em dois níveis, serão imperativas para aumentar o ganho genético por unidade de tempo.

**Palavras-chave:** Bioenergia; Sorgo; Predição genômica; Redes Bayesianas

# ABSTRACT

## Novel Bayesian networks for genomic prediction of developmental traits in biomass sorghum

Sorghum (*Sorghum bicolor* L. Moench spp.) is a bioenergy crop with several appealing biological features to be explored in plant breeding for increasing efficiency in bioenergy production. The possibility to connect the influence of quantitative trait loci over time and between traits highlight the Bayesian networks as a powerful probabilistic framework to design novel genomic prediction models. In this study, we phenotyped a diverse panel of 869 sorghum lines in four different environments (2 locations in 2 years) with biweekly measurements from 30 days after planting (DAP) to 120 DAP for plant height and dry biomass at the end of the season. Genotyping-by-sequencing was performed, resulting in the scoring of 100,435 biallelic SNP markers. We developed and evaluated several genomic prediction models: Bayesian Network (BN), Pleiotropic Bayesian Network (PBN), and Dynamic Bayesian Network (DBN). Assumptions for BN, PBN, and DBN were independence, dependence between traits, and dependence between time points, respectively. For benchmarking, we used multivariate GBLUP models that considered only time points for plant height (MTi-GBLUP), and both time points for plant height and dry biomass (MTr-GBLUP) modeling unstructured variance-covariance matrix for genetic effects and residuals. Coincidence indices (CI) were computed for understanding the success in selecting for dry biomass using plant height measurements, as well as a coincidence index based on lines (CIL) using the posterior draws from the Bayesian networks to understand genetic plasticity over time. In the 5-fold cross-validation scheme, prediction accuracies ranged from 0.48 (PBN) to 0.51 (MTr-GBLUP) for dry biomass and from 0.47 (DBN-DAP120) to 0.74 (MTi-GBLUP-DAP60) for plant height. The forward-chaining cross-validation showed a substantial increment in prediction accuracies when using the DBN model, with r = 0.6 (train on slice 30:45 to predict 120 DAP) to 0.94 (train on slice 30:90 to predict 105 DAP) compared to the BN and PBN, and similar to multivariate GBLUP models. Both the CI and CIL indices showed that the ranking of promising inbred lines changed minimally after 45 DAP for plant height. These results suggest that 45 DAP is an optimal developmental stage for imposing the two-level indirect selection framework, where indirect selection for plant height at the end of the season (first-level target trait) can be done based on its ranking with 45 DAP (secondary trait) as well as for dry biomass (second-level target trait). With the advance of robotic technologies for field-based phenotyping, the development of novel approaches such as the two-level indirect selection framework will be imperative to boost genetic gain per unit of time.

**Keywords:** Bioenergy; Sorghum; Genomic prediction; Bayesian networks

# 1 INTRODUCTION

The increasing trends of demands for food, energy, and population growth highlight the need to increase yield of crops with improvement of genetics and field management systems (Foley *et al.*, 2011; Mace *et al.*, 2013). Specially for production of bioenergy, sorghum (*Sorghum bicolor* L. Moench spp.) has several attractive biological features to be explored in breeding. The sorghum high biomass potential yield, strong resilience against biotic and abiotic stress, and also ancestor of many relevant bioenergy crops like maize and sugarcane, indicate this bioenergy crop as an outstanding resource to mitigate these challenges (Mullet *et al.*, 2014; Brenton *et al.*, 2016).

Recently, the substantial drop on the cost in genotyping large number of individuals, and the high cost of phenotyping plants in multiple trials attracted many public and private breeding companies to adopt different kinds of predictive systems (Heslot *et al.*, 2015). Among different predictive approaches, genomic prediction (GP) allows predicting unobserved phenotypes using information from Single Nucleotides Polymorphism (SNPs) over the genome (de Los Campos *et al.*, 2013). This approach is based on phenotyping and genotyping a population of related individuals (training set), and prediction of another set of related individuals only genotyped (test set) with trained model (Bernardo and Yu, 2007). For this purpose, plant breeding programs collected data over multiple environments representative of the breeding zone (Dias *et al.*, 2018). Also, genotypic data is collected either using a form of reduced representation of the genome, or based on a microarray-based SNP genotyping technology (Elshire *et al.*, 2011). Among the utilities of GP, this approach allows breeding programs to reduce the amount of financial resource spent on phenotyping within a trial, the amount of trials, and also allows predict the best parental crosses before the growing season.

Over the last years, many statistical and machine learning models have been tailored to use genomic information. Some examples are the linear mixed models (e.g. rrBLUP, GBLUP), Bayesian models (e.g. BayesA, BayesB), kernel methods and neural networks (Meuwissen *et al.*, 2001; de Los Campos *et al.*, 2013; Heslot *et al.*, 2015). Despite the large number of models developed so far, most of these tend to show similar predictive performance across different traits and species (de Los Campos *et al.*, 2013). Recently, the development of models exploiting information over multiple traits and time points have been showing substantial improvement in relation to others that assume independence (Ratcliffe *et al.*, 2015; dos Santos *et al.*, 2016; Campbell *et al.*, 2018). GP with multiple traits allows recovering information of markers linked to genes displaying pleiotropic effects. Phenotypes measured over multiple time points allows to recover the information of the trajectories of the genetic effects over time. One example of model to recover genetic information between traits and time points by modelling their genetic correlation is the Multivariate GBLUP model (Ratcliffe *et al.*, 2015; dos Santos *et al.*, 2016). Novel GP models modelling the casuaility of these effects over multiple traits and time points could improve even more the interpretation and quality of predictions from genomic prediction models.

Bayesian paradigm is a modelling approach suitable to unify information from a experiments into a likelihood probability function, from previous experiments into a prior probability distribution, and merge both sources of information into a posterior probability function us-

ing the Bayes theorem (Gelman *et al.*, 2014; Goodfellow *et al.*, 2016). Bayesian Networks (BN) is a class of probabilistic graphical models for modelling structured relationships among explanatory factors (Bishop, 2013). In genetics, BN can be used to connect information of genetic factors affecting multiple traits and time points. In this study, we main goals were: (i) perform a phenotypic analysis to eliminate non genetic experimental effects of plant height time series and dry biomass data, (ii) propose a strategy based on principal component analysis to mitigate computational burden of genomic prediction analysis, (iii) develop the Bayesian Network, Pleiotropic Bayesian Network, Dynamic Bayesian Network, Multi Time GBLUP (considering all plant height repeated measures), and Multi Trait GBLUP (considering all plant height repeated measures, and dry mass) to exploit information between traits and time points for genomic prediction, (iv) predict observed and completely non observed plant height data points across time points, and dry mass at the end of the season, (v) propose indexes based on the results of the Bayesian analysis to identify opportunities for selection before the end of the growing season, and (vi) propose a novel breeding indirect selection strategies to optimize the genetic gain per unit of time of dry mass and plant height of sorghum biomass.

## References


Bernardo, R. and J. Yu, 2007 Prospects for genomewide selection for quantitative traits in maize. Crop Science **47**: 1082–1090.

Bishop, C. M., 2013 Model-based machine learning. Phil Trans R Soc A **371**: 1–17.

Brenton, Z. W., E. A. Cooper, M. T. Myers, R. E. Boyles, N. Shakoor, K. J. Zielinski, B. L. Rauh, W. C. Bridges, G. P. Morris, and S. Kresovich, 2016 A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. Genetics **204**: 21–33.

Campbell, M., H. Walia, and G. Morota, 2018 Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. Plant Direct **2**: e00080.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345.

Dias, K. O. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da Costa e Silva, S. N. Parentoni, P. E. de Oliveira Guimarães, C. de Oliveira Anoni, J. M. V. Pádua, M. de Oliveira Pinto, R. W. Noda, C. A. G. Ribeiro, J. V. de Magalhães, A. A. F. Garcia, J. C. de Souza, L. J. M. Guimarães, and M. M. Pastina, 2018 Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity .

dos Santos, J. P. R., R. C. De Castro Vasconcellos, L. P. M. Pires, M. Balestre, and R. G. Von Pinho, 2016 Inclusion of dominance effects in the multivariate GBLUP model. PLoS ONE **11**: 1–21.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE **6**: e19379.

Foley, J. a., N. Ramankutty, K. a. Brauman, E. S. Cassidy, J. S. Gerber, M. Johnston, N. D. Mueller, C. O'Connell, D. K. Ray, P. C. West, C. Balzer, E. M. Bennett, S. R. Carpenter, J. Hill, C. Monfreda, S. Polasky, J. Rockström, J. Sheehan, S. Siebert, D. Tilman, and D. P. M. Zaks, 2011 Solutions for a cultivated planet. Nature **478**: 337–342.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2014 *Bayesian Data Analysis*.

Goodfellow, I., Y. Bengio, and A. Courville, 2016 *Deep Learning*. MIT Press, `http://www.deeplearningbook.org`.

Heslot, N., J. L. Jannink, and M. E. Sorrels, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Science **55**: 1–12.

Mace, E. S., S. Tai, E. K. Gilding, Y. Li, P. J. Prentis, L. Bian, B. C. Campbell, W. Hu, D. J. Innes, X. Han, A. Cruickshank, C. Dai, C. Frère, H. Zhang, C. H. Hunt, X. Wang, T. Shatte, M. Wang, Z. Su, J. Li, X. Lin, I. D. Godwin, D. R. Jordan, and J. Wang, 2013 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nature Communications **4**: 1–9.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Mullet, J., D. Morishige, R. McCormick, S. Truong, J. Hilley, B. McKinley, R. Anderson, S. N. Olson, and W. Rooney, 2014 Energy sorghum–a genetic model for the design of C4 grass bioenergy crops. Journal of Experimental Botany **65**: 3479–89.

Ratcliffe, B., O. G. El-Dien, J. Klapste, I. Porth, C. Chen, B. Jaquish, and Y. A. El-Kassaby, 2015 A comparison of genomic selection models across time in interior spruce ( Picea engelmannii × glauca ) using unordered SNP imputation methods. Heredity **115**: 547–555.

# 2 LITERATURE REVIEW

## 2.1 Sorghum as a bioenergy crop

The rising levels of population growth, polluted gases and the reduction of non-renewable energy resources have been suggesting a strong demand for food, fibers, grains, bioproducts and bioenergy (Foley *et al.*, 2011; Mace *et al.*, 2013). Among the several possibilities to handle a variety of challenges in the future, the genetic improvement of crops may directly or indirectly help to deal with many of those demands. After partial degrees of domestication and breeding through history, some crops display many different biological values for bioenergy production. Some examples of bioenergy crops include the species switchgrass (*Panicum virgatum*), sugarcane (*Saccharum* spp.), *Miscanthus* ( *Miscanthus spp.*), napier grass (*Pennisetum purpureum*) and sorghum (*Sorghum bicolor* L. Moench spp.) (Mullet *et al.*, 2014).

Among the mentioned crops, sorghum is a quintessential bioenergy crop. Sorghum belongs to the family Poaceae and subfamily Panicoideae, as the majority of grasses useful for bioenergy production (Calviño and Messing, 2011). In terms of evolution landmark, some archeological evidences indicate Ethiopia and Sudan as the initial domestication center of sorghum more then 8000 years ago (Mace *et al.*, 2013). The migration of sorghum happened especially in trade routes from Africa to Asia. Over this migration time, sorghum had run through a selective bottleneck and natural selection process that differentiated its species into four well known races named Durra, Caudatum, Guinea and Kefir. In the America continent, the introduction of sorghum germplasm happened at the 19th century (Mullet *et al.*, 2014).

Recently, after these domestication events, the sorghum breeding process has developed thought artificial selection multiple kinds of sorghum. Those were categorized based on its biochemical content. The most important categories are known as sweet sorghum, grain sorghum and photoperiod sensitive sorghum (Rooney *et al.*, 2007). The oligosaccharides sucrose, glucose, fructose and starch are the main composition of sweet sorghum. On the other side, the grain sorghum is enriched specially with a higher proportion of starch deposited on the grains. The photoperiod sensitive sorghum has higher proportion of cell walls and lignin content (Mullet *et al.*, 2014; Regassa and Wortmann, 2014). Considering the vast plasticity of sorghum biomass composition, this crop can be considered as a critical and imperative resource for breeding to balance novel needs of biomass content. The unification of these features with its strong resilience against biotic and abiotic stress turn sorghum an outstanding model organism to extrapolate new findings to other syntenic bioenergy crops (Regassa and Wortmann, 2014).

In terms of genetics and physiology, sorghum is diploid (2n=20) and fixes carbon through the $C_4$ pathway (Vermerris, 2011; Lawrence and Walbot, 2007). The non-complex genome and efficient photosynthetic system, also, ancestor of many relevant bioenergy crops like maize and sugarcane, enforces sorghum as a important model organism. Sorghum has a predominantly self-pollinating reproduction system. The possibility to easily make artificial crosses and natural self-fertilization can also been seen as attractive features for implementing well-established breeding methods in industry. The natural intellectual protection of cultivars by hybridization of inbred lines to obtain single-cross hybrids also favour the interest of companies

to invest in sorghum breeding (Mullet *et al.*, 2014).

## 2.2 Novel sequencing technologies

In last century, maize showed an incredible example of ∼4-fold yield increase by mutual effort of improvements in genetics and field management systems (Hammer *et al.*, 2009). Recently, Mullet *et al.* (2014) speculated that the same achievements observed in maize are equally possible for sorghum. The potential yield of sorghum ideotypes is ∼55-60 dry Mg ha$^{-1}$ under optimal water regime and ∼15-25 dry Mg ha$^{-1}$ in non-water supplied conditions (Mullet *et al.*, 2014). For leveraging the genetics of sorghum to achieve such biomass production, the adoption of modern breeding tools are a stepping stone towards the path for fast and efficient success of bioenergy breeding programs.

The whole-genome sequencing is a modern breeding tool to obtain information at the nucleotide level over the genome. The first effective and widespread technology to reveal nucleotides sequence as units of DNA (deoxyribonucleic acid) is known as Sanger sequencing. In short, this procedure is based on the generation of DNA fragments with variable length, terminated with a labeled nucleotide, which gives signals of the identity of the sequence using some capillary gel electrophoresis system (Shendure and Ji, 2008). Despite its importance in the past, Sanger sequencing have lost relevance especially due its low automation capacity because of its difficulties for parallel sequencing (Shendure and Ji, 2008).

More recently, novel second-generation sequencing technologies have been proposed to overcome the limitations of Sanger sequencing. Many of these are based on obtaining small DNA fragments, tagged nucleotides and parallel sequencing on high-tech physical environments. Two popular approaches widely used are the sequencing by synthesis on flowcells (Illumina technology), and SMRT sequencing on zero-mode waveguides (PacBio technology) (Sims *et al.*, 2014; Goodwin *et al.*, 2015). Some available platforms are the Solexa technology, SOLiD platform, Polonator and others (Goodwin *et al.*, 2015). Genomic data obtained by these platforms allows detection of single-nucleotide polymorphism (SNP), small insertions and deletions (indels), larger structural variants and copy number variants (CNVs) (Sims *et al.*, 2014). Using Illumina technology, Mace *et al.* (2013) could identity 4,946,038 SNPs, 1,982,971 indels, and 120,929 CNVs by sequencing the DNA of a panel composed by 44 sorghum lines.

Despite the power of second-generation technologies to reveal genomic information, the sequencing process is still limited by the high financial cost to be applied routinely in plant breeding (Peterson *et al.*, 2012). Over the breeding course, several breeding populations are generated dynamically over breeding cycles, which hampers its usage due its high operational cost. In order to be able to extend this technology for large scale sequencing, the preparation of libraries of fragmented DNAs with the reduced representation of the genome with restriction enzymes, or some physical fragmentation technique, have been shown to be a robust, efficient, and cheap approach for large-scale genotyping (Elshire *et al.*, 2011). In general, the genotyping methods vary in according to the library preparation steps, which usually influence costs and applications. Some example includes the genotyping-by-sequencing (GBS) (Elshire *et al.*, 2011), Double Digest Restriction-site Associated DNA sequencing (RADseq) (Peterson *et al.*, 2012) and Diversity Array technology (DArT) (Jaccoud *et al.*, 2001). In sorghum breeding,

DNA polymorphisms found during genotyping can be used as molecular markers for applications like artificial crosses designs, gene mapping and predictions (Baird *et al.*, 2008; Morris *et al.*, 2013).

