Tree Genetics & Genomes (2012) 8:573–582
DOI 10.1007/s11295-012-0497-4

REVIEW

Genomics of a phylum distant from flowering plants: conifers

Kermit Ritland

Received: 16 December 2011 / Accepted: 14 February 2012 / Published online: 30 March 2012

Abstract Conifers are evolutionarily distant from angiosperms, separated by 300 million years of evolution. The genomes of coniferous species are very large, among the largest of any nonpolyploid plant species. Their genomes are characterized by reduced evolutionary rate for coding genes, accumulation of noncoding DNA, and evolutionarily distance from angiosperms. I highlight both the advantages and disadvantages for conifers as model organism for genomics. With advances of new high-throughput sequencing technologies, we are at a watershed in conifer genomics.

Keywords Conifer genomics

Introduction

Conifers are a keystone species in most boreal forest ecosystems and even in some tropical ecosystems and have an economic market share in many countries (Neale and Kremer 2011). While conifers seem rich in species diversity (witness cedar, pine, spruce, fir, larch, redwood, cypress, juniper, yew), when compared to flowering plants, the number of conifer species is low (ca. 65 genera and 600 species in conifers vs. ca. 450 genera and 300,000 species in angiosperms). Conifers are also separated from angiosperms by over 300 million years of evolution (Bowe et al. 2000). Conifers also retain many features of primitive land plants and have extremely large genome sizes. The combination of large genome size, ecologic importance, and evolutionary distance makes conifers a unique phylum for studies of genome evolution.

"Genomics" involves the sum total of accounting for how all genes contribute to the phenotype and adaptation of an organism. Before proceeding, I note a new book, "Genetics, Genomics and Breeding of Conifers," where there are several relevant chapters to conifer genomics. These include the integration of molecular markers in breeding (Burdon and Wilcox 2011) transcriptomics (Mackay and Dean 2011), proteomics and metabolomics (Dauwe et al. 2011), genetic mapping (Ritland et al. 2011), and prospects for conifer genome sequencing (Morgante and De Paoli 2011). With this caveat, I proceed with a narrower objective below.

I review the current status of genome sequencing in conifers and its immediate application for the development of genotyping platforms, and the sequence-based studies of the unique nature of the conifer genome. I will then describe the current challenges and opportunities to enhance our understanding of conifer genomics These challenges and opportunities include (1) the genome size and complexity of conifers, (2) the rates of evolution and levels of diversity of conifers, (3) the uniqueness of gene content relative to angiosperms, and (4) the recent advances of sequencing technology that will give genome sequences of several large conifer genomes.

Current genomic resources for conifers

Like most crop and animal species research, from 1980 until about 2000, research in conifers was directed at identifying genetic markers, mapping these markers to create genetic maps, and using this information to identify genes for genetic transformation. Early investigations into the nature of conifer genomes focused on their large size putatively

Communicated by A. Abbott

A contribution to the Special Issue "The genomes of the giants: a walk through the forest of tree genomes"

K. Ritland (✉)
Department of Forest Sciences, University of British Columbia, Vancouver, British Columbia V6T1Z4, Canada
e-mail: Kermit.Ritland@ubc.ca

due to the presence of significant levels of repetitive elements in the genome Miksche and Hotta (1973). This ultracentrifugation study and others showed that conifer genomes are large and likely very repetitive (Kinlaw and Neale 1997). Other studies utilizing more recent genomic technologies substantiate the highly repetitive nature of the conifer genome and are presented below.

EST sequences

Expressed sequence tags (ESTs) are the best resource for characterizing the gene space of a large genome. While conifer genomes can be extremely large in size, data from expressed genes reduce the complexity to a manageable level. The pioneering research on sequencing mRNAs (ESTs) in conifers was done by Claire Kinlaw and associates (Kinlaw et al. 1996). Originally, this work was directed towards finding gene-based genetic markers for constructing genetic maps, but as sequencing throughput increased, random EST sequencing approaches were used as means to characterize conifer genome composition. In addition to providing possible biochemical functions encoded by individual random cDNAs, their work allowed identification of classes of genes actively transcribed in tissues from actively growing seedlings or developing phloem and cambium and also provided the first glimpse into the molecular nature of complex gene families within pine genomes.

The Forest Biotechnology Group at North Carolina State University extended this work; a first-pass sequence analysis for 1,097 sequences from differentiating xylem of loblolly pine identified 833 unique expressed sequences (Allona et al. 1998). Since these seminal studies were published, a large number of ESTs and unigene sets have been collected for several important coniferous species. The numbers of currently available ESTs for conifers and representative angiospersms are given in Table 1. Loblolly pine (*Pinus taeda*), white spruce (*Picea glauca*), and Sitka spruce (*Picea sitchensis*) dominate the group. *P. glauca* and *P. sitchensis* are at opposite ends of the *Picea* genus, with about 4 % EST nucleotide divergence (Ritland, unpublished data) so that their numbers cannot be combined. *Picea abies* (Norway spruce), *P. glauca* (white spruce), and *P. sitchensis* (Sitka spruce) all have large EST collections, and their joint analysis should reveal lineage-specific insights into conifer evolution, as the three species are about equally related (Ritland, unpublished data).

