Potential for Marker-Assisted Selection for forest tree breeding: lessons from 20 years of MAS in crops

Helene Muranty, Véronique V. Jorge, Catherine Bastien, Camille Lepoittevin, Laurent Bouffier, Leopoldo Sanchez Rodriguez

Potential for Marker-Assisted Selection for forest tree breeding: lessons from 20 years of MAS in crops

H. Muranty[1,2,5], V. Jorge[1], C. Bastien[1], C. Lepoittevin[3,4], L. Bouffier[3,4], L. Sanchez[1]

[1] INRA, UR0588 Amélioration Génétique et Physiologie Forestières (AGPF), F-45075 Orléans, France.

[2] present address: INRA, UMR1345 Institut de Recherche en Horticulture et Semences, F-49071 Beaucouzé, France

[3] INRA, UMR1202 BIOGECO, F-33610 Cestas, France.

[4] Univ. Bordeaux, BIOGECO, UMR1202, F-33400 Talence, France.

[5] corresponding author: Helene.Muranty@angers.inra.fr, Tel : 33 (0)2 41 22 57 88, Fax : 33 (0)2 41 22 57 55

Abstract

For the most part, molecular markers and detection of quantitative trait loci have been developed for forest tree species in view to performing marker assisted selection (MAS). However, MAS has not been applied to forest trees until now. In parallel, some success stories of MAS in crop breeding have been reported. Recently, genotyping techniques have undergone a tremendous increase in throughput, moving the trend from MAS to genomic selection. We analyzed 250 papers reporting the use of MAS in plant breeding and found that the most popular schemes used were gene pyramiding and marker-assisted backcross manipulating a single or very few genomic regions which have a major impact on crop value. We reviewed theoretical and simulation studies to identify the parametric space in which MAS is expected to bring about significant advantages over phenotypic selection. Then, we tried to explain why MAS has not been applied to forest trees and discuss the opportunities offered by recent advances in these species.

Key words:

Molecular markers, QTL, genomic selection, breeding strategies, genetic diversity management

The breeding of plant and animal species of interest to human societies is based on the use of the genetic variation that underlies phenotypic variation for traits of interest. Breeding is a sequential process that starts with the evaluation of candidates for economically important and/or adaptive traits, and continues with the selection of the best candidates. The selected individuals are then used either for mass propagation of genetic gain, or to create new candidates by recombination. Traditionally for most breeding programs, the evaluation step is performed on the basis of phenotypes obtained from a genetically structured population, where inferences can be made on the genetic value underlying each phenotype. With the development of molecular genotyping techniques, it has become possible to obtain marker data on selection candidates and their relatives used to indirectly evaluate their breeding values. Molecular markers are defined as heritable traits which enable the characterization of the underlying genotype, whatever the age at evaluation and independently of the specific environment (Prat et al. 2006). One of their major properties is that each marker is a simply inherited trait affected by only one genetic locus, without epistasis.

The use of molecular marker data to assist the breeding process is known as "marker-assisted breeding". Marker-assisted selection (MAS) is restricted to the use of markers as a substitute for or to assist genetic prediction and genetic screening. The basic idea of MAS is to exploit linkage disequilibrium (LD) between markers and quantitative trait loci (QTLs), i.e. non-random association between marker and QTL alleles. Other uses of markers in breeding like fingerprinting and pedigree verification or reconstruction do not rely on LD.

Several MAS breeding strategies have been proposed for plants, particularly those for which breeding depends on the development and use of inbred lines. For example, Hospital (2009) distinguished 5 groups of methods: (i) population screening, (ii) gene pyramiding, (iii) marker-assisted backcrossing, (iv) marker-based recurrent selection and (v) selection based on an index combining molecular and phenotypic scores. In the first three MAS methods, markers at loci of interest act directly as classification variables in the selection and screening processes. In population screening (i), genotypes of interest are identified on the basis of marker data, but the population screened is not necessarily built to maximize segregation on the basis of marker data, and the process is not recurrent. In gene pyramiding (ii), a population is recursively built from crosses to produce individuals that carry a maximum number of alleles of interest. The process can last several generations, depending on the number of alleles to accumulate, particularly if the ultimate goal is to obtain a single individual carrying all the alleles of interest. In marker-assisted backcrossing (iii), one (or a few) gene(s) from a

donor line is introgressed into the genetic background of a recipient line by repeated backcrossing. In this process, markers are used either to control the presence of the target allele or to accelerate the return to the recipient genetic background. For the last two MAS alternatives (iv and v), marker information is a variable in the prediction equations used to rank candidates. With marker-based recurrent selection (iv), also known as marker-assisted recurrent selection, breeding values are obtained solely on the basis of an explicit list of markers and when phenotypes for candidates are not yet available. With index selection (v), molecular and phenotypic data are combined to predict breeding values and to rank candidates. Marker information can be obtained long before phenotypic evaluations are available for all the candidates. In this case, selection can proceed in two separate steps, as in traditional independent culling level selection. In the first step, markers are used to determine the subset of candidates that will be subsequently evaluated by phenotyping in the second step.

In recent years techniques revealing genetic polymorphisms have evolved rapidly. This has opened up new perspectives for genome-wide screening of polygenic traits in breeding programmes, with the emergence of "genomic selection", making current efforts in QTL-based marker-assisted selection obsolete. Before we embark on this new revolution, a detailed review of what has been accomplished and what is still underway is needed. This is the intention of this paper, which gives particular emphasis to forest tree species.

The objectives of this review are (1) to review the implementations of MAS reported in different plant species; (2) to define the parametric space in which MAS can be implemented advantageously; and (3) to explain why MAS has been rarely applied to forest trees and to discuss opportunities offered by recent progress in these species.

1. When has MAS been applied in plant and trees? Lessons from the literature

Several reviews and opinion papers on MAS in plants have pointed out that only a few examples have been effectively implemented, even though thousands of QTLs have been reported in various plant species and for various traits (Bernardo 2008). Although the conditions for efficient MAS have been studied quite extensively in theory, as reviewed in the next section, several challenges must be overcome in order to use this technique operationally (Grattapaglia 2007; Hospital 2009), explaining the present lag between theory and practice.

To obtain a clearer and more objective idea of the use of MAS in plants, we undertook a meta-analysis of papers on MAS selected from two of the most popular bibliographic

databases, namely Web of Science® and CAB Abstracts®. The objective was to analyze the reasons and modalities of MAS implementation in a wide range of plant species.

1.1 Selection of scientific publications

The meta-analysis presented was derived from 750 peer reviewed papers potentially related to MAS experiments and selected on the basis of keywords and phrases describing MAS strategies (see Supplementary documents 1 and 2, and Supplementary Table 1). We searched the expression 'Marker Assisted Selection' (or its 25 derivatives) solely in the title of the articles to limit the number of hits when searching.

1.2 Meta-analysis results

A preliminary step before the meta-analysis was to classify the papers into one or several of the following five classes: "*Application*", "*Development*", "*Simulations*", "*Review*" and "*others*". In the "*others*" class, we included papers that did not deal with MAS in plants but with QTL detection, the development or validation of markers linked to QTLs or genes of interest, the evaluation of QTL effects in various environmental and genetic backgrounds, the development or review of methods for high throughput DNA extraction and genotyping, the management of genetic diversity and core-collection building, and with MAS in animals. The results of this classification are given in Table 1. The relatively high number of papers classified into "others" (about one third) reveals the misuse of the term "marker assisted selection" by some authors.

In the next step, we focused on the Application class and recorded the species, the trait(s), the MAS scheme used and the number of genomic regions involved in the MAS process.

The 250 papers concerned 37 species or groups of species of the same genus. It was sometimes difficult to identify the precise species when interspecific introgression was used (Figure 1). The species most represented was rice (74 papers), followed by wheat (37 papers) and maize (22 papers). Twenty species were represented by only one paper. The family most represented was Poaceae (159 papers), followed by Fabaceae (25 papers) and Solanaceae (22 papers). Among the 37 species or species groups, we classified ten as trees, shrubs and perennials (*Carica papaya*, *Manihot esculenta*, *Coffea arabica*, *Malus* sp., *Prunus* sp., *Pyrus pyrifolia*, *Rubus idaeus*, *Rosa* sp., *Simmondsia chinensis*, *Vitis vinifera*) but none was a forest tree. The tree or perennial class was represented in 20 papers among which 7 concerned the apple tree.

We classified the application papers in seven broad categories on the basis of trait(s) targeted by MAS (Figure 2a). The category most represented was "biotic stress resistance", with more than half the studies, followed by product quality (around 20%). Most often, MAS concerned a single trait or closely related traits.

The application papers were also sorted according to the five categories of MAS strategies defined above (Figure 2b). The most common scheme was marker-assisted backcross, whereas the use of a molecular and phenotypic score was very rarely employed. In 14 papers, two kinds of schemes were either compared to each other or used one after the other.

We were able to identify the number of genomic regions targeted by MAS in 194 papers, either as the number of genes or QTLs or as the number of markers used in selection. Whenever the position in the genome of the markers used in selection was given, we counted as the same genomic region, markers located less than 20 cM apart from each other. The number of genomic regions targeted by MAS varied between 1 and 25, but most often, very few genomic regions were targeted, with 57% of the papers targeting one or two genomic regions (Figure 3). Only 6 papers comprised MAS targeting 10 or more genomic regions. The highest number of genomic regions targeted by MAS (25) was found in the study by Mayor and Bernardo (2009), which aimed at comparing 5 molecular and phenotypic scores in a multi-trait selection context.

1.3 Motivation for implementing MAS

Motivation for using MAS was explicitly expressed in only 66 of the 250 application papers. These motivations fall within one of the following four categories: 1) reduction of the difficulty and/or the cost of phenotypic evaluation (including problems linked to GxE interaction and low heritability); 2) gain in terms of time; 3) the possibility of applying greater selection intensity; and 4) the increased precision of genetic evaluation. These benefits were not quantified in most cases.

When MAS targets a biotic stress resistance trait, motivations most often fall within categories 1 and 2. Markers enable piling up over several recombination cycles (gene pyramiding) several resistance genes controlling qualitative resistance in a single genotype, which is difficult to obtain by conventional phenotyping methods. Indeed, it would involve the evaluation of plants with several pathogen strains/isolates with known virulence by procedures that are usually extremely labour-intensive, while it is not always possible to find differential pathogen strains to distinguish the resistance genes (Vida et al. 2009). Moreover,

some interesting resistance genes are recessive, e.g. *xa13* for bacterial blight resistance in rice, and co-dominant markers linked to this gene enable the identification of heterozygous plants without an additional generation of selfing needed for resistance testing (Basavaraj et al. 2010). Quantitative resistance, which often hints at epistatic QTLs, can be evaluated and monitored more easily with molecular markers than with phenotypic tests (Ahmadi et al. 2001). Finally, DNA-marker assays can be non-destructive, which is not always the case for phenotypic resistance tests (Perumalsamy et al. 2010). Thus, resistance gene pyramiding is one of the "success stories" of MAS. Susceptibility to plant pathogens has a significant impact on the economics of crop production and selection for resistance can add value to the products of breeding programs.

