Final Report Summary - PROCOGEN (Promoting a functional and comparative understanding of the conifer genome - implementing applied aspects for more productive and adapted forests.)

Executive Summary:
ProCoGen (PROmoting a functional and COmparative understanding of the conifer GENome - implementing applied aspects for more productive and adapted forests) was a four year European Commission’s Seventh Framework project running from 2012 to 2015. It brought together 23 partners from research, development and technology centres as well as from universities from across the EU and North America.

The main aim of ProCoGen was to develop an integrative and multidisciplinary genomic research in conifers, using high-throughput platforms for sequencing, genotyping and functional analysis, and to unravel genome organization and identification of genes and gene networks controlling important ecological and economic traits, such as those related to the control and the reduction of climatic change impact in relation to growth and adaptation.

For the first time at European scale, ProCoGen brought together many genetic resources of the key species, relevant genetic material, advanced experimental tests as well as pre-existing genomic tools to achieve its goals. ProCoGen focused its research efforts on four model conifer species of high economic and ecologic importance across Europe: (1) Maritime pine (Pinus pinaster, Southwest Europe); (2) Scots pine (Pinus sylvestris, Central and Northern Europe); (3) Norway spruce (Picea abies, Central and Northern Europe); Sitka spruce (Picea sitchensis, of high economic interest in the UK).

Research activity was articulated in five main areas: Development of genomic resources; discovery of adaptive capacities to ensure basic and applied outcomes; comparative genomics to understand conifer evolution and achieving effective management and breeding; computational genomics; and translational genomics (quantitative genomics enabling breeding and resource management). Of course, and crosswise, much effort was devoted to the integration with conifer genomics initiatives worldwide.

Objectives of ProCoGen were completed successfully within budget and on time. Promisingly, outcomes of the project have opened potential and encouraging research lines for future cooperation. Thus, ProCoGen has contributed to improve knowledge of conifer genomics, their genome structure and function, optimizing or developing tailor-made molecular tools, strategies and advanced bioinformatics, to generate new information to improve forest productivity through targeted breeding programs based on better adapted material to regional climatic threats, forest stewardship in response to environmental change as well as conservation efforts. Main impacts of ProCoGen results are related to reinforcing European research competitiveness, integrating European research in the international network for conifer genome analysis, preparing society to better face climate change and improving tree breeding and sustainable forest management. The research, results and deliverables from ProCoGen have been disseminated widely both locally, in the countries of respective ProCoGen partners, and internationally through journal, conference papers and posters as well as training and dissemination workshops and online media finalizing the project with an open final conference were the main outcomes of ProCoGen were highlighted and shared with both scientific community and forest-based sector.

Project Context and Objectives:
ProCoGen context:
Conifers have immense ecological importance, dominating many terrestrial landscapes and representing the largest terrestrial carbon sink. Present today in a large number of ecosystems, they have evolved very efficient physiological adaptation systems. However, our understanding of their potential for such adaptive responses to projected climate changes remains limited, especially taking into account the fact that their genome has diverged from that of angiosperms over 300 million years ago, and the study of their genome is revealing unique information which cannot be inferred from currently sequenced angiosperm genomes.

Conifers are also of great economic importance, as they are primarily used for timber, paper and biomass production worldwide. Domestication of some of these species started about 70 years ago through traditional genetic improvement programmes. It has resulted in advances in growth, wood quality, pest resistance and adaptation, but breeding still remains a slow process because of long generation intervals typical of most conifers and because most traits cannot be correctly evaluated at an early stage. During the past 25 years, more and more sophisticated high throughput analysis have been developed to describe the variability and plasticity of these species at different levels of integration (from genes up to phenotypes) and are now being integrated into breeding to accelerate the domestication process by a more precise exploitation of genetic diversity.

Application of genomics is playing an important role in understanding the evolution, patterns of nucleotide variation and the molecular basis of quantitative traits and adaptation. The use of high throughput analysis makes it possible to consider the study of large and complex genomes, such as those of conifers which, due their large genome sizes (15-50 Gbp), previously have been out of reach for whole-genome sequencing and assembly. In this sense, conifer genome sequencing is providing better knowledge of the genome structure of these very long-lived organisms, enabling long-term evolution studies. The first conifer genome drafts (from Picea abies, Picea glauca, Pinus taeda species) have recently become available.

Conifer genomes are highly conserved, so genomic sequence information for pine and spruce model species could be used to study other conifers of interest. Functional genomics based on high throughput analysis allows addressing studies on conifer productivity and adaptation to environmental change. Massive identification of gene sequences involved in the adaptive response provides information about complete gene families as well as genes of interest. Intact gene identification combined with association studies is helping to dissect productive and adaptive traits, not only improving our knowledge about basic conifer biology, but also to address practical problems for the forest industry as well as problems related to the adaptation and management of conifer forests.

In this context, the ProCoGen project, a FP7-funded enabling project, was launched with the main goal of developing an integrative genomic research programme in conifers, and transform discoveries derived from genome technologies into practical applications in breeding and forest management. Thus, ProCoGen has developed a multidisciplinary genomic research in conifers of economic and or ecologic importance in Europe, using high-throughput platforms for sequencing, genotyping and functional analysis, to unravel genome organization and to identify genes and gene networks controlling important ecological and economic traits, such as those related to adaptation to climatic change effects on growth, drought and cold responses and, thus, provide tree breeders with tools for precise selection. This objective has also allowed developing a cooperative research and joint efforts based on different methodological focuses and tailor-made tools to obtain a translation of basic knowledge generated to breeding. The development of high throughput genotyping tools to implement in forest tree breeding programmes allows earlier genetic evaluation, higher selection intensity, increased accuracy in genetic prediction and better monitoring of genetic diversity along generations, which are important components of sustainable genetic gain per time unit. Comparative studies integrating information from other conifer genomics initiatives with ProCoGen outcomes has allowed to explore molecular bases of different adaptive or productive responses showed by different conifer species and populations. In the last 5 years a number of initiatives have been developed to study conifer genomes of different species: Picea abies (Sweden), Pinus taeda, Pinus lambertiana, Pseudotsuga menziesii (USA), Picea glauca (Canada), Pinus sibirica and Larix sibirica, (Russia), Pinus radiata (New Zealand) and Cryptomeria japonica (Japan).

Four model tree species of high economic importance in different European regions, Pinus pinaster, Pinus sylvestris, Picea abies and Picea sitchensis, have been studied by ProCoGen for which advanced breeding programs and a number of molecular tools were available at the beginning of the project. ProCoGen adopted a coordinated and collaborative approach to unify the efforts and resources developed at the European level in conifer genomics, integrating the fragmented activities developed by
European research groups involved in the ongoing international conifer genome initiatives, and strengthening international collaboration with initiatives promoted by teams from Russia, North America, New Zealand and Japan. ProCoGen research has allowed designing and developing tools and strategies to construct core collections and trace genetic resources to maintain adapted forest ecosystems in the face of global climate change, as well as to sustain productive working forests that provide for the European wood and energy needs of the future. Beyond its scientific and economic interests, the project has reinforced the European leadership in conifer genomics, and produced a competitive advantage to the European forest-based sector in the global market.

ProCoGen specific objectives grouped according R&D and dissemination work packages are:

- Develop an integrative genomic research programme in model conifer species. To achieve this goal, we proposed the following objectives: 1) de novo genome sequencing of two European pine species, Scot's pine (Pinus sylvestris) and Maritime pine (Pinus pinaster), as well as 2) construction of extensive catalogues of their genetic variability (in the form of SNPs) through exome re-sequencing (WP1) of trees sampled to span the natural distribution range of the two species. The work package "Discovery of adaptive capacities to ensure basic and applied outcomes (WP2) proposed the study of transcriptome dynamics and regulatory switches of gene expression in order to generate new tools not only to gain significant advances in basic information on pine structural and functional genomics and epigenetics involved in regulation of growth and adaptive responses to abiotic stresses, but also to translate basic results into selection tools to assist breeding and forest management. Four are the studies proposed: 1) Transcriptome Dynamics associated to growth and adaptive responses; 2) Small non-coding RNA (sRNA) discovery and identification; 3) Understanding the transcription regulatory network associated to key processes; and 4) Epigenetic dynamics associated with growth and adaptive responses.

- Develop conifer comparative studies for understanding evolution and evaluate efficiency of information and resources transferring to other conifer species (WP3). This objective is based on the integration of genomic resources developed in ProCoGen with genomic resources developed in different model conifer species not only in Europe but also in North America. Three are the objectives proposed: 1) Analyse the level of synteny and conservation of gene order in Pinaceae using comparative mapping; 2) Evaluate both macro- and micro-synteny in Pinaceae comparing genome-based gene space in conifers using different datasets; and 3) Infer gene and sRNA function based on comparative transcriptomics, using coding and non-coding transcripts.

- Develop new bioinformatics solutions to optimally transform the large amount of data produced into useful knowledge (WP4). There are three main activities proposed to achieve this goal: 1) Genome assembly, annotation, and analysis. Genome assemblies proposed required the application of adapted tools for large and complex genomes. Genome annotation has to be designed to include both a structural and functional annotation. Structural annotation includes identification and mapping of different elements such as protein-coding genes, gene families, transposable elements and non-coding genes on the assembled genome. 2) Genome wide systems biology analyses. Transcriptome data is used to infer transcriptional regulatory networks through the use of software tools that link potential regulators and transcriptional modules. 3) Database generation that includes a fast, accurate processing of data and efficient and user-friendly data management based on state-of-the-art computational infrastructure and software architecture.

- Establish an effective molecular-assisted pre-breeding European capacity by transforming discoveries into tools to assist, identify and quantify associations in at the genomic and phenotypic levels to enable genome-assisted breeding and natural resource management (WP5). Four are the studies proposed to achieve this goal: 1) Establish statistical association between markers and phenotypes and between markers; 2) Model approaches for optimizing the integration of genomic data into practical breeding; 3) Evaluate economic efficiency of integration of genomic information into previously investigated breeding scenario; and 4) Implement the use of molecular markers for core collection definition and forest reproductive material management.

- Integrate efforts in this project with similar large-scale initiatives worldwide. Provide training in emerging technological and translational approaches disseminate the knowledge to different stakeholders and transfer technology to the forest-based sector (WP6). Dissemination, collaboration and training activities include: 1) Integration with conifer genomics initiatives worldwide, including those working on Picea abies, Picea glauca and Pinus taeda; 2) Technology transfer within and beyond
consortium; 3) Transfer of knowledge beyond consortium partners (with main focus on non-participating countries) - dissemination workshops; 4) Public relations with the scientific community, the forest-wood chain and the biotechnology industry; and 5) Construction and progressive update of the ProCoGen webpage.

Project Results:
WP1: Development of Genomic Resources

Summary of WP objectives

The introduction of next-generation sequencing technologies has opened up unprecedented possibilities of de-novo sequencing of novel plant genomes. It has also made it possible to approach the complex genomes of conifers that because of their genome size (15-30 Gbp) have previously been out of reach. The first conifer genomes have recently become available (Picea abies, Picea glauca and Pinus taeda). The goal of ProCoGen Task 1.1 was to initiate de novo sequencing and assembly of two additional conifer genomes, Scot's pine (Pinus sylvestris) and Maritime pine (Pinus pinaster). Building on results from the earliest conifer genomes, sequencing was performed using material from haploid tissues (eg. megagametophytes), to minimize problems of distinguishing closely related paralogs from polymorphic variants.

