Plant Mol Biol (2012) 80:555–569
DOI 10.1007/s11103-012-9961-7

REVIEW

Towards decoding the conifer giga-genome

John Mackay · Jeffrey F. D. Dean · Christophe Plomion · Daniel G. Peterson · Francisco M. Cánovas · Nathalie Pavy · Pär K. Ingvarsson · Outi Savolainen · M. Ángeles Guevara · Silvia Fluch · Barbara Vinceti · Dolores Abarca · Carmen Díaz-Sala · María-Teresa Cervera

Received: 13 April 2012 / Accepted: 24 August 2012 / Published online: 9 September 2012

Abstract Several new initiatives have been launched recently to sequence conifer genomes including pines, spruces and Douglas-fir. Owing to the very large genome sizes ranging from 18 to 35 gigabases, sequencing even a single conifer genome had been considered unattainable until the recent throughput increases and cost reductions afforded by next generation sequencers. The purpose of this review is to describe the context for these new initiatives. A knowledge foundation has been acquired in several conifers of commercial and ecological interest through large-scale cDNA analyses, construction of genetic maps and gene mapping studies aiming to link phenotype and genotype. Exploratory sequencing in pines and spruces have pointed out some of the unique properties of these giga-genomes and suggested strategies that may be needed to extract value from their sequencing. The hope is that recent and pending developments in sequencing technology will contribute to rapidly filling the knowledge vacuum surrounding their structure, contents and evolution. Researchers are also making plans to use comparative analyses that will help to turn the data into a valuable resource for enhancing and protecting the world's conifer forests.

J. Mackay · N. Pavy
Center for Forest Research, Institute for Integrative and Systems Biology, Université Laval, Québec, Québec G1V 0A6, Canada

J. F. D. Dean
Department of Biochemistry and Molecular Biology, Warnell School of Forestry and Natural Resources, Davison Life Sciences, University of Georgia, Complex B314B, Athens, GA 30602, USA

C. Plomion
UMR BIOGECO, INRA, 69 Route d'Arcachon, 33612 Cestas, France

D. G. Peterson
Institute for Genomics, Biocomputing and Biotechnology (IGBB), Mississippi State University, 2 Research Blvd., Box 9627, Mississippi State, MS 39762, USA

F. M. Cánovas
Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus Teatinos, 29071 Málaga, Spain

P. K. Ingvarsson
Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, Linneaus väg 6, 901 87 Umeå, Sweden

O. Savolainen
Department of Biology and Biocenter Oulu, University of Oulu, 90014 Oulu, Finland

M. Á. Guevara · M.-T. Cervera (✉)
Genomics and Forest Ecology, Forest Research Centre INIA-CIFOR, National Research Institute for Agricultural and Food Technology (INIA), Carretera de La Coruña km 7,5, 28040 Madrid, Spain
e-mail: cervera@inia.es

S. Fluch
Health and Environment Department, Platform for Integrated Clone Management (PICME), AIT Austrian Institute of Technology GmbH, Konrad-Lorenz-Straße 24, 3430 Tulln, Austria

B. Vinceti
Bioversity International, Via dei Tre Denari 472/a, 00057 Maccarese (Fiumicino) Rome, Italy

D. Abarca · C. Díaz-Sala
Department of Plant Biology, University of Alcalá, Ctra. Madrid-Barcelona Km. 33,600, 28871 Alcalá de Henares, Madrid, Spain

Keywords Conifers · Giga-genome sequencing · Functional genomics · Comparative genomics · Integrative studies

Introduction

The conifers are classified with the seed plants, which include four living groups of gymnosperms, i.e. cycads (Cycadales), *Ginkgo* (Ginkgoales), gnetophytes (Gnetales), and conifers (Pinales), and the extant flowering plants or angiosperms (Magnoliophyta) (Raven et al. 2005; Gernandt et al. 2011). While angiosperms underwent tremendous adaptive radiation, reaching some 250,000 species (Kenrick 1999), the extant gymnosperms number fewer than 1,000 species (Farjon 2008). Despite the global success of angiosperms, one division of the gymnosperms, the conifers still dominate many of the world's temperate and boreal forest ecosystems where they play a major role in global carbon cycles, are widely used in reforestation programs, and are critical to preventing soil erosion, among other functions.

Fully sequenced plant genomes are rapidly growing in number but, to date, they do not include a representative from the gymnosperm lineage. Until very recently, sequencing of conifer genomes had not been attempted owing to their extremely large sizes. Genome sizes ranging from 18,000 to over 35,000 Mbp have been estimated for economically and ecologically important conifers, such as many pines and spruces (Murray et al. 2004), making them on average $> 200\times$ the size of the *Arabidopsis thaliana* genome, and close to $24\times$, $10\times$ and $7\times$ the genomes of rice, maize and human, respectively (Fig. 1). It is now well known that large genomes among angiosperms are the consequence of multiple genome duplications and triplications, polyploidization events complemented with periods of transposon multiplication (Bennetzen 2002; Gaut and Ross-Ibarra 2008). Considerably less is known about the mechanisms that have led to the expansion of conifer genomes. For conifers, there is no evidence of polyploidization, but retrotransposons have been shown to be abundant and widespread in conifer genomes (Morse et al. 2009; Magbanua et al. 2011; Morgante and De Paoli 2011).

In addition to their unusually large sizes, early insights into the composition and structure of conifer genomes strongly suggest that they are very different from angiosperms, and that their organization cannot readily be predicted or deduced based on lessons learned from other plant genomes. For example, glimpses into the unique biology of conifer genomes have come from attempts to estimate and catalogue their protein coding gene content. Studies have identified a large overlapping set of sequences between conifers and angiosperms, including both herbaceous annuals and woody perennials (Ralph et al. 2008; Pozo et al. 2011; Rigault et al. 2011). However, the actual total number of genes in conifers and the proportion of the genome that they represent remain elusive (MacKay and Dean 2011). Results reviewed in this report suggest that a complete genome sequence will hold the key to resolving these questions among many others.

The extreme genome expansion due to retrotransposon insertions reported for conifers is likely to have affected gene structure in ways that remain to be described, and has resulted in large distances between genes distributed amid a vast ocean of non-coding DNA. We may also expect (see Lynch 2007) that it has favored the accumulation and retention of many neutral or slightly detrimental mutations, which in turn could have contributed further to genome inflation, in addition to providing a potential store of within-species genetic diversity. Understanding the relationships between the very large genome sizes, gene structure and genome organization in conifers will undoubtedly provide new insights into genome evolution and function. Along with advances in our understanding of conifer physiology and ecology, information gained from conifer genome sequencing will help us describe the potential for conifer species and populations to adapt to environmental change or respond to selective breeding, which will help us to better protect and improve them in the future.

Technology advancements have now made the sequencing of conifer genomes feasible, and, far from representing just another plant group of interest, it is expected that such analyses will significantly expand basic knowledge of plant genomes into new areas because of the uniqueness of the genomes themselves as well as the taxonomic position of conifers relative to plants that have been sequenced previously. Furthermore, many opportunities for applied research will result owing to the ecological importance of conifers combined with their potential role in solutions to global warming as well as their economic significance in meeting the global demand for wood and other biomaterials. The purpose of this review is to outline the current state of knowledge concerning conifer genomes and to highlight not only the opportunities for knowledge creation and application associated with conifer genome sequencing, but also some of the underlying challenges that may be expected.