Another example of a motivation falling into categories 1 and 3 is when MAS targets a product quality trait, like 'basmati' specific features in rice (Myint et al. 2009), absence of anti-nutritional factors in soybean seeds (Alves de Moraes et al. 2006), protein quality in maize (Babu et al. 2005) and malt quality in barley (Igartua et al. 2000). The advantage of using markers associated with the trait of interest is to avoid a phenotypic evaluation that often needs a large number of seeds. The relatively high costs of quality assessment tests set limits to sample sizes that can be phenotypically evaluated, and thus limit the genetic gain that can be obtained with phenotypic selection (Han et al. 1997). Crop quality traits are also often of great importance to create commercial value in crop products.

Traits related to abiotic stress tolerance like nitrogen use efficiency (Dolstra et al. 2007), boron tolerance (Emebiri et al. 2009b) and drought tolerance that can involve several features such as osmotic adjustment (Levi et al. 2009) and root traits (Steele et al. 2006) also require laborious and time-consuming evaluation procedures, thereby making them very attractive for MAS. In these cases, the motivation for using MAS again belongs to category 1.

A huge effort has been made to identify QTLs controlling yield genetic variation, a typical complex trait with moderate to low heritability, but few applications in MAS have been described. We found only 29 papers where yield or its components were targeted by MAS. A specific motivation for using MAS was not explicit in these papers and could be simply described as "it improves the efficiency of selection in crop improvement" (Mahmood et al. 2005).

Gain of time and precision (categories 2 and 4) are also expected from MAS in backcross breeding, where the objective is to expedite recovery of the recurrent parent genome: the use of background markers can reduce to two or three generations the time necessary to obtain

introgressed lines (Tanksley et al. 1989). MAS highlights the use of specific genetic variation present in non-domesticated populations and in related species.

To sum up, the motivation for using MAS in plant breeding is most often related to foreground selection in which the breeder selects plants with marker alleles of the donor parent at the target loci while avoiding costly phenotypic assays (Hospital and Charcosset 1997). The most popular MAS methods (marker-assisted backcrossing and gene pyramiding) are inadequate for handling economic target traits controlled by many small-effect genes, such as most yield traits in crops and in forest trees. Moreover, these strategies require several generations to obtain satisfying recombinants, which make them of limited interest for forest trees which generally reach sexual maturity late. Consequently, we will concentrate on the conditions for using MAS advantageously in the form of marker-based recurrent selection or selection on an index combining molecular and phenotypic scores.

2. Parametric space of interest for the beneficial implementation of MAS

The theoretical bases of MAS have been studied comprehensively since the seminal paper of Fernando and Grossman (1989). In this section, we review more recent theoretical studies to draw a general picture of the optimal conditions where MAS is expected to bring significant benefits over phenotypic selection.

The theoretical genetic responses to MAS vary noticeably between studies. One of the main reasons is that there are no universal assumptions concerning underlying QTL effects, levels of genetic variation or breeding scenarios. Many of these theoretical studies were conceived with a livestock or annual crop species in mind. Only a few studies have evaluated the prospects of MAS for perennial crops (Kumar and Garrick 2001; Wong and Bernardo 2008), of which only the former corresponds to a forest tree species. Grattapaglia and Kirst (2008) reviewed in detail some of the pros and cons of MAS for *Eucalyptus*.

Predictions of the benefits of MAS are generally based on simulations that assume a mixed inheritance for the phenotypic variation (Goddard 2001). It is denoted as “mixed” because true breeding values are the result of the effects of one or a few identified biallelic QTLs and a polygenic background of unidentified loci that are frequently modelled as a quantitative deviate from a normal distribution of known mean and variance.

At the risk of oversimplification, we will consider only a few of the key factors in the literature relevant to the efficiency of MAS. These factors can be classified into: i) population

size and recruitment for selection; ii) level of variation of the targeted traits; iii) genetic architecture of the targeted traits; and iv) marker density.

2.1. Population size and recruitment factors

Two different populations of interest can be defined: (i) the population composed of related individuals which have been genotyped and phenotyped in order to identify the positions and effects of the QTLs and from which the effects of the selected markers will be derived, and (ii) the population composed of the candidate individuals on which selection will be carried out either on markers solely or on both markers and phenotypic evaluations. Unless otherwise stated, simulation studies reported in this review considered the population size (N) of the former population. Lande and Thompson (1990) studied MAS efficiency using a deterministic approach, assuming an infinite N. Under this hypothesis, the error in the estimation of QTL effects leading to bias in the estimation of MAS efficiency is negligible. However, sample finiteness is a crucial assumption. Moreau et al. (1998) evaluated MAS efficiency using analytical approaches with validation by simulation with finite N. The relative efficiency of MAS with respect to phenotypic selection increases with N, with already detectable nonzero relative efficiencies (>12%) in the range of N = 200~500. Increasing N increases the power of detection of underlying QTLs and the accuracy of the estimation of the marker effects, and thus favors MAS efficiency. Xu (2003) showed that decreasing N below 500~1000 leads to fewer detections of QTLs and an upward bias in the estimation of the effects of the detected QTLs. However, these previous studies assumed a classical single cross between two homozygous lines. When assuming outcrossing populations, minimum N requirements are expected to be raised, as pointed out by Grattapaglia and Kirst (2008). One example in Hayes et al. (2007) for abalone breeding showed detectable improvements of genetic gain over standard methods by 5 to 15%, with N varying between 1000 and 1500 individuals grouped into 20 unrelated full-sib families. Increasing family size favors accuracy in the estimation of QTL effects, as more observations per QTL can be used.

The effect of varying *N* within attainable values in forestry (in the range 200~8000) has been considered among other key factors by Grattapaglia and Resende (2011). Although the study concerned genomic selection, some of the simplest scenarios were close to MAS. Increasing *N* leads to step improvements in accuracy whenever *N* is beyond 1000, with levels similar to classical evaluation with *N* equal to 2000. These values are intended for breeding populations with a relatively low effective population size of less than 30. Doubling the effective

population size (60) would require a roughly six-fold *N* to reach accuracies similar to those of the best classic BLUP evaluation. This raises another aspect related to population size and recruitment factors, which is that of effective population size (*Ne*). The population under evaluation most often comprises related groups, so that $N >> Ne$. In general, with decreasing *Ne*, the LD between markers and QTLs is expected to increase and therefore enhance the effectiveness of marker-based selection. Twenty-one studies implementing a molecular score and/or a molecular and phenotypic score in operational situations are comparable to the simulation studies. But we could find the size of the population used to detect QTLs and estimate their effects in only 4 of them (Flint-Garcia et al. 2003a; Han et al. 1997; Mahmood et al. 2005; Moreau et al. 2004), and it varied between 92 and more than 300 in populations derived from crosses between inbred lines. The studies with the lowest population sizes were at the lower bound of what theoretical studies recommend as optimal.

2.2 Level of variation for the targeted traits

Heritability (h^2) is an expectation derived at a population level expressing how a phenotype is a good predictor of the underlying breeding value (see Visscher et al. 2008). Most often, in simulation studies on MAS, the only genetic component of variation of use is the additive variance, which is the numerator of the narrow-sense heritability. As h^2 increases, the benefits of using markers to infer genotypic values decreases, limiting MAS to traits with low h^2. Most of the simulation studies covered here assumed a single value for h^2 within the range of 0.025~0.35 and with an average of 0.24. Not surprisingly, this value is representative of many quantitative and complex traits.

Over a full range of h^2 (0.05~0.95), Moreau et al. (1998) found that the domains with the highest efficiency for MAS correspond to the lowest third of the h^2 range. However, with very low h^2, QTL effects are poorly estimated, which reduces MAS efficiency. With higher h^2, MAS tends to perform like phenotypic selection. An optimal range of h^2, between 0.15~0.2, and confirmed by other studies, leads to the highest efficiencies for MAS compared to phenotypic selection (Dekkers 2007; Grattapaglia and Resende 2011). Thanks to an increase in statistical power, this optimal h^2 point decreases with increasing N (Moreau et al. 1998). Ollivier (1998) and Xie and Xu (1998) obtained a similar range of optimal h^2 (0.1~0.3) in a within-family and a two-stage MAS scheme respectively. With a multiparental mating design similar to a partial factorial mating design, Blanc et al. (2008) confirmed previous findings with medium to low heritabilities ($h^2 \leq 0.25$) resulting in the highest efficiencies for MAS over

phenotypic selection at a given time horizon, leading to an additional gain from 25% up to 150%. Among the 21 studies reporting the implementation of MAS with a molecular score or a phenotypic and molecular score, we found an indication of the estimated heritability of traits under selection in only 10 publications, and its range (0.07 to 0.9) was too wide to draw conclusions.

2.3 Genetic architecture for the targeted traits

QTL numbers and *QTL effects on the targeted trait* reflect the extent to which we know the underlying genetic model. In addition to and associated with the numbers and effects of QTLs, the percentage of trait variation accounted for by the known QTLs is relevant for MAS. Most MAS schemes assumed in theoretical studies comprise a single QTL or a small number of them, often with equal effects. This simple genetic architecture reflects the limits in QTL detection success for most quantitative traits at the time of the studies, in which the detected QTLs corresponded to those with large and purely additive effects. The traits for which variation is due to few genes with large effects would promptly respond to directional selection. When genes are few, each of them would take a larger share of selection pressure than what would happen under a multigenic scheme, correspondingly leading to faster changes in gene frequencies.

This reasoning was demonstrated in the first simulation studies on MAS that considered, to a limited extent, a variation in QTL numbers between alternative scenarios. Edwards and Page (1994) compared MAS with 10 and 25 QTLs having equal effects, with the result that the responses with the 10 QTLs scenario were 5 to 20% faster. Similar trends were shown by Bernardo and Charcosset (2006) for scenarios with 10, 40 and 100 QTLs. Moreau et al. (1998) also explored different numbers of QTLs but sampled them in geometric distributions. For a given number of QTLs, a scenario with even QTL effects led to lower MAS efficiency compared to QTL effects in geometric distributions, with differences being particularly noticeable within an h^2 range of 0.15~0.2. With a pedigree-based dairy-cattle dataset, Guillaume et al. (2008) found that MAS was more beneficial when there are fewer QTLs with large effects, with situations close to the infinitesimal model being less favorable for MAS. Geometrical or *L-shape* distributions of QTL effects are thought to reflect what is observed for many quantitative traits (Kearsey and Farquhar 1998; Bernardo 2008), in livestock (Silva et al. 2011), and in crops (Truntzler et al. 2010).