## 2.3 Genomic prediction

Among the challenges of breeding crops, the choice of crosses to obtain superior progenies is critical for the success of breeders (Bernardo and Yu, 2007). The limited budget, people, resources, and field facilities across multiple locations make this challenge even harder during the breeding process (Heslot *et al.*, 2015). As an example, only 100 lines are capable of generating 4950 unique hybrids combinations. The empirical knowledge of germplasm from breeders, and statistical analytical tools to identify the best parental combinations can help to mitigate these challenges.

To help predict the performance of unobserved progenies, the genomic prediction (GP) approach can leverage the information obtained by high density SNPs of evaluated and non-evaluated plants in the field. This approach build predictors using as input the phenotypes and SNPs of evaluated progenies (training set), and predict the related non-evaluated progenies only using the SNPs (test set) (Heslot *et al.*, 2015). The GP process usually is done by either training a linear or nonlinear statistical model using phenotypes as response variables, and genotypes as covariates. In plant breeding, the GP approach have been used extensively to optimize resources by reducing the number of candidate cultivars tested, trials, and prediction of promising crosses (Heslot *et al.*, 2015). Specially in sorghum, GP have been boosting the management of gene banks by making possible the prediction of unevaluated accessions (Yu *et al.*, 2016).

Some popular models for GP are linear mixed models, hierarchical Bayesian models with informative priors, kernel methods, and neural nets (de Los Campos *et al.*, 2013; Heslot *et al.*, 2015; dos Santos *et al.*, 2016a). In the linear mixed model framework the parameters are modelled with the population mean considered as a fixed effect, and implement either as a form of regression directly on markers with effects modelled as random and following an isotropic Gaussian distribution (rrBLUP), or using markers to build a identity-by-state realized kinship matrix assuming genetic effects as random effects and following a multivariate Gaussian distribution (GBLUP). The hierarchical Bayesian models differ by the choice of the informative priors, varying from isotropic Gaussians (e.g. BayesA, Bayesian Linear Regression), mixture of isotropic Gaussians with known (e.g. BayesB) or unknown mixture parameters (e.g. BayesC$\pi$), to isotropic Laplace distributions (e.g. Bayesian Lasso) (de Los Campos *et al.*, 2013). The kernels methods are based on covariance or distance matrices constructed applying some kernel function on markers codification (de Los Campos *et al.*, 2013). Neural nets are based on different representations of the input (markers) data on hidden layers composed by several hierarchical nonlinear functions of unknown weights optimized via supervised learning approaches (Goodfellow *et al.*, 2016).

After several empirical and *in silico* studies comparing the performance of genomic prediction models, few differences have been observed between models across different traits and species. The multicollinearity between the columns of the design (marker) matrix, and the

applied NP problem nature of these may justify few differences. Genetically, the multicollinearity may be justified by the strong linkage disequilibrium of markers linked to a single or multiple genes controlling the trait, bringing redundant information to be explored in this regression process. Also, most of these models just use as target variable only one trait, and data collected in the end of the growing season. To handle these situations, specially for multiple traits, some extension of the multivariate linear mixed model have been proposed (Calus and Veerkamp, 2011; dos Santos *et al.*, 2016b; Dias *et al.*, 2018; Fernandes *et al.*, 2018). So far, few studies have proposed models to exploit information over time points covering different crop growth stages.

As advantages, GP models for multiple traits can recover the information of pleiotropic genes instead of restricting for genetic effects influencing only one main trait. Measures of stages of plants during growth can leverage the information of the trajectory of genetic effects over time. In rice, Campbell *et al.* (2018) have developed nonlinear random regression models to predict the sum of pixels covering the plant on the image as a measure of shoot biomass. The time effect was modelled with second-order Legendre polynomials. The plants were evaluated daily during the initial growth stage (13 to 33 DAP) in the greenhouse. The nonlinear random regression models improved prediction accuracy up to 11.6% compared to others that do not share information across time points. Despite the importance of such approach to predict data from plants at initial growth stages, the second-order polynomials works well specially when the growth curve has an exponential shape, and may not generalize for data points in the end of the season. In tree breeding, Ratcliffe *et al.* (2015) evaluated GP models to predict height using the BC$\pi$, rrBLUP, and generalized ridge regression models using repeated measures covering six sparse measures over 3$\sim$40 years. Few differences of prediction performance were observed between models. Their evaluated models did not recover information over time. Novel genetic models recovering information between traits, or time points, can be a opportunity to improve substantially the effectiveness of genomic prediction models (dos Santos *et al.*, 2016b; Campbell *et al.*, 2018).

## 2.4 Bayesian data analysis: example for genomic prediction

In statistics, the frequentist paradigm assumes that exists only one true value of a parameter, it must be fixed and unknown, and estimated as a function of a data set - data set treated as random (Bishop, 2013; Goodfellow *et al.*, 2016). Another important field in statistics is known as Bayesian paradigm. This line believes the point value of a parameter is never know, and it should be treated as unknown and represented by a random variable learned directly on the data set - data set treated as not random and observed. The Bayesian paradigm learns the likely values of a unknown parameter using the probability theory rules (Gelman *et al.*, 2014; Goodfellow *et al.*, 2016). In this framework, previous knowledge is mapped on a prior probability function, the information from experiments into a likelihood probability function, and the Bayes theorem is used to merge both into a posterior probability distribution (Gelman *et al.*, 2014),

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2.1}$$

where $p(y|\theta)$ is the conditional likelihood probability function, $p(\theta)$ is the marginal prior probability of the parameter $\theta$, $p(y)$ is the marginal probability function of the data.

In GP, we are interested in learning the effects ($\beta$) of a set $\{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ik}\}$ of $k$ SNPs on the phenotype ($y_i$) of the $i^{th}$ line from a population of $n$ lines with mean $\mu \in \mathbb{R}$, allelic dosage $x_{ij} \in \{2, 1, 0\}$, SNP effect $\beta \in \mathbb{R}^k$, and phenotype $y \in \mathbb{R}^n$, where $\mathbb{R}^k$ and $\mathbb{R}^n$ denotes real space with $k$ and $n$ dimensions, respectively.

The first step in Bayesian data analysis requires specify a likelihood probability function reflecting a probabilistic mechanism that generates the data given the parameters. This probability distribution will concatenate the information from the phenotypic data obtained by a plant breeding trial. In the case of quantitative traits, a widely accepted likelihood probability function is the normal distribution (Mackay *et al.*, 2009; Barton *et al.*, 2017),

$$y_i \mid \; x_i, \mu, \beta, \sigma \sim \mathcal{N}(y_i \mid \mu + x_i\beta, \sigma) \tag{2.2}$$

where $\sigma \in \mathbb{R}^+$ represent the scale parameter controlling the uncertainty around the expected value $\hat{y}_i = \mu + x_i\beta$.

The next step is to specify the prior probability functions reflecting our current beliefs about the unknowns parameters. In the example case, for $\mu$ and $\beta_j$, the normal probability function may be a wise prior choice in the light of Fisher the infinitesimal model, which states that genes have small and linear cumulative contribution on the phenotype (Fisher, 1918; Mackay *et al.*, 2009; Barton *et al.*, 2017). In this perspective, the conditional probability prior distributions for these two parameters are given by (Feller, 1968),

$$\mu \mid 0, s_\mu \sim \mathcal{N}(\mu \mid 0, s_\mu) \tag{2.3}$$

$$\beta_j \mid 0, s_\beta \sim \mathcal{N}(\beta_j \mid 0, s_\beta) \tag{2.4}$$

where the probability point of mass a priori is centered around zero, and the $s_\mu \in \mathbb{R}^+$ and $s_\beta \in \mathbb{R}^+$ represents scale hyperparameters describing the uncertainty regarding the likely values of the parameter around zero. In non hierarchical Bayesian models, we can simply create a uninformative prior with high entropy by setting the scale hyperparameters to a known high positive value depending on the scale of the phenotype (Goodfellow *et al.*, 2016). However, this approach may have the disadvantage of subjectiness when choosing the high scalar value (Bishop, 2013). Also, this approach may cause posterior mean discalibration, that is, the situation where bias may be introduced in the model if the likely values assumed a priori is too much different from the one expected by the true mechanism that generates the data (Gelman *et al.*, 2014).

As outlined above, the scale parameter $\sigma$ represent our uncertainty about the phenotypic values $y_i$. Due to this parameter restricted to positive real numbers, it may be modeled by a half Cauchy prior probability function to avoid the presence of unexpected large predicted values after the learning process (Feller, 1968; Gelman *et al.*, 2008, 2014),

$$\sigma \mid 0, s_\sigma \sim \text{Cauchy}^+(\sigma \mid 0, s_\sigma) \tag{2.5}$$

where the probability point of mass a priori is centered around zero, and the $s_\sigma \in \mathbb{R}^+$ represents a scale hyperparameter describing the uncertainty regarding the likely values of the parameter around zero. Similar to the normal prior distribution case, an uninformative half Cauchy prior can be constructed by assuming the value of $s_\sigma$ as a known high scalar value. However, a strategy to eliminate the subjectiveness in defining those scale hyperparameters can be obtained by using a hierarchical structure instead of a non hierarchical formulation. In this approach, hyperpriors are added to the hyperparameters - prior distribution of the hyperparameters, which the information from the data can be used to learn their most likely values in a full probabilistic model only controlled by a global known hyperparameter. In the hierarchical formulation, the hyperparameters are treated as unknown parameters instead of known subjective parameters set by the user. In this hierarchical Bayesian model formulation, we can define the hyperpriors for the scale hyperparameters using half Cauchy priors (Gelman *et al.*, 2008, 2014),

$$s_\mu \,|\, 0, \phi \sim \text{Cauchy}^+(s_\mu \,|\, 0, \phi) \tag{2.6}$$

$$s_\beta \,|\, 0, \phi \sim \text{Cauchy}^+(s_\beta \,|\, 0, \phi) \tag{2.7}$$

$$s_\sigma \,|\, 0, \phi \sim \text{Cauchy}^+(s_\sigma \,|\, 0, \phi) \tag{2.8}$$

where the hyperpriors are centered around zero and have a scale global hyperparameter $\phi$ for creating a parameter sharing mechanism to recover information between hyperpriors. One strategy to handle this global hyperparameter is to set to a known but not large value to build a weakly informative prior (Gelman, 2006; Gelman *et al.*, 2008, 2014). One choice may be a measure of the size of the vector of the phenotypes, for instance, we may define $\phi$ to build a weakly informative prior by using the $L^\infty$ norm (Goodfellow *et al.*, 2016) of the phenotypic vector times a constant of order 10,

$$\phi = ||y||_\infty \times 10 = \max_i |y_i| \times 10 \tag{2.9}$$

where max represent the maximum value in the vector.

This known hyperparameter will create hyperpriors that will have weakly information without introducing dramatically discalibration at the posterior means, and eliminate the subjectiveness in determining known values of hyperparameters in the priors (Gelman *et al.*, 2008, 2014). The main advantage of using the global hyperparameter as a function of the $L^\infty$ norm of the data is that the hyperprior will most of the time be invariant to the scale, if the dataset represent most values of the true distribution that generates the data.

After defining the likelihood, priors, and hyperpriors distribution, the next step in Bayesian data analysis is specify the joint distribution of all model unknowns, which can be obtained by the product of the conditional probability distributions by the chain rule (Bishop, 2013),

$$p(y, \mu, \beta, \sigma, s_\mu, s_\beta, s_\sigma \,|\, X, \phi) = \prod_{i=1}^{n} \mathcal{N}(y_i \,|\, \mu + x_i\beta, \sigma)\, \mathcal{N}(\mu \,|\, 0, s_\mu)\, \mathcal{N}(\beta \,|\, 0, s_\beta)$$
$$\text{Cauchy}^+(\sigma \,|\, 0, s_\sigma)\, \text{Cauchy}^+(s_\mu \,|\, 0, \phi)$$
$$\text{Cauchy}^+(s_\beta \,|\, 0, \phi)\, \text{Cauchy}^+(s_\sigma \,|\, 0, \phi) \quad (2.10)$$

To merge the prior information with the information of the experiments, we can do ancestral sampling on the joint distribution using some Markov Chain Monte Carlo (MCMC) method (Goodfellow *et al.*, 2016). One important point is that the joint distribution has the same functional form as the unnormalized joint posterior distribution, that is, the numerator of the Bayes Theorem expression - we can ignore the denominator of the Bayes theorem to obtain the solutions because they are not a function of the parameters, but only function of the data (marginal distribution of the data or normalization constant to the probabilities sum to one) (Murphy, 2013; Bishop, 2007).

In most of the current Bayesian GP models, usually it is needed to have the conditional posterior distribution of each parameter conditioned on all other parameters for performing the MCMC engine, which requires both likelihood and prior distributions with mathematically tractable probability functions. This Bayesian approach is known as conjugate analysis. After obtaining the analytical conditional posterior distributions, the parameters can be integrated using a MCMC integration algorithm known as Gibbs sampler (de Los Campos *et al.*, 2013). However, one of the disadvantages of this approach is that the priors usually are required to be chosen subjectively to be able to conduct the conjugate analysis, and it may result in posterior distributions displaying discalibration at the posterior mean with unlikely high values (Gelman *et al.*, 2008, 2014).

As mentioned before, a solution to avoid discalibration of the posterior mean for the scale components can be accomplished with half Cauchy priors (Gelman *et al.*, 2008). However, the Gibbs sampler algorithm can not be implemented because the conditional posteriors can not be obtained algebraically as required by the conjugate analysis. A convenient solution for this non-conjugate problem is to use a general approach based on the Hamiltonian Monte Carlo (HMC) algorithm, which on the analytical posterior distributions are not required for sampling (Gelman *et al.*, 2014; Hoffman and Gelman, 2014). The HMC algorithm uses the dynamics of the samples and the gradient of the posterior distribution estimated with the backpropagation algorithm with computational graphs, to learn directions of the parameter space with the most likely values of the parameters controlling the joint distribution. The No-U-Turn Sampler (NUTS) algorithm is a form of HMC method with parameters automatically tuned. This algorithm usually shows the same performance (or even better) then the Gibbs sampler. Also, it does not show random walk behavior, usually it does not concentre the samples around the modes, and conjugate priors are not required (Hoffman and Gelman, 2014; Goodfellow *et al.*, 2016). The NUTS algorithm is available in many high level languages like python (PyMC3, pystan) and R (rstan) (Team, 2018; Stan Development Team, 2018; Carpenter *et al.*, 2017). These open source probabilistic programming platforms can be easily used for developing novel GP models for research and production.

Once the training process is over and posterior samples are available, predictions can be obtained by obtaining an estimate of the posterior mean of the parameters by averaging over the MCMC samples. Predictions of the phenotypes in the test set are obtained by,

$$\hat{y}_i^{[\text{test}]} = \hat{\mu} + x_i^{[\text{test}]}\hat{\beta} \tag{2.11}$$

where $\hat{y}_i^{[\text{test}]}$ is the phenotype of the i$^{\text{th}}$ in the test set and $x_i^{[\text{test}]}$ is its marker vector, $\hat{\mu}$ the posterior mean estimate of the population mean, and $\hat{\beta}$ is the posterior mean effects of the markers.

## 2.5 Bayesian networks

As mentioned before, Bayesian theory is a powerful modelling approach to unify prior information with the ones obtained by experiments to update some state of knowledge(Gelman *et al.*, 2014). A Bayesian model can be understood by factoring a joint distribution of random variables into a product of conditional probability distributions by the application of the chain rule,

$$p(x_1, x_2, x_3, x_4) = p(x_4|x_3, x_2, x_1)p(x_3|x_1, x_2)p(x_2|x_1)p(x_1) \tag{2.12}$$

On the other hand, non-structured models like the ones shown on equation (1), with fully flexible joint distribution, can become an intractable problem as the number of unknown variables increases. To mitigate this problem, a better strategy may be obtained by using some structured representation, based on the knowledge about relationships between the unknown variables (Bishop, 2013). For example,

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3)p(x_4|x_2, x_3) \tag{2.13}$$

With structured representations, a powerful way to understand the relationship among known and unknown factors can be achieved by a graphical representation called probabilistic graphical models (Bishop, 2013). The most widely used probabilistic graphical models are known as directed graphical models or Bayesian networks, undirected graphical models or Markov random fields, chain models and factor graphs (Hamelryck, 2012). Among these structured representations, Bayesian networks (BN) have several interesting properties to develop new statistical genetic models to help better improve crops. Some advantages of BNs includes: (i) generalization of any statistical or machine learning model that can be represented by a directed acyclic graph, (ii) model can be tailored to a specific problem without changing the inference algorithm (numerical optimization or integration), (iii) the structured representation of the network naturally shows the nature of the problem, (iv) missing data are handled naturally, and (v) a *priori* knowledge can be included in the model (Moreau *et al.*, 2003; Bishop, 2013).