Accomplishments in conifer genomics

Transcriptome sequencing and analysis

Although genome sequencing has not been attempted in conifers, large-scale investigations of expressed gene sequences have being ongoing in conifers for over a decade through cDNA analysis, and clustering of expressed

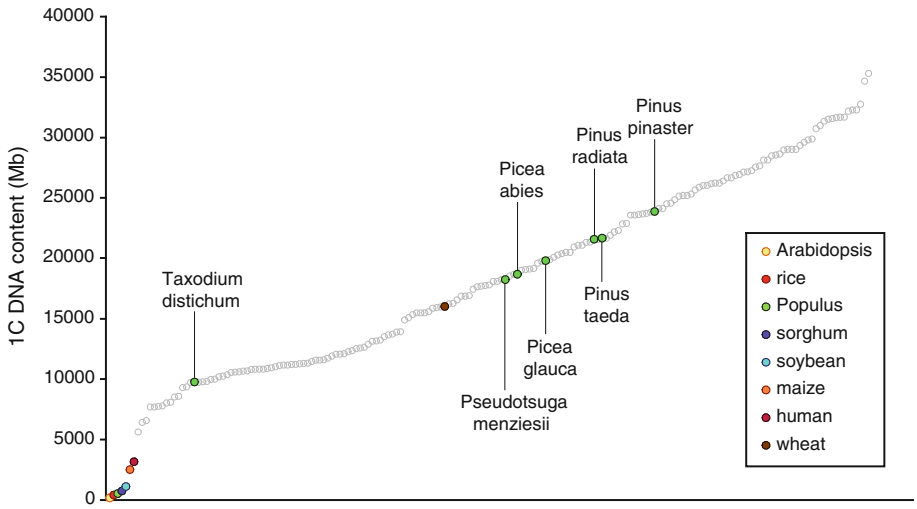

Fig. 1 Genome sizes of 181 conifers (Murray et al. 2004) and select angiosperms (Bennett and Leitch 2005) with complete- or partially-sequenced genomes. Each angiosperm is represented by a colored circle (see KEY) while conifers are represented by *white circles* having *gray outlines*, except for a few socioeconomically important conifer species, which are represented by *black circles* with labels

sequence tags (ESTs) to infer putative unigenes or transcript sets (reviewed by MacKay and Dean 2011). Knowledge of protein coding sequences has thus become rather extensive and has been fundamental to enhancing our understanding of a variety of biological processes and evolutionary mechanisms in conifers.

Over 90 % of the approximately 1 million gymnosperm ESTs found in dbEST are from conifers, mostly representing pines (*Pinus*) and spruces (*Picea*) (Pinaceae family) (MacKay and Dean 2011). Wood formation has received the most attention in gene discovery efforts, but other biological processes have also been targeted, including somatic embryogenesis, responses to defoliating insects, and root responses to water stress (for a partial list of larger projects, see Conifer Genome Network CGN; http://www.pinegenome.org/). Currently available expressed sequence resources include 328,662 ESTs and 17,379 unigenes (GenBank) for loblolly pine (*Pinus taeda*), as well as 313,110 and 186,637 ESTs with 27,848 and 19,944 unigenes for white spruce (*Picea glauca*) and Sitka spruce (*Picea sitchensis*), respectively (Table 1). Other conifers having EST datasets containing >10,000 sequences include *Cryptomeria japonica*, *Pinus contorta*, *Pinus banksiana*, *Pinus pinaster*, *P. radiata*, *Pseudotsuga menziesii*, *Picea engelmannii* × *P. glauca*, and *Picea abies*. Large full-length (FL) cDNA collections are available for *P. glauca* (23,000 unique fully sequenced cDNA insert sequences) and *P. sitchensis* (13,005 FL-cDNAs). Extensive gene catalogues have been derived from these data (Ralph et al. 2008; Rigault et al. 2011; Pozo et al. 2011) making feasible a wide variety of functional genomic approaches previously unavailable for conifer studies.

Conifer transcriptome sequencing using the 454 life sciences platform (Roche) recently added more than 10 million reads for over a dozen conifer species drawn for all seven families of conifers (Lorenz et al. 2012). Similar studies have developed resources specifically for *P. contorta* (Parchman et al. 2010), and *P. glauca* (Rigault et al. 2011). The 1,000 Plant Transcriptome (1KP) project (http://www.onekp.com/) set its sights on short-read sequence transcriptome scans from every plant family and numerous genera. Its current goals include generation of transcriptome scans comprising approximately 2 Gb of short-read (Illumina) sequences for at least 48 conifer species representing all seven conifer families, as well as several other gymnosperms and non-flowering plants. Once released, these sequences will constitute an enormous resource for comparative genomic studies of conifers.

Species-specific, as well as multi-species conifer EST assemblies are to be found in several public databases including NCBI (Table 1). The recently developed database, EuroPineDB, integrates ESTs from both dideoxy and 454 sequencing for several pine species (Pozo et al. 2011). MacKay and Dean (2011) discussed the approaches used to generate the wide ranging numbers of unigenes and assemblies found in some of these databases. Next-generation sequencing (NGS) greatly reduces the sequencing cost, but assembly fidelity can be difficult due to repeats, insertions/deletions, highly conserved paralogs, and other sequence variations that occur on scales ranging from a few tens of bases to complete genes (Wall et al. 2009). The large FL-cDNA datasets available for *P. glauca* and *P. sitchensis* are thus extremely important for their potential to serve as templates for guiding assembly of short-read sequences (Rigault et al. 2011).

Insights into gene and protein families and non-coding small RNAs

Phylogenetic analyses of EST and cDNA sequences have highlighted distinct evolutionary trajectories for gene families in conifers and angiosperms, including genes

Table 1 Transcriptomics resources currently available for the five best-studied conifer genera

Genus	ESTs	454/ Illumina	Unigene	FLcDNAs	Database	Website/reference
Pinaceae						
Pinus	3,983,264		35,550		PineDB	http://bioinfolab.muohio.edu/txid3352v1/
	1,445,703	79,823	283,479		TreeGenes	http://dendrome.ucdavis.edu/treegenes/
	382		954,411		EuroPineDB	http://www.scbi.uma.es/pindb/gene_libraries
		686,170			NCBI SRA	http://www.ncbi.nlm.nih.gov/sra
		586,732	17,000		NCBI SRA	Parchman et al. (2010)
	452,256		77,326	2,225	The Gene Index Project	http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=pine
	369,187				NCBI dbEST*	http://www.ncbi.nlm.nih.gov/nucest/
			17,379		NCBI Unigene	http://www.ncbi.nlm.nih.gov/unigene/statistics/
Picea	209,386		27,720	23,589	Arborea	http://www.arborea.ulaval.ca/
	542,511	19,026			TreeGenes	http://dendrome.ucdavis.edu/treegenes/
	541,490		79,409	9,636	The Gene Index Project	http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=spruce
	75,000		46,745	6,464	Treenomix	http://www.treenomix.ca/
	505,719				NCBI dbEST	http://www.ncbi.nlm.nih.gov/nucest/
			47,792		NCBI Unigene	http://www.ncbi.nlm.nih.gov/unigene/statistics/
				12,375	NCBI SRA	http://www.ncbi.nlm.nih.gov/sra
	656				NA	Prunier et al. (2011)
Pseudotsuga	3,788	64,250	51,548	209	TreeGenes	http://dendrome.ucdavis.edu/treegenes/
	18,142				NCBI dbEST*	http://www.ncbi.nlm.nih.gov/nucest/
		9,050			NCBI SRA	http://www.ncbi.nlm.nih.gov/sra
Larix		572,403	65,115		GenBank	Zhang et al. (2012)
	966				NCBI dbEST	http://www.ncbi.nlm.nih.gov/nucest/
	899				TreeGenes	http://dendrome.ucdavis.edu/treegenes/
Cupressaceae						
Cryptomeria	57,658		24,299		TreeGenes	http://dendrome.ucdavis.edu/treegenes/
	36,011		10,463		NCBI dbEST	Futamura et al. (2008)

dbEST entries in 2008–2009, according to MacKay and Dean (2011)

According to Zhang et al. (2012)

related to cell wall and wood formation, such as arabinogalactan proteins (*AGP*s), cellulose synthases (*CesA*), and expansins, defense-related genes, such as dirigent proteins and the cytochrome P450 monooxygenases of the terpenoid oxygenase superfamily, and transcription factors, such as auxin response factors (*Aux/IAAs*), *KNOTTED-LIKE HOMEOBOX Class 1* (*KNOX1*), and R2R3-MYBs (see MacKay and Dean 2011). For example, the expressed *KNOX1* genes in conifers were found to belong to only one of three sub-families found in angiosperms, but had more recent gene duplications (Guillet-Claude et al. 2004). The AGP, Aux/IAA, and dirigent protein families were comprised of clades representing numerous duplications that had occurred since the angiosperm-gymnosperm split (Li et al. 2010). And the PAL gene family was shown to have undergone gene duplication and loss events that have resulted in an expanded gene clade specific to gymnosperms (Bagal et al. 2012).