In addition, the goal for Task 1.2 was to develop catalogs of existing genetic variation (in the form of SNP databases) for the P. sylvestris and P. pinaster. The work was based on Next Generation Sequencing (NGS) technologies combined with sequence-capture arrays to target selected genome regions for re-sequencing. Re-sequencing involved pools of trees sampled from the natural distribution range to recover most of the natural variability in each species (ie single nucleotide polymorphisms, SNPs). The material utilized for re-sequencing was also sampled from available breeding material of the two species.

Description of the work done in each Task

Task 1.1: De novo sequencing of key European conifer species (P. sylvestris)

Libraries for sequencing were constructed from a single megagametophyte from an elite clone of Pinus sylvestris (W4009). After some initial quality checks, sequencing was performed using Illumina HiSeq2500 technology to genome coverage of 60x. Following sequencing, reads were trimmed based on their quality scores. Overlapping reads from the 180 bp library were aligned to form longer single-end reads and all quality filtered reads were assembled using CLC Assembly Cell on a 2TB RAM computer. Due to the low quantity of DNA used for as starting material for library constructions, some read redundancy was detected and instances where multiple reads mapped to the exact same coordinates were filtered away so that on one such read remained and was used for assembly. Mapping of the original data back to the contigs in final assembly was used to remove contaminants. Scaffolds with <1X coverage of mapped reads were removed (~58 Mbp). Likewise, scaffolds representing the chloroplast were also removed (those with >99% sequence identity over >85% of their length). Potential non-plant scaffolds (with top megablast hits to non-plant sequences in nt) were also filtered away. To validate the contiguity and completeness of the assembly we employed the same approach as described for P. abies. Briefly, a set of random full-length cDNA sequences (including 5' and 3' UTRs) from P. sitchensis were mapped to the P. sylvestris genome assembly revealing a wee coverage of the gene space of P. sylvestris despite its fragmented nature. The draft genome assembly of Scots pine is available through The Gymno-Plaza website, hosted by VIB.

Task 1.1: De novo sequencing of key European conifer species (P. pinaster)

The de novo sequencing of the Pinus pinaster genome was conducted by construction of genomic paired-end libraries from P. pinaster and DNA sequencing using the 454 GS FLX Technology. DNA was isolated from callus haploid tissue of a single individual from the Oria provenance. The genetic structure of the DNA extracted from the selected callus line was determined using molecular markers. This material was used to minimize problems of distinguishing closely related paralogs from polymorphic variants.

The haploid genomic DNA was used to prepare 454 GS FLX sequencing libraries using long PE. The output sequencing data generated true PE reads with two end tags averaging over 140 bp and separated by 3.5 kb, 6kb, 8kb or 10kb. Additional information of P. pinaster genome sequence has been generated in the frame of PineGenSeq, a Spanish project funded by the MICINN (Spain). Haploid genomic DNA from the selected callus line was also used to construct PE and MP genomic libraries for Illumina sequencing. Sort and long reads were obtained from 3 PE libraries of 10KB insert size (Roche 454/GS – FLX
sequencing), 2 PE libraries of 350 & 450 bp (Illumina - HiSeq2000 sequencing) and 5 MP libraries, 2 x 3 Kb & 3 x 5 Kb (Illumina - HiSeq2000 sequencing). Both datasets were used for P. pinaster genome shotgun sequencing and assembly. The total raw sequence coverage was estimated > 65X. The draft genome assembly of P. pinaster genome is also available through The Gymno-Plaza website, hosted by VIB.

The shotgun sequencing approach has been complemented by the sequencing of isolated bacterial artificial chromosomes (BAC clones) and the targeting and sequencing of gene-rich regions in the genome. This sequencing information covering gene coding regions have been generated in the frame of Nitrogenofor, a Spanish project funded by the MICINN (Spain). Probes for maritime pine transcripts were designed for sequence gene capture from genomic DNA. Haploid DNA extracted from the selected callus line was hybridized to 120-mer probes derived from the corresponding cDNA sequences and the captured DNA was sequenced using the FLX-Titanium platform. Gene models for maritime pine genes were established and exon/intron boundaries determined. The gene structures of relevant gene families in the maritime pine genome were analysed. The assembly was also studied for repeat content and annotation has been performed on it. The results are available via the ORCAE portal.

Task 1.2: SNP discovery through re-sequencing (P. sylvestris)
The goal of this task is to develop a large set of SNPs for Pinus sylvestris representing genetic variability of the species. This task was performed using an exome sequencing approach. Goals were initially achieved using Pinus taeda transcriptome, and once the P. sylvestris transcriptome was available, it was used as reference transcriptome. An additional goal was to develop suitable bioinformatics tools to help efficient discovery of SNPs and assessment of their quality for population genetics and other analyses.

Exome capture was initially conducted using baits designed based on the P. taeda transcriptome. This pilot work was conducted in collaboration with the laboratory of Matias Kirst (University of Florida, US). The analysis of 60 samples allowed detecting about 16,000 SNPs, also used in WP5.

The final aim was to be able to develop an exome capture system with baits developed based on the Pinus sylvestris transcriptome. The transcriptome was made available based on ProCoGen information, and the next set of 160,000 baits was developed based on the unigenes, trying to avoid repetitive sequences. Optimization of the wet lab procedures was required. A total of 60,000 out of 160,000 baits were selected based on the bioinformatics mapping against the P. taeda genome sequences, using as selecting criteria both read depth and mapping to single sites. Much of the work was performed using on haploids to detect paralogous sequences.

We aimed at detecting SNPs species wide, and included in the final experiments a geographically wide set of samples. After capture, sequencing of the samples was carried out on Illumina HiSeq 2500 instrument. The final approach to filter was as follows: The 100 base-long paired end reads were first filtered with Trimmomatic software to trim low quality reads. To remove sequence originating from outside target areas, reads were, then, aligned to the P. sylvestris transcriptome with BWA-MEM software. On-target sequence was then extracted from the transcriptome alignments, and realigned to P. taeda reference genome v. 1.01 with BWA-MEM. SNP calling was, then, performed with FreeBayes software.

Task 1.2: SNP discovery through re-sequencing (P. pinaster)
The material used for re-sequencing in P. pinaster was sampled from available breeding material, using pools of trees sampled from the natural distribution range to recover most of the natural variability (ie, single nucleotide polymorphisms, SNPs).

Transcriptome characterization of different maritime pine (Pinus pinaster) provenances has been performed using the Illumina next-generation sequencing platform.

The generated sequencing data permitted the identification of polymorphisms and the establishment of robust single nucleotide polymorphism (SNP) data sets for genotyping applications and integration of translational genomics in maritime pine breeding programmes. This catalog of SNP markers was added to those already existing in SustainpineDB, the expression database of maritime pine species. Seeds from 10 maritime pine populations from the natural range of the species were analysed. Total DNA was extracted from whole seedlings sequencing using the Illumina platform. SNP calling was performed using trimmed reads. The results are available via the Gymno-Plaza website.
WP2: Discovery of adaptative capacities to ensure basic and applied outcomes

Summary of WP objectives

The general objective of this Work Package was to develop a wide range of genomic-based approaches to unravel the functional regulation of growth and adaptive responses to abiotic stresses. These approaches have been used to identify genes and gene networks controlling important ecological and economical traits, such as those related to the control and the reduction of climatic change impact in relation to growth, drought and cold stress and, thus, providing tree breeders with tools for precise selection.

To accomplish this general objective the following specific research activities were proposed:

• Transcriptome sequencing for protein coding gene annotation, transcript expression profiling and transcriptome dynamics
• Small non-coding RNA profiling and the discovery of novel small RNA genes
• Understanding the transcription regulatory network
• Epigenetic dynamics of candidate genes associated with traits of interest

Pinus pinaster and Pinus sylvestris have been used as experimental models to analyse “Transcriptome for function-trait relationship of growth and adaptive responses”. P. pinaster, P. sylvestris and Picea abies have been used to study “Functional switches in gene regulation: Molecular plasticity associated with growth and adaptive responses”.

Different adaptive traits have been functionally dissected using species that have developed specific adaptive mechanisms: growth and drought response (P. pinaster), and cold acclimation (P. sylvestris and Picea abies). Results have been analysed not only in the frame of this proposal but also integrated with those generated in other European and national projects.

Description of the work done in each Task

SubTask 2.1.1: Transcriptome Dynamics associated to growth and adaptive responses

Metabolic processes in conifers are organized in a complex number of tissues and specialized cells. It is critical to know the transcriptome of these tissues and cells to fully understand these processes. Laser-capture microdissection (LCM) combined with massive parallel sequencing by 454 GS FLX+ (Roche) was performed to characterize cell- and tissue-specific libraries from maritime pine (Oria provenance). A map of transcriptional activity of the maritime pine tissues has been established. This map of transcriptional activity completes previous studies on the maritime pine transcriptome, and represents a valuable new resource for exploration of transcriptome dynamics in trees exhibiting contrasted phenotypes for adaptive traits. We have established the composition and function of some gene families.

Transcriptome dynamics has been studied in maritime pine trees (25-year-old) growing under natural conditions. Overall, the results show that seasonal changes in the metabolomics pattern were only affected by the needle age and acclimatization to winter, but changes in transcript profiles were mainly depending on climatic factors. Gene network analysis revealed relationships between 14 co-expressed gene modules and development and adaptation to environmental stimuli. Novel Myb transcription factors were identified as candidate regulators during needle development. In summary, the results strongly suggest that environmental changes modulate the transcriptome for fine regulation of the metabolome during development. According to this model, adaptive responses in maritime pine would influence the developmental program through the maintenance of a metabolic homeostasis.

The transcriptome dynamics and the identification of differentially expressed genes in response to drought stress have been studied using clonally propagated ramets from genotypes showing a contrasted response to drought stress. RNAs were extracted and transcriptome libraries were sequenced using 454 GS FLX+. Morphological variables, growth rates and physiological parameters regarding plant water status were also studied. Using genetic and transcriptomic approaches, sets of genes associated with drought stress in specific organs were identified and validated using qRT-PCR. Additionally, they were tested for response to dehydration and ABA treatments in seedlings. Specifically, members of gene families, representing aquaporin and dehydration-related gene families, aquaglyceroporins, LEA proteins, dehydrins, embryo-abundant proteins and RD22-like proteins, exhibited differential expression in P. pinaster clones with contrasting response to drought.

Transcriptome dynamics has also been studied in two full-sib families contrasted by their drought stress response. cDNA libraries were constructed and sequenced. (1) Genes regulated between Low vs. High MDS phenotypes: In moderate or intense water stress, specific genes were identified in the Landes*Marocco family or in the Landes*Corses family. (2) Genes regulated
between moderate vs. severe stress: A higher number of genes were identified in the Landes*Marocco family than in the Landes*Corses pedigree. Among the differentially expressed genes, a set of genes were shared between both families.