MAS efficiency greatly depends on the extent to which detected QTLs have a large effect and are able to explain the variation of the trait. In the classical paper by Lande and Thompson (1990), the dependencies of MAS efficiency on sample size, heritability and percentage of explained variation by selected QTLs were mathematically derived in the context of MAS. Increasing the percentage of variation accounted for by known QTLs leads to increasing MAS efficiency, with the dependency between these two parameters becoming greater as heritability decreases. To maintain a given level of MAS efficiency when the percentage of variation accounted for by known QTLs is low, it is necessary to increase the size of the sample. Numerical examples were provided by Moreau et al. (1998), where MAS efficiency for a trait of intermediate heritability ($h^2 = 0.3$) increased from 10 to 30% when the percentage of explained variation by a QTL rose from 30 to 50%.

To sum up, MAS efficiency is highest when selection targets one or a small number of QTLs of large effect and that account for a large percentage of the variation in the selected trait, ideally 30% or more. High numbers of QTLs make MAS inefficient particularly when segregating populations are of small size and favorable alleles are at low frequencies. Optimal selection strategies dealing with multiple QTLs have been proposed, mostly based on the explicit management of selection intensities (Chakraborty et al. 2002; Sánchez et al. 2006), which facilitate the implementation of MAS recurrently in complex scenarios.

In 57% of the 250 application papers of our survey, the number of genomic regions targeted by MAS was one or two, which, according to simulations, facilitates the successful implementation of MAS. The most frequent MAS scheme behind such low QTL numbers was gene pyramiding or marker-assisted backcross while most of the simulation works assumed a molecular and phenotypic score scheme.

2.4 Marker density

MAS efficiency depends on the use of molecular markers to tag the QTLs during the selection process and particularly on the precision of QTL positions. This tagging process can be made more effective when genomic regions of interest are sufficiently covered by evenly spaced markers, with a given *marker density* (markers / cM, markers / base pairs [bp]).

Increasing marker densities in MAS is expected to increase the likelihood of tagging close to a causal mutation and thus make the association between markers and QTLs more persistent at the population level. This association is defined as *linkage disequilibrium (LD)*, which exists in natural populations and can be generated by hybridization between genetically

differentiated parents. LD is continually decreased by recombination. As a consequence of LD, markers are able to capture a portion of the quantitative variation that is accounted for by their neighboring QTLs. Therefore, MAS efficiency is based on the density of markers at the vicinity of relevant QTLs and the extent of LD between markers and causal QTLs. Both factors have been extensively considered in many simulation and theoretical studies on MAS but the amount of LD that is captured through a given level of marker density is population-specific. Although it is assumed that LD declines with distance, this decline is known to be different within regions of the genome, between populations due to distinctive population histories and among species (Flint-Garcia et al. 2003b; Gupta et al. 2005; Sorkheh et al. 2008).

The key element when tagging already detected QTLs is their recombination distance to the closest flanking markers or the flanking bracket size, which is the interval in cM between markers containing a QTL. The shorter this bracket size the greater the potential superiority of using markers as proxies of QTLs in selection. Spelman and Bovenhuis (1998) considered a simple setup with two QTLs representing between 5 and 10% of phenotypic variation, with the rest coming from an infinitesimal polygenic model, and a nucleus of a thousand individuals structured into a close hierarchical family system. Decreasing the flanking bracket size of 15 cM down to 2 cM could result in a substantial gain of up to 30% for the first MAS generation, with better capture of the polygenic response in later generations for the narrower bracket size. In the best case of 2 cM, the advantage of MAS over a conventional system was between 5 and 15%. Bracket sizes down to 0.1 cM were considered by Villanueva et al. (2002) who assessed the benefits of recurrent MAS over a range of different parameters. At the smaller bracket size, the benefits over phenotypic selection were slightly lower than those obtained by Spelman and Bovenhuis (1998), because smaller population sizes, lower heritabilities and lower QTL effects were assumed. Three to four extra cycles were needed to attain the maximum accumulated benefit of 15%. Nevertheless, in the best case, benefits were still far from the theoretical maximum attainable by selecting directly on the QTL (20%), implying that there is room for improvement at higher densities. However, extra gains should be balanced against the cost of achieving proximity to the causal QTL by marker densification.

The benefits of high densities are seen more clearly when assuming very large populations involving several families, as illustrated by Guillaume et al. (2008) in a dairy cattle dataset. They found a substantial gain in reliability, by replacing microsatellite markers by ten SNPs

within a 1cM bracket around the QTL, from 43% to 79% in comparison to classical selection. Here, reliability refers to the degree to which QTL transition and positioning is followed across the population, which is relevant to the selection accuracy of MAS. At the other extreme, sparse marker density or extensive linkage disequilibrium might lead to MAS inefficiency through uncertainties in the QTL effect estimates and inaccurate QTL positions. Mapping precision has been greatly improved by the combination of findings from linkage studies with QTL congruency investigations. Congruency studies result in positioning markers and QTLs on single consensus maps through the use of meta-analytical approaches. Tools have been developed based on this principle (e.g. Veyrieras et al. 2007). Although the impact of consensus maps on MAS efficiency has not yet been fully evaluated, it is reasonable to assume that selection accuracy would be improved and therefore benefit MAS.

At the risk of oversimplification, we conclude this section by assuming three basic scenarios. Within the first scenario, MAS is based on QTLs detected for the first time, with large confidence intervals of more than 10 cM. In this situation, MAS is not expected to be of benefit in comparison to classical alternatives, unless the targeted QTLs have very large effects, selection is applied over one cycle and candidates result from a single crossing event. The second scenario is applied when fine-mapping over one or several lineages leads to confidence intervals of 1 to 5 cM. This situation is the most favorable for MAS according to theoretical predictions and is currently feasible in many species. In our MAS literature survey, we observed that the main plant species for which MAS is already in use meet these requirements: rice, wheat, maize, barley and soybean linkage maps are very dense (one marker every 2.2 cM or less, Harushima et al. 1998; Hyten et al. 2010; Somers et al. 2004; Szücs et al. 2009; Varshney et al. 2007) and close linkages have been observed between QTLs of interest and adjacent markers (Emebiri et al. 2009a; Jena and Mackill 2008; Kuchel et al. 2005; Prasanna et al. 2010; Sebastian et al. 2010). The third scenario is applied after positional cloning. This will provide the upper bound in MAS efficiency, although at a very high cost in resources, allowing implementation over extended pedigrees and over several selection cycles. In our meta-analysis, we found several papers using a gene sequence-based marker for *Xa21*, a previously cloned rice gene for bacterial blight resistance (Basavaraj et al. 2010; Bhatia et al. 2011; Chen et al. 2001; Chen et al. 2000; Huang et al. 1997)

2.5 Additional parametric and time considerations in MAS efficiency

We reviewed some of the key parameters and their optimal ranges for efficient MAS: a population size of at least 1000 evaluations, selected traits of low to intermediate heritability, simple genetic architecture, and targeted QTLs flanked by marker brackets of less than a few cMs. This parametric exercise helps to set a *baseline* beyond which MAS is expected to be a breeding alternative as efficient as that of classical schemes. However, such a reduced set of parameters may lead to an oversimplified and even optimistic view of MAS. For instance, most MAS evaluations consider QTLs with pure additive effects (no dominance or epistasis). Non-additive genetic components may be of relevance for some traits and be successfully exploited through clonal selection. Although difficult to put into practice, unfavorable epistatic interactions between QTLs can eventually be dissected and broken down by the use of mapping and MAS (Causse et al. 2007; Knaap et al. 2002). Often, simulation studies consider even distributions of LD over the genome and even coverage of markers and QTLs. Such scenario could be too optimistic, particularly when breeding concerns specific genomic regions with close clusters of relevant QTLs (Mayor and Bernardo 2009). Finally, breeding often involves several traits of interest, which might present adverse genetic correlations that impair simultaneous gains. The latter situation was not considered explicitly when evaluating MAS by simulations in previous studies, although multivariate approaches of MAS have been proposed by Dekkers et al (2002) and Chakraborty et al (2002).

Another factor of great importance is MAS efficiency per unit time. The use of markers has the potential to shorten the evaluation lag, for instance, in traits expressed only at mature ages, like grain yield, fruit quality traits in fruit trees or in traits with low juvenile-mature correlations such as production traits in forestry. Although selection accuracy can be affected by the use of molecular scores, the resulting selection could be very effective both cost-wise and time-wise. For species with long breeding cycles, such as those in tree breeding, this time factor can easily become very important.

3. The past and future of marker assisted breeding in forest tree species

3.1 Motivation to apply MAS in trees

Genetic gains to increase the quantity and the quality of wood products have been achieved by conventional breeding in advanced breeding programs. Compared to traditional breeding, MAS for forest trees appears particularly attractive for four main reasons.

The first one is the possibility of early selection, which could highly reduce selection age. Two biological aspects of most forest trees are the fact that the full expression of economically important traits generally occurs late and that this late expression often shows poor correlations with the corresponding juvenile traits. Consequently, selection is currently performed at about one third of the rotation age (which means 2-3 years for *Eucalyptus* to 6-12 years for *Pinus* spp.). However, the first step of selection on the basis of markers alone could be applied to seedlings as soon as material can be collected for DNA extraction, provided that reliable associations between markers and QTLs have been established previously.

The second reason is the limitation of phenotyping costs. Trees are big organisms requiring large areas and a high number of replicates to manage stand development and periodical thinning operations. Moreover, some important traits in forest tree breeding are highly expensive to evaluate (wood quality) or / and difficult to obtain under natural conditions (pest resistance, drought tolerance). These phenotyping activities rely heavily on human labor, whose costs are not expected to drop. Genotyping, on the other hand, tends to get cheaper or, at least, to comprise larger numbers of assays per monetary unit. However, rigorous cost efficiency evaluations for MAS in forest tree breeding are still lacking.

The third reason is the possibility of evaluating the genotypes of the breeding population more accurately by combining marker and phenotype information. This is of particular relevance for traits with low heritability, especially traits related to tree growth, like height and diameter (Cornelius 1994), and for new target traits for which phenotypic evaluations cannot be obtained with the required accuracy or population coverage within a reasonable time. Additionally, experimental tests in forestry tend to cover large areas, spreading most often over heterogeneous environments, which tend to decrease the signal to noise ratio and, therefore, the heritability estimates for traits affected by environmental variation. Markers also offer the invaluable possibility of tracking minor resistance genes, whose phenotypic expression can be masked by the epistatic effects of major resistance genes, so that genotypes with more durable resistance can be developed.

The fourth reason is the possibility offered only by the use of molecular markers of monitoring inbreeding at the individual level and managing explicitly genetic diversity at the population level. Indeed, most commercially important forest tree species are still undomesticated, comprising huge outcrossing populations that harbor large levels of genetic diversity. Only a fraction of this genetic diversity has been *captured* in pedigrees, associated

with markers and evaluated, the *rest* being background diversity with unknown but potentially useful functions. Markers can be used to manage part of this background resource while proceeding with breeding, for instance, in a way similar to the optimum contribution selection proposed by Nielsen et al (2011).

3.2 Understanding the lack of MAS implementations in forest trees

We could not find any paper describing the implementation of MAS in forest trees. Xu and Crouch (2008) also reported that the annual number of articles on plants with the term "marker-assisted selection" has consistently lagged behind the number of articles with the term "quantitative trait locus". The first hypothesis is that some MAS applications have been developed with no publication associated, for example in breeding programs managed by private companies. A second hypothesis is that until now various constraints have limited the application of MAS in plants. For forest trees, constraints relating to the development of MAS can be biological, socio-economical or technical.

