Novel Tree Breeding

MONOGRAFIAS INIA: SERIE FORESTAL nº 24-2013

Novel Tree Breeding

Editors:
Steve Lee
John Woolliams

NovelTree

SEVENTH FRAMEWORK PROGRAMME

GOBIERNO DE ESPAÑA

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

INIA
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
INSTITUTO NACIONAL DE INVESTIGACIÓN
Y TECNOLOGÍA AGRARIA Y ALIMENTARIA

Novel Tree Breeding

Editors:
Steve Lee
The Roslin Institute near Edinburgh, Scotland

John Woolliams
Forest Research based near Edinburgh, Scotland

2013
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria
Ctra. A. Coruña, km 7,5. 28040 Madrid (España)
Tel.: 91 347 39 16 - Fax: 91 357 22 93 - publinia@inia.es

La responsabilidad por las opiniones emitidas en esta publicación corresponde exclusivamente a los autores de las mismas

Las Monografías INIA están indexadas en Latindex, ICYT y Catálogo de Revistas CSIC

Imagen de portada: Plus trees, and genetic test from different forest tree species (Noveltree project)
Author: S. Lee

Edita: Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria
Ministerio de Economía y Competitividad

ISBN (papel): 978-84-7498-557-3
ISBN (eBook): 978-84-7498-558-0
ISSN: 1575-6116
NIPO (papel): 730-13-007-4
NIPO (en línea): 730-13-008-X
Depósito Legal: M-22560-2013

Imprime: DiScript Preimpresión, S. L.

En esta publicación se ha utilizado papel reciclado libre de cloro de acuerdo con los criterios medioambientales de la contratación pública

INDEX

ABSTRACT

Novel Tree Breeding

Tree breeding is at a crossroads. Many countries have completed their first generation of traditional breeding by selecting superior phenotypes and estimating breeding values for characteristics of interest following long-term and costly field-based progeny trials. Now there are new technologies around. Tree breeders are looking at advancements made in the areas of crop and animal breeding to see how DNA-markers can be used to speed the process up, increase the genetic gains and lower the overall costs. The challenges facing tree breeders are changing too. In additional to the original reasons for selection and breeding there is often the need for selected populations to remain resilient in the face of a changing climate and increased incidents of damaging outbreaks of new or previously benign diseases. In February 2012 a number of tree breeders from across Europe gathered together in South Scotland for one week to compose different 'Chapters' for this monograph. Their objective was partially retrospective to consider where tree breeding has got to but then to consider how new breeding technologies can help the tree breeder in theory and in practice. The book is a presentation of the current state of the art and how it might develop. Importantly, the book identifies the need to continue high through-put yet accurate phenotyping as well as genotyping, and the need to consider genetic variability well into the future. The book and its component Chapters are presented as a conduit to further reading. They are presented here for the students of tree breeding past, present and future as well as others interested in the possible new directions of tree breeding.

Key words: genomic; genome-wide selection; association; quantitative traits; adaptation.

RESUMEN

Nuevos métodos en Mejora Genética Forestal

La mejora genética forestal está en una encrucijada. Muchos países han completado su primera generación de mejoramiento tradicional por selección de fenotipos superiores y la estimación de valores genéticos para caracteres de interés, a partir de costosos y a largo plazo ensayos de progenie sobre el terreno. Actualmente hay nuevas tecnologías disponibles. Los mejoradores genéticos forestales están analizando los avances realizados en las áreas agrícolas y mejora animal para ver cómo se pueden utilizar los marcadores de DNA para acelerar el proceso, aumentar las ganancias genéticas y reducir los costes generales. Los desafíos a los que se enfrentan los mejoradores genéticos forestales están también cambiando. Además de las razones originales para la selección y mejora, a menudo existe la necesidad de que las poblaciones seleccionadas sean resilientes ante el cambio climático y ante el aumento de los casos de brotes de enfermedades nuevas o previamente benignas. En febrero de 2012 una serie de mejoradores de árboles de toda Europa se reunieron en el sur de Escocia durante una semana para componer diferentes capítulos de esta monografía. Su objetivo era parcialmente retrospectivo para considerar lo que había conseguido la mejora genética forestal y posteriormente considerar cómo las nuevas tecnologías pueden ayudar a la mejora genética forestal en la teoría y en la práctica. El libro es una presentación del estado actual del conocimiento y como podría desarrollarse. Es importante destacar que el libro identifica la necesidad de conseguir un fenotipado preciso de alto rendimiento, así como el genotipado, y la necesidad de considerar la variabilidad genética en el futuro. El libro y sus capítulos se presentan como un conducto para la lectura adicional. Se dirige a los estudiantes de mejora genética forestal del pasado, presente y futuro, así como otros interesados en las nuevas orientaciones posibles en la mejora genética de árboles.

Palabras clave: genómica; selección genómica; asociación; caracteres cuantitativos; adaptación.

Foreword
Successes of Tree Breeding: What has been achieved and does it need help?

Steve Lee
Forest Research, Northern Research Station, Roslin, Midlothian EH25 9SY, U.K.

Questions to be addressed in this Foreword:

- Why bother with tree breeding?
- What's the problem with the traditional approach to tree breeding?
- How can new genomic techniques help tree breeding?
- What are the new challenges for tree breeders?
- What is the purpose of this book?

F.1. Summary

This foreword explains the successes of tree breeding that have been achieved over the last 60 years. In particular it identifies good first-generation genetic gains but also the long-term up-front investment required before any financial return is achieved. Breeders of progressive programmes would like to increase efficiencies and get more gain deployed to the field at a faster rate and for less total cost. The objective remains to select the very best genotypes in terms of adaptation and economic value within each generation, and quickly deploy these selections to the forest. Deployment of just a few genotypes may increase genetic gain but may impact upon genetic diversity, and the resilience of the improved population in reacting to changes in the climate or surrounding environment will become increasingly important. This brings into question the traits that need to be considered in future generations. Traditional traits such as growth-rate, and stem straightness may be joined by traits such as plasticity to a changing environment, specific internal wood quality traits, drought and disease resistance, and carbon mitigation issues. Molecular markers represent a new opportunity for tree breeding taking it from the traditional locations of nursery and forest, into the laboratory. These new techniques have the potential to reduce costs and increase gains. We can learn a lot from crop and animal breeders who are leading the field but the long-lived static nature of trees means we need to be aware of the potential pit-falls and the particular challenges of forest trees.

F.2. Why bother with tree breeding?

Compared to the breeding of agricultural crops and farm animals, tree breeding is relatively new, dealing with undomesticated species. It was only in the second half of the 20th century, during a period of rapid plantation expansion that most tree improvement programmes commenced worldwide (see Zobel & Talbert, 1984). Forest managers became aware that if an investment was to be made by planting a crop, then the source of seed as well as subsequent silviculture should be the best possible, if financials returns are to be optimised.

Early objectives were to improve the adaptability of the species to the site, to increase the number of trees able to meet the crop objectives, and so to improve the economic return on the investment. This means an increase in proportion of trees per unit area that grew faster and had better stem quality

giving a greater volume of higher quality saw-logs over a shorter rotation compared to wild material. Sometimes the trait under selection might be resistance to a particular pathogen (insect or fungus) in which case successful selection and breeding might mean the difference between retaining or losing a crop (see Carson & Carson, 1989).

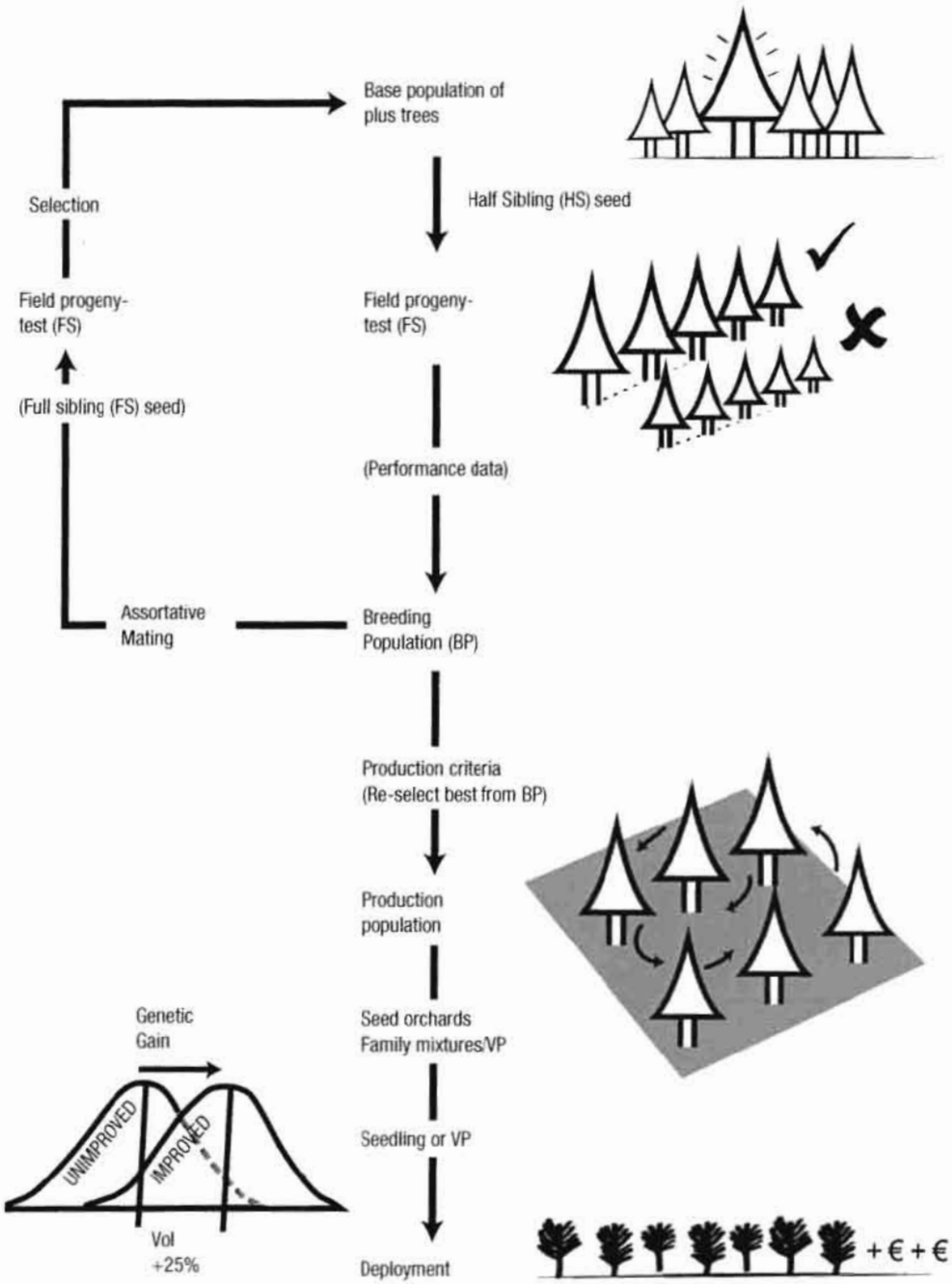
Trees differ from agricultural crops in that rotation lengths and generation intervals are much longer and they are generally out-crossing species. Trees have the advantage of being close to the wild state with a vast amount of genetic variation evolved as an aid to species-survival over their long lives. Breeding of plantation crops is particularly important today as trees are seen as a means to help mitigate the increase in global temperatures by sequestering CO_2 from the atmosphere. The more carbon that is bound-up in long-term end-products from a given standing volume (*e.g.* construction), the less is the carbon-waste from the site (slash - or fire-wood) and the longer the time interval before the bulk of the carbon re-enters the atmosphere. Forest managers and saw millers want crop uniformity which is well fitted to their end-purpose; tree breeding has a major role to play in meeting that objective.

Breeding programmes the world over followed traditional lines through the 1950s to 1990s. Whether it was pines in south east USA, spruce in Britain, eucalyptus in Brazil, or birch in Finland, the approach was to find hundreds or even thousands of superior individuals (**plus- or candidate-trees**) in the forest, raise half-sibling seedlings from open- or wind- pollinated seed collected from these superior trees which are then planted in appropriately-labelled family groups in field trials for comparison (**progeny tests**). These trials would also be replicated across a range of representative sites in order to investigate interactions with different environments and usually contained a few standard controls (unimproved material) against which progress (genetic gain) could be measured (see Williams & Matheson, 1995).

Sometimes **untested** seed orchards, which yield the improved seed for deployment, were planted composed of **graftet**-copies of plus-trees based on **phenotypes** alone. Otherwise **tested** orchards were established following collection of data in the progeny trials and re-selection of the best plus-trees after estimation of their breeding value. Rates of genetic gain per year would depend on the biology of the species; mainly when it flowers and how early final selections can be made. Shorter rotation species which flower early would allow for a faster turnover of generations than long-rotation species that flower later in life. Thus genetic gains per unit time tends to be higher for southern pines (Li *et al.*, 1999) compared to northern spruces (Lee & Matthews, 2004), although genetic gains per rotation tend to be similar. See Mullin *et al.* (2011) for a review of the economics and achievements of conifer breeding programmes from across the globe.

Many studies have found tree breeding to be an excellent research investment giving highly favourable cost benefit returns (*e.g.* Talbert *et al.*, 1985) and the justification of the work is now rarely questioned (White *et al.*, 2007). They are however, expensive to start and require considerable up-front investment mainly in the form of expensive field trials. They are long-term in nature and are slow to give a financial return due to the time required to carry out the breeding and then harvest to the trees grown from the improved seed gathered in seed orchards. However, the benefit is spread over many hectares so that if many thousands of hectares are planted annually, the return can be huge. Breeding programmes are usually undertaken either by government research organisations on behalf of the country (generally longer rotation species) or a mix of government and industry partners in a co-operative partnership (generally shorter rotation species; *e.g.* Southern Tree Breeding Association in Australia and Radiata Pine Breeding Company in NZ).

Figure F.1 outlines the steps taken in the breeding of Sitka spruce (*Picea sitchensis* (Bong) Carr.) in Great Britain as carried out by Forest Research which is a Government Agency. The programme follows traditional lines and is typical of the first generation of breeding for many different species. Sitka spruce has a generation turn-over of around 25 to 30 years and so although the programme started in the early 1960s, the first improved material became available in the early 1990s; mid-rotation gains for volume



Note: Final selection in field trials are made when the trees are10 to 12-years old. Sitka Spruce is a late flowering species and the generation interval is around 25 – 30 years. Selection has been for growth rate; stem straightness and preventing a reduction in wood density. See (Lee and Connolly; 2010) for more details.

Figure F.1. The steps in selection and breeding of Sitka spruce as carried out by Forest Research in Great Britain.

have been predicted to be between 20 and 29%; quality gains in terms of better quality logs are predicted to be between 20 to 40% (Lee & Watt, 2012). The success of this programme has led to it supplying nearly 100% of the Sitka spruce trees planted in Britain. This is approximately 30 million trees per annum. (Visit www.forestry.gov.uk/treebreeding for an overview of this programme.)

Tree breeders have learnt that turning over a generation is expensive mainly due to the cost of establishing, maintaining and measuring field trials. The return on the investment is the genetic gain and economic value identified by selecting the best genotypes in the field trials for deployment to the field, usually after cross-breeding in seed orchards. The traditional approach of harvesting seed from a tested seed orchard limits gain to the number of parents in the seed orchard and the accuracy with which those parents were selected in the field trial. This can lead to a low **effective population size (Ne)** but the genetic gain decreases as the number of parents in the orchard increases. Breeders and forest managers often prefer to reduce the population size in order to maximise gain, but they have to strike a balance between genetic gain and diversity. There are also practical considerations such as frequency and compatibility of flowering. There is little point composing an orchard of two or only a few of the best parents if they do not flower on a regular basis or female and male flowers are not synchronous.

Some organisations have tried to address these problems by carrying out artificial pollinations between the very best selections followed by the mass propagation of the resulting seed. This is known as **family forestry**. Unfortunately as the physiological age of the donor-plant from which cuttings are taken increases, so the rooting success of the cuttings decreases. Typically donor-plants are replaced after around 6 years (Mason, 1984). Up to 1,500 rooted cuttings can be achieved from each genotype in this way, which depending on stocking rate in the forest will account for less than one or up to two hectares. And although the parents of the propagated genotypes may have known merit, the propagated genotypes themselves remain untested.

Genetic gain per generation is maximised if the best ***tested* genotypes** are deployed following evaluation of those same genotypes in field-based clonal trials. This is true clonal forestry. The system involves genotype replication (usually by cuttings or tissue culture) and planting on a number of representative sites. Establishment costs depend on the number of genotypes to be tested, the number of replicates (trees) per genotype on each site, and the number of sites but are likely to be comparable or even more than first generation field trials with the requirement for similar testing periods (see Isik *et al.*, 2004).

Clonal testing does have the advantage of increasing the accuracy with which genotypes are selected; especially for those traits with low **individual tree heritability**. These genotypes could then be mated together according to a pre-arranged design to give the next generation of material for testing and selection, or planted in seed orchards to inter-mate and give improved seed for ultimate deployment. Deployment of the actual tested genotype is not usually possible unless a system of **cryopreservation** has been developed to arrest ageing whilst the testing period is proceeding. This will have to run in parallel to a tissue culture system developed to mass produce the genotype tissue taken from the cryo-store (see Park, 2002). Organisations that have operational tissue-culture and cryopreservation systems to back up field-based clonal testing will be maximising the genetic gain available in each generation by more accurately identifying the very best genotypes before their mass replication and deployment.

F.3. What is the problem with the traditional approach to tree breeding?

The problem remains that the field trial process is expensive, the testing phase takes a long time and in the case of progeny testing, genotypes are imprecisely selected. Ideally, breeders would like to find a way of eliminating expensive field trials and identify the best genotypes at a very early age

followed by direct deployment to the field. Whilst some organisations are tackling part of this problem using clonal forestry and cryopreservation (see www.arborgen.com) the barrier to enabling very early selection remains.

F.4. How can new genomic techniques help tree breeding?

If superior genotypes can be found in the laboratory by looking at the genetic markers linked to DNA extracted from seed within months of fertilisation, the efficiencies of the processes would be improved massively. Tissue culture techniques could then be used to mass-replicate the selected genotypes for direct deployment. If just a few genotypes were selected by screening the DNA of hundreds or thousands of seeds, the genetic gain could be large and costly field trials would be avoided.

To be cost effective the early screening techniques would have to be targeted at traits which are difficult to select using traditional methods. Traits which express high heritability, or are already selected at an early age using cheap non-destructive procedures should not be the focus of attention; apart from proving the principle works. Rather, this should be reserved for traits which are hard or expensive to measure, or only express themselves late in a rotation, or are of low heritability (see Grattapaglia, 2007).

The area of genomics technology is moving fast. First there was the mapping of the human genome followed by rice, mouse and chicken. More recently the first tree has been sequenced (*Populus trichocarpa* in 2006) with other species such as *Eucalyptus grandis*, American chestnut and Loblolly pine (the first conifer) and Norway spruce expected soon. It is thought that many other tree species will follow thereafter (see Neale & Kremer, 2011 for a review).

F.5. What are the new challenges for tree breeders?

Markers for all traits? Most tree breeding programmes are multi-trait in nature *e.g.* growth rate and wood quality traits, or disease resistance and survival. Final selection is usually a balance across these traits based on their respective economic importance, heritability and genetic inter-relationships. Breeders would require genetic markers for the full suite of traits under selection or alternatively screen using genetic-markers for the low heritability traits in the laboratory and then field test the re-selected population for other traits in the field in the traditional manner using progeny or clonal tests. This may not have the advantage of speeding up generation turnover, but is likely to improve genetic gain per year due to the high selection intensity, and the two-stage screening process leading to a more accurate assessment of breeding value than could be achieved by field testing alone.

There is much discussion over what sort of DNA-markers breeders need to search for. Breeders are often looking for **Quantitative Trait Loci (QTL)** which are regions on the chromosomes that contribute towards a genotype's performance for a trait controlled by many genes such as growth rate. Chapter 2 in this book is devoted to searching for QTL; how it might be done and what their importance might be. It is expected that around 10 QTLs need to be located for each quantitative trait; each QTL being responsible for around 5% to 10% of the total phenotypic variation.

Verification trials? Until confidence in the new techniques is high, it is likely that some sort of verification field trials will be required to make sure what foresters observe in the forest is what the lab-based scientist predicted. There remains the danger that very early selection based on a few markers in the laboratory may lead to some indirect link with an unknown deleterious characteristic which may affect the long-term survival or economic value of the crop, or may simply waste selection intensity on unrepeatable results.

Markers to be site specific or stable across sites? It is possible that some markers including QTLs may be site-specific or family-specific and are not transferable across pedigrees or environments. This would make life very difficult for breeders to the extent of questioning the QTLs value. Breeders should either be aware of this or alternatively find sufficient markers which are stable across environments and pedigrees and yet still control a high proportion of the total variation.

Adaptation? Any clones for deployment, whether selected in traditional field trials or through new genomic techniques, need to be future-proofed *i.e.* plastic enough to adapt to changes in climate or site condition over their rotation. This might mean new adaptation selection criteria according to the species and predictions of climate change affecting the sites on which the species is grown. Drought tolerance, frost tolerance, and resistance to occasional water logging are all traits which may become increasing important in the future. Selected genotypes of long-lived species such as trees need to have built-in resilience to anticipated environmental changes.

Other phenotyping problems? Breeders need to be capable of screening non-destructively for other new traits of interest. Wood quality traits such as micro-fibril angle, lignin, cellulose content or modulus of elasticity are becoming increasingly important as our knowledge of what determines the strength of a piece of wood improves. Variation between genotypes in the energy lost due to competition between neighbouring trees in a clonal plantation may also become more important in the future. This is something animal breeders have begun to recognise but is perhaps an under-researched area of Silviculture as the genetic base of the deployed population decreases (Brotherstone *et al.*, 2011). Genotypic response to the tissue culture techniques subsequently used to bulk-up selected genotypes may also be an important selection trait in the future.

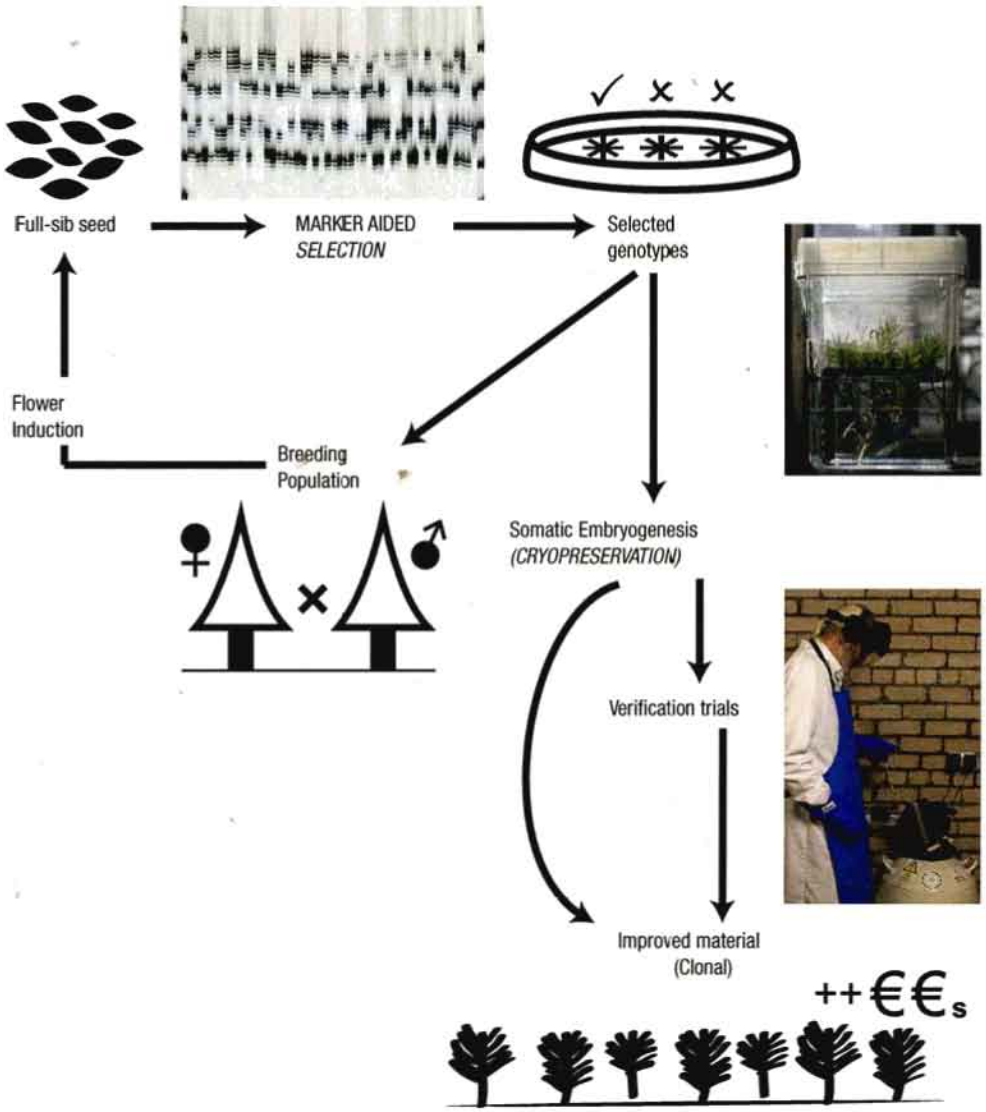
Full incorporation of these new techniques could have a big impact on the traditional breeding of Sitka spruce as described in Fig. F.1. Use of genetic markers and tissue culture could speed up breeding programmes and the rate with which genetic gain gets to the field significantly. Figure F.2 is an attempt to demonstrate the possible impact on the Sitka spruce breeding programme.

F.6. What is the purpose of this book?

This book tries to encourage the tree breeder to stand back for a moment. Think about tree breeding; what has it done, where is it going, what can be done better, cheaper, quicker, more efficiently using existing and new techniques. For the first time, this book brings together a discussion on the techniques, advantages, disadvantages, technological barriers and other challenges which face tree breeders as they consider the use of genomic selection, climate change, competition, and phenotyping as part of their breeding programme.

Chapter 1 looks at the type of markers being developed in trees, whilst Chapter 2 investigates the importance of finding QTL for use in selection. The experiences of the animal breeders will be investigated as a possible model for the way forward in Chapter 3. The need for fast through-put phenotyping (non-molecular) screening of a range of traits which may be of greater importance in the future is looked at in Chapter 4, before looking at the genetics of competition and adaptation in the light of climate change in Chapter 5. The need for genetic diversity and how much should be retained in highly selected deployment populations is discussed in Chapter 6.

This book was originally written as an output from Work Package 5 of the EU 7th Framework contract known as 'Noveltree breeding strategies' (contract FP7-211868; May 2008 to October 2012). The authors would like to thank the EU and the respective institutes for providing the funding to produce this book. Chapters are also downloadable in PDF and friendly versions from www.noveltree.eu.



Note: It is assumed that flower induction techniques advance the age of flowering to around 10-years from seed. Following control pollination between selected parents the resultant seed is screened for markers correlated to phenotypic traits (MAS) The MAS individuals may then either be deployed directly using tissue culture techniques or alternatively tested in verification trails and then deployed having held tissue in cryo-storage. Final deployment is of clonal material. The generation interval now is around 15 years.

Figure F.2. The possible impact of Marker Aided Selection and clonal forestry on the selection and breeding of Sitka spruce by Forest Research in Great Britain.

References

Brotherstone S., White I.M.S., Sykes R., Thompson R., Connolly T., Lee S.J., Woolliams J., 2011. Competition Effects in a Young Sitka Spruce (*Picea sitchensis*, Bong. Carr) Clonal Trial. Silvae Genet 60: 3-4.

Carson S.D., Carson M.J .1989. Breeding for Resistance in Forest Trees-A Quantitative Genetic Approach. Annu Rev Phytopathol 27: 373-395.

Isik F., Li B., Frampton J., Goldfarb B. 2004. Efficiency of Seedlings and Rooted Cuttings for Testing and Selection in *Pinustaeda*. Forest Sci 50: 44-53.

Grattapaglia D., 2007. Marker-Assisted Selection in Eucalyptus. In: Marker assisted selection: Current status and future perspectives in crops, livestock, forestry and fish (E.P. Guimaraes, J. Ruane, B.D. Scherf, A. Sonnino, and J.D. Dargie, eds.). FAO, Rome. pp: 251-281.

Lee S.J., Matthews R., 2004. An indication of the likely gains from improved Sitka spruce planting stock. Forestry Commission Information Note 55, Edinburgh, Scotland, 6 pp.

Lee S.J., Connolly T., 2010. Finalizing the selection of parents for the Sitka spruce (*Picea sitchensis* (Bong) Carr.) breeding population in Britain using mixed model analysis. Forestry 83: 423-431.

Lee S.J., Watt G., 2012. Improved Sitka spruce planting stock: seedlings from a clonal seed orchard or cuttings from full-sibling families. Scottish Forestry 66: 18-25.

Li B., McKean S., Weir R., 1999. Tree improvement and sustainable forestry - impact of two cycles of loblolly pine breeding in the USA. Forest Genetics 6: 229-234.

Mason B.L., 1984. Vegetative propagation of conifers using stem cuttings. 1. Sitka spruce. Forestry Commission Research Information Note 90/84/SILN, Forestry Commission, Edinburgh.

Mullin T.J., Andersson B., Bastien J.C., Beaulieu J., Burdon R.D., Dvorak W.S., King J.N., Kondo T., Krakowski, J., Lee S.J *et al.*, 2011. Economic importance, breeding objectives and achievements. Chapter 2. In: Genetics, Genomics and Breeding of Conifers (C. Plomion, J. Bousquet and C. Kole, eds.). Science Publ Inc. Enfield, NH, USA; Edenbridge Ltd., UK. pp: 40-127.

Neale D., Kremer A., 2011. Forest tree genomics: growing resources and applications. Nature Reviews 12: 111-122.

Park, Y.S. 2002. Implementation of conifer somatic embryogenesis in clonal forestry: technical requirements and deployment considerations. Ann Forest Sci 59: 651-656.

Talbert J.T., Weir R.J., Arnold R.D., 1985. Costs and Benefits of a Mature First-Generation Loblolly Pine Tree Improvement Program. J Forestry 83: 162-166.