SubTask 2.1.2: Small non-coding RNA (sRNA) discovery and identification
The workplan implemented within ProCoGen SubTask 2.1.2 aimed at the discovery of pine small non-coding RNAs (sRNAs) and the identification of those involved in the regulation of adaptive responses, in growth and development, as well as the identification of their target genes. In order to achieve these goals, the work was structured in three main parts (1) plant material preparation for construction of sRNA libraries, (2) sequencing and sRNA identification, (3) validation of sRNAs and their target genes, and comparative analyses.

Following the optimization of protocols for isolation of the small noncoding RNAs, different libraries were prepared from samples of P. pinaster and P. sylvestris for high throughput sequencing using Illumina technology. Over 900 million reads were generated which were analysed using an in-house bioinformatics pipeline. Approximately 96% of the distinct obtained reads were discarded, both for the P. pinaster and P. sylvestris libraries, and this was mainly due to low abundance (<5 copies) or sequence length outside the 18-26nt range. The sRNAs present in the different libraries were identified and classified into several classes including the conserved and novel micro RNAs (miRNAs) and the trans-acting short interfering RNAs (ta-siRNAs). In P. pinaster and P. sylvestris, a relevant number of conserved miRNAs were identified, respectively, as well as a high number of putative novel miRNAs and ta-siRNAs.

Experimental validation of a set of sRNAs, of their expression profiles and corresponding target genes was then performed. Analyses of the expression of these elements in specific libraries allowed identifying a list of over 1800 sRNAs and their corresponding target genes involved in the regulation of adaptive traits, namely in drought responses. Additionally, based on their differential expression profiles, a shortlist of interactions sRNA:target gene was prepared, including the most promising candidates to be considered in future functional studies.

The results represents a valuable resource for advancing knowledge on the regulation of gene expression in a variety of pine tissues, developmental stages and environmental conditions, and a basis for implementing tools for translational genomics.

SubTask 2.2.1: Understanding the transcription regulatory network associated to key processes
The SubTask has been divided into three subprojects. The aim of Subproject 1 has been to set up a ChIP-Seq protocol and to study the functions of selected transcription factors (TFs) for further functional analyses in transgenic Pinus pinaster. A ChIP-Seq protocol has been established. The immunoprecipitated DNA prepared according this protocol has been used to prepare libraries for next-generation sequencing. A comparison between immunoprecipitated and non-immunoprecipitated DNA revealed that 80-90% of the data corresponded to genomic non-transcribed sequences. The distribution of transcribed sequences in functional categories revealed abundant reads for genes involved in photosynthesis, cell-wall formation and secondary metabolism. New antibodies against specific members of the Myb family involved in phenylalanine metabolism and wood formation, or involved in UV responses have been raised and are currently available for ChIP-Seq.

The aim of Subproject 2 was to establish transgenic embryogenic cell lines and to regenerate plants of P. pinaster for testing at least 10 TFs (over-expressed or silenced). The selected TFs are involved growth and wood formation. Embryogenic cultures were transformed with 16 gene constructs designed for constitutive overexpression or silencing of 9 TFs (MYB1,8,14,20,23, DOF5, MADBOX4, NACx, NACataf) and 3 other genes (CAD, GS1a, GS2). It was not possible to establish transgenic cell lines, which over-expressed MYB20, MYB23, MADBOX4, NACx and NACataf, suggesting that an appropriate expression of these TFs are crucial during embryo development. A collection of cryopreserved transgenic lines and/or growing plants were established for all other constructs. Embryo production and/or germination rate were modified in DOF5 RNAi lines and in GS1a and GS2 OE lines, suggesting specific roles of DOF5 and GS during germination. Low plant survival observed in one CAD RNAi line might be explained by pleiotropic, genome-wide effects of the transgene. Transcriptomic analyses suggested that strong down-regulation of a target gene could be associated with significant differences in growth of plants regenerated from MYB1 and MYB8 RNAi lines. Characterization of the wood phenotype in 6-year-old transgenic plants from CAD and MYB14 RNAi lines is underway.

The aim of Subproject 3 was to study the regulation of early embryo development in P. sylvestris. We have performed a genome-wide high-throughput transcriptome sequencing for early stages during zygotic embryogenesis, and by transcriptome
analyses we have identified candidate genes. Around 80,000 transcripts (RPKM > 0) were detected, out of which ca 35% showed homology with genes in the Arabidopsis TAIR database. In total, 6,595 TFs were detected, out of which 2,779 were differentially expressed with a fold change greater than two. A total of 2,014 TFs were up-regulated at different developmental stages of the embryos. GO enrichment analyses pointed out the importance of cell cycle, cell differentiation, cell growth and developmental processes. We have identified critical processes during early embryo development in P. sylvestris, including phase change from the morphogenic stage to the maturation stage, auxin transport, apical-basal-polarization and radial patterning.

SubTask 2.2.2: Epigenetic dynamics associated with growth and adaptive responses

The epigenetic changes at both total DNA and candidate genes levels were analysed in trees from genotypes showing a contrasted response to drought stress or epitype-inducing conditions.

In Maritime pine, the very first insight on epigenetic events during droughts trees has been achieved independently by two complementary studies: 1. Analysing clonally propagated ramets from genotypes showing a contrasted response to drought stress using control conditions. Variation in DNA methylation of hundreds of anonymous DNA regions was assessed by MSAP. In addition, variation in DNA methylation was analyzed in candidate genes involved in response to drought stress. 2. Analysis of genes associated with needle wax synthesis involved in protecting the tree from water loss. Candidate genes responding to water stress according to individual tree drought tolerance were detected and samples produced for in depth DNA methylation analysis.

Additionally, it has been shown that epitype-inducing conditions were accompanied by marked transcriptomic changes depending on culturing temperature in Norway spruce. Variations in temperature-dependent gene expression during embryo formation might be associated with chromatin modifications leading to the induction of epigenetic memory. Remarkable similarities between the gymnosperm and angiosperm epigenetic machinery were identified. A total of 736 genes putatively involved in epigenetic regulation were identified, 309 of which were differentially expressed (DEGs) responding to the epitype inducing temperature conditions. Specific genes had a putative role in the generation of small RNAs. This seems of high significance since small RNAs were also shown to vary according to tissue and epitype. Expression studies from microdissected plant tissues shows differences among close cells within a same organ (bud) according to epigenetic mechanism, meaning that high spatial resolution studies will be needed in future investigations. Additional samples in maritime pine were produced for future comparative studies of epigenetic events occurring during embryogenesis.

WP3: Comparative Genomics to understand conifer evolution and achieving effective management and breeding

Summary of WP objectives

The main objective of WP3 is to integrate genomic resources developed in ProCoGen with genomic resources developed in different model conifer species on both continents to perform comparative studies for understanding their evolution and effectively transfer information to enable the study of other conifer species. Specific complementary objectives are pursued by the three Tasks:

• Analyse the level of synteny and conservation of gene order in Pinaceae using comparative mapping. This study involves the generation of dense genetic linkage maps based on SNP markers associated with orthologous genes in targeted conifer species, including markers associated with conifer Conserved Orthologous Set (COS) genes.
• Evaluate both macro- and micro-synteny in Pinaceae comparing gene space in conifers using different datasets: (1) BACs and genome sequences from different conifer species; (2) genome annotations based on P. pinaster and P. sylvestris genomes, to be sequenced in ProCoGen, as well as data from recently sequenced genomes (P. taeda, P. glauca and P. abies), and (3) gymnosperm COS protein-coding genes.
• Infer gene and sRNA function based on comparative transcriptomics, using coding and non-coding transcripts, respectively, to improve the understanding of complex relationships between genome sequences and phenotypes and how they evolve across the species.

Description of the work done in each Task

Task 3.1: Comparative mapping of highly dense genetic maps (SNP-based maps) of representative model conifer species
The objective of Task 3.1 was the evaluation of the level of synteny and conservation of gene order in Pinaceae, the most important among the eight families of the order Coniferales (conifers), by comparative mapping based on orthologous genes. This involved the generation of dense SNP genetic maps of different Pinus and Picea species. With this aim, two alternative but complementary strategies have been followed. The first one involved the generation of a common exome capture system that would be used to genotype a set of mapping populations covering a wide representation of Pinus and Picea genus, and the use of this information for the construction of genetic maps for each mapping population and a composite map for each species. The bioinformatic comparisons between the transcriptomic and genomic information available for five Pinus and Picea species allowed to design an exome capture system with approximately 9K probes targeting conifer orthologous sequences. However, the use of this genotyping tool did not generate sufficient number of markers to enable an efficient comparative genetic mapping. Nevertheless, the second approach, based on a 9K SNP Infinium array designed for P. pinaster and on the use of high-density gene-based linkage maps for conifers for which sequence information was available, allowed the construction and comparison of dense SNP-based maps for different species, genera, and families. Moreover, this strategy allowed not only the proposed comparative genetic mapping in the Pinaceae family (Pinus and Picea species), but also to expand the analysis to include the Cupressaceae family.

The comparative analysis developed suggests the existence of contiguous ancestral regions that may have shaped the 12 chromosomes of modern Pinaceae species and the 11 chromosomes of modern Cupressaceae species through a different number of fusions. While these findings suggest an intense shuffling of orthologous linkage group blocks during the evolution of Pinaceae and Cupressaceae, a higher conservation of gene order was found within these blocks. In conclusion, the results obtained in this project revise the static view of conifer genome evolution, which was inferred essentially from comparisons of Pinaceae species. These findings support a new hypothesis that substantial chromosome rearrangements have occurred between conifer families.

Task 3.2: Comparative genome analysis

Using genomic resources generated within the ProCoGen project, most notably the genome sequences for P. pinaster and P. sylvestris, a number of comparative genomics studies have been performed within Task 3.2. The availability of both genome sequences for a selected set of conifers as well as transcriptome data sets from a larger number of species have been used to reveal both distinctive genome features and well as differences in the evolutionary rates in conifers compared to angiosperms. These analyses highlight some important differences between conifers and angiosperms such as the large size of conifer genomes, the very high occurrence of repeat elements, and the stability of macrostructure over long periods of time, and the substantially lower mutation rate are all distinctive features of conifers that are lacking in angiosperms. An interesting example is that detailed analyses of conifer gene families have shed light on the factors influencing gene family sizes in gymnosperms and how these differ from angiosperms. In conifers there is a negative association between gene family size and gene expression and codon bias, and a positive association with rates of sequence divergence, with large gene families having, on average, a lower expression level and breadth, lower codon bias, and higher rates of sequence divergence than single-copy gene families. Analyses also show that gene family expansion has occurred differentially in conifers and angiosperms, with conifers showing differential expansions in gene families primarily involved in metabolic and biosynthetic pathways, transport and response to stimulus while angiosperms show an over-representation of gene families involved in regulatory processes, signal transduction and immunity and apoptosis.

Studies also have demonstrated the presence of higher evolutionary constraints in gymnosperms than in angiosperms protein-coding genes, resulting in lower substitution rates in conifers compared to angiosperms. Conifers also harbor a higher proportion of conserved sites and a lower proportion of neutral sites than angiosperms.