Small non-coding RNAs, including microRNAs (miRNA) and short-interfering RNAs (siRNA), have distinct functions and modes of formation in conifers, and contribute to transcriptional and post-transcriptional gene regulation (Morin et al. 2008). The miRNAs, which include many negative regulators of gene expression, predominantly accumulate as both 21- and 24-nucleotide RNAs in angiosperms, but only 21-nucleotide miRNAs were found in gymnosperms (Dolgosheina et al. 2008; Morin et al. 2008; Yakovlev et al. 2010). Specific miRNAs previously described in other plants have also been found in several conifers species (Dolgosheina et al. 2008), where they have been linked to important processes such as response to pathogens in *P. taeda* (see MacKay and Dean 2011). Novel miRNAs of uncharacterized function (51 in total) were identified in *P. contorta* (Morin et al. 2008), and in *P. abies*, miRNAs were shown to be involved in regulating temperature-dependent epigenetic memory (Yakovlev et al.

2010). Thus, current knowledge indicates that miRNAs are likely important for conifer biology, but the mechanics for their synthesis is unlike that in angiosperms plants, while siRNAs have not yet been described in conifers.

Transcript profiling

Custom cDNA arrays have been developed using sequences from various Pinaceae species (reviewed in MacKay and Dean 2011). Large cDNA arrays (>20,000 spots) using *P. glauca* (Holliday et al. 2008) and *P. taeda* (Lorenz et al. 2011) sequences, as well as an oligonucleotide microarray representing nearly 24,000 distinct *P. glauca* genes (Beaulieu et al. 2011), have enabled truly genome-scale analyses of conifer gene transcription.

Vascular tissue development and wood formation have been a major focus of transcript profiling investigations in conifer trees to date (MacKay and Dean 2011). Studies have characterized different stages of xylem differentiation and development, the response to mechanical stress, the transcripts that accumulate in secondary xylem relative to secondary phloem or needles (Pavy et al. 2008a), and transcripts, such as AGPs, lignin biosynthetic enzymes, alpha-tubulins, and various cell wall proteins, that are up-regulated in compression wood (see MacKay and Dean 2011). The latter paper also reviewed transcriptional profiling investigations of biotic factors affecting conifers, such as pathogens, insects and symbionts, as well as abiotic factors, such as drought, photoperiod and temperature as determinants of phenology, development and dormancy (El Kayal et al. 2011). These studies have only scratched the surface and a comprehensive understanding of conifer transcriptome dynamics and structure remains to be developed.

Genetic mapping: SNP resources and comparative mapping

Genetic linkage maps have been developed for the most economically important conifers using a variety of molecular markers (reviewed in Ritland et al. 2011). For example, as described by Echt et al. (2011) for loblolly pine (*P. taeda* L.), nuclear microsatellites (nSSR), microsatellites from expressed sequence tags (EST-SSRs), restriction fragment length polymorphisms (RFLPs) and expressed sequence tags polymorphisms (EST-Ps) were all combined to generate a consensus map of 460 markers covering 1,515 cM across 12 linkage groups. During the past 5 years, EST sequencing and the large-scale resequencing of amplicons gave rise to a scale change in SNP marker development (Le Dantec et al. 2004; Pavy et al. 2006) so that the pace of development for in silico SNP resources is speeding up dramatically. High-throughput genotyping methods have greatly benefitted from the accelerated rate of nucleotide polymorphism discovery, and in vitro SNPs verified by resequencing of genomic DNA using Sanger sequencing as well as SNPs discovered in silico, have been used successfully to construct gene-based linkage maps for *P. glauca* (Pavy et al. 2008b; Pelgas et al. 2011), *P. taeda* (Eckert et al. 2009a), and *P. pinaster* (Chancerel et al. 2011). Comparative analysis of EST sequences across species has enabled the development of comparative orthologous sequence (COS) markers that can be used across a broad spectrum of conifer species. Such markers showed high levels of macro-synteny between *P. taeda* and *P. sylvestris* (Komulainen et al. 2003), *P. taeda* and *P. pinaster* (Chancerel et al. 2011), and across multiple genera in the Pinaceae family (Pelgas et al. 2006). Liewlaksaneeyanawin et al. (2009) developed a set of 239 COS markers that have proven valuable for synteny analyses between *Picea*, *Pinus* and *Pseudotsuga* species. Together with the reference cytomolecular map of loblolly pine (Islam-Faridi et al. 2007), the emerging high-density gene-based maps for several conifer species will provide further opportunities for studying and comparing the organization and evolution of these genomes.

Applications: association studies and QTL analyses

Dense linkage maps can now be used to narrow down the location of loci that influence quantitative traits through classical quantitative trait locus mapping (Pelgas et al. 2011) (Tables 2, 3). So far, limited studies of candidate gene associations have been conducted with respect to wood-quality genes in *P. taeda* (González-Martínez et al. 2007), *P. radiata* (Dillon et al. 2010) and *P. glauca* (Beaulieu et al. 2011). Drought-related traits (associated with carbon-isotope discrimination) were also analyzed in *P. taeda* (González-Martinez et al. 2008) (Table 3). Associations between candidate genes and timing of growth cessation or cold tolerance initiation were examined in *P. menziesii* (Eckert et al. 2009b) and *P. sitchensis* (Holliday et al. 2010). More recently, Eckert et al. (2010) analyzed associations between a large set of SNPs (independent of function) and aridity-related environmental variants as a surrogate for the phenotype of each tree.

These candidate gene-based studies have, in general, confirmed earlier results of QTL mapping, i.e. the effects of individual loci on quantitative traits are mostly small, and the total detected effects are still far from accounting for all of the heritability for a given trait. In maize, in a very large study with high power, the effects of individual SNPs on flowering time were even smaller, but the whole-genome associations accounted for most of the additive genetic variance (Buckler et al. 2009). Additional genomic sequence characterization in conifers will provide more

Table 2 Saturated linkage maps in conifers

Species	Mapping population	Marker type	Linkage maps	No. of mapped markers	No. of LG	Reference
C. japonica	Three-generation outbred pedigree (♀ YI96 ('♀ Yabukuguri × ♂ Iwao') × ♂ YI38 ('♀ Yabukuguri × ♂ Kuji34')), n = 150	SNP, CAPS, RFLP, SSR, EST-SSR and ALP	One linkage map	1,261	11	Moriguchi et al. (2012)
P. abies	Two-generation outbred pedigree (♀ TH787F × ♂ Sire5), n = 73	AFLPs, SSRs, ESTPs and rDNA	Two parental linkage maps	461, 360	12, 16	Acheré et al. (2004)
			One consensus map	755	12	
P. glauca	Two-generation outbred pedigree C9612856 (♀ 80112 × ♂ 80109), n = 292	ALFP, ESTP, SSR and SNP	Two individual linkage maps	483, 523	12, 13	Pavy et al. (2008b)
			One composite map	821	12	
P. glauca	Two unrelated two-generation outbred pedigree: P C96-1-2856 (♀ 80112 × ♂ 80109, n = 260) and D C94-1-2516 (♀ 77111 × ♂ 2388, n = 500)	SNPs	Four individual linkage maps	471–570	12	Pelgas et al. (2011)
			One composite map	836		
Picea mariana	Three-generation outbred pedigree [♀ 643 × (♀ 32 × ♂ 40)], n = 90	AFLPs, SAMPL, SSRs and ESTPs	Two parental linkage maps	816, 743	12	Kang et al. (2010)
			One consensus map	1,111		
Picea mariana × *P. rubens*	Backcross-like BC1 [9920002: ♀ 11307-03 (♀ 83 × ♂ 425) × ♂ 425], n = 283	ALFP, ESTP, SSR and SNP	Two individual linkage maps	461, 479	12	Pavy et al. (2008b)
			One composite map	835		
P. lambertiana	Two unrelated open-pollinated haploid pedigrees (5,701 and 6,000) and a two-generation outbred pedigree (♀ 5038 × ♂ 5500)	SNPs	One consensus linkage map	399	19	Jermstad et al. (2011)
P. pinaster	Two pedigrees: (1) a three-generation outbred pedigree (G2), n = 201, and (2) a three-generation inbred pedigree (F2), n = 500	SNPs, AFLPs, SSRs, ESTPs	Three individual linkage maps	906, 179	13, 14, 21	Chancerel et al. (2011)
			One consensus map	354	24	
P. pinaster	Two related two-generation outbred pedigrees (♀ C14 × ♂ C15; ♀ C15 × ♂ C14), n = 106 and 55	SSR, SNPs, SAMPLs, ESTs	Two parental linkage maps	211, 215	13, 14	De Miguel et al. accepted
P. taeda	Two unrelated two-generation outbred pedigrees: base pedigree (♀ 20-1010 × ♂ 11-1060, n = 98) and qtl pedigree (♀ 6-1031 × ♂ 8-1070, n = 172)	RAPD, RFLP, AFLP, ESTP, SSR and SNPs	One consensus linkage map	373	12	Eckert et al. (2009b)