White T.L., Adams W.T., Neale D.B., 2007. Forest Genetics. CABI publishing, 682 pp.

Williams and Matheson, 1994. Design and Analysis of Field trials for use in tree Improvement. CSIRO, Melbourne.

Zobel B., Talbert J., 1984. Applied Forest Tree Improvement. John Wiley and Sons Inc. USA and Canada, 505 pp.

Chapter 1
Genomics: Next generation sequencing: opportunities and challenges for forest genetics and breeding

Ulf Lagercrantz[1], Francesca Bagnoli[2], Thomas Källman[1] and Giovanni Giuseppe Vendramin[3]

[1] Department of Ecology and Genetics, EBC, Uppsala University, Norbyv. 18D, 752 36, Uppsala, Sweden

[2] Plant Protection Institute, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino (FI), Italy

[3] Plant Genetics Institute, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino (FI), Italy

Questions to be addressed in this chapter:

- Why sequence genomes?
- How do tree genomes differ from other taxa?
- What is next generation sequencing?
- How do we handle the complexity in a genome?
- What are the priority objectives for next generation sequencing in trees?

1.1. Introduction

The ability to determine the DNA sequence of genes and whole genomes has greatly advanced our understanding of the biology and evolution of plants. During recent years the technologies for reading DNA sequences has gone through a dramatic development, implying that it will soon become possible to sequence the genome of not only a few model species, but multiple individuals of virtually any plant species of interest. These technologies collectively referred to as next generation sequencing (NGS), have been made possible with advancements in microfluidics, nanotechnology and bioinformatics. These advances will undoubtedly have a huge impact on all fields of plant biology including tree management and breeding.

The extensive generation of DNA sequence data provides vital information on the structure and function of tree genomes, an area that has so far lagged behind that of other plant species. This is mainly due to the complexity of tree genomes, but also due to the long life cycle of trees making experiments time consuming and expensive, and the fact that the forest-tree research community is relatively small (Neale & Kremer, 2011).

Furthermore, the advances in sequence technologies will make it feasible to obtain massive data sets on the level and distribution of genetic diversity in virtually any forest tree species. This will improve the management of the existing diversity, including diversity at genes controlling traits of importance for breeding and adaptation. The information will also advance understanding of the evolutionary history of the species.

Due to the steep reduction in cost for sequencing, it will eventually become possible to sequence whole genomes of multiple individuals of a species on a large-scale, even in species with large genomes such as conifers. Meanwhile, different so-called genome enrichment techniques are being actively developed, so allowing a focus on specific regions of the genome. This approach will be of immediate interest for many forest tree species over the foreseeable future as their genomes tend to be very large,

but might also be important in an even longer time frame. This is because, for certain applications, there is a need to study more limited parts of the genome and instead increase the number of individuals assayed for a fixed cost.

Therefore, in the short term, NGS can provide genetic markers for use in marker-assisted selection and sequence data to study gene diversity. The increased throughput in terms of generating and scoring genetic markers by direct sequencing, might allow genome-wide approaches, referred to as genomic selection, to be adopted in forest trees. Together the progress in molecular technologies, such as NGS is expected to play a major role in unravelling the mechanisms of adaptation and solving practical problems related to the management and breeding of forest trees.

1.2. Trees and their genomes

Forest trees represent a major source of wood, biomass and many other products and play a dominant role in global carbon fixation. Today there is an urgent need to understand the mechanisms that control important economic and adaptive traits in forest trees. Trees play an important role from the economic and ecological point of view, especially in relation to the problems caused by global warming. The effects on forests that will result from climate change will be considerable resulting in significant changes in species range, ecosystem functionality and interactions among species. Trees are sessile and long-living organisms, therefore under global changes they will either disappear and disperse to other places via their seeds and pollen, or adapt *in situ*. It is most likely that tree adaptation strategies will rely primarily on standing genetic variation and recombination, rather than on the occurrence of new mutations, but the strategies will also depend on phenotypic plasticity, which is crucial to predict a species response to climate change (Rehfeldt *et al.*, 2001). In this regard, national and European networks have been established to assess the level of adaptation of most commercial forest tree species under various environmental conditions.

Many angiosperm tree species are characterised by comparatively small genomes (*e.g. Populus* ≈ 485 megabase pairs, Mb), while genomes of gymnosperms range from 2.2 to 39.3 gigabase pairs (Gb). The largest gymnosperm genomes are found among conifers in the *Pinaceae*, in particular in the genus *Pinus* (Burleigh *et al.*, 2012). Thus, many conifer genomes are almost seven times the human one, and 170 times the size of the *Arabidopsis* genome. In most cases, when angiosperm plants have such large genomes (*e.g.* wheat ~17 Gb) it is partly due to polyploidization (whole genome duplication events), a process that is quite infrequent in conifers. The large genome sizes of conifers have rather been attributed to the accumulation of repetitive DNA.

Gymnosperms vary little in diploid chromosome number of 2n=18-24, particularly within the *Pinaceae* family (2n=24). The high levels of synteny and macro-colinearity, and the lack of evidence for large duplicated linkage groups, as determined from genetic mapping data (Krutovsky *et al.*, 2004), supports the hypothesis of only limited large scale genomic rearrangements in conifer genomes.

Reliable estimates of the number of genes in trees are available only from *Populus trichocarpa* and *Eucaluptus grandis* (around 45,000 genes; Tuskan *et al.*, 2006; Myburg *et al.*, 2011). For other forest tree species, the number of genes has mainly been calculated based on EST clustering, resulting in considerable uncertainty. Recent large-scale attempts to sequence expressed genes, with traditional techniques as well as NGS, suggests that conifers contains a similar number of transcribed genes to *e.g.* Arabidopsis and rice.

Due to their large population sizes, high outcrossing rates and the dispersal of pollen and seed over wide areas, forest trees are expected to maintain high levels of genetic diversity within a species. However, available data on DNA sequence variation suggest that trees may rather harbour less genetic variation than most herbaceous species. Even those conifer species that are separated by more than 100 million

years have only moderate divergence at synonymous sites and low divergence at non-synonymous sites, and conifer species that are more recently diverged have high levels of shared polymorphisms (Chen *et al.*, 2010). This data suggests that the yearly mutation rates in trees are low, and perhaps that the effects of past bottlenecks (strong reductions in population size) will persist for a longer time in trees. The recovery from the loss in genetic variation following a bottleneck, will take much longer, and many tree species may still suffer from reduced variability from such historical events. However, the slow rate of divergence also means that species belonging to different environments such as the Mediterranean *Pinus pinaster* and the boreal *Pinus sylvestris* may have adapted to rather different climatic conditions using similar genomic resources. Within Europe, most nucleotide diversity of forest species such as *P. sylvestris* (Garcia-Gil *et al.*, 2003) and *Populus tremula* (Ingvarsson, 2005) is found within populations and less than 5% is found between populations. In contrast, studies conducted on a group of Central European populations of perennial herb *Arabidopsis lyrata* showed that 17% of the variation resided between populations, thus indicating a much higher divergence (Savolainen & Kuittinen, 2011).

As regards linkage disequilibrium (LD), studies focusing on genic regions in forest trees such as *Pinus taeda, P. tremula* and *Picea abies* showed a general pattern of low and rapid decay of LD (Savolainen & Pyhajarvi, 2007), in contrast to what has been observed in highly inbred species such as *Arabidopsis thaliana*. However, recent estimates from non-genic regions in the conifer *Cryptomeria japonica* suggest that LD in such regions might be high and extend over large distances comparable to those in *A. thaliana* (Moritsuka *et al.*, 2011). The scale of data provided by NGS will help to resolve this issue.

1.3. Advances in DNA sequence technologies

DNA sequencing technologies are currently going through an impressive and rapid development, with an immense increase in throughput and importantly a great reduction in price. The main invention allowing this development is a shift from traditional Sanger sequencing, where sequencing is generated one base at a time moving along the DNA, to massively parallel sequencing of millions of short DNA pieces simultaneously. This avoids the tedious cloning of individual templates. The new methods and approaches are called collectively NGS, and a few companies currently dominate the market for NGS machines. These platforms currently include Illumina's, HiSeq, ABI's SOLiD, and Roche's 454 but others are also emerging. In addition to an impressive increase in information content (in terms of produced sequences, from one Mb to several Gb), the new approaches are about 50- to 200-fold cheaper than Sanger sequencing. Concurrently, several additional techniques are in a development or release phase, promising further improvements in throughput and reduction of costs (see below). See Table 1.1 for a list of companies currently competing in developing NGS. Several recent reviews have described the rapidly evolving NGS technologies (*e.g.* Metzker, 2010; Munroe & Harris, 2010) but this chapter will not go into details.

Table 1.1
List of current NGS platforms and manufacturers

Roche 454	GS FLX+	http://www.454.com/
Illumina	HiSeq, MiSeq	http://www.illumina.com/
ABI	SOLiD 5500	http://www.appliedbiosystems.com/
Pacific Biosciences	PACBIO RS	http://www.pacificbiosciences.com/
Helicos Biosciences	HeliScope	http://helicosbio.com/
Oxford Biosciences	GridION, MinION	http://www.nanoporetech.com/
Life Technologies	Ion Torrent, Proton, PGM	http://www.iontorrent.com/

1.3.1. *Genome complexity: reduction by targeting and pooling*

Although the cost of DNA sequencing is rapidly declining, whole genome sequencing is still a costly endeavour. Meanwhile, a short-term alternative for many applications is to reduce the complexity of assaying and assembling a whole genome by target sequencing to specific regions of the genome. A number of alternatives are available depending on the purpose of the study (Mamanova *et al.*, 2010). Initially, the most widely used method was based on PCR, in which specific regions are amplified before sequencing. This method is inevitably low throughput, although specific techniques based on micro-droplet PCR (Raindance Technology) have been developed allowing a few thousand parallel PCR reactions.

A commonly used alternative is to concentrate on the transcribed part of the genome and sequence cDNA. Beside the possibility to measure gene expression (see below), this approach allows analysis of the sequence of an important part of the genome. Due to greatly varying expression levels of genes, this approach has drawbacks in terms of uneven sequence coverage and additional uncertainty in SNP calling. Methods to alleviate this problem, such as capture through oligonucleotide-base hybridisation, have been developed. Using this approach, assays to capture all human exons, have been successfully developed and exploited (Bamshad *et al.*, 2011).

To fully utilise the capacity of the NGS technology when reducing the complexity, methods have been developed to multiplex several samples in one sequencing lane. These methods are based on the addition of a specific short tag sequence (bar code) to each sample during preparation of the sequencing libraries (Meyer *et al.*, 2007). After pooling and sequencing of the libraries, sequences from the different samples can be separated out again by using the specific tag. As an example, a project aiming to sequence the regions representing 6.6 Mb of the transcriptome of *Taxus baccata,* corresponding to about 500bp in each of ~13,000 genes, in up to 100 individuals is in progress. Assuming a conservative abundance of 1 SNP every 200 bp, this "exon capture" approach, using oligoarrays, could simultaneously detect 33,000 segregating SNP at an affordable cost per sample (S.G. Gonzalez-Martinez and G.G. Vendramin, personal communication).

1.3.2. *Genomic complexity: reduction using restriction enzymes*

A technique that holds promise for detecting large numbers of genetic markers for use in association mapping or genomic selection in forest trees is based on fragmentation of DNA with restriction enzymes. The new approaches based on the use of restriction enzymes can be grouped into three main classes: reduced-representation sequencing, including reduced-presentation libraries (RRLs); restriction-site-associated DNA sequencing (RAD-seq); low coverage genotyping, including multiplexed shotgun genotyping (MSG) and genotyping by sequencing (GBS) (Davey *et al.*, 2011). All NGS methods based on restriction enzymes include (i) the digestion of multiple samples with one or more restriction enzymes, (ii) selecting a subset of fragments, and (iii) sequencing those fragments with NGS platforms. Polymorphisms in the original sequences can then be used as genetic markers, constituting either SNPs or presence/absence of fragments as a result of variation at the restriction site. These methods promise to deliver genomic data on populations with high resolution —thousands of sequenced markers across many individuals— for any organism at reasonable costs. It has found applications in species lacking a well-assembled reference genome, including Sitka spruce (P. Utrilla-Fuentes, J. Woolliams and S.Lee, personal communication), and promises to become an important technology for ecological population genomics (see Parchman *et al.*, 2012).

These so-called RAD-markers are produced by ligating adapters including tag sequence to the digested DNA, which is size selected and sequenced in multiplex. By choice of restriction enzyme and size of fragments, variable portions of the genome can be sampled to generate a suitable number of markers. One variant of this technique has been coined genotyping by sequencing (GBS; Elshire *et al.*, 2011), and

is currently applied on a large scale on maize and other crop species. GBS can provide a cost efficient alternative to the development and use of SNP chips, avoiding the initial SNP identification step and SNP genotyping assay construction. Thus, marker identification, and genotyping is performed simultaneously. Current protocols allow multiplexing 384 samples in one sequencing lane reducing the cost to around $9 per sample for many thousands of markers per sample. Ideally, the resulting sequence read are aligned to a reference genome, but in its absence, the tag sequences can be analysed *de novo* by clustering orthologous reads before identification of sequence variants.

This type of genetic marker is particularly suited for mapping in pedigrees with large amount of linkage disequilibrium. In this setting, only low read coverage is needed, as all markers need not to be genotyped in all individuals. However, these methods might have a wider and an important role in population genetics, association mapping and genomic selection in the foreseeable future. Eventually, these complexity reduction techniques may be replaced by whole genome resequencing, but for huge conifer genomes this may still be several years from now, and may only be necessary for species characterised by high effective population sizes leading to low levels of linkage disequilibrium.

1.3.3. Transcriptome sequencing (RNA-seq)

One important application of NGS is sequencing of cDNA obtained from isolated RNA. This is called RNA-seq. Such transcriptome sequencing can be used to catalogue most transcripts, including mRNA, non-coding RNA and small RNA, to determine gene structure and to estimate expression levels of each transcript in different tissues, developmental stages or different environmental conditions (Wang *et al.*, 2009). Large-scale analysis of transcriptomes has previously mainly been pursued using hybridisation approaches relying on labelled cDNA hybridised to microarrays containing sequences representing different genes. The development of such microarrays is expensive, and they are not robust to sequence variants so NGS technologies have advantages particularly where species lack micro-array platforms. RNA-seq also offers the possibility to identify different transcript originating from a specific gene, and expression can be estimated over a large dynamic range.

RNA-seq is an efficient tool to produce much data at the functional genomic level also for non-model organisms as forest trees. This methodology is a good alternative to whole genome sequencing for gene identification. Expressed sequence tags (ESTs) have a high functional information content, which is very important for gene annotation and discovery, comparative genomics, development of molecular markers, and for population genomic studies. NGS technologies have facilitated the sequencing of ESTs, removing many time-consuming steps involved in Sanger techniques.

1.3.4. Downstream handling of sequence data – Bioinformatics

Even though NGS technologies promise read length that surpass that of traditional Sanger sequencing in the future, most data currently generated by NGS is characterised by massive amounts of sequences of short length. Depending on the application, this puts great challenges on computational approaches and hardware to produce usable data. Even with current technologies and costs vast amounts of data is being generated, and the throughput is expected to continue to grow exponentially. Thus, adopting NGS technology requires considerable investments in infrastructure, including hardware (computing capacity and disk storage) and competence in bioinformatics.

Handling of NGS data poses several computational challenges. The raw data is made up of sequence reads of small segments of the genome and assembling these reads into a genome sequence (*de novo* assembly) is still a difficult task in particular for large eukaryotic genomes. Genomes with a high propor-

tion of repetitive DNA and high heterozygosity, characteristic of many plants and forest trees in particular, require high read coverage coupled with additional techniques to bridge these regions.

Due to the massive amounts of sequence data needed to assemble genomes *de novo*, new algorithms are constantly being developed. The major problem arising from using NGS data is currently the short read length. This complicates the assembly across repetitive regions, which contain identical or highly similar sequences, causing ambiguity in the assembly, since unique placement becomes impossible. Where repeats are longer than the read length gaps will appear in the assembly, and consequently assemblies based on NGS data generally contain a higher fraction of gaps. Furthermore, similar repetitive regions can falsely be collapsed causing artificial rearrangements. To tackle these problems several new *de novo* assemblers have emerged. The basic idea is to create graphs from the reads, where repeats result in branches in these graphs. The assemblers then travel the graphs and try to choose the correct branch to follow to reconstruct the genome. One way to aid the assemblers is to produce so-called mate-pair reads, where only the ends of long fragments are sequenced. These paired reads may then be used to find the correct path in the graph and bridge complex regions.

A further complication, in common with traditional Sanger sequencing, arises from high levels of heterozygosity, gene and genome duplications. These confound sequence contig assembly. A high level of heterozygosity could lead to independent assembly of haplotypes in hyper-variable genomic regions. Likewise, duplicated regions could cause errors in assembly due to the presence of genomic regions of high sequence similarity at multiple locations within the genome.

1.3.5. Genotype and SNP calling from NGS data

To identify genetic variation from NGS data, the generated randomly sequenced fragments are typically aligned to a reference sequence (potentially of the entire genome), after which SNPs or structural variants are identified. Due to the inherent problems of sequencing errors, variable read coverage and difficulties in alignment of short reads; accurate SNP calling can be difficult (Nielsen *et al.*, 2011). Although high read coverage can resolve many of these problems, the increasing demand for larger samples of individuals being sequenced puts a limit to the amount of sequence that can be produced per locus. Thus, there is a great interest in applying advanced algorithms using a probabilistic framework to increase accuracy and estimate uncertainty.

Besides errors in base calling, which are often NGS platform specific, errors in the alignment of short reads to the reference sequence may result in inaccuracy in the calling of SNPs and indels. Several alignment algorithms have been developed, including those based on hashing algorithms *e.g.* Stampy and Novoalign.

1.4. An overview of applications of NGS

1.4.1. Understanding the content and structure of tree genomes

Genome sequences will certainly supply vital information on the general structure, function and evolution of tree genomes, but also a better basis for the development of resources for genetic mapping and marker assisted selection. The characterization of whole genomes through sequencing is a relatively young search field, particularly for forest trees. The possibility of characterizing whole genomes can reveal the extent of diversity among species, and provide information on evolutionary aspects, such as duplications and expansions of gene families, dynamics of transposable elements, and identification of conserved non-coding functional elements.

Furthermore, a genomic sequence will provide the basis for studying genetic variation among individuals. This implies searching for variation in, for instance, individual nucleotides (SNPs), insertions, deletions or duplication of nucleotide stretches, and/or other large-scale genomic changes. Understanding genomic structure of forest tree will more easily allow us to correlate important traits to specific regions of the genome and lead to precise genes or nucleotide changes responsible for variation.

Until recently, the complete sequencing of the genome of forest trees, especially conifers, has been thought to be out of reach because the costs were prohibitively expensive. The advent of NGS technologies has now made possible the gathering of genome-wide sequence information from forest trees at a reasonable cost and speed. So far, with the exception of a few angiosperm tree species, most sequencing efforts on forest tree species have focused on the transcribed parts, *i.e.* sequencing of cDNA. Thus, there is a need to determine also the parts not captured by cDNA-sequencing in trees with very large genomes. For reasons of cost this has not been feasible for the huge conifer genomes, but with the increasing maturity of NGS techniques genome sequencing projects have recently been initiated for several conifers, including *Picea* spp. (http://www.congenie.org; http://www.smartforests.ca) and *Pinus taeda* (http://dendrome.ucdavis.edu/Ipgp/). The repeat content is very high in many plant genomes, and the success of genome assemblies depend on the degree of similarity of these repeats to a large extent. Recent studies of conifers with large genomes indicate that the fraction of highly similar repeats is not as large as previously expected, for which one explanation may be that repeats are old and have diverged in sequence (Kovach *et al.*, 2010).

1.4.2. Understanding the genetic diversity of tree genomes

Forest trees have been proposed as excellent experimental systems for understanding the relationship between naturally occurring genotypic and phenotypic diversity (Neale & Savolainen, 2004). From a biological point of view, genetic diversity is fundamental to ensure present and future adaptability of species. Environmental change across a wide range of temporal and spatial scales is the rule rather than the exception. Therefore, understanding patterns and interactions with environmental changes, across the range of a species, as well its historic demography and evolution is essential for developing future management strategies (see Chapter 6).

Progress in genomics of forest trees will result in large SNP datasets, extended not only to the coding regions of the genome, but also including regulatory and inter-genic areas. This collection of SNP markers will be very significant in studying the distribution of genome-wide variability within and between populations. In addition, they will provide the possibility to study the role of rare variants in breeding programmes, and to study and compare the organisation and evolution of forest tree genomes.

However, studies of sequence diversity in large forest tree genomes have so far focused on a limited portion of the genome, both in terms of the amount of sequence as well as the type of sequence assayed. In general, the focus has been on genic regions, either based on cDNA sequencing or sequencing of amplicons of genomic DNA from genes. There is thus limited information on genetic diversity and linkage disequilibrium in inter-genic regions. Such information is of great importance to design association mapping experiments and marker density needed in genome wide selection projects. The application of NGS methods, which in some cases allow the rapid whole genome sequencing of hundreds of individuals, will undoubtedly facilitate the identification of genetic variants.

1.4.3. Understanding the function of tree genomes

Despite considerable efforts to investigate gene expression under various conditions in different species, we are far from understanding the complexity of gene regulation. This remains the case even

in well-studied species such as humans and mice in animals, and *Arabidopsis* in plants. However, with or without a reference sequence at hand, NGS technologies will also facilitate more functionally oriented studies. Due to the complex nature of most traits and a lack of mutants, forward-genetics approaches have dominated the dissection of genetic variation in tree species. Still, advances in sequencing technologies may widen our understanding of gene function also in complex tree genomes. NGS technologies have the potential to facilitate studies of gene expression, identification and analysis of non-coding RNA, as well as epigenetic aspects of gene regulation.

In plant and animals several classes of short non-coding RNA (ncRNA or sRNA) also accumulate. These small RNA (sRNA) molecules that are not translated into a protein product contribute to transcriptional and post-transcriptional gene regulation (Parent *et al.*, 2012). The lengths of sRNAs make NGS instruments ideal for genome-wide sRNA discovery. Examples of this class of molecules are: transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear and small nucleolar RNA, microRNA (miRNA), and small interfering RNA (siRNA). miRNAs have been recently identified as post-transcriptional regulators of gene expression both in plants and animals, while siRNA has a major role in defending against pathogens and endogenous transposable elements. miRNA have for instance been reported to control vegetative phase change (juvenile-to-adult transition) not only in annuals but also in perennial angiosperm trees (Wang *et al.*, 2011).

Another important aspect of gene regulation is DNA and chromatin modifications. Such epigenetic phenomena, in the sense of developmental regulation of gene activity, are important for understanding tree development, but also of practical interest for forestry (Fossey, 2009). NGS technologies are well suited to assess epigenetic states (*e.g.* ChIPseq), which can increase our understanding and control of for example maturation in trees.

1.4.4. Developing genomic tools for breeding of trees - marker development and genotyping

With the advent of NGS technologies, geneticists can deal more easily with research aimed at mapping genes for quantitative phenotypes or at the identification of genes conferring specific characters also in plants with un-sequenced genomes. Many important traits are determined by several loci that act in a quantitative manner. Up to now the method used to identify such alleles was QTL mapping, which statistically associates genetic markers with specific phenotypes (see Chapter 2). Recent advances in theory and genomic tools may enable the progressive replacement of marker-based mapping approach with genome-wide high-throughput strategies, which may better predict full genetic merit (see Chapter 3). Such advances depend upon the availability of large numbers of marker loci and genotypes to associate with phenotypes.

The technology of SNP-chips has evolved rapidly to make genotyping hundreds of individuals per week with hundreds of thousands of markers commonplace. Nevertheless, the recent development of NGS technologies represents a methodological leap forward and may offer further reductions in cost.

Short term, RAD-seq approaches is an attractive option for discovery markers in large tree genomes. RAD-seq has been used successfully to develop polymorphic SNPs for high throughput genotyping of *Eucalyptus grandis, Eucalyptus globulus* (Grattapaglia *et al.*, 2011) and *Picea sitchensis*. Although data is only now emerging from using RAD-seq techniques in trees, this source of markers seems to be a promising tool to study phylogeography and phylogenetics, as recently shown in other organisms (*e.g.* Emerson *et al.*, 2010). Using high throughput sequencing of RAD fragment, these authors were able to resolve fine scale genetic divergence among populations that have been separated for less than 20,000 years, without large-scale prior investment to develop genomic resources of the studied organism.

1.5. Conclusion

Next-generation sequencing technologies is revolutionising the field of genomics allowing for generation of enough quality data to sequence and assemble thousands of genomes and gene catalogues. These extraordinary opportunities are currently limited by challenges in data analysis, thus delaying the translation from raw sequencing data to applications in research. The ability to produce gene catalogues for thousands of samples calls for a new generation of tools to be targeted towards forest trees to conduct large-scale comparative genomics studies between and within species. Moreover, as emphasised by Neale & Kremer (2011), NGS provides an important tool for increasing the molecular knowledge of complex life histories and adaptations to the environment of forest tree species.

References

Bamshad M.J., Ng S.B., Bigham A.W. *et al.*, 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12: 745-755.

Chen J., Källman T., Gyllenstrand N., Lascoux M., 2010. New insights into the speciation history and nucleotide diversity of three boreal spruce species and a tertiary relict. Heredity 104: 3-14.

Burleigh J.G., Barbazuk W.B., Davis J.M., Morse A.M., Soltis P.S., 2012. Exploring diversification and genome size evolution in extant gymnosperms through phylogenetic synthesis. J Botany, doi:10.1155/2012/292857.

Davey JW, Hohenlohe PA, Etter PD *et al.*, 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12: 499-510.

Elshire R.J., Glaubitz J.C., Sun Q., *et al.*, 2011. A robust, simple genotyping-by-sequencing approach for high diversity species. PLoS One 6: e19379.

Emerson K.J., Merz C.R., Catchen J.M., Hohenlohe P.A., Cresko W.A., Bradshaw W.E., Holzapfel C.M., 2010. Resolving postglacial phylogeography using high-throughput sequencing. PNAS 107: 16196-16200.

Fossey A., 2009. Epigenetics: beyond genes. Southern Forests 71: 121-124.

Garcia-Gil M.R., Mikkonen M., Savolainen O., 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Mol Ecol 12: 1195-1206.

Grattapaglia D., de Alencar S., Pappas G., 2011. Genome-wide genotyping and SNP discovery by ultra-deep restriction-associated DNA (RAD) tag sequencing of pooled samples of *E. grandis* and *E. globulus*. BMC Proceedings 5: P45.

Ingvarsson P.K., 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* 169: 945-953.

Kovach A., Wegrzyn J.L., Parra G., *et al.*, 2010. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11: 420.

Krutovsky K.V., Troggio M., Brown G.R., Jermstad K.D., Neale D.B., 2004. Comparative mapping in the Pinaceae. Genetics 168: 447-461.

Mamanova L., Coffey A.J., Scott C.E., *et al.*, 2010. Target-enrichment strategies for next-generation sequencing. Nat Methods 7: 111-118.

Metzker M.L., 2010. Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.

Meyer M, Stenzel U, Myles S, *et al.*, 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35: e97.

Moritsuka E., Hisataka Y., Tamura M., Uchiyama K., Watanabe A., Tsumura Y., Tachida H., 2011. Extended linkage disequilibrium in non-coding regions in a conifer, *Cryptomeria japonica*. Genetics doi: 10.1534/genetics.111.136697.

Munroe D.J., Harris T.J.R., 2010. Third-generation sequencing fireworks at Marco Island. Nat Biotechnol 28: 426-428.

Myburg A., Grattapaglia D., Tuskan G., Jenkins J., Schmutzm J., Mizrachi E., Hefer C., Pappas G., Sterck L., van De Peer Y., Hayes R., Rokhsar D., 2011. The *Eucalyptus grandis* genome project: genome and transcriptome resources for comparative analysis of woody plant biology. BMC Proceedings 5: I20.

Neale D.B., Savolainen O., 2004. Association genetics of complex traits in conifers. Trends Plant Sci 9: 325-330.

Neale D.B., Kremer A., 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet 12: 111-122.

Nielsen R., Paul J.S., Albrechtsen A., Song Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443-451.

Parchman T.L., Gompert Z., Mudge J., Schilkey F.D., Benkman C.W., Buerkle C.A., 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol doi: 10.1111/j.1365-294X.2012.05513.x.

Parent J.S., Martínez de Alba A.E., Vaucheret H., 2012. The origin and effect of small RNA signalling in plants. Front Plant Physiol 3: 179.

Rehfeldt G.E., Wykoff W.R., Ying C.C., 2001. Physiologic plasticity, evolution, and impacts of a changing climate on *Pinus contorta*. Climatic Change 50: 355–376.

Rigault P., Boyle B., Lepage P., Cooke J., Bousquet J., MacKay J.J., 2011. A white spruce gene catalog for conifer genome analyses. Plant Physiol 157: 14-28.

Savolainen O., Kuittinen H., 2011. *Arabidopsis lyrata* genetics. In: Genetics and Genomics of the Brassicaceae. Plant Genetics and Genomics: Crops and Models 9: 347-372.

Savolainen O., Pyhajarvi T., 2007. Genomic diversity in forest trees. Curr Opin Plant Biol 10: 162-167.

Tuskan G.A., Difazio S., Jansson S., *et al.*, 2006. The genome of black cottonwood *Populus trichocarpa* (Torr. & Gray). Science 313: 1596.

Wang J.W., Park M,Y., Wang L.J., *et al.*, 2011. miRNA control of vegetative phase change in trees. PLoS Genetics 7: e1002012.

Wang Z., Gerstein M., Snyder M., 2009. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

Chapter 2
Genetic architecture of quantitative traits in trees: evolution of tools and methods

Véronique Jorge[1] and Patricia Faivre-Rampant[2]

[1] INRA, UR0588, Amélioration, Génétique et Physiologie Forestières, 2163 Avenue de la Pomme de Pin, CS 40001 ARDON, 45075 ORLEANS Cedex 2, France

[2] INRA, UMR1279, Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, 2 rue Gaston Crémieux, 91057 EVRY, France

Questions that will be answered in this chapter:

- Why do we need to dissect complex traits?
- Why QTL mapping?
- How QTLs are detected in trees?
- How NGS will change the strategy identifying important genes in trees?

