Ref. Ares(2021)2898115 - 30/04/2021

B4EST

Adaptive BREEDING for productive, sustainable and resilient FORESTs under climate change

Deliverable D3.2

Simulation and evaluation tools for integration of genomic information into breeding schemes

Planned delivery date (as in DoA): M30

Actual submission date: 30/04/2021, M36

Workpackage: WP3

Workpackage leader: INRAE

Deliverable leader: INRAE

Version: 1.0 and final

Project co-funded by the European Commission within the Horizon 2020 Programme (2014 - 2020)	
Dissemination Level	
PU Public	**PU**
CI Classified, as referred to Commission Decision 2001/844/EC	
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation action: GA no. 773383

Start date of the project: May 1st, 2018

TABLE OF CONTENTS

1 Summary 3
2 Methodologies, tools and case studies 4
2.1 Simulation tools for selected populations following a genomic evaluation scheme 4
2.2 Genomic evaluation tools corresponding to background knowledge in use for WP3 9
2.3 Genomic evaluation tools able to account for longitudinal data across environments 13
2.4 Genomic evaluation tools able to account for different omics data 17
3 Conclusions and perspectives 21
4 Partners involved in the work 22
5 Bibliography 22

1 Summary

This document presents a series of methodologies aimed at improving breeding performance in a context of integrating genomic information. These methodologies can be classified according to the level at which the improvement operates, as well as on the type of extra information that is included. According to this, the first series of methods concern **simulation tools**, which can improve breeding performance by allowing to identify *a priori* the best scenarios of integration of genomic information. The second type are those methodologies that belong clearly to background knowledge, with a proven benefit, and concerning notably **evaluation methods** to conduct **genomic selection**. The third type are methodologies that, although based on known approaches, are implemented in the novel context of **integrating environmental information** for genomic selection in forest tree breeding. The fourth type concerns methodologies still under heavy development and aimed at **integrating extra biological information** in the form of omics in the context of predicting phenotypes.

For all the cases, methods are briefly introduced with the required concepts and bibliographical context and complemented by the ongoing case studies in B4EST. Technical descriptions have been kept to a minimum, for simplicity, avoiding the mathematical description and formulations. All methods were conveniently referenced by bibliographical support, in case detailed methodological descriptions are required.

2 Methodologies, tools and case studies

2.1 Simulation tools for selected populations following a genomic evaluation scheme

2.1.1 Concepts and typologies

Breeding is based on simple principles emulating evolution, but its practical implementation involves often many interacting variables owing basically to operational and genetic factors. Often, setting those multiple factors becomes a strategic issue when establishing or upgrading a breeding program, impacting the costs and efficiency of the whole system in complex ways. Before entering into costly praxis, breeders can often predict the outcomes of breeding by using more or less complex simulation models that feed on some of the strategic factors at stake. Such outcomes can be useful guidance pinpointing the best compromises between conflicting factors when designing breeding. In forest tree breeding, where time investment is huge, simulations have often been used whenever disruptive technologies of potential use point on the horizon (Grattapaglia and Resende 2011).

Basically, simulation models used for breeding fall into two main categories, infinitesimal models and finite locus models, depending on where in the data flow the stochasticity is applied (Plate *et al.* 2019). Infinitesimal models use stochasticity to draw genetic and environmental deviates from known gaussian distributions. Finite locus models dive into a more functional principle in genetics, by using stochasticity to mimic Mendelian segregation of virtual alleles of known effects from generation to generation. The infinitesimal model is simple to set, without the need of extra assumptions on the genetic nature of traits, and appropriate over a few selection cycles of predictions. One example of its use in the framework of B4EST comes in the quantification of the usefulness of pedigree reconstruction with marker fingerprinting (Bouffier *et al.* 2019) (POPSIM, https://www.skogforsk.se/popsim). One main limitation of these models, however, is when questions are to be addressed on how the underlying nature of quantitative traits affects selection outputs, or over a long-term (Plate *et al.* 2019).

The finite locus model addresses that particular limitation of infinitesimal models by working with virtual genomes and genetic architectures. These virtual entities comprise a finite number of genes spread across a predefined gene map, with each of those genes bearing explicit allelic effects on a given trait or set of traits. The quantitative genetic variation appears when allele effects and their eventual interactions are integrated across genomes at population level for each simulated trait. The stochastic process happens when alleles are sampled at each locus, mimicking recurrent recombination and segregation processes within and between loci. As a direct consequence, drift and linkage disequilibrium can be evaluated with precision over generations.

Finite locus simulations have been extensively applied to evaluate the benefits of genomic selection, as one of the most disruptive methodologies (Meuwissen *et al.* 2001) of recent times in the breeding context. Most often, simulations are used to compare different marker-based evaluation approaches (Campos *et al.* 2013; Daetwyler *et al.* 2013). Another area of application is the study of the optimal population conditions in which to implement successfully genomic evaluation, typically focussing on the structure and composition of the population to train the genomic evaluation model (Rincent *et al.* 2017), or addressing the benefits of diversity constrains alongside the genomic evaluation (Gómez-

Romano *et al.* 2016). More rarely, simulations are applied to ponder cost and benefits of genomic selection in an economic context (Ben-Sadoun *et al.* 2020 p.).

2.1.2 Applications in B4EST

It is in this latter less explored area of economic impacts where finite locus simulations will be applied in the framework of B4EST, explicitly linking activities of WP3 and WP5 (Task 5.2). Briefly, before describing the simulation tool in itself, the objective will be to predict breeding outcomes in terms of genetic gains and diversity from alternative breeding schemes designed to run at the same operational costs. Some of those schemes mimic current breeding programs, some others are genome-aided counterparts under development (Bartholomé *et al.* 2016; Pegard *et al.* 2020). For this, costs functions for all the operational activities are being derived in WP5 for the maritime pine and hybrid poplar French breeding programs. The simulation stage will run all the alternative schemes at stake to obtain the predictions of gain and diversity required to identify the best cases in terms of gain per cost efficiency.

The simulation tool to be used is publicly available at https://github.com/mathieutiret/meiosis.git. It has been originally developed within the EU Gentree project ('676876') to study optimal ways to constrain diversity loss in long-term selected populations (Tiret, *et al.* 2021). Given its simplicity, flexibility and generality, the tool needs no adaptations to be used for the present economic analysis. The simulator comprises three modules working in close sequence over generations: one handling the genomic data, the second dealing with mating and segregation, and a third solving the selection for next generation. The first module is able to feed on real genotypic data, like the one being produced by SNP arrays in WP3, at the start of the simulation. The second module assumes a Wright-Fisher framework with selection and drift, and produces a stochastic realisation of the genotypes of the offspring following a given mating regime and structure of parental contributions. Although, the simulator is able to perform a generic formulation of optimum contribution selection (Gebregiwergis *et al.* 2020), this feature can be easily bypassed in the third module to do classical unconstrained selection. The simulator can be called repeatedly over a number of generations, while keeping track of all genomic data at stake for further analyses. Any genetic or genomic evaluation methodology can be applied at any generation, as a complementary module, like to ones referenced in the next section.

A typical output of the simulator can be seen in **Figure 1** (Tiret, *et al.* 2021), representing multiple realizations of selected populations for which gain and coancestry is plotted, and according to a constrain denoted by *alpha* that sets the importance of gain over coancestry in the output. The global pattern across populations delineates a Pareto optimum for gain and coancestry realisations.

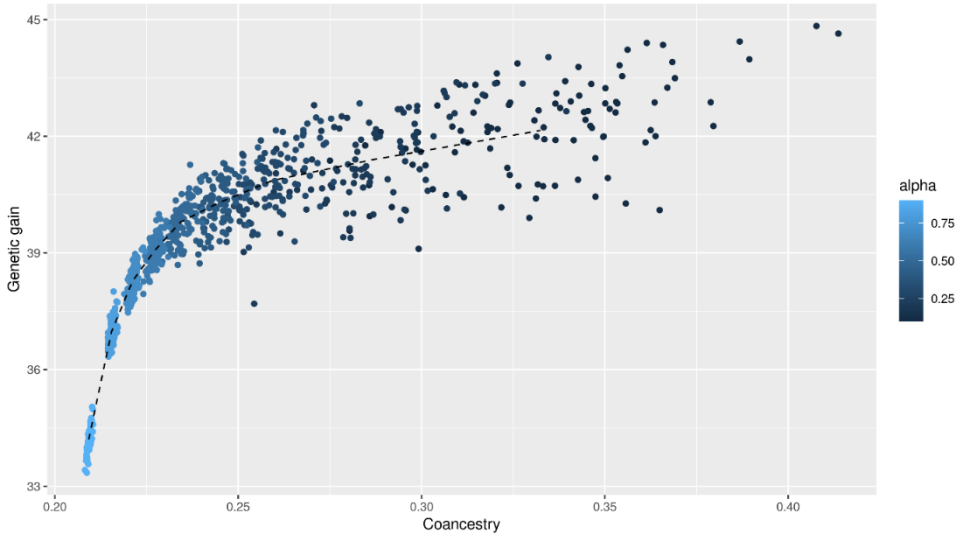

Figure 1. Output in terms of coancestry (abscissa) and genetic gain (ordinate) of multiple selected populations (dots) undergoing optimum contribution selection, differing in a constrain "alpha" setting the relative importance of gain with respect of coancestry. Dashed line delineates the expected Pareto optimum (Tiret, *et al.* 2021).