In short, the Bayesian network is defined as a graphical representation of a structured joint distribution factored into a set of conditional probability distributions, where nodes represent variables - known variables by shaded nodes, and unknown variables by unshaded nodes,

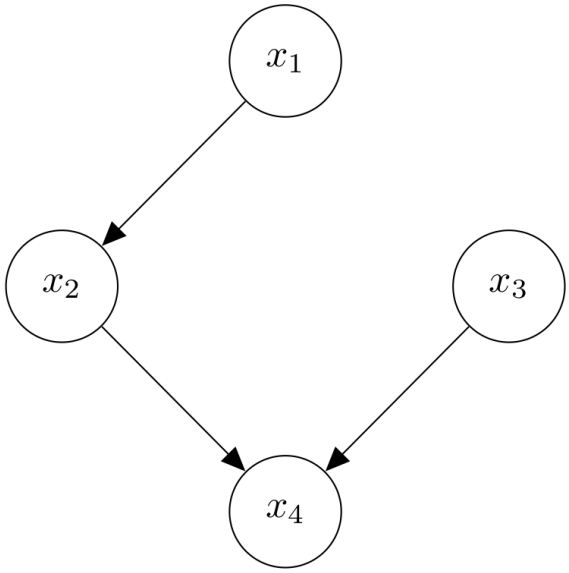


**Figure 2.1:** Factorization of a joint probability distribution in a set of conditional distributions using Bayesian networks as graphical syntax notation.

and arrows the dependence between them. An example of BN structure of the right-hand side of the equation (2) can be seen at the Figure (2.1).

In the Figure (2.1), the node $x_1$ has an independent relationship with the nodes $x_2$, $x_3$ and $x_4$, as a result, there is no arrows from them to the node $x_1$. The node $x_2$ has a dependent relationship with $x_1$, since there is a direct arrow linking the node $x_1$ to the node $x_2$. The node $x_3$ is independent to the others, similar to what was observed with the node $x_1$. Finally, the node $x_4$ has a dependent relationship with nodes $x_2$ and $x_3$, once both nodes are connected with arrows to the node $x_4$.

In a BN representation, arrows dictates if nodes are parent or child of others. For instance, if there is an arrow expressing a conditional relationship of the node $x_2$ with the node $x_1$, we can say that $x_1$ is a parent of the node $x_2$, and that $x_2$ is a child of the node $x_1$. In BNs, no loop structure of arrows must connect a specific subset of nodes in the net. The factorization of a joint probability distribution into a set of conditional distributions can also be given by,

$$p(\boldsymbol{x}) = \prod_{x_j}^{J} p(x_j|pa_j) \tag{2.14}$$

where $pa_j$ represents the parents of the node $x_j$, and $x = \{x_1, ..., x_J\}$

This key property of BNs is known as Markov condition, and it states that a variable (child) is only dependent on the information of its parents in the net. In terms of inferential algorithms, this can simplify substantially computations with multiple variables, once only local inferences can be performed with a reduced set of variables, and each modularity may be run in parallel with multiple processors (Su *et al.*, 2013; Bishop, 2013).

The joint distribution of the parameters controlling the Bayesian hierarchical regression model for genomic prediction described in the section 2.4 can also be shown by a probabilistic graphical model (Figure 2.2). The Markov condition and the Bayesian network visualization are enough to reconstruct the joint distribution shown on equation 2.10. The elaboration of

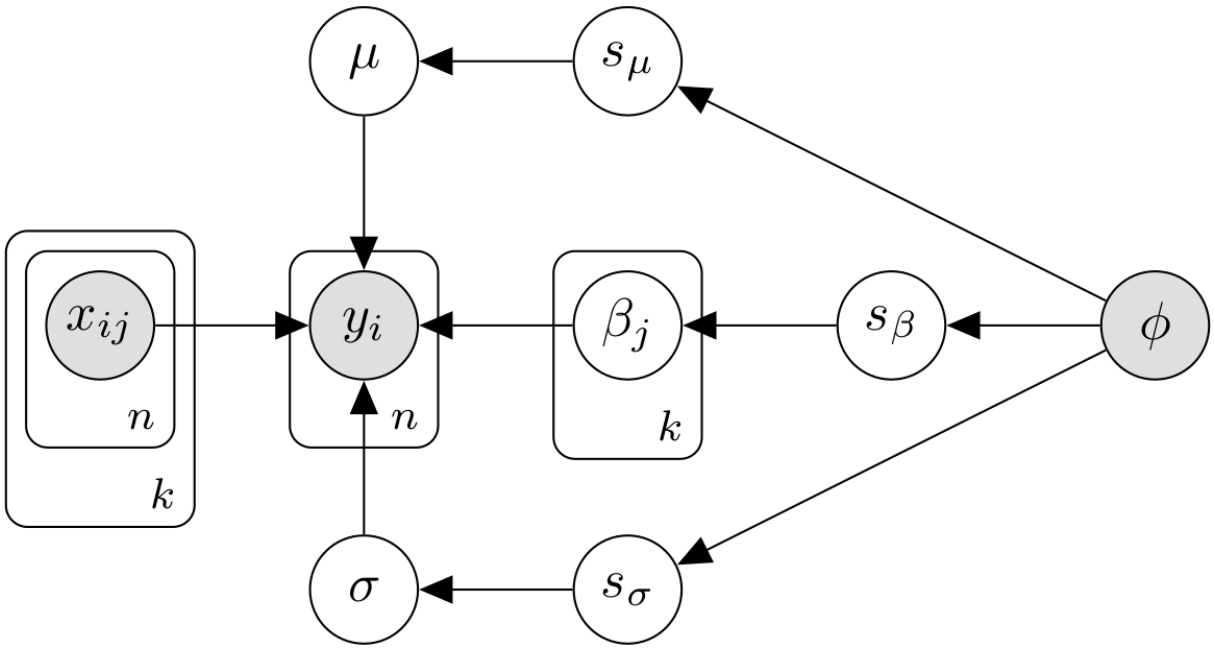


**Figure 2.2:** Hierarchical Bayesian linear regression model joint distribution network representation. $y_i$: phenotype of the $i^{\text{th}}$ line, $x_{ij}$: allelic dosage for the $i^{th}$ line in the $j^{th}$ locus, n: number of lines, k: number of markers, $\mu$: population mean, $\beta_j$: effect of the $j^{th}$ marker, $\sigma$: residual standard deviation, $s_\mu$: scale hyperparameter of the population mean, $s_\beta$: scale hyperparameter of the marker effects, $s_\sigma$: scale hyperparameter of the standard deviation, $\phi$: global known hyperparameter.

more complex Bayesian network architectures connecting phenotypes collected over multiple time points and traits can be derived using as building blocks the probabilistic structure shown on Figure (2.2).

Specially with high-dimensional molecular marker data, the possibility to speed up computations using the Markov condition property turn BNs approach extremely useful for the development of computational efficient procedures in different genetic applications. The possibility to use any probability distribution to model unknown variables, and also the high versatility of the directed acyclic graph structure for mapping the relationship between nodes make BN a powerful approach to solve many challenges in genetics science. For instance, there are applications of BNs in genome assemblies (Loman *et al.*, 2015), genotyping (Serang *et al.*, 2012; Garcia *et al.*, 2013), gene interactions (Han *et al.*, 2012), gene-environment interactions (Su *et al.*, 2013), gene expression patterns (Neapolitan *et al.*, 2013), and many others.

In genetics, one of the great challenges is understanding the genetics of polyploids, specially to discover the genotypic state of a given loci using data derived from second-generation sequencing technologies. One powerful application of BNs was performed by elucidating the genotypic states of loci with variables degrees of polyploidy (Serang *et al.*, 2012; Garcia *et al.*, 2013). In this study, the BN had as variables (nodes): (i) the level of ploidy ($P$), (ii) the genotypic distribution of the population ($G$) that depends on the level $P$, (iii) the number of individuals ($C$) assigned for a given $G$, and (iv) the distribution of allelic frequencies ($T$) depending on $P$. This BN uses a generative process by sampling hierarchically from the BN, and identify the optimal joint genotypic configuration by maximizing the joint probability of the

observed data ($D$). The implementation of this BN is available on the software SuperMASSA (Serang *et al.*, 2012; Garcia *et al.*, 2013). Nowadays, the routinely genotyping of polyploids is possible due this great advance with probabilistic graphical models (Garcia *et al.*, 2013).

The flexibility to use categorical and continuous distributions in the same modelling framework allows BNs to aggregate different powerful machine learning algorithms as building blocks for the development of novel GP models. The unification of the advantages from different models and avoidance of disadvantages could be used to reshape novel GP models with BNs. The understanding of the uncertainty, possibility to compute indexes conditioned on the observed data, and also probabilities for inference motivates even more BNs. The employment of this approach to exploit the information from pleiotropic genes influencing multiple traits, and also to leverage the signals from trajectories of genetic influences triggered by the effects of genes over time could substantially improve genomic prediction models for predicting and understanding complex multiple traits over different growth stages.

## References

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson, 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE **3**: e3376.

Barton, N., A. Etheridge, and A. Véber, 2017 The infinitesimal model: Definition, derivation, and implications. Theoretical Population Biology **118**: 50 – 73.

Bernardo, R. and J. Yu, 2007 Prospects for genomewide selection for quantitative traits in maize. Crop Science **47**: 1082–1090.

Bishop, C. M., 2007 *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, first edition.

Bishop, C. M., 2013 Model-based machine learning. Phil Trans R Soc A **371**: 1–17.

Calus, M. P. and R. F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution **43**: 26.

Calviño, M. and J. Messing, 2011 Sweet sorghum as a model system for bioenergy crops. Current Opinion in Biotechnology **23**: 323–329.

Campbell, M., H. Walia, and G. Morota, 2018 Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. Plant Direct **2**: e00080.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, 2017 Stan: A probabilistic programming language. Journal of Statistical Software, Articles **76**: 1–32.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345.

Dias, K. O. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da Costa e Silva, S. N. Parentoni, P. E. de Oliveira Guimarães, C. de Oliveira Anoni, J. M. V. Pádua, M. de Oliveira Pinto, R. W. Noda, C. A. G. Ribeiro, J. V. de Magalhães, A. A. F. Garcia, J. C. de Souza, L. J. M. Guimarães, and M. M. Pastina, 2018 Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity .

dos Santos, J. P., L. Paulo, M. Pires, R. Coelho, and D. C. Vasconcellos, 2016a Genomic selection to resistance to Stenocarpella maydis in maize lines using DArTseq markers. BMC Genetics **17**: 1–10.

dos Santos, J. P. R., R. C. De Castro Vasconcellos, L. P. M. Pires, M. Balestre, and R. G. Von Pinho, 2016b Inclusion of dominance effects in the multivariate GBLUP model. PLoS ONE **11**: 1–21.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE **6**: e19379.

Feller, W., 1968 *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley.

Fernandes, S. B., K. O. G. Dias, D. F. Ferreira, and P. J. Brown, 2018 Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theoretical and Applied Genetics **131**: 747–755.

Fisher, R., 1918 The correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburgh **52**: 399–433, cited By 93.

Foley, J. a., N. Ramankutty, K. a. Brauman, E. S. Cassidy, J. S. Gerber, M. Johnston, N. D. Mueller, C. O'Connell, D. K. Ray, P. C. West, C. Balzer, E. M. Bennett, S. R. Carpenter, J. Hill, C. Monfreda, S. Polasky, J. Rockström, J. Sheehan, S. Siebert, D. Tilman, and D. P. M. Zaks, 2011 Solutions for a cultivated planet. Nature **478**: 337–342.

Garcia, A. a. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva, M. L. C. Vieira, R. Vicentini, E. a. Costa, M. C. Mancini, M. O. S. Garcia, M. M. Pastina, R. Gazaffi, E. R. F. Martins, N. Dahmer, D. a. Sforça, C. B. C. Silva, P. Bundock, R. J. Henry, G. M. Souza, M.-A. van Sluys, M. G. a. Landell, M. S. Carneiro, M. a. G. Vincentz, L. R. Pinto, R. Vencovsky, and A. P. Souza, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific Reports **3**: 3399.

Gelman, A., 2006 Prior distributions for variance parameters in hierarchical models. Bayesian Analysis **1**: 515–534.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2014 *Bayesian Data Analysis*.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su, 2008 A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics **2**: 1360–1383.

Goodfellow, I., Y. Bengio, and A. Courville, 2016 *Deep Learning*. MIT Press, `http://www.deeplearningbook.org`.

Goodwin, S., J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. Schatz, and W. R. McCombie, 2015 Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. bioRxiv p. 013490.

Hamelryck, T., 2012 *Bayesian Methods in Structural Bioinformatics*.

Hammer, G. L., Z. Dong, G. McLean, A. Doherty, C. Messina, J. Schussler, C. Zinselmeier, S. Paszkiewicz, and M. Cooper, 2009 Can Changes in Canopy and/or Root System Architecture Explain Historical Maize Yield Trends in the U.S. Corn Belt? Crop Science **49**: 299–312.

Han, B., X.-w. Chen, Z. Talebizadeh, and H. Xu, 2012 Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. BMC Systems Biology **6**: 1–12.

Heslot, N., J. L. Jannink, and M. E. Sorrels, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Science **55**: 1–12.

Hoffman, M. D. and A. Gelman, 2014 The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research **15**: 1593–1623.

Jaccoud, D., K. Peng, D. Feinstein, and A. Kilian, 2001 Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research **29**: 1–7.

Lawrence, C. J. and V. Walbot, 2007 Translational Genomics for Bioenergy Production from Fuelstock Grasses: Maize as the Model Species. The Plant Cell **19**: 2091–2094.

Loman, N. J., J. Quick, and J. T. Simpson, 2015 A complete bacterial genome assembled de novo using only nanopore sequencing data. bioRxiv **12**: 015552.

Mace, E. S., S. Tai, E. K. Gilding, Y. Li, P. J. Prentis, L. Bian, B. C. Campbell, W. Hu, D. J. Innes, X. Han, A. Cruickshank, C. Dai, C. Frère, H. Zhang, C. H. Hunt, X. Wang, T. Shatte, M. Wang, Z. Su, J. Li, X. Lin, I. D. Godwin, D. R. Jordan, and J. Wang, 2013 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nature Communications **4**: 1–9.

Mackay, T. F. C., E. a. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nature reviews. Genetics **10**: 565–77.

Moreau, Y., P. Antal, G. Fannes, and B. De Moor, 2003 Probabilistic graphical models for computational biomedicine. Methods of Information in Medicine **42**: 161–168.

Morris, G. P., P. Ramu, S. P. Deshpande, C. T. Hash, T. Shah, H. D. Upadhyaya, O. Riera-Lizarazu, P. J. Brown, C. B. Acharya, S. E. Mitchell, J. Harriman, J. C. Glaubitz, E. S. Buckler, and S. Kresovich, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences of the United States of America **110**: 453–458.

Mullet, J., D. Morishige, R. McCormick, S. Truong, J. Hilley, B. McKinley, R. Anderson, S. N. Olson, and W. Rooney, 2014 Energy sorghum–a genetic model for the design of C4 grass bioenergy crops. Journal of Experimental Botany **65**: 3479–89.

Murphy, K. P., 2013 *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

Neapolitan, R., D. Xue, and X. Jiang, 2013 Modeling the altered expression levels of genes on signaling pathways in tumors as causal bayesian networks. Cancer Informatics **13**: 77–84.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012 Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE **7**: e37135.

Ratcliffe, B., O. G. El-Dien, J. Klapste, I. Porth, C. Chen, B. Jaquish, and Y. A. El-Kassaby, 2015 A comparison of genomic selection models across time in interior spruce ( Picea engelmannii × glauca ) using unordered SNP imputation methods. Heredity **115**: 547–555.

Regassa, T. H. and C. S. Wortmann, 2014 Sweet sorghum as a bioenergy crop: Literature review. Biomass and Bioenergy **64**: 348–355.

Rooney, W. L., J. Blumenthal, and J. E. Mullet, 2007 Designing sorghum as a dedicated bioenergy feedstock. BIofuels, Bioproducts, Biorefining **1**: 147–157.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS ONE **7**: 1–13.