3.2.1 Biological issues

Forest trees as well as most fruit trees reach sexual maturity quite late, compared to most domesticated crops and livestock. This long lag prevents breeders from gathering the full benefits from early marker selection and excludes multi-generation MAS schemes like gene pyramiding and marker-assisted backcross, i.e. those that are the most popular for crops. This is probably why one of the most advanced examples of MAS in apple trees is based on a transgenic early flowering line making it possible to obtain one generation in roughly one year under controlled conditions (Flachowsky et al. 2011). The first steps toward genomic selection were reported in tree species for which the generation lag can be shortened by artificial means: top-grafting in loblolly pine (Resende Jr et al. 2012b) and growth regulator application in Eucalyptus (Grattapaglia 2007). When generation lag cannot be shortened beyond biological constraints, markers can still be used in a multistage selection scheme, with the first step of selection on markers only identifying trees allowed to reach sexual maturity. This initial step is completed by the second step which consists in selecting on molecular and phenotypic scores based on phenotypic tests performed on these trees just before flowering.

Breeding in forest trees differs from that of many crops due to the early stage of domestication in the former. Although this can be an advantage in terms of the genetic variation available for selection, it presents the drawback of low levels of LD (Neale and

Savolainen 2004) and large effective population sizes in base breeding populations, both factors being disadvantages for MAS.

Due to these limitations, MAS is expected to be efficient after considerable narrowing of genetic diversity (founding effect) which would increase linkage disequilibrium. With the marker densities that were reachable until recently, MAS would have been valuable for selection within full-sib or half-sib families. This can be adequate for species for which clonal deployment obtained by cuttings is common (*Populus* spp. and *Eucalyptus* spp.) Indeed, this would involve the selection of a few promising families in which to assess markers – QTL associations for use in MAS. On the other hand, MAS would be of lesser interest for most pine species for which deployment is mainly based on varieties with a large genetic base (synthetic varieties produced in open-pollinated seed orchards comprising 30-40 unrelated genotypes), because it would entail determining marker – QTL associations in 30-40 pedigrees. A large genetic base of deployed varieties is a necessary precaution to reduce the economic and ecological risks associated with long rotation age, and to maintain the stability and resilience of production populations under varying environmental conditions.

3.2.2 Socioeconomic issues

In most of the papers in the literature review, the cost/benefit advantage of MAS was absent. Morris et al. (2003) carried out one of the few cost/benefit analyses comparing conventional and MAS methods for a particular breeding application in a specific socioeconomic environment. They found that neither method showed clear superiority in terms of both cost and speed and that the optimal choice depends on the availability of operating capital. A few studies carried out an economic evaluation of MAS for forest trees (Johnson et al. 2000; Plomion et al. 1996; Wilcox et al. 2001). However, these studies limited the use of markers to a narrow within-family context, and did not provide a clear demonstration of MAS's profitability compared to conventional breeding. Moreover, none of these studies considered the difference between the scenarios regarding the management of genetic diversity, whereas it can affect future genetic gains.

This lack of sound cost/benefit analyses is one of the issues preventing the implementation of MAS. This can be at least partially explained by the difficulty of the task, given the number of parameters to be taken into consideration. These parameters are highly dependent on the species, and phenotyping and genotyping costs vary significantly over time and between programs, which makes generalizations difficult. Moreover, the increases in genetic gains

expected from MAS must offset the investments required to develop the genomic resources necessary, although these investments far precede the possible genetic gains. The unpredictability of the economic context over several decades in a typical rotation period in forestry is another difficulty in this analysis.

On top of these difficulties, the forest tree research and breeding community is smaller than that of crops, and its limited financial and human resources are spread over a large number of species. Moreover, the community more directly concerned by the development of MAS in forest trees is somehow divided into breeders and geneticists on one side and genomic researchers on the other side, with this division often leading to competition for funding resources and a lack of awareness of each other's scientific challenges. There are a few exceptions, for example the two European initiatives NovelTree (http://www.noveltree.eu/) and ProCoGen (http://bfw.ac.at/rz/bfwcms2.web?dok=9020) and the American initiative Conifer Translational Genomics Network (https://dendrome.ucdavis.edu/ctgn/). In these three initiatives, breeders, geneticists and genomic researchers are well integrated in interdisciplinary MAS projects.

3.2.2 Technical issues

3.2.2.1 Genomic resources

The first technical issue with MAS implementation is the availability of markers to ensure good genome coverage, with adequate polymorphisms and transferability between pedigrees, and the availability of genetic or physical maps. Due to the large genomes in conifers, genomic resources are still a limitation for the development of MAS.

Adequate marker numbers and adequate polymorphisms are those that enable placing QTLs in marker brackets less than a few cM, which is one of the minimum requirements for MAS efficiency. Until now, the densities of markers have remained relatively low on tree genetic maps. In a survey of 48 papers concerning 11 species, we found that the average distance between markers varied between 2.7 and 17.6 cM without any mode in the distribution (Supplementary figure 1). These densities do not reach the optima described for MAS.

Transferability is also a concern in taxa with poor genomic resources. By transferability we mean that markers genotyped in various populations represent the same genomic region in each population. This is a condition for validating QTLs by comparative mapping between populations within a species or even between species. It is based on the sequence data behind the markers i.e. primer or probe sequences (SSR, short sequence repeat; SNP, single

nucleotide polymorphism; DArT, diversity array technology). Anonymous markers obtained by RAPD (randomly amplified polymorphic DNA) and AFLP (amplified fragment length polymorphism) techniques generally lack transferability and, unfortunately, most of the pioneering QTL detection studies in forest trees have used these latter markers (see Annex table 7 in Prat et al. 2006).

Although the availability of genomic resources seems essential for implementing a MAS program, we nonetheless identified a few examples where MAS could be used without the costly and time-consuming development of linkage maps: bulk segregant analysis has been used successfully to detect sex-specific RAPD and SSR markers in dioecious plants (Gill et al. 1998; Parasnis et al. 1999; Sharma et al. 2008; Silva et al. 2007). RAPD and their derived SCAR (sequence-characterized amplified regions) markers have been used to predict quantitative traits such as polyphenol content in the aerial parts of the medicinal plant *Echinacea purpurea* (Chen et al. 2009), sprouting resistance in rye (Twardowska et al. 2005), neutral detergent fiber fraction in smooth bromegrass (Stendal et al. 2006) and oleic acid content in spring turnip rape (Tanhuanpää and Vilkki 1999). Applications of MAS in species for which genomic resources are still scarce are however limited (only 8 examples out of 250 papers reporting the application of MAS in our survey).

3.2.2.2 Genetic variation explained by DNA markers

The second technical issue involved in implementing MAS is the availability of DNA markers explaining a high proportion of genetic variation for traits under selection. The pedigree-based QTL approach allows the detection of chromosome regions related to genetic variation within the family studied. More recently, population-level association mapping revealed alleles of genes (SNPs) linked to the phenotype.

We report below several QTLs (Table 2) and SNPs (Table 3) related to major selection traits detected in the main forest tree taxa for which breeding programs are implemented: *Eucalyptus*, major conifers (*Pinus*, *Picea*, *Pseudotsuga*, *Cryptomeria*) and *Populus*.

In QTL studies, the genetic diversity explored is very narrow as few pedigrees were studied, due to the cost of managing and evaluating such populations. Moreover, the pedigrees used for QTL mapping studies were sometimes chosen to maximize genetic variation in the targeted trait, for example by selecting a pedigree whose grandparental pairs displayed divergent values for the trait in focus (Jermstad et al. 2001; Sewell et al. 2000). This kind of pedigree is rarely interesting in a breeding context. Cost is also a constraint for the size of

families used for QTL mapping, reducing the accuracy of QTL detection and overestimating the variation explained. The proportion of variance explained by each QTL is generally quite low: in a sample of 17 papers given in Table 2, it varied between 0.03 to 36.4%, with most values being between 4 and 12%. The number of QTLs reported in each study can sometimes be considered as low (2 to 6 for growth in *P. taeda*, 5 for growth in *P. menziesii*). If these are summed over studies for a given species or genus, the number can become quite large (603 for growth in *Populus*), but we did not attempt to analyze the congruency of these QTLs. We suspect that traits of interest in forest trees are not in the favorable situation of a simple genetic architecture. Due to the limited pedigree sizes and low proportion of variance explained, the confidence interval of the positions of QTLs were generally quite large: we calculated the length of this confidence interval using the formula of Darvasi and Soller (1997) in the same set of 17 papers and found that it varied between 9.3 and more than 200 cM, most values being between 20 and 60 cM.

A way to explore more genetic diversity was to perform association studies, but only few and small effect QTLs have been reported until now (Table 3). The low number of associations reported is probably a consequence of the low genome coverage obtained in these studies focusing on candidate genes, whereas the small size of QTL effects reflects the genetic architecture of the traits studied in populations with large genetic variation. Additionally, few studies reported validations of detected QTLs (e.g. Devey et al. 2004). Thus most species were in the first scenario of QTLs detected for the first time and in which MAS was not expected to be of benefit compared with classical schemes (see above).

3.2.2.3 Phenotyping constraints

Now that genotyping costs are decreasing, phenotyping costs are becoming the main limiting factor for population size, particularly when addressing new traits like ecophysiological and functional traits or new wood-related traits. These new traits have come to the forefront to take into account climate change and the foreseen increased use of tree biomass as a biofuel source. Moreover, to take into account climate change, genotype by environment interaction is attracting growing interest whereas previously it was considered as a nuisance. In order to predict the behavior of genotypes in a future climate, evaluation trials should be established in a large number and range of environments. This limitation points out the need to develop high throughput phenotyping methods to analyse large populations in several environments to detect QTLs.

3.3 Prospects in the short-term

3.3.1 Uses of molecular markers for purposes other than MAS

Some marker-based tools can already be used in the management of breeding populations. These tools essentially use molecular marker polymorphisms and not their positions (Step1 and 2 in Fig. 4). One of these uses is fingerprinting: markers are used to assess or control the genetic identity of individuals. This is of great interest to check controlled crosses in an experimental design, to avoid mislabelling when handling large numbers of genotypes or to monitor the deployment of improved material.

Another use of marker polymorphism is paternity and maternity analyses. Full pedigree information enables reliable estimates of genetic parameters and breeding values from phenotypic tests (El-Kassaby et al. 2011). The traditional way of obtaining this information is by making controlled paired-crosses, which is often costly, requiring highly experienced staff. There are cheaper and simpler alternatives, either without the use of controlled crosses like open (unrecorded) pollination, or with man-made crosses but with a mixture of known pollen donors (polymix breeding). For these two simpler alternatives, pedigree information can be retrieved *a posteriori* with the help of DNA markers, sometimes at lower costs than by making controlled crosses. This is one of the basic ideas of the "breeding-without-breeding" strategy proposed by El-Kassaby and Lstiburek (2009).