2.1. Introduction

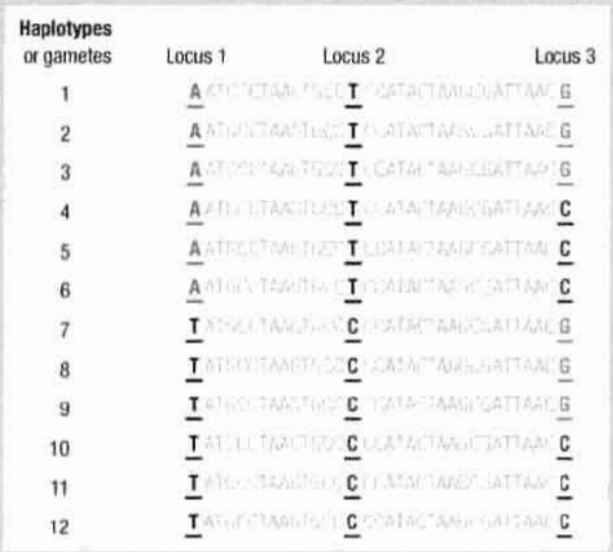
Cooper *et al.* (2009) give a precise definition of what is the genetic architecture of a quantitative trait: it "*[...] is defined by the set of genes contributing to genetic variation within a reference population of genotypes together with information on their location in the genome and the effects of their alleles on traits, including intra-locus and inter-locus interactions, environmental dependencies, and pleiotropy*". Of course, this level of information has been reached for only a very few traits such as phenology and yield in model and crop plants (Price, 2006), and even less in forest trees, because it requires a huge amount of resources to dissect all traits at this detailed level. Nevertheless, the evolution of technologies may contribute to a jump in our knowledge of the genetic architecture of traits and we will try to demonstrate it in this chapter.

There are three main objectives to the dissection of complex traits in trees, one is fundamental and the other two are more applied. Genetics of complex trait contribute to: (i) our knowledge and understanding of genome structure and functioning, and what is specific to trees; (ii) identification of markers that may be used to assist selection in **Marker Assisted Selection** (MAS); and (iii) give insight on adaptation of tree populations to their environment, their role as drivers in terrestrial biodiversity, and so help reasoning in conservation and management strategies.

There are many ways to dissect complex traits and to identify important underlying genes, going from genomics to genetics. Most important traits in trees are **complex traits**, *i.e.* quantitative and controlled by many loci. One way to dissect this kind of traits is Quantitative Trait Loci (QTL) mapping (a form of forward genetics). Basically, the tools needed for QTL mapping are a mapping population, with genotyped molecular markers and quantitative traits that have been measured and recorded in this same population. Variation must exist for the markers and phenotype. The principle of QTL mapping relies on determining a statistical association between an allele at a marker locus and a quantitative trait on a population-based manner. This association depends on the genetic distance (in cM) between the marker and the QTL, and on the effect of the QTL. The shorter the distance, the more statistically significant the association will be because recombination events breaking the association are less probable in short distances. The non-random association between alleles at two loci is called **linkage disequilibrium** (LD, see Box 2.1).

Box 2.1. Linkage disequilibrium (LD)

LD is defined as non-random association of alleles at two or more loci. The pattern and extent of LD in a genomic region is affected by mutation, recombination, genetic drift, natural selection and demographic history. In the figure below, we show an example of sequences (theoretical) for 12 gametes with 3 polymorphic loci (SNP). Allele at locus 1 and locus 2 are not randomly associated, so these two loci are in LD. In contrast, loci 2 and 3 are in equilibrium.



Contingency Tables

	Locus 1	
Locus 2	**C**	**T**
T	6	0
C	0	6

	Locus 3	
Locus 2	**C**	**G**
T	3	3
C	3	3

Statistics for measuring LD

All statistics used to quantify LD are based on the following quantity, the coefficient of linkage disequilibrium: $D_{AB} = p_{AB} - p_A \cdot p_B$; Where p_{AB} is the frequency of gametes carrying the pair of alleles A and B at two loci and p_A and p_B the frequencies of each alleles. The value of this parameter depends of the allele frequencies at loci compared, and thus cannot be compared across different pairs of loci. To standardize it, several formulas have been used and one commonly applied is

$$r^2 = \frac{D^2}{p_A(1 - p_A)\, p_B(1 - p_B)}$$

This is the square of the correlation coefficient and varies from 0 to 1. For loci 1 and 2, $D_{AT} = p_{AT} - p_A \cdot p_T = 0.5 - 0.5 \times 0.5 = 0.25$, and $r^2 = (0.25 \times 0.25)/(0.5 \times 0.5 \times 0.5 \times 0.5) = 1$. Note the value of r^2 will be the same if the alleles used had been A & C, or T & T, or T & C instead of A & T at loci 1 and 2 respectively.

The level of LD depends on several factors; one major factor is the number of generations, *i.e.* number of successive meiosis, that separate the mapping population from the initial event causing LD. This event may be a mutation or, commonly in forest trees, a crossing of two previously separate populations. We can say that this LD is at its maximum at the first generation and decreases with subsequent generation.

In the following paragraphs, we will explain what has been achieved in terms of mapping populations, markers and mapping strategies in trees. We will also show what we could expect from the recent technological advances in genomics for the knowledge of genetic architecture of complex traits.

2.2. Mapping population structure

What is a mapping population? A mapping population is the result of one or several crosses where markers and QTLs are supposed to segregate, *i.e.* the selection of the parents used for crossing is a crucial point.

To simplify the segregation and linkage analysis, ideally the cross should be made between inbred lines and different kind of populations could be generated for which standard linkage algorithms exist (Fig. 2.1). Except for a few cases (Wu *et al.*, 1998), there are no inbred lines in trees because of a large **inbreeding depression** observed in the majority of species. To overcome this biological constraint, several strategies have been used. One advantage of forest trees is that individuals are generally highly heterozygous, thus segregation for many loci is expected in the first generation when crossing two individuals randomly (see Fig. 2.1). Nevertheless, two differences between this 'two individuals' strategy and crosses between inbred lines exist: (i) not all loci segregate in the same way *e.g.* the individuals may both be heterozygous at a locus unlike inbred lines, and (ii) the **linkage phase** between two loci is *a priori* unknown. These differences are constraints but strategies to overcome them will be discussed hereafter.

One important parameter is the number of individual progeny observed since this determines the number of meiosis observed, and thus the quality of the genetic map generated and the accuracy of QTLs detected. Generally, there are no biological limits in trees to create large families (except for some incompatibilities between parents). The main constraint is linked with costs of managing/conserving large tree progenies because of the space needed. Basically, 70-300 individuals per families have been used so far, with a few exceptions in specific fine-mapping studies (Stirling *et al.*, 2001; Zhang *et al.*, 2001; Bresson *et al.*, 2011). These constraints can be also overcome using families generated in quantitative genetics/breeding programs where more than two parents were used. Theoretically, all natural tree populations would potentially be useful for genetic mapping because all resulted from crosses, except that genotype of parents and relatedness between individuals are not *a priori* known (but can be recovered *a posteriori* to some extent).

Two or three generations full-sib families. In a tentative approach to approximating the segregation patterns of populations used in inbred crops and to get the information on linkage phase, three generation pedigrees have been developed in several tree species. One or two pairs of grand-parents are crossed, and grand-parents of each pair should differ at least for the phenotype. In the second generation, only one or two parents were used to obtain an "F2" progeny. These parents were chosen based on their intermediate phenotype or based on molecular markers to approximate the configuration of real F1 individuals, *i.e.* heterozygous at almost all loci. The third generation was obtained by crossing two F1 individuals. The information of the linkage phase can be retrieved from the genotypes of the grand-parents. This strategy has been successfully used in only a few species because of the time constraints in generating these kinds of populations (Bradshaw *et al.*, 1994; Groover *et al.*, 1994; Chagné *et al.*, 2002).

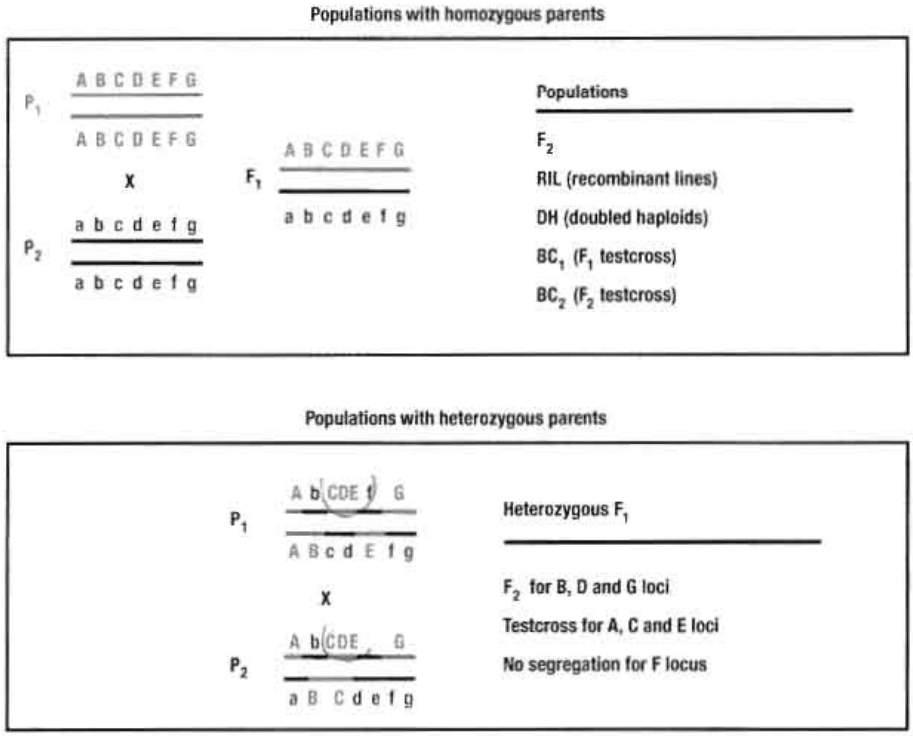


Figure 2.1. Main differences between segregation in inbred lines and in outbred species. In heterozygous parents, the linkage phase is not known *a priori* as in the F_1 from homozygous parents where all markers are linked in coupling. In the example shown, the P_2 heterozygous parent has marker A and B in repulsion phase, and D and E in coupling phase. This information can be retrieved *a posteriori* from the segregation in the progeny of the cross.

The two-generation full-sib progenies (also called F1 progenies) have been widely used as it is an easier and more rapid way of obtaining a segregating population in trees. These progenies are obtained by crossing two parents differing for their phenotype. These two parents are usually highly heterozygous and thus, segregation can be observed in the progeny for markers and for the phenotype. The constraint is that different segregation types are observed in the progeny (Fig. 2.1) and it is difficult to perform linkage analysis.

The expectations from these tree bi-parental progenies are the same as for crop plants. Firstly a limited number of QTL alleles are detected as only a small part of the existing variability of the species is explored, namely the variability existing in the two parents used for crossing. Evidence of this can be highlighted when we compare studies on different segregating populations: most of the QTLs detected do not co-localise (Blanc *et al.*, 2006). Secondly QTL effects are estimated in a specific background, and in the context of MAS it would be useful to test its effect in different backgrounds.

Populations from controlled crosses including more than two ancestral parents. Specific mating designs (factorial or cross-classified designs, Lynch & Walsh, 1998) including more than two parents were used for decades in quantitative genetics in livestock and plants to estimate genetic parameters

as **additive and dominance variances**, **heritability** and **general/specific combining ability**. Families obtained in these mating designs were often connected (half-sibs). The interest in these populations is (i) to explore more variation in the trait of interest, so detecting more QTLs and more alleles at QTLs and (ii) to test for **epistatic interactions** with genetic background thanks to the connection existing between multiple families (half-sibs).

These populations have recently received much interest in crop plants to overcome many of the limitations of bi-parental QTL mapping and association mapping populations: improving map resolution, population structure and the unknown frequency of causative mutations. The structure of these populations has been considered and optimised, and they are now called "next generation mapping populations" (Morrell *et al.*, 2012).

In forest trees, the value of these mapping populations has been explored through simulations (Muranty, 1996) but never applied. With a reasonable total population size (*i.e.* 300), the number of different parents and the connections between families give a significantly high power to the QTL detection. The main constraint to using these populations for genetic mapping in forest trees has been the difficulty and cost of development of transferable markers (*i.e.* markers that are polymorphic for most of the progenies).

Natural populations. Natural populations are also an obvious result of crosses, but the relationships between individuals in terms of ancestry are much more complex (and most of the time unknown!) than in populations resulting from few controlled crosses. Nevertheless, we can infer that these populations resulted from a large number of crosses between numerous parents, thus with a lot of recombination events and a high number of segregating alleles. The lower extent of LD that results, compared to controlled crosses, makes detecting a QTL more difficult but should result in a tighter physical linkage between QTL and the marker when a significant association is found. However, population structure as a result of demographic history (drift) can introduce a bias that can give spurious marker-phenotype associations without validation (Sneller *et al.*, 2009).

For association mapping, forest trees appear to have two main advantages compared to crop or to some herbaceous plants, providing the set of markers is sufficiently dense: natural forest populations are generally not significantly structured and the extent of LD is much lower (Neale & Savolainen, 2004). It has been hypothesised that if a significant association is found in forest trees, it may lie in the gene that controls trait variation due to this low LD.

One crucial parameter for discussion of this strategy is the size of population (number of individuals) to use, which determines the power to detect an association. In published works on forest trees, the sizes of the association populations used ranges from 290 to 961 genotypes (individuals or families): however no clear empirical relationship between the number of associations detected and the size of the population emerges from these studies. However, the structure of these populations varied a lot, from half-sibs to unrelated individuals.

2.3. Genotyping and constructing genetic maps in trees

Why do we need molecular markers? Molecular markers are binary traits which have simple segregation that follows Mendel's laws. They are used to track the segregation of loci that control quantitative traits and to localise them on genetic maps (QTLs). Molecular markers are also characterised by their abundance compared to the few existing morphological markers. The evolution of molecular biology techniques since the early 80's has given rise to a repertoire of molecular markers differing by the polymorphism revealed (protein mobility, DNA sequence, DNA repetition ...) and the principle of the

assay technique (DNA hybridization, PCR, etc.). In trees, the use and development of molecular markers has followed the same dynamics observed in crop plants, nevertheless efforts have been concentrated on major economic or model species. For example, species of tropical trees have much less abundant molecular markers.

2.4. Development and use of molecular markers in trees

We can basically identify three generations of markers:

- the pre-PCR era (before 1986), when markers were primarily **isozymes** and **RFLPs (Restriction Fragment Length Polymorphisms)** which were relatively limited in number, with laborious and time-consuming techniques;
- the advent of **PCR (Polymerase Chain Reaction)** gave rise to a multitude of techniques to reveal abundant anonymous markers (*i.e.* not based on DNA sequence information) that generally give multiple band patterns [**RAPD (Randomly Amplified Polymorphic DNA)**, **AFLP (Amplified Fragment Length Polymorphism)**] and
- the development of techniques requiring DNA sequence information [**SSR (Simple Secuence Repeat)** and **SNP (Single Nucleotide Polymorphism)** markers] and which relied on the growing resources provided by genomics. These marker types are abundant, particularly SNPs, and the sequence information make them useful for comparative mapping studies since they can be placed on a genome map.

Until recently, resources like the genome sequence and the transcriptome for trees were developed in only a few model or economically important families (Pinaceae, Salicaceae, Myrtaceae, Fagaceae). The number of SNP officially published in databases (*e.g.* dbSNP) varies between several hundred to several thousand (Neale & Kremer, 2011), but this number is rising exponentially thanks to new sequencing technologies. The SNP discovery has focused in recent years on re-sequencing candidate genes to give more insight in nucleotide diversity, patterns of selection, and linkage disequilibrium. The main features specific to trees are high levels of nucleotide diversity, the short extent of LD and some patterns of selection (Savolainen & Pyhajarvi, 2007).

The advantages of SNP are their abundance and their even distribution throughout the genome compared to other polymorphism (SSR for example). Technically, they also have also the feature of being genotyped in a highly automated way, opening the way to high throughput genotyping. Today the available techniques for SNP genotyping allows customising SNP chips containing from just a few tens to millions of SNPs. Nevertheless, these technologies have some constraints mainly linked to a step of oligonucleotide hybridisation that impede the genotyping of all SNPs that are discovered.

Molecular markers based on sequence knowledge (SSR, SNP) cited here are informative for only two kinds of polymorphisms: repetition and single nucleotide change. Other "anonymous" markers (RFLP, AFLP, etc.) can also mark **INDELs (INsertions or DELetions)**, but are not fully informative on the nature of polymorphism (SNP or INDEL). It has become obvious in other biological systems that much of the variation in the phenotype could be caused by DNA structural variation (INDEL, translocations, etc.) and epigenomic modifications such as methylation of DNA (Conrad & Hurles, 2007; Maher, 2008).

2.5. Genetic maps

The road to genetic maps is full of potholes. After marker or polymorphism discovery, markers should be tested for polymorphism in parents and Mendelian segregation in progeny, and/or tested for

"genotypability" (see remarks concerning SNP in previous paragraph). This step is time consuming and represents a bottleneck which reduces significantly the number of useful markers. It is also a step differing from crop species where the screening of polymorphism between parents guarantees the segregation in progenies. As an example, for *Populus*, we estimated that less than one third of the ~4 000 SSR markers available from the *P. trichocarpa* genome sequence could be useful for mapping in an unrelated *Populus* pedigree (Jorge *et al.*, unpublished). Additionally, we needed to throw away about 15% of the genotyped markers because they show segregation distortion and are not suitable to construct a robust framework genetic map for QTL mapping.

There are some specific problems for constructing maps of tree genomes: to overcome the problem of heterogeneous segregation types in outbred tree populations (Fig. 2.1), several strategies have been used. In one of them, only markers that follow the backcross segregation type (1:1) were retained, *i.e.* markers that are heterozygous in one or the other parent. Segregations are followed separately from each parent leading to the construction of two parental genetic maps. This strategy has been called the double-pseudo-test-cross strategy (Box 2.2; Grattapglia & Sederoff, 1994). The expectation from this strategy is that not all QTLs will be detected, especially those segregating from both parents.

One of the perspectives for linkage mapping is to shift to linkage disequilibrium mapping (see for ex. Yu *et al* 2008, Nested Association mapping). No map is needed to perform the association mapping, but background information on the position of the markers on a genetic map or on a genome would help to avoid redundancy in the set of marker used in the analysis. A second perspective arises from the Next Generation Sequencing technologies. NGS technologies have two major impacts on genetic mapping in general and of course for forest trees. First, they would accelerate marker discovery when applied for re-sequencing several individuals of the same species, even without a reference genome. As an example in *Eucalyptus*, more than 23,000 SNP have been discovered on a single experiment of transcriptome sequencing (Novaes *et al.*, 2008).

The second impact of NGS is on the reduction of steps to genetic mapping: developing, testing and genotyping markers are now possible in a single step. Genotyping by sequencing (Chapter 1) allows the discovery and genotyping of a high number of markers in a mapping family in a single step.

2.6. Strategies and perspectives and for QTL mapping

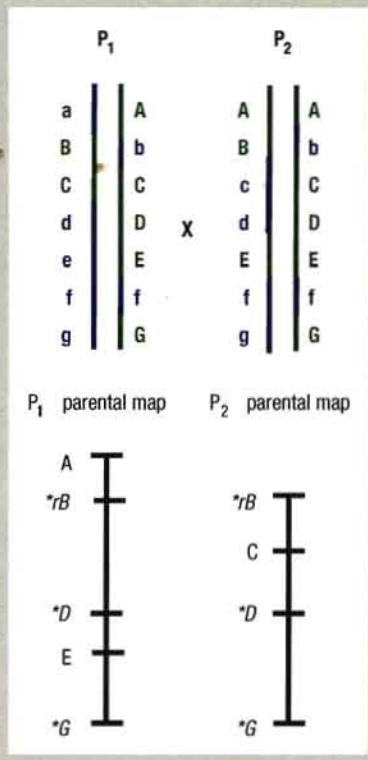
Why do we want to go to genes? In crops, it is commonly held that QTL information, *i.e.* markers flanking the QTL, will be used for marker assisted selection; in trees, due to the rapid decay of LD, the association between marker alleles and QTL alleles are expected to be broken/lost when transferring the marker from the 'QTL population' to the breeding population. Thus it is of crucial importance to identify the causative mutation(s) to implement MAS in trees.

Another important point in trees is that despite the many QTL studies reported, the number of different pedigrees and species for QTL studies can be counted on the fingers of one hand. So it means that we have explored only a very small part of the variation existing in forest trees. Using cross-classified designs, and overcoming problems of marker transferability with the large amount of marker generated by NGS could be one of the strategies for exploring more of the variation.

In genomic selection, molecular markers are used to predict breeding values, but knowledge on their association with the genes controlling the trait variation is not needed. How can QTL and gene information contribute to the improvement of genomic selection? This will be discussed in the second paragraph.

Box 2.2. Two-way pseudo-testcross mapping strategy

In 1994, Grattapaglia and Sederoff proposed a mapping approach as a general strategy to generate parental linkage maps quickly for heterozygous organisms. At this time, the type of marker used (RAPD) were frequently heterozygous in one parent, null (homozygous) in the second parent and segregating 1:1 in the progeny as in a testcross. The linkage phase of markers in the parents (coupling/repulsion) are not known unlike in inbred crop plants, so the authors used the term "pseudo"; the configuration could retrieved from the segregation analysis. As two maps are constructed, one for each parent, it is termed a two-way pseudo-test-cross strategy.



In the configuration shown on the figure, two heterozygous parents P_1 and P_2 have been tested for seven biallelic markers (A to G), the tests must include a sample of the progeny to infer the genotype of parents. After genotyping the whole progeny with all markers except F which is homozygous in the two parents, 1:1 segregating markers (A, C and E) but also 1:2:1 segregating markers (B, D and G) could be used for genetic mapping. The latter have not been extensively used in tree genetic mapping because they are not so frequent and because they are only partially informative. For example, half of the progeny are uninformative as the origin of the alleles in heterozygotes is unknown. As the phase configuration of alleles is not known *a priori*, the strategy used to recover this information is to duplicate all marker data and recode it inverted. Then when the linkage analysis is performed, linkages in repulsion phase could be detected and parental configuration recovered (marker B in repulsion with A and D).

Strategies to go from QTL to genes. A first step would be to refine our vision of QTL in trees, *i.e.* identifying more QTLs and alleles, estimating more effects than additive effects (dominance, epistasis, etc.). Tools for that are available for forest trees: breeders have developed complex pedigrees for quantitative genetics (cross-classified designs, see 2.2) and as forest trees are only very recently domesticated, a large amount of natural diversity is still available and can be used in association mapping approach. Statistical tools are also available (*e.g.* Bink *et al.*, 2008).

Given the rapid decay of LD in forest trees, it has been assumed that a too large number of markers is needed to perform efficient **association genetics**. A strategy has been to focus the genotyping on candidate genes for the trait studied. Re-sequencing candidate genes in discovery panels and in population samples has been performed not only to identify SNP but also to have insight into patterns of diversity and selection and estimate the extent of LD in these candidate genes (*e.g.* see Ingvarson (2008) or Eveno *et al.* (2008)). In trees, significant marker-trait associations explain only a very small amount of the trait variation (Muranty *et al.*, submitted), however this is not very different from many human traits. It is likely that not all the variation can be caught with the actual density of markers in these studies, but also more knowledge on candidate genes or on non-additive effects could be useful tracks to follow.

In the process of identifying important genes, the availability of a reference genome could contribute by complementing QTL studies. Thanks to the sequences of *Populus trichocarpa* and *Eucalyptus grandis* (www.phytozome.net), and with the availability of molecular markers with sequence information, we have already an idea of how many genes could be found in QTL intervals (from several hundred to several thousand). Several strategies could be used to reduce confidence intervals and to make a reasonable list of candidate genes. One classical way is fine mapping and positional cloning but it is laborious and time-consuming work, and has been scarcely applied in trees (Bresson *et al.*, 2011). Nevertheless, the accumulation of QTL data for a species provides an opportunity for comparative mapping and QTL meta-analyses (Veyrieras *et al.*, 2007). QTL meta-analysis reduces significantly the size of region of interest (*e.g.*: Truntzler *et al.*, 2010; Danan *et al.*, 2011).

Next generation sequencing (NGS) technologies are providing new ways to accelerate fine mapping and gene isolation in many species. Even in species where a reference genome sequence is not available, a strategy combining Bulk Segregant Analysis and NGS (BSA-NGS) has been proved to be efficient in identifying markers closely linked to loci of interest (Trick *et al.*, 2012).

Recently QTL mapping has been applied to transcript levels of genes, so called expression QTL (eQTL). The strategy has been called genetical genomics (reviewed in Holloway and Li, 2010). Comparison of QTL and eQTL positions gives more insight in the gene networks controlling the trait of interest. Until now, transcript levels are measured in each individual of a family using microarray technology and remain relatively costly. Today, NGS technologies can be applied quantify RNA levels in a reliable way (RNA-seq, see Chapter 1; Martin & Wang, 2011).

2.7. Why do we still want to know the genes in the context of GS?

As mentioned previously, for genomic selection the knowledge on genes controlling the trait is not needed. But the genetic models generally used in GS do not take into account potential epistatic effects or QTL clustering. QTL studies in trees have shown many examples of large effect QTLs, with a clustered pattern in the genome (especially for resistance to disease). Non-additive effects have also been described particularly in species were heterosis is explored through inter-specific crosses. Thus QTL studies and gene discovery could bring information to refine the models of genomic selection and make prediction more efficient.

References

Bink M., Boer M.P., ter Braak C.J.F., Jansen J., Voorrips R.E., de Weg W.E.V., 2008. Bayesian analysis of complex traits in pedigreed plant populations. Euphytica 161: 85-96.

Blanc G., Charcosset A., Mangin B., Gallais A., Moreau L., 2006. Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor Appl Genet 113: 206-224.

Bradshaw H.D., Villar M., Watson B.D., Otto K.G., Stewart S., Stettler R.F., 1994. Molecular Genetics of Growth and Development in Populus. 3. A Genetic Linkage Map of a Hybrid Poplar Composed of RFLP, STS, and RAPD Markers. Theor Appl Genet 89: 167-178.

Bresson A., Jorge V., Dowkiw A., Guerin V., Bourgait I., Tuskan G.A., Schmutz J., Chalhoub B., Bastien C., Rampant P.F., 2011. Qualitative and quantitative resistances to leaf rust finely mapped within two nucleotide-binding site leucine-rich repeat (NBS-LRR)-rich genomic regions of chromosome 19 in poplar. New Phytologist 192: 151-163.

Chagné D., Lalanne C., Madur D., Kumar S., Frigerio J.M., Krier C., Decroocq S., Savoure A., Bou-Dagher-Kharrat M., Bertocchi E., Brach J., Plomion C., 2002. A high density genetic map of maritime pine based on AFLPs. Ann Forest Sci 59: 627-636.

Conrad D.F., Hurles M.E., 2007. The population genetics of structural variation. Nature Genetics 39:S30-S36.

Cooper M., van Eeuwijk F.A., Hammer G.L., Podlich D.W., Messina C., 2009. Modeling QTL for complex traits: detection and context for plant breeding. Curr Opin Plant Biol 12: 231-240.

Danan S., Veyrieras J.B., Lefebvre V., 2011. Construction of a potato consensus map and QTL meta-analysis offer new insights into the genetic architecture of late blight resistance and plant maturity traits. BMC Plant Biol 11: 16.

Eveno E., Collada C., Guevara M.A., Leger V., Soto A., Diaz L., Leger P., Gonzalez-Martinez S.C., Cervera M.T., Plomion C., Garnier-Gere P.H., 2008. Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. Mol Biol Evol 25:417-437

Grattapaglia D., Sederoff R., 1994. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a Pseudo-Testcross: Mapping strategy and RAPD markers. Genetics 137:1121-1137.

Groover A., Devey M., Fiddler T., Lee J., Megraw R., Mitchelolds T., Sherman B., Vujcic S., Williams C., Neale D., 1994. Identification of quantitative trait loci influencing wood specific-gravity in an outbred pedigree of loblolly-pine. Genetics 138: 1293-1300.

Holloway B., Li B.L., 2010. Expression QTLs: applications for crop improvement. Mol Breeding 26: 381-391.

Ingvarsson P.K., 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics 180: 329-340.

Lynch M., Walsh J.B., 1998. Genetics and analysis of quantitative traits. Sinauer Assoc Inc, Sunderland, MA, USA, 980 pp.

Maher B., 2008. Personal genomes: The case of the missing heritability. Nature 456: 18-21.

Morrell P.L., Buckler E.S., Ross-Ibarra J., 2012. Crop genomics: advances and applications. Nat Rev Genet 13: 85-96.

Martin J.A., Wang Z., 2011. Next-generation transcriptome assembly. Nat Rev Genet 12: 671-682.

Muranty H., 1996. Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity 76: 156-165.

Muranty H., Jorge V., Bastien C., Lepoittevin C., Bouffier L., Sanchez L., 2012. Marker-assisted selection in forest tree breeding: which realistic scenarios thanks to lessons from crop plant experience? Submitted.

Neale D.B., Savolainen O., 2004. Association genetics of complex traits in conifers. Trends Plant Sci 9: 325-330.

Neale D.B., Kremer A., 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet 12: 111-122.

Novaes E., Drost D.R., Farmerie W.G., Pappas G.J., Grattapaglia D., Sederoff R.R., Kirst M., 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9: 312.

Prat D., Faivre Rampant P., Prado E., 2006. Analyse du génome et gestion des ressources génétiques forestière. INRA QUAE, Paris. 456 pp.

Price A.H., 2006. Believe it or not, QTLs are accurate! Trends Plant Sci 11: 213-216.

Sneller C.H., Mather D.E., Crepieux S., 2009. Analytical Approaches and Population Types for Finding and Utilizing QTL in Complex Plant Populations. Crop Sci 49: 363-380.

Savolainen O., Pyhajarvi T., 2007. Genomic diversity in forest trees. Curr Opin Plant Biol 10: 162-167.