Shendure, J. and H. Ji, 2008 Next-generation DNA sequencing. Nature Biotechnology **26**: 1135–1145.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics **15**: 121–132.

Stan Development Team, 2018 RStan: the R interface to Stan. R package version 2.18.2.

Su, C., A. Andrew, M. R. Karagas, and M. E. Borsuk, 2013 Using Bayesian networks to discover relations between genes, environment, and disease. BioData Mining **6**: 1–21.

Team, S. D., 2018 *PyStan: the Python interface to Stan, Version 2.17.1.0.*.

Vermerris, W., 2011 Survey of Genomics Approaches to Improve Bioenergy Traits in Maize, Sorghum and Sugarcane. Journal of Integrative Plant Biology **53**: 105–119.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S. E. Mitchell, K. L. Roozeboom, D. Wang, M. L. Wang, G. A. Pederson, T. T. Tesso, P. S. Schnable, R. Bernardo, and J. Yu, 2016 Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants **2**: 1–7.

# 3 NOVEL BAYESIAN NETWORKS FOR GENOMIC PREDICTION OF DEVELOPMENTAL TRAITS IN BIOMASS SORGHUM

**Keywords:** Bioenergy; Sorghum; Genomic Prediction; Bayesian Networks, Indirect Selection, Probabilistic Programming.

## 3.1 Abstract

The ability to connect information between traits over time allow Bayesian networks to offer a powerful probabilistic framework to construct genomic prediction models. In this study, we phenotyped a diversity panel of 869 biomass sorghum lines, which had been genotyped with 100,435 SNP markers, for plant height (PH) with biweekly measurements from 30 to 120 days after planting (DAP) and for end-of-season dry biomass yield (DBY) in four environments. We developed and evaluated five genomic prediction models: Bayesian network (BN), Pleiotropic Bayesian network (PBN), Dynamic Bayesian network (DBN), multi-trait GBLUP (MTr-GBLUP), and multi-time GBLUP (MTi-GBLUP) models. In 5-fold cross-validation, prediction accuracies ranged from 0.48 (PBN) to 0.51 (MTr-GBLUP) for DBY and from 0.47 (DBN, DAP120) to 0.74 (MTi-GBLUP, DAP60) for PH. Forward-chaining cross-validation further improved prediction accuracies (36.4-52.4%) of the DBN, MTi-GBLUP and MTr-GBLUP models for PH (training slice: 30-45 DAP). Coincidence indices (target: biomass, secondary: PH) and a coincidence index based on lines (PH time series) showed that the ranking of lines by PH changed minimally after 45 DAP. These results suggest a two-level indirect selection method for PH at harvest (first-level target trait) and DBY (second-level target trait) could be conducted earlier in the season based on ranking of lines by PH at 45 DAP (secondary trait). With the advance of high-throughput phenotyping technologies, statistical approaches such as our proposed two-level indirect selection framework could be valuable for enhancing genetic gain per unit of time when selecting on developmental traits.

## 3.2 Introduction

The development of renewable energy resources from biomass crops is a key step towards the establishment of a sustainable agroecosystem (Foley *et al.*, 2011; Mace *et al.*, 2013). Among the plant species amenable to bioenergy production, sorghum [*Sorghum bicolor* (L.) Moench] is a prominent candidate for genetic improvement because it is diploid ($2n{=}2x{=}20$), fixes carbon through the $C_4$ pathway, predominantly autogamous, and resilient against biotic and abiotic stresses (Vermerris, 2011; Lawrence and Walbot, 2007). Although substantially more economic investments have been made in crops such as maize, which resulted in improvements of up to 4-fold in grain yield in the last century, proportional gains in yield might be achievable in biomass sorghum (Mullet *et al.*, 2014). Currently, sorghum has an average biomass yield of 12-15 dry Mg ha$^{-1}$ under rainfed conditions, with a predicted potential yield of 55-60 dry Mg ha$^{-1}$ for ideotypes (Mullet *et al.*, 2014). Additionally, because of its biology sorghum has the potential to become a model organism to understand the genetic basis of growth traits related to biomass production (Brenton *et al.*, 2016).

Genomic prediction (GP) is a statistical approach that predicts the unobserved phenotypes of individuals using genomic information (Meuwissen *et al.*, 2001). Because of its potential to enrich for promising selection candidates, GP is increasingly becoming an important component of plant breeding and genetic resources conservation programs. In sorghum, Yu *et al.* (2016) showed that GP can optimize the management and evaluation of accessions from gene banks through the prediction of different traits. Briefly, the GP procedure involves two steps: (i) phenotyping and genotyping a reference population (training set) to train statistical models, and (ii) genotyping of unevaluated individuals (test set) for predicting their unobserved phenotypes with the trained models (Heslot *et al.*, 2015). To support this procedure, plant breeding programs collect phenotypes from training set individuals evaluated in multi-environment trials. In parallel, high density single-nucleotide polymorphism (SNP) markers are scored on individuals in the training and test sets with skim sequencing or SNP arrays (Elshire *et al.*, 2011; Davey *et al.*, 2011; Buckler *et al.*, 2016).

Mixed linear models, hierarchical Bayesian models with informative priors, kernel methods, and neural nets are modeling approaches used for GP, but minimal differences in predictive performance are typically seen across these approaches (de Los Campos *et al.*, 2013; Heslot *et al.*, 2015; dos Santos *et al.*, 2016a). This outcome may be explained by the high density of covariates (SNP markers) compared to the population size used for training the models. This scenario is known as the large $p$ and small $n$ problem ($p >> n$) (Gianola *et al.*, 2009). In statistical models, this may lead to the problem of multicollinearity, i.e., multiple covariates with redundant information. As it relates to GP models, markers in complete or near complete linkage disequilibrium provide redundant information and will not contribute to enhancing statistical power (Gianola, 2013). Dimensionality reduction techniques, such as the artificial bins approach, may help circumvent the challenge of multicollinearity with minimal information loss, as well as mitigate the computational cost often associated with GP (Xu, 2013).

The vast majority of GP studies conducted in crop species have only tested models for predicting individual traits. However, recent studies have shown the advantages of combining multiple correlated traits in a GP model (Calus and Veerkamp, 2011; Jia and Jannink, 2012; Fernandes *et al.*, 2018), allowing genetic correlations among secondary traits to be leveraged for improving predictions of a target trait (dos Santos *et al.*, 2016b; Okeke *et al.*, 2017). Most of these efforts used multi-trait GBLUP - a type of multivariate mixed linear model that incorporates a genomic relationship matrix (Gianola *et al.*, 2015). Despite the advances obtained so far, the use of genetic models that exploit information between traits using other parametrizations beyond those reliant on genetic correlations under multivariate normal distribution assumptions have yet to be addressed. Indeed, novel genetic models with parametrizations to partition genetic effects influencing only a single trait from those acting on multiple traits (i.e., pleiotropy) may help to better understand the genetic architecture of correlated traits.

There have been significant advances in field-based high-throughput phenotyping (HTP) technologies for the rapid measurement of plant traits over the growing season (Bao *et al.*, 2019; Pauli *et al.*, 2016). Measuring phenotypes at multiple time points over the life cycle of a plant can better describe the progression of growth and development (Muraya *et al.*, 2017). Hav-

ing collected phenotypic information on a time axis may help to identify key environmental stress events during the growing season, which might be masked if phenotypic data are only obtained at harvest (Campbell *et al.*, 2018). Furthermore, the underlying genetic signals of these phenotypic responses are additional sources of information to more powerfully predict and dissect the genetic architecture of developmental plant traits (Muraya *et al.*, 2017; Campbell *et al.*, 2018). Statistical models that exploit temporal genetic trends are especially needed for longitudinal (repeated measure) data collected by field-based HTP systems. Such models could be used to reduce generation time and prioritize which breeding populations to evaluate.

Among the models available for analyzing traits in a time series, probabilistic graphical models (PGMs) offer a versatile, efficient, and intuitive approach for drawing inferences (Murphy, 2013; Bishop, 2013). Popular PGMs include directed graphical models or Bayesian networks (BNs), undirected graphical models or Markov random fields, chain models, and factor graphs (Hamelryck, 2012). In particular, BNs provide the flexibility to model repeated measure and correlated trait data, as would be important for the study of developmental traits. A BN is defined as a structured graphical representation of joint distributions factored into a set of conditional probability distributions, where shaded and unshaded nodes represent known and unknown variables, respectively, and arrows showing dependence between them (Bishop, 2013). The Markov condition is a key property of the BN, ensuring that a variable (child) is only dependent on the information of its parents in the network (Su *et al.*, 2013; Bishop, 2013). Through their ability to connect joint probability distributions, BNs enable the aggregation of advantages from multiple machine learning approaches under a directed acyclic graph structure. Notably, BNs have been diversely applied in genetic and genomic studies (Loman *et al.*, 2015; Serang *et al.*, 2012; Garcia *et al.*, 2013; Han *et al.*, 2012; Su *et al.*, 2013; Neapolitan *et al.*, 2013), but to our knowledge have never been used for modelling trends of genetic effects considering repeated measures and correlated traits.

There are several features of BNs that enable them to recover information from correlated data types such as multiple correlated traits scored at a single time point or the repeated measurement of a single trait across multiple time points (Bae *et al.*, 2016). Several different GP models could be unified for leveraging pleiotropy or temporal genetic effects in a single BN to improve prediction accuracies. This is because these genetic effects can be modeled with a BN through connections between likelihood functions. Also, BNs offer the possibility to use general Markov chain Monte Carlo (MCMC) methods to obtain solutions for complex time series and multiple trait models that otherwise would have been mathematically intractable to derive analytically. Furthermore, the posterior samples of genomic estimated breeding values (GEBVs) may be used to create indices for understanding the uncertainty of selecting promising lines either earlier in the season or through indirect selection based on the ranking of the lines at other measurement time points or with correlated traits.

With sorghum as a model biomass crop, we developed PGMs for the GP of developmental traits in a sorghum diversity panel of nearly 900 lines. Herein, we aimed to (i) develop PGMs for the GP of plant height (PH) and dry biomass yield (DBY) traits by connecting genetic effects across multiple developmental time points and traits, and (ii) describe growth dynamics based on the change of the ranking of lines across multiple time points and correlated traits to

design novel breeding strategies to genetically improve biomass sorghum.

## 3.3 Materials and Methods

### 3.3.1 Plant material, field experiments and phenotypic data

In this study, we evaluated a biomass sorghum diversity panel consisting of 869 lines (Valluru *et al.*, 2018). The diversity panel was grown at three field locations only a few km from the main campus of the University of Illinois Urbana-Champaign in 2016 (Fisher and Energy Farms) and 2017 (Maxwell and Energy Farms). Each of the four environments had one complete replication of the field experiment that contained 960 four-row plots laid out in an 40-row by 24-column arrangement. The experimental field design consisted of 16 incomplete blocks, and each block was augmented with a common set of six lines (four shared between years): Pacesetter, PI276801, PI148089, PI524948, NSL50748, and PI148084 in 2016 and Pacesetter, PI276801, PI148089, PI524948, PI525882, and PI660560 in 2017. Plots were 3 m in length, with a 1.5 m alley at the end of each plot. Plots had a spacing between rows of 0.76 m. The plant population had a targeted density of 270,368 plants ha$^{-1}$. Experiments were planted in late May and harvested in early October. PH was measured in centimeters (cm) from the soil line to the topmost leaf whorl. A single plant was measured in each plot on a biweekly basis from 30 to 120 days after planting (DAP). Plots were harvested for above ground biomass using a four-row Kemper head attached to a John Deere 5830 tractor. Wet weight of total biomass (lbs) and biomass moisture (%) in the center two rows of each 4-row plot were measured using a plot sampler that had a near infrared sensor (model 130S, RCI engineering). DBY in dry metric tons per hectare was calculated as follows: dry metric tons ha$^{-1}$ = total plot wet weight (kg) × (1-plot moisture) / (plot area in square meter/10,000).

### 3.3.2 Phenotypic data analysis

Phenotypic measurements for DBY (dry metric tons per ha$^{-1}$; one measurement at harvest) and PH (cm; one measurement on each of seven plant developmental stages) were analyzed individually with the following mixed linear model:

$$\boldsymbol{y} = \mathbf{1_n}\mu + \boldsymbol{X_1 g} + \boldsymbol{Z_1 b} + \boldsymbol{Z_2 e} + \boldsymbol{Z_3 ge} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\boldsymbol{y}$ $(n \times 1)$ represents the phenotypic vector with $n$ entries, $\mathbf{1_n}$ a unit vector, $\mu$ a scalar to map the population mean, $\boldsymbol{X_1}$ $(n \times q)$ the design matrix of the $q$ fixed genetic effects (number of lines), $\boldsymbol{Z_1}$ $(n \times l)$ the design matrix of the $l$ random block within environment effects, $\boldsymbol{Z_2}$ $(n \times s)$ the design matrix of the $s$ random environment (location x year combination) effects, $\boldsymbol{Z_3}$ $(n \times m)$ the design matrix of the $m$ random genotype-by-environment effects; and $\boldsymbol{g}$, $\boldsymbol{b}$, $\boldsymbol{e}$, and $\boldsymbol{ge}$ are column vectors mapping the design matrices effects, respectively, and $\boldsymbol{\epsilon}$ $(n \times 1)$ the vector of errors. The model random effects $\boldsymbol{b}$, $\boldsymbol{e}$, $\boldsymbol{ge}$, and $\boldsymbol{\epsilon}$ were assumed to follow a MVN$(0, \boldsymbol{I_l}\sigma_b^2)$, MVN$(0, \boldsymbol{I_s}\sigma_e^2)$, MVN$(0, \boldsymbol{I_m}\sigma_{ge}^2)$, and MVN$(0, \boldsymbol{I_n}\sigma_\epsilon^2)$, respectively.

Heritability on an line-mean basis was estimated for each phenotype. Variance component estimates were obtained by refitting model (1) with all terms as random effects in ASReml-R version 3.0 (Butler *et al.*, 2009). The variance component estimates from each model for a

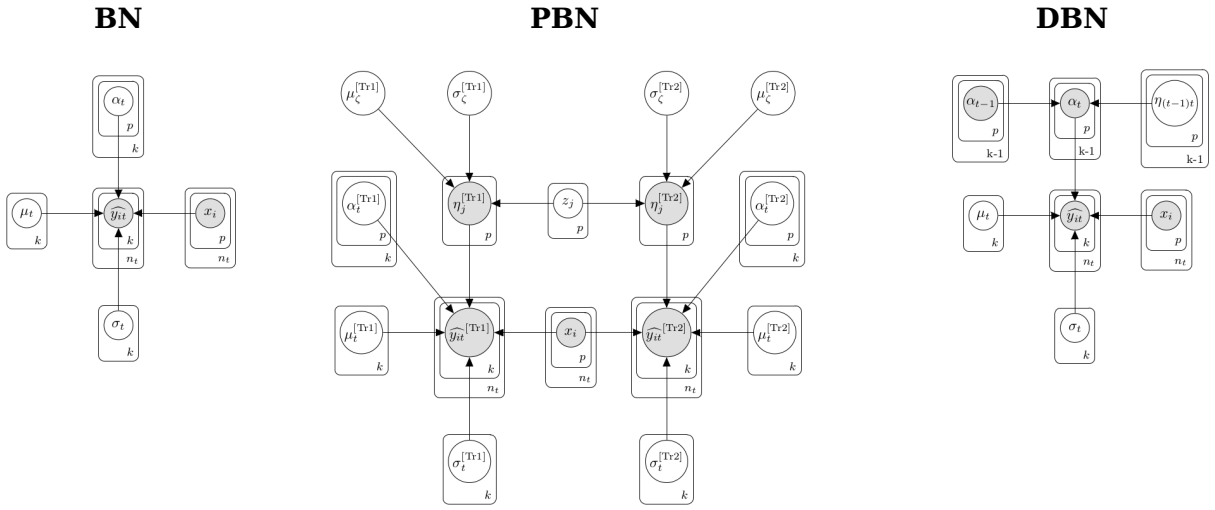


**Figure 3.1:** Bayesian network (BN), pleiotropic Bayesian network (PBN), and dynamic Bayesian network (DBN) probabilistic graphical models. $k$: number of time points; $n_t$: number of lines within a time point; $p$: number of artificial bins; $\widehat{y_{it}}$, $\widehat{y_{it}}^{[\text{Tr1}]}$, $\widehat{y_{it}}^{[\text{Tr2}]}$: Adjusted means for the i$^{\text{th}}$ line evaluated in the t$^{\text{th}}$ time point, which can be for trait 1 (Tr1) or trait 2 (Tr2); $\mu_t$, $\mu_t^{[\text{Tr1}]}$, $\mu_t^{[\text{Tr2}]}$: population means; $x_i$: row vector with artificial bins; $\alpha_{t-1}$, $\alpha_t$, $\alpha_t^{[\text{Tr1}]}$, $\alpha_t^{[\text{Tr2}]}$: column vector with artificial bins effects; $\eta_j^{[\text{Tr1}]}$, $\eta_j^{[\text{Tr2}]}$: pleiotropic j$^{\text{th}}$ bin effect; $z_j$: standardized pleiotropic j$^{\text{th}}$ bin effect; $\mu_\zeta^{[\text{Tr1}]}$, $\mu_\zeta^{[\text{Tr2}]}$: pleiotropic means hyperparameters; $\eta_{(t-1)t}$: bin effects between the current and previous time point; $\sigma_\zeta^{[\text{Tr1}]}$, $\sigma_\zeta^{[\text{Tr2}]}$: pleiotropic standard deviations hyperparameters; $\sigma_t$, $\sigma_t^{[\text{Tr1}]}$, $\sigma_t^{[\text{Tr2}]}$: standard deviations.

phenotype were used to estimate heritability on a line-mean basis as the ratio of genetic variance to phenotypic variance following (Holland *et al.*, 2003; Hung *et al.*, 2012). Standard errors of the heritability estimates were calculated with the delta method (Lynch *et al.*, 1998; Holland *et al.*, 2003) in the *nadiv* R package (Wolak, 2012). The Pearson's correlation coefficient ($r$) was used to assess the degree of relationship between adjusted means for each pair of traits.