Full pedigree information is also extremely valuable to maintain genetic diversity and manage inbreeding. The objective of genetic diversity management is to capture genetic gain and simultaneously maintain genetic variability along the selection process in the breeding population, which should safeguard future genetic gains (Burdon and Wilcox 2007). Even when pedigree information is known, DNA markers can give a much more precise picture of genetic relatedness, for example within a full-sib family, where sib-pairs depart from expected sib similarity due to Mendelian sampling and linkage (Hill and Weir 2011). This can be used to improve a pedigree-based BLUP evaluation, by using a marker-based realized relationship matrix instead of the pedigree-based relationship matrix (VanRaden 2008). This process is known as GBLUP evaluation, which is a possible form of genomic selection. A marker-based relationship matrix can also be the basis for estimating genetic parameters such as heritability in wild populations without recorded pedigrees (Sillanpää 2011).

Even if most of these previous marker-based tools do not fall exactly within previous definitions of marker-assisted selection, they could be considered as a first step to introduce

markers in forest breeder's everyday life with a real prospect of higher genetic gains and better diversity management. We suggest that these marker-based tools might help breeders and genomic researchers to get closer and pave the way for easier MAS integration in the near future. Indeed, the tools included in step 1 of molecular marker use (fingerprinting, genealogy control, Fig. 4) are already available for several species (Massah et al. 2010; Plomion et al. 2001; Ribeiro et al. 2011; Van de Wen and McNicol 1995), while those needed for step 2 (genetic drift control, GBLUP evaluation) are only prospects for most forest tree species.

3.3.2 Redefinition of breeding populations

Many of the most advanced forest tree breeding programs involve a few hundred elite genotypes being used in evaluation and deployment stages. For instance, for the Swedish Scots pine breeding program, the target is to keep effective population size to not less than 50 for about 20 unrelated populations across the country (Rosvall 2011). Another example in the *Pinus pinaster* breeding program in France shows effective population sizes above 130 (A. Raffin, personal communication). These relatively high numbers are mostly intended to cover the high levels of genetic variation present in conifer populations and, to a certain degree, the genetic structure of original populations in the natural distribution. These high numbers also respond to the need for deployed varieties with limited relatedness. In any case, these sizes represent a challenge for MAS implementation. Gains are to be expected from scaling up from these preliminary experiments, but current breeding sizes are still too big for cost effective marker approaches. Therefore an effort must be made to redefine elite populations with a narrower census while maintaining genetic representativeness. Moreover, such compact elite populations could allow the concentration of phenotyping efforts over reasonable numbers of candidates, to explore variation of new traits and provide training populations for QTL detection.

3.3.3 Development of genomic tools

DNA sequencing is one of the better ways to obtain genomic resources useful for marker development. In the past ten years, sequencing costs have been reduced by a factor of over 10000 and further reduction in time and cost for DNA sequence generation is expected in the next few years. Today, an improved high-quality draft genome sequence (in the sense of Chain et al. 2009) is available for *Populus trichocarpa*, one of the species of interest in *Populus* spp.. A draft sequence has been obtained for *Eucalyptus grandis*

(http::://www.phytozome.net/eucalyptus) and for *E. camaldulensis* (Hirakawa et al. 2011). These three angiosperm species have quite modest genome sizes (485, 691 and 650 Mb, respectively). Acquisition of draft genome sequences for conifer species seemed a "daunting task" until recently (Hamberger et al. 2009), because of their large genome sizes (18-40 Gb), their very high content in repetitive DNA and the absence of an inbred genotype. Consequently, reduced-representation approaches have been in use to alleviate the complexity of these genomes. Targeted re-sequencing of candidate genes, pointed out by functional genomics studies, was an initial step in developing new markers, quickly followed by the re-sequencing of unigene sets developed on the basis of previous EST libraries (Chancerel et al. 2011; Eckert et al. 2010). However, Feuillet et al. (2011) qualified the acquisition of draft genome sequences for conifer species as achievable at reasonable costs due to advances in next-generation DNA sequencing technologies. Indeed, draft assemblies of the 20 Gb genome of Norway spruce (*Picea abies*), of the 20.8 Gb genome of white spruce (*Picea glauca*) and of the 22 Gb genome of loblolly pine (*Pinus taeda*) were recently presented (Birol et al. 2013; Neale et al. 2014; Nystedt et al. 2013; Zimin et al. 2014), whereas several projects to sequence the genome of other economically important conifer species are underway: *Pinus pinaster* and *Pinus sylvestris* (in Europe ProCoGen, http://bfw.ac.at/rz/bfwcms2.web?dok=9020).

On the basis of these sequence data, high-throughput genotyping tools like Illumina GoldenGate or Infinium assays, Affymetrix Axiom assays, Sequenom MassArray assays and TaqMan OpenArray assays can be constructed, with the result that the number of markers available will soon no longer be a limiting factor for MAS. Genotyping costs are being progressively reduced with the standardization of the use of these assays, notably when observing the number of data points attainable per monetary unit. In parallel, genotyping-by-sequencing techniques are under rapid development thanks to the possibility they offer of obtaining high marker densities without prior knowledge of a species genome (Elshire et al. 2011). Such a technique was used in an association study concerning an adaptive trait in lodgepole pine (Parchman et al. 2012) and GBS performance was evaluated in a pilot study on lodgepole pine and white spruce (Chen et al. 2013).

The acceleration in the development of genomic markers currently observed provides an optimistic perspective of reaching enough LD between markers and QTLs for the application of MAS in most forest tree species in the near future. The cost of initial investment to integrate MAS in breeding strategies will greatly decrease. In the mid- or maybe long-term,

high throughput methodologies for developing genetic markers could change the perspectives of marker use in forest tree breeding schemes as has been observed for cattle breeding with the possibility of genome wide evaluation.

3.4 Prospects in the mid-term

With chip-based and genotyping-by-sequencing technologies, data are obtained "en masse" compared to the almost step-wise discovery of previous technologies. Genomic selection, or more appropriately genome wide evaluation (GWE), was thus designed to use very dense genetic data chosen to cover the entire genome. With coverage being dense enough, all the QTLs underlying the trait of interest are expected to be in LD with some of the genotyped markers. In such situations, most of the genetic variation for the evaluated trait could eventually be captured. GWE enables predicting genetic values in a given candidate set without the need for its phenotypic information. To do this, a *training* population closely related to the candidate set provides beforehand the phenotypes and genotypes required to fit the prediction equations (Meuwissen et al. 2001).

GWE has already been widely embraced by dairy cattle breeding programs, with a few more examples of on-going implementations in livestock: small ruminants and pigs to cite two of the most advanced (Stock and Reents 2013). The picture appears somehow less brilliant in crop plants, as the success of GWE in dairy cattle cannot be simply translated into crop breeding (Jonas and de Koning 2013). However, many implementations are underway, notably in crops like maize and wheat. The very first steps of GWE in forest trees were reported in loblolly pine (Resende Jr et al. 2012b; Resende Jr et al. 2012a) and in two independent breeding populations of *Eucalyptus* (Resende et al. 2012). These were devised as proof-of-concept studies and provide first-hand assessments of prediction accuracies for various traits of economic interest and assuming several alternative statistical models. Accuracies can be at levels comparable to those already obtained by pedigree-based genetic predictions.

Given the preliminary and somehow down-scaled nature of these first studies, certain questions still remain unanswered concerning the implementation of GWE in forest trees, and the same questions also concern crop plants. First, as stated in previously, most forest tree and crop breeding programmes involve populations of large effective sizes, comprising several landraces or subpopulations that would require correspondingly large genomic resources for efficient coverage. Previous proof-of-concept studies in trees showed that accuracies

comparable to BLUP-selection are attainable with down-scaled genomic resources (i.e. involving a few thousands SNP), but at the cost of reducing the representativeness of the training populations. This situation is parallel to that of mapping pedigrees for MAS, where representativeness and connections to breeding populations were also an issue. Solutions will certainly come from further developments in genomic resources for these species as well as from an optimum choice of training populations.

Another potential issue with GS is that of the faster increase of inbreeding on the time scale, as compared to phenotypic selection, because of the accelerated turnover in generations. This rapid increase can be exacerbated in populations with narrow genetic bases, supposed to facilitate GWE implementation. However, unlike pedigree-based BLUP, GWE is able to capture the Mendelian sampling term, which in turn allows greater accuracy and genetic gains without increasing the rate of inbreeding (Pryce et al. 2012). Thus the issue is still contentious, as what can be gained from tracking the Mendelian sampling term could be lost by a rapid generation turnover. Solutions were available for BLUP-selection (known to be prone to high inbreeding rates) and recent derivations have been proposed for GWE (Nielsen et al. 2011).

A third issue, that is also relevant for crops, concerns the expected decrease in GWE accuracies for candidates after several generations of recombination without prediction updating with a new training population. This loss in accuracy has not been sufficiently addressed throughout case studies, as these often concern candidates that are closely related to the members of the training population. Solutions would certainly come from a continuous renewal of the training population, in a *rolling front* fashion, so that the genetic *distance* between training and evaluation sets never increases beyond a point that downgrades accuracy. The phenotypic data obtained in deployment could be used to update the genome-wide prediction models (Iwata et al. 2011).

Another last issue that particularly affects crops and forest trees is the pervasiveness of G×E interactions for many key physiological and production traits. In order to track these effects conveniently, phenotypic evaluations need to be multiplied across biologically relevant environments. This need brings new demands to the design of training populations for GWE, up to the extreme point of needing population-specific alternatives. One of the few studies considering GWE in forest trees (Resende Jr et al. 2012a) shows examples of G×E severely affecting the transferability of models across environments. Apart from population-specific alternatives, other solutions could be model-based and focus on the explicit incorporation of

G×E terms (Resende et al. 2012). There is therefore a great need to develop GWE models beyond the additive paradigm.

Because phenotyping costs are becoming the limiting factor, available phenotypic data are valuable resources and archiving DNA of all or part of the candidates that provided these data, particularly the founders, could be a wise way of preparing a future implementation of GWE and/or MAS. This means preparing biobanks of DNA samples and phenotypic databases relating these samples to phenotypic data. However, new phenotypic data will be needed to cope with new traits and to update genome-wide prediction models. Consequently, research devoted to increasing the throughput of phenotyping activities, using for example robotics or automatic image capture and analysis, are needed.