As more data will accumulate we should not only be able to verify these results but also start understanding the mechanistic bases of some of these distinctive features that make conifers so unique among seed plants.

Task 3.3: Interspecific transcript analysis (based on coding and non-coding RNAs)

The main objectives for this Task were the integration of the transcriptome data available for conifers and the use of them for comparative analyses within and between the conifers. To that end we started with collecting all available transcript data both generated within the ProCoGen project as well as data from the Canadian partners (P. glauca) and already publically available
data (e.g. P. abies and P. taeda). In a first step all this data were integrated to result in a reference transcriptome for each of the species involved. Next the transcriptome dataset were subjected to an analysis in which we compared the gene content for each species to each other. For the non-protein-coding transcripts we investigated in detail the lncRNAs and miRNAs. For the analysis of the miRNAs we adapted the MIRFINDER software to take into account the specificities of the conifer miRNAs. We successfully build a reference transcriptome for each of the species ranging from approximately 40,000 to 200,000 transcripts. Since not all these transcripts code for protein sequences, we conducted an ORF finding approach for all transcripts to identify the ones that are coding for proteins. This resulted in a high quality protein dataset for each species. The proteins were subsequently subjected to a comparative analysis within the conifers. Not only the conservation of transcript within the conifers was analysed, we also compared the datasets to angiosperms to broaden our results to also conservation between angiosperms and conifers. The results and data are also available at the PLAZA platform. In PLAZA, the broader community has access to not only the transcripts and proteins but also to the obtained gene families (both interspecies and species specific ones) and synteny analysis results. The results derived from analysing the non-protein-coding fraction of the transcripts includes lists and descriptions of potential miRNAs identified in the conifers as well as a list of their potential target genes. These results are put online and are accessible to the general public. These results will serve as a basis for further research on sequence conservation within and between the conifer species and can be used to shed light on the proteins, and their function, that are specifically conserved within the conifers.

WP4: Computational Genomics

Summary of WP objectives

There are three main Tasks in WP4: genome annotation, genome wide systems biology analyses and data management. Genome annotation provides both a structural and functional annotation. Structural annotation identifies and maps protein-coding genes, transposable elements and non-coding genes on the assembled genome. Gene family identification allows downstream bioinformatics analyses, such as phylogenomics and comparative genomics. Transcriptome data generated in WP2 is used to infer transcriptional regulatory networks through the use of software tools that link potential regulators and targets. State-of-the-art computational infrastructure and software architecture ensures fast and accurate processing of data. Data is managed to facilitate the exchange among partners of the consortium and, in a second step, to become available to the general public through user-friendly and highly integrated environments. To that end, existing portals (e.g. PLAZA, PhylomeDB, ORCAE) are expanded to accommodate the data and results generated within the frame of the ProCoGen project.

Description of the work done in each Task

Task 4.1: Conifer genome sequence management: assembly, annotation and analysis

The main goal of this Task has been the generation of a draft assembly of the P. pinaster genome, along with accompanying annotations and phylogenetic studies using different types of omics data. Due to the particularly large size of P. pinaster genome (estimated to be around 25 and 30 Gbases) and the high coverage accomplished during the sequencing, the assembly constituted a computational challenge with no precedents. Starting from haploid tissue, a total estimated coverage of 65x was divided into (1) Illumina sequencing and (2) 454 GS FLX+ Roche sequencing. The assembly of genomes can be tackled through several strategies and using different programs available. In order to cover the wider range of possibilities, we used three different programs: MaSurCa, AbySS and SOAPdenovo, that we run in parallel. Alternatively, as AbySS can be parallelized, it was also used to run the assembly. A large part of the work done focused on the identification of the repeat content of conifer genomes, which is the basis of a high quality gene-annotation. Over the course of the Project this has been extensively worked on. With the custom build repeat libraries the genome assemblies, have been repeat-masked and subsequently annotated using our genome annotation pipeline based on the gene prediction tool EuGene. EuGene has been specifically trained and optimized for each of the Pinus genome assemblies in order to perform as accurately as possible for each species. For the final prediction, all available data have been taken into account, which included the transcriptome data produced by the other ProCoGen partners as well as the
phylogenomic results. The final annotation results are available for the ProCoGen partners and are also been put at the availability of the larger public via the ORCAE portal.

Finally, we have completed the integration of omics data into the PhylomeDB resource using gene predictions derived from P. pinaster sequence analyses. We integrated a total of six other conifer species Pinus taeda, Picea abies, Picea glauca, Picea sitchensis, Pinus pinaster and Pinus sylvestris, and seven other fully sequenced plant species to generate phylogenetic trees using PhylomeDB phylogenetic pipeline developed in by the CRG partner. This set of trees has been used to predict and date duplication events and to infer orthology and paralogy relationships between the species in the study.

Task 4.2: Computational System Biology
The main objective in this Task was to compile an expression matrix based on RNAseq mapping. This matrix was then the input for a system biology analysis to construct modules of transcriptional regulators.

In the course of the ProCoGen project RNAseq datasets on important traits for conifers became available: embryogenesis (P. sylvestris) and drought stress (P. pinaster). For each dataset the produced RNAseq was mapped onto the corresponding (genome derived) transcripts using the BWA software. The mapping result was subsequently transformed into RPKM values for each gene by using the samtools package and some in-house scripts.

The resulting lists of RPKM values provided the input for a differential expression analysis (R-package DEseq). These results are available as big matrices listing all the genes with logFoldChange in expression for all the comparisons described above. The above described lists of differential expressed genes were then used to build regulatory modules. To achieve this goal the LeMoNe and ENIGMA software were used. Both are softwares that apply a top-down approach to analyse the provided gene expression data. The LeMoNe software, which is capable of identifying potential regulators controlling a gene expression network, was fed with the list of DEGs and an additional list of transcription factors as identified from the genome annotation.

The result is a number of lists (and figures) listing candidate transcription factors with their corresponding network of genes they control. The ENIGMA software works solely on the DEGs as input and does not require any additional data. The output is network of genes that make modules of expressed genes controlling specific biological aspects, in our case for instance the changes in expression of genes when the tree is under drought stress.

Task 4.3: Database generation: Genome portal and crosstalk with databases of ongoing projects
This Task comprises the management of the various types of genomics data produced within the ProCoGen project in public databases, the creation of a genome portal, and the integration of the produced data with other databases and ongoing projects. This Task is central as it ensures not only the long-term maintenance of the data produced by the project but also it facilitates its exploitation by the scientific community as a whole. For this, we have been working during the whole project in setting up the different resources and their seamless integration. As the genomic data became available, these have been analysed through several annotation and comparative pipelines that serve to relate the genomic data produced within this project to those of other organisms and from other projects. In particular, we have established links with protein and genome data from other sequenced conifers and plants, and with functional databases (Gene Ontology), and protein (i.e. Uniprot) and domain (i.e. PFAM) databases. All objectives can be considered fully accomplished.

The genomic data for P. pinaster and P. sylvestris have been integrated in the ORCAE portal. This portal provides access to all the genomic and transcriptomic data produced within the framework of ProCoGen, in the form of genome browser, downloadable files, and annotations. The portal is accessible on the following URL: http://bioinformatics.psb.ugent.be/orcae. The resource includes a genome browser where the genomic data can be navigated, and download sections where the data can be retrieved in bulk. For each predicted gene, annotations, and DNA and protein sequence data is available.

Comparison of these genomes with those of other gymnosperms has been done in both a comparative genomics framework and a phylogenomics framework, which provide complementary information. Specifically the comparative framework has been implemented in the plant comparative genomics portal PLAZA, where a new subsection specifically focused on gymnosperms species has been created. This “gymno-plaza” section is accessible at the following URL: http://bioinformatics.psb.ugent.be/plaza/versions/gymno-plaza/. It contains all the results that have been produced within this project, but it also allows other people to query and investigate specific questions one might have concerning comparative analysis in the gymnosperms and, specifically, the conifers.
Finally, the genomic sequence and annotation has been integrated within the phylogenomics framework provided by PhylomeDB database (www.phylomeDB.org). This includes the reconstruction of phylomes (i.e. the complete collection of phylogenetic histories of all genes predicted to be encoded in a genome) for the genomes and transcriptomes produced by ProCoGen, as well as genomes and transcriptomes from other sequenced species. For each gene identified in the genomes, an alignment and evolutionary relationships with homologous genes in other species is available to conifer researchers, together with the functional information, and predictions of orthology and paralogy relationships. All the data is fully accessible in the form of a gene tree, alignment browser at phylomeDB (www.phylomedb.org) as well as a web API that facilitates cross linking trees from external sources. Finally, phylomeDB provides an advanced tree visualization interface based on the ETE toolkit, which integrates tree topologies, taxonomic information, domain mapping and alignment visualization in a single and interactive tree image.

WP5: Transnational Genomics: Quantitative genomics enabling breeding and resource management

Summary of WP objectives
The main objective of WP5 was to translate discoveries and developments from WP1 to WP4 into applications that enable genome-assisted breeding and forest genetic resources assessment and management. In more specific terms, WP5 aimed at developing and providing a comprehensive set of modelling tools, methodologies and guidelines to optimize the use of available genomic information for improving, managing and monitoring forest tree genetic resources. The use of genomic information was studied in praxis and demonstrated in WP5 via several case studies involving three main conifer species and covering breeding proof-of-concept studies, conservation and genetic diversity monitoring activities.

Some key activities comprised the review of existing methodologies and resources. Among the methodologies, there are those that able to identify and quantify statistical associations between genetic polymorphisms and phenotypic trait variation. One key parameter for revealing these associations is the extent and importance of linkage disequilibrium (LD) in conifer populations. The assessment of LD was also a key aim of this WP, as it can help in the dimensioning of the efforts that are required in terms of sampling and markers densities for capturing relevant variation.

The WP aimed at evaluating the feasibility of genome-wide evaluation in praxis and in theory. Genomic selection was planned in two case studies operated in the framework of breeding programs (maritime pine and Sitka spruce). Theoretical evaluation aimed a wider range of scenarios than those close to praxis, and for this, a number of simulation tools were developed and tested under diversifying conditions. The WP wanted also to add an economic perspective, and this was to be attained by the development of a cost analysis model adapted to one of the main breeding species.

Last but not least, genomic resources are also central for the assessment of existing diversity in the natural ranges of our species and to make decisions in terms of collection and conservation of populations, as well as for operational uses when it comes to track back the use of afforestation varieties. These aspects were also to be covered by case studies in two species.

Description of the work done in each Task

Task 5.1 Statistical association between markers and phenotypes and between markers (LD patterns)

Aims of the Task comprised the review of existing methodologies that able to identify and quantify statistical associations between genetic polymorphisms and phenotypic trait variation. The assessment of LD was also a key aim of this Task, as it can help in the dimensioning of the efforts that are required in terms of sampling and markers densities for capturing relevant variation. Another objective was the elaboration of prediction models based on marker information and relevant for breeding cases studies.

Review of approaches for statistical analyses of genotype-phenotype relationships: A review of currently available approaches for statistical analyses of genotype-phenotype relationships that could be considered in conifers was performed. Simultaneously, some of the case study specific issues that might affect the implementation of these approaches were studied. A review of methodologies pinpointed the adequacy of linkage disequilibrium mapping for detecting genetic associations between markers and quantitative traits, and given the available genomic resources and the dimensions of the problems to tackle conifer genome.