### 3.3.3 Genotypic data

We genotyped the sorghum diversity panel using the genotyping-by-sequencing (GBS) procedure (Elshire *et al.*, 2011) based on the PstI-HF/HinP1I and PstI-HF/BfaI restriction enzymes. A total of 367 million sequence reads were generated (100 bp length) on a HiSeq 4000 sequencer. Sequence reads were aligned to the *Sorghum bicolor* genome v3.1 (`www.phytozome.jgi.doe.gov`) using Bowtie2 (Langmead and Salzberg, 2012). The TASSEL 3 GBS pipeline (Glaubitz *et al.*, 2014) was then used to call variants. Only biallelic SNPs were retained. Additionally, lines with >80% missing data (sample call rate) and SNPs with >60% missing data (SNP call rate) were removed. Also, SNPs with a minor allele frequency less than 5% were discarded. Missing genotypes of SNP markers were imputed using *Beagle 4.1* (Browning and Browning, 2016) with default parameters and an Ne of 150,000. In total, 100,435 SNP markers were scored and converted to dosage format (0,1,2). Of the 869 total lines, 839 had both phenotypic and genotypic data; therefore, the GP analyses focused on only these 839 lines.

### 3.3.4 Artificial bins

Due to the high dimensionality of the SNP marker matrix (100,435 loci), we developed a strategy similar to that proposed by Xu (2013) for obtaining artificial bins. In our approach, after centering (subtracting) the marker scores 2 (MM), 1 (Mm), and 0 (mm) by $2p$, instead of averaging the columns from equally sized slices of the marker matrix, we conducted a principal component analysis (PCA) to calculate the first PC of each artificial bin. The procedure was based on the following steps: (i) subdivision of the centered marker matrix into 1000 column-slices, with each slice comprising ∼100 columns; (ii) singular value decomposition of each matrix column slice into singular vectors and values; and (iii) construction of each artificial bin as the first PC coordinates of its respective matrix column slice, given by $\boldsymbol{u_1}\lambda_1$, where $\boldsymbol{u_1}$ is the first left singular vector of the matrix slice and $\lambda_1$ its first singular value. This procedure resulted in 1,000 artificial bins. This number of artificial bins was selected as a balance between model run time and predictive performance for the computationally intensive Bayesian models. In theory, the first PC (artificial bin) will retain as much information as possible from the matrix slice in one dimension in a least-square reconstruction error sense (Goodfellow *et al.*, 2016).

### 3.3.5 Probabilistic graphical models

We developed three different Bayesian models for genomic prediction of the PH and DBY traits (Figure 3.1). The first model, Bayesian network (BN), neither recovered information between traits nor time points. The BN has the following conditional normal likelihood form:

$$\widehat{y_{it}} \,|\, \text{all} \sim \mathcal{N}(\mu_t + x_i\,\alpha_t, \sigma_t)$$

where $\widehat{y_{it}}$ is the adjusted mean related to the i$^{\text{th}}$ line evaluated in the t$^{\text{th}}$ time point (or DAP), $\mu_t$ the unknown population mean in the t$^{\text{th}}$ time point, $x_i$ the known row vector $(1 \times p)$ of the $p$ artificial bins of the i$^{\text{th}}$ line, $\alpha_t$ the column vector $(p \times 1)$ of the unknown $p$ artificial bins effects within in the t$^{\text{th}}$ time point and $\sigma_t$ the unknown residual standard deviation mapping the uncertainty around the expected value in the t$^{\text{th}}$ time point.

The BN has an unnormalized joint posterior distribution (hyperpriors omitted from Figure 3.1 for simplicity),

$$\begin{aligned} p(\,\text{Bayesian network}|y,\phi\,) \propto \\ \prod_{t=1}^{k}\prod_{i=1}^{n_t} \mathcal{N}(\widehat{y_{it}}|\mu_t + x_i\alpha_t, \sigma_t)\ \mathcal{N}(\mu_t|0, s^{\{\mu_t\}})\ \mathcal{N}(\alpha_t|0, s^{\{\alpha_t\}}) \\ \text{Cauchy}^{+}(\sigma_t|0, s^{\{\sigma_t\}})\ \text{Cauchy}^{+}(s^{\{\mu_t\}}|0, \pi^{\{\mu_t\}})\ \text{Cauchy}^{+}(s^{\{\alpha_t\}}|0, \pi^{\{\alpha_t\}}) \\ \text{Cauchy}^{+}(s^{\{\sigma_t\}}|0, \pi^{\{\sigma_t\}})\ \text{Cauchy}^{+}(\pi^{\{\mu_t\}}|0, \phi)\ \text{Cauchy}^{+}(\pi^{\{\alpha_t\}}|0, \phi) \\ \text{Cauchy}^{+}(\pi^{\{\sigma_t\}}|0, \phi) \end{aligned}$$

where $k$ is the total number of time points, $n_t$ is the total number of lines in the t$^{\text{th}}$ time point, $\mathcal{N}(\theta|\mu^{\{\theta\}}, \sigma^{\{\theta\}})$ and $\text{Cauchy}^{+}(\theta|\mu^{\{\theta\}}, \sigma^{\{\theta\}})$ denotes the normal probability density function, and Cauchy probability density function truncated to the real positive space ($\mathbb{R}^{+}$) of the random variable $\theta$ (general notation), respectively, parametrized by the mean ($\mu_\theta$), standard deviation ($\sigma_\theta$). The joint distribution was parameterized as second ($s^{\{\theta\}}$) and third ($\pi^{\{\theta\}}$) level scale hyperparameters. The known global hyperparameter was defined by $\phi = ||y||_\infty \times 10 = \arg\max(y) \times 10$,

resulting in weakly informative second-level hyperpriors that eliminated the subjectiveness to define hyperparameters when choosing first-level prior hyperparameters (Gelman *et al.*, 2014). This same approach was used for the next set of described models.

The second Bayesian model, pleiotropic Bayesian network (PBN), exploited information between PH and DBY (Figure 3.1). This model has two conditionally dependent normal likelihood functions that characterized the observed adjusted means distribution for each trait as follows:

$$\widehat{y_{it}}^{[\mathrm{Tr1}]} \,|\, \mathrm{all} \sim \mathcal{N}(\mu_t^{[\mathrm{Tr1}]} + x_i\,(\alpha_t^{[\mathrm{Tr1}]} + \eta^{[\mathrm{Tr1}]}), \sigma_t^{[\mathrm{Tr1}]})$$
$$\widehat{y_{it}}^{[\mathrm{Tr2}]} \,|\, \mathrm{all} \sim \mathcal{N}(\mu_t^{[\mathrm{Tr2}]} + x_i\,(\alpha_t^{[\mathrm{Tr2}]} + \eta^{[\mathrm{Tr2}]}), \sigma_t^{[\mathrm{Tr2}]})$$

where all variables are the same from the previous model, except the column vectors $\eta^{[\mathrm{Tr1}]}$ $(p \times 1)$ and $\eta^{[\mathrm{Tr2}]}$ $(p \times 1)$, that represent the pleiotropic effects of known bins with continuous space corrected by the transformation of an unknown pleiotropic standardized random variable $z_j$ for the j$^{\mathrm{th}}$ bin,

$$\eta_j^{[\mathrm{Tr1}]} = \mu_\zeta^{[\mathrm{Tr1}]} + \sigma_\zeta^{[\mathrm{Tr1}]}\, z_j$$
$$\eta_j^{[\mathrm{Tr2}]} = \mu_\zeta^{[\mathrm{Tr2}]} + \sigma_\zeta^{[\mathrm{Tr2}]}\, z_j$$

with $\mu_\zeta^{[\mathrm{Tr1}]}$, $\sigma_\zeta^{[\mathrm{Tr1}]}$, $\mu_\zeta^{[\mathrm{Tr2}]}$, and $\sigma_\zeta^{[\mathrm{Tr2}]}$ being unknown random variables. The PBN model has an unnormalized joint posterior density function (Figure 3.1),

$$p(\,\text{Pleiotropic Bayesian network}|y,\phi\,) \propto$$
$$\prod_{t=1}^{k}\prod_{i=1}^{n_t} \mathcal{N}(\widehat{y_{it}}^{[\mathrm{Tr1}]}|\mu_t^{[\mathrm{Tr1}]} + x_i(\alpha_t^{[\mathrm{Tr1}]} + \eta^{[\mathrm{Tr1}]}), \sigma_t^{[\mathrm{Tr1}]})$$
$$\mathcal{N}(\mu_t^{[\mathrm{Tr1}]}|0, s^{\{\mu_t^{[\mathrm{Tr1}]}\}})\ \mathcal{N}(\alpha_t^{[\mathrm{Tr1}]}|0, s^{\{\alpha_t^{[\mathrm{Tr1}]}\}})\ \mathcal{N}(\sigma_t^{[\mathrm{Tr1}]}|0, s^{\{\sigma_t^{[\mathrm{Tr1}]}\}})$$
$$\mathrm{Cauchy}^+(s^{\{\mu_t^{[\mathrm{Tr1}]}\}}|0, \pi^{\{\mu_t^{[\mathrm{Tr1}]}\}})\ \mathrm{Cauchy}^+(s^{\{\alpha_t^{[\mathrm{Tr1}]}\}}|0, \pi^{\{\alpha_t^{[\mathrm{Tr1}]}\}})$$
$$\mathrm{Cauchy}^+(s^{\{\sigma_t^{[\mathrm{Tr1}]}\}}|0, \pi^{\{\sigma_t^{[\mathrm{Tr1}]}\}})\ \mathrm{Cauchy}^+(\pi^{\{\mu_t^{[\mathrm{Tr1}]}\}}|0, \phi)$$
$$\mathrm{Cauchy}^+(\pi^{\{\alpha_t^{[\mathrm{Tr1}]}\}}|0, \phi)\ \mathrm{Cauchy}^+(\pi^{\{\sigma_t^{[\mathrm{Tr1}]}\}}|0, \phi)$$
$$\mathcal{N}(z|0,1)\ \mathcal{N}(\mu_\zeta^{[\mathrm{Tr1}]}|0, s^{\{\mu_\zeta^{[\mathrm{Tr1}]}\}})\ \mathcal{N}(\sigma_\zeta^{[\mathrm{Tr1}]}|0, s^{\{\sigma_\zeta^{[\mathrm{Tr1}]}\}})$$
$$\mathcal{N}(\mu_\zeta^{[\mathrm{Tr2}]}|0, s^{\{\mu_\zeta^{[\mathrm{Tr2}]}\}})\ \mathcal{N}(\sigma_\zeta^{[\mathrm{Tr2}]}|0, s^{\{\sigma_\zeta^{[\mathrm{Tr2}]}\}})$$
$$\mathcal{N}(\widehat{y_{it}}^{[\mathrm{Tr2}]}|\mu_t^{[\mathrm{Tr2}]} + x_i(\alpha_t^{[\mathrm{Tr2}]} + \eta^{[\mathrm{Tr2}]}), \sigma_t^{[\mathrm{Tr2}]})$$
$$\mathcal{N}(\mu_t^{[\mathrm{Tr2}]}|0, s^{\{\mu_t^{[\mathrm{Tr2}]}\}})\ \mathcal{N}(\alpha_t^{[\mathrm{Tr2}]}|0, s^{\{\alpha_t^{[\mathrm{Tr2}]}\}})\ \mathcal{N}(\sigma_t^{[\mathrm{Tr2}]}|0, s^{\{\sigma_t^{[\mathrm{Tr2}]}\}})$$
$$\mathrm{Cauchy}^+(s^{\{\mu_t^{[\mathrm{Tr2}]}\}}|0, \pi^{\{\mu_t^{[\mathrm{Tr2}]}\}})\ \mathrm{Cauchy}^+(s^{\{\alpha_t^{[\mathrm{Tr2}]}\}}|0, \pi^{\{\alpha_t^{[\mathrm{Tr2}]}\}})$$
$$\mathrm{Cauchy}^+(s^{\{\sigma_t^{[\mathrm{Tr2}]}\}}|0, \pi^{\{\sigma_t^{[\mathrm{Tr2}]}\}})\ \mathrm{Cauchy}^+(\pi^{\{\mu_t^{[\mathrm{Tr2}]}\}}|0, \phi)$$
$$\mathrm{Cauchy}^+(\pi^{\{\alpha_t^{[\mathrm{Tr2}]}\}}|0, \phi)\ \mathrm{Cauchy}^+(\pi^{\{\sigma_t^{[\mathrm{Tr2}]}\}}|0, \phi)$$

The third Bayesian model, dynamic Bayesian network (DBN), recovered information from PH measurements across multiple time points (Figure 3.1). This network architecture has a specific conditionally, dependent normal likelihood function for each time point as follows:

$$\widehat{y_{it}} \,|\, \mathrm{all} \sim \mathcal{N}(\mu_t + x_i\,\alpha_t, \sigma_t)$$
$$\alpha_t = \alpha_{t-1} + \eta_{(t-1)t}$$

where the column vector $\alpha_t$ $(p \times 1)$ is the known artificial bins effects at time $t$, that are a linear combination of the $\alpha_{t-1}$ $(p \times 1)$ known artificial bins effects displayed in the previous time point $(t-1)$ plus the unknown $\eta_{(t-1)t}$ $(p \times 1)$ random noise mapping the bin effect between the current and previous time points, such that genetic information is propagated over time. The artificial bins effects were treated as unknown random variables only at the first time point. The DBN model has an unnormalized joint posterior distribution (Figure 3.1),

$$
\begin{aligned}
&p(\text{ Dynamic Bayesian network}|y, \phi\,) \propto \\
&\prod_{t=1}^{k}\prod_{i=1}^{n_t} \mathcal{N}(\widehat{y_{it}}|\mu_t + x_i\alpha_t, \sigma_t)\ \mathcal{N}(\mu_t|0, s^{\{\mu_t\}})\ \mathcal{N}(\alpha_0|0, s^{\{\alpha_0\}}) \\
&\mathcal{N}(\eta_{(t-1)t}|0, s^{\{\eta_{(t-1)t}\}})\ \text{Cauchy}^{+}(\sigma_t|0, s^{\{\sigma_t\}})\ \text{Cauchy}^{+}(s^{\{\mu_t\}}|0, \pi^{\{\mu_t\}}) \\
&\text{Cauchy}^{+}(s^{\{\alpha_0\}}|0, \pi^{\{\alpha_0\}})\ \text{Cauchy}^{+}(s^{\{\eta_{(t-1)t}\}}|0, \pi^{\{\eta_{(t-1)t}\}}) \\
&\text{Cauchy}^{+}(s^{\{\sigma_t\}}|0, \pi^{\{\sigma_t\}})\ \text{Cauchy}^{+}(\pi^{\{\mu_t\}}|0, \phi)\ \text{Cauchy}^{+}(\pi^{\{\alpha_0\}}|0, \phi) \\
&\text{Cauchy}^{+}(\pi^{\{\eta_{(t-1)t}\}}|0, \phi)\ \text{Cauchy}^{+}(\pi^{\{\sigma_t\}}|0, \phi)
\end{aligned}
$$

The joint distributions of the BN, PBN, and DBN models were integrated using the No-U-Term sampler algorithm available in the probabilistic programming language Stan (Hoffman and Gelman, 2014). We used the implementation available in the Python package *pystan 2.17.1.0* (Team, 2018). Stan compiles the probabilistic programming code in C++ and has a user interface within the Python environment. The probabilistic programming language saved time during customization of the C++ code, allowed rapid implementation during model design and training, and facilitated the manipulation of posterior draws after fitting the model in Python (Carpenter *et al.*, 2017). We set up the No-U-Term sampler to iterate 400 times and used as warm up 50% of the samples from four Markov chains. The number of iterations (400) was selected in consideration of runtime and predictive performance.