Another example of use of finite locus simulations in the framework of B4EST is the study of the reliability of GS across generations, accounting for different scenarios of training update. For this, the software Metagene, first developed at INRAE (Sanchez *et al.* 2007) and later used and redeveloped for several different studies on breeding strategy (Hallingback *et al.* 2014; Wu *et al.* 2016; Hall *et al.* 2016) was used as platform for the ongoing study.

Metagene is able to mimic most of the major features of inheritance, e.g. additive and dominance effects, locus segregation, major effect locus architectures or alternatively architectures closely resembling classical infinitesimal genetic models, linkage-disequilibrium (LD, historical and breeding generated) and genetic correlations between traits. To incorporate such a genetic framework into a breeding context, Metagene mimics the processes of directed or random selection, crossings and regulated family sizes plus genetic diversity management strategies. In addition, pedigree-based and genomic model predictions could be employed within Metagene by calling the GS3 breeding value prediction software https://github.com/alegarra/gs3 .

Within GS research at UPSC, SLU, the main objective was to determine the extent to which genomic predictions made across generations remain reliable and accurate (e.g. F1 constituting the training population while their progenies - F2 - constitute the population where GS without phenotyping is applied). Moreover, the impacts of different parameters associated with the breeding strategy or the genetic architecture on across-generational reliability and accuracy were investigated.

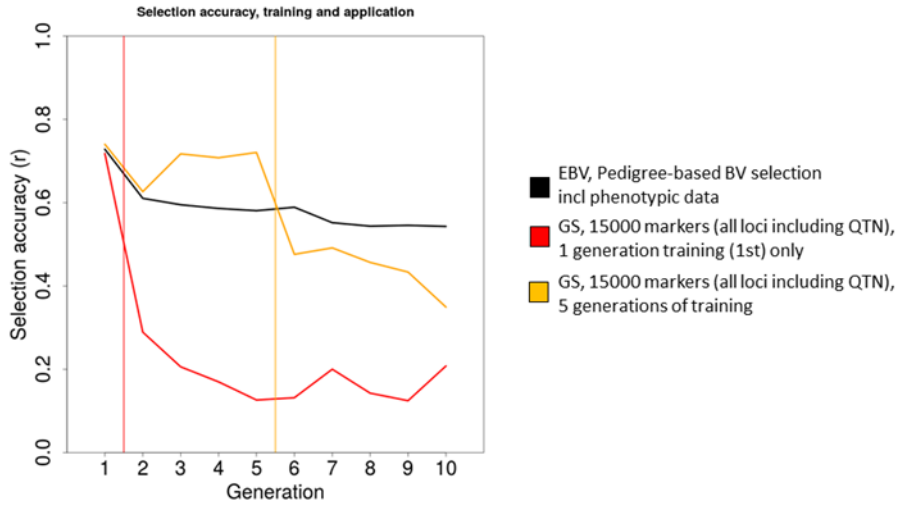

Figure 2. Simulation pilot study results showing the development of the selection accuracy across generations for a conventional pedigree-based breeding value selection approach (EBV) and for two genomic selection strategies. The genetic architecture (12 chromosomes, 15000 loci, 500 causal loci, 3000 cM long genome, trait heritability at 0.25) and the breeding population characteristics (2000 individuals per generation, 100 selected parents crossed in a single-pair-mating design) were designed to roughly mimic that of Norway spruce (*Picea abies*) (Hall *et al.* 2016; Bernhardsson *et al.* 2021). The vertical lines of the respective genomic selection approach signify the point in time where phenotypic data is no longer made available to the prediction model, i. e. the border between model training phase and model application phase.

Preliminary results from Metagene simulations (**Figure 2**) showed that, for forest trees with low levels of LD and high effective population sizes, high prediction accuracies across generations cannot be taken for granted unless considerable efforts are made with respect to model training, preferably by collating phenotypic data from multiple generations of breeding. It should be emphasized that improved genotyping efforts (more markers) would not compensate for a paucity of phenotypic data in the training phase because the marker set used for model training in the simulations included even the causal loci. It was also shown that strong initial LD among founders did have a positive effect on across-generational prediction accuracy but this effect was visible only for strong initial LD (and thus assumed very low effective population size). Furthermore, it appeared that breeding selection conducted from one generation to the next did per se have a negative impact on across-generational accuracy. This is not entirely unsurprising since a selection of future parents would also imply a change of the population characteristics (allele frequencies, LD, pedigree structures et c.) which would make prediction of advance generation breeding values more difficult to achieve with retained accuracy.

Although these results are preliminary and need corroboration, the conclusion appears to be that the development of genomic prediction models robust across generations requires a considerable amount of high-quality phenotypic data, preferably from a variety of populations, generations and breeding stages. It also appears that there is a limited scope to trade these requirements by increasing the density of markers. The requirements on phenotypic data moreover seems more demanding than what is usually indicated by GS-studies that validate their models by classical cross-validation approaches within one and the same generation of trees. The most concise message from these simulations are therefore that in order to make genomic selection a success, high quality and quantity of phenotypic data for the training of robust prediction models is paramount.

Many other simulation tools of use in breeding design are available. Making a comprehensive list here would be out of the scope of this deliverable. Among those, *xbreed* R package (Esfandyari and Sørensen) or *AlphaSimR* R package (Gaynor *et al.* 2021) are good examples of fully featured simulators able to adapt to multiple breeding designs, while providing many parametrizations to set up underlying finite locus genetic architectures. The reason for our simpler choice here is that of the need of a flexible and light alternative that could be easily combined with other analytical tools.

2.2 GS tools corresponding to background knowledge in use for WP3

2.2.1 Context, concepts and typologies

One of the objectives of WP3 in B4EST is to conduct genomic selection (GS) case studies that are close to breeding praxis for seven of the eight species in the project. Although the WP3 is also invested in methodological developments (Task 3.3), which will be the object of subsequent sections, most of the case studies rely on background knowledge in terms of methodological tools for their GS applications. The novelty here is to provide newly selected resources by using the best among already existing tools. The past two decades, since the breakthrough paper proposing the first explicit concept of GS (Meuwissen *et al.* 2001), have been characterized in the quantitative genetics context by numerous methodological developments aimed at improving the marker-based statistical fitting of phenotypes. A comprehensive review of these developments and their corresponding tools is definitely out of the scope of the present document. Excellent reviews already exist in the specialized literature (Campos *et al.* 2013; Crossa *et al.* 2017). Instead, a selection of the most widely used tools representing the different typologies will be presented briefly (**Table 1**), together with a few conceptual elements. These are also the GS tools most frequently used among B4EST partners.

By a GS tool, we mean here any modelling approach aimed at regressing phenotypes on massive numbers of molecular markers, with the disruptive application of unchaining the selection process from that of phenotyping. Trees too young to yield relevant phenotypes can be selected based on genotyping alone. From the modelling point of view, the key element is that of a *massive* number of markers that typically characterises GS. In fact, overcoming the problem of regressing a relatively low number of phenotypes (*n,* typically in the hundreds) on a massive number of markers (*p,* usually in the tens of thousands), a *big-p little-n* problem, has been the driving factor in the diversification of efficient developments for GS. Indeed, there are different ways to address a *big-p little-n* problem and such distinction determines the basic typology of strategies. Firstly, there is a big cluster of methodologies that use **explicit models** to explain the GS data, while implementing some strategy to *focus* on a reduced number of relevant explanatory factors. We will find within this first cluster most of the classical statistical approaches in GS, like linear models. A second cluster of methodologies comprise those using heuristics, with **no explicit model**, to classify and rank the relevant factors by seeking for hidden structural patterns in the data. These are well represented my recent applications in machine and deep learning (Breiman 2001; Ma *et al.* 2018).

Within the first cluster of classical statistical approaches, we can distinguish two basic strategies to overcome the *big-p little-n* problem (Campos *et al.* 2013). One is to use a penalty factor to render negligible the effect of many markers, also denoted regularization. The second is to do a selection of the relevant markers to be used in the fit. The use of a penalty factor is a way to solve the trade-off between goodness of fit and model complexity, i.e. between explicability and number of unknowns, with an effect that produces a shrinkage of the estimates towards zero, which simplifies the fitting for the remaining unknowns. One of the most common methods based on the penalized strategy is the **ridge regression** (Endelman 2011). Several tools referenced below propose ridge regression methods with the possibility of optimizing the penalty function.

Selection of markers occurs often as a result of a special penalty implementation, which induces some zeroing out of a number of marker effects. Often, penalties and selection of variables are combined in a single continuous procedure, like in the case of the **LASSO** methodology (Legarra *et al.* 2011),

switching off many markers and shrinking the effect of the remaining non-zeroes. One inconvenient of LASSO, however, is that the upper bound for non-zero markers must be equal to the number of observations for solvability, which in many cases might not be biologically sounding. This motivated the development of a derivative that removes such a constrain, denoted **elastic net** (Li and Sillanpää 2012).

The Bayesian framework offers also different alternatives to deal with the *big-p little-n* problem in the context of GS. Functionally speaking, the same strategies with penalization and selection of effects of previous frequentist tools also apply. This comes through the use of different priors for the assumption of the marker effect density, where the shape determines the degree of shrinkage and selection that is applied. In that sense, a gaussian prior mimics a ridge regression output inducing only shrinkage, while densities combining big masses close to zero and thick tails produce both shrinkage and marker selection. A whole Bayesian alphabet (**BayesRR**, **BayesA**, **BayesB**, **BayesC**, …) is then available by the use of different, more or less sophisticated priors (Habier *et al.* 2011).

Finally, the most common GS tool overcomes the *big-p little-n* problem in a very elegant and simple way, by regressing phenotypes observing a marker-based relationship matrix between individual effects, which basically equals unknowns and number of observations. This, commonly denoted **G-BLUP** model, is methodologically equivalent to the pedigree mixed-model counterpart, the classical "animal model" used worldwide. While the latter uses a pedigree-based relationship matrix, the former is based on markers. Given the universality of these two mixed model approaches, multiple options to accommodate different extra effects are available, like specific designs, spatial effects or longitudinal or multivariate data.

Table 1. Some of the most common GS tools available as R libraries (free opensource codes), representing all the typologies described previously in the section.

R package	GS tool, functions	Main reference	Other features
Rrblup	Ridge regression, G-BLUP, alternative kernels to account for effect covariates	(Endelman 2011)	Also used for GWAS
Regress	Ridge regression, G-BLUP, alternative kernels to account for effect covariates	(McCullagh and Clifford 2006)	Dominance and epistatic effects available
Glmnet	Ridge regression, Lasso, Elastic net	(Qian *et al.* 2020)	K-fold validation, grid search for penalty
Bglr	Ridge regression, BayesA, BayesB, BayesC, …	(Pérez and de los Campos 2014)	Multiple prior densities available, uses Gibbs sampler
randomForest	Machine learning GS	(Breiman 2001)	Learning and unsupervised modes
deepGS	Deep learning GS	(Ma *et al.* 2018)	GitHub repository

Ultimately, the choice of one of these methods over the others would depend on many factors, including how well the underlying genetic architecture of the trait under study fits the tool's built-in assumption for the complexity and distribution of effects. Typically, traits with underlying polygenic

nature (i.e. close to the infinitesimal model) tend to be well predicted with G-BLUP and ridge regression models, while Bayesian methods usually accommodate better some departures from a polygenic architecture (Campos *et al.* 2013). Other factors, like marker density, can be differently handled by the different tools, with usually those exerting greater selection and shrinkage being able to exploit the most out of the highest marker densities. In any case, it is wise to test several tools for each case study, and for each trait, in a cross-validation scheme.

There are other tools, of more generic use in B4EST, that are also able to accommodate G-BLUP GS, while fitting other useful effects. This is the case of ***breedR*** (https://github.com/famuvie/breedR), which accommodates spatial modelling for heterogeneous trials (Cappa *et al.* 2019) and competition modelling (Pablo Cappa *et al.* 2015). Another example is the standalone software ***wombat*** (http://didgeridoo.une.edu.au/km/wombat.php), able to handle efficiently complex models fitting longitudinal data to produce norms of reaction (Meyer and Kirkpatrick 2005).

2.2.2 Examples of applications in B4EST

Two examples will illustrate here the use of background knowledge GS tools in B4EST: one in Eucalyptus and another in Norway spruce.

In eucalyptus, GS has already been used in several proof of concept case studies. Most of them were designed for classical phenotypic traits, as growth or wood density, using large sets of families from large-scale factorial mating designs (Resende et al. 2012; Cappa et al. 2017). In general, the accuracy of these models suggest that GS is a competitive alternative to pedigree-based evaluation for classical traits. In the framework of B4EST, an ongoing initiative is genotyping a relevant part of ALTRI's breeding population in order to implement GS, not only for classical traits like growth but also for novel derivatives including time series of annual and daily growth increments. For this, briefly, a multispecies 72K SNPs commercial array designed for Eucalyptus is used to genotype the *E. globulus* breeding population, with 185 clones already assessed and 400 more expected 2021. The genotyping of the first batch yielded 56% of good quality markers out of the 72K array (40151 SNPs), covering well the species genome (**Figure 3**).

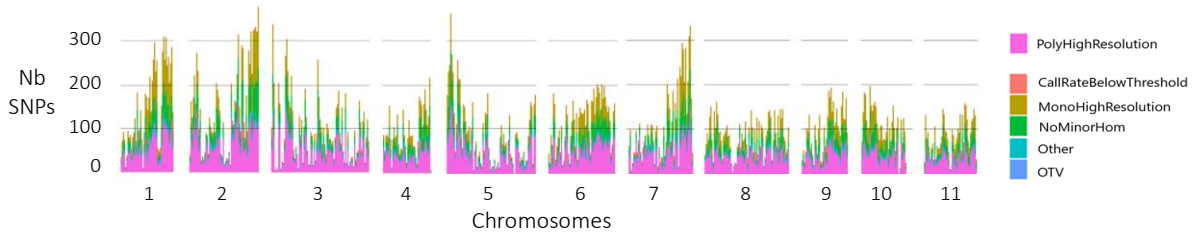

Figure 3. Distribution of the different classes of the genotyped SNPs along the 11 chromosomes of the *E. grandis* genome used as reference. *PolyHighResolution* and *MonoHighResolution* classes amount to 56% of the total number of available markers in the array.

This genotyping constitutes an excellent starting point to characterize the breeding population. This population comprises clones from three successive generations, with founders belonging to different genetic pools (Portugal, Spain, USA and Australia), but preliminary analyses show that the genetic diversity decreases across generations due to the high variability of the parental genetic contributions, typical of intensive selection (**Figure 4**). This characterization and the subsequent use of the genotyping

effort for GS in the breeding population are the subjects of a PhD thesis (G. Haristoy), supervised by J-M Gion (CIRAD) and L. Bouffier (INRAE). The next steps include the implementation of **G-BLUP** approaches on growth and survival data, accounting for different treatments of drought and frost, together with longitudinal data on growth involving different ages. Both sources of information will provide a solid framework to quantify genotypic by environment interactions across treatments and norms of reaction, for which the **G-BLUP** presents the required modelling flexibility.

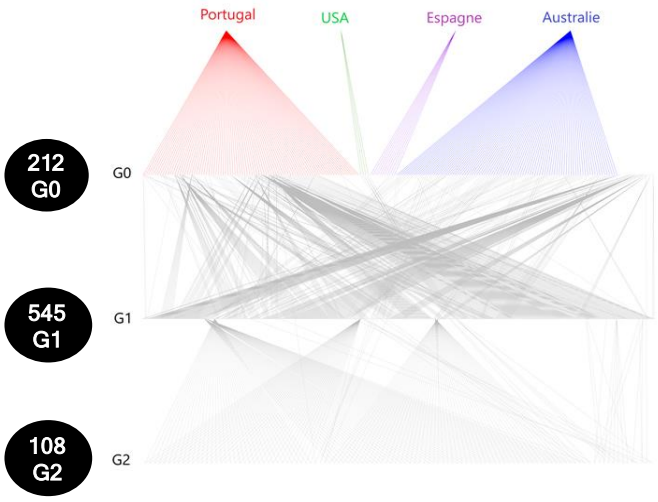

Figure 4. Pedigree founder and provenance structure (coloured contributions) and subsequent parental genetic contributions (grey) in ALTRI's breeding population sample used in B4EST. Generation sizes are given by the encircled figures.

Another example of application of well-known GS tools comes from a recently published work funded by B4EST (Chen *et al.* 2019). In this study, the benefits of GS using a **G-BLUP** approach were evaluated against a pedigree counterpart for two important traits in the context of the Swedish Norway spruce breeding program: tree height and solid wood quality. The latter trait was assessed by two complementary field measurements: Pilodyn penetration and acoustic velocity. The flexibility of the mixed model framework that underlies G-BLUP, and the pedigree-based BLUP, allowed the exploration of different ways to capture non-additive variation. This extra source of information can be important for advanced breeding, for instance, in forward selection and for traits showing dominance and epistatic genetic variation. Results show that G-BLUP models accounting for dominance for height can provide extra gain. On the other hand, for late assessment traits like those involving solid wood quality, G-BLUP yielded the highest gains per unit of time, in as much as 90% improvement compared to the pedigree counterpart.

2.3 Genomic evaluation tools able to account for longitudinal data across environments

2.3.1 Context and concepts

Trees are among the longest-living organisms and, as plants, they are constitutively attached to their substratum. Trees, unlike domestic species that benefit from controlled growing conditions, face their nursing environment with little or no husbandry. Time and immobility in an ever changing and challenging environment pinpoint the importance for trees of adaptable plastic responses, as a key for their adaptive success (Kramer 1995; Petit and Hampe 2006). In that sense, height and diameter, the classical traits used for genetic evaluation, are extremely integrative assessments, i.e. the cumulative result of numerous underlying functional processes. Such level of integration with single point assessments like height prevents any dissection of the way environmental cues interact with tree growth through the underlying functional processes (Lagercratz, Ulf *et al.* 2013). The use of longitudinal or function traits over environmental gradients open multiple possibilities in the understanding of how trees react and adapt to the changing environment, and in the prediction of their adaptation to future climates. Technically, this reaction function corresponds to the concept of **phenotypic plasticity** (PP), where the phenotype for a given genotype reflect environmental changes in a vectorial or longitudinal fashion. PP is a central concept in many B4EST studies, requiring specific analytical methodologies. Below, we present two methodological examples reflecting this importance.

2.3.2 Example of functional estimation of time-varying QTLs

In the context of QTL detection when assessing marker effects globally, the number of parameters to be accounted for during the fitting of the model can easily become a limiting factor, and such limitation can be exacerbated when multiplying the environmental levels at which the quantification is applied. This is the *big-p little-n* problem already mentioned in previous section, but expanded into a new dimension of multiple environments. A recent publication funded by B4EST (Heuclin *et al.* 2021) investigates new methodological developments to overcome such limitation, when the succession of environments are ages across which measurements are taken (PhD thesis, B. Heuclin). The authors propose a variable selection method in the Bayesian context that uses a special kind of prior denoted *pike-and-slap* by its shape, a mixture of two distributions combining a conspicuous peak neighbouring zero and thick long tails towards the extremes. The consequence in terms of estimation is that it produces both shrinkage and marker selection, in order to deal with the *big-p little-n* problem.

The output is a function describing time-related or age-related effects of *selected* QTLs resulting associated to traits, as shown in the example in **Figure 5** below.

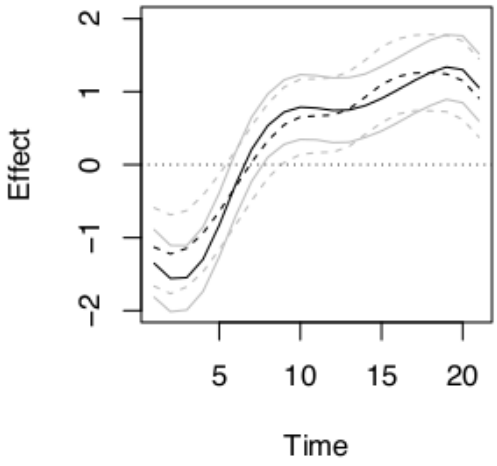

Figure 5. Genetic effects over time (ages) of markers selected in two contrasting environmental conditions, corresponding to a well-watered experimental setting (solid line), and a moderate water deficit counterpart, with median, upper and lower bounds lines (Heuclin *et al.* 2021).

The benefits of this methodology with respect to other variable selection counterparts were assessed through simulations in an age-dependent functional context. Results show that the newly proposed method is competitive, outperforming the other methods in terms of estimation and prediction qualities. Authors also tested the impact of the number of individuals and ages available, concluding that the combination of high number of individuals with extended longitudinal data improves the variable selection process even for small effect QTLs, while at the same time the confidence associated to large effect QTLs results strengthened. Another aspect analysed in the paper is the impact of collinearity between markers resulting from linkage disequilibrium, by using a real dataset from *Arabidopsis thaliana* as a starting point for a forward simulation approach. While at intermediate levels of collinearity between markers the new method is not affected in its detection capabilities, it is at high levels of collinearity that discrimination between zero and non-zero effects of adjacent markers can be substantially impacted. The authors pinpoint the need to do supplementary research on real data to quantify the extent of that problem.

In conclusion, the proposed method pinpoints the fact that approaches based on longitudinal data allow better detections compared to single-time approaches. One limitation of the present study would be the use of a time or age varying environmental condition instead of an environmental quantitative covariable, of more relevance when prospecting new conditions for the evaluated material. Such an approach, even if limited to the age-to-age analyses, could be of higher relevance for some of the case studies in B4EST, like the one of Eucalyptus cited in previous sections, a maritime pine case that will be presented in the next section, or in the black poplar breeding population where several series of longitudinal data are available.

2.3.3 Example of functional estimation of environmentally driven individual genetic effects

Another dimension where the knowledge of the impact of environmental gradients is of high relevance is that of individual level, at which selection takes place. Wood records in trees are excellent candidates for such functional assessments, as they can reflect precisely how individuals react to climatic and other environmental clues. Particularly, wood density has been found to be strongly associated to resistance to mechanical failure in water conduction (Cochard *et al.* 2009; Choat *et al.* 2012), and

even to mortality as a consequence of heatwaves and extreme draught (Martinez-Meier *et al.* 2008; Sergent *et al.* 2014). Conceptually speaking, the principle of functional estimation of environmentally driven individual genetic effects is similar to the one described for the previous study (Heuclin *et al.* 2021): deriving a function describing how a genetic component behaves across a succession of environments, either taken as time or space related qualitative levels, or as a quantitatively assessed variable like a climatic index. In the following application, both ways of characterization of the environment are considered. This application concerns a subset of the maritime pine breeding population and it is the object of an ongoing PhD thesis (V. Papin), supervised by L. Bouffier (INRAE) and L. Sanchez (INRAE).

In a nutshell, one of the main objectives of this thesis is to derive genomic predictions of wood-based plasticity for breeding purposes. For this, thanks to a single modelling approach, three sources of information will be combined:

- longitudinal wood records contained in increment cores (see **Figure 6** for an example of wood record) and dendrometers from a subset of relevant elite trees growing in two contrasting testing sites;
- time series of weather records corresponding to growing sites and periods of growth;
- genomic information provided by dense genotyping of elite trees using B4EST 4TREE SNP array.

Obtaining individual plasticity functions will be possible by the use of random regression approaches (Meyer and Kirkpatrick 2005) in the mixed-model framework of **G-BLUP**, the classic GS tool already described in previous section. Basically, random effects can be modelled as a function of a covariate representing the environmental gradient. Although this kind of random regression modelling has already been implemented in livestock in the past (Schaeffer 2004), notably to stablish lactation functions in dairy cattle, it has rarely been used in forest trees (ChunKao *et al.* 2009; Alves *et al.* 2020), despite the promising availability of wood records. One of those rare studies (Marchal *et al.* 2019) opened up this possibility using increment cores in hybrid larches. The study, however, had only pedigree information available for the approach. The novelty in the maritime pine implementation is then the use of random regression of wood records in a genomic prediction context using **G-BLUP** modelling.

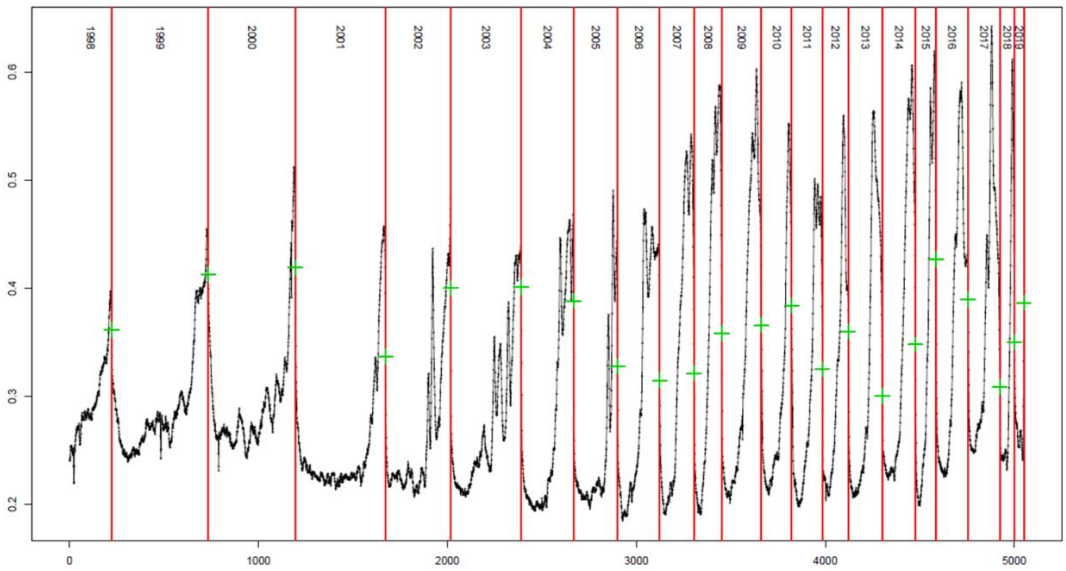

Figure 6. Profile corresponding to a single maritime pine tree showing wood micro-density variations across years (thick black pattern), and obtained by X-raying a specially prepared increment core. On the horizontal axis, the figures correspond to number of cumulated points from pit to bark, a measure of radial growth. The vertical axis measures density (kg/m^3). Years are delimitated by vertical red lines with corresponding year indicated in the upper part, from 1998 to 2019. Green symbols correspond to the boundary between successive years.

The environment will be characterized following two alternative approaches. The simplest will be to consider each year in the profile as a qualitative level of environment, in an age-to-age trend similar to the one described previously for the QTL functional mapping. Climate records will be summarized at annual levels to serve as covariates in the regression. Similarly, wood records will be summarized at annual levels, as annual increments in diameter or mean levels of density (**Figure 6**). Therefore, the pairing of wood records and climate will be done by year, across the longitudinal trend. A more complex alternative will be to consider within year variation in the wood record, given that this level harbours most of the phenotypic variation and climatic records at daily or weekly levels can be easily obtained for the experimental sites. The challenge here is to synchronise wood and climate records, as only the later is labelled on a time axis. The anchoring of each wood variable to the corresponding level of climate record will be done by using a machine learning approach (*randomForest*, **Table 1**), with a training subset where daily growth was recorded via dendrometers, and involving one fourth of the individuals to be used for the subsequent GS prediction. Such a synchronization step needs to be done at a separate step for simplicity of the modelling, and before the final step of genomic evaluation.

The thesis has just finished the data curation step and the first modelling involving random regression have already started. The thesis is expected to provide a methodological framework to apply genomic evaluation to longitudinal data, notably for wood records. Such data is known to be costly and difficult to obtain, but could help to address some of the key questions on tree adaptation to climate.

2.4 Genomic evaluation tools able to account for different omics data

2.4.1 Context and concepts

In this section, we will introduce one last series of initiatives aimed at improving the prediction of phenotypes. As with previous cases, the improvement factor of these initiatives is the incorporation of extra information, which is of biological nature here and aimed at describing the functional complexity occurring between sequences and phenotypes. One substantial part of such complexity is made of interactions, most often of unknown effect. Gene-to-gene and gene-to-environment interaction networks are a pervasive phenomenon underlying the observable phenotype (Tyler *et al.* 2009). Understanding how these interactions interplay at different biological levels to produce global phenotypes has become the new emerging field of **systems biology** (Mackay *et al.* 2009; Ritchie *et al.* 2015). Techniques like genome-wide association (GWAS) and GS have paved the way towards reliable phenotypic prediction and knowledge of underlying genetic architectures. However, these approaches rely heavily on the paradigm that genetic variation can be additively dissected into the contributions of relevant DNA polymorphisms. Although there are numerous attempts to improve our understanding of genetic variation by accounting for interaction terms like epistasis (Wei *et al.* 2014; Jiang and Reif 2015), often the underlying methodologies rely on statistical approaches rather than on functional and biological knowledge. Behind biological information, we find new data generally denoted ***omics***, like transcriptomics from RNA sequencing, metabolomics or proteomics, which start to populate the gap between phenotypes and DNA sequences. Numerous studies use *omics* like RNA co-expression profiles to infer gene networks, notably in humans. Still few of those studies address the problem of predicting with *omics* in tree species (Chateigner *et al.* 2020). At the end, it is a question of integrating different layers of biological information to gain understanding of genetic architecture and predictability for phenotypes. Two B4EST case studies illustrate here methodologically the use of extra *omics* information for the improvement of prediction in the GS context. One based on the use of GWAS with validation by differential RNA expression in Norway spruce, and another one based on the integration of two *omics* into GS schemes in black poplar.

2.4.2 Implementations in B4EST

In Norway spruce, the objective of the SLU team was to improve GS predictability by incorporating as fixed effects in the prediction model the most important markers, either QTLs or validated QTNs, resulting from a previous step of GWAS. Indeed, some authors have already pinpointed the benefits of incorporating major gene effects to prediction models (Bernardo 2014; Lopes *et al.* 2017). To validate such a hypothesis for Norway spruce, the main results from a GWAS using a large dataset were verified by an analysis of differential expression for contrasting early *versus* late bud flush phenotypes at a given time of early spring (**Figure 7**). Several resulting QTNs from the differential analysis were also associated to important traits, like growth, wood quality, disease and phenology, when using the most important gene within the gene model. The most significant QTN, located in contig MA_12842, was added as a fixed effect to a G-BLUP prediction model, resulting in improved predictability following a cross-validation scheme for several traits. The authors also investigated the effects of extra additions coming from QTNs of lesser effects according to GWAS, resulting in less conclusive improvements for GS predictability. As a conclusion, GS models could be easily improved with current methodological approaches and some extra biological information from GWAS and transcriptomic analyses. This

empirical study is currently being complemented by a simulation approach to further verify the usefulness of including QTNs with relatively large effect in GS.

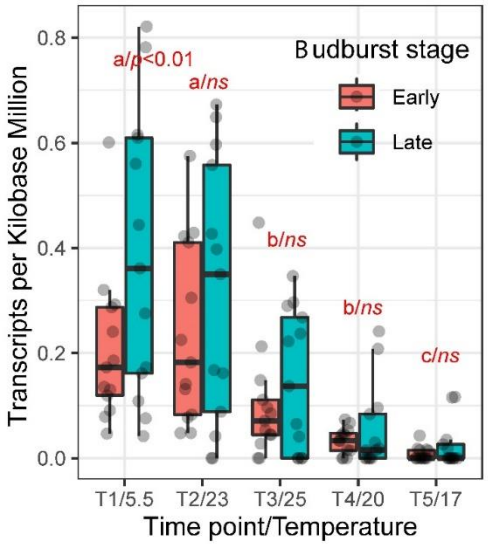

Figure 7. Difference in gene expression during budburst stages. The expression of RNA-Seq in Transcripts Per Kilobase Million (TPM) for gene MA_12842g0030 from six early bud burst and six late bud burst clones, sampled five times in May 2018. The x-axis is the five sampling time points (T1–T5) and their corresponding temperature of sampling time (°C). Ns represents non-significant between early and late budburst clones and letters represent the levels of significance between different sampling times.

As for the black poplar case study, the INRAE team had the objective of developing a methodology to integrate the data coming from genotyping and that from transcriptomics into a single prediction approach. This objective is part of PhD project that started two years ago within the framework of B4EST (Task 3.3) (A. Wade, supervised by V. Segura, H. Duruflé and L. Sanchez). The transcriptomic data used in the thesis has already been the object of a previous study on the role of different typologies of genes, according to their co-expression network importance, in the prediction quality of phenotypes (Chateigner *et al.* 2020). The genotyping information came from RNA sequence (Rogier *et al.* 2018), providing roughly 400 000 SNPs. In the present thesis, the transcriptomic layer is combined to the genotyping information for the same set of 241 black poplar trees, corresponding to a collection that represents part of the diversity of the species in western Europe. Besides genotypes and transcripts, these trees were assessed for 21 traits including growth, phenology, rust resistance and several biochemical quantifications, in two sites with contrasting conditions.

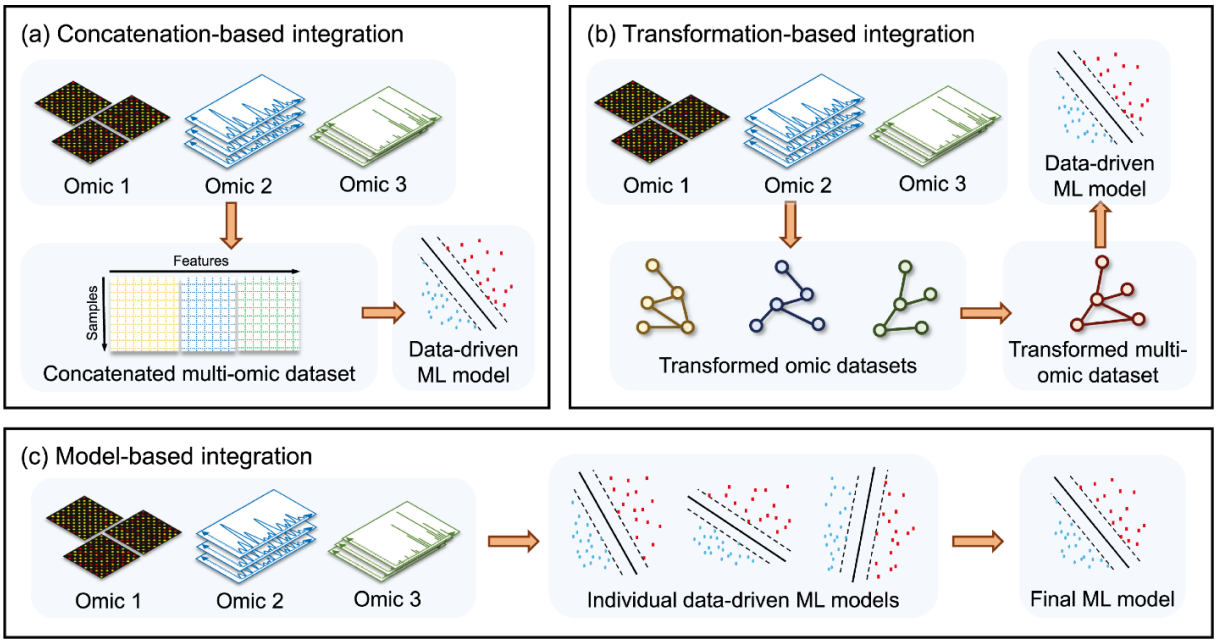

Figure 8. Different typologies of omics data integration for modelling purposes (Zampieri *et al.* 2019).

Briefly, *omics* could be integrated following three different strategies according to the level at which the integration takes place (Zampieri *et al.* 2019) (**Figure 8**). The simplest of the integrations happens at the level of the raw data by doing a concatenation of the different datasets, which participate without any further transformation into the modelling step. The second strategy involves some data transformation previous to the integration process, so that the different sets can be differentially scaled and weighted. Finally, the third strategy involves integration at the level of the models, each individually obtained from one of the data sets. This latter option, not covered in the thesis, is typically handled with machine learning approaches.

For the thesis, the team originally chose to do integration at the level of transformed data, as the aim was to weight the different sources of information. This was obtained by creating a kernel describing resemblances between individuals, in a way similar to the genomic relationship matrix for G-BLUP. The markers contributing to the kernel were firstly grouped into blocks corresponding to genes associated to transcripts. Then, each block was weighted according to the importance of the transcript count for the corresponding gene. Different ways of summarising markers into blocks and weighting strategies were developed. Results did not show any clear advantage of integration versus single data models. A second integration used the simpler concatenation strategy, in order to study how the different features, markers and genes, behaved at the integration step. Unlike previous strategy, the new simpler integration showed benefits for some of the most important traits in terms of predictability (**Figure 9**), like for growth and rust resistance. In general, the most benefited traits were those already showing the best predictabilities under simpler GS prediction models. The integration showed losses in accuracy mostly for those traits with already poor performances under simpler models, in general for the biochemical assessments.

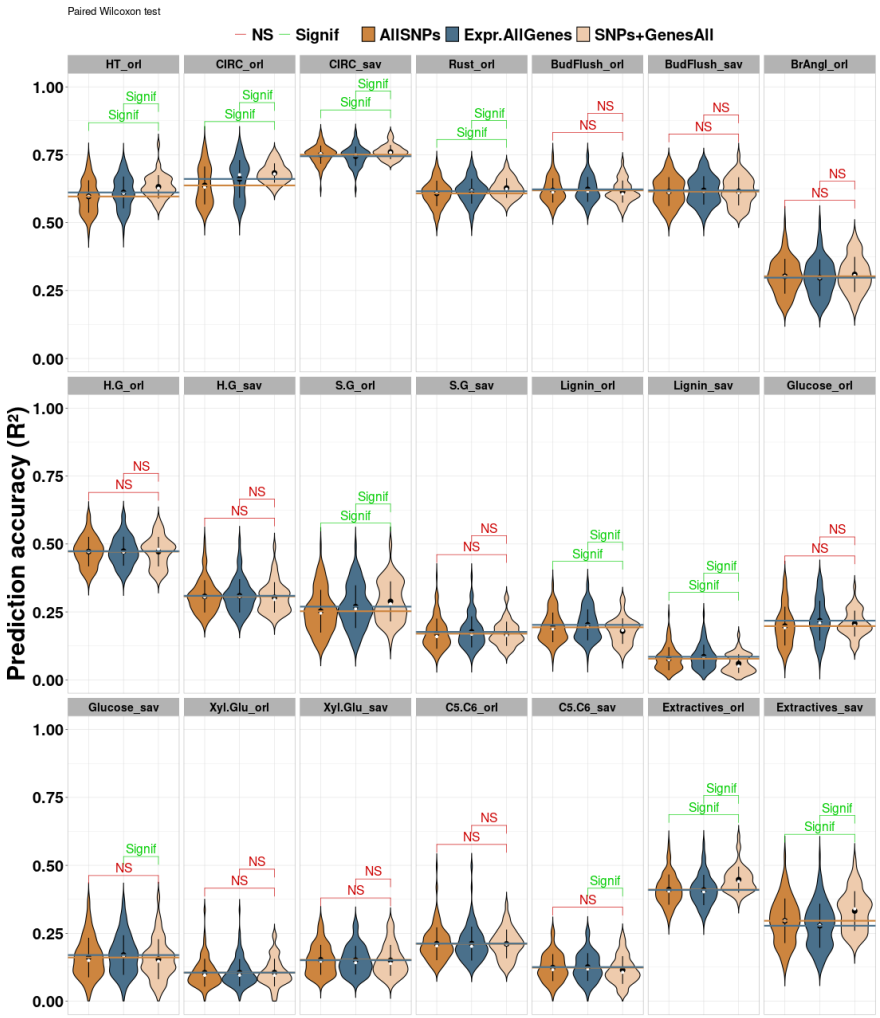

Figure 9. Violin plots of prediction accuracies for 21 traits in the poplar dataset according to three models, using only SNPs (dark brown to the left in the panels), using only gene transcripts as in (Chateigner *et al.* 2020) (dark blue), and concatenating both SNP and gene information (light brown to the right). Each distribution of accuracies resulted from a cross-validation scheme. Signification from paired tests is shown for comparisons between models. Some traits were evaluated in two sites, denoted by the ending of the name of the trait ("*orl*" standing for Orléans in France and "*sav*" for Savigliano in Italy).

In order to understand the reasons underlying the benefits of concatenation, some extra analyses were undertaken on the role that particular typologies of genes and SNPs had on the gains in accuracy. Firstly, an association analysis for transcripts based on SNPs revealed numerous eQTLs, QTLs associated to expressional patterns, where some have relevant *hub* regulator roles with large impacts on the expression of many genes across the genome. Based on positional proximity to genes, the different eQTLs could be classified as CIS regulatory elements (non-coding DNA regulating the transcription of neighbouring genes), and/or as TRANS regulatory elements (regulating the transcription of distant genes). Similarly, genes were classified either as targets of the different regulators or showing no link. Secondly, the change in estimated effects of genes and SNPs from the models with only one data type to those involving concatenated data were minutely analysed for all traits. Results showed that the traits most benefited by concatenation presented also substantial changes in estimated effects for some of the genes and SNPs. This was not observed for the remaining traits with no benefits after integration. Among the SNPs most affected by changes in estimated effects, i.e. for traits benefited by integration, it was found a high proportion of TRANS eQTLs, with

generally downsized effects after concatenation. On the gene side, the opposite pattern was generally observed and, again, only for traits showing benefits after concatenation. The genes being regulatory targets were also those showing the most important changes in estimated effects, with mostly increases in effect with concatenation.

In conclusion, beneficial integration appears to operate through a change in the weighting of important genes and SNPs, notably those showing already mutual associations through eQTLs. Looking retrospectively at the first attempt of integration in the thesis, these results throw some light on the reasons why that approach might have been unsuccessful, i.e. the weighting after transformation was almost opposite to the one found after concatenation. These results have not been published yet, but they constitute already a promising way to explore data integration and to devise future methodologies combining in an optimal way different omics.

3 Conclusions and perspectives

Through the different types of methodologies presented here, we intended to make a wide panorama of the tools and methodologies available for WP3. Many correspond to background knowledge being implemented *straight away* from bibliography in our case studies. Some are derivatives with novel applications in the context of forest species. Finally, a few are ongoing new developments. In any case, all of them provided or are providing novel results, even those using known methodologies with abundant bibliography.

The four types of tools cover comprehensively most of the relevant aspects of data integration in GS, from strategic design of new breeding programs to experimental integration of omics. Simulations, by their power to evaluate rapidly numerous scenarios, provide a way to improve the design of our experiences implementing GS. This is the case for economic impacts in poplar and maritime pine breeding of new GS schemes, or to explore the degree of updating of prediction models across generations. Trusted GS methods, with ample bibliographical support, offer essential baselines for GS in our species, some of which being analysed for the first time under a GS scheme. For the methodologies involving the inclusion of the environment as covariable, the role of genomic information can be essential to render models more robust. We have seen how the environment needs to be considered as a consubstantial part of tree evaluation and not as a nuisance to be systematically removed from predictions. The proposed methodologies allow implementations in that sense, with the added support of genomic information. Finally, in a context of rapidly growing genomic resources for forest species, it is essential to benefit from data integration and feed a more systemic view of tree behaviour. The last models presented here pinpoint into that direction, with some developments that incorporate function into prediction, adding explicability to prospection.

It is expected that most of the developments referenced here will be the object of **scientific publications**, as it is already the case of some of them (Chen *et al.* 2019; Heuclin *et al.* 2021). The most relevant output, however, related to the implementation and development of these methodologies is unquestionably the **training of junior scientists** in a field of research in need of new talents. Indeed, genetics, genomics and statistics are highly demanded expertises by many sectors, including big attractors like human health. Four **PhD thesis**, at different phases of development, are underway and will further contribute to the scientific production of B4EST. Ultimately, some of the methodologies could be transferred into **publicly available software**. This could be the case of the last development

on the integration of omics for prediction. However, such a valorisation will surely fall beyond the time frame of B4EST, given the other urgent phases to accomplish before the end of the project.

4 Partners involved in the work

The following scientists and partners have been actively involved in the elaboration of this deliverable.

Partners: UU, Skogforsk, SLU, CIRAD and INRAE.

- In the section 3.1, M. Tiret (UU), L. Bouffier (INRAE), H. Hallingbäck (Skogforsk), H. Wu (SLU) and L. Sanchez (INRAE).
- Section 3.2, G. Haristoy (**PhD**, CIRAD), J.M. Gion (CIRAD), L. Bouffier (INRAE), Z. Chen (SLU), H. Wu (SLU) and L. Sanchez (INRAE).
- Section 3.3, M. Denis (CIRAD), B. Heuclin (**PhD**, CIRAD), V. Papin (**PhD**, INRAE), L. Bouffier (INRAE) and L. Sanchez (INRAE).
- Section 3.4, Z. Chen (SLU), H. Wu (SLU), A. Wade (**PhD**, INRAE), H. Duruflé (INRAE), V. Segura (INRAE) and L. Sanchez (INRAE).

5 Bibliography

Alves R. S., M. D. V. de Resende, C. F. Azevedo, F. F. e Silva, J. R. do A. S. de C. Rocha, *et al.*, 2020 Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. Tree Genet. Genomes 16: 38. https://doi.org/10.1007/s11295-020-01431-5

Bartholomé J., J. Van Heerwaarden, F. Isik, C. Boury, M. Vidal, *et al.*, 2016 Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17: 604. https://doi.org/10.1186/s12864-016-2879-8

Ben-Sadoun S., R. Rincent, J. Auzanneau, F. X. Oury, B. Rolland, *et al.*, 2020 Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality. Theor. Appl. Genet. 133: 2197–2212. https://doi.org/10.1007/s00122-020-03590-4

Bernardo R., 2014 Genomewide Selection when Major Genes Are Known. Crop Sci. 54: 68–75. https://doi.org/10.2135/cropsci2013.05.0315

Bernhardsson C., Y. Zan, Z. Chen, P. K. Ingvarsson, and H. X. Wu, 2021 Development of a highly efficient 50K single nucleotide polymorphism genotyping array for the large and complex genome of Norway spruce (*Picea abies* L. Karst) by whole genome resequencing and its transferability to other spruce species. Mol. Ecol. Resour. 21: 880–896. https://doi.org/10.1111/1755-0998.13292

Bouffier L., J. Klápště, M. Suontama, H. S. Dungey, and T. J. Mullin, 2019 Evaluation of forest tree breeding strategies based on partial pedigree reconstruction through simulations: Pinus pinaster and Eucalyptus nitens as case studies. Can. J. For. Res. https://doi.org/10.1139/cjfr-2019-0145

Breiman L., 2001 Random Forests. Mach. Learn. 45: 5–32. https://doi.org/10.1023/A:1010933404324

Campos G. de los, 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. https://doi.org/10.1534/genetics.112.143313

Cappa E. P., Y. A. El-Kassaby, F. Muñoz, M. N. Garcia, P. V. Villalba, *et al.*, 2017 Improving accuracy of breeding values by incorporating genomic information in spatial-competition mixed models. Mol. Breed. 37: 125. https://doi.org/10.1007/s11032-017-0725-6

Cappa E. P., F. Muñoz, and L. Sanchez, 2019 Performance of alternative spatial models in empirical Douglas-fir and simulated datasets. Ann. For. Sci. 76: 53. https://doi.org/10.1007/s13595-019-0836-9

Chateigner A., M.-C. Lesage-Descauses, O. Rogier, V. Jorge, J.-C. Leplé, *et al.*, 2020 Gene expression predictions and networks in natural populations supports the omnigenic theory. BMC Genomics 21: 416. https://doi.org/10.1186/s12864-020-06809-2

Chen Z.-Q., J. Baison, J. Pan, J. Westin, M. R. G. Gil, *et al.*, 2019 Increased Prediction Ability in Norway Spruce Trials Using a Marker X Environment Interaction and Non-Additive Genomic Selection Model, (J. R. Stommel, Ed.). J. Hered. 110: 830–843. https://doi.org/10.1093/jhered/esz061

Choat B., S. Jansen, T. J. Brodribb, H. Cochard, S. Delzon, *et al.*, 2012 Global convergence in the vulnerability of forests to drought. Nature. https://doi.org/10.1038/nature11688

ChunKao W., B. Andersson, and P. Waldmann, 2009 Genetic analysis of longitudinal height data using random regression. Can. J. For. Res. 39: 1939–1948.

Cochard H., T. Holtta, S. Herbette, S. Delzon, and M. Mencuccini, 2009 New insights into the mechanisms of water-stress-induced cavitation in conifers. Plant Physiol. 151: 949–954. http://dx.doi.org/10.1104/pp.109.138305

Crossa J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, *et al.*, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. 22: 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Daetwyler H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193: 347–365. https://doi.org/10.1534/genetics.112.147983

Endelman J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4. https://doi.org/10.3835/plantgenome2011.08.0024

Esfandyari H., and A. C. Sørensen, xbreed: An R package for genomic simulation of purebred and crossbred populations. 32.

Gaynor R. C., G. Gorjanc, and J. M. Hickey, 2021 AlphaSimR: an R package for breeding program simulations. G3 GenesGenomesGenetics 11. https://doi.org/10.1093/g3journal/jkaa017

Gebregiwergis G. T., A. C. Sørensen, M. Henryon, and T. Meuwissen, 2020 Controlling Coancestry and Thereby Future Inbreeding by Optimum-Contribution Selection Using Alternative Genomic-Relationship Matrices. Front. Genet. 11: 345. https://doi.org/10.3389/fgene.2020.00345

Gómez-Romano F., B. Villanueva, J. Fernández, J. A. Woolliams, and R. Pong-Wong, 2016 The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. Genet. Sel. Evol. 48: 1–17. https://doi.org/10.1186/s12711-015-0172-y

Grattapaglia D., and M. D. V. Resende, 2011 Genomic selection in forest tree breeding. Tree Genet. Genomes 7: 241–255. http://dx.doi.org/10.1007/s11295-010-0328-4

Habier D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. 12.

Hall D., H. R. Hallingbäck, and H. X. Wu, 2016 Estimation of number and size of QTL effects in forest tree traits. Tree Genet. Genomes 12: 110. https://doi.org/10.1007/s11295-016-1073-0

Hallingback H. R., L. Sanchez, and H. X. Wu, 2014 Single versus subdivided population strategies in breeding against an adverse genetic correlation. Tree Genet. Genomes 10: 605–617. https://doi.org/10.1007/s11295-014-0707-3

Heuclin B., F. Mortier, C. Trottier, and M. Denis, 2021 Bayesian varying coefficient model with selection: An application to functional mapping. J. R. Stat. Soc. Ser. C Appl. Stat. 70: 24–50. https://doi.org/10.1111/rssc.12447

Jiang Y., and J. C. Reif, 2015 Modeling Epistasis in Genomic Selection. Genetics 201: 759–768. https://doi.org/10.1534/genetics.115.177907

Kramer K., 1995 Phenotypic plasticity of the phenology of seven European tree species in relation to climatic warming. Plant Cell Environ. 18: 93–104.

Lagercratz, Ulf, Badnoli, Francesca, Källman, Thomas, G. G. Vendramin, V. Jorge, *et al.*, 2013 *Novel Tree Breeding*(S. Lee, and J. Woolliams, Eds.). INIA - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Madrid.

Legarra A., C. Robert-Granie, P. Croiseau, F. Guillaume, and S. Fritz, 2011 Improved Lasso for genomic selection. Genet. Res. 93: 77–87. http://dx.doi.org/10.1017/S0016672310000534

Li Z., and M. J. Sillanpää, 2012 Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor. Appl. Genet. 125: 419–435. https://doi.org/10.1007/s00122-012-1892-9

Lopes M. S., H. Bovenhuis, M. van Son, Ø. Nordbø, E. H. Grindflek, *et al.*, 2017 Using markers with large effect in genetic and genomic predictions. J. Anim. Sci. 95: 59–71. https://doi.org/10.2527/jas.2016.0754

Ma W., Z. Qiu, J. Song, J. Li, Q. Cheng, *et al.*, 2018 A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta 248: 1307–1318. https://doi.org/10.1007/s00425-018-2976-9

Mackay T. F. C., E. A. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10: 565–577.

Marchal A., C. D. Schlichting, R. Gobin, P. Balandier, F. Millier, *et al.*, 2019 Deciphering Hybrid Larch Reaction Norms Using Random Regression. G3 Genes Genomes Genet. 9: 21–32. https://doi.org/10.1534/g3.118.200697

Martinez-Meier A., L. Sanchez, M. Pastorino, L. Gallo, and P. Rozenberg, 2008 What is hot in tree rings? the wood density of surviving Douglas-firs to the 2003 drought and heat wave. For. Ecol. Manag. 256: 837–843.

McCullagh P., and D. Clifford, 2006 Evidence for conformal invariance of crop yields. Proc. R. Soc. Math. Phys. Eng. Sci. 462: 2119–2143. https://doi.org/10.1098/rspa.2006.1667

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.

Meyer K., and M. Kirkpatrick, 2005 Up hill, down dale: quantitative genetics of curvaceous traits. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360: 1443–1455. http://dx.doi.org/10.1098/rstb.2005.1681

Pablo Cappa E., F. Munoz, L. Sanchez, and R. J. C. Cantet, 2015 A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. Tree Genet. Genomes 11: 120. https://doi.org/10.1007/s11295-015-0917-3

Pegard M., V. Segura, F. Munoz, C. Bastien, V. Jorge, *et al.*, 2020 Favorable Conditions for Genomic Evaluation to Outperform Classical Pedigree Evaluation Highlighted by a Proof-of-Concept Study in Poplar. Front. Plant Sci. 11: 581954. https://doi.org/10.3389/fpls.2020.581954

Pérez P., and G. de los Campos, 2014 Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198: 483–495. https://doi.org/10.1534/genetics.114.164442

Petit R. J., and A. Hampe, 2006 Some evolutionary consequences of being a tree. Annu. Rev. Ecol. Evol. Syst. 37: 187–214.

Plate M., R. Bernstein, A. Hoppe, and K. Bienefeld, 2019 Comparison of infinitesimal and finite locus models for long-term breeding simulations with direct and maternal effects at the example of honeybees, (O. Rueppell, Ed.). PLOS ONE 14: e0213270. https://doi.org/10.1371/journal.pone.0213270

Resende M. D. V., M. F. R. Resende, C. P. Sansaloni, C. D. Petroli, A. A. Missiaggia, *et al.*, 2012 Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. 194: 116–128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Rincent R., E. Kuhn, H. Monod, F.-X. Oury, M. Rousset, *et al.*, 2017 Optimization of multi-environment trials for genomic selection based on crop models. Theor. Appl. Genet. 130: 1735–1752. https://doi.org/10.1007/s00122-017-2922-4

Ritchie M. D., E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, 2015 Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. 16: 85–97. https://doi.org/10.1038/nrg3868

Rogier O., A. Chateigner, S. Amanzougarene, M.-C. Lesage-Descauses, S. Balzergue, *et al.*, 2018 Accuracy of RNAseq based SNP discovery and genotyping in Populusnigra. BMC Genomics 19: 909. https://doi.org/10.1186/s12864-018-5239-z

Sanchez L., A. Yanchuk, and J. King, 2007 Gametic models for multitrait selection schemes to study variance of response and drift under adverse genetic correlations. Tree Genet. Genomes.

Schaeffer L. R., 2004 Application of random regression models in animal breeding. Livest. Prod. Sci. 86: 35–45. https://doi.org/10.1016/S0301-6226(03)00151-9

Sergent A.-S., N. Bréda, L. Sanchez, J.-C. Bastein, and P. Rozenberg, 2014 Coastal and interior Douglas-fir provenances differ in growth performance and response to drought episodes at adult age. Ann. For. Sci. 71: 709–720. https://doi.org/10.1007/s13595-014-0393-1

Tiret, M., M. Pegard, and L. Sanchez, 2021 How to reach a higher selection plateau with optimal contribution selection and compensatory mating? bioRxiv 2021.02.15.431233. https://doi.org/10.1101/2021.02.15.431233

Tyler A. L., F. W. Asselbergs, S. M. Williams, and J. H. Moore, 2009 Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays 31: 220–227. http://dx.doi.org/10.1002/bies.200800022

Wei W.-H., G. Hemani, and C. S. Haley, 2014 Detecting epistasis in human complex traits. Nat. Rev. Genet. 15: 722–733. https://doi.org/10.1038/nrg3747

Wu H. X., H. R. Hallingbäck, and L. Sánchez, 2016 Performance of Seven Tree Breeding Strategies Under Conditions of Inbreeding Depression. G3 GenesGenomesGenetics g3.115.025767. https://doi.org/10.1534/g3.115.025767

Zampieri G., S. Vijayakumar, E. Yaneske, and C. Angione, 2019 Machine and deep learning meet genome-scale metabolic modeling. PLOS Comput. Biol. 15: e1007084. https://doi.org/10.1371/journal.pcbi.1007084