Compilation of phenotypic and genotypic data for the selected case studies: For maritime pine, an optimum sampling of a training population was determined to carry out a proof-of-concept experiment for a genomic selection approach, to provide
high GBLUP precision and to have a limited effective population size. The maritime pine training population was genotyped with the new 9k Illumina Infinium SNP array. For Sitka spruce, two full-sibling crosses of 500 individuals each were analyzed using RAD sequencing. RAD sequence data from a third family previously analyzed was also included in the study. Wood density measurements from all three of these families were collected at one site. For Scotch pine, a set of ten populations (30 individuals each) spanning the European range was examined based on about 400 SNPs. The effect of sample size was assessed in a large population of 500, by using a similar set of 400 SNPs. Phenotypic data of 778 Scots pine trees for growth, phenology and wood quality traits including wood density, MFA, MOE and fiber dimension traits were compiled.

Models fitting and genetic evaluation on final data: For maritime pine, GS analyses were carried out on two data sets. A first data set (G0 and G1 trees covering a large genetic basis and genotyped in a previous project) was analysed with several statistical models (genomic BLUP, Bayesian Ridge and Bayesian Lasso). The average predictive ability was between 0.4 and 0.5 depending on traits and models considered. These results were promising despite low linkage disequilibrium and low marker coverage of the genome. A second analyses was carried out with the addition of a larger number of individuals from third generation and the use of the new 9K SNP array. These second data set was designed in the ProCoGen project and optimized for GS analyses. A collaborative effort comprised the initial selection of Maritime pine G1 half sib families, whose G2 offspring was to be used for genomic prediction studies. The strategy consisted of selecting G1 families based on BLUPs of their G0 founders, while limiting the maximum founder contribution to guarantee sufficient genetic diversity. This approach was used to select families whose offspring would be subjected to paternity analysis for pedigree reconstruction. Preliminary results showed promising predictive abilities as high as 0.7. An additional GWAS analysis for height and stem straightness was carried out in a sample of 669 breeding individuals, which were genotyped for 2600 SNP markers. Markers were used to create a genome-wide relationship matrix that served to correct for spurious associations in a linear mixed model of genetic association. Preliminary results indicate that for the traits considered there is no strong evidence for association with any of the 2600 markers.

For Scots pine, the number of QTL and distribution of their effects from pedigree based QTL and population based association studies were estimated for eight quantitative traits to understand the genetic base of forest tree traits for effective MAS or genomic selection. These inferences were used for simulation of genomic selection.

For Sitka spruce, the objective was for tree breeders to work with animal breeders to investigate the latest techniques for development of GBLUP prediction models and to explore how these could be made to work within the British Sitka spruce tree breeding program. The combined phenotypic and RAD genotypic datasets have been used to construct and test GBLUP prediction models. The Sitka Spruce studies were also valuable to examine how much variance is actually captured by all the RAD sequence markers. In these analyses, the genomic relationships were calculated by different methods: (1) simple interpretation of offspring reads with no thresholds on read number; (2) using genotype probabilities conditional on parent genotypes and offspring reads; and (3) using novel relationship matrices derived for GBS conditional on both coverage and reads at each locus. Mean predictive correlations of phenotypes were around 0.2 for 6-year height and 0.25 for wood density. This was considered somehow disappointing as earlier correlations of 0.4 for 6-year height and 0.8 for 5-year budburst for just one family were obtained at one site in the EU NovelTree contract.

LD pattern analyses: Patterns of linkage disequilibrium in the genome of Pinus sylvestris have been analysed with multiple approaches. A preliminary analysis based on different populations, with limited number of samples (30) and limited number of SNPs, showed that the long range LD in the overall data set was very limited (at cM scale). Further, the individual loci level showed a rapid decay in short fragments. A larger survey based on a set of about 400 SNPs showed hardly any variation in decay of LD within populations. They also showed that combining samples results in a more rapid decay of LD. This study showed also that within the scaffold level, with pairwise distances of up to 100,000 bp, LD decayed quite rapidly. However, the scaffolds may have higher gene content than the genome on average (as gene areas facilitate assembly). Thus other areas may still have lower LD than what was shown by the very preliminary analysis. An additional study using extensive Scots pine samples from Punkaharju, Finland (500 trees) was used to examine LD-based associations of SNPs and timing of growth and wood quality and phenology traits and the results indicated that these association effects were very small.

Task 5.2 Modeling approaches for optimizing the integration of genomic data into practical breeding
The aims of the Task 5.2 were to develop modelling tools to guide the implementation of genomic information into breeding
strategies, and to test these tools in a variety of circumstances other than those already encountered in the praxis of the case studies. Simulation studies comprised two key aspects: factors affecting prediction quality and impacts on diversity. Two series of simulation studies were developed in parallel here. The first concerns an exploratory study on the potential for genomic prediction in conifer breeding by using a Monte-Carlo approach with explicit virtual genomes. One of the main objectives of this first simulation study was to evaluate the effect of genomic prediction on the selection of progenies in representations of current breeding programs in Europe. The choice of breeding programs was made in a way to encompass the range of strategies typically found in the continent for advance breeding in conifers. Results suggest that there is a modest amount of additional gains in accuracy achieved by the use of molecular markers under the conditions of the scenarios assumed: low LD, larger and diversify training populations and relatively low marker densities. Given the training size and marker density used in this study, no significant gain is observed in the largest population sizes. The results for prediction using true identity by descent (IBD) at marker positions however, suggests that increased marker density could potentially cause a gain in accuracy in this case. In all other cases, markers offer at least some improvement in accuracy, with gains increasing with smaller population size and more pedigree depth, as expected. This translates to molecular markers offering substantial gains in accuracy for breeding programs that use small breeding populations with deeper pedigrees, as the main guideline.

By using also a Monte-Carlo approach with explicit virtual genomes and real data, a second series of studies explore two additional aspects to that of the previous study. One is the benefit of imputation on the quality of predictions made with genomic evaluation, and the other is the impact of genomic selection on relatedness among selected candidates. The benefits of imputation were shown in simulated sets and in the maritime pine dataset, with sensible gains to be had in terms of quality of predictions compared to already published results (7%). It is suggested that with the development of alternative tools for genotyping at different densities with different costs, there is scope for optimization of genotyping efforts. The second aspect on the impact of genomic selection on relatedness is often neglected in the literature. We showed using simulation sets and the maritime pine dataset that for the pedigrees usually found in conifer breeding like the one available for maritime pine there is no advantage in terms of a better management of relatedness by using a genomic selection strategy. The prospects of shortening breeding cycles would exacerbate the problem. It is suggested that the use of larger segregation families and explicit constraints on relatedness could be beneficial.

Task 5.3 Evaluation of economic efficiency of integration of genomic information into previously investigated breeding scenario

The aim of the Task was to provide a sounding assessment of the costs of different breeding scenarios, including current breeding and breeding incorporating genome-wide evaluation. For this, an intermediate step was to develop a parametric simulator based on a growth and yield model. Some of the basic assumptions on breeding scenarios ware based on a review of breeding program typologies and on discussions with partners from previous Task 5.2 although the main input came from the maritime pine breeding program setting.

Costs analyses are often absent when it comes to assessing the benefits of genomic selection. The last study presents a first attempt of economic evaluation. Three alternative forward breeding strategies (polymix breeding with or without genotyping use for pedigree reconstruction, and Genomic Selection) have been compared to the backward breeding strategy currently used by the French Maritime Pine Breeding Cooperative (FMPBC). The FMPBC breeders agreed on achievable genetic gain within a single generation for the four breeding strategies to be studied in the case of three breeding targets: improving (1) tree volume alone, (2) tree straightness alone (considered as a direct proxy for the relative quantity of timber grade wood) or (3) a combination of the two parameters. The corresponding breeding activities and seed orchard establishment operations were assessed in detail in terms of time schedule and costs. Corresponding tree growth and silviculture were simulated by deterministic approaches with time scheduled costs and incomes. The increased growth speed led to very significant shortenings of the optimal cropping cycle. A full financial benefit cost analysis was performed on these bases, revealing the relative difference of the 12 strategies x targets scenarios. Sensitivity analyses were performed with wood price (per wood grade).

The cost analysis brings some important conclusions. Firstly, the costs of all breeding strategies appear to have a negligible potential impact on seedling price and costs incurred by seed orchard managers and forest nurseries. However, the shorter
duration of genomic selection could constitute a difficulty for the institutes performing the breeding work and their direct funders because their total cost is quite similar per cycle, which mean a higher cost per year. Backward selection strategies appeared clearly less performing for the landowner in any case. The different kind of forward strategies performed similarly at constant breeding target, but this is mainly due to the assumptions of identical genetic gains. Even at similar genetic gains, the faster achievement of genomic selection could lead to definitive advantage over other breeding strategies (including forward non genomic selection) considering that it would mean a higher genetic gain in a total time amount (e.g. one century, or shorter time for large forest area with reforestation each year) because a higher number of cycles would be performed. Our sensibility analysis shows that the foreseen economic performance of future varieties produced using the innovative breeding methodologies considered here would lead to a significant additional value, which could fuel the prosperity of forest owners and industry while allowing them a far higher involvement into the funding of breeding.

Task 5.4 Implementation of molecular markers for core collection definition and forest reproductive material management

SubTask 5.4.1: Assessment of genomic diversity at the natural range scale and redefinition of core collections with Scots pine as model species.

Core collections represent strategic repositories of gene pools for key species, and are often a baseline for advanced breeding. They must cover efficiently both distinct demographic units and the adaptive genetic structure at the natural range of those key species. Using molecular markers (SNPs and SSRs) and applying an exome capture and subsequent sequencing approach for high-throughput SNP genotyping, we screened several individuals across a European-wide natural distribution core collection of Scots pine. After join planning of sampling and providing of haploid or diploid DNA samples, the library preparation, exome capture and first bioinformatics and populations genetics analyses were conducted. Several individuals across a European-wide natural distribution core collection of Scots pine were screened. An initial data analysis provided information on the polymorphism levels of the different populations, overall divergence and pair-wise divergence of populations. The principal component analysis demonstrated some genetic structure. Two southern European populations, CDP (Italy) and Baza (Spain) clearly separate from the others, as expected due to their location in Southern European peninsulas. Further, at larger scale, genetic division following East-West and South-North axes was observed. Proportion of polymorphic loci varied from 0.12 to 0.28. Average gene diversity varied from 0.050 to 0.075. The Eastern cline appeared to have relatively high genetic diversity compared to the others. Pairwise FST for all populations was calculated to look at individual populations’ relationships with each other. Baza (Spain) and CDP (Italy) populations are the most divergent. In addition, some Eastern cline populations seem to have low genetic divergence among each other.