### 3.3.6 Multivariate GBLUP model

For comparison to the three Bayesian models, we also evaluated two different formulations of the multivariate GBLUP model (Henderson and Quaas, 1976; dos Santos *et al.*, 2016b; Fernandes *et al.*, 2018) that recovered information between traits and/or time points. In the first formulation, only PH measurements over time were used (MTi-GBLUP). In the second, PH measurements over time and DBY (MTr-GBLUP) were jointly analyzed. Both formulations share the same linear model as follows:

$$\widehat{y_{it}} = \beta_t + g_{it} + e_{it}$$

where $\widehat{y_{it}}$ corresponds to the adjusted mean related to the i$^{\text{th}}$ line evaluated for the t$^{\text{th}}$ time point or trait, $\beta_t$ is the fixed population mean effect for the t$^{\text{th}}$ time point and/or trait, $g_{it}$ is the genomic estimated breeding value (GEBV) of the i$^{\text{th}}$ line evaluated for the t$^{\text{th}}$ time point and/or trait, and $e_{it}$ is the residual.

Considering the structure of multivariate GBLUP models with stacked trait and/or time subvectors like $\boldsymbol{g} = [\boldsymbol{g_1}, \boldsymbol{g_2}, \ldots, \boldsymbol{g_k}]^T$ and $\boldsymbol{e} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_k}]^T$, in which $k$ is the number of traits and/or time points, we assume that $\boldsymbol{g} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{G} \otimes \boldsymbol{A})$ and $\boldsymbol{e} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{I_{n_t}} \otimes \boldsymbol{R})$, where $\boldsymbol{A}$ $(n_t \times n_t)$ is the additive relationship matrix (VanRaden, 2008) between $n_t$ lines, $\boldsymbol{G}$

($k \times k$) and $\boldsymbol{R}$ ($k \times k$) are unstructured variance-covariance matrices for genetic and residual effects, respectively. The $\boldsymbol{A}$ matrix was constructed using the 100,435 SNP markers with the A.mat function in the R package *rrBLUP 4.6* (Endelman, 2011). Spectral decomposition was performed to transform the $\boldsymbol{A}$ matrix into positive definite. The procedure is based on the singular value decomposition of the $\boldsymbol{A}$ matrix, substitution of the negative values by a decreasing small constant ($10^{-4}$), and reconstruction of the $\boldsymbol{A}$ matrix. Additional details on the spectral decomposition procedure are available in Caliński *et al.* (2005); dos Santos *et al.* (2016b). The MTi-GBLUP and MTr-GBLUP models were fitted using the R package *EMMREML 3.1* (Akdemir and Godfrey, 2018).

### 3.3.7 Cross-validation schemes

Two different cross-validation (CV) schemes were used to evaluate the predictive accuracy of the GP models. The first scheme used was stratified 5-fold CV for each individual trait (i.e., DBY or PH measured at a single time point). This procedure was based on stratifying the phenotypic and genotypic data of the lines into five non-overlapping folds, training the model with four folds (training set), and predicting the phenotypes of lines in the fold not included for training (test set) with only their genotypes as predictors in the trained model. This procedure was repeated until phenotypes from all five folds were predicted. Forward-chaining CV was used as a second scheme. In this scheme, data were split into time point subsets. The initial training set of five total was based on data from the first two time points (30 and 45 DAP), with the remaining time points (60, 75, 90, 105, and 120 DAP) comprising the test set. This procedure was repeated four more times to build new training sets until all time points except the last one (120 DAP) were included in the training set. The forward-chaining CV scheme was used to assess the accuracy of GP models to predict and identify the best set of lines for PH (tallest) prior to harvest.

The correlation (Pearson's r) of adjusted means with predicted values was used to estimate predictive accuracy in both CV schemes. In the forward-chaining CV scheme, the predicted values were always obtained with the artificial bins effects from the previous time point used for training the DBN model. For the MTi-GBLUP and MTr-GBLUP models, the predicted values from the previous time point used for training were used as predicted values. For the BN and PBN models, which do not share information across time, the effects of artificial bins for each time point were used to compute the predicted values.

### 3.3.8 Coincidence index

For the five GP models, coincidence indices (CIs) were constructed to evaluate the capacity for selecting the top 20% best performing lines for DBY when considering the rank of the PH adjusted mean values at each time point. The posterior values of the CI were calculated as the rate of successes between the top 20% best lines for PH at each time point and DBY. The CI was computed by assigning a '1' to lines in the top 20% best lines for PH and DBY, or '0' otherwise, then dividing the total number of successes (sum of '1's) by the total number of lines at each posterior sample.

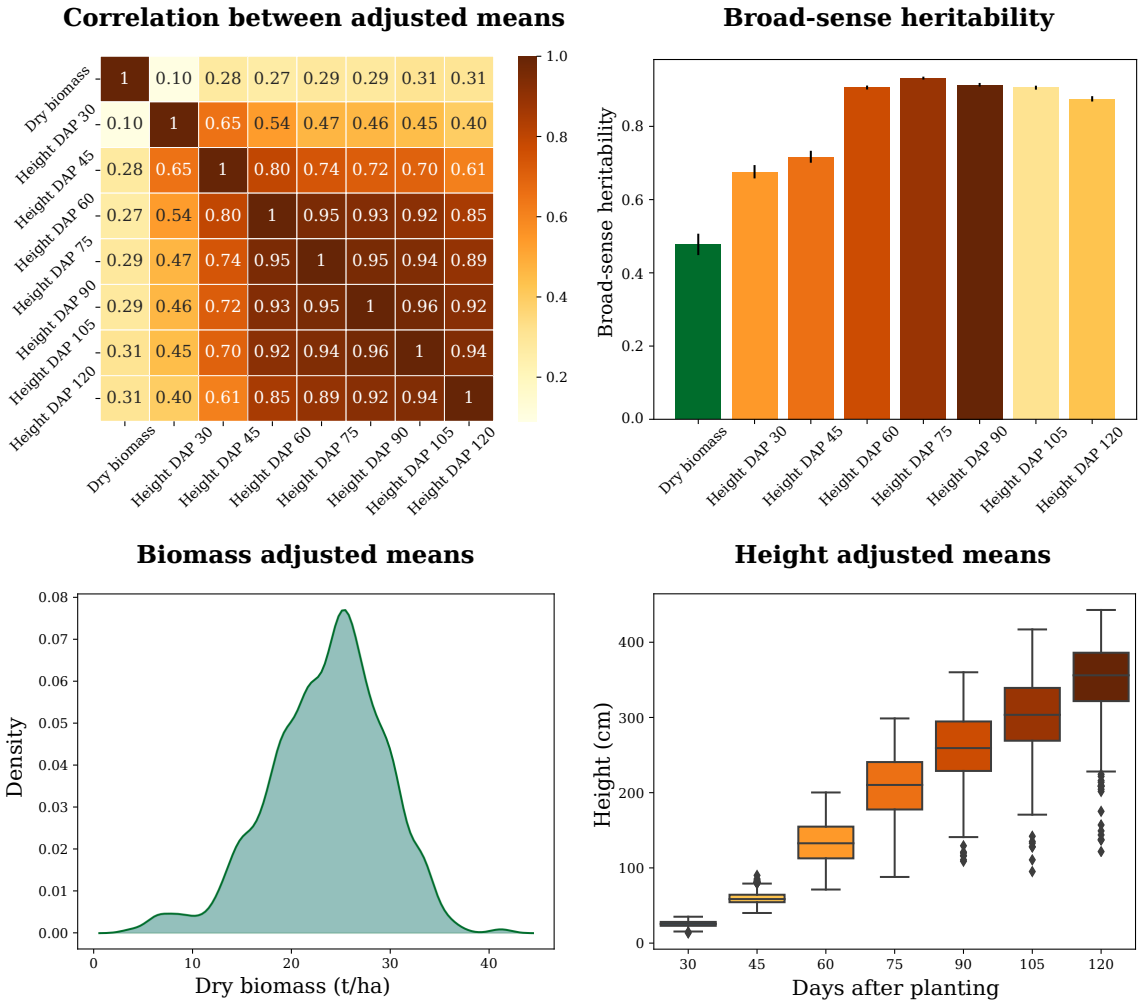


**Figure 3.2:** Correlation among adjusted means, heritabilities, and distribution of adjusted means for all traits.

### 3.3.9 Coincidence index based on lines

For the DBN model, we constructed a coincidence index based on lines (CIL) that used posterior samples from the adjusted mean values of PH across the seven measurement time points. This CIL was used to determine how early selection could be performed within-season to optimally reduce the length of the breeding cycle. Calculation of the CIL was based on the following steps: (i) identify the top 20% best lines for each posterior sample of the PH adjusted mean values at each time point; (ii) create for each posterior sample a one-hot vector encoding, assigning a '1' to the best lines in the top 20% in the evaluated time point and at the end of the season (120 DAP), or '0' otherwise; and (iii) compute the division of total successes by the total number of posterior samples that each line appeared in the top 20% for the predicted (evaluated time points) and observed adjusted mean values (end of season).

## 3.4 Results

### 3.4.1 Phenotypic variation

We used a mixed linear model that accounted for the influence of environment and genotype-by-environment interaction to generate adjusted means for end-of-season DBY and PH measured at seven developmental time points over the growing season. Both DBY and the multiple PH measures had moderately high estimates of heritability on a line-mean basis ($h^2 > 0.48$) across the four environments (Figure 3.2). The distribution of adjusted mean values for DBY was slightly skewed towards the left tail, with values centered on 23.55 tons ha$^{-1}$ (std=5.79). In comparison, the adjusted mean values for PH showed an expected growth pattern across the seven time points, with the population mean for PH changing from 25.6 (std=3.7) at 30 DAP to 350.7 cm (std=48.0) at 120 DAP. Indicative of an autoregressive trend of correlation over time, the weakest correlation was observed between PH measures collected at 30 and 120 DAP ($r$=0.40), while 90 and 105 DAP were the two most strongly correlated time points ($r$=0.96). Correlations between DBY and PH varied from 0.10 to 0.31 across time points, suggesting an opportunity for recovering information across time points and/or between traits to improve the predictive accuracy of GP models.

### 3.4.2 Predictive accuracies from stratified 5-fold CV

**Table 3.1: Prediction accuracies obtained from the 5-fold cross-validation scheme by training the Bayesian network (BN), pleiotropic Bayesian network (PBN), dynamic Bayesian network (DBN), multi time GBLUP (MTi-GBLUP) and multi trait GBLUP (MTr-GBLUP) models with dry biomass yield (DBY) collected at harvest and plant height (PH) measured across different days after planting (DAP). The standard deviations of the prediction accuracies obtained by the Bayesian models are represented within parenthesis.**

| Trait | Accuracy of the Genomic Prediction Models | | | | |
|---|---|---|---|---|---|
| | **BN** | **PBN** | **DBN** | **MTi-GBLUP** | **MTr-GBLUP** |
| **DBY** | 0.49 (0.021) | 0.48 (0.009) | - | - | 0.51 |
| **PH-30** | 0.53 (0.021) | 0.52 (0.020) | 0.47 (0.021) | 0.56 | 0.57 |
| **PH-45** | 0.59 (0.018) | 0.59 (0.017) | 0.57 (0.016) | 0.62 | 0.62 |
| **PH-60** | 0.72 (0.013) | 0.72 (0.014) | 0.51 (0.016) | 0.74 | 0.74 |
| **PH-75** | 0.70 (0.015) | 0.69 (0.015) | 0.53 (0.013) | 0.72 | 0.72 |
| **PH-90** | 0.67 (0.016) | 0.67 (0.016) | 0.51 (0.013) | 0.70 | 0.70 |
| **PH-105** | 0.67 (0.016) | 0.66 (0.017) | 0.52 (0.013) | 0.70 | 0.69 |
| **PH-120** | 0.61 (0.019) | 0.60 (0.019) | 0.47 (0.015) | 0.65 | 0.65 |

We evaluated the accuracy of the BN, PBN, DBN, MTr-GBLUP, and MTi-GBLUP models for predicting DBY and PH measured throughout the growing season with a stratified 5-fold CV scheme. Prediction accuracies (0.48-0.51) of DBY were nearly identical for the BN, PBN, and MTr-GBLUP models (Table 3.1). When predicting PH at each of the seven developmental stages with the BN, PBN, MTi-GBLUP and MTr-GBLUP models, we found that accuracies gradually increased from 30 to 60 DAP, peaked at 60 DAP, and incrementally decreased from 60

to 120 DAP. Of these four models, MTi-GBLUP (0.56-0.74) and MTr-GBLUP (0.57-0.74) had prediction accuracies comparable to each other and slightly higher than those of BN (0.53-0.72) and PBN (0.52-0.72). Comparatively, the DBN model showed a randomly fluctuating trend of relatively slightly lower prediction accuracies for PH at all seven time points. The minor variation in predictive accuracy, especially after 45 DAP, suggests that early season prediction could be possible for PH.

### 3.4.3 Predictive accuracies from forward-chaining cross-validation

We performed a forward-chaining CV procedure to evaluate the accuracy of the five models to predict PH at unobserved time points. In general, the models showed high accuracy to predict the phenotypic values of the lines observed at the last time point (120 DAP) even when trained only with data from both 30 and 45 DAP (Figure 3.3). The BN and PBN models had similar prediction accuracies across all scenarios, ranging from 0.42 (BN, training: 45 DAP; predicting: 120 DAP) to 0.86 (PBN, training: 60 DAP; predicting: 75 DAP). In contrast, the DBN, MTi-GBLUP, and MTr-GBLUP models had substantially higher predictive accuracies compared to the BN and PBN models. The DBN model showed predictive accuracies varying from 0.6 (training slice: 30-45 DAP, predicting: 120 DAP) to 0.95 (training slice: 30-60 DAP; predicting: 75 DAP). The MTi-GBLUP prediction accuracies ranged from 0.63 (training slice: 30-45 DAP; predicting: 120 DAP) to 0.94 (training slice: 30-90 DAP; predicting: 105 DAP), and comparably, the MTr-GBLUP varied from 0.64 (training slice: 30-45 DAP; predicting: 120 DAP) to 0.94 (training slice: 30-90 DAP; predicting: 105 DAP). These results did not suggest any advantage for modelling dependence between PH and DBY in the PBN and MTi-GBLUP models; however, the results did suggest that the dependence between time points accounted for in the DBN, MTi-GBLUP and MTr-GBLUP models improved the prediction accuracy of PH.

### 3.4.4 Coincidence indexes

The high predictive performance of the DBN, MTri-GBLUP, and MTr-GBLUP models, which exploited multiple PH measurements over initial growth stages, incentivized us to investigate how the rank of the lines varied across the different time points. To that end, we evaluated how PH measures over time (secondary traits) could be informative for performing indirect selection of DBY (target trait) through the calculation of coincidence indices (CIs). The posterior distribution of the CIs showed an overlapping pattern over time for the three Bayesian models (Figure 3.4), with most of them ranging from 0.18 to 0.34. This implied that the ranking of the lines for PH did not change significantly from early to late growth stages. Additionally, the MTi-GBLUP CI (target: DBY; secondary: PH) ranged from 0.25 (training slice: 30-105 DAP) to 0.27 (training slice: 30-45 DAP), and MTr-GBLUP varied from 0.26 (training slice: 30-105 DAP) to 0.28 (training slice: 30-45 DAP). These results suggest that the initial growth stage ranging from 30 to 45 DAP could be an optimal stage of development for the early selection of PH and indirect selection of DBY based on the ordering of the PH adjusted means over the growing season.