Another sequence resource is full-length cDNAs (FL-cDNAs), which span the entire length of coding sequences plus possibly 5′ and 3′ noncoding regions. In terms of functional characterization and marker development, FL-cDNAs are best suited for deciphering the conifer genome. Additionally, "unigene sets" can be identified from collections of ESTs; these are groups of singleton ESTs and contigs of ESTs which mutually are inferred to be distinct genes (ncbi.nlm.nih.gov/unigene). The use of conifer FL-cDNAs has been instrumental in drawing inferences about conifer genome evolution as presented in Fig. 2.

Table 1 EST numbers and genome sizes (*C* value) for conifers and some angiosperms

	No. ESTs	*C* value
Pines		
P. taeda	328,662	22.10
P. contorta	40,483	18.90
Pinus banksiana	36,379	17.20
P. pinaster	34,044	24.35
Pinus radiata	34,044	24.35
P. radiata	8,717	22.00
Pinus densiflora	3,316	21.50
Spruce		
P. glauca	313,110	20.20
P. sitchensis	186,637	N.A.
Picea engelmannii × *P. glauca*	28,174	19.45
P. abies	14,345	20.01
Other gymnosperm phyla		
C. japonica	56,645	11.05
Ginkgo biloba	21,590	9.95
Gnetum gnemon	10,724	3.87
Welwitschia mirabilis	10,129	7.20
Major angiosperms of interest		
Z. mays	2,019,105	2.73
Arabidopsis thaliana	1,529,700	0.16
O. sativa	1,251,304	0.50
T. aestivum	1,071,335	17.33
Brassica napus	643,884	1.15
Vitis vinifera	362,392	0.43
Mimulus guttatus	231,095	0.37

Numbers of EST that were present in GenBank as of February 1, 2011 and were for taxa with more than 1,000 deposited; *C*-genome sizes are from Leitch et al. (2001) and from Kew Plant *C* values database (data.kew.org/cvalues). *C* values approximately equal the number of gigabases (billions of bases) in the genome

Bacterial artificial chromosomes

The size of the conifer genome warrants millions of bacterial artificial chromosomes (BACs) to cover the genome; in this light, BACs are only of value for characterization of genome structure (clustering of genes, nature of repetitive DNA). Recent achievements in BAC conifer genomics include a 1.8 million clone library that was constructed for loblolly pine (100-kb average insert size) (Magbanua et al. 2011). In bald cypress, a 600,000 clone library was constructed (113 average insert size) (Liu et al. 2009). In white spruce, a 1.1 million clone library was constructed (140-kb average insert

size) (Hamberger et al. 2009). In maritime pine (*Pinus pisaster*), an arrayed library of 72,192 clones was achieved (average insert size of 107 kb). (Bautista et al. 2007).

SNPs from genomic resources in conifers

The first major database for conifer SNPs was that for white spruce, where an automated in silico approach found 12,264 SNPs from 6,459 EST contigs (Pavy et al. 2006). These and other SNPs discovered in later studies are currently being used on an Illumina 13,680 SNP format Illumina chip for various studies (Bousquet, personal communication). In "ADEPT2" (dendrome.ucdavis.edu/NealeLab), a unigene set of roughly 20,500 contigs was identified in loblolly pine, from which 7,424 amplicons were successfully resequenced. Of those, 6,178 amplicons yielded high-quality SNP data from which a panel of SNPs for current use are available and being utilized in the laboratory of D. Neale at UC Davis. So far, this work represents the first initial practical genomic-scale application of EST resources.

More interestingly, these primers were tested in five additional conifer species, as listed in Table 2. As expected, since the primers were designed from loblolly pine, SNPs were more easily transferred to more closely related species; 84 % of the primers were successful in the closely related *Pinus radiatia* (a hard pine as is loblolly) but only 30 % for *Pinus lambertiana* (a soft pine in the other major section of *Pinus*). Both spruce and Douglas fir had about 10 % success. While this might sound low, this still provides hundreds of SNPs that have a value for comparing the overall genome structure and evolution in these related species and for providing cross-species markers for breeding applications. SNP transfer success with more distantly related species such as redwood, which is not a pine family member, was very low, 0.5 %.

Chloroplast genomes

Chloroplast genomes are relatively conserved among plants and are small (100–150 kb) with few genes (ca. 140). These genes are mainly involved with major metabolic activities.

Table 2 Ability of primers designed in *P. taeda* to amplify other conifer species (data of D. Neale and associates)

Species	No. of successful resequenced amplicons	Percent total
P. taeda	7,424	100.0
P. radiata	6,429	84.2
P. lambertiana	2,234	30.1
P. abies	1,024	13.8
P. menziesii	750	10.1
Sequoia sempervirens	40	0.5

Classically, the chloroplast genome has been used for many studies of plant systematics. rbcL and matK seem to be the current focus for "DNA barcoding" (Group et al. 2009). As of September 2010, complete chloroplast sequences have been deposited in GenBank for 164 angiosperm and 12 conifer species (*Pinus koraiensis*, *Pinus krempfii*, *Pinus gerardiana*, *Pinus contorta*, *Pinus nelsonii*, *Pinus monophylla*, *Pinus longaeva*, *P. lambertiana*, *Pinus thunbergii*, *P. sitchensis*, *Keteleeria davidiana*, and *Cryptomeria japonica*).