SubTask 5.4.2: Fingerprinting and traceability of genetic resources

Picea abies was selected as a good example of a species for which deployment in Europe is done through different strategies, natural regeneration, seed stands, seed orchards and clonal mixtures. At each of these levels of genetic diversity, geographical transfer and molecular fingerprints of forest reproductive material from areas in Austria was assessed using different molecular markers. A genotyping by sequencing approach has been used to genotype several thousand of SNPs in the selected material. Work on markers development and genotyping, also for fingerprinting and traceability purposes, of a large set of SNP markers has been performed in P. pinaster. A new 9K Illumina Infinium SNP array has been used to genotype natural populations, sampled across the distribution range of the species, including populations from Landes and Portugal. Population structure analyses using approximately 4K shared SNP markers between the Portuguese and French allowed identifying a set of highly differentiated SNPs that could replace the biochemical assay, currently used to test the origin of adult forest stands in Aquitaine, before seeds can be collected and then distributed for commercial purposes in France. The Bayesian clustering identified only two gene pools (K=2), with the natural and base breeding French populations forming a unique gene pool clearly and significantly separated from the Portuguese population.

SubTask 5.4.3: Polymix breeding with paternity analysis

First application of molecular markers in tree breeding could come through polymix breeding with paternity analysis (PBPA). High-throughput genotyping brings the possibility of genotyping several thousands of unpedigreed individuals for pedigree reconstruction. Breeding programs can benefit from this technology first to re-analyse polycross trials without any bias due to neglected relatedness, and second to integrate new breeding strategies as substituting full-sib mating with PBPA. Most of tree breeding programs involve polycross progeny trials. As the identity of the male parent is unknown, the genetic evaluation of
this design assumes that families are true half-sib families. This assumption can be easily broken whenever the pollen mix comprises a limited number of contributing males with probably differential reproductive success, leading to the eventuality of full- and half-sib mixtures. Identifying male parent allows re-estimating BLUP with an individual tree model taking into account the full pedigree of each tree. This has been done selecting P. pinaster as a case study, comprising up to 170 half-sib families (with known pollen mix composition) and the use of 12 SSRs markers for the reconstruction of the relatedness matrix. Based on this paternity analysis, simulations have been conducted to evaluate the feasibility to implement PBPA under more general circumstances.

Important achievements in terms of definition of core collections of populations and of sets of informative molecular markers for genotyping according to specific purposes (assessment of genomic diversity at the natural range and traceability of collectible genetic resource) have been obtained. Using already existing and newly developed molecular markers, detailed information about the distribution of genetic diversity in natural populations of P. sylvestris and P. pinaster has been obtained. Similarly, a combination of different markers allowed tracing the origin of some P. abies stands in Austria, and of P. pinaster French and Portuguese populations (using SNPs only), showing high efficiency of the selected approach for fingerprinting and traceability. New SNPs markers were efficiently used to reconstruct full pedigrees in the maritime pine breeding population and to evaluate the efficiency in terms of genetic gains of different strategies derived from polymix breeding with paternity analysis.

The large-scale analysis of genomic variation in Scots pine was initially planned to be conducted using a limited set of SNPs and microsatellite (SSR) markers. The same holds for the Norway spruce experiment. Instead, technology has developed so rapidly that now it is feasible to aim for an analysis of a significantly higher number of SNP markers using targeted sequencing. For conducting this analysis, there is the need to capture the desired parts of the Scots pine genome and then sequence these parts in a set of populations. This required first designing suitable baits for the capture, which in turn depends on the availability of transcriptomic resources. The baits were designed based on the transcriptome data delivered. The exome capture experiment samples produced the expected very good results generating a much larger data set in terms of markers. The higher number of markers allows to more precisely and efficiently dissect the role of demography and selection in shaping diversity in this species.

WP6: Integration with conifer genomics initiatives worldwide: Dissemination and Training

Summary of WP objectives

One of the aims of this Work Package was to interact and strengthen collaborations with other international conifer genomic initiatives. Further, to establish links with previous projects on conifers in Europe, and integrate outcomes to ensure a continuity of research.

Other objectives included, organisation of training workshops for transfer of knowledge within the consortium. Coordination of staff exchange programmes, to provide opportunities to young researchers to benefit from expertise of other scientists and to acquire new skills. Organisation of dissemination workshops in non-participant EU countries to bridge knowledge gaps, establish new networks and create awareness about research being conducted on ProCoGen. To show case the outcomes of ProCoGen, a final open conference at the end of the project term was also proposed.

Creating a directory of biotechnology technique providers for facilitating technological access to the research community. Interacting with various stakeholders was also one of the important objectives of this Work Package. Today the internet is quintessential medium of dissemination. For this reason setting up: project homepage, blog, facebook and twitter were essential goals of this Work Package.

Description of the work done in each Task

Task 6.1: Integration with conifer genomics initiatives worldwide, including those working on Picea abies, Picea glauca and Pinus taeda

The ProCoGen consortium members as such are involved in numerous other projects and initiatives. While each of these projects and initiatives have their own agenda and focus, it is clear that many synergies are retrieved from the personal involvement of the ProCoGen member in the aforementioned projects. Additionally, different contacts have been established among ProCoGen partners and researchers involved in other conifer
genome analysis & application initiatives, to improve the scope of their studies. Specific contacts and collaborations have been established to integrate efforts in a broad sense. Thus, ProCoGen coordinators and partners have actively participated in the Conifer Genome Sequencing Summits held in Sweden (2013), Canada (2014) and Sweden (2015) as well as in the organization of the last Conifer Genome Sequencing Summit, held in Sweden in 2015. The Conifer Genome Sequencing Summit is an annual meeting open to all members of the reduced forest genomics community. The researchers involved have a unique opportunity to update each other on the most advanced approaches in conifer genome sequencing, assembly and annotation, as well as on the interpretation and practical application of the information obtained from conifer genomes. Also ProCoGen has maintained a fluid communication with other European initiatives such as COST Actions as well as previous (NovelTree, EvolTree) and ongoing (Trees4Future) projects, participating in their events and meetings. It is important to mention the ProCoGen final open conference held in Orléans, France, last November 30th - December 4th, 2015. This meeting promote collaboration among ongoing initiatives on conifer genomics (listed below) in a broad sense. Discussions on future joint research efforts subjects on specific topics were addressed.

ProCoGen has actively participated in conferences organised by other international conifer initiatives like the Alpine Forest Genomics Network (AForGeN) meetings held in Italy in 2012 and Austria in 2013, and the IUFRO “Genetics and Conservation of White Pine Species” in the USA in 2014. These meetings continue to offer interactive platforms for ProCoGen with ongoing conifer genomic initiatives worldwide. ProCoGen has promoted and organized round tables, workshops and discussions about collaborative work on conifer genomics and the application of resulting information in the frame of meetings with a broader spectrum of the Plant Genomics community, such as at the IUFRO Tree Biotechnology Conferences held in Brazil in 2011, the USA in 2013 and Italy in 2014, as well as the PAG meetings (USA) and the IUFRO conference on Integrative Vegetative Propagation, Biotechnologies and Genetic Improvement for Tree Production and Sustainable Forest Management held in Czech Republic in 2012.

ProCoGen scope, goals and advances have been actively advertised in different scientific meetings and dissemination events for stakeholders (see also Task 6.3: Transfer of knowledge beyond consortium partners - Dissemination workshops and list of events) in order to promote collaborations. An example is the transfer of information to the European Technology Platform for the Forest-based Sector (FTP), that has lead us to prepare, at their request, a document in 2016, summarizing ProCoGen achievements of interest and those ones that can be applied to the sector.

In the frame of ProCoGen, different collaborations have been established with other ongoing initiatives on conifer genomics worldwide, such as those ones in:

- **Picea abies**, in the frame of the Norway spruce genome project (ConGenie), coordinated by a Swedish consortium. One of the teams coordinating this project (P6) is beneficiaries of the ProCoGen, facilitating the integration of activities.
- **Pinus taeda**, in the frame of the PineRefSeq, coordinated by UC Davis (USA). Dr. Jill Wegrzyn has actively been involved in the collaboration with ProCoGen, providing not only updated information about P. taeda sequence drafts but also information about progress on Pinus labmertiana and Pseudotsuga menziesii genome sequencing.
- **Picea glauca**, in the frame of the SMarTForests project and SIRI Canadian project “Génomes, adaptation climatique et valorisation des conifères forestiers”, coordinated by Canada. Teams coordinating these projects are beneficiaries of the ProCoGen, making easier the integration of activities and development of joint activities.
- **Pinus radiata**, coordinated by New Zealand. Research teams coordinating this proposal have been collaborating in comparative analysis and actively participating in the ProCoGen open final meeting to discuss future joint actions.
- **Cryptometia japonica** (coordinated by Japanese research teams) have been also involved in the comparative mapping of conifer species and participated in the ProCoGen open final meeting.
- **Pinus sibirica** and Larix sibirica, coordinated by Russian research teams, that have also participated in the discussions developed in the frame of the ProCoGen open final meeting.

The project has benefited from collaborations with other initiatives in terms of problem shooting, data exchange for setting probes, genome and transcriptome comparisons, data interpretation, etc. The list of publications shows some of the fruitful collaborations.

**Task 6.2: Technology Transfer within and beyond consortium**

**SubTask 6.2.1: Training Workshops (TWS)**
During the entire course of this project four TWS were held to introduce new findings and techniques within the consortium:

TWS1 “Genome Sequencing and Gene Discovery”. January 30th - February 1st, 2013, Umeå (Sweden)


TWS4 “Bioinformatic and trees”. March 11th -12th 2015, Vienna (Austria)

TWS3 “Practicalities of markers and genome-assisted selection”. December 3rd 2015, Orléans (France)

SubTask 6.2.2: Staff Exchange Programme (SEP)

A total of twelve SEPs were offered at the beginning of the project but only half have been utilised. Nevertheless, it can be gathered from the reports submitted by the participants that young students and early stage researchers benefited from this opportunity.

Task 6.3: Transfer of knowledge beyond consortium partners (with main focus on non-participating countries) - dissemination workshops

Four dissemination workshops were held during the course of this project. It was an effort to transfer the know-how generated in this project to researchers in a broad sense and stakeholders. Moreover, DWS also proved to be a networking platform. All the workshops were received with great interest, some of the participants travelled to participate in more than one event on their own expenses.

DWS1 “Conifer sequencing: basic concepts in conifer genomics”. November 5th – 6th 2013, Riga (Latvia)

DWS2 “Genomics and the conservation of conifer genetic resources”. September 1st – 3rd 2014, Szombathely (Hungary)

DWS3 “From our labs to your forests” December 3rd 2015, Orléans (France)

DWS4 “Transfer of genomic tools to breeding programs”. December 4th 2015, Orléans (France)

Task 6.4: Public relations with the scientific community, the forest-wood chain and the biotechnology industry

SubTask 6.4.1: PR Material

The PR materials like project leaflets, general poster and presentations etc. containing general information about the Project have been available for free downloading. Apart from this, information about workshops and presentations of the workshop, and publications in which ProCoGen funding has been acknowledged has also been posted on the homepage. Roll-up displays have also been used at various events as PR material.

SubTask 6.4.2: Interaction with stakeholders

Scientific community:

ProCoGen support has been already acknowledged in three book chapters, twenty five articles in peer reviewed journals and one IUFRO proceeding. Two articles and two book chapters are in press and two have been submitted. Fifteen manuscripts are in preparation and nine are being planned to be published. The ProCoGen workshops have also helped transfer of knowledge to the young scientific community. The homepage and blog have also made workshop presentations and publications of the project available on the internet.