Table 2 continued

Species	Mapping population	Marker type	Linkage maps	No. of mapped markers	No. of LG	Reference
P. taeda	Two unrelated two-generation outbred pedigrees: base pedigree (♀ 20-1010 × ♂ 11-1060, n = 98) and qtl pedigree (♀ 6-1031 × ♂ 8-1070, n = 172)	AFLPs, RAPDs, RFLPs, ESTPs and SSRs	Two pedigree linkage maps One composite map	275, 338 460	12	Echt et al. (2011)

comprehensive sets of markers that also account for gene promoters and non-genic regions of the genome. These expanded sets of markers, which will include more rare (and possibly larger effect) variants, will provide powerful tools for breeding programs.

Links to updated results from ongoing and concluded research programs and projects in the area of tree genomics have been collected as part of the FoResTTraC project and can be found at http://www.foresttrac.eu/index.php/resources-database as well as in Plomion et al. (2007, 2011).

Pilot studies in genome sequencing

Sequence structure: BAC sequencing

With the aim of characterizing the structure of transcriptional units as well as non-coding intergenic regions, large-insert "bacterial artificial chromosome" (BAC) genomic libraries have been constructed for *P. pinaster*, *P. glauca, P. taeda*, and *Taxodium distichum* (Bautista et al. 2007; Hamberger et al. 2009; Magbanua et al. 2011; www.mgel.msstate.edu).

Targeted screening of a small number of conifer BACs harboring gene or gene-like sequences has indicated that pseudogenes may be a frequent feature within conifer genomes (Kovach et al. 2010; Magbanua et al. 2011). Of two *P. pinaster* BACs hybridizing to a sequence for ferredoxin-dependent glutamate synthase (Fd-GOGAT), one contained a sequence coding for an intact Fd-GOGAT polypeptide while the other contained what appeared to be a pseudogene (Bautista et al. 2007). Of eight *P. taeda* BACs containing sequences hybridizing to known genes, most contained apparent pseudogenes (Kovach et al. 2010). *P. glauca* BAC clones harboring intact coding sequences for genes encoding a terpenoid synthase and a cytochrome P450 were successfully isolated using PCR-screening and amplicon sequencing to confirm identity (Hamberger et al. 2009).

In all of these cases, no more than one intact gene sequence was present in any given BAC assembly (>100 kb), suggesting that large intergenic regions separate the coding sequences in conifer genomes. Consistent with this interpretation and the very large size of conifer genomes, sequencing of several random BACs proved unsuccessful in identifying multiple coding sequences (unpublished data).

Conifer genomes contain very large amounts of repetitive DNA (Morse et al. 2009; Magbanua et al. 2011). For example, the *P. taeda* LEA3 BAC was composed of 18.8 % LTR retroelement sequences (Magbanua et al. 2011). Similarly, three conifer-specific LTR retroelements (*PtIFG7*, *PtGypsyX1* and *PtCopiaX1*), as well as direct and tandem repeats from putative uncharacterized LTR retrotransposons, were abundant in *P. taeda* and *P. pinaster* BACs (Kovach et al. 2010; Cánovas et al. unpublished). The distribution of the *Gypsy* retrotransposons, *Gymny* and *PtIFG7,* and the *Copia* retroelement, TP1, suggest that pine genomes contain highly abundant and diverse repetitive DNA (Morse et al. 2009; Magbanua et al. 2011; Kovach et al. 2010), and it has been hypothesized that accumulation of retrotransposon derivatives could explain the tremendous size and complexity of conifer genomes (Morse et al. 2009). There is even evidence that the low-copy, non-genic portion of the pine genome is primarily composed of extremely diverged mobile elements (Morse et al. 2009; Kovach et al. 2010; Magbanua et al. 2011). These patterns seen in pine genomic sequences were consistent with major features observed in the spruce (*P. glauca*) genome (Hamberger et al. 2009). This information holds important ramifications for efforts to develop reference genome sequences for conifers since the quantity, complexity and divergence of repetitive sequences in genomic DNA profoundly affect the speed and quality of output from available sequence assembly tools.

Genome composition: genome filtration or reduced representation sequencing, and snapshot sequencing

The tools and methods used in genomic analyses are changing rapidly, driven in large part by recent advances in sequencing technologies. The new generation of sequencing technologies, represented by HiSeq/Illumina, SOLiD/ABI, GS-FLX Titanium-FLX +/Roche, HeliScope/Helicos, PACBIO RS/Pacific Biosciences, and GridION-/Oxford Nanopore, are

Table 3 QTL and association genetics studies in conifers

Species	Molecular markers	Population structure	Traits analyzed	QTL-association	Reference
P. glauca	SNPs	Open-pollinated families	Wood quality traits	Association	Beaulieu et al. (2011)
	SNPs	2 unrelated two-generation outbred pedigrees	Bud flush, bud set and growth	QTL	Pelgas et al. (2011)
P. sitchensis	SNPs	Open-pollinated families	Budset timing and autumn cold hardiness	Association	Holliday et al. (2010)
P. contorta	SNPs	Three populations from three mountain ranges	Serotiny	Association	Parchman et al. (2012)
Pinus elliottii	RAPDs	Two-generation outbred pedigree	Aluminum tolerance	QTL	Kubisiak et al. (2000)
Pinus elliottii × *Pinus caribaea*	AFLPs, SSR	Two-generation outbred pedigree	Wood-quality and growth traits	QTL	Shepherd et al. (2002)
	SSR	Two-generation outbred pedigree and a three-generation inbred pedigree	Adventitious rooting	QTL	Shepherd et al. (2006)
(*Pinus palustris* × *Pinus elliottii*) × *Pinus elliottii*	RAPDs	Two-generation outbred pedigree	Growth traits	QTL	Weng et al. (2002)
P. pinaster	ESTs, AFLPs	Three-generation outbred pedigree	Wood-quality and growth traits	QTL	Pot et al. (2006)
	ESTs, AFLPs	Three-generation outbred pedigree	Wood density and cell wall components	QTL	Chagné et al. (2003)
	AFLPS, SSR, ESTS	Two-generation outbred pedigree	Wood-quality and growth traits	QTL	Markussen et al. (2003)
	RAPDs	Three-generation inbred pedigree	Growth traits	QTL	Plomion et al. (1996a)
	AFLPs	Three-generation outbred pedigree	Water use efficiency and growth	QTL	Brendel et al. (2002)
	SSR, SNPs, SAMPLs, ESTs	Two-generation outbred pedigree	Water use efficiency and growth	QTL	De Miguel et al. accepted
	RAPDs	Three-generation inbred pedigree	Secondary metabolism	QTL	Plomion et al. (1996b)
P. radiata	RAPD, AFLP, SSR	Three-generation inbred pedigree	Wood-quality traits	QTL	Ball (2001)
	RAPDs, AFLPs, SSR	Two-generation outbred pedigree	Wood-quality traits	QTL	Kumar et al. (2000)
	SNPs	Natural populations and land races	Wood quality traits	Association	Dillon et al. (2010)
	RFLPs, SSR	3 unrelated two-generation outbred pedigree	Wood-quality and growth traits	QTL	Devey et al. (2004a)
	RFLPs, SSR	6 related three-generation outbred pedigree	Wood-quality and growth traits	QTL	Devey et al. (2004b)
	RAPDs	Full-sib cross (megagamethophyte analysis)	Stem growth	QTL	Emebiri et al. (1997)
	RAPDs	Full-sib cross (megagamethophyte analysis)	Growth traits	QTL	Emebiri et al. (1998a)
	RAPDs	Full-sib cross (megagamethophyte analysis)	Stem growth	QTL	Emebiri et al. (1998b)
P. sylvestris	RAPDs	Three-generation half-sib design	Bud set and frost hardiness	QTL	Hurme et al. (2000)
	AFLP	Two-generation outbred pedigree	Growth, wood-quality and frost hardinesss	QTL	Lerceteau et al. (2000)
	RAPD	Two-generation outbred pedigree	Growth and cold acclimation	QTL	Yazdani et al. (2003)

Table 3 continued

Species	Molecular markers	Population structure	Traits analyzed	QTL-association	Reference
P. taeda	RFLPs, ESTS	Three-generation outbred pedigree	Wood-quality traits	QTL	Brown et al. (2003)
	SNPs	Trees from ex situ clone banks and seed orchards	Wood property traits	Association	González-Martínez et al. (2007)
	RFLP	Three-generation outbred pedigree (QTL pedigree)	Wood-quality traits	QTL	Groover et al. (1994)
	RFLP	Three-generation outbred pedigree (QTL pedigree)	Wood-quality traits	QTL	Knott et al. (1997)
	RFLPs	Three-generation outbred pedigree (QTL pedigree)	Physical wood properties	QTL	Sewell et al. (2000)
	RFLP	Three-generation outbred pedigree (QTL pedigree)	Wood-quality traits	QTL	Sewell et al. (2002)
	SSR	Three-generation outbred pedigree (base pedigree)	Growth traits	QTL	Gwaze et al. (2003)
	RFLPs, RAPDs, Isozymes	2 unrelated three-generation outbred pedigrees: BASE and QTL pedigrees	Annual height- and diameter-increment growth	QTL	Kaya et al. (1999)
	SNPs	Unrelated trees sampled across the natural range	Environnemental association study (aridity)	Association	Eckert et al. (2010)
	SNPs	Unrelated trees sampled across the natural range	Analysis of metabolites	Association	Eckert et al. (2012)
	SNPs	Sixty-one families using a partial diallel mating design	Carbon isotope discrimination	Association	González-Martinez et al. (2008)
P. menziesii	AFLPs	Eight full-sib families F2	Wood quality and growth	QTL	Ukrainetz et al. (2008)
	RFLPs	Three-generation outbred pedigree	Bud flush	QTL	Jermstad et al. (2001)
	RFLPs	Three-generation outbred pedigree	Bud phenology	QTL	Jermstad et al. (2003)
	RFLPs	Three-generation outbred pedigree	Spring cold hardiness	QTL	Wheeler et al. (2005)
	SNPs	Unrelated families that represent an extensive distribution and a diversity of environmental conditions	Spring cold hardiness	Association	Eckert et al. (2009a)

providing new opportunities for high-throughput functional genomic research. Yet even with the vastly improved throughput of these new technologies, large genomes, such as those of *P. abies* (1.8×10^{10} bp) or *P. taeda* (2.7×10^{10} bp) compared to *A. thaliana* (1.6×10^{8} bp) (Fig. 1), demand novel approaches in order to reduce genome complexity to levels permitting sequence assembly with reasonable fidelity.

Among several methods available for creating reduced-representational libraries for genome sequencing (Springer et al. 2004), restriction enzymes sensitive to DNA methylation at the ′5 position of cytosines in CpG dinucleotides can be used to generate genome fractions that are enriched for genes. This approach can be used prior to sequencing to help eliminate the most highly repetitive elements, which can constitute >75 % of conifer genomes (Kovach et al. 2010; Morse et al. 2009). However, one pilot study using this approach did not find the expected degree of enrichment using *P. taeda* DNA, suggesting that methylation patterns in pine may be somewhat different from those typically seen in angiosperms (Rabinowicz et al. 2005). Sequence capture methods developed to enrich for exons in resequencing projects (Ng et al. 2009), as well as methods based on thermal renaturation kinetics (Peterson et al. 2002), represent alternative approaches for reducing the complexity of conifer genomic DNA for sequencing projects. Because of their tendency to enrich for gene space sequences, these techniques are particularly useful for targeting protein coding regions that correspond to the expressed sequences available from EST projects.

Genome scans have been performed at low coverage on diploid genomic DNA from *P. taeda* (Kovach et al. 2010), *P. glauca* (Rigault et al. 2011), and *P. abies* (Morgante and De Paoli 2011). Early analyses of these studies too have indicated that a significant proportion of the genome is comprised of uncharacterized repeats (Morgante and De Paoli 2011). However, while these authors reported that 3 % of the genome of *P. abies* (approximately 600 Mb) is comprised of sequences similar to genes, their finding was at odds with studies based on cDNAs that estimated the *P. glauca* transcriptome at between 40 and 50 Mb (Rigault et al. 2011). On the other hand, the finding that a large portion of the genome may be comprised of gene-like sequences could fit with earlier results suggesting an unexpectedly complex transcriptome in *P. taeda* (Lorenz and Dean 2002). These discrepancies could indicate that conifer genomes could contain more non-coding pseudogene sequences than genes encoding expressed functional proteins or that novel transcriptional mechanisms may exist in conifers. Additional studies are underway using haploid genomic DNA from *P. abies* (Invargson et al. unpublished data) and *P. pinaster* (Cervera et al. unpublished) to help test these hypotheses. Additional efforts using low-coverage genome scans will provide important preliminary evidence for the potential of using NGS in shotgun sequencing approaches to characterize conifer genomes.

Future directions, prospects, and barriers

Opportunities and strategies for sequencing conifer genomes

In the last 2 years, genome sequencing projects for seven different conifer species representing three different genera have been launched: (a) the *P. abies* genome project, a European consortium led by Sweden to sequence the genome of Norway spruce (http://www.upsc.se/Networks/Networks/sprucegenome.html); (b) PineRefSeq, a USDA-funded project to sequence the genomes of loblolly pine (*P. taeda*), sugar pine (*Pinus lambertiana*), and Douglas-fir (*P. menziesii*) (http://dendrome.ucdavis.edu/NealeLab/pinerefseq); (c) SMarTForests, a Genome Canada project to characterize the white spruce (*P. glauca*) genome; (d) a Spanish consortium, funded by MICINN, with INRA contribution, to generate an initial draft of *P. pinaster* genome sequence; and (e) ProCoGen, a 7FP-KBBE project from the European Commission, to sequence the maritime (*P. pinaster*) and Scots pine (*P. sylvestris*) genomes.

Recent advances in sequencing technology have made it a straightforward exercise to achieve 50–100× coverage from conifer giga- genomes using whole-genome shotgun sequencing (WGS). WGS is typically performed using paired-end sequencing of fragments from a number of different size classes (from 100 to 800 bp); however, these fragment sizes are too short to span most repetitive elements in conifers. WGS therefore needs to be complemented with jumping (mate pair) libraries generated from fragments ranging from a few kb to tens of kb. Because even 10–20 kb fragments are too short to span some repeat regions, other methods, for example, end-sequencing of fosmids or BACs, as well as sequencing of fosmid pools, are actively being researched (Philippe et al. 2012). The silver lining to conifer genome sequencing with respect to repeat sequences is that a significant portion of the repeats appear to be rather old and as a consequence are highly diverged in their sequences (Kovach et al. 2010; Morgante and De Paoli 2011). Therefore, the highly repetitive nature of conifer genomes may in the end not be an impassable obstacle for sequence assembly. Combined with appropriate approaches to reduce genome complexity (genome filtration or BAC/fosmid pooling), the assembly of a conifer reference genome sequence appears more and more feasible. Assembling data from WGS from diploid tissue may lead to problems with allele splitting due to sequence polymorphisms and polymorphic repetitive elements. However, conifers have the advantage that bulk haploid

tissue is available from megagametophytes, and WGS from haploid tissue may mitigate a lot of these assembly problems. The only drawback is that a single megagametophyte might not provide enough DNA to achieve the desired sequence coverage given the seed size in most of the conifer species currently targeted by genome projects. Regardless of what different strategies that are pursued, a critical and difficult step for most de novo conifer genome projects will be merging of multiple assemblies produced from different kinds of data.

A more significant short-term obstacle for conifer genome assembly may be accessing adequate computing resources, particularly computers with sufficient random-access memory (RAM), to handle the most advanced sequence assembly methods. Successful assembly of these reference genome sequences will require high-density genetic maps that can be used to position scaffolds along genetic linkage groups. The genetic maps currently available for conifers have relatively low marker density, given the very large numbers of sequence scaffolds expected from WGS assemblies. Construction of saturated genetic maps and expanded mapping populations for higher map resolution are currently in progress (e.g. Neale and Kremer 2011; Pelgas et al. 2011; Table 2).

Application of genomic information to advanced breeding

While the conifer species used in commercial production worldwide have undergone a degree of artificial selection, which has improved productivity and quality of plantations (White et al. 2007), most are regarded as undomesticated and still occur in large, broadly distributed populations representing large pools of untapped genetic potential. These obligate out-crossing species are highly heterozygous and harbor diverse alleles that await discovery through high-throughput surveys. Provided with an understanding of how different combinations of alleles determine economic performance as well as fitness, future tree breeding and conservation programs will likely use cataloged individual genomes for purposes of controlled matings, for preserving diversity, and for selecting optimal combinations of alleles (Nelson and Johnsen 2008; Neale and Kremer 2011) using techniques such as genomic selection (Grattapaglia and Resende 2011). Iwata et al. (2011) have, in fact, used a modeling approach to demonstrate how genomic approaches may be optimized for greater efficiency in conifer tree breeding. The high degree of synteny between conifer species and the increasing availability of high-density genetic maps should bring the power of these genomics-based tree improvement approaches even to species that previously received relatively little attention (Liewlaksaneeyanawin et al. 2009; Chancerel et al. 2011; Jermstad et al. 2011; Pavy et al. 2012).

Improved understanding of conifer biology through genomics

The key to realizing the vision of both enhancing and preserving conifers is a deep understanding of the relationships between specific alleles and phenotypic traits, as well as the influence of environmental conditions on trait expression. Nowadays, knowledge of the function and relationship to phenotype for specific genes and alleles is growing rapidly in other plants and important strides have been made toward understanding these relationships in conifers. The possibilities for linking this information via conifer genomic sequences hold enormous potential for rapidly improving our understanding of fundamental conifer biology and for identifying genes and gene networks critical to the performance of these trees for commercial purposes as well as in the wild (Neale and Ingvarsson 2008; Dean 2011).

Unlocking the conifer genome will have several economical and environmental benefits. Fundamental understanding of the genes that control wood formation, the responses to the attack of invasive pests or to environmental risks, and the capacity of a better and faster growth as well as an efficient carbon sequestration will result in the availability of new tools to breed robust and well-adapted trees that assimilate CO_2 from the atmosphere, or produce renewable bio-based products, including bio-fuels, more efficiently in a changing climate. This will accelerate the selection process and the discovery of new ways to capture the ecological and economic value contained in the genetic information of conifers.

These resources will also make it possible to proceed with genome-scale investigations that have so far only resided in our imaginations. For example, is there a biological basis for the rarity of polyploidy conifers? What constitutes the microbiome of a conifer and to what degree does this microbiome influence conifer responses to the environment? Have conifers found a way to prevent viral infections or have we just overlooked conifer viruses as the agents of pathologies of unknown cause? Have the long life spans and generation times of conifers led their genomes to retain genes whose expression is only needed to guard against selective processes that occur on the order of millennia? Could such ‘hidden’ genes allow conifers to adapt more rapidly than expected to challenges such as climate change? In addition to such large-scale organismal questions, it stands to reason that having reference genome sequences for conifers will also make it possible to study a variety of small-scale molecular questions as well. For

example, to what extent do uncharacterized types of genotypic variation, including copy number variation, regulatory non-coding RNAs, transposable elements and epigenetic imprinting contribute to the diversity significant for adaptation.

The sequencing of conifer genomes is expected to rapidly fill these key gaps in our knowledge. We will thus be able to more accurately identify the natural stores of genetic diversity in conifers and using this knowledge, devise plans to protect and preserve this key genetic resource. This knowledge could facilitate the identification of genetic mechanisms and their association with traits of interest provided that databases and other information resources are designed with foresight. The transfer of knowledge will be facilitated by adopting standard nomenclature and data structure conventions that already exist in the genomic research communities. Development of these information resources for conifers requires new investment, and will benefit greatly from the wide range of genomic tools and resources for diverse organisms.

Perhaps the most critical investment going forward will be in the human capital associated with conifer genomics. Even with great technological advances, conifer genome projects are large-scale and have long time horizons. To fully realize the goals of this work will require the coordinated efforts of many researchers across national and international boundaries, through continuous assessment of progress and planning of future work.

Acknowledgments The preparation of this article was supported by FoResTTraC (2440096 FP7) and ProCoGen (289841-FP7) UE projects. The authors are very grateful to Susana Ferrándiz for helping with the preparation of the manuscript.

References

Acheré V, Faivre-Rampant P, Jeandroz S, Besnard G, Markussen T, Aragones A, Fladung M, Ritter E, Favre JM (2004) A full saturated linkage map of *Picea abies* including AFLP, SSR, ESTP, 5S rDNA and morphological markers. Theor Appl Genet 108:1602–1613

Bagal UR, Leebens-Mack JH, Lorenz WW, Dean JFD (2012) Phylogenomic analysis of the phenylalanine ammonia Iyase gene family in loblolly pine (*Pinus taeda* L.). BMC Genomics 13:S1

Ball RD (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. Genetics 159:1351–1364

Bautista R, Villalobos DP, Diaz-Moreno S, Canton FR, Canovas FM, Claros MG (2007) Toward a *Pinus pinaster* bacterial artificial chromosome library. Ann Forest Sci 64:855–864

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

Bennett MD, Leitch IJ (2005) Angiosperm DNA *C*-values database. http://www.kew.org/cvalues/. Release 6.0, accessed Oct 2005

Bennetzen JL (2002) Mechanism and rates of genome expansion and contraction in flowering plants. Genetica 115:29–36

Brendel O, Pot D, Plomion C, Rozenberg P, Guehl JM (2002) Genetic parameters and QTL analysis of d13 in maritime pine. Plant Cell Environ 25:945–953

Brown GR, Bassoni DL, Gill GP, Fontana JR, Wheeler NC, Megraw RA, Davis MF, Sewell MM, Tuskan GA, Neale DB (2003) Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda*, L.). III. QTL verification and candidate gene mapping. Genetics 164:1537–1546

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Villeda HS, Sofia da Silva H, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325: 714–718

Chagné D, Brown G, Lalanne C, Madur D, Pot D, Neale D, Plomion C (2003) Molecular breeding—comparative genome and QTL mapping between maritime and loblolly pines. Mol Breed 12:185–195. doi:10.1023/a1026318327911

Chancerel E, Lepoittevin C, Le Provost G, Lin YC, Jaramillo-Correa JP, Eckert AJ, Wegrzyn JL, Zelenika D, Boland A, Frigerio JM, Chaumeil P, Garnier-Géré P, Boury C, Grivet D, González-Martínez SC, Rouzé P, Van de Peer Y, Neale DB, Cervera MT, Kremer A, Plomion Ch (2011) Development and implementation of a highly multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics 12:368

Dean JFD (2011) Future prospects. In: Plomion C, Bousquet J, Kole C (eds) Genetics, genomics and breeding of conifers trees. Edenbridge Science Publishers and CRC Press, New York, pp 404–438

Devey ME, Carson SD, Nolan MF, Matheson AC, Te Riini C, Hohepa J (2004a) QTL associations for density and diameter in *Pinus radiata* and the potential for marker-aided selection. Theor Appl Genet 108:516–524

Devey ME, Groom KA, Nolan MF, Bell JC, Dudzinski MJ, Old KM, Matheson AC, Moran GF (2004b) Detection and verification of quantitative trait loci for resistance to Dothistroma needle blight in *Pinus radiata*. Theor Appl Genet 108:516–524. doi:10.1007/s00122-003-1446-2

Dillon SK, Nolan M, Li W, Bell C, Wu HX, Southerton SG (2010) Allelic variation in cell wall candidate genes affecting solid wood properties in natural populations and land races of *Pinus radiata*. Genetics 185:1477–1487

Dolgosheina EV, Morin RD, Aksay G, Sahinalp SC, Magrini V, Mardis ER, Mattsson J, Unrau PJ (2008) Conifers have a unique small RNA silencing signature. RNA 14:1508–1515

Echt CS, Saha S, Krutovsky KV, Wimalanathan K, Erpelding JE, Liang C, Nelson CD (2011) An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. BMC Genet 12:17

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, Neale DB, Clair JB (2009a) Association genetics of coastal Douglas fir (*Pseudotsuga menziesii var. menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics 182:1289–1302

Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, Neale DB (2009b) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). Tree Genet Genomes 5:225–234

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martinez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185:969–982

Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, Tolstikov V, Fiehn O, Neale DB (2012) Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. New Phytol 193:890–902. doi:10.1111/j.1469-8137.2011.03976.x

El Kayal W, Allen CC, Ju CJ, Adams E, King-Jones S, Zaharia LI, Abrams SR, Cooke JE (2011) Molecular events of apical bud formation in white spruce, *Picea glauca*. Plant Cell Environ 34:480–500

Emebiri LC, Devey ME, Matheson AC, Slee MU (1997) Linkage of RAPD markers to NESTUR, a stem growth index in radiata pine seedlings. Theor Appl Genet 95:119–124

Emebiri LC, Devey ME, Matheson AC, Slee MU (1998a) Age-related changes in the expression of QTLs for growth in radiata pine seedlings. Theor Appl Genet 97:1053–1061

Emebiri LC, Devey ME, Matheson AC, Slee MU (1998b) Interval mapping of quantitative trait loci affecting NESTUR, a stem growth efficiency index of radiata pine seedlings. Theor Appl Genet 97:1062–1068

Farjon A (2008) A natural history of conifers. Timber Press, Portland 304 p

Futamura N, Totoki Y, Toyoda A, Igasaki T, Nanjo T, Seki M, Sakaki Y, Mari A, Shinozaki K, Shohara K (2008) Characterization of expressed sequence tags from a full-length enriched cDNA library of *Cryptomeria japonica* male strobili. BMC Genomics 9:383

Gaut BS, Ross-Ibarra J (2008) Selection of angiosperm genomes. Science 320:484–486

Gernandt DS, Willyard A, Syring JV, Liston A (2011) The conifers (Pinophyta). In: Plomion C, Bousquet J, Kole C (eds) Genetics, genomics and breeding of conifers trees. Edenbridge Science Publishers and CRC Press, New York, pp 1–39

González-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity 101:19–26

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. Genetics 175:399–409

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

Groover AT, Devey ME, Lee JM, Megraw R, Mitchell-Olds T, Sherman B, Vujcic S, Williams C, Neale DB (1994) Identification of quantitative trait loci influencing wood specific gravity in an outbred pedigree of loblolly pine. Genetics 138:1293–1300

Guillet-Claude C, Isabel N, Pelgas B, Bousquet J (2004) The evolutionary implications of *knox-I* gene duplications in conifers: correlated evidence from phylogeny, gene mapping, and analysis of functional divergence. Mol Biol Evol 21:2232–2245

Gwaze DP, Zhou Y, Reyes-Valdés MH, Al-Rababah MA, Williams CG (2003) Haplotypic QTL mapping in an outbred pedigree. Genet Res 81:43–50

Hamberger B, Hall D, Yuen M, Oddy C, Hamberger B, Keeling CI, Ritland C, Ritland K, Bohlmann J (2009) Targeted isolation, sequence assembly and characterization of two white spruce (*Picea glauca*) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. BMC Plant Biol 9:106

Holliday JA, Ralph SG, White R, Bohlmann J, Aitken SN (2008) Global monitoring of autumn gene expression within and among phenotypically divergent populations of Sitka spruce (*Picea sitchensis*). New Phytol 178:103–122

Holliday JA, Ritland K, Aitken SN (2010) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytol 188:501–514

Hurme P, Sillanpaa MJ, Arjas E, Repo T, Savolainen O (2000) Genetic basis of climatic adaptation in scots pine by Bayesian quantitative trait locus analysis. Genetics 156:1309–1322

Islam-Faridi MN, Nelson CD, Kubisiak T (2007) Reference karyotype and cytomolecular map for loblolly pine (*Pinus taeda* L.). Genome 50:241–251

Iwata H, Hayashi T, Tsumura Y (2011) Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. Tree Genet Genomes 7:747–758

Jermstad KD, Bassoni DL, Jech KS, Wheeler NC, Neale DB (2001) Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. I. Timing of vegetative bud flush. Theor Appl Genet 102:1142–1151

Jermstad KD, Bassoni DL, Jech KS, Ritchie GA, Wheeler NC, Neale DB (2003) Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas fir. III. Quantitative trait loci-by-environment interactions. Genetics 165:1489–1506

Jermstad KD, Eckert AJ, Wegrzyn JL, Delfino-Mix A, Davis DA, Burton DC, Neale DB (2011) Comparative mapping in *Pinus*: sugar pine (*Pinus lambertiana* Dougl.) and loblolly pine (*Pinus taeda* L.). Tree Genet Genomes 7:457–468

Kang BY, Mann IK, Major JE, Rajora OP (2010) Near-saturated and complete genetic linkage map of black spruce (*Picea mariana*). BMC Genomics 11:515

Kaya Z, Sewell MM, Neale DB (1999) Identification of quantitative trait loci influencing annual height- and diameter-increment growth in loblolly pine (*Pinus taeda* L.). Theor Appl Genet 98:586–592

Kenrick P (1999) The family tree flowers. Nature 402:358–359

Knott S, Neale DN, Sewell MM, Haley C (1997) Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. Theor Appl Genet 94:810–820

Komulainen P, Brown GR, Mikkonen M, Karhu A, Garcia-Gil MR, O'Malley D, Lee B, Neale DB, Savolainen O (2003) Comparing EST-based genetic maps between *Pinus sylvestri*s and *Pinus taeda*. Theor Appl Genet 107:667–678

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, Neale DB (2010) The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11:420

Kubisiak TL, Nelson CD, Nowak J, Friend AL (2000) Genetic linkage mapping of genomic regions conferring tolerance to high aluminum in slash pine. J Sustain Forest 10:69–78

Kumar S, Spelman R, Garrick D, Richardson TE, Wilcox PL (2000) Multiple marker mapping of quantitative trait loci on chromosome three in an outbred pedigree of radiata pine. Theor Appl Genet 100:926–933

Le Dantec L, Chagné D, Pot D, Cantin O, Garnier-Géré P, Bedon F, Frigerio JM, Chaumeil P, Léger P, Garcia V, Laigret F, de Daruvar A, Plomion C (2004) Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Plant Mol Biol 54:461–470

Lerceteau EC, Plomion C, Andersson B (2000) AFLP mapping and detection of quantitative trait loci (QTLs) for economically important traits in *Pinus sylvestris*: a preliminary study. Mol Breed 6:451–459

Li X, Wu HX, Southerton SG (2010) Comparative genomics reveals conservative evolution of the xylem transcriptome in vascular plants. BMC Evol Biol 10: Article 190

Liewlaksaneeyanawin C, Zhuang J, Tang M, Farzaneh N, Lueng G, Cullis C, Findlay S, Ritland CE, Bohlmann J, Ritland K (2009)

Identification of COS markers in the Pinaceae. Tree Genet Genomes 5:247–255
Lorenz WW, Dean JFD (2002) SAGE profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (*Pinus taeda*). Tree Physiol 22:301–310
Lorenz WW, Alba R, Yu Y-S, Bordeaux JM, Simões M, Dean JFD (2011) Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.). BMC Genomics 12:264
Lorenz WW, Neale DB, Jermstad KD, Howe GT, Rogers DL, Bordeaux JM, Ayyampalayam S, Dean JFD (2012) Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for twelve diverse conifer species. Tree Genet Genomes. doi:10.1007/s11295-012-0547-y
Lynch M (2007) The origins of genome architecture. Sinauer Associates Inc, Sunderland
MacKay JJ, Dean JFD (2011) Transcriptomics. In: Plomion C, Bousquet J, Kole C (eds) Genetics, genomics and breeding of conifers trees. Edenbridge Science Publishers and CRC Press, New York, pp 323–357
Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, Cronn RC, Nelson CD, Peterson DG (2011) Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. PLoS One 6:e16214
Markussen T, Fladung M, Achere V, Favre JM, Faivre-Rampant P, Aragones A, Da Silva P, Havengt L, Ritter E (2003) Identification of QTLs controlling growth, chemical and physical wood property traits in *Pinus pinaster* (Ait.). Silvae Genetica 52:8–15
Morgante M, De Paoli E (2011) Toward the conifer genome sequence. In: Plomion C, Bousquet J, Kole C (eds) Genetics, genomics and breeding of conifers trees. Edenbridge Science Publishers and CRC Press, New York, pp 389–403
Moriguchi Y, Ujino-Ihara T, Futamura N, Saito M, Ueno S, Matsumoto A, Tani N, Taira H, Shinohara K, Tsumura Y (2012) The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in in *Cryptomeria japonica* D. Don. BMC Genomics 19:95
Morin RD, Aksay G, Dolgosheina E, Ebhardt HA, Magrini V, Mardis ER, Sahinalp SC, Unrau PJ (2008) Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. Genome Res 18:571–584
Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carloson JE, Nelson CD, Davis JM (2009) Evolution of genome size and complexity in *Pinus*. PLoS One 4:e4332
Murray BG, Leitch IJ, Bennett MD (2004) Gymnosperm DNA *C*-values database. http://www.kew.org/cvalues/. Release 3.0, accessed Dec 2004
Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11:149–155
Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111–122
Nelson CD, Johnsen KH (2008) Genomic and physiological approaches to advancing forest tree improvement. Tree Physiol 28:1135–1143
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276
Parchman T, Geist K, Grahnen J, Benkman C, Buerkle A (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation and marker discovery. BMC Genomics 11:180
Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle C (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol 21:2991–3005
Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. BMC Genomics 7:174
Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H, Cooke J, Mackay J (2008a) Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. New Phytol 180:766–786
Pavy N, Pelgas B, Beauseigle S, Blais S, Gagnon F, Gosselin I, Lamothe M, Isabel N, Bousquet J (2008b) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 9:21
Pavy N, Namroud MC, Gagnon F, Isabel N, Bousquet J (2012) The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. Heredity 108:273–284
Pelgas B, Beauseigle S, Acheré V, Jeandroz S, Bousquet J, Isabel N (2006) Comparative genome mapping among *Picea glauca*, *P. abies* and *P. mariana* × *rubens*, and correspondence with other Pinaceae. Theor Appl Genet 113:1371–1393
Pelgas B, Bousquet J, Meirmans PG, Ritland K, Isabel N (2011) QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. BMC Genomics 12:145
Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12:795–807
Philippe R, Choulet F, Paux E, van Oeveren J, Tang J, Wittenberg AH, Janssen A, van Eijk MJ, Stormo K, Alberti A, Wincker P, Akhunov E, van der Vossen E, Feuillet C (2012) Whole genome profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. BMC Genomics 13:47
Plomion C, Durel CE, O'Malley DM (1996a) Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. Theor Appl Genet 93:849–858
Plomion C, Yani A, Marpeau A (1996b) Genetic determinism of δ3-carene in maritime pine using RAPD markers. Genome 39:1123–1127
Plomion C, Chagné D, Pot D, Kumar S, Wilcox PL, Burdon RD, Prat DG, Paiva J, Chaumeil P, Vendramin GG, Sebastiani F, Nelson CD, Echt CS, Savolainen O, Kubisiak TL, Cervera MT, de Maria N, Islam-Faridi MN (2007) Pines. In: Genome mapping and molecular breeding in plants: volume 7, Forest Trees. Springer, Berlin, pp 29–92
Plomion C, Bousquet J, Kole C (2011) Genetics, genomics, and breeding of conifers. CRC Press, Boca Ratón p 456
Pot D, Rodrigues JC, Rozenberg P, Chantre G, Tibbits J, Cahalan C, Pichavant F, Plomion C (2006) QTLs and candidate genes for wood properties in maritime pine (*Pinus pinaster* Ait.). Tree Genet Genomes 2:10–24
Pozo NF, Canales J, Fernandez DG, Villalobos D, Moreno SD, Bautista R, Monterroso AF, Guevara MA, Perdiguero P, Collada C, Cervera MT, Soto A, Ordas R, Canton FR, Avila C, Canovas FM, Claros MG (2011) EuroPineDB: a high-coverage web

database for maritime pine transcriptome. BMC Genomics 12:366

Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. Mol Ecol 20:1702–1716

Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA (2005) Differential methylation of genes and repeats in land plants. Genome Res 15:1431–1440

Ralph S, Chun H, Kolosova N, Cooper D, Oddy C, Ritland C, Kirkpatrick R, Moore R, Barber S, Holt R, Jones SJM, Marra MA, Douglas CJ, Ritland K, Bohlmann J (2008) A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). BMC Genomics 9:484

Raven PH, Evert RF, Eichhorn SE (2005) Biology of plants, 7th edn. W.H. Freeman and Co., New York p 686

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

Ritland K, Krutovsky KV, Tsumura Y, Pelgas B, Isabel N, Bousquet J (2011) Genetic mapping in conifers. In: Plomion C, Bousquet J, Kole C (eds) Genetics, genomics and breeding of conifers trees. Edenbridge Science Publishers and CRC Press, New York, pp 196–238

Sewell MM, Bassoni DL, Megraw RA, Wheeler NC, Neale DB (2000) Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). I. Physical wood properties. Theor Appl Genet 101:1273–1281

Sewell MM, Davis MF, Tuskan GA, Wheeler NC, Elam CC, Bassoni DL, Neale DB (2002) Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). II. Chemical wood properties. Theor Appl Genet 104:214–222

Shepherd M, Cross M, Dieters MJ, Henry R (2002) Branch architecture QTL for *Pinus elliottii* var. *elliottii* x *Pinus caribea* var. *hondurensis* hybrids. Ann Forest Sci 59:617–625

Shepherd M, Huang S, Eggler P, Cross M, Dale G, Dieters M, Henry R (2006) Congruence in QTL for adventitious rooting in *Pinus elliottii* × *Pinus caribaea* hybrids resolves between and within-species effects. Mol Breed 18:11–28

Springer NM, Xu XQ, Barbazuk WB (2004) Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. Plant Physiol 136:3023–3033

Ukrainetz NK, Ritland K, Mansfield SD (2008) Identification of quantitative trait loci for wood quality and growth across eight full-sib coastal Douglas-fir families. Tree Genet Genomes 4: 159–170

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang HY, Landherr L, Tomsho LP, Hu Y, Carlson JE, Ma H, Schuster SC, Soltis DE, Soltis PS, Altman N, dePamphilis CW (2009) Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10:347

Weng C, Kubisiak TL, Nelson CD, Stine M (2002) Mapping quantitative trait loci controlling early growth in a (longleaf pine × slash pine) × slash pine BC1 family. Theor Appl Genet 104:852–859

Wheeler NC, Jermstad KD, Krutovsky K, Aitken SN, Howe GT, Krakowski J, Neale DB (2005) Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. IV. Cold-hardiness QTL verification and candidate gene mapping. Mol Breed 15:145–156

White TL, Adams WT, Neale DB (2007) Forest genetics. CABI, Cambridge p 682

Yakovlev IA, Fossdal CG, Johnsen O (2010) MicroRNAs, the epigenetic memory and climatic adaptation in Norway spruce. New Phytol 187:1154–1169

Yazdani R, Nilsson JE, Plomion C, Mathur G (2003) Marker trait association for autumn cold acclimatation and growth rhythm in *Pinus sylvestris*. Scand J For Res 18:29–38

Zhang Y, Zhang S, Han S, Li X, Qi L (2012) Transcriptome profiling and in silico analysis of somatic embryos in Japanese larch (*Larix leptolepis*). Plant Cell Rep. doi:10.1007/s00299-012-1277-1