Although GWE can be considered as a variant of MAS, there is a *key* difference in that in the former all available marker data are used anonymously. GWE remains a *black box* predicting process similar to that of phenotypic evaluation, but using all sources of available information to improve predictability. On the contrary, MAS picks relevant sources by their explanatory value: identified QTLs or markers pointing to them as QTL proxies. Lorenz et al. (2011) go further in opposing MAS to GWE. MAS would rely on a *breeding by design* engineering approach, requiring fine (prior) knowledge of biological functions (of the functional polymorphisms underlying QTLs). This qualifies well for genotype building (Step 3 of marker use, Fig. 4), by gene pyramiding or marker-assisted backcrossing, which were the most successful applications of MAS in annual crops, but it does not seem adequate for marker-based recurrent selection. GWE overlooks explanations with its *black box* approach. However, these two approaches do not necessarily need to be opposed and could well become complementary (Resende Jr et al. 2012a). Thus, once discovered and their effects accurately estimated, functional polymorphisms underlying QTLs could be incorporated in GWE models to further improve prediction accuracy. Such a strategy is already applied in the French dairy cattle GWE program (Boichard et al. 2012), in which a panel of a few hundred trait-dependent genomic regions and SNP-based haplotypes are routinely used in a classic QTL-BLUP model, which presents accuracies as high as those of *blind* GWE. This panel of relevant markers is selected by using classical linkage disequilibrium – linkage analysis (LD-LA) or by the use of variable selection models. Although a promising route, as it involves prediction and explanation into a single model, it is far from being directly applicable to crops and forest trees.

Given the extent to which explanation and prediction advance in parallel in our model species, we foresee a convergence between GWE and MAS in the future, when knowledge on the underlying genetics of key traits is well advanced, with identified functional and regulatory QTLs adding to the masses of anonymous markers of GWE. To conclude, a definitive solution to the G×E issue will certainly involve careful dissection of QTL effects and QTL×QTL interactions over multiple environments with the help of genome-wide evaluations.

Acknowledgements

This work was funded by the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreement n° 211868 (Project Noveltree).

References

Ahmadi N, Albar L, Pressoir G, et al. (2001) Genetic basis and mapping of the resistance to *Rice yellow mottle virus*. III. Analysis of QTL efficiency in introgressed progenies confirmed the hypothesis of complementary epistasis between two resistance QTLs. Theor. Appl. Genet. 103:1084–1092.

Alves de Moraes RM, Bastos Soares TC, Colombo LR, et al. (2006) Assisted selection by specific DNA markers for genetic elimination of the kunitz trypsin inhibitor and lectin in soybean seeds. Euphytica 149:221–226. doi: 10.1007/s10681-005-9069-0

Babu R, Nair SK, Kumar A, et al. (2005) Two-generation marker-aided backcrossing for rapid conversion of normal maize lines to quality protein maize (QPM). Theor. Appl. Genet. 111:888–897. doi: 10.1007/s00122-005-0011-6

Basavaraj SH, Singh VK, Atul S, et al. (2010) Marker-assisted improvement of bacterial blight resistance in parental lines of Pusa RH10, a superfine grain aromatic rice hybrid. Molecular Breeding 26:293–305.

Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48:1649–1664. doi: 10.2135/cropsci2008.03.0131

Bernardo RC, Charcosset A (2006) Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. Crop Sci. 46:614–621. doi: 10.2135/cropsci2005.05-0088

Bhatia D, Sharma R, Vikal Y, et al. (2011) Marker-assisted development of bacterial blight resistant, dwarf, and high yielding versions of two traditional basmati rice cultivars. Crop Sci. 51:759–770. doi: 10.2135/cropsci2010.06.0358

Birol I, Raymond A, Jackman SD, et al. (2013) Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497. doi: 10.1093/bioinformatics/btt178

Blanc G, Charcosset A, Veyrieras JB, et al. (2008) Marker-assisted selection efficiency in multiple connected populations: a simulation study based on the results of a QTL detection experiment in maize. Euphytica 161:71–84.

Boichard D, Guillaume F, Baur A, et al. (2012) Genomic selection in French dairy cattle. Anim. Prod. Sci. 52:115–120.

Burdon RD, Wilcox PL (2007) Population management: potential impacts of advances in genomics. New Forests 34:187–206. doi: 10.1007/s11056-007-9047-6

Causse M, Chaib J, Lecomte L, et al. (2007) Both additivity and epistasis control the genetic variation for fruit quality traits in tomato. Theoretical and Applied Genetics 115:429–442. doi: 10.1007/s00122-007-0578-1

Chain P, Grafham D, Fulton R, et al. (2009) Genome project standards in a new era of sequencing. Science 326:236.

Chakraborty R, Moreau L, Dekkers JC (2002) A method to optimize selection on multiple identified quantitative trait loci. Genet. Sel. Evol. 34:145–170. doi: 10.1186/1297-9686-34-2-145

Chancerel E, Lepoittevin C, Provost GL, et al. (2011) Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics 12:368. doi: 10.1186/1471-2164-12-368

Chen C, Mitchell SE, Elshire RJ, et al. (2013) Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9:1537–1544. doi: 10.1007/s11295-013-0657-1

Chen CL, Chuang SJ, Chen JJ, Sung JM (2009) Using RAPD markers to predict polyphenol content in aerial parts of *Echinacea purpurea* plants. J. Sci. Food Agric. 89:2137–2143. doi: 10.1002/jsfa.3704

Chen S, Lin XH, Xu CG, Zhang Q (2000) Improvement of bacterial blight resistance of "Minghui 63", an elite restorer line of hybrid rice, by molecular marker-assisted selection. Crop Sci. 40:239–244.

Chen S, Xu CG, Lin XH, Zhang Q (2001) Improving bacterial blight resistance of "6078", an elite restorer line of hybrid rice, by molecular marker-assisted selection. Plant Breed. 120:133–137. doi: 10.1046/j.1439-0523.2001.00559.x

Cornelius J (1994) Heritabilities and additive genetic coefficients of variation in forest trees. Canadian Journal of Forest Research 24:372–379. doi: 10.1139/x94-050

Darvasi A, Soller M (1997) A Simple Method to Calculate Resolving Power and Confidence Interval of QTL Map Location. Behavior Genetics 27:125–132. doi: 10.1023/A:1025685324830

Dekkers JC, Chakraborty R, Moreau L (2002) Optimal selection on two quantitative trait loci with linkage. Genet. Sel. Evol. 34:171–192. doi: 10.1186/1297-9686-34-2-171

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331–341.

Devey ME, Groom KA, Nolan MF, et al. (2004) Detection and verification of quantitative trait loci for resistance to Dothistroma needle blight in *Pinus radiata*. Theor. Appl. Genet. 108:1056–1063. doi: 10.1007/s00122-003-1471-1

Dolstra O, Denneboom C, Vos ALF de, Loo EN van (2007) Marker-assisted selection for improving quantitative traits of forage crops. In:(ed) Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish ed. Food and Agriculture Organization of the United Nations (FAO), Rome Italy, pp59–65

Eckert AJ, van Heervaarden J, Wegrzyn JL, et al. (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185:969–982. doi: 10.1534/genetics.110.115543

Edwards MD, Page NJ (1994) Evaluation of marker-assisted selection through computer simulation. Theor. Appl. Genet. 88:376–382.

El-Kassaby YA, Cappa EP, Liewlaksaneeyanawin C, et al. (2011) Breeding without Breeding: Is a Complete Pedigree Necessary for Efficient Breeding? PLoS ONE 6:e25737. doi: 10.1371/journal.pone.0025737

El-Kassaby YA, Lstiburek M (2009) Breeding without breeding. Genet. Res. 91:111–120. doi: 10.1017/S001667230900007X

Elshire RJ, Glaubitz JC, Sun Q, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6:e19379. doi: 10.1371/journal.pone.0019379

Emebiri L, Michael P, Moody DB, et al. (2009a) Pyramiding QTLs to improve malting quality in barley: gains in phenotype and genetic diversity. Mol. Breed. 23:219–228. doi: 10.1007/s11032-008-9227-x

Emebiri LC, Michael P, Moody DB (2009b) Enhanced tolerance to boron toxicity in two-rowed barley by marker-assisted introgression of favourable alleles derived from Sahara 3771. Plant Soil 314:77–85. doi: 10.1007/s11104-008-9707-0

Fernando RL, Grossman M (1989) Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467–477.

Feuillet C, Leach JE, Rogers J, et al. (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci. 16:77–88. doi: 10.1016/j.tplants.2010.10.005

Flachowsky H, Le Roux P-M, Peil A, et al. (2011) Application of a high-speed breeding technology to apple (*Malus* × *domestica*) based on transgenic early flowering plants and marker-assisted selection. New Phytol. 192:364–377. doi: 10.1111/j.1469-8137.2011.03813.x

Flint-Garcia SA, Darrah LL, McMullen MD, Hibbard BE (2003a) Phenotypic versus marker-assisted selection for stalk strength and second-generation European corn borer resistance in maize. Theor. Appl. Genet. 107:1331–1336. doi: 10.1007/s00122-003-1387-9

Flint-Garcia SA, Thornsberry JM, S E, Iv B (2003b) Structure of linkage disequilibrium in plants. Annual Review of Plant Biology 54:357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Gill GP, Harvey CF, Gardner RC, Fraser LG (1998) Development of sex-linked PCR markers for gender identification in *Actinidia*. Theor. Appl. Genet. 97:439–445. doi: 10.1007/s001220050914

Goddard ME (2001) The validity of genetic models underlying quantitative traits. Livest. Prod. Sci. 72:117–127.

Grattapaglia D (2007) Marker-assisted selection in *Eucalyptus*. In:Guimaraes EP, Ruane J, Scherf BD, et al.(ed) Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish ed. Food and Agriculture Organization of the United Nations (FAO), Rome Italy, pp251–281

Grattapaglia D, Kirst M (2008) Eucalyptus applied genomics: from gene sequences to breeding tools. New Phytologist 179:911–929. doi: 10.1111/j.1469-8137.2008.02503.x

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet. Genom. 7:241–255.

Guillaume F, Fritz S, Boichard D, Druet T (2008) Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40:91–102. doi: 10.1051/gse:2007036

Gupta P, Rustgi S, Kulwal P (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Molecular Biology 57:461–485. doi: 10.1007/s11103-005-0257-z

Hamberger B, Hall D, Yuen M, et al. (2009) Targeted isolation, sequence assembly and characterization of two white spruce (Picea glauca) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. BMC Plant Biol. 9:106. doi: 10.1186/1471-2229-9-106

Han F, Romagosa I, Ullrich SE, et al. (1997) Molecular marker-assisted selection for malting quality traits in barley. Mol. Breed. 3:427–437.

Harushima Y, Yano M, Shomura P, et al. (1998) A high-density rice genetic linkage map with 2275 markers using a single F-2 population. Genetics 148:479–494.

Hayes B, Baranski M, Goddard ME, Robinson N (2007) Optimisation of marker assisted selection for abalone breeding programs. Aquaculture 265:61–69.

Hill W, Weir B (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93:47–64.

Hirakawa H, Nakamura Y, Kaneko T, et al. (2011) Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*. Plant Biotechnology 28:471–480.

Hospital F (2009) Challenges for effective marker-assisted selection in plants. Genetica 136:303–310. doi: 10.1007/s10709-008-9307-1

Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485.