ProCoGen Final Open Conference was an event that showcased the project outcomes. It was also an important event for networking and strengthening collaborations with other international initiative. Ten posters and 17 oral presentations were made by project members highlighting their research findings. 18 oral presentations were made by invited external speakers mostly from other conifer genomic initiatives.

The ProCoGen partners have presented/participated in 172 national and international events, out of these oral presentations have been made at 92 events and posters have been presented at 73 events. Participation in meeting organization completes the number of events. Out of all the presentations ten oral and 25 posters were presented at various IUFRO meetings. Other important events included PAG (Plant and Animal Genome) conference and Genome Summits.

ProCoGen is linked to all the present major conifer genomic networks around the world now. Some project members have also joined and are actively involved in the Alpine Forest Genomics network, which is still at an initial stage.

Forest based sector:

Project information was posted on the FTP homepage. A dissemination committee was set up in the early stages of the project for promoting ProCoGen at national levels.
Austria: Project folders were distributed at various FTP, EFI and COST events. The FTP, IFSA, IUFRO, EFI were informed about various project events through mailings. Events were also posted on the CORDIS wire network. ProCoGen was also promoted in a forestry event in Sardinia and in a lecture series in Mexico by Austrian partners. Waldwissen network proved to be an important link referring to the project website and blog. Project information was circulated through BFW mailing list to over 5,500 addresses, which included the forest based sector in Austria and neighbouring countries. FTP Austria was updated about the project through newsletters and folders.

The Netherlands: EUCARPIA, EUFORGEN and Seed and Plant Committee of Industrial Board for Forest and Nature were informed by project members in the Netherlands.

Norway & Sweden: Web communication, press articles and radio interviews were carried out in Norway. ProCoGen was promoted along with the Norway spruce project in Sweden. Presentation was made at the Nordic Primary Industry meeting on climate change by Finish partners.

Portugal: Along with web news releases, a presentation was also made for the AIFF (centre for forest industry) in Portugal. ProCoGen was promoted along with the Norway spruce project in Sweden. Presentation was made at the Nordic Primary Industry meeting on climate change by Finish partners.

Italy: Italian partners were also intensively involved in the organisation of IUFRO Tree Biotechnology meet 2015 in Italy. This conference was also attended by members of EFSA and forest based industry representatives. The second dissemination workshop included topics for forestry sector. This was organised together with the Italian partners. It was attended by representative of the Austrian Federal Forests ÖBf) and by director of one of the largest forest based company (LIECO) in Austria.

Spain: The following organizations were informed about the project through presentation in Spanish and folders: Aspapel (Spanish Association of Pulp and Paper Manufacturers), Confemadera (Confederation of Wood Industry), Aidima (Technology Institute Furniture Wood & Packaging), Accion Forestal (Forest Action), Altim (Association of Technological Investigation for Wood Industry), Anfia (National Association of Board Manufacturers), Asemfo (National Association of Forest Companies), AEMCEM (Association of Castile Wood), COSE (Confederation of Forest Owners of Spain), Colegio de Ingenieros de Montes (College of Forestry Engineers), Invegen (Association for the Promotion of R & D Technology in Plant Genomics). Spanish Forest Based Sector was also updated about ProCoGen. Apart from this press release and interviews were also conducted in Spain.

France: The final open conference was organised and hosted by INRA. Project information was published on the XYLOFOREST website.

UK: Project information and events were posted on the homepage of the Forestry Commission UK.

SubTask 6.4.3: Making biotechnological tools available to all: within and beyond the consortium
A directory of biotech service providers was posted on the project website to enable people to access some of the technology service providers without endorsing for any of the companies.

SubTask 6.4.4: Technology transfer cooperate group
The ProCoGen blog was set up in 2013 to post some interesting findings mainly related to conifers from ProCoGen and other international initiatives. The blog attracted viewers from over 70 nations across the globe within the time span of this project. So far over 3,000 viewings have been recorded for 30 posts on this blog. This form of media has successfully spread the word about ProCoGen across the world.

Task 6.5: Project Website
The ProCoGen website statistic records after “facelifting” of the homepage in Sep. 2013 show over 11,000 viewings of the website across hundred countries. The homepage was intended to provide information about the project on a global scale and it has successfully accomplished this Task.

WP7: Management
Summary of WP objectives
The project management provides the means to ensure the correct fulfilment of the project scientific and dissemination objectives, activities, tasks or work packages according with the DoW and with the EC rules and procedures, including quality standards. It also fosters the legal involvement of the participants and participating organizations, enables legal decision making and elaborates the contractual and legal ending of the project. It will steer the project to address all unexpected
situations, be these of scientific, technological, legal or financial nature. The management plan designed in this project was based on three objectives:

• Administrative and financial coordination according to the work plan and within the financial constraints
• Survey of activities to evaluate the correct fulfilment of the planned objectives, optimizing the use of fast evolving technologies and the development of new strategies, to adapt activities, task or working packages, to ensure an optimal use of resources.
• Efficiency of the organization setup to support the project, with special attention paid to financial accountancy, logistics, communication, coordination issues, quality and conformity to EC rules and procedures.

Description of the work done in each Task
Task 7.1: Administrative and financial management
Based on the description of WPs activities, it can be concluded that the project evolution has been appropriate, including project management. The unexpected delay in specific activities during the 3rd and 4th year was overcome, and the objectives of the project have been successfully achieved in due time. Important efforts have been successfully carried out during the lifetime of the project for:

• Implementing sophisticated mechanisms to communicate among partners, aimed to encourage a faster flow of information to complete the experimental work.
• A comprehensive plan for dissemination, focusing on national/regional targets and on the external scientific community, and the Dissemination Workshops.
• Improvements in the project reporting tools.
• Encourage partners to activate and participate in the project outcomes.

Tools and devices for communication and information
• The mailing lists have been the main system to communicate with and between the partners. The project Secretariat has been in charge to update eventually a set of different mailing lists, depending on the target achieved: PIs, Financial staff of the beneficiary institutions, legal staff, members of the General Assembly, EAB, WP Leaders, Task Leaders, etc. To that end, a List of personnel from each institution has been updated to monitor the staff directly involved in the project.
• Doodle polling system has been very useful to set dates for meetings, especially the EB meetings as well as WP meetings. These meetings were also held via Skype. The intranet has become the essential workplace communication where to upload relevant documents, such as Deliverables, legal project documents, Dissemination tools, etc. With the aim to classify all these documents, a set of rules for naming and saving files have been implemented and described in the ProCoGen Rules for Project Reporting and the Rules for Deliverables.

Project progress and meetings of decision-making bodies
Project progress has been monitored by the organization of several meetings: Annual meetings General Assembly, Executive Board, and Scientific and Technical Advisory Board meetings. The project scientific progress and the dissemination activities were presented by the WP and Task Leaders during the Annual Meetings. Four Annual Meetings were held. The 4th Annual meeting was open to external attendees. The Executive Board (EB) has become the main decision-making body for several scientific and management decisions. The involvement and implication of the EB has been necessary to implement several choices. In addition, it has been the most useful tool to monitor and update the project progress and, if required, assess to take decisions that may have avoided any harmful impact on the project. Thirteen EB meetings were held that, along with the annual meetings, EB met approximately every two months during the lifetime of the project. The General Assembly has been the final decision-making body of the project. Four GA meetings were held. General Assembly approved all the relevant scientific, legal and administrative proposals made by the EB and the coordination. The External Advisory Board actively contributed and collaborated in all activities they were required for. The outstanding support given by the External Advisory Board was a key element to find solutions to unexpected situations which could have affected the regular project progress.

Deliverables and Milestones. General overview
Since the beginning of activities, the project coordination has prepared and updated a set of templates for the preparation and submission of each type of Deliverable, as well as several rules to harmonize edition. These rules can be found in the ProCoGen Rules for Deliverables available in the project intranet. A calendar of information request about the progress of
deliverables has also been used. Deliverables and Milestones have been achieved.

Risk Assessment
The Risk Assessment Plan has been regularly updated in relation with the Project report. Procedures to rate both damage and probability for any prospective risk, with the aim to obtain a better understanding and to optimize the information provided for each WP, was implemented.

Task 7.2: Reporting to the European Commission

Project Reporting
Coordinator prepared the ProCoGen Guidelines for Project Reporting available in the project intranet. This document details the main rules and conditions for both scientific and financial reporting. A Power Point system was used to become the guidelines in a more interactive tool instead of using the PDF format. The information included in the guidelines focus on how to use the scientific reporting templates, explanations to use the participant portal and the Person Month Excel file. Three periodic reports were prepared, as stated in the DoW.

Scientific reporting
The project coordination had established a set of templates for technical reporting, classified per project role, WP and beneficiary, and personalized for each case. The aim was to request only the required information for each case, as well as to simplify the work for the scientific staff. These templates were also structured under a clear schedule of deadlines, and based on a bottom-up hierarchy for their submission. These templates were continuously improved to simplify the procedure without missing the quality of the information provided for the technical reporting.

Financial reporting
The financial reporting was sent to the European Commission in each periodic report. This report included the financial information for the corresponding period. Once the Commission approved the financial reporting, Coordinator transferred the corresponding funding to the partners within 10 days from reception of the funding from the European Commission. Partners confirmed the reception of their corresponding funding.

Communication with the European Commission
The Coordinator has established a regular communication with the Project Officer to inform him about the project progress at real time. The Coordinator informed the Project Officer about scientific, dissemination or management activities regarding the project every three months as average. Minutes of the EB meetings, as well as other meetings, such as the Annual Meeting, General Assembly Meetings or External Advisory Board Meetings, were sent in this flow of communication. Information about Deliverables or Milestones achieved as well as scientific, technological, legal or financial issues were also requested and provided.

Potential Impact:

Potential impact
ProCoGen has contributed towards the expected impact of the call by developing an integrative research programme on conifer species. ProCoGen has combined the use of the most appropriate high-throughput (HT) technologies with advanced bioinformatics to generate new knowledge to improve forest productivity through targeted breeding programs based on better adapted material to regional climatic threats, forest stewardship in response to environmental change as well as conservation efforts. ProCoGen has been strongly committed towards advancement of knowledge in conifer genome structure and function, including regulation mechanisms of genes controlling economically and ecologically important traits, as well as technology transfer. A great amount of effort and resources have been directed to technology transfer. Enormous efforts have been applied to translate basic genomics results into practical applications in order to enable genome-assisted breeding and resource management in conifers in a broad sense.