Stirling B., Newcombe G., Vrebalov J., Bosdet I., Bradshaw H.D.J., 2001. Suppressed recombination around the *MXC3* locus, a major gene for resistance to poplar leaf rust. Theor Appl Genet 103:1129-1137.

Trick M., Adamski N.M., Mugford S.G., Jiang C.C., Febrer M., Uauy C., 2012. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. BMC Plant Biology 12: 14.

Truntzler M., Barriere Y., Sawkins M.C., Lespinasse D., Betran J., Charcosset A., Moreau L. (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.

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.

Wu H.X., Matheson A.C., Spencer D., 1998. Inbreeding in *Pinus radiata*. I. The effect of inbreeding on growth, survival and variance. Theor Appl Genet 97: 1256-1268.

Yu J.M., Holland J.B., McMullen M.D., Buckler E.S., 2008. Genetic design and statistical power of nested association mapping in maize. Genetics 178:539-551.

Zhang J., Steenackers M., Storme V., Neyrinck S., Van Montagu M., Gerats T., Boerjan W., 2001. Fine mapping and identification of nucleotide binding site/leucine-rich repeat sequences at the *Mer* locus in *Populus deltoides* 'S9-2'. Phytopathology 91: 1069-1073.

Chapter 3
Genome wide selection: a radical re-think or more of the same?

John Woolliams
The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG, U.K.

Questions that will be answered in this chapter:

- What is genome wide selection?
- Why should this be of interest?
- What are the technical issues?
- What are the prospects for forest breeding?

3.1. Using the genome for selection

3.1.1. Background

From the earliest history of domestication the breeding of species of plants and animals has focused on observing the phenotypes of populations and selecting those individuals with the most desirable characteristics. This remains the approach of tree breeders today, albeit with sophisticated techniques for organising, measuring and evaluating populations. The Introduction has described some of the characteristics of a typical modern tree breeding scheme. This has been described variously as 'black-box breeding', or 'bean-bag genetics' as it treats the genome like a sack of beans without opening the bag to discover and evaluate what is inside.

The discovery of DNA prompted attempts to move towards its use in the selection process. The broad idea is that when phenotypes, such as tree height, are measured the result is a combination of the genes and the environment, whereas genotyping a locus that is known to make one tree taller than another allows genetically superior trees to be selected for directly. This would avoid the need for complex models to account for differences within a site *e.g.* whether or not a tree was in good soil, at a low altitude, on a south facing slope, in a warm climate, or subject to shade from neighbouring trees. Technically, this is summarised by the heritability (h^2) which is the proportion of the variance between individuals measured for a trait that is genetic, and genetic progress is related to h^2. When observing phenotypes only a fraction of the variation observed is genetic, typically $h^2 < 0.4$, whereas all the variation in the genotypes is genetic *i.e.* $h^2 = 1$.

3.1.2. Development of MAS

As described in Chapter 1, the tree genome can vary from 3 billion bases up 20 Gb, and so finding which loci influence quantitative traits (termed Quantitative Trait Loci, or QTL) is a challenging task. The finding of QTL, termed QTL mapping, relies heavily upon genetic markers and this topic was developed in detail in Chapter 2. Briefly markers are loci scattered throughout the genome, whose genotypes can be routinely measured and whose position and order in the species genome is known. The usefulness of markers is because alleles at loci that are close together tend to be inherited together, and so if a marker is associated with a trait then it is an indication that there is a QTL close to the marker.

The use of this information in selection is known as Marker-Assisted Selection (MAS). In such schemes the phenotypic information that is routinely recorded is supplemented by the routine genotyping of marker

loci close to QTL that have been previously identified as being in LD. From a breeding perspective, the strength of this approach is that the variation explained by the markers used is captured with $h^2 = 1$, and so the accuracy of selection will, which in turn will increase genetic progress. This is particularly useful when it is difficult or costly to measure a trait as some discrimination among candidates is possible even if they have no phenotypes, by looking to see which individuals have inherited the favourable marker allele. In livestock breeding a number of weaknesses in MAS have undermined its application in breeding schemes, and a similar situation exists in tree breeding. These barriers are outlined in Box 3.1. In summary, MAS is only useful for improving genetic progress when markers are close to a QTL that dominates the genetic variation, and this situation is unlikely to occur for most quantitative traits in tree breeding. However, the discovery of the causative mutations underlying QTL is of considerable value for advancing our scientific understanding of tree biology and acting as springboards for further development in tree breeding and management.

3.1.3. Genome Wide Selection

The aim of QTL mapping is to localise and identify the precise variants that have a large effect on a trait, and ultimately explore their biology. In contrast, genome wide selection sets out with a different aim, by setting out to explain the full genetic variance observed in a trait through the simultaneous fitting of all the observed genotyped markers. This difference is important in that there is no *direct* implication that any single marker is a QTL, close to a QTL, or in high LD with a QTL. This concept is made feasible by the innovation of low-cost, high-throughput SNP genotyping allowing 100,000's of SNP genotypes obtained on a single individual – either by sequence methods or 'SNP chips'. Therefore it becomes plausible to assume that every QTL is close to a large number of these markers, and that the genetic variation contributed by the QTL can be almost entirely captured by the *set* of SNPs in its vicinity, through the linkage disequilibrium of the QTL with these SNP. This can be viewed as saying that the simple regression of the trait on the QTL is closely approximated by the multiple regression of the trait on the SNP markers neighbouring the QTL.

The ideas of genome wide selection and its feasibility were established by Meuwissen *et al.* (2001). The statistical framework for fitting many tens or hundreds of thousands of marker effects to a much smaller number of phenotypic records is more complex than multiple regression, and uses the statistical concept that the marker effects come from an underlying distribution. This concept is similar to the one underpinning BLUP, which is widely used in forestry for predicting breeding values by associating phenotypes with pedigree. An estimate of the effect of any single marker included in the prediction may have a wide standard error, but these errors are negatively correlated between markers, so the predicted breeding value obtained by summing over all the marker effects has the potential to be precise. It is genomic, but it remains a 'black box'.

This section was introduced with the idea that traditional selection treated the genome as an unopened sack of beans. QTL mapping and MAS pursue the idea that it is possible to open the bag and identify some beans in the bag that are of special value and then use them in selection. In contrast, genome-wide selection opens the bag but does not attempt to pre-judge the value of the beans, but instead utilises all of them, each according to their potential for predicting the breeding value.

3.2. Genome-wide selection and its potential as a disruptive innovation

The power of genome-wide selection comes from the idea that marker density will (soon) be sufficiently large that almost all the genetic variance is captured by the set of marker genotypes. In species where the availability of markers is still limited, then this will be overcome by the rapid development of SNP technology described in Chapter 1. The degree to which this can be justified for tree species is

Box 3.1. Issues related to the application of MAS in breeding schemes

1. *QTL discovery.* Experience has shown that identifying QTL is statistically challenging and many QTL reported in the scientific literature turn out to be false positives. These false positives are an extreme case of a more general problem whereby the process of statistical inference underlying QTL discovery will tend to overestimate the size of the effect. To overcome this, newly discovered QTL must be validated by testing the markers in a new population sample. A further problem is that QTL, even when confirmed, often have wide confidence regions for their location in the genome. Historically, the reason for these wide confidence regions are (a) the low density of markers, (b) linkage disequilibrium extending over large distances of the chromosome either due to the use of crossbred populations for mapping, or to low effective population size within populations. The use of new technologies to overcome these problems is discussed in Chapter 2.
2. *Estimating the variance explained.* The variance in many important quantitative traits is explained by more than a single locus. Therefore a single QTL explains only a fraction of the total genetic variance. Consequently, to account for the unmarked genetic variance, the information from the markers must be combined with the information from phenotypes. This is statistically difficult unless the true QTL genotype is known precisely. Uncertainty over how much variance is truly explained by the marker and the underlying QTL is serious because placing too great an emphasis on the marker can result in a loss of accuracy and progress, and *not* the anticipated gain!
3. *The number of loci explaining genetic variance in quantitative traits.* The experiences of studies in medicine and livestock have shown that most quantitative traits are explained by many loci explaining only a small fraction of the variance. Therefore each QTL requires large resources to identify it, validate it, and estimate its effect. For example, the 30 loci that have been discovered that affect the risk of developing colorectal cancer in humans collectively explain only 10% of the known genetic variance for risk. These studies in colorectal cancer have required billions of genotypes, on tens of thousands of individuals for a disease that is considered moderately heritable.
4. *Examples of effective application.* There are exceptions to the observation that a single QTL explains only a small fraction of the genetic variance, and in such cases MAS has been used with some success. In cattle, the DGAT locus was found to explain close to 50% of the genetic variance for fat % in several dairy breeds, and variants in the myostatin locus explain considerable genetic variance for muscling within some beef breeds. In these cases the causative mutation has been identified and breeders have used the genotype information in breeding schemes. In salmon, a single locus appears to explain 90% of the genetic variance for resistance to IPN and although the causative mutation is not known (at the time of writing) the marker information has been used routinely in commercial breeding schemes. In trees, several loci have been identified as conferring resistance to specific diseases but usage has been avoided for fear of co-evolution of the pathogen.

examined in the next section, but it has already been established in livestock species such as cattle and pigs.

The implication is that given sufficient phenotypic information then the accuracy of the genomic predictor used for selection can increase to 1. Further, since (i) the markers are sufficient to estimate the breeding value, and (ii) the high density of markers slows down the break-up of the linkages between markers and true QTL by recombination, it may be possible to dramatically reduce the scale of

Box 3.2. Factors controlling the rate of gain, ΔG

The most commonly used formula for predicting ΔG in a trait is widely referred to as the breeder's equation, with the form:

$$\Delta G = i r \sigma_A / L$$

where i is the selection intensity, r is the selection accuracy, σ_A is the additive genetic standard deviation, L is the generation interval, and ΔG is measured in trait units per year. These terms are described in more detail below.

Selection Intensity, i. The intensity describes the degree of selection that is practised. If a group of contemporaries are selected on the basis of an index of merit, then i is equal to the superiority of the indices of the selected parents compared to the average of the group, standardised by the standard deviation of the index for the group. If selection is random then $i = 0$ and intense selection in forestry will have $i > 2$. Whilst ΔG increases linearly with i, the rate of inbreeding increases with i^2, so breeding schemes may face more problems associated with inbreeding if the intensity is made too high.

Selection Accuracy, r. The accuracy is the correlation between the index used for selection and the true breeding value for the trait. It is therefore a value between 0 and 1. More information on the candidates will increase the accuracy.

Genetic standard deviation, σ_A. This is the standard deviation of the breeding values. Breeders cannot change this and so can be regarded as a biological constant for the trait.

Generation Interval, L. The generation interval is how quickly generations can be turned over. Heuristically it is the time taken to renew the gene pool. It can be calculated as the average age of the parents when their replacement offspring are born.

Among these parameters breeders can influence the intensity, accuracy and generation interval. However there are trade-offs between these parameters, and the trade-off between accuracy and generation interval is particularly important for understanding the potential impact of genomic predictors. In traditional breeding more information on a tree's merit can be obtained as the trees grow older, so r increases and ΔG increases, but if selection is delayed beyond the minimum breeding age then L increases and ΔG decreases. So if selection intensity is held constant, the maximum progress is obtained when r/L is maximised. Therefore accumulating more accurate information more rapidly is very important for breeders. Genomic predictors have the potential to make accurate predictions from seed or from young plants, so making them very attractive to breeders.

phenotypic recording once the adequacy of the genomic predictor has been established, perhaps for several generations. In this case the breeding scheme can evolve from the structure of Fig. F1 in the Foreword, to an alternative Fig. 3.1. Here an initial effort of phenotyping is carried out on a set of trees, called the training set, which are then all genotyped for the set of SNPs on the SNP chip. This (large) set of genotypes and phenotypes are then analysed to produce a genomic predictor that can be used for all individuals in the wider population providing they have been genotyped. When the predictor is of sufficient accuracy, selection can be made at the earliest opportunity without any need for measurement of phenotypes, which has the potential for greater accuracy or shorter generation intervals, and both of these will increase progress. Furthermore there would be no need to repeat this training procedure over generations.

All methods of increasing the accuracy of selection are of potential importance because of its relationship to rate of gain, but their value depends on the benefits relative to the cost, and whether or not

they extend the generation interval (see Box 3.2). However it is common in breeding schemes to have traits with one or more of the following problems.

— *Age-limited.* The trait for selection can be measured on an individual but only late in life, well after breeding age, for example tree volume at rotation age. This slows genetic progress by lengthening the generation interval, and is commonly overcome by using measurements of related traits at younger ages as 'indicator traits', however this has a penalty in reducing the accuracy of selection.

— *Low heritability.* If a trait is of low heritability (h^2) the signal to noise ratio is low, consequently for a given amount of recording the accuracy increases as h^2 increases. The impact of this is that it takes more resources and/or more generations to achieve a desired gain in a trait of low h^2 and in many schemes the economic progress tends to be driven by the traits that have highest h^2.

— *Difficult-to-measure.* Some traits are difficult to measure on an individual for breeding. Examples of such traits:
  - those that are destructive for the individual *e.g.* some assessments of resistance to pests;
  - those that are costly because of the type of measurement *e.g.* requiring expensive equipment such as infra-red, see Chapter 4; or
  - those that are costly because of the scale of measurement involved.

  Overcoming genotype by environment interactions is a very important example of this last case. The relative performance of individual tree genotypes can vary greatly depending on the environment, *i.e.* one tree may have the genetic merit to grow much better than another in one environment, but much worse in a different environment. Since the improved trees will be planted out in many different environments it is important to obtain a profile of a tree's genetic merit across representative environments. For this reason it is common for tree breeders to carry out tests on clones or offspring of the candidates. The numbers of trees that need to be planted and monitored, and long-term management of large areas of land make this an expensive option but necessary for establishing the robustness of performance to environmental influences. Genetically, the testing offers high accuracy but only at the expense of lengthening the generation interval.

— *Sex-limitation.* A further problem that is less common in forest trees than in livestock species is sex-limitation, when the trait is only measurable on one sex *e.g.* milk production in dairy cows.

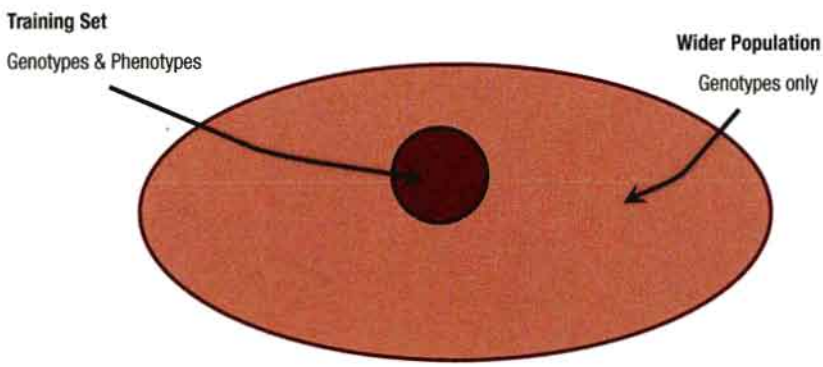


Figure 3.1. An ideal outcome for genomic selection, whereby a training set of genotypes and phenotypes provides sufficient information on the impact of each marker to predict breeding values in the wider population without the need for phenotypes.

In each of the scenarios, it is clear that if a very accurate genomic predictor was available for these types of traits before breeding age then this would be a very attractive proposition for tree breeding. It would remove the need for routine testing of clones or progeny testing, and open possibilities of directly addressing traits that to date have been considered intractable. The "burden" of phenotyping also increases when dealing with several traits, and there are examples where this burden has precluded breeding programs from considering new relevant traits. The potential savings to be gained with genome wide evaluation are therefore greater with multiple trait breeding.

In this context tree breeding has some similarity with dairy cattle breeding where historically there has been a reliance on selecting bulls by progeny testing. In dairy breeding the key traits are associated with milk production, mastitis during lactation, and the speed of re-breeding of cows after calving, and all these are age- and sex-limited. To generate selection intensity, a large number of bulls need to be tested. The cost of testing per bull is large since it requires producing and recording as many as 200 daughters per candidate bull on many different farms. Bulls can only be selected after their daughters have themselves have matured and lactated, making the generation interval six years or more. In contrast, the genomic predictor from a scheme similar to Fig. 3.1 can be obtained at (or before) birth, and the costs of genotyping bulls can be recouped by either reducing the scale of progeny testing by only testing those bulls with competitive genomic predictors, or stopping it altogether as a routine procedure. The latter case is feasible if the genomic predictor has sufficient accuracy, and the generation interval can be reduced down to its biological limit of two years. There is then the prospect of obtaining a major boost in genetic progress.

For these reasons, genomic selection is a disruptive technology, as it prompts a complete re-thinking of how to generate genetic progress. Providing the necessary investment is made at the outset to produce and record an adequate training population, then there is a *vision* encompassing:

- Reduced costs whilst maintaining or increasing genetic progress through reducing or removing the resources used for routine testing each generation.
- Breeding goals that are either better addressed or widened in their scope because it is feasible to obtain more accurate predictors of genetic merit for difficult-to-measure traits by focusing phenotyping resources upon individuals in a single training population.
- Reduced pressures on inbreeding rates arising from the use of information on collateral relatives such as half- and full-sibs, since the genomic predictor improves the accuracy of prediction of the Mendelian sampling term (see Box 3.3) of a candidate, and this in turn generates more progress for a pre-specified rate of inbreeding.

What has been described is already embraced by breeders of dairy cattle, but to explore the practicalities of this vision for tree breeders it is necessary to examine in more technical detail what is required to make the genomic predictions of genetic merit accurate, both in terms of technology and records.

3.3. A technical appraisal

Figure 3.2 shows the range of factors that can influence the accuracy of predicting a breeding value using genomic selection techniques. The scope includes attributes of the phenotypes, attributes of the genotypes, and the methods used for evaluation. This section explores these factors in more detail.

3.3.1. Assessment of accuracy

It is common when using BLUP to assess accuracy by examining the standard errors obtained for the estimated breeding values (EBV) in the data set. Often these accuracies are for individuals that

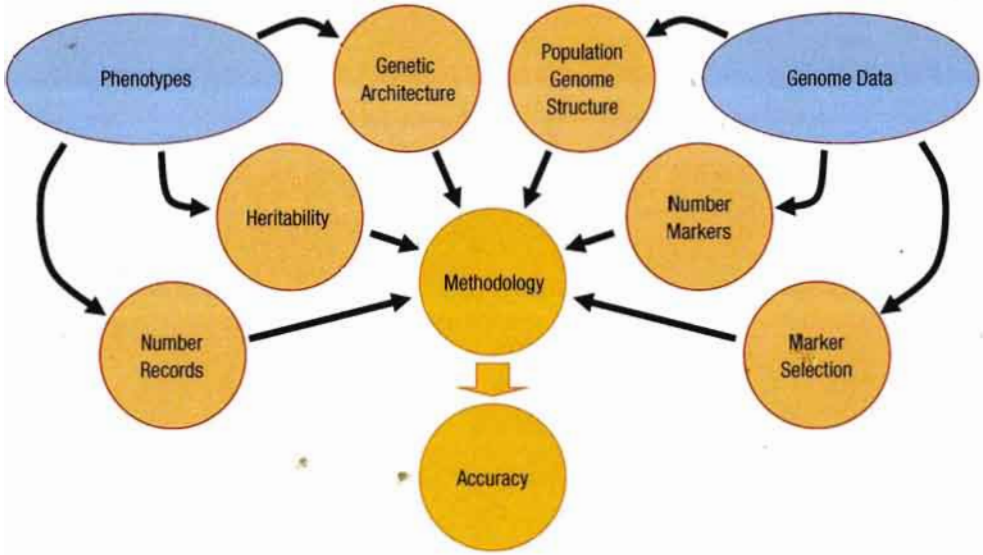


Figure 3.2. Factors affecting accuracy of genomic selection.

have phenotypes measured on them, or progeny test information, and the information on the individual will have contributed to the estimates of the parameters. A consequence of this is that the EBV will appear more accurate (*i.e.* have lower error) on average than if the individual's phenotype had not contributed the data - as the parameters have been optimised to fit the data. This overestimation of accuracy can be a serious problem when comparing models. Using the phenotypes of the selection candidates in model comparisons offers a degree of self-calibration and prevents major re-ranking among contemporaries even when the predictive qualities of the models differ substantially. With genomic selection the interest lies in its disruptive potential to predict breeding values on candidates without phenotypes, and therefore a careful assessment of accuracy is absolutely critical. For that reason, validation approaches must be used for assessment, where the data used for assessing the accuracy of prediction is not included in the data used for obtaining estimates. Cross validation is one approach to this: in k-fold cross-validation the data is split into k different sets and the accuracy of prediction is assessed in each of the sets using the remaining k-1 sets to estimate the parameters. The 'best' value of k will vary from problem to problem, but 5-fold cross-validation has often been used in assessing genomic predictors of livestock.

3.3.2. Evaluation methods

What follows is a very brief introduction to two broad approaches to genomic evaluation. Further details of the evaluation methods are found in Meuwissen *et al.* (2001), as a starting point for the considerable scientific effort in this area.

The first approach is similar to BLUP. In BLUP, the genetic variances and covariances among the breeding values of candidates are derived from the pedigree, and summarised in the numerator relationship matrix, **A**. This matrix is then used in statistical models to predict the breeding value. The pedigree provides information on the *expected* relationships among candidates; *e.g.* two full sibs of base generation parents have an expected relationship of ½ but their true relationship can vary widely and will depend on

the similarity of their Mendelian sampling terms (see Box 3.3). If the marker density is sufficiently large then the markers are capable of providing a more precise estimate of relationship between two individuals than the pedigree. Therefore the markers can be used to construct a marker based relationship matrix **G**, such that $E[\mathbf{G}] = \mathbf{A}$, where **G** describes the true relationships better than **A**. Consequently if **G** is used

Box 3.3. Mendelian sampling terms and their role in sustainable breeding

1. The breeding value of an individual i, A_i, can be expressed as the average of the breeding values of its parents plus a deviation called the Mendelian sampling term i.e. as $A_i = \frac{1}{2} A_{sire(i)} + \frac{1}{2} A_{dam(i)} + a_i$, where the mean value of $a_i = 0$. This deviation arises because each parent carries two alleles at a locus but only passes one to its offspring, and the allele that is passed is selected at random. Therefore whilst the best estimate of the value of the allele passed is the average of the values of the two alleles carried by the parent, the true value of the allele passed will be different. This Mendelian sampling term makes individuals unique – if it did not exist all full-sibs would be identical!
2. The additive genetic variance (σ_A^2) is the variance of the breeding values. The partition of the breeding value in (1) decomposes σ_A^2 into 3 terms: between sires i.e. $var(\frac{1}{2} A_{sire(i)})$, between dams, and the Mendelian sampling variance. With simplifying assumptions of random mating between sires and dams, and no selection the magnitudes of these variances are $\frac{1}{4} \sigma_A^2$, $\frac{1}{4} \sigma_A^2$, and $\frac{1}{2} \sigma_A^2$ respectively, so the Mendelian sampling is the biggest component of the genetic variance.
3. When an estimate of breeding value is made (an EBV), more information leads to more accuracy in the EBV, assuming the information is relevant. Relevant information is obtained from a range of sources including relatives. However different sources give information on different components as shown below.

Variance Component	Information Source						
	Sire	Dam	Paternal Half Sibs	Maternal Half Sibs	Full Sibs	Self	Offspring
Sires	+		+		+	+	+
Dams		+		+	+	+	+
Mendelian						+	+

The source of information can have an important impact on the rate of inbreeding (ΔF). For example, information on sire (dam) components is shared between all paternal (maternal) sibs. Therefore using such information will tend to increase the co-selection of sibs, i.e. the probability an individual's sib is also selected given an individual is selected. Co-selection of sibs will in turn increase the ΔF. Information obtained on Mendelian sampling terms does not carry this penalty on inbreeding.

4. If a parent and its offspring are genotyped then it is possible to identify which allele a parent passes to it offspring, which means that the marker provides information on the Mendelian sampling term of the offspring. Daetwyler *et al.* (2009) show that much of the increase in accuracy arising from genomic selection, as proposed by Meuwissen *et al.* (2001), comes from the increase in accuracy of the Mendelian sampling term. This contrasts with BLUP where the increase in accuracy comes primarily from including information on ancestors and collateral relatives and consequently increasing the accuracy of the sire and dam components. Avendano *et al.* (2007) show that when maximising progress for a fixed ΔF the selective advantage of an individual is the estimate of its Mendelian sampling term. Therefore increasing the accuracy of this estimate increases the gain without increasing the ΔF. This is precisely what is done by genomic selection.

in the same way that **A** is used in BLUP, the genomic EBV (gEBV) will be more accurate than BLUP. This methodology is often termed GBLUP.

It is possible to show that GBLUP is analogous to obtaining estimates of marker effects for a very large number of loci and assuming that the effects at all the loci come from the same distribution. Consider if only a relatively small number of QTL described the whole of the genetic variance. In that case most marker effects are zero and only a relatively small number are different from zero. Furthermore, for a fixed amount of data, the accumulated prediction errors increase as more effects are fitted so if it were feasible to avoid estimating the zero effects then more accurate estimates of breeding value would be anticipated. This can be achieved by adopting a Bayesian approach, and assuming a prior distribution on the number of marker loci with non-zero effects. This is the idea behind 'Bayes B' of Meuwissen *et al.* (2001). There have been many variations on this theme but they have the common principle of adopting a Bayesian approach to estimating the number of loci with small or with large effects, and allowing the information to be used more effectively when the true number of loci affecting a trait is small.

The accuracies of predicting breeding values from these two approaches can be broadly summarised:

— GBLUP approaches are indifferent to the true distribution of marker effects;
— Bayes B approaches provide greater accuracy when the number of loci contributing genetic variance is small, but declines to broadly similar values to GBLUP when this number increases.

What is meant by small or large number of loci depends very much on the genome of the species being evaluated, and this is explored below. In many sections below the discussion will focus on GBLUP justified on the basis that (i) it is a conservative assumption, and (ii) some assessment can be made of how accuracy relates to resources using deterministic predictions of accuracy. These ideas are explored in more detail by Daetwyler *et al.* (2010).

3.3.3. Attributes of the phenotypes and populations

The impact of attributes of the phenotype that influence the accuracy (r) of genomic evaluation of a trait is summarised following a formula of Daetwyler *et al.* (2009) which shows to first order $r^2 = (\lambda h^2)/(1 + \lambda h^2)$, where h^2 is the heritability of the trait and λ is the ratio of the number of phenotypes to the number of locus effects being estimated (*i.e.* n_P/n_G). Strictly, this applies if it is known that all the loci being estimated explain all the genetic variation associated with the trait, but this assumption is considered in Section 3.3.4. It also assumes that the n_G loci are independent, but this can be addressed by replacing n_G with M_e, defined as the number of independent segments in the genome (or to extend the analogy, as M_e beans in the bag). Theoretical work has shown that M_e can be approximated of by $\Sigma\ 2N_eL/\ln(4N_eL)$ where N_e is effective population size (see Box 3.4) and L is the map length of a chromosome and the sum is over chromosomes. So for a genome of say 30 chromosomes each of 1 Morgan in a population which has an effective population size of 100, Me is 1001. The value of Ne might be obtainable from modelling of family structures within a generation, or through the change in linkage disequilibrium as a function of distance between loci, or pedigree.

Using this formula some broad principles emerge:

— The accuracy depends upon the product of n_P and h^2, so if h^2 halves the number of records required must double.
— The accuracy reduces as M_e increases, and doubling M_e doubles the requirement for records.
— M_e depends upon the structure of the host genome. Species and populations maintained with high N_e have many independent segments and so require more records to achieve the same accuracy.

Therefore a target accuracy will be more easily achieved in loblolly pine with N_e ~40, M_e ~473 than with Norwegian Spruce with N_e ~1000, M_e ~7234 (Neale & Kremer, 2011).

As an example, for a trait with $h^2 = 0.4$, to obtain a target accuracy of 0.7 for loblolly pine will require approximately 1140 records, whereas 17,300 records will be required for Norwegian spruce. There is some injustice here, as those populations which have made more efforts to conserve diversity require more effort to exploit genomic selection techniques!

Referring to Fig. 3.2, one aspect of phenotypes that has not been considered is genetic architecture. One reason for this is that we have considered the accuracy of GBLUP which is indifferent to the number of loci. A second reason for this is that if dominance or epistasis were to be exploited using crossbreds, then the training data can simply consist of a crossbred population, with the data used to provide genomic predictors for each of the populations being crossed, providing the parental origin of alleles can be identified. However deployment does not always use crossbreds, and given the long rotation age of trees the possibility of cloning within a population also allows the capture of non-additive variation. Models for this are being explored (*e.g.* by the group of Gianola) but this will not be described here.

3.3.4. Attributes of genotypic data

Section 3.3.3 describes the number of records required to achieve a pre-determined accuracy in GLUP assuming the set of markers captures the full extent of the genetic variance. However a key ques-

Box 3.4. Inbreeding, rate of inbreeding and effective population size

1. Inbreeding occurs when mates have one or more common ancestors. When this occurs there is a probability > 0 that the offspring will inherit two copies of one of the alleles carried by a common ancestor. The two alleles inherited by the offspring are identical by descent (IBD), ignoring any intervening mutations. The inbreeding process causes random drift in allele frequencies and the population mean, is expected to reduce heterozygosity and genetic variance within lines, and increase differentiation between lines. It can lead to inbreeding depression in performance. It is a natural process as all populations are finite in size and some inbreeding cannot be avoided. It is measured by the inbreeding coefficient (F) which is a probability of IBD, and so lies between 0 and 1.
2. The rate of inbreeding (ΔF) is a measure of how rapidly F is increasing per generation of inbreeding. Many studies have shown that the impact of inbreeding on populations is largely related to ΔF rather than to F. For example inbreeding depression becomes more severe as ΔF increases. Therefore sustainable management of populations in relation to inbreeding should be concerned with the magnitude of ΔF not F.
3. A tree population has effective size N_e, if a population of N_e individuals randomly selected and mating at random (including selfing) would produce the same ΔF as observed in the tree population. In fact, $N_e = (2\Delta F)^{-1}$. The justification for this transformation is its conceptual use for visualising the accumulation of inbreeding in a population.
4. In managed populations of livestock, breeders are strongly encouraged to maintain $N_e > 50$ to promote long term sustainability of their populations. In contrast, wild species are unmanaged and interventions on a population-scale are extremely difficult, and Ne dropping below 500 is regarded as a serious cause for concern. Of primary concern here, and as seen in Section 3, N_e and/or ΔF are important in describing properties of the genome.

tion is how dense must markers be to expect to capture all (or nearly) all the genetic variance? This is a difficult question to answer as a monogenic trait only requires a single locus to explain all the genetic variance. However for most complex traits affecting timber production we may assume that there are a large number of true QTL explaining some of the variance. In this section an empirical approach will be adopted by looking at what has been learned from dairy and scaling this to trees.