To gain more insight and empirical evidence to support the hypothesis of early selection for PH with measures from 30 and 45 DAP, we developed a coincidence index based on lines

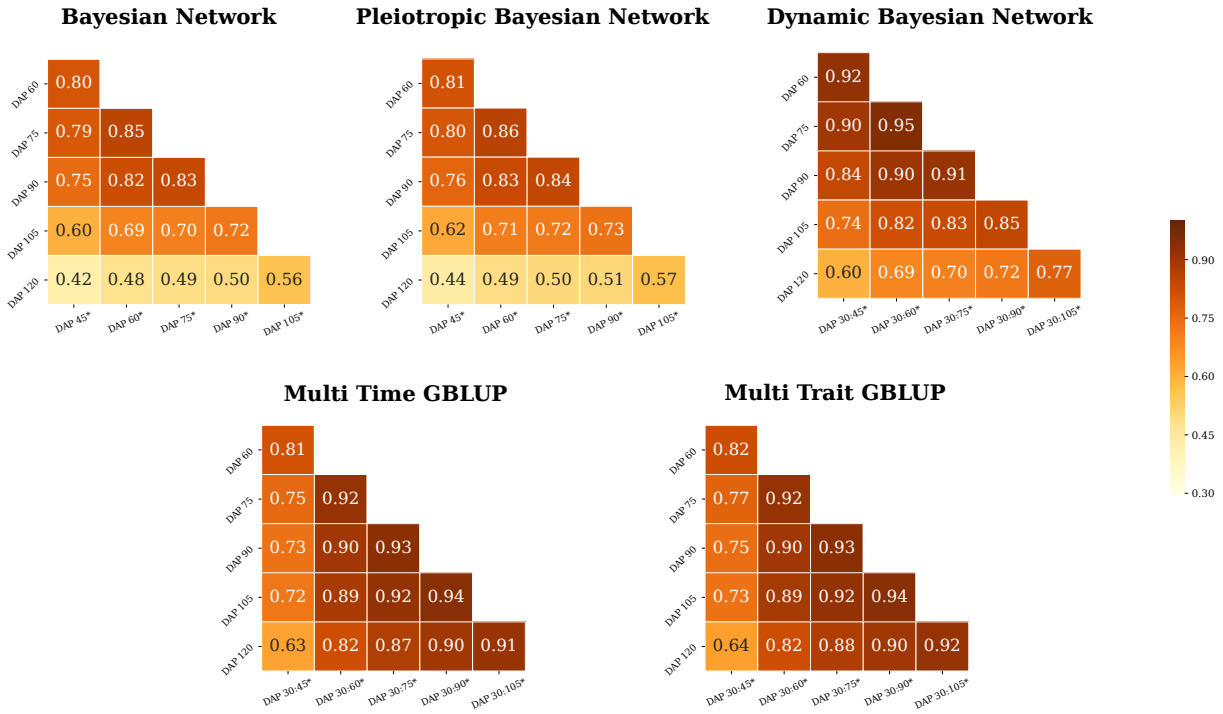


**Figure 3.3:** Prediction accuracies obtained by the forward-chaining cross-validation to evaluate genomic prediction models exploiting single (Bayesian Network and Pleiotropic Bayesian Network) or multiple time points (Dynamic Bayesian Network, Multi Time GBLUP, and Multi Trait GBLUP). The horizontal axis represents the slice (:) of the time interval used for training the models with multiple time points and the vertical axis the testing data. The '*' symbol tags the days after planting (DAP) time point used to obtain the adjusted means.

(CIL) using the posterior values from the DBN model that achieved optimal performance among the Bayesian models tested in the forward-chaining CV. The CIL allows us to better understand phenotypic plasticity through assessing how the expected rank of the lines at the end of the season agrees with their ranking at earlier growth stages. The closer that the CIL is to one, the more likely the line is expected to be at the top 20% for PH at the end of the season. We plotted lines with CIL $> 0.5$, fixed their ordering (training slice: 30-45 DAP), and displayed the CILs from other time slices in the same order (Figure 3.5). The CILs showed the expected trend of increasing the chance of lines to be in the top 20% after 45 DAP, which indicated that the ranking of the top lines had not majorly changed over time.

## 3.5 Discussion

Biomass sorghum is a promising bioenergy feedstock because of its extensive genetic diversity, high biomass yield potential, and strong tolerance to environmental stress. Sorghum is evolutionarily related to key bioenergy grasses, including maize, sugarcane, switchgrass, and Miscanthus spp., making it a potentially important diploid model to inform the genetic improvement of these other bioenergy crops (Morris *et al.*, 2013; Brenton *et al.*, 2016). Despite sorghum's appealing features as both a crop and model species, few studies have focused on genetically modelling its growth patterns and leveraging this information for breeding optimization. In this study, we investigated a diverse panel of 839 sorghum lines genotyped with 100,435

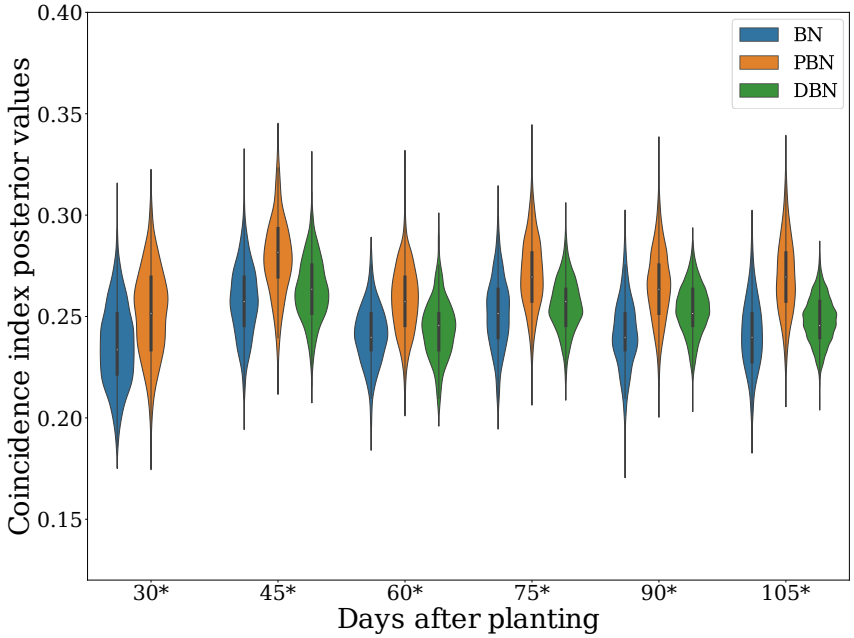


**Figure 3.4:** Coincidence index for selecting the top 20% for dry biomass yield using as reference the adjusted mean values obtained by training the Bayesian network (BN), pleiotropic Bayesian network (PBN), and dynamic Bayesian network (DBN) models with plant height data across the seven developmental time points. The '*' symbol denotes the time point estimates used to obtain the adjusted means as expected values for indirect selection. For the DBN model that leveraged multiple time points, the symbol * denotes the last time point used for training with the earlier time points also considered in the model.

SNP markers that was evaluated for a PH time series and DBY in four environments. With these collected data, we evaluated several GP models for exploiting genetic information over time and/or between correlated traits to improve prediction accuracies compared to models that assumed independence. Our implemented Bayesian models allowed us to estimate the level of uncertainty in determining optimal time points when developing breeding strategies for early selection of PH within season, as well as the indirect selection of DBY in combination with the repeated measures of PH as secondary traits.

To conduct GP of DBY and PH, we used both PGM and multivariate mixed linear model approaches to better model growth dynamics (Bishop, 2013; Henderson and Quaas, 1976; dos Santos *et al.*, 2016b). Due to the high computational cost of the PGM approach, we modified the artificial bins method of Xu (2013) to reduce the dimensionality of the SNP marker matrix through a PCA. This modified approach reduced the number of parameters needed to train the different PGM architectures by a 100-fold. Also, this procedure in other scenarios has the flexibility for predictions even when the number of loci pooled is different between the training and testing sets. Indicative of minimal information loss, there were negligible differences in prediction accuracies achieved by the MTi-GBLUP and MTr-GBLUP models that used the 100,000 SNP markers to compute the relationship matrix compared to those of the BN and PBN models reliant on the 1,000 artificial bins. Also, the artificial bins approach did

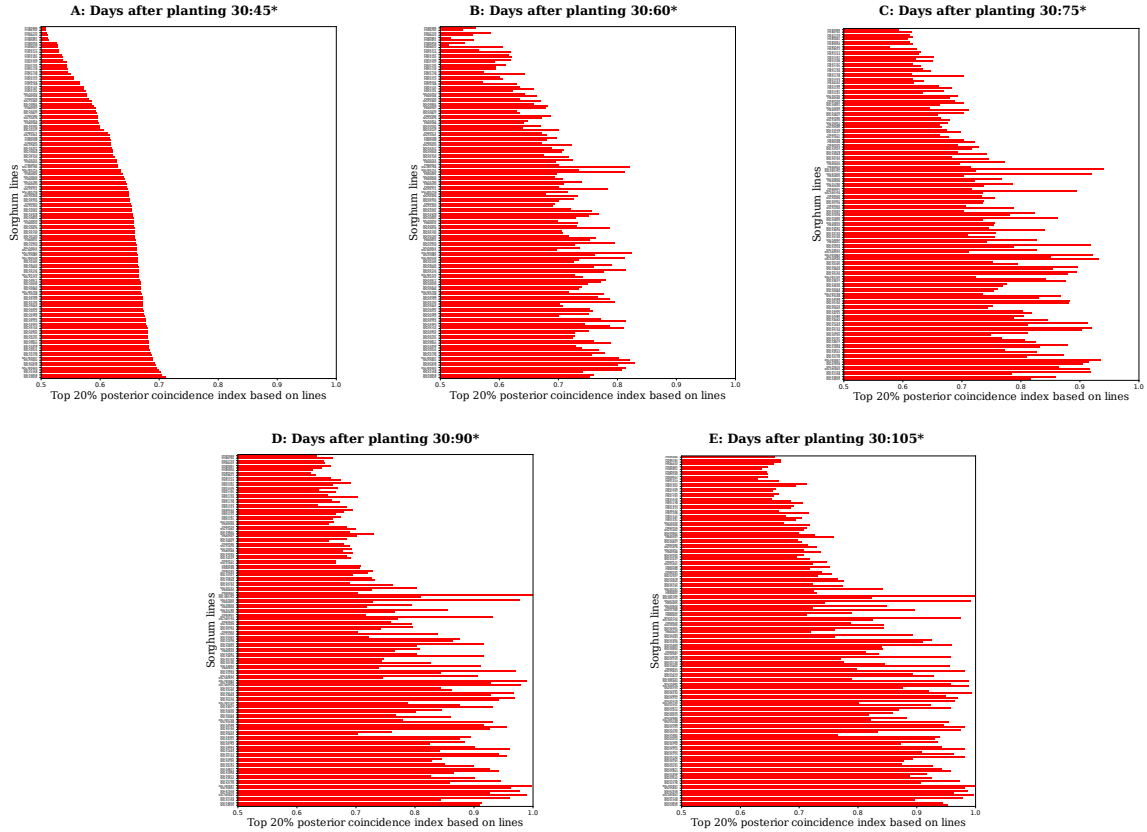


**Figure 3.5:** Top 20% posterior coincidence index based on lines (CIL) from the results of the dynamic Bayesian network. The rank order of the lines in subplot A was fixed for subplots B, C, D and E to understand phenotypic plasticity over time. Only lines with CIL>0.5 were plotted. The '*' symbol tags the days after planting (DAP) time point used to obtain the adjusted means.

not compromise the results from the DBN model, as indicated by the similarity of its obtained prediction accuracies with those from the multivariate GBLUP models in the forward-chaining CV scheme.

We initiated a model-based machine learning approach for GP analysis by first defining the baseline of PGMs (Bishop, 2013). Several studies have shown a similar level of predictive performance between PGMs that assume either a common or specific normal prior for each marker effect (de Los Campos *et al.*, 2013; Heslot *et al.*, 2015; Ferrão *et al.*, 2018). Therefore, we parsimoniously used a common normal prior for the effects of all artificial bins, resulting in a BN model that had less unknown parameters but a slower MCMC process to obtain posterior draws. The BN model can be considered a non-conjugate form of the Bayesian linear regression model (de Los Campos *et al.*, 2013) that automatically learns the hyperparameters of priors from the data. This model also has Cauchy priors truncated to the $\mathbb{R}^+$ on the scale components, which avoids sampling implausible standard deviation values (Gelman *et al.*, 2014). Although the main disadvantage of the BN model formulation is that its architecture cannot recover information between traits or time points, the BN model can be useful analyzing highly unbalanced data as is frequently observed for large-scale field trials in the plant breeding industry.

To improve the performance of the BN model, we novelly developed the PBN model

by connecting two representations of the BN model with hidden variables (nodes in the graph representing artificial bins effects influencing both traits), allowing a conditional relationship between the likelihood functions of PH and DBY to be established. Despite our efforts to estimate pleiotropic effects with the PBN model, the interpretation of the effects as pleiotropic should be carefully interpreted, especially due to the challenge of differentiating between pleiotropy and tight linkage. Gianola *et al.* (2015) theorized that the linkage disequilbrium (LD) between markers and quantitative trait loci (QTL), LD between QTL controlling different traits, or LD between markers linked to QTL controlling different traits can make it difficult to partition pleiotropic effects in GP models. The presence of such complex LD structure could explain the minor difference in predictive accuracy between the BN and PBN models when attempting to partition genetics effects influencing single versus multiple traits. This interpretation is supported by the finding that prediction accuracies obtained by the MTi-GBLUP (PH time series) and MTr-GBLUP (DBY and PH time series) models were similar to each other. Additionally, the lower than expected performance of the PBN model relative to the BN model could be attributed to the modest correlations observed between DBY and the multiple PH traits (r = 0.10 - 0.31). Simulation studies have shown that genetic correlations weaker than 0.5 do not provide marked improvements to the prediction accuracy of GP models used for multiple traits (Calus and Veerkamp, 2011; dos Santos *et al.*, 2016b). Further studies analyzing traits having strong genetic correlations with DBY are needed to better understand how to further improve the prediction accuracy of the PBN model.

The DBN model, a variant of hidden Markov models that is intended for modeling continuous variables (Murphy, 2013), was used to exploit the effects of artificial bins from the prediction of PH at earlier time points. This connection of genetic information between time points was crucial for dramatically improving the performance of the Bayesian framework in the forward-chaining CV scheme. Moreover, the predictions using the posterior mean of the DBN model allowed us to obtain predictions as precise as the point estimates from the MTi-GBLUP model and construct indices with posterior samples to identify optimal time points for indirect selection. The strong correlation between multiple time points for PH is quite possibly the main factor that favored the improvement of prediction accuracy for the DBN model compared to the BN and PBN models that did not leverage information over time. In contrast to the high predictive performance achieved in the forward-chaining CV scheme, relatively lower predictive accuracies were observed for the DBN model in the 5-fold CV scheme, especially when removing lines across all time points to split the data into training and testing sets. This is because the splitting did not allow the DBN model to learn with precision the effects of the artificial bins between time points and added noise to the artificial bin estimates. Despite the sharing of genetic signals over time, these findings suggest that the DBN model should not be used when lines are completely unobserved across all time points (Burgueño *et al.*, 2012; Dias *et al.*, 2018; Fernandes *et al.*, 2018). This reduced level of prediction performance because of unobserved lines has also been previously reported for multivariate GP models tested with a 5-fold CV scheme (Burgueño *et al.*, 2012; Dias *et al.*, 2018; Fernandes *et al.*, 2018).

Even though modeling growth dynamics is an important area of research, there have been a limited number of studies in plants that have analyzed longitudinal phenotypes related

to growth rate with GP models. In a greenhouse study of 357 diverse rice (*Oryza sativa* L.) accessions, Campbell *et al.* (2018) developed GP models with a first- or second-order Legendre polynomial to predict the sum of "plant pixels" from image-based phenotyping as a daily estimate of shoot biomass during initial growth stages (13 to 33 days after transplant), resulting in an 11.6% improvement in prediction accuracies relative to a single time point analysis. Despite the advantage of fitting a nonlinear random regression model with phenotypic data that showed an exponential curve over 20 days of early vegetative growth, this procedure is limited to only early developmental stage phenotypes such as shoot biomass that follow an exponential growth curve. In our study, PH was measured throughout the entire growing season without a focus on the early vegetative stage, thus the collected PH data did not have an exponential shape. The modelling of genetic effects as either a linear additive function over time or though the $G$ var-cov matrix might be the main factor causing the 36.4-52.4% (training slice: 30-45 DAP; predicting: 120 DAP) improvements in prediction accuracy of the DBN, MTi-GBLUP, and MTr-GBLUP models. When analyzing repeated measures of height collected from an interior spruce (Picea engelmannii $\times$ glauca) population of 769 trees at six sparse time points over a period of 37 years, Ratcliffe *et al.* (2015) observed minor differences in prediction performance among the evaluated BC$\pi$, rrBLUP, and generalized ridge regression models. These findings are contrary to our evaluated PGMs that recovered genetic information between time points that showed substantial improvement compared to the BN model - a Bayesian formulation of the rrBLUP model.