As a general view, the information generated in the frame of this project has allowed unravelling a) the genome structure of several model conifer species of interest in Europe, identifying b) the molecular bases underlying their plasticity and adaptation c) to compare gene and gene networks governing the response of conifers to climate change and infer the information that can be transferred to other conifer species of high importance in European forest economy and (d) identify and quantify associations between genomic and phenotypic variants to design a pre-breeding and selection tools that enable
ProCoGen has specifically impacted on:

- Reinforcing European Research competitiveness
- Integrating European research in the international network for conifer genome analysis
- Preparing society to better face climate change
- Improving tree breeding and sustainable forest management

Reinforcing European Research competitiveness. Integrating European research in the international network for conifer genome analysis

ProCoGen has built upon previous networking efforts from European initiatives on forest ecosystems and integrated previous fragmented activities developed by European research groups to build an international conifer genomics network made of representatives of relevant projects and initiatives in order to ensure a fluid communication and correct canalization of information to optimally exchange results and ideas. For this purpose, ProCoGen linked up with the previous initiatives, such as the EC-funded EVOLTREE, NOVELTREE, Trees4Future, or FoResTTrac projects, and other projects on conifer genetics, genomics, breeding, and forest management, with a view to ensuring wide and continued data availability. As aforementioned, ProCoGen integrated efforts in this project with similar large-scale initiatives in North America, to use synergies, to create added value and to ensure proper dissemination of project results in the scientific community worldwide. This enabling integrative project has greatly contributed to strongly reinforce the forefront position of the European research on conifer genomics and bioinformatics. Pioneering developments have been achieved in the frame of ProCoGen in different fields of basic research, such as structural genomics of large and complex plant genomes, providing new reference pine genomes, genome variability based on SNP discovery, and a vast catalogue of genes involved in adaptive responses. Functional analysis of these long-lived species provided information not only about the genetic control of adaptive and productive traits, but also information about the potential role of epigenetic mechanisms in the molecular control of the phenotypic plasticity, which is a crucial issue when boosting resilience to climate change. Comparative analysis allowed studying macro and micro-synteny of conifer species at different levels. The resulting information is accelerating the use of genomics tools in other conifer species. Impacts in translational genomics, based on multiple-goal breeding modelling and case studies, are crucial to select basic results to apply in molecular-assisted breeding and conifer resources management. Pre-breeding tools (i.e. genotyping arrays for pedigree reconstruction, etc) and exome capture and Genotyping By Sequencing (GBS) approaches for high-throughput SNP genotyping have been used to assess genomic diversity at the natural range scale of the species and redefine core collections. Impact in bioinformatics is mainly associated with new developments required to analyse such large and complex genomes. The development of this project has allowed improving de novo assembly purely based on NGS data. State-of-the-art computational infrastructure and software architecture has ensured fast and accurate processing of data. Data will be provided to the scientific community and to the general public through user-friendly and highly integrated environments. To that end, existing portals (eg. PLAZA, PhylomeDB, ORCAE) has been expanded and adapted to accommodate the data and results generated by ProCoGen project. Of course, the knowledge gained in both collecting, processing, analysing and using next-generation sequence data, as well as genome assembly have proven invaluable for future projects analysing complex genomes. Multiple conifer species studies performed in this project has been an ideal test case to develop a software package capable of managing the high-level NGS information, which then can be applied to a multitude of genome analysis of other species.

Improving tree breeding and sustainable forest management. Preparing society to better face climate change

Forest trees are among all species under some degree of domestication the ones most likely to benefit from the use of genome-assisted breeding and marker-based monitoring for conservation purposes. One of the most evident benefits is the expected reduction in the long generation intervals that are to be found in most conifer breeding programs. There has been a substantial development in conifer genomics, and this effort has to be translated into applications that enable genome-
assisted breeding and forest genetic resources assessment and management. This latter effort has been one of the main aims of ProCoGen. Some key activities comprised the review of existing methodologies and the assessment of key population parameters. Among the methodologies, linkage disequilibrium mapping was identified as one of the most appropriate approaches for detecting genetic associations between markers and quantitative traits in typical conifer populations, as illustrated by Genome Wide Association Analysis (GWAS) on a maritime pine population. One of the key parameters playing a fundamental role in our capacity to reveal these associations is the extent and importance of linkage disequilibrium (LD) in conifer populations. We have provided LD patterns at different scales and for two of the main conifer species (Scots pine and maritime pine). These findings help to gauge the efforts that are required in terms of sampling and markers densities for capturing relevant variation in our conifer populations. The project comprised proof-of-concept studies on the feasibility of genomic selection for two of the main conifer species in Europe for the first time. The development of the new 9K SNP array for maritime pine in the course of the project has considerably raised marker densities and, thus, opened up the possibility to improve substantially the predictive ability. Sitka spruce proof-of-concept has brought a complementary approach to raise marker densities by adopting RAD-sequencing. Predictive ability has to be further studied, but there are strong indications on ways to improve the quality of genotyping outcome by the use of imputation. Proof-of-concept studies are somehow limited to the experimental conditions, often far from operational situations. In ProCoGen, there has been a substantial effort in developing simulation tools and in applying them to a series of typological breeding scenarios to broaden the assessment of benefits and effects associated to the implementation of genomic selection. One of the simulation tools was based on original ‘gene-dropping’ techniques. This tool was used to mimic existing pedigrees, while considering current conditions in the maritime pine breeding population: low LD, larger and diversify training populations and relatively low marker densities. The results suggested that increased marker density beyond what is typically available could potentially cause a competitive gain in accuracy. In current conditions, however, genomic selection would be beneficial only with smaller population sizes and deep pedigrees. Another series of simulation studies tackled the assessment of imputation and the impacts of genomic selection of diversity. Imputation techniques were found to be highly efficient and beneficial in the conditions of maritime pine pedigrees, with a resulting gain in terms of quality of predictions compared to already published results (7%). Also, concerning the management of diversity under genomic selection, it is shown that for the pedigrees usually found in maritime pine there is no real advantage compared to pedigree-based evaluation. It is suggested that the use of larger segregation families and explicit constraints on relatedness could be beneficial. The project wanted also to add an economic perspective, and this was to be attained by the development of a cost analysis model. Results suggest that despite heavy investments, genomic selection is a competitive option at time-wise basis. Last but not least, genomic resources are also central for the assessment of existing diversity in the natural ranges of our species and to make decisions in terms of collection and conservation of populations, as well as for operational uses when it comes to track back the contribution of varieties. A complete profiling of diversity patterns across the Scots pine range by the use of exome capture and genotyping-by-sequencing has allowed the identification of key constituents of a core collection, with some populations at the fringe of the range being of special signification. It has also been demonstrated that low density marker sets can be of great help for management of breeding populations, with an example in maritime pine of a posteriori pedigree reconstruction that substantially simplified the handling of crossings during breeding while improving the evaluation of candidates. The project has provided many guidelines for a rational use of genomic information into breeding and conservation activities, and that covering a wide range of initial requirements: from comparatively few resources in the traceability and pre-breeding activities, to denser resources in the case of genomic selection and association studies.

Main dissemination activities
Additionally, ProCoGen has established programmes to ensure an active and efficient training and dissemination at different levels: a) Transfer of knowledge within the consortium, which includes organization of Training Workshops and a Staff Exchange Program, to achieve scientific excellence during the project and ensure an outstanding formation of researchers (both students and seniors); b) transfer of knowledge to the scientific community through publications in scientific journals and communication in scientific meetings; c) activities to integrate European and North American research in Forest Genomics; d) transfer of knowledge beyond the consortium, which includes organization of Dissemination Workshops and transfer activities to the sector, to ensure that relevant contribution of genomics in genome-assisted breeding and forestry sustainable
management is made available not only to European policy makers and stakeholders but to the international community at large facing a global challenge such as climate change; e) Newsletters describing the project progress; f) PR materials. Details about the aforementioned activities:

a) Transfer of knowledge within the consortium:
Training Workshops (TWS): Four TWS were organised covering different research fields and expertise to improve the consortium skills in cutting-edge research and technologies. See description in WP6 at the section “A description of the main Science and Technology results”.
Staff Exchange Program: This tool was designed to ensure effective training of young scientist in cutting-edge technologies in the pioneering groups within the consortium. A gender policy was developed using it as one of the main selection criteria (75% eligible candidates were females).

b) Transfer of knowledge to the scientific community:
Different publications have been already published during the lifetime of the project (see description in WP6 at the section “A description of the main Science and Technology results”). Considering that most activities have been accomplished at the end of the project, additional publications are in preparation or outcomes will be prepared to be published.

c) Activities to integrate European and North American research in Forest Genomics:
ProCoGen is an enabling integrative project that brings together not only European groups involved in the ongoing international Conifer Genome Initiative (CGI), but also North American groups actively participating in this initiative. Thus, ProCoGen has allowed articulating and strengthening this transatlantic collaboration with North American initiatives, as well as has served as a tool to facilitate collaboration with other worldwide conifer genome initiatives from Russia focusing on Pinus sibirica and Larix sibirica, from New Zealand focusing on Pinus radiata, from Japan focusing on Cryptomeria japonica. The multi-species comparative study at the core of ProCoGen strategy has produced useful information for the entire international community of scientists and users.

d) Transfer of knowledge beyond the consortium
Dissemination Workshops (DWS): Four DWS were held during the course of this project. It was an effort to transfer the know-how generated in this project to researchers in a broad sense and stakeholders. Moreover, DWS also proved to be a networking platform. See description in WP6 at the section “A description of the main Science and Technology results”.
Transfer activities to the sector: A dissemination committee was set up in the early stage of the project for promoting ProCoGen at national levels (see description in WP6 at the section “A description of the main Science and Technology results”). Different initiatives at national and regional level were organized in order to properly and progressively disseminate ProCoGen outcomes to the sector. Additionally, project information was posted on the Forest-based Sector Technology Platform (FTP) homepage. A final article has been requested by FTP with the final outcomes of the project which was delivered in early February 2016.

e) Newsletters: Information about the ProCoGen progress has been progressively made available in the form of Newsletters, which can be downloaded from the project website (http://www.procogen.eu/).

f) PR materials: project leaflets, general poster and oral presentations etc. containing general information about the Project have been available for free downloading from the project website (http://www.procogen.eu/). Outcomes of the project will be delivered following the rules described in the Consortium Agreement. Databases with genomics information generated in the frame of the project are public and available in the websites described in the section “A description of the main Science and Technology results”.

Exploitation of results
ProCoGen is an enabling integrative project that developed basic information about conifer genomes, and translated these basic genomics results into practical applications in order to enable genome-assisted breeding and resource management in conifers in a broad sense. Databases with genomics information generated in the frame of the project are public and available in the websites described in the section “A description of the main Science and Technology results”, fostering open access according with the Consortium Agreement rules. Outcomes of the project will be delivered following the rules described in the Consortium Agreement. Because of this “enabling” nature, no patents, trademarks, registered designs or utility models were initially planned or are expected.
List of Websites:
ProCoGen website: http://www.procogen.eu/

Project Coordinators:

Prof. Carmen Diaz-Sala
Department of Life Sciences
University of Alcalá
Alcalá de Henares
Madrid
Spain
Email: carmen.diazsala@uah.es
Web: http://www.uah.es

Dr. Maria-Teresa Cervera
INIA-CIFOR
Carretera de La Coruña Km7.5
Madrid
Spain
Email: cervera@inia.es
Web: http://www.inia.es

Related information

| Result In Brief | More sustainable pine forests in the making |

Reported by

UNIVERSIDAD DE ALCALA
Spain

Subjects

Scientific Research

Last updated on 2016-10-11
Retrieved on 2019-06-27

© European Union, 2019