3.3.4.1. Lessons from dairy cattle

In dairy cattle, several different SNP chips have been developed, moving from 10k SNP to 800k SNP per animal, with the majority of data obtained with 50k SNP. The experiences of the use of these chips suggest that when 20k SNP or more have been used the accuracy of genomic EBV obtained from GBLUP has been relatively stable when assessed within a single breed or population. This implies that 50k SNP are sufficient to capture all the information on QTL that obtained from a SNP chip using GBLUP. This does not imply that this would be sufficient for tree populations.

3.3.4.2. Scaling rules

In a population in mutation-drift balance the linkage disequilibrium between two markers as measured by r^2 is a function of $N_e c$ where c is the recombination fraction. Ignoring complicating factors such as selection, this implies that to obtain approximately the same amount of information when N_e is doubled the marker density must be doubled. This suggests an empirical approach to comparing populations is to scale marker densities by N_e for the breeding population. The question on what is the relevant generation for estimating N_e is immediate, and some evidence from livestock suggests the most recent 5 or 6 generations are the most relevant. Therefore to revisit the cattle example, it will be assumed that 50k SNP are adequate to capture the relationships on a 30 M genome and that N_e ~50 per generation. This implies the density required to capture relationships adequately is ~17 SNP per cM, or ~$(0.3N_e)$ SNP/cM. Translating this to Norwegian Spruce with N_e=1000 and a 24M genome is ~720,000 SNP. Such numbers are feasible for producing genotyping chips but it requires an advance on the current scale of SNP libraries (at the time of writing in spring 2012).

3.3.4.3. Capturing genetic variance on SNP chips

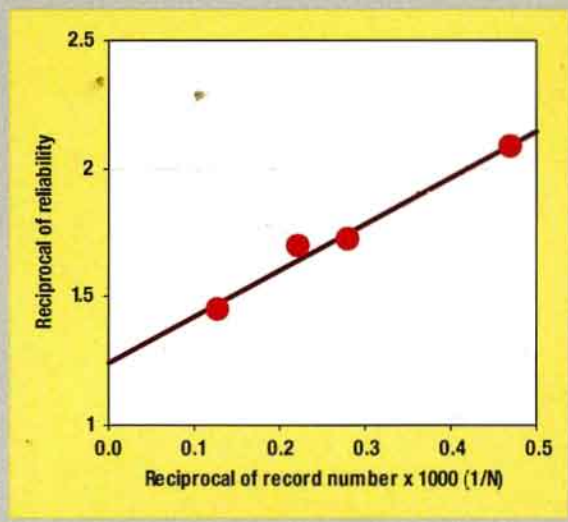
In much of the above it has been assumed that a set of SNP has the capacity to capture all the genetic variance. Box 3.5 suggests that this is not the case: using dairy cattle as an analogy the 50k SNP chip is capable of capturing only slightly more than 80% of the genetic variance, which implies the maximum accuracy is ~0.9. In turn, this demonstrates that a set of SNP of poor quality for explaining the genetic variance of a population is a structural problem that *cannot* be overcome by accumulation of data. If q is the maximum accuracy of the SNP chip then the predicted accuracy will be $q\, r_D$ where r_D is the accuracy defined in section 3.3 above. There are a number of reasons for this failure, and one of these is that the density may be insufficient to capture LD with all QTL (which may be unknown SNP or copy number variations, including deletions).

3.3.4.4. Re-estimating predictions

Genomic evaluations rely on LD with QTL, but each generation the LD between a marker locus and a QTL will decay, on average, as recombinations will move the loci towards equilibrium. If QTL are closer together then the rate of decay of LD will be slower. As the patterns of LD change, so the prediction equa-

Box 3.5. The extent of genetic variance captured by a SNP chip

The fraction (q^2) of additive genetic variance (σ_A^2) explained by SNP chips is a topic of high interest since: (i) it determines the accuracy of genomic selection and so is relevant to all breeding schemes; (ii) understanding its origin helps to set targets for technology development; and (iii) it places into perspective q^2 derived from GWAS – related to 'missing heritability'. Daetwyler, Van Raden and Woolliams developed an estimate of q^2 from the expression for squared accuracy (r^2) of genomic evaluation discussed in section 3. They showed that for a series of evaluations for a trait with different numbers of phenotypes, the reciprocal of r^2 is linearly related to the reciprocal of the number of records used and in this regression, the intercept is the reciprocal of q^2. In testing this with USDA cattle data, it was estimated that for the Illumina Bovine SNP50 BeadChip q^2=0.8 (s.e. 0.034), leading to a maximum accuracy ~0.9.



q^2 is an expectation and an attribute of the chip. For particular traits the ultimate r^2 using a chip may be more or less than the q^2, for example if a single SNP is included in the chip and explains all σ_A^2, then r^2 will ultimately be 1. However for most traits used with a chip this will not be the case, and the theory provides a prediction of the adequacy of the chip and a realistic assessment of the impact of increasing the number of records on achieved accuracy.

tion using the SNP will change and the predictions using the data from the original training set will become less accurate. Therefore one expected benefit of using marker SNP at very high density, so that many SNP will be close to a QTL, is that the accuracy of predictions will be more robust over generations. If this is not the case then the training data will need to be refreshed after a small number of generations. However, given the relative lack of experience in how genomic evaluations change over time in real populations, it would be advisable to update and review the quality of the predictions regularly.

3.3.4.5. Minimising genotyping costs

To carry out genomic selection amongst a large number of candidates it is necessary to genotype all the candidates. This can become expensive if all the genotyping is done at high density, but work in

livestock has shown that it is possible to reduce these costs by only genotyping a small subset of loci for each candidate providing their parents are densely genotyped. This is possible if the SNP genotyped on the candidates are distributed evenly across the genome and are chosen to have high allele frequencies. The genotypes for the untyped loci are then inferred using the parent data. This process of imputation is an important tool but further details are beyond the scope of this chapter.

3.3.5. Summary

In summary the existing theory on the resources required for effective genomic selection indicate that they will depend on the species, in particular the effective population size over recent generations. Large effective population sizes require more records to be collected for the training population to achieve a particular accuracy and require denser markers to capture the genetic variance through linkage disequilibrium between the markers and the QTL. Using too small a number of SNPs will fail to capture the full genetic variance, and the predictive accuracy will deteriorate more rapidly over generations – these problems cannot be overcome by using more records.

3.4. Prospects for tree breeding

Given the outcome of the technical review of the previous section, it is now possible to assess how realistic the opportunities for genomic selection are. One of the pervasive factors in the technical review is the extent of linkage disequilibrium, which in turn is governed by the history of N_e over generations, since it determines the density of SNP and the magnitude of the recording task to obtain adequate training data. It is broadly accepted that most silvicultural species have large N_e, which is not the outcome that is most promising for application of genomic selection. The number of generations of managed breeding is small, perhaps one or two, so much linkage disequilibrium over small distances is determined by the N_e found in the wild forest, which is assumed to be large. This is supported by evidence of N_e estimated in some studies using pair-wise linkage disequilibrium for small numbers of SNP found within genes, although studies of large numbers of anonymous SNP distributed throughout the genome have yet to be carried out and these might tell a different story. This is not true for all species, for example the N_e of Loblolly Pine has been estimated as low as 32. Therefore some species appear well placed to make use of genomic selection technology.

However the implication of this observation is that for typical forest species, such as Norwegian Spruce or Maritime Pine, it would be feasible but costly to obtain sufficient genotype/phenotype records for a training set to achieve a high predictive accuracy for a complex trait (*i.e.* with large numbers of QTL). This would likely preclude the immediate adoption of full-scale genomic evaluation in such populations. However USDA kick-started genomic selection in cattle by providing funding for 50 k SNP data for many thousands of bulls, which serves as core data for later initiatives. However even though the cost of obtaining >500 k SNP data now is similar to obtaining 50 k data a few years ago, this is still a significant cost. One solution might be cost sharing through international collaboration but a potential problem with this option is the large potential for genotype by environment interactions in observed within many forest species – the set of genotypes promoting good performance in Scots Pine on the west coast of Scotland may be very different from those promoting good performance in Finland. In dairy cattle the genetic correlation in bull performance across USA and Europe is very high (>0.80) for most important traits, but can be much lower in forest trees as they are static and subject to less management of the production environment. Therefore cost sharing through international sampling and funding may not be an effective way forward in many cases.

Another solution for forest species might be to reduce the effective population size for the elite populations of the key species, where they are not already low. In livestock, sustainability over the long term would argue that N_e should be maintained above 50, and in trees the arguments are likely to be whether this is high enough. Nevertheless, to this extent it would be appropriate to reduce N_e for elite forest populations for future generations, although the benefits from this will take some time to emerge given the relatively long generation intervals of forest trees

One influential strategy in the history of tree breeding was the 'multiple population' strategy which argued for many breeding populations each targeted to particular environments. This has been followed to the extent that national breeding schemes are largely distinct, but within species the approach has been to target robust performance across environments. In principle genomic selection opens opportunities for the multiple population strategy to be followed within national populations due to the reduction in long-term requirements in testing resources for a single population. This would have benefits in that multiple sub-populations each sustainable in their own right will increase N_e for the national populations. The drawback is the large scale training set that would be required to establish the sub-populations.

Recent work has suggested some new possibilities through the use of within-family genomic evaluation *e.g.* predicting the ranking of breeding or phenotypic merit among full-sibs in the absence of phenotypes – and in large families, as in trees, substantial selection intensities can be achieved even if selection is only utilising the genetic variance found within families. In a recent study in Sitka spruce the *phenotypes* in bud burst and height were predicted with accuracies 0.40 and 0.58 from genotypic and phenotypic data of 497 of their full-sibs – this corresponds to much higher accuracies for predicting the breeding values. The reason for this accuracy is the long extent of disequilibrium that exists within a full-sib family. However it remains to be assessed how best this can be integrated into breeding structures.

In summary, if tree populations have been through several generations where N_e has been relatively low (say <100), then genomic selection offers substantial opportunities now. These opportunities are radical in offering expanded breeding objectives and a re-design of breeding schemes. As recent N_e increases so these opportunities are overtaken by the start-up costs of developing a training set of sufficient scope and size in an industry where margins are tight. In such cases genomic selection will become more feasible over time as the costs of genotyping reduce. Even if practicalities postpone implementation, important preparatory steps can be undertaken in: (i) reviewing N_e of elite populations to keep it at the minimum required for sustainability; and (ii) collecting a biobank of material from which phenotypes, both traditional and novel, can be associated with genotypes as soon as technology and costs allow.

3.5. Conclusions

At the start of the chapter, the question was posed whether genome-wide selection was a radical rethink or more of the same. In a sense, both alternatives are true. It is "more of the same" with the image of the bean-bag, in the sense that our evaluation remains a black box, or a black bean-bag. However it is radical when aiming at supplanting the heavy load of regular phenotyping with all the consequences that this could bring in terms of re-thinking the breeding scheme.

References

Avendano S., Woolliams J.A., Villanueva B., 2004. Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. Genet Res 83: 55-64.

Daetwyler H.D., Villanueva B., Woolliams J.A., 2009. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3: e3395.

Daetwyler H.D., Pong-Wong R., Villanueva B., Woolliams J.A., 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185: 1021-1031.

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

Neale D.B., Kremer A., 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet 12: 111-122.

Chapter 4
Phenotyping for the future and the future of phenotyping

Ricardo Alia[1,3], Juan Majada[2]

[1] INIA Forest Research Centre. Department of Forest Ecology and Genetics,
Avda A Coruña km 7.5, 28040 Madrid. Spain

[2] Forest and Wood Technology Research Centre (CETEMAS),
Finca Experimental «La Mata» Grado 33820, Spain

[3] Sustainable Forest Management Research Institute. University of Valladolid-INIA,
Avda A Coruña km 7.5, 28040 Madrid. Spain

Questions that will be answered in this chapter:

- Why are high-throughput phenotyping technologies needed?
- What makes forest trees unique in genetic evaluation programs?
- Why is there a need to characterize phenotypes by biotic and abiotic factors?
- Which traits could be of interest in the future breeding programs?
- What attributes are needed for a technology for phenotyping in the future?
- Why do we need phenomics platforms in Forest tree breeding?

4.1. Summary

The focus of genomic-based breeding is changing from genotyping to phenotyping, due to the decreasing costs and increasing availability of genomic resources. This change raises the challenge of developing new and more efficient methods of phenotyping. Emerging traits, related to adaptation and/or new products, need to be addressed in breeding programs, and new trait definitions, based on the knowledge of the trait architecture and also on the development of high-throughput phenotyping methods should be incorporated into genomic-based breeding programs. Here we describe the main aspects related to phenotying in future breeding programs, especially those more novel traits that could be of interest for inclusion into future breeding objectives, and some new methods for high-throughput phenotyping that are being developed.

4.2. Introduction

A key goal of biology is to understand phenotypic characteristics, such as growth, production, pest and disease tolerance, or adaptation, in order for them to be used in breeding and conservation programs. It is possible to define a 'genotype-phenotype' map produced through a complex web of interactions between genotype and environment, which is not possible to decipher without detailed phenotypic data (Houle *et al.*, 2010). But also, we need to know in detail the multi-dimensional variables that form the environment (as a combination of biotic and abiotic external factors) that allow these interactions to be studied objectively.

There is a great need to meet the human demand for forest tree products (timber and non-timber forest products) and genomics-based breeding approaches can accelerate traditional approaches that are based solely on the phenotype. *Understanding genotype - phenotype relationships*, and *developing genomics-based tools to assist in the genetic improvement of production populations* have been pointed out as 2 out of the 4 major objectives of forest tree genomic research (Neale & Kremer, 2011). However, genomic-based breeding approaches need a change of scale: the time and money required for the

collection of genotype data were critical in the past, but the increasing availability of inexpensive DNA sequencing and genotyping methods should prompt researchers to shift their attention to more traditional concerns of experimental design and collecting phenotypes. Nowadays, our ability to characterise phenomes, *i.e.* the full set of phenotypes of an individual, begins to lag behind our ability to characterise genomes. Therefore, we are now at the point where the phenotypic information will be the limiting factor in deciphering this 'genotype - phenotype' map, and it is necessary to develop cost-effective phenotyping with the aim of high-throughput technologies recording multiple traits in parallel.

In tree breeding programs we can define the economically important traits that constitute the aggregate breeding value or economic genetic value. Those traits may or may not be measured on the trees in the field. For example, rotation-age stem volume, and pulp yield may be the target traits in the breeding program, while juvenile height, estimated wood density, health and straightness may be the measured traits upon which selection is based (White *et al.*, 2007). In the forthcoming years these breeding objectives will be revised because of new emerging demands and constraints. We will need to re-think many different aspects to take advantage of the genomic-based approaches: the traits to measure at our experimental sites or in our labs or both; the design of our experimental sets by combining the testing of both training populations and the and wider population; better choice and characterisation of the environments to test the material for both outdoor and controlled testing; and search for new cost-effective technologies to improve and speed the phenotyping process.

4.3. Forest trees in genetic evaluation programs

There are some specific questions that need to be addressed when working with forest trees. Forest trees are difficult to phenotype, as they are big and long-lived. This fact poses some problems in the evaluation programs, because a huge amount of the resources is devoted to the establishment, tending and measurement of experimental tests over a long period of time. For example, a typical experimental site may include 8,000 individuals occupying an area of 3.5 ha for a period of 20 years, assuming a spacing of 2m × 2m. For genomics-based tree breeding, we will need to increase the sampling size (usually more than 1,000 unrelated individuals), and a different evaluation strategy needs to be applied.

Breeding objectives always involve consideration of multiple traits, even though in some situations the weight given to single trait may dominate. Plant breeders find the concept of deriving relative economic weights of traits difficult to apply and often replace the optimal index by restricted or desired gains indices (Sölkner *et al.*, 2008). So, as pointed out by White *et al.* (2007), tree improvement programs have not been as effective as animal improvement programs in the explicit specification of breeding objectives. A general idea of the overall goal is often formulated in tree improvement programs as a list or high-priority traits.

Evaluation of genetic material is a crucial phase in all the breeding programs. We need to establish many genetic trials to estimate breeding values for the different traits of interest for our breeding objectives. Usually, because of economic constraints, we will manage a very limited set of experimental sites (<10), for evaluating many genetic candidates for selection (>100). The progress in recent years, by using more efficient field designs (*e.g.* row-column designs, α-lattices, see (Williams *et al.*, 1996), and data analysis methods (including spatial analysis as a routine, post-blocking) have improved the efficiency of field evaluation for many traits (Dutkowski *et al.*, 2006), but we do not expect any further significant improvement in the forthcoming years.

Another aspect of evaluation is the general low number of breeding generations available for most of the species. Therefore, in most of the cases, the pedigree is not precisely known, and cannot be efficiently used to improve estimation of the breeding values. With the new approaches, we will have a very precise estimation of the molecular pedigree of the trees, summarized in a matrix of molecular relationships (**G**,

see Chapter 3; Fernández *et al.*, 2005; Blonk *et al.*, 2010), as a result of the information from numerous genetic markers used to characterise each individual. Also, controlled crosses and progeny testing can be simplified by obtaining structured pedigrees for the breeding and testing from wind-pollinated progenies and using high-resolution markers instead of the conventional approaches described in the foreword ('breeding without breeding' or the BWB- approach, see El-kassaby & Lstibůrek, 2009).

The growth and phenotyping of trees in laboratory environments, or at early ages in the field, has in many cases demonstrated only limited value (Neale & Kremer, 2011). Usually, the interest is in the gain during the full rotation, but only a fraction of what is possible has been obtained by using very early testing in many different traits (Apiolaza, 2009). We usually do not know precisely the pattern of age-age correlations for the different traits of interest, and how the environmental factors are affecting the trait under evaluation. Also, the correlations for the same traits measured in climate chambers and field tests are not high enough to be routinely applied in traditional breeding programs (Jansson *et al.*, 2005), at least for some that are most studied traits such as height. However, in a genomic-based breeding, we should expect an increase in genetic gain per unit of time, according to some predictions. Under such a scenario, we might expect an increase in the use of early testing under controlled conditions (greenhouse, glasshouse, Phytotron), combined with new technologies and bioassays, where we can test many candidates for selection at a reasonable cost.

Therefore, we need to define a clear evaluation strategy, by using different types of evaluation: outdoor for traits related to the genetic economic value, and mainly for checking the effectiveness of genomic selection, and early testing experiments to advance in the scope of the breeding program (Fig. 3.1, Chapter 3). As stated in Chapter 3 "With genomic selection the interest lies in its disruptive potential to predict breeding values on candidates without phenotypes, and therefore a careful assessment of accuracy is absolutely critical".

Foresters have been measuring phenotypes in test plantations for a long time, but usually in a limited number of traits (for example, tree height, tree diameter, presence or absence of disease), *i.e.* easy to measure, and non-destructive traits. This is a clear limitation if we need to define the 'genotype-phenotype' map, taking advantage of the new technologies. In this case, the use of as many traits as possible will increase our power to predict the performance of some traits, but also complicate the analysis of the quantitative variation in multiple dimensions and make demands on the size of the datasets upon which the predictions are based.

Under outdoor conditions, evaluation is made over a different, usually limited, set of environments, and over a long period. Therefore, the cost limits resources and it is difficult to disentangle the effect of the age and the climatic conditions. In agronomic studies, replication over space usually allow good estimates of response functions, and replication over time can usually be achieved by establishment of new genetic trials; whereas in forest-tree breeding, replication over space is usually much reduced, and replication over time often means evaluation of the same tree over its lifetime. The extensive information in broad-range field experiments has been used for estimating response curves along different climatic conditions (Rehfeldt *et al.*, 2002). However, the auto-correlation arising from measuring the same tree at different ages has sometimes been neglected, resulting in a bias in the estimates (Apiolaza, 2009). As a result, in forest trees we need to characterize a tree under a complex combination of biotic and abiotic factors interacting with ontogenetic effects.

4.4. Phenotyping for the future

Which characteristics do we need for the future forest, and more specifically for the plantations of the future? We will focus upon two types of constraints: climate change and the future demands of end-

users. Climate change will affect the growth, distribution, among other established aspects of performance species (Kolström *et al.*, 2011), but it will also interact with stressors (*e.g.* wind, drought, invasive plants, animals, pathogens, and altered fire regimes) to raise new challenges that could be addressed by tree breeding programs. Among future demands, there is expected to be increased demands placed on minimising processing costs (*e.g.* by reducing the lignin content for the pulp and paper industry), or increasing the quality (*e.g.* by increasing the stiffness in solid wood). According to Vanhanen *et al.* (2007) maturing European markets for forest products will place pressure on the forestry sector to renew itself through the development and adoption of technical and social innovations. The generation of further added value and new innovative products and business models will move forestry and forest-based industry towards higher profitability with less emphasis on volume growth. Also, some objectives such as producing bioenergy or biofuel, and the use of non-timber products will open up new possibilities for tree breeding.

Up to date, there is a complete set of traits of interest in breeding programs (*e.g.* www.noveltree.eu), which can be classified in different categories, and they have been identified of importance in the following (Box 4.1):

- cold tolerance: bud phenology, bud flush, bud set, growth onset, growth cessation;
- disease resistances: avoidance (related to phenology), resistance (amount of tissue affected, laboratory resistance bioassays);
- drought tolerance: water-use-efficiency, or avoidance: cavitation;
- volume production: total height, diameter or circumference, fresh biomass;
- tree architecture: branchiness (number, thickness, angle), stem straightness, crown architecture; and
- wood characteristics: wood density, mechanical properties, lignin content, spiral grain

However, operationally, most of those traits are not being used in breeding programs, or they have only recently been introduced. In many cases, the success of the other "new" traits has been limited. As an example, the lack of substantial progress on breeding for wood quality could be interpreted as being due to a range of issues: the autoregressive nature of selection criteria, where the breeding values of new measurements depend on previous assessments plus an innovation term; ignoring age-related trends; using rotation age rather than technical thresholds to define objective traits; and ignoring the observation that not all grades of traits (*e.g.* density) have identical marginal economic value (Apiolaza, 2009).

During the last few years, there has been a significant advance in the biological understanding of complex traits such as growth, biomass production or adaptation to local conditions, due to the joint efforts of different disciplines (*e.g.* physiology, genomics, etc.). However, as previously described for density, we still need to have a better understanding of the trait architecture for it to be used more efficiently in breeding programs (see Chapter 5 and Box 4.2).

To date, phenotypic plasticity has not been considered as an independent trait in breeding programs. Commonly the main objective of the breeders has been to reduce the GE interaction by selecting against the most unstable (sensitive) genotypes or by identifying the environments that result in more extreme re-ranking. However, phenotypic plasticity has recently become a hot topic for discussion in evaluations, especially with the need for adaptation to different biotic and abiotic factors, and as new theoretical models propose plasticity as playing a central role in the evolution of a species. Also, by studying reaction norms it is possible to describe variation in genotype response under a range of environments described by one or more continuous variables. However, to determine these reaction norms more precisely, it is necessary to independently estimate different biotic or abiotic factors characterising the experimental sites. Technically, this is not a problem since there are technologies available for estimating many different

Box 4.1. Present and future breeding objectives and evaluation traits in some European forest tree species. (Source: D1.2-Evoltree. Author: Skogforsk)

The European project "Noveltreebreeding" has compiled information on the breeding programs in Europe for the most important species of forest tree. Over all the species many traits have been measured. From among these traits, we can distinguish selection traits (S), or complementary (C). In face of the new challenges, some of these traits could be incorporated as selection traits, but if so then more efficient evaluation methods will be required. The table below reports narrow heritabilities (h^2) and broad heritabilities (H^2) for four species.

Trait	Type	Pinus pinaster	Pinus sylvestris	Picea abies	*Populus* spp
		h^2	h^2	h^2	H^2
Height	S	0.13 - 0.65	0.21 - 0.70	0.17 - 0.31	0.09 - 0.76
Diameter	S	0.10 - 0.35	0.17 - 0.21	0.15 - 0.37	0.09 - 0.73
Survival	S		0.06	0.05 - 0.05	
Stem-straightness	S*	0.22 - 0.34	0.16 - 0.56	0.16 - 0.56	0.16 - 0.43
Wood density	C	0.30 - 0.61	0.40	0.39 - 0.50	0.47 - 0.75
Spiral grain	C	0.60 - 0.63	0.29		
Lignin content	C	0.25 - 1.00			
Slenderness	C		0.22 - 0.27		
Wood colour	C				0.13 - 0.22
Spike knots	C		0.09		
Branch angle	C		0.26		0.45 - 0.51
Branch diameter	C		0.13 - 0.21		
Crown dimensions	C		0.19		
Nb branches	C				0.21 - 0.45
Delta 13C	C	0.26 - 0.30			0.16 - 0.49
Onset of shoot elongation	C		0.15 - 0.46		0.26 - 0.76
Cessation of shoot elongation	C		0.00 - 0.64		0.18 - 0.52
Resistance to *Gremmeniella*	C		0.30		
Resistance to pine stem rust	C		0.40		
Field leaf rust resistance	S				0.12 - 0.77
Resistance to *Marssonina*	S				0.20 - 0.44
Bacterial canker resistance	C				0.47 - 0.75

* Stem straightness was regarded as C for *Pinus sylvestris* and *Populus* spp.

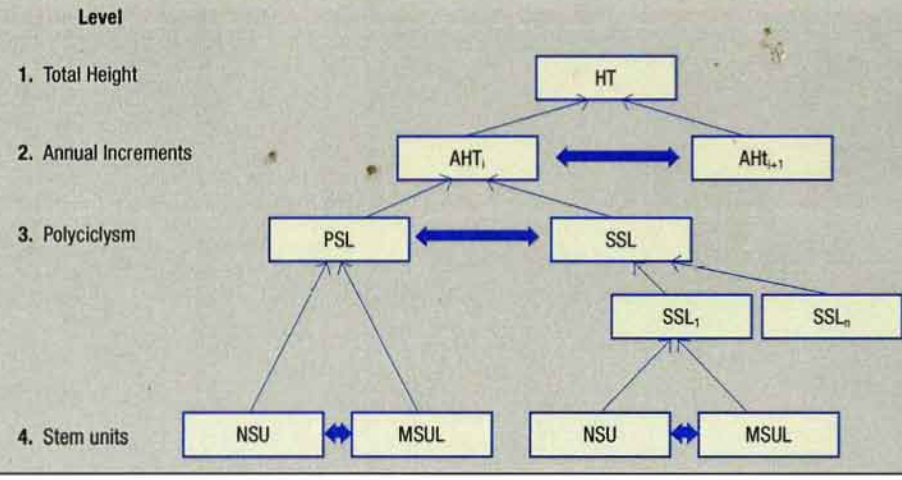
environmental variables at the required precision. However, estimating norms for multiple environmental variables simultaneously for all candidate trees is challenging and economic limitations reduce the possibility of achieving this during the evaluation period.

4.5. Future of phenotyping

Phenomics is an emerging trans-discipline dedicated to the systematic study of phenotypes on a genome-wide scale, which is still not fully developed in forest trees. The basic requirements of an ideal phenomics effort are easy to state but difficult to achieve: genomic information on many loci for a large sample of individuals, which are each exposed to a range of environments; extensive and intensive phenotyping across the full range of spatial and temporal scales; and low cost (Houle *et al.*, 2010). Therefore,

Box 4.2. An example of decomposing complex traits: height growth in maritime pine

Traditionally, growth of tree height can be divided into different components that have different genetic control, and this observation could be used more effectively in evaluation programs. As an example, we can distinguish four levels to decompose the growth (modified from Kremer & Lascoux, 1988). Total height (HT) could be divided into the annual growth increments (AHT_i). Each of these increments, can be divided into Primary or secondary shoot elongation (PSL and SSL), corresponding to the different components in a polycyclic species (SSL_i). And we can consider a fourth level of division, into the number of stem units (NSU) and the mean elongation of the units (LSUL).



there are initiatives related to developing new infrastructures for phenotyping traits (*e.g.* automatic watering systems for drought tolerance at INRA-Nancy), imaging technologies, metabolomics, or special techniques for particular traits (Table 4.1 and Box 4.3), sensors (to characterise genotypes and environments), and also to develop pipelines for data analysis and new methods for the analysis of multi-site and multi-dimensional phenotypes.

One of the ways to open up this new approach is to use as many traits as possible each associated with the different processes that determine the trait of interest. For example, drought resistance can be determined by different avoidance or tolerance mechanisms, with differing genetic controls and extent of genetic variation, and correlations with other factors. We can explore this multi-dimensional space of traits as WUE, osmotic potential, root/shoot allocation, growth phenology etc. Each of these factors could be approached by different medium- or high-throughput phenotyping technologies (Table 4.1).

Growth, one of the most important traits, could be analysed in a more efficient way by including temporal variation in relation to the environment. Until now, LIDAR (for Light Detection and Ranging) or T-LIDAR (Terrestrial LIDAR) has been used in forest inventories, but it would improve the estimation both of height/diameter growth and tree architecture (branching, stem form).