Seven of the pine and spruce chloroplast genomes were done in a single study with the Solexa sequencer (Cronn et al. 2008). This study obtained a mean coverage per genome of 55× to 186×, with sequence runs made from pools of four species. With this approach, genomes were not completely assembled; the number of contigs ranged from 9 to 183, and assembly strategy relied upon previously sequenced conifer chloroplast genomes. This study previews what can be done at the genome level in spruce and pine.

Mitochondrial genomes

Unlike the chloroplast genome, the mitochondrial genome of plants is highly variable in organization. This genome can be more than 100 times larger in plants than in animals and is structurally complex due to frequent recombination (Knoop et al. 2011). For conifers, a complete genome sequence exists only for *Cycas taitungensis* (fern palm) (Chaw et al. 2008), with a size of 414.9 kbp that is similar to angiosperm mitochondrial genome sizes but much larger than those of Charophytes and Bryophytes. Unlike the chloroplast genome of conifers, the conifer mitochondrial genome is as yet uncharacterized.

Transcriptome and protein profiling

Transcriptome profiling in forest trees, using a variety of microarray technologies, is a very active area of research. In conifers, most profiling studies are focused on growth, wood properties, biotic stress, and abiotic stress. As *Pinus* and *Picea* have the largest EST collections, most published studies have focused upon resources from these species. Transcript profiling can be done digitally by comparing EST abundance among libraries constructed from RNAs isolated from somatic embryogenic tissues (Cairney et al. 2000), from roots responding to water stress (Lorenz et al. 2006), or with cDNA microarrays constructed from tissues responding to defoliation by insects (Ralph et al. 2006). EST and FL-cDNA databases have also been very useful for large-scale identification of expressed spruce transcripts (Lippert et al. 2005). Recent and more comprehensive reviews of the profiling of transcripts, metabolites, and proteins are given by Dauwe et al. (2011) and Mackay and Dean (2011).

Websites for conifer genomics

Databases are needed to deposit and manage genome resources. Unlike species such as *Arabidopsis* with the established TAIR database (arabidopsis.org), there is no single comprehensive database for conifers. Currently, the most complete database for conifers (and tree species) is Dendrome (dendrome.ucdavis.edu); others include the Conifer Genomics Network (pinegenome.org) and Conifer-EST (Liang et al. 2007). The major goal of a current European Union project (foresttrac.eu) is how to coordinate databases between Europe and North America.

Obstacles for conifer genomics

The presence of large, repetitive, and often polyploid genomes in many plant species presents challenges for genomics and genome sequencing (Paterson 2006). Before discussing conifers, we must note recent achievements made in two crop plant species with large genomes: maize and wheat (*Zea* and *Triticum*, respectively, in Table 1). In maize, a draft genome sequence found nearly 85 % of the genome to be composed of transposable elements (Schnable et al. 2009). The even larger genome of wheat was recently examined by sequencing different regions of its largest chromosome; gene distribution was not random, with 75 % of them clustered into small islands containing three genes on average (Choulet et al. 2010). But concomitant with the writing of this review, the introduction of next-generation sequencing technologies is changing the whole-genome sequencing landscape for many complex genomes, and thus, many of the previous obstacles in obtaining whole-genome sequences in conifers may no longer exist.

Genome size and traditional hierarchal sequencing approaches

Conifers are famous for their large genome size. Genome size can be roughly gauged by *C* values, as measured by flow cytometry. Table 1 gives *C* values for conifers derived in major gymnosperm EST sequencing projects and those of some representative angiosperms.

As evident in Table 1, the genomes of spruce and pine have sizes of 19–24 billions of bases (gigabases or gb). This is over six times the size of the human genome but at least comparable to the genomes of some angiosperms such as *Zea mays* and *Triticum aestivum*. With a conifer genome size of 20 gb, with a BAC insert of 120 kb, just for a 1× coverage, 166,000 BAC clones are needed. The size of conifer genomes precludes traditional "hierarchical" sequencing projects, which use tiled BAC maps, since this would require a prohibitive number of large insert clones for fingerprinting and tiling path construction for sequencing. In conifers, no library has been fingerprinted for the purpose of constructing tiling paths.

Genome complexity

Duplicate genes and nearly identical paralogues Southern hybridization patterns suggested that genomes of gymnosperms include complex families of genes (Kinlaw and Neale 1997). The presence of multiple hybridizing fragments to probe in analyses of conifer DNAs compared to single hybridizing fragments in parallel samples of representative angiosperms was considered an evidence to support this hypothesis. The multiple hybridizing fragments may represent nonfunctional pseudogenes or duplicated loci. García-Gil (2008) showed that in a specific gene family, the phytochromes, such genes add to complexity of the phytochrome family in *Pinus sylvestris*, and psuedogenes evolve neutrally, while functional genes have signatures of natural selection. Another recent study that compared genome complexity in conifers to angiosperms involved *C. japonica*. Futamura et al. (2008) found that the numbers of transcripts that encoded certain protein families or domains, such as NAD-dependent epimerase/dehydratase, the C3HC4-type zinc finger, the WD domain, aspartyl proteases, and aldo/keto reductases, were larger than those that encoded the corresponding protein families or domains in the *Arabidopsis* genome. They found an increased complexity of gene families in *C. japonica* as compared to *Arabidopsis*.