The implemented PGMs provided a powerful modelling framework to infer uncertainty based on well-established probability theory (Murphy, 2013; Bishop, 2013), allowing us to define optimal time points for earlier selection of a target trait within season or indirect selection through secondary correlated traits. To this end, we used the posterior distribution of the CIs and observed that the ordering of the lines changed minimally after 45 DAP, which suggested an opportunity to indirectly select for DBY based on early season PH measures as secondary traits. The Bayesian CIs allowed the direct assessment of overlapping time points with credible intervals for identying optimal time points contrary to the point estimate approach of Hamblin and Zimmermann (1986) applied to genomic prediction of a PH time series and DBY in sorghum by (Fernandes *et al.*, 2018). The Bayesian CIs also did not require resampling to build the index. In addition, we used the DBN model to evaluate the phenotypic plasticity of PH with the CILs at the population level. The pattern of phenotypic plasticity shown by the CILs confirmed the general findings of the CIs–the ranking of lines by PH changed minimally from early (45 DAP) to late (120 DAP) measures. Early within-season selection for PH could help to efficiently accelerate the breeding process. Considering the breeder's equation $\Delta \mathrm{G} = \mathrm{i}\,\mathrm{h}\,\sigma_g/\mathrm{L}$ (Li *et al.*, 2018), reducing the length of the generation (L) for given a selection intensity (i) could provide a new avenue for increasing genetic gain per unit of time through early indirect selection of PH and DBY at 45 DAP in these tested environments.

Given that the order of lines ranked by PH minimally changed across the measured plant developmental stages and had a moderate coincidence with rankings based on DBY, we propose a two-level indirect selection framework: (i) fit the DBN model–Bayesian model with the best performance for predicting future measures–to learn posterior values of the GEBVs in initial

growth stages for PH repeated measures and obtain a precise estimate of the ranking of each line at the end of the season; (ii) compute CI and CIL to evaluate the extent to which the rank order of lines change; (iii) use GEBVs of the previous time point as a secondary trait; (iv) perform indirect selection for PH at the end of the season as first-level target trait; and (v) perform indirect selection for DBY as the second-level target trait. This selection approach together with the trait-assisted genomic selection approach proposed by FERNANDES *et al.* (2018) may allow the end of season rank order of observed and eventually unobserved lines to be accurately predicted early in the growing season. When combined with high-throughput phenotyping platforms, the two-level indirect selection framework has the potential to further accelerate selection cycles and support a larger number of evaluated families. This could be accomplished by deploying low-cost ground rovers (e.g., Earthsense, `https://www.earthsense.co`) for repeated measurements of height and other traits genetically correlated with DBY on field-grown plants in combination with off-season winter nurseries and greenhouses with automated phenotyping systems (e.g., Lemnatec, http://www.lemnatec.com; Photon Systems Instruments, `http://www.psi.cz`) and optimized LED lights for speed breeding (WATSON *et al.*, 2018).

## 3.6 Conclusion

We analyzed phenotypic measures over time for PH and DBY at the end of the season to design a novel indirect selection scheme. To that end, we developed and evaluated novel Bayesian networks for GP that were used to better model and understand phenotypic plasticity of PH across different plant developmental stages. The GP models showed minor differences in prediction accuracies for the 5-fold CV scheme. In stark contrast, in the forward-chaining CV scheme, we observed a 36.4-52.4% improvement in prediction accuracy of the DBN and multivariate GBLUP models (train on 45 DAP, predict 120 DAP) compared to the BN and PBN models that assumed independence over time. The Bayesian models were used to show that the ranking of lines changed minimally after 45 DAP based on the CI and CIL, serving as novel approaches to understand ranking dynamics with repeated measures. These results suggest that in these environments 45 DAP is an optimal developmental stage for imposing a two-level indirect selection framework for biomass sorghum. Such that indirect selection for end of season PH (first-level target trait) and DBY (second-level target trait) could be performed based on the ranking of lines by PH at 45 DAP (secondary trait).

## Acknowledgements

This work was supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) Grant 2017/03625-2 and 2017/25674-5. CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Finance Code 001. CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000598. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

## Author Contributions

J. P. R. dos Santos and M. A. Gore co-wrote the manuscript; J. P. R. dos Santos led data analysis; P. J. Brown and S. B. Fernandes designed and managed field trials; P. J. Brown collected sequencing and phenotyping data; S. B. Fernandes collected phenotypic data; R. Lozano analyzed sequencing data; J. P. R. dos Santos developed genomic prediction models, codes, and statistical analysis methods; E. S. Buckler and A. A. F. Garcia. consulted on data analysis and provided support in the development of genomic prediction models and project coordination; M. A. Gore overall project management, design, coordination, oversaw data analysis; all authors contributed to the critical review of the manuscript.

## References


Akdemir, D. and O. U. Godfrey, 2018 *EMMREML: Fitting Mixed Models with Known Covariance Structures (Version 3.1).*

Bae, H., S. Monti, M. Montano, M. H. Steinberg, T. T. Perls, and P. Sebastiani, 2016 Learning Bayesian Networks from Correlated Data. Scientific reports **6**: 1–14.

Bao, Y., L. Tang, M. W. Breitzman, M. G. Salas Fernandez, and P. S. Schnable, 2019 Field-based robotic phenotyping of sorghum plant architecture using stereo vision. Journal of Field Robotics **36**: 397–415.

Bishop, C. M., 2013 Model-based machine learning. Phil Trans R Soc A **371**: 1–17.

Brenton, Z. W., E. A. Cooper, M. T. Myers, R. E. Boyles, N. Shakoor, K. J. Zielinski, B. L. Rauh, W. C. Bridges, G. P. Morris, and S. Kresovich, 2016 A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. Genetics **204**: 21–33.

Browning, B. and S. Browning, 2016 Genotype imputation with millions of reference samples. The American Journal of Human Genetics **98**: 116 – 126.

Buckler, E. S., D. C. Ilut, X. Wang, T. Kretzschmar, M. A. Gore, and S. E. Mitchell, 2016 rampseq: Using repetitive sequences for robust genotyping. bioRxiv .

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa, 2012 Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Science .

Butler, D. G., B. R. Cullis, A. R. Gilmour, and B. J. Gogel, 2009 *ASReml-R reference manual.*

Caliński, T., S. Czajka, Z. Kaczmarek, P. Krajewski, and W. Pilarczyk, 2005 Analyzing multi-environment variety trials using randomization-derived mixed models. Biometrics **61**: 448–455.

Calus, M. P. and R. F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution **43**: 26.

Campbell, M., H. Walia, and G. Morota, 2018 Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. Plant Direct **2**: e00080.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, 2017 Stan: A probabilistic programming language. Journal of Statistical Software, Articles **76**: 1–32.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics **12**: 499–510.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345.

Dias, K. O. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da Costa e Silva, S. N. Parentoni, P. E. de Oliveira Guimarães, C. de Oliveira Anoni, J. M. V. Pádua, M. de Oliveira Pinto, R. W. Noda, C. A. G. Ribeiro, J. V. de Magalhães, A. A. F. Garcia, J. C. de Souza, L. J. M. Guimarães, and M. M. Pastina, 2018 Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity .

dos Santos, J. P., L. Paulo, M. Pires, R. Coelho, and D. C. Vasconcellos, 2016a Genomic selection to resistance to Stenocarpella maydis in maize lines using DArTseq markers. BMC Genetics **17**: 1–10.

dos Santos, J. P. R., R. C. De Castro Vasconcellos, L. P. M. Pires, M. Balestre, and R. G. Von Pinho, 2016b Inclusion of dominance effects in the multivariate GBLUP model. PLoS ONE **11**: 1–21.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE **6**: e19379.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with r package rrblup. Plant Genome **4**: 250–255.

Fernandes, S. B., K. O. G. Dias, D. F. Ferreira, and P. J. Brown, 2018 Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theoretical and Applied Genetics **131**: 747–755.

Ferrão, L. F. V., R. G. Ferrão, M. A. G. Ferrão, A. Fonseca, P. Carbonetto, M. Stephens, and A. A. F. Garcia, 2018 Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. Heredity .

Foley, J. a., N. Ramankutty, K. a. Brauman, E. S. Cassidy, J. S. Gerber, M. Johnston, N. D. Mueller, C. O'Connell, D. K. Ray, P. C. West, C. Balzer, E. M. Bennett, S. R. Carpenter, J. Hill, C. Monfreda, S. Polasky, J. Rockström, J. Sheehan, S. Siebert, D. Tilman, and D. P. M. Zaks, 2011 Solutions for a cultivated planet. Nature **478**: 337–342.

Garcia, A. a. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva, M. L. C. Vieira, R. Vicentini, E. a. Costa, M. C. Mancini, M. O. S. Garcia, M. M. Pastina, R. Gazaffi, E. R. F. Martins, N. Dahmer, D. a. Sforça, C. B. C. Silva, P. Bundock, R. J. Henry, G. M. Souza, M.-A. van Sluys, M. G. a. Landell, M. S. Carneiro, M. a. G. Vincentz, L. R. Pinto, R. Vencovsky, and A. P. Souza, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific Reports **3**: 3399.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2014 *Bayesian Data Analysis*.

Gianola, D., 2013 Priors in whole-genome regression: the bayesian alphabet returns. Genetics **194**: 573–96.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics **183**: 347–363.

Gianola, D., G. de los Campos, M. A. Toro, H. Naya, C.-C. Schön, and D. Sorensen, 2015 Do molecular markers inform about pleiotropy? Genetics **201**: 23–29.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler, 2014 Tassel-gbs: A high capacity genotyping by sequencing analysis pipeline. PLOS ONE **9**: 1–11.

Goodfellow, I., Y. Bengio, and A. Courville, 2016 *Deep Learning*. MIT Press, `http://www.deeplearningbook.org`.

Hamblin, J. and M. d. O. Zimmermann, 1986 Breeding common bean for yield in mixtures. Plant Breeding Reviews **4**: 245–272.

Hamelryck, T., 2012 *Bayesian Methods in Structural Bioinformatics*.

Han, B., X.-w. Chen, Z. Talebizadeh, and H. Xu, 2012 Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. BMC Systems Biology **6**: 1–12.

Henderson, C. R. and R. L. Quaas, 1976 Multiple trait evaluation using relatives' records. Journal of Animal Science **43**: 1188–1197.

Heslot, N., J. L. Jannink, and M. E. Sorrels, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Science **55**: 1–12.

Hoffman, M. D. and A. Gelman, 2014 The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research **15**: 1593–1623.

Holland, J., W. Nyquist, and C. Cervantes, 2003 Estimating and interpreting heritability for plant breeding: An update. plant breeding reviews vol. 22. Technical report.

Hung, H., C. Browne, K. Guill, N. Coles, M. Eller, A. Garcia, N. Lepak, S. Melia-Hancock, M. Oropeza-Rosas, S. Salvo, *et al.*, 2012 The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity **108**: 490.

Jia, Y. and J.-L. Jannink, 2012 Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics **192**: 1513–1522.

Langmead, B. and S. L. Salzberg, 2012 Fast gapped-read alignment with bowtie 2. Nature methods **9**: 357.

Lawrence, C. J. and V. Walbot, 2007 Translational Genomics for Bioenergy Production from Fuelstock Grasses: Maize as the Model Species. The Plant Cell **19**: 2091–2094.

Li, H., A. Rasheed, L. T. Hickey, and Z. He, 2018 Fast-forwarding genetic gain. Trends in Plant Science **23**: 184 – 186.

Loman, N. J., J. Quick, and J. T. Simpson, 2015 A complete bacterial genome assembled de novo using only nanopore sequencing data. bioRxiv **12**: 015552.

Lynch, M., B. Walsh, *et al.*, 1998 *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Mace, E. S., S. Tai, E. K. Gilding, Y. Li, P. J. Prentis, L. Bian, B. C. Campbell, W. Hu, D. J. Innes, X. Han, A. Cruickshank, C. Dai, C. Frère, H. Zhang, C. H. Hunt, X. Wang, T. Shatte, M. Wang, Z. Su, J. Li, X. Lin, I. D. Godwin, D. R. Jordan, and J. Wang, 2013 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nature Communications **4**: 1–9.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Morris, G. P., P. Ramu, S. P. Deshpande, C. T. Hash, T. Shah, H. D. Upadhyaya, O. Riera-Lizarazu, P. J. Brown, C. B. Acharya, S. E. Mitchell, J. Harriman, J. C. Glaubitz, E. S. Buckler, and S. Kresovich, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences of the United States of America **110**: 453–458.

Mullet, J., D. Morishige, R. McCormick, S. Truong, J. Hilley, B. McKinley, R. Anderson, S. N. Olson, and W. Rooney, 2014 Energy sorghum–a genetic model for the design of C4 grass bioenergy crops. Journal of Experimental Botany **65**: 3479–89.

Muraya, M. M., J. Chu, Y. Zhao, A. Junker, C. Klukas, J. C. Reif, and T. Altmann, 2017 Genetic variation of growth dynamics in maize (zea mays l.) revealed through automated non-invasive phenotyping. The Plant Journal **89**: 366–380.

Murphy, K. P., 2013 *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

Neapolitan, R., D. Xue, and X. Jiang, 2013 Modeling the altered expression levels of genes on signaling pathways in tumors as causal bayesian networks. Cancer Informatics **13**: 77–84.

Okeke, U. G., D. Akdemir, I. Rabbi, P. Kulakow, and J.-L. Jannink, 2017 Accuracies of univariate and multivariate genomic prediction models in african cassava. Genetics Selection Evolution **49**: 88.

Pauli, D., S. C. Chapman, R. Bart, C. N. Topp, C. J. Lawrence-Dill, J. Poland, and M. A. Gore, 2016 The quest for understanding phenotypic variation via integrated approaches in the field environment. Plant Physiology **172**: 622–634.

Ratcliffe, B., O. G. El-Dien, J. Klapste, I. Porth, C. Chen, B. Jaquish, and Y. A. El-Kassaby, 2015 A comparison of genomic selection models across time in interior spruce ( Picea engelmannii × glauca ) using unordered SNP imputation methods. Heredity **115**: 547–555.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS ONE **7**: 1–13.

Su, C., A. Andrew, M. R. Karagas, and M. E. Borsuk, 2013 Using Bayesian networks to discover relations between genes, environment, and disease. BioData Mining **6**: 1–21.

Team, S. D., 2018 *PyStan: the Python interface to Stan, Version 2.17.1.0.*.

Valluru, R., E. E. Gazave, S. B. Fernandes, J. N. Ferguson, R. Lozano, P. Hirannaiah, T. Zuo, P. J. Brown, A. D. Leakey, M. A. Gore, E. S. Buckler, and N. Bandillo, 2018 Leveraging mutational burden for complex trait prediction in sorghum. bioRxiv .

VanRaden, P., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414 – 4423.

Vermerris, W., 2011 Survey of Genomics Approaches to Improve Bioenergy Traits in Maize, Sorghum and Sugarcane. Journal of Integrative Plant Biology **53**: 105–119.

Watson, A., S. Ghosh, M. J. Williams, W. S. Cuddy, J. Simmonds, M.-D. Rey, M. A. M. Hatta, A. Hinchliffe, A. Steed, D. Reynolds, *et al.*, 2018 Speed breeding is a powerful tool to accelerate crop research and breeding. Nature plants **4**: 23.

Wolak, M. E., 2012 nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. Methods in Ecology and Evolution **3**: 792–796.

Xu, S., 2013 Genetic mapping and genomic selection using recombination breakpoint data. Genetics **195**: 1103–1115.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S. E. Mitchell, K. L. Roozeboom, D. Wang, M. L. Wang, G. A. Pederson, T. T. Tesso, P. S. Schnable, R. Bernardo, and J. Yu, 2016 Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants **2**: 1–7.