It would be also possible to develop more efficient bioassays for pest resistance, by combining existing methods of evaluation and novel technologies of detection (sensors, 'omics disciplines). In all the cases, geneticists and breeders will need to work more closely with physiologists, ecologists and engineers to develop informative, precise and standardized high-throughput phenotyping technologies (Neale & Kremer, 2011). We also need to revisit the concept of the environment of evaluation. It is possible

Table 4.1
A summary of some non-destructive high-throughput phenotyping technologies

Trait & characteristics	Non-destructive method	T'put	Technical challenges	References
Drought stress: Decreased stomatal conductance	Thermal infrared	H	Influence of environmental conditions (air temperature, wind speed etc.) Orientation of leaves to light source and camera	Jones *et al.* 2009, Chaerle *et al.*, 2009
Drought stress: Decreasing leaf water content	Near infrared (NIR)	H	Low sensitivity Absolute measurements not possible at this stage, only monitoring of changes over time	Seelig *et al.* 2008, 2009
Drought stress: Decrease in photosynthetic activity (only under severe stress conditions)	Fluorescence	H	Pulsed imaging setups currently limited to about 100 cm^2 and only optimized for planophyll plants	Baker, 2008, Woo *et al.* 2008, Jansen *et al.* 2009
Drought stress: Water use efficiency(WUE)	Carbon isotope discrimination	M	Can be used in leafs or wood. In leaves, dependent of the time of collection. High price per unit.	Correia *et al.* 2008
Frost hardiness	VIS plus NIR spectroscopy	H	In combination with multivariate modeling. Fast and nondestructive method for measuring frost hardiness (tested in Scots pine and Norway spruce seedlings)	Sundblad *et al.* 2001
Efficiency in use of input resources, such as nitrogen, phosphorus, water and solar radiation	Active canopy spectral reflectance sensors	H	less influenced by environmental conditions (varying solar radiation or soil color). Measure few wavelengths (might limit the prediction of complex traits, such as nitrogen-use efficiency)	Teal *et al.* 2006
Physical Wood Characteristics: Basic density	Gravimetric evaluation Pilodyn Resistograph	H	Species and age dependent. Use for early testing and low accuracy (mainly for Resistograph)	Raymond, 2002
Physical Wood Characteristics: Density variation, Density gradient, growth	X-ray diffraction Confocal microscopy X-ray micro-densitometry NIR	M	Dissection of annual growth	Saren *et al.* 2006, Isik and Li, 2003
Physical Wood Characteristics: Fibre length	Optical measurement of separated fibres	M		
Air-dry density, microfibril angle, stiffness and tracheid coarseness, specific surface and wallthickness	Silviscan	M	Costly preparation of samples. Not possible at very early ages	Beaulieu *et al.* 2011
Chemical wood characteristics: Cellulose content Lignin content Extractives Wood sugars	Chemical analysis of ground wood Near infrared reflectance analysis Raman spectroscopy	H	Species and age dependent	Raymond. 2002
Mechanical wood characteristics: Strength and stiffness of structural lumber	Acoustic methods	H	Rapid and cost-effective assessment methods	Wang *et al.* 2007
Tree Architecture: Height,	T-LIDAR	H	Affected by object occlusion and wind. Not applied to genetic trials.	Côté *et al.* 2011
Branching, Form, Knot size	T-LIDAR	H		

[1] H-High throughput phenotyping method, M: Medium

Box 4.3. Use of NIR in determining wood properties

Tree breeders have been reluctant to include traits associated with wood quality in tree improvement programs owing to logistic difficulties and the costs associated with assessing them. Now, several non-destructive techniques are available, including acoustics (velocity, stiffness). However near infrared (NIR) spectra have become a standard technique in recording for genetic evaluation, particularly in Eucalyptus (Raymond, 2002). The techniques has been used extensively to calibrate prediction models for (air-dry density, microfibril angle, modulus of elasticity, coarseness and wall thickness). NIR can also be used to estimate a more general chemical composition. NIR analyses are conducted at the increment core level, but they can also be split to get heartwood and sapwood estimates. However the most important fact is the high heritabilities suggesting that acoustic and NIR-based methods can efficiently be used for screening progeny tests for surrogate wood traits (*e.g.* Isik *et al.*, 2011).

to characterise the environment very precisely, but it is probably necessary to make a better selection of the set of environments (laboratory or outdoors) in which the traits need to be evaluated. For instance, to evaluate frost tolerance it is necessary to use different and standardised conditions. It would be possible to select more precisely different experimental sites for mimicking broad environmental conditions in short-term experiments.

The necessity of developing more efficient phenotyping methods (and more technological dependent), and also to reduce costs of calibrating the methods for different species, make necessary to have some large-scale phenotyping infrastructures that can be used for a large group of scientist and breeders.

4.6. Conclusions

Genomics-based breeding will change our approach to phenotyping, prompting an interest in new traits for selection and a need for multi-disciplinary approaches (genetics, physiology, etc.). This will lead to a better understanding of the architecture of key traits and make phenotyping more efficient.

The costs associated with these new developments, and the sourcing of the wide ranging technical expertise, will require co-operation among different organisations. The goal will be to develop new phenomics platforms leading to a more integrative approach to the study of phenotypes.

References

Apiolaza L.A., 2009. Very early selection for solid wood quality: screening for early winners. Ann Forest Sci 66: 601.

Baker NR., 2008. Chlorophyll fluorescence: a probe of photosynthesis in vivo. Annu Rev Plant Biol 59, 89-113.

Beaulieu J., Doerksen T., Boyle B., Clement S., Delasciers M., Beauseigle S., Blais S., Poulin P.L., Caron S., Rigault P., Bicho P., Bousquet J., Mackay J., 2011. Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. Genetics 188: 197-214.

Blonk R.J.W., Komen H., Kamstra A., van Arendonk J.A.M., 2010. Estimating breeding values with molecular relatedness and reconstructed pedigrees in natural mating populations of common sole, *Solea solea*. Genetics 184: 213-219.

Chaerle L, Lenk S, Leinonen I, Jones HG, Van Der Straeten D, Buschmann C., 2009. Multi-sensor plant imaging: towards the development of a stress-catalogue. Biotechnol J 4: 1152-1167.

Chevin L.M., Lande R.. Mace G.M., 2010. Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. PLoS Biology 8: e1000357.

Correia I., Almeida M.H., Aguiar A., Alia R., David T.S., Pereira J.S., 2008. Variations in growth, survival and carbon isotope composition (delta(13)C) among *Pinus pinaster* populations of different geographic origins. Tree Physiol 28: 1545-1552.

Côté J.F., Fournier R.A., Egli R., 2011. An architectural model of trees to estimate forest structural attributes using terrestrial LiDAR. Environ Model Softw 26: 761-777.

Dutkowski G.W., Costa J., Gilmour A.R., Wellendorf H., Aguiar A., 2006. Spatial analysis enhances modelling of a wide variety of traits in forest genetic trials. Can J Forest Res 36: 1851-1870.

El-kassaby Y.A., Lstibůrek M., 2009. Breeding without Breeding. Genet Res (Camb) 91: 111-120.

Fernández J., Villanueva B., Pong-Wong R., Toro M.A., 2005. Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics 170: 1313-1321.

Houle D., Govindaraju D.R., Omholt S., 2010. Phenomics: the next challenge. Nat Rev Genet 11: 855-866.

Isik F., Li B., 2003. Rapid assessment of wood density of live trees using the Resistograph for selection in tree improvement programs. North 2435: 2426-2435.

Ivkovic M., Wu H.X., McRae T.A., Powell M.B., 2006. Developing breeding objectives for radiata pine structural wood production. I. Bioeconomic model and economic weights. Can J Forest Res 36: 2932-2942.

Jansen M, Gilmer F, Biskup B, Nagel KA, *et al.*, 2009. Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via GROWSCREEN FLUORO allows detection of stress tolerance in *Arabidopsis thaliana* and other rosette plants. Funct Plant Biol 36: 902–914.

Jansson G., Jonsson A., Eriksson G., 2005. Use of trait combinations for evaluating juvenile-mature relationships in *Picea abies* (L.). Tree Genet Genome 1: 21-29.

Jones HG, Serraj R, Loveys BR, Xiong LZ, Wheaton A, Price AH, 2009. Thermal infrared imaging of crop canopies for the remote diagnosis and quantification of plant responses to water stress in the field. Funct Plant Biol 36: 978–989.

Kolström M., Vile T., Lindner M., 2011. Climate Change Impacts and Adaptation in European Forests: Management. EFI, Joensuu, Finland.

Kremer A., Lascoux M. 1988. Genetic architecture of heigh growth in Maritime pine (*Pinus pinaster* Ait.). Silvae Genet 37: 1-8.

McRae T.A., Apiolaza L.A., Dutkowski G.W., Kerr R.J., Pilbeam D.J., Powell M.B., Tier B., 2003. TREEPLAN - A genetic evaluation system for forest trees. 27th Southern Forest Tree Improvement Conf, Stillwater, OK, USA, 25-27 Jun. [CD ROM].

Neale D.B., Kremer A., 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet 12: 111-122.

Raymond C.A., 2002. Genetics of Eucalyptus wood properties. Ann Forest Sci 59: 525-531.

Rehfeldt G.E., Tchebakova N.M., Parfenova Y.I., 2002. Intraspecific responses to climate in *Pinus sylvestris*. Global Change Biol 8: 912-929.

Saren M.P., Serimaa R., Tolonen Y., 2006. Determination of fiber orientation in Norway spruce using X-ray diffraction and laser scattering. Holz als Roh- und Werkstoff 64: 183–188.

Seelig HD, Hoehn A, Stodieck LS, Klaus DM, Adams WW, Emery WJ, 2008. The assessment of leaf water content using leaf reflectance ratios in the visible, near-, and short-wave-infrared. Int J Remote Sens 29: 3701-3713.

Seelig HD, Hoehn A, Stodieck LS, Klaus DM, Adams WW, Emery WJ, 2009. Plant water parameters and the remote sensing R (1300)/R (1450) leaf water index: controlled condition dynamics during the development of water deficit stress. Irrig Sci 27: 357–365.

Sundblad L, Andersson M., Geladi P., Sjöström M., 2001. Fast, nondestructive measurement of frost hardiness in conifer seedlings by VIS + NIR spectroscopy. Tree Physiol 21: 751-757.

Sölkner J., Grausgruber H., Mwai A., Peter O, 2008. Breeding objectives and the relative importance of traits in plant and animal breeding : a comparative review. Euphytica 161: 273-282.

Teal R.K., Tubana B., Girma K., Freeman K.W., Arnall D.B.,Walsh O., Raun W.R., 2006. In-Season Prediction of Corn Grain Yield Potential Using Normalized Difference Vegetation Index. Agron J 98(6): 1488-1494.

Vanhanen H., Rometsteiner E., Mery G., Lorenz M., Tikkanen I., Toppinen A., Paavilainen L., Flies R, 2007. Making European Forests Work for People and Nature: Policy. EFI, Joensuu, Finland.

White T.L., Adams W.T., Neale D.B., 2007. Forest Genetics. CAB Int, Wallingford.

Wang X., Carter P., Ross R., Brashaw B.K., 2007. Acoustic assessment of wood quality of raw forest materials: a path to increased profitability. Forest Prod J 57(5): 6-14.

Woo NS., Badger M.R., Pogson B.J., 2008. A rapid, non-invasive procedure for quantitative assessment of drought survival using chlorophyll fluorescence. Plant Methods 4: 14.

Chapter 5
Shifting from growth to adaptive traits and competition: the prospect of improving tree responses to environmental stresses

Leopoldo Sanchez, Philippe Rozenberg and Catherine Bastien
INRA, UR0588, Amélioration, Génétique et Physiologie Forestières, 2163 Avenue de la Pomme de Pin, CS 40001 ARDON, 45075 ORLEANS Cedex 2, France

Questions that will be answered in this chapter:

- What is adaptation? Is response to competition an adaptive feature?
- Why are adaptive traits becoming relevant now to scientists?
- Do we need to look at and breed for new traits?
- Trees show plastic responses; what is plasticity and how is it revealed?
- What do wood records tell us on reactions to changing environments? A case study
- How should we analyse genetic variation for adaptive traits?
- Do we need to re-think our field test evaluations?

5.1. What is adaptation? Is response to competition an adaptive feature?

As any other living organism, trees grow, compete with each other and reproduce to form populations in a changing and often challenging environment. Some individuals are able to face these changes and challenges with a greater success than their counterparts. A greater success directly means or ultimately leads to larger numbers of descendants for the next generation. Descendants of successful individuals are better *equipped* to face the changes and challenges in which their parents have excelled. **Adaptations** are in this manner spread out over time through the population, which changes its genetic composition accordingly and correspondingly to the environmental pressures. But adaptation not only refers to a biological functionality or phenotypic variant that improves the survival and reproductive success of its carrier, as described above, it also refers to the dynamic process by which populations evolve phenotypically. This phenotypic evolution results ultimately from **natural selection**. Often, authors qualify a trait evolving under natural selection as being **adaptive**. This means that certain phenotypic variations in that trait confer advantages upon carriers, favouring them as genetic contributors for the next generation over those lacking the variations.

Adaptations and adaptive traits can be defined in a multitude of ways. Classically, traits that show conspicuous clines over climatic gradients are considered of adaptive value, like those related to phenology, bud set and growth cessation (see Savolainen *et al* 2007 for a clear exposition on adaptive clines). Also resistances and tolerances to pests are assumed to be adaptive. Growing fast at the first stages of a tree life can be advantageous in a place where competition for resources with other counterparts is hard. Often, trees seek actively for light, leading in dense forest to arms races for height. Conversely, tolerating shade from taller competitors can be found to be advantageous for those that *seek* the eventuality of free resources left behind by a decaying neighbour. In that sense, the ability to compete can be seen as a component of adaptation. Although competition is often meant to happen between conspecifics, and

this will be our focus here, we should be aware of other components arising from interacting species or, at lower levels, from antagonistic biological functions within individuals.

We have just introduced, somewhat simplistically, the concept of adaptation in the context of biological evolution. Biological evolution is universal, in the sense that it affects all living organisms through the same mechanisms. Trees present, however, two features that jointly characterise the distinctiveness of these life forms among other taxa. The first of them is the fact that trees are most often among the **longest-living organisms**. A lifespan of a century is common for most forest tree species and several centuries and even millennia are not rare among oaks, pines or Douglas-firs to cite a few examples. The second feature, although not specific to trees, is the fact of being **sessile**, thus, immobile and permanently attached to their soil subtract. Any given successful tree grows, stands and dies at the *very same spot*, and this can cover a considerably long lifespan. This strongly suggests that trees have distinctively built-in adaptive strategies in order to cope *in-situ* with the large succession of environmental eventualities of their lives *i.e.* daily changes, seasonal changes and inter-annual trends. These strategies fall within the general phenomenon of phenotypic plasticity, which will be described in the following sections. Eventually, trees also succeed in modifying at least partially their immediate environment for their own benefit *e.g.* by retaining soil, nutrients and humidity.

These two features of trees are still of relevance when considering reforestation or afforestation for commercial purposes. **Rotation age**, the period spanning between planting and harvesting, although optimized for maximum return of investment, is usually long for trees: twenty to hundred years depending on species, latitudes and production scenarios. This is considerably larger than in any other crop or livestock species. Also, unlike many other domestic species where there is a considerable degree of control by the grower or farmer of the crop environment, think of fertilizers for crops and health-care in livestock, with planted trees the degree of control remains comparatively limited, often circumscribed to the choice of the plantation site or, at most, partly handled by silvicultural activities.

As a palliative to challenging global changes, experts have recently suggested the concept of **assisted migration**. By this concept of environmental engineering, populations are shifted geographically from their current positions, whenever these are expected to become unfavourable, to alternative locations that are foreseen to become suitable according to prospective trends. This shift takes the form of siblings or seeds collected from the original population and used for afforestation in the new suitable environment. Although undeniably attractive, assisted migration is at the centre of experts' turmoil (see for instance Hewitt *et al* 2011). Whatever the measures, understanding adaptation and its expression through phenotypes remain central.

5.2. Why are adaptive traits becoming relevant now to scientists?

To what extent do these distinctive built-in adaptive strategies of trees make them well prepared to the challenge of global and climatic changes? Indeed, this is a major challenge for whole ecosystems, including terrestrial ecosystems where trees play often the role of key species. We have to remind that trees represent a large percentage of the continental biomass and harbour substantial parts of the terrestrial biodiversity. There is growing evidence about the unfavourable effects of global change on biodiversity and distributional areas of many species. Discussing the nature of this **global change** is, however, beyond the scope of this chapter but is a hot topic of the moment among specialists. What is an undeniable fact is the change itself, with fast rates of changes in climatic parameters like temperature and precipitation and larger amplitudes in their variation over seasons than those seen in past climate records. Authors agree on the fact that long-lived species, with late reproductive maturation and sessile in nature are at higher risk of maladaptation and ultimately *condemned* to extinction under the conditions of global change (Thomas *et al.*, 2004).

This idea of "*environmental changes going too fast for trees to catch up*" has motivated recently a growing interest among scientists in tree adaptation studies. We have said that trees appear to be well *equipped* to cope *in-situ* with changes in their immediate environment. Evolution seems to have modelled trees to be *plastic* in their long lifespans, and we will in subsequent sections give more details on this **concept of plasticity**. Experience from tree breeding in the last decades tells us that many trees show indeed this plastic capacity in some of their key phenotypic features when reacting to a changing environment. Field testing networks have been set over wide geographical ranges for many valuable forest trees, comprising in some cases contrasting test sites that share genotypes or clonal copies of the same individuals. From them we learn that some genotypes, families or populations grow well over a wide range of evaluation conditions (climate, photoperiod, soil). The question that remains is whether this capacity will show enough genetic variability within and across tree species to withstand the pressure of global change. Seeking answers to that question motivates the rising interest in adaptation.

5.3. Do we need to look at and breed for new traits?

Therefore, there appears to be great potential in the adaptability of trees, as shown by performances in field trials. Does this help us to answer the question of what the fate of many tree species will be in the next century? Not really. Although impressive in numbers, often with tens of thousands of records, many observations coming from field tests are based exclusively on **growth traits**, typically height and diameter of trees at half their rotation age. Variation in these growth traits reflects well the general breeding objective of many forest tree selection programs, basically to increase wood volume production. Additionally, with a little intuition, it is easy to see that growth reflects the well-being of a tree, at least in the particular conditions that have affected the development of that tree. Other traits are indeed studied in field trials, like architectural traits, phenology or survival, but none of these traits reach the amount of data across sites of heights and diameters.

There is one basic objection to the consideration of "*classical*" growth traits in the study of breeding for tree adaptation. Height and diameter are intrinsically complex and extremely integrative traits. They represent the *visible* result of the contribution of multiple underlying functional processes: *e.g.* the kinetics of cell division in meristematic tissues and its phenology, transpiration and photosynthesis, and the hydraulic dynamics of sap conducting vessels in the wood. These processes react directly to environmental cues, like temperature or water availability, and can interact to each other in synergistic or antagonistic ways. Some of these functional traits present desirable features for breeding, like being variable and highly heritable. For instance, the isotopic ratio $\delta^{13}C$, a proxy of transpiration efficiency, has already shown large levels of variation in maritime pine and poplars. Others, like **wood density** have relatively high heritabilities, compared to that for height. The density of wood is known to be related to hydraulic dynamics, as variations in number and size of conducting vessels in the wood can be detected by variations in wood density. Understanding the way trees grow is based ultimately in the comprehension of how environmental cues interact with these functional underlying processes. Of course, tree height often responds to variation in temperature and water availability, but the resulting relationship might be loose, with little scope for further interpretation if we do not know the elementary processes underlying growth.

The assessment of traits is not limited to field trials. These allow for limited control of growth conditions, other than those expected from local climate and soil quality, and preclude regular monitoring. Many functional traits have been studied in climatic chambers, where fine control of growing conditions and intense monitoring is at hand. Sánchez-Gómez *et al.* (2011) carried one of these intensive studies on the response of a Mediterranean pine species to drought stress for several functional traits, like transpiration efficiency and photosynthetic rates. Studies like this are of great value for the comprehension of the

functional components of growth as well as for revealing the differential expression of the genes that underlie these physiological responses. However, they are limited in the variation that can be screened.

Unlike classical breeding, where selection emphasis is usually placed on a few integrative traits, breeding for a suite of functional traits brings the difficulty of multiplying costly measurements and managing multiple selection thresholds. The advent of new evaluation methods that rely entirely on genome-wide screenings (see Chapter 3) can circumvent previous difficulty, as they can substantially reduce the reliance on phenotyping.

Therefore, growth traits, as seen classically, are good global indicators of tree well-being and productivity, but are limiting for us to understand how trees react to environmental changes. We will see with next question that some of these functional processes can be interpreted as plastic responses to environmental cues, in what has been termed as plasticity. Only this knowledge will give us the tools to breed for better adapted trees.

5.4. Trees show plastic responses; what is plasticity and how is it revealed?

The last decade of scientific literature has been very prolific in defining plasticity (see two general and comprehensive books on plasticity by DeWitt & Scheiner (2004) and Schlichting & Pigliucci (1998)). There is, however, an inconsistency in the different meanings of the phenomenon as it is perceived from the literature, due certainly to the fact that plasticity in living organisms has been looked at from many different disciplines: ecology, physiology or genetics. Plasticity or, more strictly, **phenotypic plasticity** is a *phenotypic change that results from the exposition of a given genotype to a variation in the environment.* It is always associated to a trait and built up over an environmental cline, for instance, the variation in phenological traits in replicates over latitudinal and climatic gradients (Kramer, 1995). Unlike other *broad-sense* definitions in which the genotype is *loosely* interpreted as a family or even a population, our definition here is a *narrow-sense* one *i.e.* the plasticity phenomenon is to be observed for a given tree or for copies of a given genotype.

Phenotypic plasticity implies that there are multiple phenotypic states per genotype, either corresponding to different measurements taken over time or over different test sites at a given maturation stage. For the latter case, genotypes can still be the same through clonal copies, a *natural* feature of many plants, and it results in what is known as **spatial** or **site-related plasticity**. This is the most common situation in field tests where genotypes are often shared amongst plots and sites. Whenever measurements are repeated over time for the same individual, the resulting reaction is called **temporal** or **time-related plasticity**. Spatial plasticity presents the desirable statistical property of comprising independent measurements, albeit an arbitrary choice of environments might reduce the adaptive relevance of the spatial experiment. In temporal plasticity, the risk comes from the fact of having repeated measures linked by autocorrelation, *i.e.* independence is lost, or flawed by **ontogeny**, the underlying process of development driven by aging. However, it shows an interesting feature in the fact that the recorded individual reaction is closer to what might be *perceived* by natural selection at individual level. One example of this temporal plasticity will be given in the next section.

In both temporal and spatial views, phenotypic plasticity can be seen as a phenotypic trajectory over a heterogeneous environmental landscape. This trajectory can be modelled statistically, for instance through a regression equation, as a function of one or several explanatory environmental variables. The resulting function is then called **norm of reaction** (NoR). Several individuals of a given population might show different NoRs over the same environmental cline, resulting for instance in intercrossing curves. This is typically a **genotype by environment interaction**, a component of a factorial analysis of variance with a genotype versus environment layout. Genotype by environment interaction measures variation in NoR. The Fig. 5.1 explains some of these latter concepts with the help of four simplified scenarios.

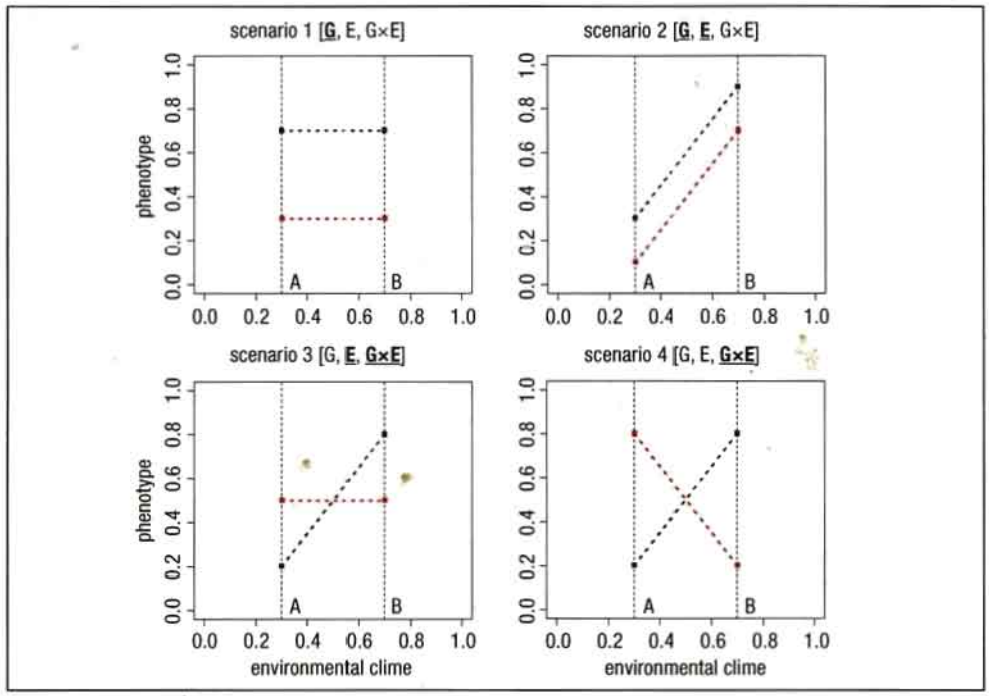


Figure 5.1. Let be two distinct environments A and B, for which a climatic parameter differs to a certain quantifiable amount, and represented here arbitrarily by the two vertical lines in the horizontal axis (environmental cline). Four scenarios represent NoRs for two genotypes (red and black), whose phenotypic differences are shown at the vertical axis (phenotype). "G", "E" and "G×E" stand for different components of variation in phenotypic units (in **bold** and underlined when relevant): between genotypes (G), between environments (E) and between NoRs (G×E) i.e. interaction between genotypes and environments. **Scenario 1** shows no plasticity in both genotypes, with one outperforming the other by the same amount in both environments, which results in G being the only relevant component of variation. **Scenario 2** shows the same plastic response for both genotypes, with phenotypic differences (G) and B environment showing higher phenotypic levels than A (E). **Scenario 3** shows interaction between NoRs (G×E), A environment results on average in lower performances than in B (E), and there is not net G variation as genotypic effects level out between environments and genotypes. Finally, **scenario 4** shows again interaction between NoRs (G×E), but genotypic and environmental effects both level out resulting in no net variation at population level in G and E. This is not an exhaustive enumeration of all possible scenarios.
[Adapted from DeWitt and Scheiner 2004]

5.5. What do wood records tell us on tree reactions to changing environments? A case study

Wood can be seen as a record of tree anatomical reaction to environmental changes. Here, we will present how this record is formed and we will introduce some of the hypotheses that link this process to adaptive functionality.

During wood formation, the specialized tissue called **vascular cambium** reacts to environmental cues by adjusting *more or less* promptly the anatomy of the newly formed xylem cells to the *perceived* external changes. This continuous modification leaves a permanent anatomical trace in the wood that follows seasonal changes and weather events (see Fig. 5.2). In a temperate climate, where seasonal changes are well marked, the most conspicuous anatomical traces in the wood, by their intensity and reiteration, are **tree rings**. They are a type of *landmark* for seasonal growth. If there is a key adaptive role in the wood, this is mostly related to the fact that it mediates the transport of sap among the different functional parts

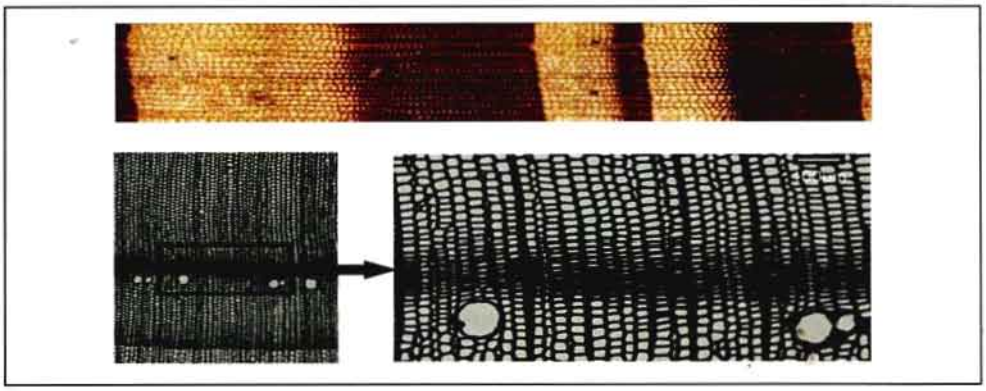


Figure 5.2. Top picture illustrates a wood anatomy radial transect showing the succession of tree rings in a coniferous species (Douglas-fir), where darker bands correspond to late wood and lighter bands to early wood. In the **bottom pictures,** a cross-section in a *Pinus* tree-ring that shows a false ring, darkened band, with a close-up to the right. [Adapted from Wimmer 2002]

of the tree. Often, this transport functionality is reserved to the most recently formed rings, while older ones progressively become non-conductive.

The anatomy of newly formed xylem ducts must withstand the pressures that result from the balance between water demand up in the leaves and water availability down at the roots. Large diameter xylem ducts are better suited for the large flows of water that fast growth needs. This configuration occurs typically in spring when growth conditions are the most favorable over the year. Such diameters, however, come with the drawback in some temperate climate species of being particularly susceptible to **wood embolism**. Embolism happens when the water column in the ducts has been stretched beyond its physical limits and breaks, often during a severe water deficit event. Other anatomical features like inter-duct pits might also be of relevance for embolism initiation. The resulting consequences of embolisms are hydraulic dysfunction, which in turn can lead to loss of growth and, ultimately if dysfunction is generalized, to mortality (Breda *et al.*, 2006). Yet xylem is a living tissue, constantly renovated by vascular cambium. If it reacts promptly to stress cues, adequate cells are produced with higher resistance to hydraulic dysfunction. When this happens early in the growing season, **false rings** usually appear in some species (see Fig. 5.2). All this functioning recalls explicitly the phenotypic trajectory of a NoR as described previously. Some qualities of these NoRs would have an attached adaptive value, for instance when showing promptness under stressing conditions in producing cells with narrow-lumens and thicker walls.

Wood density across a radial section reflects well changes in wood anatomy: denser wood contains smaller ducts, while larger ducts result in lighter wood. Strong associations have been found between wood density and resistance to mechanical failure across species (Cochard *et al.*, 2009). Fewer are, however, the number of studies showing these associations at within-species level. The Fig. 5.3 illustrates one of these examples.

Wood records appear as a promising source of information on the ways trees react to stressing environmental events, like severe droughts. However, the adaptive role of the dynamic change in anatomy of sap conducting xylem has not been fully elucidated yet. Of particular importance is the identification of what a favourable wood formation NoR looks like when trees face stressing events. For this, studies comparing NoRs of trees surviving a severe weather event to those of non-surviving counterparts are a promising starting point. Of no less importance is the assessment of genetic variation for these plastic traits, both across and within species. This brings us to the next question.