Huang N, Angeles ER, Domingo J, et al. (1997) Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. Theor. Appl. Genet. 95:313–320. doi: 10.1007/s001220050565

Hyten DL, Choi IY, Song QJ, et al. (2010) A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci. 50:960–968. doi: 10.2135/cropsci2009.06.0360

Igartua E, Edney M, Rossnagel BG, et al. (2000) Marker-based selection of QTL affecting grain and malt quality in two-row barley. Crop Sci. 40:1426–1433. doi: 10.2135/cropsci2000.4051426x

Iwata H, Hayashi T, Tsumura Y (2011) Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. Tree Genet. Genom. 7:747–758. doi: 10.1007/s11295-011-0371-9

Jena KK, Mackill DJ (2008) Molecular markers and their use in marker-assisted selection in rice. Crop Sci. 48:1266–1276. doi: 10.2135/cropsci2008.02.0082

Jermstad KD, Bassoni DL, Jech KS, et al. (2001) Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. I. Timing of vegetative bud flush. Theor. Appl. Genet. 102:1142–1151. doi: 10.1007/s001220000505

Johnson GR, Wheeler NC, Strauss SH (2000) Financial feasibility of marker-aided selection in Douglas-fir. Can. J. For. Res. 30:1942–1952. doi: 10.1139/x00-122

Jonas E, de Koning D-J (2013) Does genomic selection have a future in plant breeding? Trends in Biotechnology 31:497–504. doi: 10.1016/j.tibtech.2013.06.003

Kearsey MJ, Farquhar AGL (1998) QTL analysis in plants; where are we now? Heredity 80:137–142. doi: 10.1046/j.1365-2540.1998.00500.x

Knaap E van der, Lippman ZB, Tanksley SD (2002) Extremely elongated tomato fruit controlled by four quantitative trait loci with epistatic interactions. Theor Appl Genet 104:241–247. doi: 10.1007/s00122-001-0776-1

Kuchel H, Ye G, Fox R, Jefferies S (2005) Genetic and economic analysis of a targeted marker-assisted wheat breeding strategy. Mol. Breed. 16:67–78. doi: 10.1007/s11032-005-4785-7

Kumar S, Garrick DJ (2001) Genetic response to within-family selection using molecular markers in some radiata pine breeding schemes. Can. J. For. Res. 31:779–785.

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Levi A, Paterson A, Barak V, et al. (2009) Field evaluation of cotton near-isogenic lines introgressed with QTLs for productivity and drought related traits. Mol. Breed. 23:179–195. doi: 10.1007/s11032-008-9224-0

Lorenz AJ, Chao S, Asoro FG, et al. (2011) Genomic Selection in Plant Breeding: Knowledge and Prospects. In:Sparks DL(ed) Advances in Agronomy, Vol 110 ed. Elsevier Academic Press Inc, San Diego, pp77–123

Mahmood T, Rahman MH, Stringam GR, et al. (2005) Molecular markers for yield components in Brassica juncea - do these assist in breeding for high seed yield? Euphytica 144:157–167. doi: 10.1007/s10681-005-5339-0

Massah N, Wang J, Russell JH, et al. (2010) Genealogical relationship among members of selection and production populations of yellow cedar (Callitropsis nootkatensis [D. Don] Oerst.) in the absence of parental information. J. Hered. 101:154–163. doi: 10.1093/jhered/esp102

Mayor PJ, Bernardo R (2009) Doubled haploids in commercial maize breeding: one-stage and two-stage phenotypic selection versus marker-assisted recurrent selection. Maydica 54:439–448.

Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Moreau L, Charcosset A, Gallais A (2004) Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica 137:111–118. doi: 10.1023/b:euph.0000040508.01402.21

Moreau L, Charcosset A, Hospital F, Gallais A (1998) Marker-assisted selection efficiency in populations of finite size. Genetics 148:1353–1365.

Morris M, Dreher K, Ribaut JM, Khairallah M (2003) Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. Mol. Breed. 11:235–247. doi: 10.1023/a:1022872604743

Myint Y, Khin Than N, Vanavichit A, et al. (2009) Marker assisted backcross breeding to improve cooking quality traits in Myanmar rice cultivar Manawthukha. Field Crops Res. 113:178–186. doi: 10.1016/j.fcr.2009.05.006

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci. 9:325–330. doi: 10.1016/j.tplants.2004.05.006

Neale DB, Wegrzyn JL, Stevens KA, et al. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 15:R59. doi: 10.1186/gb-2014-15-3-r59

Nielsen HM, Sonesson AK, Meuwissen THE (2011) Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. J ANIM SCI 89:630–638. doi: 10.2527/jas.2009-2731

Nystedt B, Street NR, Wetterbom A, et al. (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584. doi: 10.1038/nature12211

Ollivier L (1998) The accuracy of marker-assisted selection for quantitative traits within populations in linkage equilibrium. Genetics 148:1367–1372.

Parasnis AS, Ramakrishna W, Chowdari KV, et al. (1999) Microsatellite (GATA)n reveals sex-specific differences in Papaya. Theor. Appl. Genet. 99:1047–1052. doi: 10.1007/s001220051413

Parchman TL, Gompert Z, Mudge J, et al. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular Ecology 21:2991–3005. doi: 10.1111/j.1365-294X.2012.05513.x

Perumalsamy S, Bharani M, Sudha M, et al. (2010) Functional marker-assisted selection for bacterial leaf blight resistance genes in rice (Oryza sativa L.). Plant Breed. 129:400–406.

Plomion C, Durel CE, Verhaegen D (1996) Marker-assisted selection in forest tree breeding programs as illustrated by two examples: Maritime pine and eucalyptus. Ann. For. Sci. 53:819–848.

Plomion C, LeProvost G, Pot D, et al. (2001) Pollen contamination in a maritime pine polycross seed orchard and certification of improved seeds using chloroplast microsatellites. Canadian Journal of Forest Research 31:1816–1825. doi: 10.1139/cjfr-31-10-1816

Prasanna BM, Pixley K, Warburton ML, Xie C-X (2010) Molecular marker-assisted breeding options for maize improvement in Asia. Molecular Breeding 26:339–356. doi: 10.1007/s11032-009-9387-3

Prat D, Faivre-Rampant P, Prado E (2006) Genome analysis and the management of forest genetic resources. Institut National de la Recherche Agronomique (INRA),

Pryce JE, Hayes BJ, Goddard ME (2012) Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. Journal of Dairy Science 95:377–388. doi: 10.3168/jds.2011-4254

Resende Jr MFR, Muñoz P, Acosta JJ, et al. (2012a) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol. 193:617–624. doi: 10.1111/j.1469-8137.2011.03895.x

Resende Jr MFR, Muñoz P, Resende MDV, et al. (2012b) Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus Taeda L.). Genetics 190:1503–1510. doi: 10.1534/genetics.111.137026

Resende MDV, Resende Jr MFR, Sansaloni CP, 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. doi: 10.1111/j.1469-8137.2011.04038.x

Ribeiro MM, Sanchez L, Ribeiro C, et al. (2011) A case study of Eucalyptus globulus fingerprinting for breeding. Annals of Forest Science 68:701–714. doi: 10.1007/s13595-011-0087-x

Rosvall O (2011) Review of the Swedish breeding programme. Skogforsk, Uppsala, Sweden,

Sánchez L, Caballero A, Santiago E (2006) Palliating the impact of fixation of a major gene on the genetic variation of artificially selected polygenes. Genetical Research 88:105–118. doi: 10.1017/S0016672306008421

Sebastian SA, Streit LG, Stephens PA, et al. (2010) Context-specific marker-assisted selection for improved grain yield in elite soybean populations. Crop Sci. 50:1196–1206. doi: 10.2135/cropsci2009.02.0078

Sewell M, Bassoni D, Megraw R, et al. (2000) Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). I. Physical wood properties. Theor. Appl. Genet. 101:1273–1281.

Sharma K, Agrawal V, Gupta S, et al. (2008) ISSR marker-assisted selection of male and female plants in a promising dioecious crop: jojoba (*Simmondsia chinensis*). Plant Biotechnol Rep 2:239–243. doi: 10.1007/s11816-008-0070-7

Sillanpää MJ (2011) On statistical methods for estimating heritability in wild populations. Molecular Ecology 20:1324–1332. doi: 10.1111/j.1365-294X.2011.05021.x

Silva FF da, Pereira MG, Campos WF, et al. (2007) DNA marker-assisted sex conversion in elite papaya genotype (*Carica papaya* L.). Crop Breed. Appl. Biotechnol. 7:52–58.

Silva KM, Bastiaansen JWM, Knol EF, et al. (2011) Meta-analysis of results from quantitative trait loci mapping studies on pig chromosome 4. Animal Genetics 42:280–292. doi: 10.1111/j.1365-2052.2010.02145.x

Somers DJ, Isaac P, Edwards K (2004) A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). Theor. Appl. Genet. 109:1105–1114.

Sorkheh K, Malysheva-Otto LV, Wirthensohn MG, et al. (2008) Linkage disequilibrium, genetic association mapping and gene localization in crop plants. Genet. Mol. Biol. 31:805–814. doi: 10.1590/S1415-47572008005000005

Spelman R, Bovenhuis H (1998) Genetic response from marker assisted selection in an outbred population for differing marker bracket sizes and with two identified quantitative trait loci. Genetics 148:1389–1396.

Steele KA, Price AH, Shashidhar HE, Witcombe JR (2006) Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. Theor. Appl. Genet. 112:208–221.

Stendal C, Casler MD, Jung G (2006) Marker-assisted selection for neutral detergent fiber in smooth bromegrass. Crop Sci. 46:303–311.

Stock K, Reents R (2013) Genomic Selection: Status in Different Species and Challenges for Breeding. Reproduction in Domestic Animals 48:2–10. doi: 10.1111/rda.12201

Szücs PB, Bhat VC, Chao PR, et al. (2009) An integrated resource for barley linkage map and malting quality QTL alignment. Plant Genome 2:134.

Tanhuanpää P, Vilkki J (1999) Marker-assisted selection for oleic acid content in spring turnip rape. Plant Breed. 118:568–570.

Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) RFLP Mapping in Plant Breeding: New Tools for an Old Science. Nature Biotechnology 7:257–264. doi: 10.1038/nbt0389-257

Truntzler M, Barrière Y, Sawkins M, et al. (2010) Meta-analysis of QTL involved in silage quality of maize and comparison with the position of candidate genes. Theor. Appl. Genet. 121:1465–1482. doi: 10.1007/s00122-010-1402-x

Twardowska M, Masojc P, Milczarski P (2005) Pyramiding genes affecting sprouting resistance in rye by means of marker assisted selection. In:(ed) Proceedings of the 10th International Symposium on Pre-Harvest Sprouting in Cereals, Norfolk, UK, 7-11 June 2004. ed. Norfolk, UK, pp257–260

VanRaden PM (2008) Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi: 10.3168/jds.2007-0980

Varshney R, Marcel T, Ramsay L, et al. (2007) A high density barley microsatellite consensus map with 775 SSR loci. Theor. Appl. Genet. 114:1091–1103. doi: 10.1007/s00122-007-0503-7

Veyrieras J-B, Goffinet B, Charcosset A (2007) MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. BMC Bioinformatics 8:49. doi: 10.1186/1471-2105-8-49

Vida G, Gal M, Uhrin A, et al. (2009) Molecular markers for the identification of resistance genes and marker-assisted selection in breeding wheat for leaf rust resistance. Euphytica 170:67–76. doi: 10.1007/s10681-009-9945-0

Villanueva B, Pong-Wong R, Woolliams JA (2002) Marker assisted selection with optimised contributions of the candidates to selection. Genet. Sel. Evol. 34:679–703. doi: 10.1051/gse:2002031

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era - concepts and misconceptions. Nat. Rev. Genet. 9:255–266. doi: 10.1038/nrg2322

Van de Wen WTG, McNicol RJ (1995) The use of RAPD markers for the identification of Sitka spruce (Picea sitchensis) clones. Heredity 75:126–132.

Wilcox PL, Carson SD, Richardson TE, et al. (2001) Benefit-cost analysis of DNA marker-based selection in progenies of Pinus radiata seed orchard parents. Can. J. For. Res. 31:2213–2224. doi: 10.1139/cjfr-31-12-2213

Wong CK, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116:815–824. doi: 10.1007/s00122-008-0715-5

Xie C, Xu S (1998) Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. Heredity 80:489–498. doi: 10.1038/sj.hdy.6883080

Xu SZ (2003) Theoretical basis of the Beavis effect. Genetics 165:2259–2268.

Xu YB, Crouch JH (2008) Marker-assisted selection in plant breeding: From publications to practice. Crop Sci. 48:391–407. doi: 10.2135/cropsci2007.04.0191

Zimin A, Stevens KA, Crepeau MW, et al. (2014) Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics 196:875–890. doi: 10.1534/genetics.113.159715

Table 1 Classification of 750 papers retrieved from Web of Science and CAB abstract databases selected for meta-analysis.

Class	**Number of papers**
Applications	240
Theoretical developments	32
Simulations	71
Developments and Simulations	27
Developments, Simulations and Applications	1
Simulations and Applications	6
Reviews	99
Review and Application	3
Review and Simulations	1
Others	270
QTL detection, development or validation of markers, evaluation of QTL effects in various environment or genetic backgrounds	227
Management of genetic diversity, core-collection building	13
Development or review of methods for high throughput DNA extraction or genotyping	16
MAS in animals	14

Table 2 Overview of QTLs associated with traits of interest grouped in 5 main categories (Growth/biomass (incl. rooting, water use efficiency), phenology, wood properties, abiotic/biotic stress resistance, and specific traits such as leaf secondary metabolism) within main forest tree taxa with breeding programs.

Species	**Trait**	**Nb of QTL**	**Variation explained by QTL**	**Sample size**	**CI length (cM)** (as reported or *estimated roughly from figures*)	**Reference** [c]
Cryptomeria japonica	Growth	6	18.6% - 48.8%	73	*4-30*	Yoshimaru et al. 1998
	Wood properties	23	15.7% - 34.3%	72	*10-100*	Kuramoto et al. 2000
Eucalyptus spp.	Abiotic/biotic stress resistance	1-6	3% to a major gene	75 - 994	*5-50*	Byrne et al. 1997; Dale et al. 2000; Freeman et al. 2008; Fullard and Moran 2003; Junghans et al. 2003; Mamani et al. 2010
	Wood properties	3-18	3.2% - 49%	91 - 277	*20-50*	Thamarus et al. 2004
	Growth	2-10	3%-60%	91-221	*25-40*	Grattapaglia et al. 1996
	Leaf secondary metabolism	1 major - 30	1% - 60%	52-296	*4-50*	Shepherd et al. 1999 Henery et al. 2007
Picea glauca	Growth	52	2.5%-10.3%	260 + 500	13-104	Pelgas et al. 2011
	Phenology	85	2.5 – 16.4%	260 + 500	3-123	Pelgas et al. 2011
Pinus pallustris x P. elliottii	Growth	11	3.6 – 11%	258	18.7-57.1 [a]	Weng et al. 2002
Pinus pinaster	Abiotic/biotic stress resistance	7	Up to 11.1%	53	45.4-186.9 [a]	Hurme et al. 2000
	Growth	1-40	5-20.4%	80-202	1.2-*100*	Brendel et al. 2002
	Phenology	4	Up to 12.7%	81-92	90.1-250 [ab]	Hurme et al. 2000
	Wood properties	10-54	3.7 – 18.4%	80-186	1-185	Pot et al. 2006
Pinus radiata	Growth	2 - 21	2.2 – 36.4%	50 -400	15.7-152.3 [a]	Emebiri et al. 1997; Emebiri et al. 1998a; Emebiri et al. 1998b
	Wood properties	1-8	0.78-3.58%	80 – 400	37-165.6 [a]	Kumar et al. 2000
	Abiotic/biotic stress resistance	4	0 – 4.8%	202	54.7-149.1 [a]	Devey et al. 2004
Pinus sylvestris	Growth	4-12	9.3-16.7%	94-108	33.8-60.6 [a]	Lerceteau et al. 2000, Yazdani et al. 2003
	Abiotic/biotic stress resistance	4-9	11.3 – 22.7%	94-108	24.8-54.5 [a]	Lerceteau et al. 2000; Yazdani et al. 2003
	Phenology	5	42-79%	108	40.6	Yazdani et al. 2003
Pinus taeda	Growth	2 - 6	4.7 – 30%	84 - 171	23.6-121.3 [a]	Kaya et al. 1999
	Wood properties	2 - 12	1.7 – 15.9%	172-434	19.6-64.4 [a]	Groover et al. 1994; Sewell et al. 2000
Pseudotsuga menziesii	Phenology	1-11	0.7 – 11.5%	78-460	42.3-165.7	Jermstad et al. 2001a
	Growth	5	8.6-17.7%	320	9.4-19.3	Ukrainetz et al. 2008

	Abiotic/biotic stress resistance	11 - 15	2.0 – 9.8%	170-383	*20-50*	Jermstad et al. 2001b
	Wood properties	1 - 26	0.1 – 15.7%	320	10.5-46	Ukrainetz et al. 2008
Populus spp.	Abiotic/biotic stress resistance	50	1,8 – 87%	68-336	0,9 - 46	Jorge et al. 2005; Labbé et al. 2011; Newcombe et al. 1996; Newcombe and Bradshaw 1996; Tagu et al. 2004
	Phenology	160	1-51.5	55-356	7.5 – 314.2	Bradshaw and Stettler 1995; Chen et al. 2002; Frewen et al. 2000; Marron et al. 2010; Rohde et al. 2011
	Wood properties	62	3.6 – 48.9	87-387	3.3 – 40.6	Huang et al. 2004; Novaes et al. 2009; Yin et al. 2010; Zhang et al. 2006)
	Growth	603	0 – 73.6	55-387	2.1 – 261.2	Bradshaw and Stettler 1995; Dillen et al. 2009; Marron et al. 2010; Novaes et al. 2009; Rae et al. 2009; Rae et al. 2008; Rae et al. 2007; Street et al. 2011; Street et al. 2006; Wu 1998; Wu et al. 1998; Wullschleger et al. 2005; Zhang et al. 2006
Salix spp.	Growth	73	8 - 29	87-92	19.9-72	Rönnberg-Wästljung et al. 2005; Tsarouhas et al. 2002
	Abiotic/biotic stress resistance	134	3.2 – 56.4	73-467	1 – 91.7	Hanley et al. 2011; Rönnberg-Wästljung et al. 2006; Rönnberg-Wästljung et al. 2008; Samils et al. 2011
	Phenology	21	1.0 – 24.0	87	43.1-138	Tsarouhas et al. 2004; Tsarouhas et al. 2003
	Wood properties	5	10.2 – 21.8	125	21.8 – 35.6	Brereton et al. 2010

[a] estimated through R^2 and mapping population size with Darvasi's and Soller's (1997) formula. [b] 4 values > 300 cM excluded. [c] References to be found in Supplementary document 3.

Table 3 Detection of QTLs associated with traits of interest in population-level association mapping studies in forest tree species.

Species	Trait	Nb of associated markers	Variation explained	Sample size	Reference [c]
Eucalyptus nitens	Microfibril angle	2	4.6% - 4.6%	290	Thumma et al. 2005
	Cellulose content, kraft pulp yield	1	Not estimated	420	Thumma et al. 2009
Picea sitchensis	bud set timing	26	1% - 4.3% (34.4% altogether)	410	Holliday et al. 2010
	cold hardiness	19	0.7% - 5.4% (28.1% altogether)		
Picea glauca	10 Wood quality traits	13	3% - 5%	492	Beaulieu et al. 2011
	Radial growth				
Pinus pinaster	Cellulose content	1	~10% (probably overestimated)	162	Lepoittevin et al. 2012
	Growth (height, diameter)	1	~10% (probably overestimated)	160	
Pinus radiata	Solid wood properties	10	2% - 6.5%	447	Dillon et al. 2010
Pinus taeda	Carbon isotope discrimination	4	0.45% - 3.38%	961	Gonzalez-Martinez et al. 2008
	Early wood specific gravity	2	3.27% - 3.51%	422-435	Gonzalez-Martinez et al. 2007
	Percentage of late wood	1	3.57%		
	Early wood microfibril angle	1	0.78%		
	Geoclimatic variables	48	-	682	Eckert et al. 2010
	Resistance to pitch canker (*Fusarium circinatum*)	10	< 0.98% (3.74% altogether)	498	Quesada et al. 2010
	Carbon isotope discrimination	7	0.01% - 3.8%	380	Cumbie et al. 2011
	Height	1	0.01%		
	Foliar nitrogen concentration	6	0.01% - 7%		
Populus tremula	Timing of bud set	2	1.5% - 5%	120	Ingvarsson et al. 2008
	Growth cessation	6	10%-15% altogether	120	Ma et al. 2010
Populus trichocarpa	Lignin and cellulose content	27	Not estimated	448	Wegrzyn et al. 2010
Pseudotsuga menziesii	cold hardiness	30	1% - 3.6%	700	Eckert et al. 2009

[c] References to be found in Supplementary document 3.

Figure legends

Figure 1. Distribution of MAS papers per species, involving only application studies. Boxed group corresponds to the family with the highest number of hits, *Poaceae*.

Figure 2. Percentage distribution of MAS application papers according to: a) the type of targeted trait, and to b) the type of scheme applied, according to the classification by Hospital (2009)

Figure 3. Distribution of MAS application papers according to the number of targeted genomic regions. Markers located less than 20 cM apart from each other were considered to represent the same genomic region.

Figure 4. Steps foreseen for the introduction of molecular markers in forest tree breeding

Legend for supplementary documents

Supplementary document 1: A text explaining how we selected articles for the meta-analysis and a table showing the queries used and the number of articles obtained.

Supplementary document 2: list of references of the articles used in the meta-analysis, in RIS format.

Supplementary figure 1: Distribution of marker density in 48 papers, concerning 11 species, reporting genetic maps in forest trees.

Supplementary document 3: list of references of the articles cited in Tables 2 and 3, in RIS format.