Sequencing of BACs can be much more revealing about the underlying structure of conifer genomes. Sanger sequencing of 10 loblolly pine BACs showed that the presence of both known and novel conservative repeats comprised only a small portion of the genome (Kovach et al. 2010). Computational annotation of the 10 BACs predicted three putative protein-coding genes and at least fifteen likely pseudogenes in nearly 1 mb of sequence. They found three conifer-specific LTR retroelements in the BACs and tentatively identified at least 15 others based on evidence from the distantly related angiosperms. Hamberger et al. (2009) found high-complexity repeats in two BACs from a white spruce library. Compared to angiosperms, in these two BACs that were sequenced, transposable element content was about 20 %, and high-complexity repeats comprised about 40 % of the sequences.

Implications for marker development The highly repetitive and large genome size of conifers has been a major obstacle for the development of genetic markers; however, large-scale EST collections have allowed more efficient development of markers for conifers. With traditional methods of developing microsatellites (cloning of simple sequence

repeat (SSR) motifs), the proportion of positive clones that actually lead to a reproducible, clearly resolved, diverse SSR loci is very low for conifers, about 1–4 per 100 positives (Ritland, personal communication). To avoid problems posed by the large and repetitive conifer genome, microsatellites can be developed from ESTs (denoted EST-SSRs). ESTs can also reveal polymorphisms in related taxa (Ellis and Burke 2007). EST-SSRs have been found in loblolly pine (Chagné et al. 2004) and spruce (Rungis et al. 2004), and EST-SSRs from loblolly pine amplified products in lodgepole pine (Liewlaksaneeyanawin et al. 2004). However, one disadvantage to the use of EST-SSRs is that as gene sequences, they exhibit less polymorphism than genomically derived SSRs. For example, spruce EST-SSRs had 9 % less heterozygosity than genomic-derived SSRs. EST databases can also identify "conserved orthologous set" (COS) markers (Fulton et al. 2002). Using current EST databases, a large set of COS markers were identified for loblolly pine, white spruce, Douglas fir, and sugi (Krutovsky et al. 2006). A wet-lab study however found that COS markers do suffer from reduced diversity; average nucleotide heterozygosity for 931 tested primers was ca. 0.04 % (Liewlaksaneeyanawin et al. 2009) about 10 times lower for other genes in loblolly pine (Brown et al. 2004b). Conifers also pose the same problems for next-generation genotyping methods.

A recent promising technology that is very appropriate for conifers is restriction-site associated DNA (RAD). It uses next-generation DNA sequencing to generate thousands of genetic markers across a genome, multiplexing tens of individuals in a single sequencing lane. DNA fragments assayed by RAD are generated by restriction fragment enzyme digests. "Radcounter" (wiki.ed.ac.uk) allows one to estimate the number of "RADSeq" sites (loci), and rare cutters should be used for a conifer genome. NotI is by far the best to achieve rare cutting with just 35 K loci expected in the 20 gb conifer genome.

Getting around the repetitive genome of conifers

A number of "reduced-representation sequencing" approaches have been used to enrich for the gene space by removing repetitive DNA. There are two gene-enrichment approaches: methylation filtration and high-Cot sequencing (Barbazuk et al. 2005). Springer and colleagues (Springer et al. 2004) evaluated the ability of these two strategies to reconstruct 78 full-length cDNAs in maize. Both methyl filtration and high-Cot enrichment methods provided a sevenfold to eightfold increase in gene discovery rates as compared to random genomic sequencing. Wheat researchers also realize that prior to new sequencing technologies, sequencing 17 gb of DNA requires a more targeted approach (Lamoureux et al. 2005). They concluded that Cot filtration was twice as efficient as methyl filtration at enriching for gene sequences. Although these approaches have been used in the past, next-generation sequencing technologies are expected to eclipse such technologies in the future.

Opportunities for conifer genomics

Slower rate of sequence evolution and lower diversity in conifers

Within the pine family, most of the ca. 240 species have 12 chromosomes (the exception being Douglas fir with 13), and polyploidy is rare in conifers, except in the *Cupressaceae* (redwoods, junipers, cedars). At the macrosynteny level, there is much conservation of genetic map marker order and content (Krutovsky et al. 2004). At the microsynteny level, as inferred by EST sequence comparisons of loblolly pine with white spruce, nucleotide substitution rates appear to be an order of magnitude lower in the pine family compared to angiosperms, with an average synonymous substitution rate of about 4×10^{-10} per year which is 10 times slower than that of most Angiosperms (Buschizzo et al. 2012). Low levels of nucleotide diversity have also been found in studies of single-nucleotide polymorphisms, a level consistent with a low mutation rate of 1.17×10^{-10} per year (Brown et al. 2004a). These results suggest that genomic information can be transferred among coniferous species and that species such as spruce and pine would have the same degree of sequence similarity and microsynteny as angiosperms separated by 10–20 million years of evolution.

Ancient retroelements and genome assembly

By comparing the two ends of a retrotransposon, which are genetically identical at the time of insertion, the date of transposition can be inferred by the sequence divergence of the two ends, assuming a molecular clock. In an examination of four BAC sequences from spruce, De Paoli et al. (unpublished data) demonstrated that the spruce genome was shaped by the mobilization of several retrotransposon families. They inferred that there were two waves of colonization in spruce, involving the copia and gypsy elements: 50–80 mya for copia and 5–40 mya for gypsy. This is far older than any reported angiosperm such as corn (which demonstrated retroelement movement only 10–15 mya). These researchers suggest that the retention of ancient repetitive features contributes to conservative gymnosperm genome evolution. These results bode well for genome assembly, as most "related" repetitive elements in the conifer genome will be substantially diverged eliminating problems of erroneous sequence merges due to identical

repetitive element sequences. Despite the overall low rate of conifer genome evolution, this implies that paralogy of transposons will not pose a problem for assembly of future whole-genome shotgun sequences for conifer genomes.

Low amounts of linkage disequilibrium and population structure for association mapping

The most important downstream application of genome sequences is identifying genetic variants associated with genes critical for genetic improvement and management of adaptive diversity in the face of climate change (Neale and Kremer 2011). In loblolly pine, a survey across 18 kb found that linkage disequilibrium declined within several kilobases (Brown et al. 2004a). The same pattern was found by Heuertz and colleagues in Norway spruce (Heuertz et al. 2006). Low linkage disequilibrium allows much greater power to directly associate single nucleotide polymorphisms with phenotypic traits. In contrast to the limited population structure of conifers, the presence of population structure in many crop plants, particularly inbred species, reduces the power for detecting marker–trait associations. A good example is rice, where both varieties and inbred lines contribute to population structure and family relatedness and make association studies more complicated (Wen et al. 2009).

The megagametophyte and a haploid library

A unique feature of conifers is the haploid tissue, the "megagametophyte." This is a small nutritive tissue derived from the maternal parent and has been used extensively in conifer isozyme population genetics. For genomics, most notably, it was the tissue used for the generation of the first RAPD genetic map of a conifer (Tulsieram et al. 1992). It would be ideal to construct BAC libraries and perform genome sequencing on this tissue (to avoid mistaken paralogy due to heterozygosity). While the amount of extractable DNA is small in spruce, the Swedish Norway spruce project has successfully sequenced megametophytes for their project, and the PineRefSeq projet led by David Neale has used the larger megametophyte of pine for sequencing as well. As this activity is currently in flux (as of January 2012), for further information, contact Stefan Jansson (UPSE.SE) and David Neale (UCD.EDU) for updates on this activity.

To get around the problem of small tissue available for DNA isolation, one can resort to tissue culture. Tissue cultures of haploid gametophytes have been successfully generated in spruce (Simola and Santanen 1990) and larch (Aderkas and Bonga 1993), but genetic instability (loss of haploid lines within cell cultures) at least over the longer term was observed in larch, such as various degrees of polyploidy and aneuploidy (von Aderkas and Anderson 1993). Thus, the risk of introducing chromosome abnormalities into genomic studies is too high with the current tissue culture techniques.

Unique gene space of conifers

Early, it was recognized that comparison of conifers' ESTs with sequences from angiosperms could be used as a route to gain information about the evolution of higher plant genomes (Allona et al. 1998). Various methods for focusing on the "gene space" of the genome have been investigated and deployed in conifers.

There have been a few speculations about how many genes in conifers are "unique" to this phylum. Gene

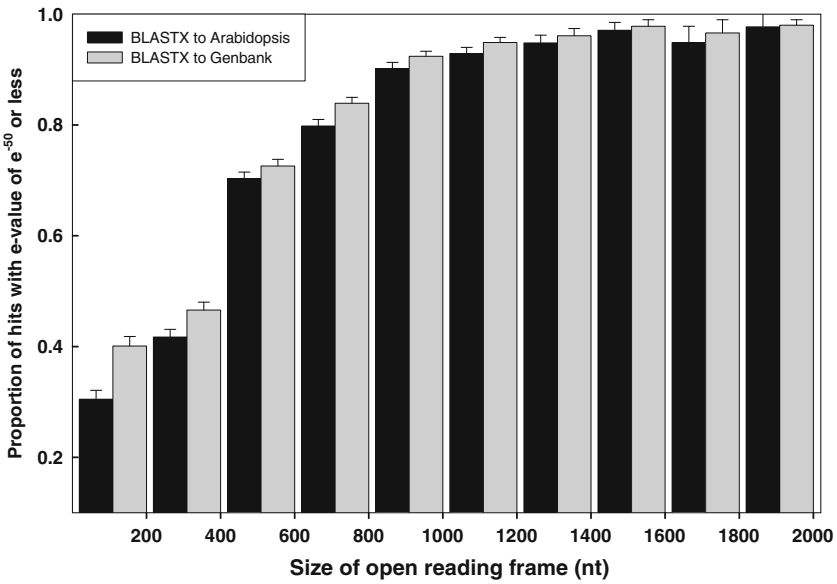

Fig. 1 Ability to annotate (via BLASTX) complete cDNAs from spruce to *Arabidopsis* and to all organisms in GenBank, as a function of open reading frame

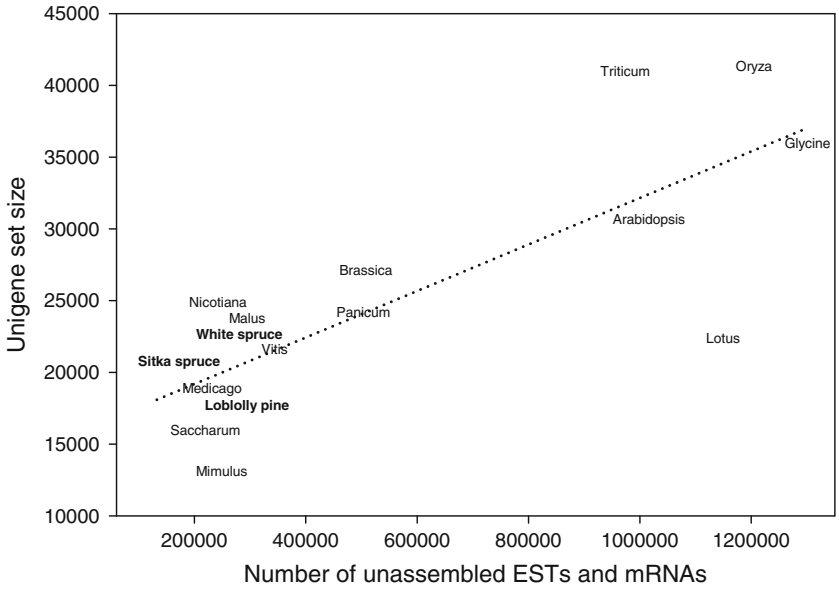

Fig. 2 Relationship of conifer unigene sets (white spruce, Sitka spruce, loblolly pine) to other unigene sets from representative plant taxa

annotation is critical for defining unique gene space in different species. Simply doing a BLAST analysis against published sequences in GenBank has pitfalls. Poor hits may be due to either rapid gene evolution or short sequence length, especially in light of the evolutionary divergence between conifers and annotated angiosperms. To illustrate the problem of gene length, we used the 6,464 complete cDNA collection of spruce full-length cDNAs (Ralph et al. 2008) to perform a BLASTX search (which compares putative protein coding codons) against the *Arabidopsis* and total GenBank databases. At a threshold of $1e^{-50}$, Fig. 1 shows that any coding sequence below about 800 bp (333 amino acids) is difficult to annotate to angiosperms, as evidenced by smaller proportions of hits below about 800 bp, compared to 1,000 bp and above. It might be that the smaller genes evolve more quickly. The asymptote suggests that the actual number of genes that are unique to conifers is about 5 %, which is much less than that suggested by a much earlier study of Allona et al. (1998) which found only 42 % of ESTs from loblolly pine to show strong similarity to public databases at rather low *e* values ($1e^{-5}$). Further research is needed to disentangle the biological role of rapid gene evolution from statistical artifacts of small sequence lengths.

The uniqueness of the conifer genome can also be gauged by comparing the size of unigene sets among angiosperms and conifers. We examined unigene sets assembled by GenBank (ncbi.nlm.nih.gov/unigene) as of February 1, 2011 and removed the effect of the number of ESTs and mRNAs used to infer the unigene set by regressing the number of ESTs and mRNAs used vs. actual numbers of unigenes inferred. The results, in Fig. 2, show that (1) the unigene sizes of conifers (white spruce, Sitka spruce, loblolly pine) are all quite similar, about 20,000 genes on average, and these results are not confounded by the number of ESTs and mRNAs used; (2) unigene sizes for conifers are close to many angiosperm species, such as *Malus* (apple), *Vitis* (grape), and *Medicago* (alfalfa); and (3) there are a number of genomes in angiosperms with much larger unigene sets (wheat, rye, soybean, due to polyploidy). These data suggest that while conifers harbor many repetitive elements and psuedogenes, the number of expressed genes in conifers is quite similar to many angiosperms species.

Conclusions

"Next-generation" sequencing is the new wave of genomics. These advances in DNA sequencing involve parallel sequencing of millions of oligonucleotides at one time, resulting in gigabases of sequences in a few days. Besides impacting the whole of plant and animal genomics and making (in my opinion) the activities of EST collections, microarrays, and SNP discovery, members of the "past generation," this new sequencing technology will make the greatest relative impact on conifers. With the typical 20 billion base conifer genome, for example, the Illumina Hiseq 2000 can sequence at a current capacity of 60 billion bases per slide, meaning each slide can do a 3× coverage of a conifer. This cost is a fraction of a percentage compared to technologies available 10 years ago.

Bioinformatics for genome assembly now becomes the major issue. In the past 10 years, the method of assembly via the "de Bruijn graph" has become predominant (Li et al. 2010b) and as well as the algorithms to handle the massive numbers of contigs (Bonfield and Whitwham 2010). While the reads are short (ca. 100 nucleotides), they are getting longer, and paired reads (mate pairs) can be generated, with

the reads separated by several hundred nucleotides which potentially can allow spanning of unreadable regions and repetitive elements (Shendure and Ji 2008). Short read assemblers are available, such as SSAKE (Warren et al. 2007) and ABySS (Simpson et al. 2009). The major issue is distributing work among processors and the available memory space in the final assembly. ABySS can efficiently assemble conifer-sized genomes at the first stage, through distributed algorithms among processors (128 at last count), but final stages of assembly require a computer with enormous RAM. The panda genome sequence (Li et al. 2010a) is an excellent example of a sequencing strategy for a conifer. In the panda genome project, a variety of library sizes were used to shotgun the genome without resorting to BAC tiling paths.

To provide longer contiguity and sequence scaffolds, new "third-generation" technology is required. Such sequencing technologies should allow us to identify differentiation at the "pan-genomic" level. By scanning nucleotide divergence between contrasting populations, we can identify specific genomic regions involved with phenotypic species differentiation (Neale and Kremer 2011). In a larger time frame, we might be able to fully catalog the genetic changes that have occurred during conifer evolution. This can be done by comparing whole-gene sequences of spruce to pine to sister groups of conifers and to representative angiosperm species. Ever since Darwin, it was speculated that the *Gnetales* (*Gnetum* spp.) and various fossil groups were sister to angiosperms. Bowe et al. (2000) using chloroplast rbcL, nuclear 18S rDNA, and three mitochondrial genes, demonstrate this relationship. Comparisons with these sister groups could reveal the uniqueness of conifers.

In the last months of 2011, there has been an avalanche of genome projects funded for conifers. As of February 2012, genome sequencing projects have been initiated in at least seven conifer species (*P. taeda, P. lambertiana, Pinus pinaster, P. sylvestris, Psuedotsuga menziesii, P. abies, P. glauca*). As noted above, this has been aided by (1) next-generation sequencing, (2) new strategies for sequencing, and (3) advances in the bioinformatics of assembling large genomes. It is difficult to write a review in such changing times.

Acknowledgments I thank David Neale and Bert Abbott for their comments on drafts, John MacKay for discussions about conifer genomics, Steven Jones and Inanc Birol for teaching me the latest about sequence assembly, Carol Ritland for support at both work and home, and Genome BC/ Genome Canada for their support of conifer genome projects.

References

Aderkas P, Bonga JM (1993) Plants from haploid tissue culture of *Larix decidua*. TAG Theor Appl Genet 87:225–228

Allona I, Quinn M, Shoop E, Swope K, Cyr SS, Carlis J, Riedl J, Retzel E, Campbell MM, Sederoff R, Whetten RW (1998) Analysis of xylem formation in pine by cDNA sequencing. Proc Natl Acad Sci U S A 95:9693–9698

Barbazuk WB, Bedell JA, Rabinowicz PD (2005) Reduced representation sequencing: a success in maize and a promise for other plant genomes. BioEssays 27:839–848

Bautista R, Villalobos DP, Díaz-Moreno S, Cantón FR, Cánovas FM, Gonzalo Claros M (2007) Toward a *Pinus pinaster* bacterial artificial chromosome library. Ann For Sci 64:855–864

Bonfield JK, Whitwham A (2010) Gap5—editing the billion fragment sequence assembly. Bioinformatics 26:1699–1703

Bowe LM, Coat G, dePamphilis CW (2000) Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and *Gnetales*' closest relatives are conifers. Proc Natl Acad Sci U S A 97:4092–4097

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A 101:15255–15260

Burdon R, Wilcox P (2011) Integration of molecular markers in breeding. In: Plomion C, Bousquet J (eds) Genetics, genomics and breeding of conifers. Science Publishers, Edenbridge

Buschizzo E, Ritland C, Bohlmann J, Ritland K (2012) Slow but not low: genomic comparisons reveals slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evol Biol 12:8

Cairney J, Xu N, Mackay J, Pullman J (2000) Special symposium: in vitro plant recalcitrance transcript profiling: a tool to assess the development of conifer embryos. In Vitro Cell Dev Biol Plant 36:155–162

Chagné D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera M, Vendramin G, Garcia V, Frigerio JM, Echt C, Richardson T, Plomion C (2004) Cross-species transferability and mapping of genomic and cDNA SSRs in pines. TAG Theor Appl Genet 109:1204–1214

Chaw S-M, Chun-Chieh Shih A, Wang D, Wu Y-W, Liu S-M, Chou T-Y (2008) The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol 25:603–615

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier M-C, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36:e122

Dauwe R, Robinson A, Mansfield S (2011) Recent advances in proteomics and metabolomics in gymnosperms. In: Plomion C, Bousquet J (eds) Genetics, genomics and breeding of conifers. Science Publishers, Edenbridge

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99:125–132

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

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

García-Gil M (2008) Evolutionary aspects of functional and pseudogene members of the phytochrome gene family in Scots pine. J Mol Evol 67:222–232

Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim K-J, Kress WJ, Schneider H, van AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim Y-D, Lahaye R, Lee H-L, Long DG, Madriñán S, Maurin O, Meusnier I, Newmaster SG, Park C-W, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi D-K, Little DP (2009) A DNA barcode for land plants. Proc Natl Acad Sci 106:12794–12797

Hamberger B, Hall D, Yuen M, Oddy C, Hamberger B, Keeling C, 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

Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics 174:2095–2105

Kinlaw CS, Neale DB (1997) Complex gene families in pine genomes. Trends Plant Sci 2:356–359

Kinlaw CS, Ho T, Gerttula SM, Gladstone E, Harry DE, Quintana L, Baysdorfer C (1996) Gene discovery in loblolly pine through cDNA sequencing. In: Ahuja MR, Boerjan W, Neale DB (eds) Somatic cell genetics and molecular genetics of trees. Kluwer Academic Publishers, Dordrecht, pp 175–182

Knoop V, Volkmar U, Hecht J, Grewe F (2011) Mitochondrial genome evolution in the plant lineage. In: Kempken F (ed) Plant mitochondria. Springer, New York, pp 3–29

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

Krutovsky KV, Troggio M, Brown GR, Jermstad KD, Neale DB (2004) Comparative mapping in the *Pinaceae*. Genetics 168:447–461

Krutovsky K, Elsik C, Matvienko M, Kozik A, Neale D (2006) Conserved ortholog sets in forest trees. Tree Genet Genomes 3:61–70

Lamoureux D, Peterson DG, Li W, Fellers JP, Gill BS (2005) The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). Genome 48:1120–1126

Leitch IJ, Hanson L, Winfield M, Parker J, Bennett MD (2001) Nuclear DNA C-values complete familial representation in gymnosperms. Ann Bot 88:843–849

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC-C, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT-Y, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T-W, Yiu S-M, Liuand S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK-S, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J (2010a) The sequence and de novo assembly of the giant panda genome. Nature 463:311–317

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010b) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Liang C, Wang G, Liu L, Ji G, Fang L, Liu Y, Carter K, Webb J, Dean J (2007) ConiferEST: an integrated bioinformatics system for data reprocessing and mining of conifer expressed sequence tags (ESTs). BMC Genomics 8:134

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. TAG Theor Appl Genet 109:361–369

Liewlaksaneeyanawin C, Zhuang J, Tang M, Farzaneh N, Lueng G, Cullis C, Findlay S, Ritland C, Bohlmann J, Ritland K (2009) Identification of COS markers in the *Pinaceae*. Tree Genet Genomes 5:247–255

Lippert D, Zhuang J, Ralph S, Ellis DE, Gilbert M, Olafson R, Ritland K, Ellis B, Douglas CJ, Bohlmann J (2005) Proteome analysis of early somatic embryogenesis in *Picea glauca*. Proteomics 5:461–473

Liu W, Magbanua ZV, Orzkan S, Chouvarine P, Bartlett BD, Peterson DG (2009) BAC libraries for two distantly related conifers, loblolly pine and bald cypress. In: Plant and animal genomes XVII conference, San Diego, USA

Lorenz WW, Sun F, Liang C, Kolychev D, Wang H, Zhao X, Cordonnier-Pratt M-M, Pratt LH, Dean JFD (2006) Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries. Tree Physiol 26:1–16

Mackay J, Dean J (2011) Transcriptomics. In: Plomion C, Bousquet J (eds) Genetics, genomics and breeding of conifers. Science Publishers, Edenbridge

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

Miksche JP, Hotta Y (1973) DNA base composition and repetitious DNA in several conifers. Chromosoma 41:29–36

Morgante M, De Paoli E (2011) Toward the conifer genome sequence. In: Plomion C, Bousquet J (eds) Genetics, genomics and breeding of conifers. Science Publishers, Edenbridge

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111–122

Paterson AH (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. Nat Rev Genet 7:174–184

Pavy N, Parsons L, 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 7:1–14

Ralph SG, Yueh H, Friedmann M, Aeschliman D, Zeznik JA, Nelson CC, Butterfield YSN, Kirkpatrick R, Liu J, Jones SJM, Marra MA, Douglas CJ, Ritland K, Bohlmann J (2006) Conifer defence against insects: microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. Plant Cell Environ 29:1545–1570

Ralph S, Chun H, Kolosova N, Cooper D, Oddy C, Ritland C, Kirkpatrick R, Moore R, Barber S, Holt R, Jones S, Marra M, Douglas C, 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

Ritland K, Krutovsky K, Tsumura Y, Pelgas B, Bousquet J (2011) Genetic mapping in conifers. In: Plomion C, Bousquet J (eds) Genetics, genomics and breeding of conifers. Science Publishers, Edenbridge

Rungis D, Bérubé Y, Zhang J, Ralph S, Ritland CE, Ellis BE, Douglas C, Bohlmann J, Ritland K (2004) Robust simple sequence repeat (SSR) markers for spruce (*Picea* spp.) from expressed sequence tags (ESTs). Theor Appl Genet 109:1283–1294

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Simola LK, Santanen A (1990) Improvement of nutrient medium for growth and embryogenesis of megagametophyte and embryo callus lines of *Picea abies*. Physiol Plant 80:27–35

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

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

Tulsieram LK, Glaubitz JC, Kiss G, Carlson JE (1992) Single tree genetic linkage mapping in conifers using haploid DNA from megagametophytes. Nat Biotechnol 10:686–690

von Aderkas P, Anderson P (1993) Aneuploidy and polyploidization in haploid tissue cultures of *Larix decidua*. Physiol Plant 88:73–77

Warren RL, Sutton GG, Jones SJM, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501

Wen W, Mei H, Feng F, Yu S, Huang Z, Wu J, Chen L, Xu X, Luo L (2009) Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L.). Theor Appl Genet 119:459–470