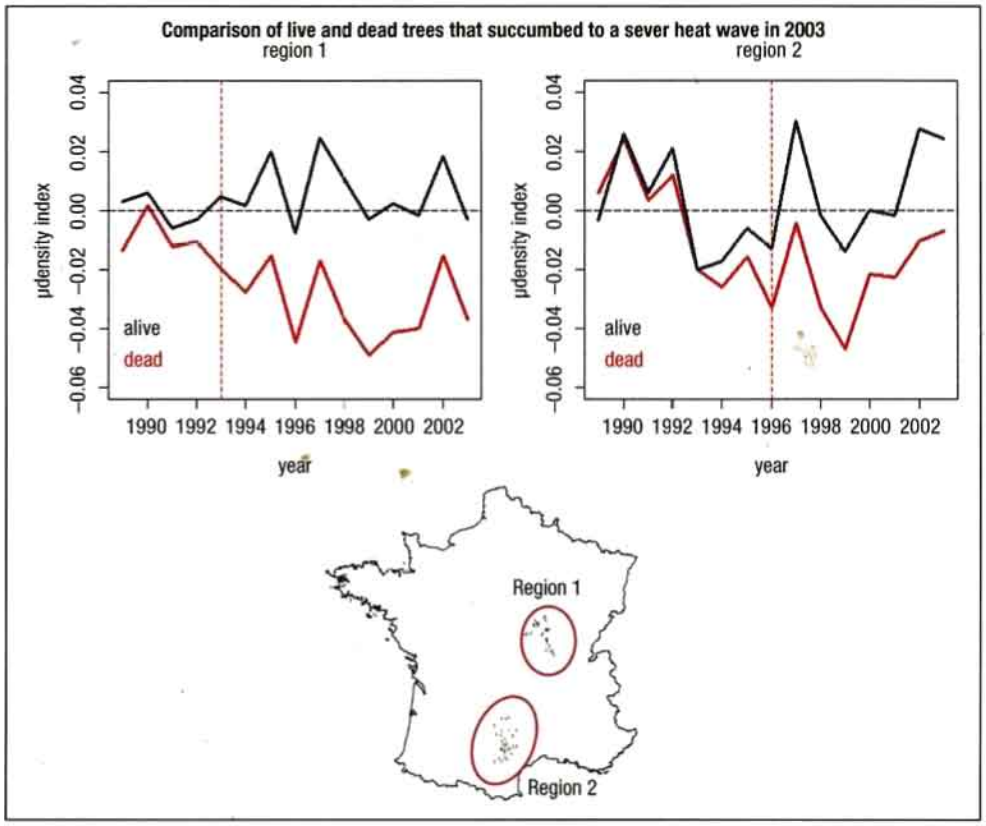


Figure 5.3. Top figures illustrate a comparison between dead and alive trees (Douglas-firs) for a µdensity index in two French regions (**map**). This **index** results from adjusting the within-ring µdensity average to ontogenic effects across years and site effects, resulting in a deviate from zero that is the population mean at region level (**horizontal dashed line**). Dead trees were those identified as such immediately after a severe heat wave that hit Europe in 2003, while alive trees were neighboring conspecifics in a healthy state after that date.
Vertical red dashed lines indicate the year for which the differences between dead and alive trees for the µdensity index were statistically significant (type I error = 0.05). These figures suggest that trees that succumbed to the weather event had a distinctive behavior in terms of wood characteristics over a long period preceding death. Although this appears as a good indicator of tree's adaptability to extreme events, it remains a statistical association with no direct causal implications. Further research is needed to identify causal components.
[Adapted from A. S. Sergent Thesis 2011]

5.6. How should we analyse genetic variation for adaptive traits?

The two main approaches to assessing and characterising genetic variation differ by whether the *starting point* for the analysis is the phenotype or, at the *opposite* extreme, the gene or a molecular marker linked to it. Assessments of underlying genetic components that are based on phenotypic variation in genetically structured populations belong to the realm of quantitative genetics. Gene-based assessments aimed at phenotype inferences have been the general objective of genomics. We will present here some of the classical approaches of quantitative genetics applied to adaptive plastic traits and to competition. To a lesser extent, we will illustrate genomic approaches to genetic variation inference by an example at the end, given that they are already treated with detail in other chapters.

Quantitative genetics of NoRs. Indeed, estimating NoRs and their variation in populations for relevant traits is only a first step before trying to answer the question whether there is a genetic component underlying the variation of phenotypic plasticity. That component would eventually make our plastic trait relevant to breeding and with adaptive potential in natural populations. Classically, quantitative genetics analyses are used to decompose phenotypic variation by experimental designs into genetic and residual components. One of the key outcomes of this kind of analyses is the **heritability**, the fraction of variation between individuals in the experiment that is due to their additive genetic values *i.e.* those that are transmissible from parent to offspring. For decomposing plasticity traits, there is no need of conceptual changes with respect to *point-based* traits, like growth at rotation age, and the same quantitative genetic framework can be readily used to obtain estimations of heritabilities.

The simplest case is for NoRs that correspond to linear regressions of phenotypes on environmental variables. Under this scenario, every measured genotype would correspond to a **straight regression line** linking phenotypes to their corresponding environmental records, and could be summarised by the resulting slope or regression coefficient. This is the case of the oversimplified example in Fig. 5.1. At this point, slopes can be treated as one of the attributes of genotypes in the quantitative analysis, as if they were the observations, and estimate the heritability of the slope and its components of variation.

Often, however, representations of biological phenomena do not follow a straight line. Many natural processes exhibit a progression that is well characterized by a **sigmoid function**. An example of this modelling is given in the Fig. 5.4 for a plasticity trait related to wood formation within a tree ring as a function of a **drought index**. As with the straight regression line example, once the sigmoid-shaped curve has been fitted at the phenotypic records, resulting parameters of the curve are analysed as if they were new traits in the *classical* quantitative genetics manner.

Although conceptually simple, both previous approaches are not exempt of drawbacks. One of them is the fact that the estimated functions (sigmoid and linear) might be affected by systematic environmental effects that were neglected and not included as explanatory variables in the fitting of the curve. Of lesser importance is the problem of being limited by the choice of curve families to fit our phenotypic trajectory.

For the analysis of any given phenotypic trajectory, several authors have proposed a general framework within the family of **mixed model analyses** (see a comprehensive technical review in Meyer & Kirkpatrick, 2005). Basically, mixed models are statistical linear models that represent the observations in terms of **fixed** and **random explanatory variables** simultaneously. By their great flexibility, mixed models have become the *workhorse* in the estimation of breeding values in selected populations, where there is a multiplicity of fixed and random explanatory variables like *sex*, *age*, *site*, *block* and *family*, to name a few of the most common. The generalization that is of concern here consists in the extension of mixed models to be able to represent phenotypic trajectories in terms of random explanatory variables that are in turn functions of time or even space coordinates. These enlarged models are known as the **random regression** (RR) approaches. A typical RR linear model describing a phenotypic trajectory (y) over time (t) would be in a simplified form as follows:

$$y(t) = M(t) + g(t) + p(t) + e(t)$$

where $M(t)$ represents an average trajectory over t, equivalent to the mean for a *point-based* trait. In this example, terms $g(t)$ and $p(t)$ correspond to genetic effects and to permanent environmental effects, respectively. The former is the heritable part of the phenotypic trajectory, while the latter corresponds to non-heritable factors that affect a given cohort, like those of nurture or weather conditions. These terms are random functions that are evaluated for each individual over time and that can be chosen among many

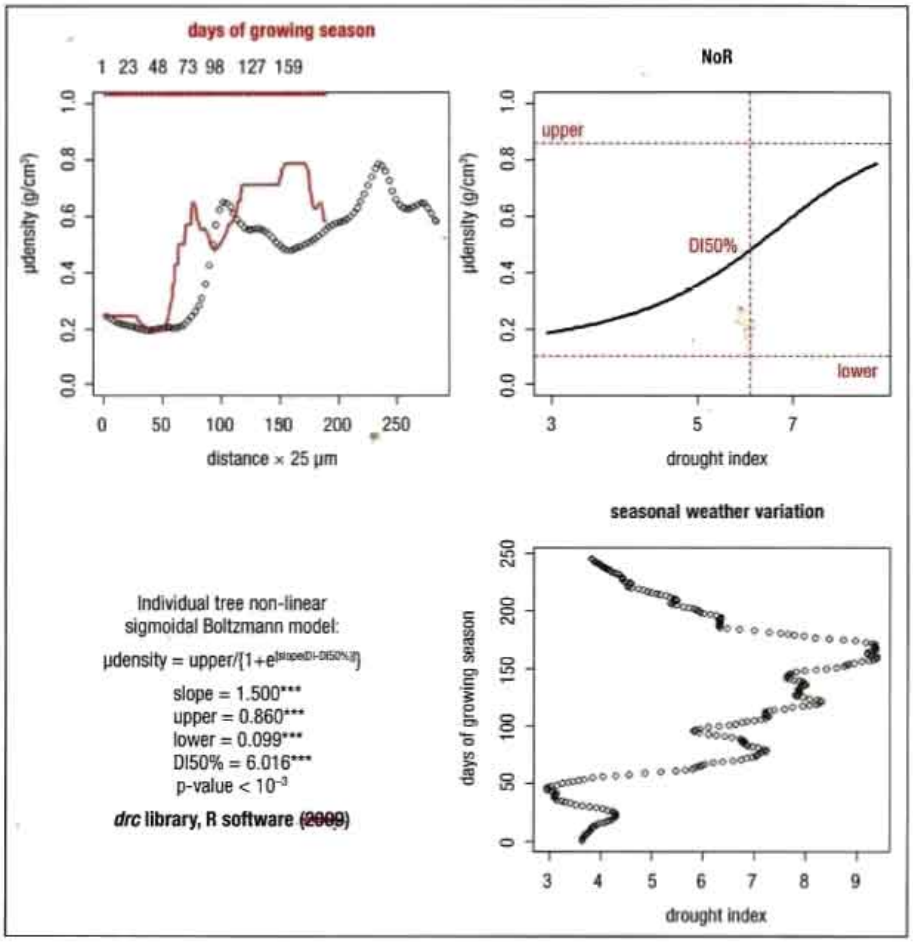


Figure 5.4. Top left figure illustrates an example of a tree ring µ-density profile (in black) in Douglas-fir. This profile comprises wood density measurements (g/cm³) every 25µm, from low density early wood at the origin to high density late wood at the end of the profile. The same profile (in red) is represented as function of days of growing season by a synchronization function based on weather variation. Weather variation for the year corresponding to depicted µ-density profile is illustrated in **bottom right figure,** in the form of a drought index (DI) with high values indicating unfavorable water availability. Although simplistic, this drought index have been found to explain large proportions of tree ring µ-density variation. The response in µ-density is modeled as a function of DI in **top right figure,** following a Boltzmann fitting. Boltzmann coefficients (**bottom left**) provide characteristic values with biologically interpretable meaning, where slope represents a linear response to DI increments, *upper* the tree response towards the end of the growing season, and DI50% an indicator of susceptibility to DI.

types of continuous functions, like for instance exponential curves. Finally, in previous equation, last term *e(t)* is the temporary environmental effect that is attached to each observation at a given *t* and for a given individual. Of importance here is the fact that any NoR can be readily modeled thanks to RR approaches, providing that suitable sets of relevant environmental variables are available as permanent environmental

effects. As with **estimated breeding values** in classical mixed models, the resulting outcome with RR approaches is a set of trajectories of estimated breeding values over a given environmental landscape.

Most of the published examples where NoRs are modeled within the RR framework come from livestock experiments, notably for inferring environmental sensibilities in production traits. De Jong & Bijma (2002) offered a comprehensive review on NoR modeling and on the strategy of selection gradients that can be used to select on NoRs. These approaches remain relatively novel in forest breeding, despite their potential in plasticity studies. One example is that of Wang *et al.* (2009), where tree height over years is modeled for a large progeny trial, resulting on average with larger heritability estimates than those obtained in single-trait analyses, which suggest that RR models are able to capture more genetic variance than that obtained from classical modeling.

Quantitative genetics of competition. So far we have seen the environment as an ensemble of physical factors, like weather, that contribute to the modeling of phenotypes over time. The Fig. 5.5 illustrates the effects of competing conspecifics on tree growth, in what constitutes a wider picture of the neighbouring environment that includes biological interactions.

Competition reflects the impairing interplay of closely growing trees, often when local resources are limiting. The incorporation of competition effects into quantitative genetic approaches needs a change of the analytical paradigm. Often, it is assumed that gene effects are to be circumscribed to the sole carrier. However, genes from one individual can impact the performances of other neighbouring individuals in such a way that variation among individuals may result substantially distorted. In a seminal paper, Griffing (1967) introduced a model that expands the concept of phenotype to include not only a term for direct effects, as classically from an individual's own genes, but also associative effects as those arising from neighbouring genotypes.

The Fig. 5.6 describes with an example this original approach. Competition can then be readily interpreted from this model as an antagonism between direct effects and associative effects. This antagonism results ultimately into impaired phenotypes. If selection proceeds by keeping only the best growers, it might also promote indirectly competitiveness and thus impair biomass in neighbouring conspecifics. This perspective was at the origin of the concept of *group selection*, as opposed to classical individual selection. By selecting on the group performance while accounting for competition, the net biomass would be improved. Group selection has, however, been a contentious subject in the field

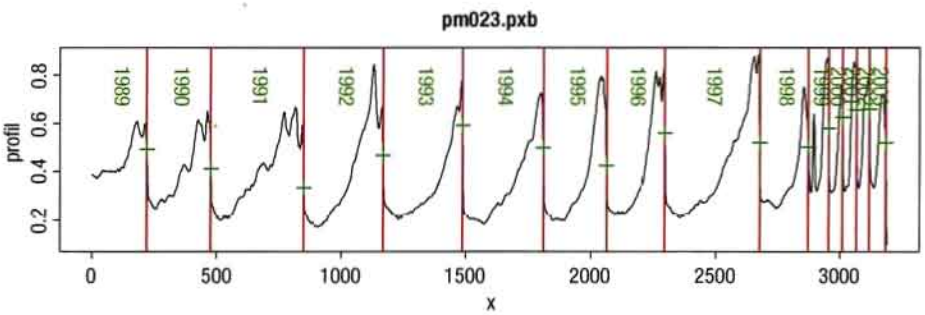
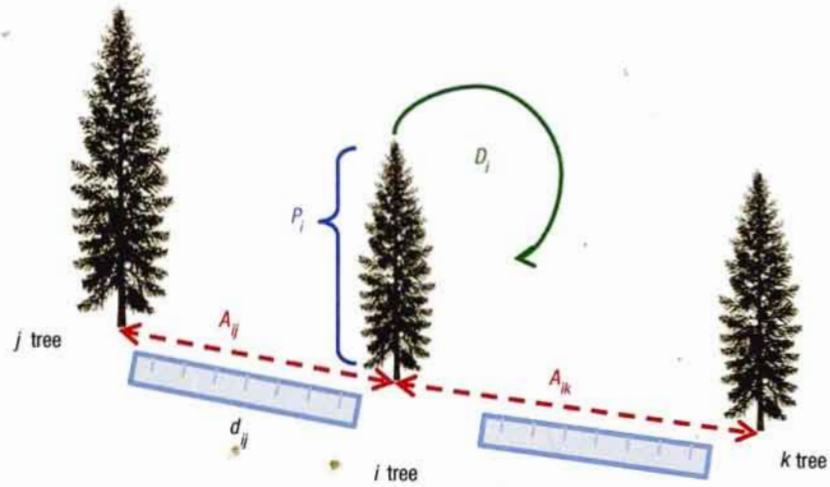


Figure 5.5. Example of **µdensity profile** (in gr/cm³) across several years (1989-2003), with vertical red lines indicating the limit between consecutive tree-rings and abscise **x** corresponding to distance from pith (left) to bark (right). This profile corresponds to a young Douglas-fir from a field trial and it shows a conspicuous narrowing of tree rings in its later years (from 1999 and onwards), when competition for light and other resources with neighboring trees started to impair growth.



$$P_i = D_i + (1/d_{ij}^2) \times A_{ij} + (1/d_{ik}^2) \times A_{ik} + e_i$$

Figure 5.6. Schematic representation of direct (***D***) and associative (***A***) effects that account for a tree's phenotype (***P***) in the field. Two trees (j and k) are in the vicinity of our focal tree (*i*), with distances between them being quantified by ***d***. Associative effects, for instance between *i* and *j*, are assumed to diminish geometrically as *d* increases (i.e. d^{-2}), and quantified by the amount A_{ij}. Competition results whenever associative effects and direct effects act antagonistically to produce a phenotype. The same reasoning is followed for any other pair involving our focal tree. The observed phenotype (P_i) is, as expressed in the equation, the net result of direct effects of the tree's own genotype and associated effects of other genotypes in the vicinity. An error term (**e**) is added to account for other phenotypic components that can not be associated experimentally to any causal factor.

of evolutionary theory, although for breeding it does not have the same relevance as grouping falls within breeders' decision.

In the framework of quantitative genetics, where the partition of observed variances into causal components relies on accounting for resemblances between relatives, unaccounted competition might result in distorted genetic estimates. For instance, resemblance amongst sibs raised in different plots could be deflated below expectations if their respective neighbours have distinct competing abilities. Resemblance and competition are both functions of a measure of distance. While in Griffing's (1967) model the distance is a physical measure between competitors (*d* in the Fig. 5.6), with resemblances the distance is the genetic closeness or the extent to which relatives are close genetic replicates. Muir (2005) proposed a model that incorporates both distances for the decomposition of pedigreed phenotypes in conditions of competition. Like previous RR approach, Muir's (2005) approach belongs to the family of flexible mixed models, where random explanatory variables include direct and associative genetic effects, both being conveniently weighted by physical and genetic distances. When this approach is used in an appropriate experimental setup, where every tree is conveniently mapped and where pedigree relationships have been assessed, such as most of the progeny field tests, estimated breeding values for direct and associative effects can be readily obtained.

Trees in field trials are excellent candidates for competition assessments, as these rely basically on readily available distance grids to quantify the sharing of resources. Unlike trees, animals present behavioural components that modulate and complexify competition relationships. Despite of that, many

examples of competition assessments exist for livestock, basically penned animals, and few are yet available for trees (Brotherstone *et al.*, 2011).

Genetic approaches based on the use of markers. Phenotypic resemblance between trees can be explained to a certain quantifiable extent by genetic resemblance or, equivalently, by gene sharing. Usually, this genetic resemblance is magnified in experimental setups of quantitative genetic approaches by using close relatives, like full-sibs in a progeny test. Genetic resemblances do exist in natural populations, although they tend to be at a comparatively lower level and with a larger variation than their artificial counterparts. Given the fact that forest trees usually form large outcrossing populations, with the resulting feature of high levels of recombination between gene variants, only small fractions of contiguous "non-recombinant" DNA containing few genes are shared between individuals in the population. In these circumstances, any genetic resemblance underlying a phenotypic similarity involves ultimately narrow intervals around the causal loci in the genome, which facilitates their identification. This logic has driven the development and implementation of what has been called **association genetics**, as a way of decomposing genetic variation into causal DNA variants. Firstly applied in humans, this technique has been amply embraced by forest geneticists. Many gene-to-phenotype associations in trees have been reported involving adaptive traits related to phenology, pest resistances and drought tolerance. It appears, however, that large fractions of genetic variation are yet to be linked to their causal loci. These approaches as well as others for assessing growth traits in trees from the gene point of view are reviewed in Neale & Kremer (2011).

However, without lessening the importance of the choice in analytical approaches, what remains of importance here is whether genetic differences between individuals in their NoRs or competing abilities can be readily detected and linked to the relevant environmental variables. This knowledge would be precious when predicting potential adaptability of trees under novel scenarios.

5.7. Do we need to re-think our field test evaluation?

Previous analyses relied ultimately on adequate field testing, where environment and silviculture are the closest to those expected in true plantations. What we have seen up to now suggests that the environmental part in the evaluation process needs also a close scrutiny. Environmental terms are no longer residual components in a *black-box*; efforts have to be made to identify those that interact actively with the genetic terms.

This can be looked at by two complementary approaches. The first comes at the phase of deciding the layout and range of field testing and concerns basically the choice of environments. We must stress here the importance of multiple sites testing, without which NoRs and adaptive capabilities cannot be gauged. The choice of environments must fall into what is *normal* for the current use of the species, but also some extra environments should go beyond *normality* and explore marginal conditions under current use scenarios. This extra effort might payback in revealing **hidden plasticity** reactions or newly expressed sections of the NoR that may represent the raw material for subsequent evolution (Schlichting & Smith, 2002).

The second approach comes after completion of field recording. This concerns different evaluation methodologies aimed at splitting more precisely some of the components of variation into additional terms. Some of these methodologies account for spatial patterns in the field test that may not be well handled by the original blocking design. Other methods aim at reflecting tree-to-tree interactions into a topological manner. The former spatial methodologies, which we won't develop here for simplicity, belong to the rapidly expanding family of **spatial statistics** (Dutkowski *et al.*, 2006; Cappa & Cantet, 2007). Of course, the use of spatial statistics goes beyond the scope of adaptive traits. Any phenotypic

measure could potentially benefit from spatial adjustments, delivering records that are less prone to bias by uncontrolled or *hidden* environmental factors.

Great efforts have already been done in indexing field test data at European level in newly created databases, which greatly helps to identify common axes across existing experiments over vast geographical ranges. It is clear also that the extra testing effort that new traits require would only be reasonable to attain in the framework of European or international initiatives.

5.8. Conclusion

Adaptation is the visible part of evolution. For trees, unique by being long-living and sessile, these adaptations can be particularly an expression of plastic responses to surrounding conditions. Growth is the commonest trait for breeding, which ultimately reflects the well-being of a tree. Growth is highly integrative, involving the interplay of multiple underlying traits that are at their turn functions of environmental clues. However, identifying the origins of plasticity and its interplay with the changing environment cannot advance from such integrative traits. We suggested the framework of phenotypic plasticity for comprehending how trees adapt to their environment. We illustrated phenotypic plasticity by the link between wood formation and weather conditions. We summarized some of the quantitative methods that are available for assessing the genetic variability in adaptive traits, and the ways to improve field assessments.

References

Breda N., Huc R., Granier A., Dreyer E., 2006. Temperate forest trees and stands under severe drought: a review of ecophysiological responses, adaptation processes and long-term consequences. Ann Forest Sci 63: 625-644.

Brotherstone S., White I.M.S., Sykes R., Thompson R., Connolly T., Lee S., Woolliams J., 2011. Competition Effects in a Young Sitka Spruce (*Picea sitchensis,* Bong. Carr) Clonal Trial. Silvae Genet 60(3-4): 149-155.

Cappa E.P., Cantet R.J.C., 2007. Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. Can J Forest Res 37: 2677-2688.

Cochard H., Holtta T., Herbette S., Delzon S., Mencuccini M., 2009. New insights into the mechanisms of water-stress-induced cavitation in conifers. Plant Physiol 151: 949-954.

De Jong G., Bijma P., 2002. Selection and phenotypic plasticity in evolutionary biology and animal breeding. Livest Prod Sci 78: 195-214.

DeWitt T.J., Scheiner S.M., 2004. Phenotypic variation from single genotypes: a primer. In: Phenotypic Plasticity (DeWitt, T.J., Scheiner, S.M., eds). Oxford Univ Press, NY.

Dutkowski G.W., Costa e Silva J., Gilmour A.R., Wellendorf H., Aguiar A., 2006. Spatial analysis enhances modelling of a wide variety of traits in forest genetic trials. Can J Forest Res 36: 1851-1870.

Griffing B., 1967. Selection in reference to biological groups. I. Individual and group selection applied to populations of unordered groups. Aust J Biol Sci 10: 127-139.

Hewitt N., Klenk N., Smith A.L., Bazely D.R., Yan N., Wood S., MacLellan J.I., Lipsig-Mumme C., Henriques I., 2011. Taking stock of the assisted migration debate. Biol Conserv 144: 2560-2572.

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

Meyer K., Kirkpatrick M., 2005. Up hill, down dale: quantitative genetics of curvaceous traits. Philos T R Soc B, Biol Sci 360: 1443-1455.

Muir W.M., 2005. Incorporation of competitive effects in forest tree or animal breeding programs. Genetics 170: 1247-1259.

Neale D.B., Kremer A., 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet 12: 111-122.

R Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sánchez-Gómez D., Velasco-Conde T., Cano-Martín F.J., Ángeles Guevara M., Teresa Cervera M., Aranda I., 2011. Inter-clonal variation in functional traits in response to drought for a genetically homogeneous Mediterranean conifer. Environ Exp Bot 70: 104-109.

Savolainen O., Pyhajarvi T., Knurr T., 2007. Gene flow and local adaptation in trees. Annu Rev Ecol Evol S 38: 595-619.

Schlichting C.D., Pigliucci M., 1998. Phenotypic evolution: a reaction norm perspective. Sinauer Assoc; Sunderland, MA, USA.

Schlichting C.D., Smith H., 2002. Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. Evol Ecol 16: 189-211.

Sergent A.S., 2011. Diversité de la réponse au déficit hydrique et vulnérabilité au dépérissement du douglas. PhD Thesis, Université d'Orléans.

Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J., Collingham Y.C., Erasmus B.F.N., de Siqueira M.F., Grainger A., Hannah L., *et al.*, 2004. Extinction risk from climate change. Nature 427: 146-148.

Wang C., Andersson B., Waldmann P., 2009. Genetic analysis of longitudinal height data using random regression. Can J Forest Res 39: 1939-1948.

Wimmer R., 2002. Wood anatomical features in tree-rings as indicators of environmental change. Dendrochronologia 20: 21-36.

Chapter 6
The importance and management of genetic diversity in forest trees

Gunnar Jansson[1] and Alan Harrison[2]

[1] Skogforsk (The Forestry Research Institute of Sweden), Uppsala Science Park, SE-751 83 Uppsala, Sweden

[2] Forest Research, Northern Research Station, Roslin, Midlothian EH25 9SY, UK

Questions that will be answered in this chapter:

- What is genetic diversity?
- How can genetic diversity be measured?
- How does genetic diversity break down between and within populations?
- How can genetic diversity be controlled in breeding and deployment?
- How can forest management affect diversity?
- What diversity is needed to meet market demands?
- How can future techniques increase our knowledge?

6.1. Summary

This chapter summarises different ways in which genetic diversity can be measured. Genetic diversity allows a species to adapt to environmental changes and is also needed for long-term progress from breeding for economically significant traits. We show how the observed diversity can be partitioned into a genetic and an environmental component. Different considerations concerning genetic diversity in breeding and deployment are discussed. We also identify how forest management can influence genetic diversity. Finally, we address how development of new molecular tools can increase our knowledge of genetic diversity.

6.2. Introduction

Within any form of life, greater genetic diversity within a population allows a species to be more adaptable and more flexible in response to any change or action that it may experience. Though some individuals may be weakened or succumb to the effects of a particular change, the population as a whole has a greater chance of survival because of the diversity in response. The consequences of this on-going process are evolutionary and ultimately speciation.

This process is true with forest trees, but there are also more immediate and economic reasons for considering the benefits of diversity. These benefits principally impact upon two broad areas of interest, namely environmental changes and economic needs. As tree-breeders, an understanding of diversity and how traits can be identified and isolated or combined, allows us to contemplate the breeding of genotypes that are better suited to a particular situation, or produce a higher quality end-product. The wider the initial diversity within a population, the more likely is the potential for identifying and developing traits that meet the adaptive and economic needs, and the subsequent response will be more rapid.

Environmental impacts may be both negative and positive, but, either way, the prime concern is that any adverse effects on tree growth and subsequent timber yield are minimised. The speed and persistence

of an environmental change is critical. Survival of short-term catastrophic events *e.g.* drought, unseasonal cold, will depend upon a population's existing level of adaptation and phenotypic plasticity to that event. Higher diversity gives a greater chance that a proportion of the population will be less damaged. To cope with more gradual change, trees can adapt in successive generations in timescales of centuries, and populations can migrate at a slower rate (millennia) if there is a suitable environmental cline (pathway).

Of particular current concern is the ability of trees to withstand changes in climate such as a gradual warming, drying or increased rainfall. The predicted changes are likely to occur on a timescale that is short in relation to the lifetime of a tree, within 100 years, making any significant species adaptation unlikely. The climate changes can be managed to some degree by change in the choice of provenance or using alternative species. The potential for increased but unknown climatic volatility (Read *et al.*, 2009) does make this an inexact science, but the retention of diversity remains an important consideration.

Positive aspects of climate change include increased temperatures and raised atmospheric CO_2, both of which have the potential to increase tree growth, assuming other factors such as nutrients, water and impact of pathogens are not limiting. Clearly, a more diverse tree population should be more likely to be able to utilise these opportunities.

The resistance to pests and diseases is another major environmental issue where the health and survival of a species can be improved if it has greater genetic diversity. Pathogen resistance can vary widely between genotypes both within and between provenances of a species. If the genetic base is narrowed and an individual in that population is found to be infected, there is a far greater likelihood of the pathological agent attacking the rest of that population. Reducing diversity down to just a few clones has been shown to have devastating effects in the past, *e.g.* the impact of *Melampsora* rust infection on poplars. Although there are strong productivity arguments for utilising clonal forestry (Libby & Rautner, 1984), the risks of a wipe-out due to a pest or disease attack needs to be carefully weighed against the benefits. The precautionary principle always argues for greater diversity.

Economic factors are driven by the product mix required to form the crop. Processors prefer product uniformity as it ensures ease of utilisation. In this case diversity can often be seen as a negative if it leads to a less uniform crop. Therefore, there needs to be a balance between achieving the correct processing standard and specification of crop while still retaining genetic diversity as a buffer against the environmental concerns. The processing requirements for a crop planted now will probably change in the future from what it is now, so a degree of diversity may prove essential to meet an unknown future.

The 'unknown', whether it is environmental or economic, is the ultimate reason for keeping genetic diversity. We cannot predict the future with any accuracy and, therefore, need to retain a wide gene pool to have a resource that is able to react or be manipulated to meet change.

6.3. What is genetic diversity?

Genetic diversity can be considered at a number of different levels:

I. between species;
II. between populations within species;
III. among trees within a population;
IV. among trees within a stand (c.f. local diversity; Woolliams & Wilmut, 1999)
V. within trees

Here we address the genetic variation between populations (ii) and within populations (iii). Diversity of trees among trees in a stand is of importance for deployment, especially in clonal forestry as discussed by Woolliams & Wilmut (1999). A first step in evaluating the evolutionary as well as the tree

breeding potential is to measure the level of genetic variation within and among populations (White *et al.*, 2007). Genetic diversity can be qualitative, such as colour, or quantitative, such as height. Quantitative traits are generally controlled by many genes, each assumed to contribute only a relatively small amount, but collectively explaining a considerable amount of the observed variance. In breeding, quantitative traits are often of greatest interest because this includes most adaptive and economic traits of significance.

Diversity can be observed and measured in different ways within a species. A useful way is to establish field or nursery experiments and to quantify genetic variation by statistical methods for analysis of variance, where the total variance is decomposed into identifiable components. If we have independent variables, the total variance is the sum of the variance of the independent variables. In this way the observable variation can be subdivided into a genetic and an environmental effect which are assumed to be independent (See Box 6.1 for more details).

6.4. Population variation

Genetic variation can be affected by genetic drift, migration, selection and mutation and it is these factors that through many generations have developed the differences between populations (*e.g.* Falconer & Mackay, 1996). These processes affect the gene frequencies and thereby the genotype frequencies. Genetic drift cause random fluctuations in populations and can lead to fixation or loss of alleles in the population over many generations, with the scale of these changes becoming larger in smaller populations. Migration of individuals moving from one population to another can bring in new alleles and so increases the variation within a population but reduces differences between populations. Studies of pollen grains from wind-pollinated species such as spruce and pine show that they may spread their pollen over hundreds of kilometres. When selection occurs individuals with favourable alleles have an advantage for the next generation and contribute a disproportionate number of alleles to the next generation. The favourable alleles will then increase in the population. Mutation is responsible for creating new variation in the population. Mutations occur at low frequency and it is only through natural or artificial selection of

Box 6.1. Subdivision of the phenotypic variance

The phenotypic value (P) that we observe can be decomposed into a genetic (G) and an environmental (E) effect, *i.e.* $P = G + E$, where P has total phenotypic variance σ_P^2 and G has total genetic variance σ_G^2 and E has total environmental variance σ_E^2. Assuming G and E are independent gives $\sigma_P^2 = \sigma_G^2 + \sigma_E^2$ σ_G^2 includes a number of different sources of genetic variation including additive, dominance and epistatic effects. The ratio of the total variance that is genetic is called the broad-sense heritability, *i.e.* $H^2 = \sigma_G^2/\sigma_P^2$. Heritability expresses the portion of variation in the population that is attributable to genetic differences among individuals. Heritability can vary from 0 to 1, with 0 meaning no genetic and totally environmental effects, and 1 meaning that there are only genetic effects and no environmental effects. The heritability of a trait is unique to each population and can therefore differ between populations. The heritability also varies between traits.

To be able to estimate H^2 we need identical genotypes which we can get from clonal trials. However, it is more common to focus on the additive genetic effects, σ_A^2, which is the variance of the breeding values in the population. The narrow-sense heritability is the ratio between the additive genetic variance and the total phenotypic variance, *i.e.* $h^2 = \sigma_A^2/\sigma_P^2$. It is this part of the variation that can be used in breeding to create changes in the population mean by selection. (See Chapter 3, Box 3.2).

favourable alleles, or chance drift, that the mutation will have any effect in the population. Here random processes play an important role, meaning that only a small fraction of the mutations will increase to such a frequency that they will be of significance in the population.

Most studies of quantitative traits are based on population (or provenance) and/or progeny trials. A provenance is a population or group of individuals of the same species originating from a defined geographic area. Seed from open pollination or from controlled crosses are sown and planted in nurseries or in field trials (see 'Foreword'). Population and family effects can be separated based on analysis of variance. We can subdivide the observed breeding value of an individual into a population effect and an individual effect. Since the populations will have different mean values we can estimate the population variance (σ_{Pop}^2) and the individuals' deviations from the population mean (σ_W^2) will also have a variance. The genetic diversity of a species can then be subdivided into a genetic component for population and a genetic component within populations. The ratio $\sigma_{Pop}^2/(\sigma_{Pop}^2 + \sigma_W^2)$ can then be used as a measure of the importance of variation between populations.

The patterns of variation observed in provenance trials show geographic variability on a broad scale. However, genetic variation is also present among offspring from natural populations within the same provenance region and among progenies from trees in the same population. This variation can often be larger than the variation between mean performance among provenances (*e.g.* Zobel & Talbert, 1984; White *et al.*, 2007).

6.5. Genetic diversity based on markers

Another way to estimate genetic diversity is to look at genetic markers. Three statistical methods based on markers are commonly used to quantify genetic variation within populations (White *et al.*, 2007):

- proportion of polymorphic loci;
- mean number of alleles per locus;
- expected heterozygosity

The proportion of polymorphic loci (P) is estimated as N_p/r where N_p is the number of polymorphic loci and r is the total number of loci sampled. The mean number of alleles per locus (A) is estimated as $\Sigma m_j/r$, where m_j is the number of alleles at the j^{th} locus. An alternative measure to number of alleles is allelic richness, which is a sample size-adjusted measure of the number of alleles in the population (*e.g.* Foulley & Olliver, 2006; Leberg, 2002). Expected heterozygosity (H_e) is the probability that two randomly sampled alleles drawn from individuals in a population differ from each other, which is the heterozygosity. The expected heterozygosity at a locus is then estimated as $1 - \Sigma x_i^2$, where x_i is the observed frequency of the i^{th} allele at the locus and is summed over all alleles at the locus. The mean expected heterozygosity is the expected heterozygosity averaged over all loci sampled.

6.6. How can genetic diversity be controlled in breeding?

Applied tree improvement programs start with selection, testing and mating activities in the breeding cycle (see 'Foreword'). The next step is then to form propagation populations for deployment of the improved varieties. The activities within the breeding cycle generally aim to improve traits of economic significance, while at the same time maintaining a broad genetic base for long-term gains over many generations. Risks for reducing genetic variation can be handled by the number of unrelated individuals in the breeding population and by keeping track of the relationship between the individuals to reduce inbreeding in the breeding population (Box 6.2, see also Chapter 3, Box 3.4).

Box 6.2. Inbreeding

Inbreeding occurs when related individuals are mated with each other, and reduces the variation in the population. The inbreeding coefficient, F, is used to measure the amount of inbreeding. Whilst inbreeding is a natural phenomenon, too rapid a rate of inbreeding increases the risk of deleterious effects appearing in the population, and in good population management this rate is controlled. As an example, the Swedish breeding programs for Norway spruce and Scots pine are divided into 20 sub-populations per species. Each sub-population has at least 50 individuals. A maximum increase in inbreeding of 1% per generation has been assumed to be an acceptable rate of inbreeding in each sub-population (Danell, 1991).

As mentioned above it is the additive variation that allows a change in the population mean by selection. The additive genetic variance $V_A(t)$ after t generations of random selection and mating for different population sizes can be calculated as:

$$V_A(t) = \left(1 - \frac{1}{2N_e}\right)^t V_A(0)$$

where N_e is the effective population size and $V_A(0)$ is the additive variance we start with. The effective population size is the number of individuals that when randomly selected and mated would be expected to have the same rate of inbreeding as the population itself. See Oldenbroek (2007) for more detail in what determines N_e. Fig. 6.1 shows how the remaining additive variance depends on the population size. Lack of genetic variation will cause stagnation in the genetic gain. However, increasing N_e will slow down the reduction in variance.

6.7. Rare alleles

Genetic variation is needed for selection and to get improvement through breeding, but it is also required for natural evolution of a population. Genetic variation depends on allelic effects and their fre-

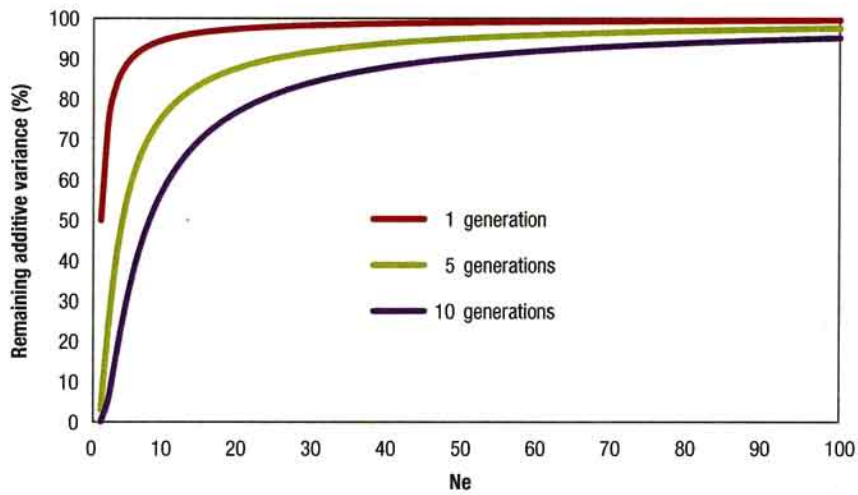


Figure 6.1. The additive genetic variance remaining after different numbers of generations as a function of effective population size, assuming random selection and mating. The calculations are based on the formula above.

quencies in the population. In a breeding program a limited number of trees are used as parents for future forests but the genetic variation can be handled without keeping all alleles in the population. Alleles in very low frequencies (*i.e.* rare alleles) do not contribute significantly to the additive variance. There are lots of loci with rare alleles in a population but each only occur in few individuals and will therefore not affect the variation in the population. They will not affect the population unless their frequency in the population increases. The frequency of such alleles can only have an effect in the long run if they have some advantage and increase in the population either by natural or artificial selection. Rare alleles may therefore only be of importance in long-term breeding and for evolution of the species in *e.g.* changing environments. New rare alleles can arise by mutations but most of the rare alleles disappear by chance, a natural process, since they only exist in few individuals, and it is not possible for a population to retain all such mutations. The lower limit of individuals to keep alleles for gene conservation has been debated. Table 6.1 shows the lowest number of individuals in the population to keep alleles at only one locus with a certain probability, based on the worst case scenario considered by Gregorius (1980).To keep alleles at a frequency of 10% with a probability of 95 % requires only 51 individuals. But to keep low-frequency alleles (1%) with a high probability (99%) around 1,000 individuals are needed. In practice the number of unrelated individuals in breeding populations will be at least 100 to 200. The Swedish breeding programs for Scots pine and Norway spruce have more than 1,000 individuals for each species. The total population is divided into some 20 subpopulation with 50 individuals in each (Danell, 1991).

Table 6.1
The number of individuals in the population to keep all alleles at a locus with a given probability. Numbers are based on worst case scenario, assuming sampled individuals are homozygotes (Gregorius, 1980). If Hardy Weinberg is assumed, numbers required are ½ those stated here, rounded up to an integer

Frequency of rarest allele	Probability for keeping rarest alleles		
	95%	99%	99.9%
20%	21	28	39
10%	51	66	88
5%	117	149	194
1%	754	916	1146
0.8%	972	1174	1462

6.8. Balancing gain and diversity

Genetic gain and genetic diversity must be balanced in a breeding program to get long-term gains. Unrestricted selection leads to selection of many individuals from a few top ranking families, which will result in loss of genetic diversity and ultimately lower gains in future generations. Therefore, selection methods that optimize selection gains with constraints on the increase of inbreeding in a population are important in breeding programs. One such method that has got much attention is the optimal contribution selection method in which algorithms maximise selection response while constraining the inbreeding in the population by optimizing mating proportions of parents (Meuwissen, 1997; Grundy *et al.*, 1998; Hallander & Waldmann, 2009).

6.9. How can genetic diversity be controlled in deployment?

The propagation population is made up of the best selections in the breeding populations and used to produce plants for the operational forestry. The genetic diversity in the deployment population depends on the genetic diversity in the selections from the breeding population and the deployment method (*e.g.*

seed orchard seed, clones). Reproduction is a necessary part of any tree improvement program and can be done by:

- seed from unimproved provenances or stands;
- open-pollinated seed orchard seed;
- seed from selected full- or half-sibs, *i.e.* family forestry; and
- vegetatively propagated clones, families or other selections by rooted cuttings or somatic embryogenesis, *i.e.* clonal forestry.

The order of the alternatives show increasing gain but this may also reduce the genetic diversity if the number of individuals is too small. How the improved genetic materials are used and how forests are planted will affect this.

Provenances. Provenances refer to a population or group of individuals from a defined geographical area. Genetic variation among provenances can be large, especially for species growing in many diverse climates (White *et al.*, 2007). Often patterns of adaptation can be found along latitude or climatic gradients. Transfer of provenances has in many cases been successful but not in others. Provenance trials have been established to match a population originating from a defined geographical area with the environment where it is going to be planted.

Many species show a large clinal variation between provenances. One example of a successful transfer of provenances is in increasing adaptation and growth. In northern Sweden, southward transfer of Scots pine provenances has led to increased survival and higher yield compared to local provenances (*e.g.* Persson, 1994). In a similar way, use of late flushing Norway spruce from Belorussia and the Baltic countries has led to reduced frost damage and higher yield on spring-frost exposed sites in southern Sweden (*e.g.* Werner *et al.*, 1991). For Norway spruce the timing of budburst and budset is important for adaptation to the climatic conditions. It will be even more important with climate change in the future.

Seed orchards. Seed from seed orchards comes from open pollinated seed among the best selections from a breeding program. There are no signs today that seed from seed orchards cause problems due to genetic variation providing a sufficient number of parent clones are included. Seedlings from seed orchards may also be better adapted to climate change than seedlings from natural regeneration, since their parents have been tested in different environments. Fewer clones will increase the gain because of higher selection intensity but the genetic diversity will be lower. Sometimes a country may have legal restrictions on the number of parent clones in an orchard. As a rule of thumb Lindgren & Prescher (2005) suggest that 20 parent clones are sufficient for Swedish conifers. Seed orchards experience a high degree of contamination from outside pollen adding to the genetic diversity, and consequently the genetic diversity in seed-orchard seed can even be higher than in a natural stand. This is because the parent trees are selected in different stands and are therefore less related with each other than those in a natural stand. Since the trees are selected over a large area there may also be a larger genetic variation among the seed orchard clones than among trees in a naturally regenerated population. Adams *et al.* (1998) found higher genetic diversity in seedling stock than in natural regeneration. This was presumably because the seedling stocks were derived from trees from many stands; so in this case, more diversity with more gain.

Family forestry. Family forestry is based on control-pollinated seedlings from the best families or tested half-sib families. If only one full-sib family is planted it includes the same variation as half the additive variance (*i.e.* the variance within a full-sib family). As breeding progress new parents will be used to produce the full-sib families. Compared to seed orchards the gain will be higher, since only a

few proven good parents are used and there will not be any pollen contamination. However, the genetic diversity in the plantations will be reduced.

Vegetative propagation can be based on multiplication of full- or half-sib families which will give the same diversity as above. Multiplication of full-sib families starts with production of seed from control pollination to get seed from full-sib families. Multiplication can also be based on half-sib families but the gain will then be lower since the father is unknown. These are then multiplied vegetatively to increase the number of plants (*e.g.* White *et al.*, 2007).

Clonal forestry. Clonal forestry means large-scale plantation of relatively few clones. Individual clones are tested and the best clones are then used for mass-propagation. The diversity depends on the number of clones used and how they are related. Monoclonal plantations or plantations with few clones will give a more uniform forest but risks are being taken regarding the consequences for the environment. Monoclonal plantations are used in intensively managed industrial plantations with short rotations *e.g.* eucalypts in Brazil and South Africa (*e.g.* White *et al.*, 2007). Whilst all the trees in a monoclonal plantation will have the same genetic effect there will be variation in the phenotype due to the environment. Nevertheless the genetic uniformity increases the possibility that a pest or disease can spread more easily in the stand. This is the issue of local diversity described by Woolliams & Wilmut (1999). However, a mix of clones can reduce risks for pests and disease.

6.10. Influence of using improved varieties on natural stands

When the improved varieties start flowering they will mate and influence the surrounding natural populations. Trees poorly adapted tend to die and will not affect the population, while surviving trees will spread their genes in the population. The genetic variation in the population should increase when selections from different origins are mixed. The question becomes "Is this good or bad?" It is the better adapted trees that will survive and spread their alleles; this may be the local or the introduced trees but most likely an evolving mix of both depending on natural and man-induced selection pressures and the relative fitness and quality of each group.

The conclusion for deployment of improved genetic varieties is that diversity of the improved plants depends on the diversity in the breeding population, the number of parents involved in generating the deployed population and how these parents are related to each other. At a landscape level different genetic varieties will be used at different times and the landscape will probably also include material from different deployment methods which will increase the genetic diversity. If the same seed source is used over large areas, diversity at landscape level will decrease. A general conclusion is that planted forests of mid- to long-rotation lengths (30 to 100+years) are as genetically diverse as the natural stand that they replace when measured as allelic or as genotypic variation (Skrøppa, 1994).

6.11. How does management affect diversity?

The management of a forest is driven by man-made objectives, which can vary from forest preservation to full-scale commercial production. Each management scenario has its own idea of a 'perfect' tree that meets all its specifications, but this search for perfection tends generally to reduce genetic diversity, as the best trees are selected during the stages of the management process (breeding, thinning etc.). This reduction in diversity due to management is not necessarily negative. Focusing on specific well-chosen traits produces a tree that is fit and fit for purpose, assuming, of course, that the specific traits are well chosen and does not change within the life-time of the tree. The combination of traits influencing the selection process, whether known or not, may also contribute to producing a population that is superior

to its parents in unspecified ways. It should also be remembered that naturally regenerated forests may also have low diversity and may also lack resilience to unusual challenges. In a naturally regenerated and managed Scots pine forest García-Gil *et al.* (2009) found high levels of inbreeding in a study of fine-scale spatial structure.

Having established a forest, the next major management impact is thinning of the crop. If done non-selectively, *e.g.* by line-thinning, effectively taking out a proportion of the crop at random, the effect upon diversity is likely to be small, and probably insignificant. However, when thinning is done selectively, the impact on diversity is potentially greater. The selection process is based on phenotypic characteristics *e.g.* height, diameter, straightness and branching, to identify the trees to be removed. From a management perspective, this is crop improvement which is targeting uniformity, and diversity will be reduced, though the level of reduction will depend upon the heritability of the genetic traits expressed in the phenotype, and the level of heterozygosity within the population. As many crops go through a series of thinning cycles, the final trees, which may be 20% or less of the original number planted, are likely to have reduced diversity to some extent, though they will be 'superior' trees as defined by the selection criteria.

If the decision is made to naturally regenerate the stand after felling, this can be done in several ways, depending upon the silvicultural system used *e.g.* felling the crop immediately after it has seeded, or retaining seed trees of superior phenotypes as a post-felling seed source. Individual trees vary considerably in their ability to produce seed, and even in a good seed year, not all trees will produce equally large quantities of seed. Therefore, when choosing to fell after a good seedling year there will be a potential bias within the regenerating trees toward the prolifically seeding genotypes, which may not best meet the crop objectives. However, given the high heterogyzosity of most tree populations and good pollen flow, particularly with wind-pollinated species, the resultant inter-breeding and recombination of genes is unlikely to seriously affect diversity, though the resultant genotypes may not fully recreate the parental mix. In regeneration using post-felling selected seed trees, the initial diversity is reduced, but recombination of genes will recover most of the initial diversity if the number of trees is not too low (cf. Fig. 6.1). However, once again the management objective is crop improvement, where a degree of diversity may be sacrificed for a better product.

Low-impact silvicultural systems (LISS) have a multi-aged structure, with natural regeneration and avoid the clear-felling of large areas *e.g.* continuous cover forestry (CCF) and shelter-wood systems. These can be managed to conserve genetic diversity by always having a large proportion of the crop, representative of its overall diversity, in close proximity to each other at an age, so that a wide sample can contribute to the regeneration population.

Conversely, in the management of a forest or woodland for more environmental or conservation objectives, there may well be a presumption against change. In the case of native woodlands, populations can become physically isolated as a consequence of decisions made either externally in removing neighbouring populations, or to protect an ancient stand, where the latter is often enforced by legislation. The effect of this on diversity will depend upon the state of the population prior to being 'protected'. Many native forests have been selectively exploited, often for several rotations, with regeneration coming from the remaining, often phenotypically inferior trees. Studies have shown that this reduces diversity and also the quality of the timber resource (dis-genetic selection). In severe cases all that may remain are remnant woodlands, which often contain only a small number of trees. As a breeding population, over successive generations, there will be an increased tendency to in-breed, with genotypes becoming more closely related and recessive gene expression greater, giving rise to the possibility of inbreeding depression, *i.e.* loss of vigour and fertility, and ultimately extinction with no built-in variation to adapt to climate change. In these situations, the informed management solution is to increase diversity by introducing external genotypes into the population by planting or direct seeding, but this is rarely carried out (Peterken, 1996).

In conclusion, management practices in forestry generally concentrate on tangible economic objectives. Thus any effects are a by-product of the systems used to achieve an end goal, and generally result in a reduction of diversity. The degree of diversity change over the rotation is dependent upon the initial stocking diversity. The management effect on genetic diversity is relatively small and planted forests are often as genetically diverse as the natural stands they replace. An exception from this is when monoclonal plantations, or plantations with very few clones, are used.

6.12. The role of diversity in meeting market demand

There are principally three main markets that forestry needs to supply:

— the more traditional markets for timber;
— wood fibre; and
— biomass, which has expanded more recently.

The requirements for each are different and shortage of supply can lead to market distortion. This is currently happening between biomass and wood fibre, where the increasing use of biomass to meet targets for reduction in carbon emissions is raising the price and reducing the availability of smaller logs that would have previously been targeted solely at the fibre market.

The basic quality criteria for each market product are:

Timber: straightness, lack of defects (spiral grain, knots, cracks), strength (tensile and elastic);
Fibre: length, structure, strength, colour; and
Biomass: calorific value, density, moisture content, non-woody content.

There are grades of quality within each product class, but timber has the more stringent criteria than fibre. In contrast, biomass is a different type of product whose criteria can be met by timber or fibre products without reference to their qualities.

Tree breeding programmes continue to look at improving the quality and productivity for all three markets across a range of species, although programmes are less developed for biomass production, except for willow. An example is the breeding programme for Sitka spruce in the UK (Lee, 1999). These programmes all commenced from a diverse base population, selecting phenotypes that matched specified criteria followed by breeding and testing. As new market needs and technologies are developed, *e.g.* the move towards the conversion of woody biomass into liquid biofuel, we need to re-examine the originally selections and check if they meet the new criteria. Currently it is possible to return to the diversity of the base population to select for these different traits if necessary.

The main role of diversity from the market viewpoint is to provide a suite of attributes that can be chosen, either individually or in combination, to improve the usability of product. Diversity provides the future proofing in today's population to make the best of whatever demands arise for tomorrow's population.

There is also the possibility of enhancing diversity by inter-specific hybrids, which may bring together required traits and amplify them through 'hybrid vigour'. Several commercially used inter-specific hybrids are regularly used in the genera *Larix, Populus* and *Eucalyptus,* for timber, fibre and biomass. This utilises the diversity of two species populations, though the hybrids are not themselves genetically diverse as they are often comprised of only a small group of clones, which may have reduced fertility (Drake, 1981). Such crossbred trees are an example of where the trees may be very similar to each other, but where each tree has a large diversity arising from the differences between the two parental species.

Markets can change rapidly, but forestry production moves much more slowly. Therefore, a more diverse general plantation has a greater flexibility to at least partially meet a need, than a selective crop managed within a very narrow range of objectives.

6.13. How can future techniques increase our knowledge?

Inbreeding will reduce the genetic variation in the population and reduce long-term gains. As discussed in Chapter 3 markers can be used to construct the relationship matrix. In the numerator relationship matrix (**A**) the pedigree provides the expected relationship among individuals. With large marker density the markers can provide a more precise relationship matrix (**G**). This will generate more gain in breeding and deployment for a pre-specified rate of inbreeding.

It is assumed that most traits under selection are polymorphic meaning a large number of alleles are affecting the trait each by a small amount. If we could look directly at the genes controlling traits of interest or markers linked to these traits that would result in increased possibilities to control the variation of the trait under selection. In the breeding populations we could then retain genetic variation, such as traits related to adaptation, to meet possible future needs. This knowledge could also focus on selecting trees with genes that are well adapted over a range of environments (*i.e.* high plasticity).

It could also be possible to see which genes affect multiple traits and thereby how selection for one trait will affect other traits. This will reduce the risk for reducing the variation in one trait while selecting for another trait.

6.14. Conclusions

The genetic variation between and within populations are important for evolutionary potential as well as for breeding. Genetic diversity can be estimated from data based on measured traits or based on genetic markers. Breeding requires a balance between gain and genetic diversity in the breeding populations for long-term improvement of traits. Regeneration after harvesting will affect the genetic diversity of the new forest. The genetic diversity in the deployment population depends on the number of selections from the breeding population and the deployment method. In general forest management will only result in small effects on genetic diversity and planted forests are often as genetically diverse as the natural stand they replace. On the other hand if only a few clones or monoclonal plantations are used this will reduce the genetic variation significantly. The future forests will consist of a mix of naturally regenerated and planted stands. The latter will come from stand seed, seed-orchard seed and vegetatively propagated forests. In addition to this stands will be set aside for nature conservation. All these activities will contribute to a high genetic diversity. A main role of diversity from the market viewpoint is to provide a suite of attributes that can be chosen in the future to improve the usability of products.

References

Adams W.T., Zuo J., Shimizu J.Y., Tappeiner J.C., 1998. Impact of alternative regeneration methods on genetic diversity in coastal Douglas-fir. Forest Sci 44: 390-396.

Danell Ö., 1991. Survey of past, current and future Swedish forest tree breeding. Silva Fennica 25: 241-247.

Drake D.W. 1981. Reproductive success of two *Eucalyptus* hybrid populations. Aust J Bot 29: 25-35.

Falconer D.S., Mackay T.F.C., 1996. Introduction to quantitative genetics, 4th edition. Longman Group Ltd, Harlow, Essex, England.

Foulley J.L., Olliver L., 2006. Estimating allelic richness and its diversity. Livest Sci 101: 150-158.

García-Gil M.R., Oliver F., Kamruzzahan S., Waldman P., 2009. Joint analysis of spatial genetic structure and inbreeding in a managed population of Scots pine. Heredity 103: 90-96.

Gregorius HR, 1980. The probability of losing an allele when diploid genotypes are sampled. Biometrics 36: 643-652.

Grundy B., Villeanueva B., Woolliams J.A., 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genet Res (Camb) 72: 159-168.

Hallander J., Waldmann P., 2009. Optimum contribution selection in large tree breeding populations with an application to Scots pine. Theor Appl Genet 118: 1133-1142.

Leberg P.L., 2002. Estimating allelic richness: Effects of sample size and bottlenecks. Mol Ecol 11: 2445-2449.

Lee S.J., 1999. Improving the timber quality of Sitka spruce through selection and breeding. Forestry 72: 123-146.

Libby W.J., Rautner R.M., 1984. Advantages of clonal forestry. The Forestry Chronicle 60: 145-149.

Lindgren D., Prescher P., 2005. Optimal clone number for tested clones. Silvae Genet 54: 80-92.

Meuwissen T.H.E., 1997. Maximizing the response to selection with a predefined rate of inbreeding. J Anim Sci 75: 934-940.

Oldenbroek J.K. (ed), 2007. Utilisation and conservation of farm animal genetic resources. Wageningen Acad Publ, Wageningen.

Persson B., 1994. Effects of climate and provenance transfer on survival, production and stem quality of Scots pine (*Pinus sylvestris* L.) in northern Sweden. Swedish Univ Agric Sci, Dept of Forest Yield Res. Report 37, 42 pp.

Peterken G.F., 1996. Natural woodland. Ecology and conservation in northern temperate regions. University Press, Cambridge, UK.

Read D.J., Freer-Smith P.H., Morison J.I.L., Hanley N., West C.C., Snowdon P. (eds), 2009. Combating climate change – A role for UK forests. The Stationery Office, Edinburgh.

Skrøppa T., 1994. Impacts of tree improvement on genetic structure and diversity of planted forest. Silva Fennica 28: 265-274.

Werner M., Karlsson B., Palmér C.H., 1991. Ortens gran – ett osäkert alternative i Götaland. [Domestic provenances of Norway spruce – A risky alternative in Southern Sweden]. Institutet för Skogsförbättring. Information Skogsträdsförädling Nr 4.

White T.L., Adams W.T., Neal D.B., 2007. Forest Genetics. CABI Publ, CAB Intnl, Wallingford, Oxfordshire, UK.

Woolliams J.A., Wilmut I., 1999. New advances in cloning and their potential impact on genetic variation in livestock. Anim Sci 68: 245-256.

Zobel B., Talbert J., 1984. Applied forest tree improvement. John Wiley and Sons, Inc.

Acknowledgements

The time involved in compiling this book received funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement n° 211868 (Project: Noveltree). We are grateful to Mrs Lisa Cormack of The Roslin Institute for organising the drafting meeting, reviewing the manuscripts for typographical errors and formatting them. We are also grateful to INIA for their offer to publish our work in this book.

Tree breeding is at a crossroads. Many countries have completed their first generation of traditional breeding by selecting superior phenotypes and estimating breeding values for characteristics of interest following long-term and costly field-based progeny trials. Now there are new technologies around. Tree breeders are looking at advancements made in the areas of crop and animal breeding to see how DNA-markers can be used to speed the process up, increase the genetic gains and lower the overall costs. The challenges facing tree breeders are changing too. In additional to the original reasons for selection and breeding there is often the need for selected populations to remain resilient in the face of a changing climate and increased incidents of damaging outbreaks of new or previously benign diseases. In February 2012 a number of tree breeders from across Europe gathered together in South Scotland for one week to compose different 'Chapters' for this monograph. Their objective was partially retrospective to consider where tree breeding has got to but then to consider how new breeding technologies can help the tree breeder in theory and in practice. The book is a presentation of the current state of the art and how it might develop. Importantly, the book identifies the need to continue high through-put yet accurate phenotyping as well as genotyping, and the need to consider genetic variability well into the future. The book and its component Chapters are presented as a conduit to further reading. They are presented here for the students of tree breeding past, present and future as well as others interested in the possible new directions of tree breeding.

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
INIA
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria