Tree Genetics & Genomes (2012) 8:1477–1485
DOI 10.1007/s11295-012-0547-y

SHORT COMMUNICATION

Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species

W. Walter Lorenz · Savavanaraj Ayyampalayam · John M. Bordeaux · Glenn T. Howe · Kathleen D. Jermstad · David B. Neale · Deborah L. Rogers · Jeffrey F. D. Dean

Received: 27 March 2012 / Revised: 20 June 2012 / Accepted: 5 July 2012 / Published online: 26 July 2012

Abstract Conifers comprise an ancient and widespread plant lineage of enormous commercial and ecological value. However, compared to model woody angiosperms, such as *Populus* and *Eucalyptus*, our understanding of conifers remains quite limited at a genomic level. Large genome sizes (10,000–40,000 Mbp) and large amounts of repetitive DNA have limited efforts to produce a conifer reference genome, and genomic resource development has focused primarily on characterization of expressed sequences. Here, we report the completion of a conifer transcriptome sequencing project undertaken in collaboration with the U.S. DOE Joint Genome Institute that resulted in production of almost 12 million sequence reads. Five loblolly pine (*Pinus taeda*) cDNA libraries representing multiple tissues, treatments, and genotypes produced over four million sequence reads that, along with available Sanger expressed sequence tags, were used to create contig assemblies using three different assembly algorithms: Newbler, MiraEST, and NGen. In addition, libraries from 11 other conifer species, as well as one member of the Gnetales (*Gnetum gnemon*), produced 0.4 to 1.2 million sequence reads each. Among the selected conifer species were representatives of each of the seven phylogenetic families in the Coniferales: Araucariaceae, Cephalotaxaceae, Cupressaceae, Pinaceae, Podocarpaceae, Sciadopityaceae, and Taxaceae. Transcriptome builds for each species were generated using each of the three assemblers. All contigs for every species generated using each assembler can be obtained from Conifer DBMagic, a public database for searching, viewing, and downloading contig sequences, the associated sequence reads, and their annotations.

Keywords Coniferales · *Pinus* · Transcriptome · Database · Annotation · Gene models · Comparative phylogenomics

Communicated by R. Sederoff

Electronic supplementary material The online version of this article (doi:10.1007/s11295-012-0547-y) contains supplementary material, which is available to authorized users.

W. W. Lorenz · J. M. Bordeaux · J. F. D. Dean
Warnell School of Forestry and Natural Resources,
University of Georgia,
Athens, GA 30602, USA

S. Ayyampalayam
Department of Plant Biology, University of Georgia,
Athens, GA 30602, USA

G. T. Howe
Department of Forest Ecosystems and Society,
Oregon State University,
Corvallis, OR 97331, USA

K. D. Jermstad
USDA Forest Service, Pacific Southwest Research Station,
Institute of Forest Genetics,
Placerville, CA 95667, USA

D. B. Neale · D. L. Rogers
Department of Plant Sciences, University of California,
One Shields Avenue,
Davis, CA 95616, USA

J. F. D. Dean (✉)
Department of Biochemistry and Molecular Biology,
Institute of Bioinformatics, University of Georgia,
Athens, GA 30602, USA
e-mail: jeffdean@uga.edu

Introduction

Conifers (Division: Pinophyta) represent an ancient and diverse branch of the gymnosperms, and many conifer species are notable for their adaptations to extreme and highly stressful environments. Among the more than 500 extant species are conifers known for their great size and others for their extreme longevity (Farjon 2008). Some conifer species dominate modern ecosystems that are repositories for large amounts of terrestrial sequestered carbon, while others exist in small and fragmented populations—some threatened with extinction. Conifer forests are among the most productive in terms of annual lignocellulosic biomass and coniferous trees are the preferred feedstock for much of the forest products industry, one of the most energy-intensive manufacturing sectors of the U.S. economy (Bowyer et al. 2007). Breeding programs for conifer genetic improvement have been in existence for more than 50 years, but progress has been slow due to constraints of both time (slow growth to sexual and economic maturity) and space (large size of the trees) (Plomion et al. 2012). Over-exploitation, rapid climate change, and exotic forest pests are threatening many conifer populations, but a general dearth of genomic resources and tools limits our capacity to use advanced genetic approaches to mitigate some of these threats.

The advent of highly cost-effective sequencing technologies has led to rapid increases in available sequence data for model and non-model organisms alike. Because of longer read lengths, Roche 454 pyrosequencing offers a particularly useful system for de novo transcriptome assembly and gene discovery in organisms that lack a complete reference genome. Transcribed sequence (cDNA) resources assembled from deep sequence datasets generated using the 454 pyrosequencing platform have been developed for a wide variety of plant species, including olive (*Olea europaea*) (Alagna et al. 2009), American chestnut (*Castanea dentata*) (Barakat et al. 2009), buckwheat (*Fagopyrum* sp.) (Logacheva et al. 2011), the stress-tolerant grain, *Amaranthus tuberculatus* (Delano-Frier et al. 2011), orchid (*Phalenopsis* sp.) (Hsiao et al. 2011), and sagebrush (*Artemisia tridentata*) (Bajgain et al. 2011). Such transcriptome assemblies are valuable for gene annotation, whole-genome assembly, identification of structural genes, marker discovery, and comparative phylogenetic analyses (Varshney et al. 2009, 2012; McKain et al. 2012).

Assembled transcriptomes have been used to characterize gene content and identify SNPs in two model tree species, *Eucalyptus grandis* (Novaes et al. 2008) and *Populus trichocarpa* (Geraldes et al. 2011). In non-model tree species, assembled transcriptomes have been used to identify fruit development genes in olive (Alagna et al. 2009) and to compare the transcriptomes of American and Chinese chestnut (*C. dentata* and *Castanea mollissima*, respectively) (Barakat et al. 2009). Because of their large genomes, conifer studies have mostly focused on transcriptome analyses, mostly on species in the Pinaceae and only one of seven (or eight) extant families in the Order Pinales. For example, lodgepole pine (*Pinus contorta*) sequences derived from a normalized, pyrosequenced library were used to assemble a transcriptome of ca. 64,000 contigs that contained a large number of retrotransposon-derived sequences (Parchman et al. 2010). Rigault et al. (2011) used more than 2.5 million reads from multiple sequencing platforms to create a gene catalog for white spruce (*Picea glauca*), while a pyrosequencing dataset of nearly one million reads was assembled and annotated for maritime pine (*Pinus maritima*) laying the foundation for the EuroPineDB database (Fernandez-Pozo et al. 2011). Mining of conifer transcriptome datasets has led to important discoveries, such as the wide extent of diversification of the terpene synthase gene family in white and Sitka spruce (*Picea sitchensis*) (Keeling et al. 2011), as well as the identification of genes responsive to fungal elicitation in Scots pine (*Pinus sylvestris*) (Sun et al. 2011).

Loblolly pine (*Pinus taeda*), another member of the Pinaceae, is the single most economically important crop species in the USA and is the source of nearly 16 % of the global annual timber harvest (Schultz 1999). We previously developed a variety of genomic resources for loblolly pine, including SAGE and expressed sequence tag (EST) datasets (Lorenz and Dean 2002; Lorenz et al. 2006), cDNA microarrays (Lorenz et al. 2011), and a variety of genetic maps and molecular markers (e.g., Eckert et al. 2009; Jermstad et al. 2011). However, these genomic resources remain insufficient for understanding gene expression associated with loblolly pine growth and development and for constructing a reference genome sequence.

Through a collaboration with the US Department of Energy Joint Genome Institute, we substantially increased the number of transcript sequences available for loblolly pine and developed the first transcriptome assemblies for species representing five conifer families: Araucariaceae (Araucaria); Podocarpaceae (Yellowwood); Sciadopityaceae (Umbrella pine); Cephalotaxaceae (Plum-yew); and Taxaceae (Yew). We used three different sequence assembly tools (Newbler, MiraEST, and NGen) to assemble the transcriptomes of 13 gymnosperm species. These transcriptome assemblies and associated annotations are available from the database reported here.

Materials and methods

Source tissues and treatments: *P. taeda*

Five loblolly pine cDNA libraries were prepared using multiple tissues from multiple genotypes. Two libraries

(CFCN and CFCP) were generated using elongating shoot tips (candles) collected from six 4-year-old trees representing three genotypes (CCLONES 40430, 41586, and 43608) originally obtained from the Forest Biology Research Cooperative at the University of Florida. Pooled RNA was used to synthesize one normalized (CFCN) and one non-normalized (CFCP) library, and these two libraries were sequenced using the Roche 454 GS-FLX platform. All other libraries characterized in this project were sequenced using the longer read GS-XLR platform.

One of the remaining loblolly pine libraries (MIXED) was prepared using a mixture tissues collected from a 40-year-old feral tree harvested at Whitehall Forest (Athens, GA). The sampled tissues included mature basal secondary xylem, mature basal secondary phloem, juvenile crown xylem, branch compression (underside) xylem, branch opposite (topside) xylem, apical tips, first-year whole female cones, and second-year whole female cones. Harvested tissues were flash-frozen in liquid N_2, transported to the laboratory, and stored at −80 °C until RNA preparation. Equal amounts of total RNA purified from each individual tissue were pooled and used as the template for cDNA synthesis.

Another library (STEM) was prepared using similarly pooled RNA samples from stems and needles of 7-month-old seedlings subjected to ten treatments (ten seedlings/treatment). In six of the treatments, stems were cut at soil level and the cut ends were submerged and incubated for 24 h in 500 mL aqueous solutions of the following compounds (final concentration): cycloheximide (0.5 mg/mL), gibberellic acid (100 μM), H_2O_2 (1 % *v*/*v*), indole-3-acetic acid (100 μM), methyl jasmonate (100 μM), and kinetin (100 μM). For a seventh treatment (desiccation), cut stems were left on an open bench for 24 h. For the remaining three treatments, whole, uncut seedlings were incubated for 24 h at −20 °C (cold-shock), 45 °C (heat-shock), or after their stems had been extensively crushed with pliers (wounding). After these incubations, stems were cut at soil level and tissues were flash-frozen until RNA preparation

The remaining loblolly pine library (CALLUS) was prepared using suspension-cultured needle cells grown in Eberhardt medium as described (Eberhardt et al. 1993; Bordeaux 2008). Freshly transferred cells were grown in shake culture (100 rpm) for 14 days at room temperature in the dark. Fourteen separate cultures were incubated for 4 h with the following compounds (final concentration): methyl jasmonate (100 μM), gibberellic acid (100 μM), kinetin (100 μM), cycloheximide (0.5 mg/mL), indole-3-acetic acid (100 μM), nitrophenol (100 μM), sodium azide (100 μM), colchicine (100 μM), sodium (meta) vanadate (100 μM), H_2O_2 (0.01 % *v*/*v*), chitin (1 mg/mL), chitosan (100 μg/mL), *Sirex noctilio* venom gland extract (125 mg/mL), and heat-precipitated *S. noctilio* venom gland extract (125 mg/mL). After the incubations, cells were harvested by filtration, flash-frozen, and stored at −80 °C prior to RNA isolation. Equal amounts of total RNA isolated from these variously treated cells were pooled and used to prepare cDNAs as described above

Source tissues and treatments: other gymnosperm species

To the extent possible, elongating shoot tips were used as source tissues for cDNA libraries prepared for the other gymnosperm species in this study (Table 1). In most cases, tissues were collected from plants growing under ambient conditions in the greenhouse or in outdoor collections. In all cases, tissues were flash-frozen, transported to the laboratory, and stored at −80 °C prior to RNA preparation.

Unless otherwise noted, RNA was extracted from trees and tissues that had not undergone experimental treatment. For those samples that did receive specific treatments (Supplementary Table 1), brief details are as follows. The sugar pine (*Pinus lambertiana*) sample contained shoot tip (candle) tissues from two parental genotypes (#5038 female and #5500 male), as well as needles from known resistant and susceptible open-pollinated progeny from #5701. Needle samples from #5701 progeny were collected from seedlings pre- and post-inoculation with *Cronartium ribicola*, the causative agent of white pine blister rust. These sugar pine genotypes have been described in previous studies (e.g., Jermstad et al. 2011). The Douglas-fir (*Pseudotsuga menziesii*) sample contained a mixture of stem, lateral and terminal buds, shoot, needle, seed, and cambium tissues collected from a single genotype (#2650). The Norway spruce (*Picea abies*) sample provided by Dr. Trevor Fenning (Forest Research, Midlothian, UK) contained aerial tissues from control and treated seedlings that had been sprayed with 100 μM methyl jasmonate in an aqueous solution of 0.05 % (*v*/*v*) Tween-20. Treated tissues were harvested at 2, 7, and 30 days posttreatment.

RNA isolation, cDNA synthesis, and sequencing

Tissue samples were pulverized under liquid nitrogen using a SPEX model 6850 freezer mill (SPEX, Metuchen, NJ). Total RNA was isolated as described previously (Lorenz et al. 2010) and RNA samples were DNase-treated using the Ambion TURBO DNA-*free*™ Kit (Applied Biosystems Inc., Foster City, CA). RNA concentrations were determined spectrophotometrically and RNA quality and integrity were verified using an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA).

cDNA synthesis, library normalization (*P. taeda* CFCN library only), and sequencing were performed at the US DOE Joint Genome Institute (JGI) using standard in-house protocols (http://www.jgi.doe.gov/sequencing/protocols/_prots_production.html). Two *P. taeda* libraries (CFCP and CFCN) were primed using oligo-dT, while all other libraries

Table 1 Species, sample, library identification, and sequence metrics

Species	Family	Tissue	Stage[a]	Age (year)	Library ID	Length[b]	Reads[c]	SRA[d]
Pinus taeda	Pinaceae	First flush shoots	IM	4	CFCP	236	469,478	SRA023533
Pinus taeda	Pinaceae	First flush shoots	IM	4	CFCN	239	415,646	SRA023533
Pinus taeda	Pinaceae	Mixed	M	40	CGIT	356	1,255,033	SRA023533
Pinus taeda	Pinaceae	Stem, treated	JV	0.6	CGIS	340	1,178,375	SRA023533
Pinus taeda	Pinaceae	Callus, treated	N/A	N/A	CGIU	396	1,012,793	SRA023533
Cedrus atlantica	Pinaceae	Mixed shoot	JV	8	CGON	347	416,216	SRA023736
Cephalotaxus harringtonia	Cephalotaxaceae	Mixed shoot	JV	1	CGTS	393	695,559	SRA023613
Gnetum gnemon	Gnetaceae	Mixed shoot	JV	1	CGSA	367	432,598	SRA023615
Picea abies	Pinaceae	Mixed shoot	IM	3	CTGN	388	630,265	SRA023567
Pinus lambertiana	Pinaceae	Shoot, needles	JV/M	1–30	CGIP	369	1,184,417	SRA023577
Pinus palustris	Pinaceae	Mixed shoot	IM	10	CGOI	317	552,384	SRA023739
Podocarpus macrophyllus	Podocarpaceae	Mixed shoot	JV	1	CGTO	370	594,502	SRA023741
Pseudotsuga menziesii	Pinaceae	Mixed	M	22–24	CGOH	354	1,256,470	SRA023776
Sciadopitys verticillata	Sciadopityaceae	Mixed shoot	JV	1	CGTP	399	484,806	SRA023758
Sequoia sempervirens	Cupressaceae	Needles	M	<5	CGPX	343	480,130	SRA023765
Taxus baccata	Taxaceae	Mixed shoot	M	>100	CGPY	331	409,750	SRA023771
Wollemia nobilis	Araucariaceae	Mixed shoot	JV	1	CGPZ	373	481,506	SRA023774

IM immature, *M* mature, *J* juvenile, *N/A* not applicable, *UK* unknown

[a] Developmental stage

[b] Mean read length for each sequenced library

[c] Total number of raw reads prior to filtering

[d] NCBI Short Read Archive (SRA) accession number

were primed with random hexamer oligonucleotides. In most cases, libraries were sequenced at half-plate scale (400,000 read target). Full-plate sequencing runs were performed for three *P. taeda* libraries (CGIT, CGIS, and CGIU), as well as one *P. lambertiana* (CGIP) and one Douglas-fir (CGOH) library.

Contig assembly

GS-FLX and GS-XLR reads for *P. taeda* libraries were initially filtered through the JGI bioinformatics pipeline to remove ribosomal and contaminating linker/adaptor sequences. These reads were subsequently filtered and trimmed again to remove additional mitochondrial and chloroplastic sequences using an in-house pipeline based on SeqClean (Chen et al. 2007). Sanger ESTs (328,662) used in the *P. taeda* hybrid assemblies were also filtered using the SeqClean pipeline. Filtered sequences were assembled using Newbler version 2.3 (454 Life Sciences, Branford, CT), MiraEST version 3.0.5 (Chevreux et al. 2004), and SeqMan NGen version 3.0 (DNAStar, Madison, WI). Newbler and NGen assemblies were performed using default settings for de novo transcriptome assembly. MiraEST hybrid assemblies were performed at default settings recommended for de novo transcriptome using the following commands: -project = projectname -job = denovo,est,accurate,454,sanger -fasta -notraceinfo -AS:ugpf = no -SK:mnr = yes:nrr = 10 -GE:not = 4.

Sequence reads for the other 12 species were screened through the SeqClean pipeline using adaptor/vector and generic trim files based on *P. taeda* and *P. abies* sequence data prior to assembly using the MiraEST and NGen tools. Prior to assembly using Newbler, builds were processed as sff files and the data were filtered using the same adaptor/vector and generic trim files.

BLAST analysis

BLASTX analysis (Altschul et al. 1990) was used to assign putative gene function(s) to each separate contig assembly generated by the MiraEST, Newbler, and NGen tools. Separate queries were made against the National Center for Biotechnology Information (NCBI) non-redundant and The Arabidopsis Information Resource (TAIR) databases. Returned information was accepted where expect values (*E* values) were $<1\times10^{-5}$, and a maximum of five high-scoring pairs (HSPs) returned for each contig were parsed into the database.

Database access to sequences, assemblies, and annotations

Raw sequence datasets were deposited in the Sequence Read Archive (SRA) database at NCBI under the accession numbers shown in Table 1. Conifer_DBMagic is a publicly accessible database housing all sequence and assembly data, including contig consensus sequences (referred to in the database as unique transcripts or "uniscripts"), reads, alignments, and annotations (http://ancangio.uga.edu/ng-genediscovery/conifer_dbMagic.jnlp).

Results

Twelve species of conifer and one Gnetales were the source of 17 cDNA libraries from which 11.95 million pyrosequencing reads were generated. The conifer samples included representative species for each of the seven existent conifer families (Table 1). *Gnetum gnemon* (Division Gnetophyta) was included in the study to develop additional nuclear genome sequence resources that could be used to address conflicting phylogenetic hypotheses variously placing this unusual group of plants as a sister clade to the Pinales (e.g., Bowe et al. 2000; Chaw et al. 2000), the Coniferophyta (Burleigh and Mathews 2004), or elsewhere in the seed plant lineage (Hilton and Bateman 2006). Libraries were typically generated from mixed current year tissues (primarily shoots and needles) with most samples also being collected from juvenile trees (Table 1). However, the age of plants from which tissues were harvested ranged from less than 1 year (*P. taeda*) to over 100 years (*Taxus baccata*). Details of the treatments applied to some tissues, genotype information (where available), sequence number and read length, and SRA accession numbers are listed in Supplementary File 1.

Six of the conifer species we sequenced are members of the Pinaceae and yielded over 7.9 million reads. More than 4.2 million reads were generated for the remaining Pinophyta species. More than four million reads were produced from *P. taeda* libraries. In the pilot phase of the project, two *P. taeda* shoot tip libraries were generated using oligo-dT priming and were either normalized (CFCN) or were non-normalized (CFCP). Because the difference in the rates of novel sequence discovery between these two libraries was only about 10 %, subsequent libraries were not normalized. The remaining libraries were primed using random hexamers to improve coverage when the sequencing pipeline shifted to the titanium pyrosequencing chemistry. Read length for the two pilot libraries (CFCP and CFCN) averaged 240 bases, while read lengths increased to near 400 bases for several libraries after the shift to titanium chemistry (Table 1). Mean read length across the entire project was 348 bases. The highest number of reads per plate was seen for the *P. taeda* mixed tissue library (CGIT, ca. 1.25 million reads), while the lowest number of reads was seen for *T. baccata* (CGYP, ca. 410,000 reads).

Assembly

Prior to assembly, *P. taeda* sequence reads were converted from sff format to FASTA and FASTA/qual formats before being filtered and trimmed using both the JGI and in-house pipelines. Reads for other species were similarly converted and filtered as described in the "Materials and methods" section. Due to differences in results reported from the different assemblers, contigs less than 100 bases in length were excluded from further analyses, as were singletons and any unassembled or "debris" data. Six metrics were used to compare assembly results: total contigs, number of bases assembled, mean contig length, longest contig, large contigs (i.e., contigs ≥1 or 2 kb), and contigs containing ≥5 reads. In general, assembly metrics for NGen and MiraEST assemblies were more similar to one another than either was to the Newbler assemblies (Fig. 1 and Supplementary File 2). For example, the total number of contigs assembled was consistently about 60 % greater using NGen and MiraEST (Fig. 1a). This was most evident in the hybrid *P. taeda* assembly that contained both Sanger and pyrosequencing data and was by far the largest and most varied dataset with respect to multiple samples and genotypes. In this case, the NGen (198,852 contigs) and MiraEST (187,374 contigs) counts were increased 3.8- and 4-fold, respectively, over the Newbler (48,751 contigs) count. NGen and MiraEST assemblies for all other species averaged 2.4-fold more contigs than Newbler with a range of 1.8- to 3.0-fold. With respect to the percentage of bases assembled, NGen performed better than either Newbler or MiraEST in all 13 datasets (Fig. 1b). Averaged across all datasets, the total percentage of bases assembled was 89.2, 82.5, and 73 % for NGen, Newbler, and MiraEST, respectively (Supplementary File 2). Unlike NGen and Newbler, which did not report singletons, MiraEST generated a large number of singletons that decreased the percentage of bases assembled into contigs.

Mean contig length was significantly greater in the Newbler assemblies across all libraries with the largest differences occurring in libraries having greater numbers of reads, e.g., *P. menziesii* and *P. taeda* (Fig. 1c). Averaged across all assemblies, the mean Newbler contig length was ca. 400 bp longer than average contig lengths returned from either MiraEST or NGen (Fig. 1d). Read alignments were inspected using the Tablet program (Milne et al. 2010), and the largest *P. taeda* contig (a 12-kb transcript generated by MiraEST) as well as many of the largest contigs produced by all three assemblers (e.g., those >10 kb) were found to be chimeras resulting from misassembly (data not shown).

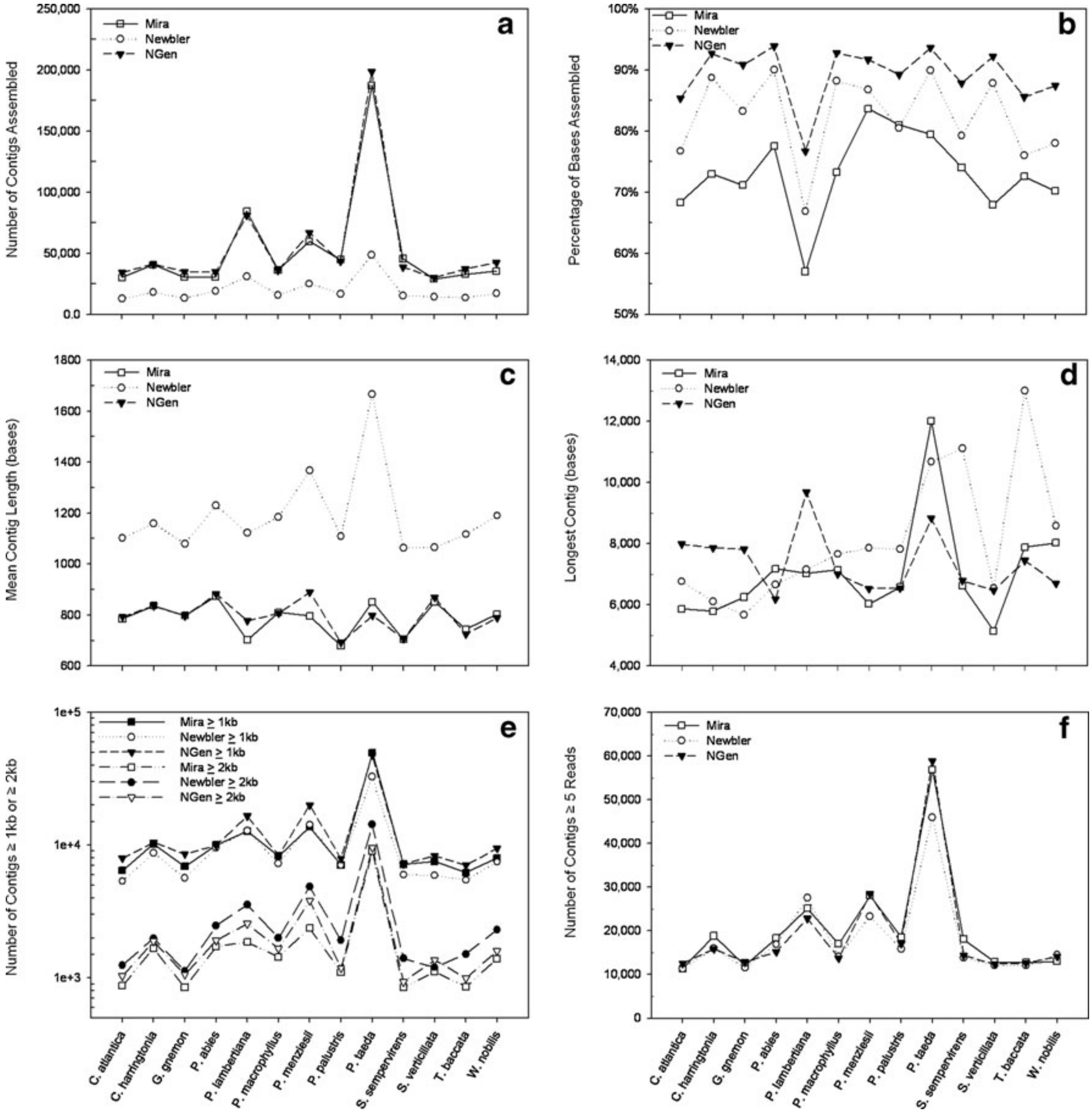

Fig. 1 Assembly metrics for 13 gymnosperms each assembled with three different assemblers. **a** Total number of contigs assembled. **b** Percentage of bases assembled into contigs. **c** Mean assembly contig length. **d** Longest contig identified in assembly. **e** Number of contigs that are either ≥2 or ≥1 kb in length. **f** Number of contigs containing at least five ESTs. Note that lines appearing to connect data points are for visualization purposes only and are not meant to convey quantitative relationships between the metrics for different species

Such chimeras were generally more prevalent in the Newbler assemblies, e.g., for *Sequoia sempervirens* and *T. baccata*, as well as the *P. taeda* hybrid assemblies. The number of large contigs that was either ≥1 or ≥2 kb was determined as an estimate for average-sized genes (Fig. 1e). With few exceptions, the NGen assemblies, followed by MiraEST, contained a greater number of contigs ≥1 kb, whereas the Newbler assemblies contained more contigs ≥2 kb. For most species, differences in the number of large contigs were slight among assemblers. However, in those libraries having the deepest sequencing coverage, particularly *P. lambertiana* and *P. menziesii*, the Newbler assemblies contained almost twice the number of contigs ≥2 kb than were seen in the MiraEST assemblies. With respect to the final assembly metric, the MiraEST assembler generated more contigs with ≥5 reads than NGen or Newbler did (Fig. 1f).

Annotation of contigs

BLASTX analysis was performed on each set of contigs generated for each species by each of the three assemblers. Every contig set was used to query the NCBI non-redundant and TAIR protein databases, and the top five HSPs with *E* values $\leq 1 \times 10^{-5}$ were returned and parsed into the database. For comparing annotation results, only the best hit having an *E* value cutoff $\leq 1 \times 10^{-15}$ was used (Supplementary File 3). As expected, a greater percentage of contig sequences returned annotations from the NCBI database than from TAIR. However, the difference between the two was typically less than 10 %, irrespective of the assembler used (Fig. 2).

For every species, the percentage of annotations returned was highest for contigs assembled using Newbler, and in 11 of the 13 species, NGen assemblies returned a higher percentage of annotations than did those from MiraEST (Fig. 2). Across all assemblies, the percentage of Newbler contigs annotated was 81 and 73 % for queries against the NCBI and TAIR databases, respectively. It is important to note, however, that while a higher percentage of Newbler contigs were annotated in every species, the total number contigs annotated from NGen and MiraEST assemblies exceeded the number from Newbler because contig counts were typically two- to threefold greater for NGen and MiraEST. More than 87 % of the *Cedrus atlantica* Newbler contigs returned annotations, followed closely by *P. abies* and *Sciadopitys verticillata* for which >85 % of contigs were annotated. The lowest percentage of contigs returning annotations occurred in the NGen assembly for *P. lambertiana* (39 %); however, all *P. lambertiana* assemblies fared poorly (<60 %) relative to other species. Storage issues resulting in lower RNA quality may have contributed to the poor results for *P. lambertiana*.

Database

Conifer DBMagic (http://ancangio.uga.edu/ng-genediscovery/conifer_dbMagic.jnlp) is a relational database. The schema is modeled after the .ace file format for storing assembly alignments and database access is through a Java front end. It is based on NGMagic (http://sourceforge.net/apps/trac/ngmagic/), which, unlike its predecessor MAGIC-SSP (Liang et al. 2006), is oriented towards handling of next-generation sequencing data and assemblies. Once a species and assembly are selected, transcriptome builds can be queried by assembly characteristics or by annotation keyword searches. The contig consensus sequences identified through these queries can be downloaded in FASTA format, with or without the corresponding reads. A short tutorial for using the Conifer DBMagic database is provided in Supplementary File 4.

Discussion and conclusions

The nearly 12 million pyrosequencing reads generated in this study, along with accompanying transcriptome builds for each species, provide a robust new dataset to aid our understanding of conifer genomics. Prior to this work, only *P. taeda*, *P. abies*, *P. menziesii*, and *G. gnemon*, among the

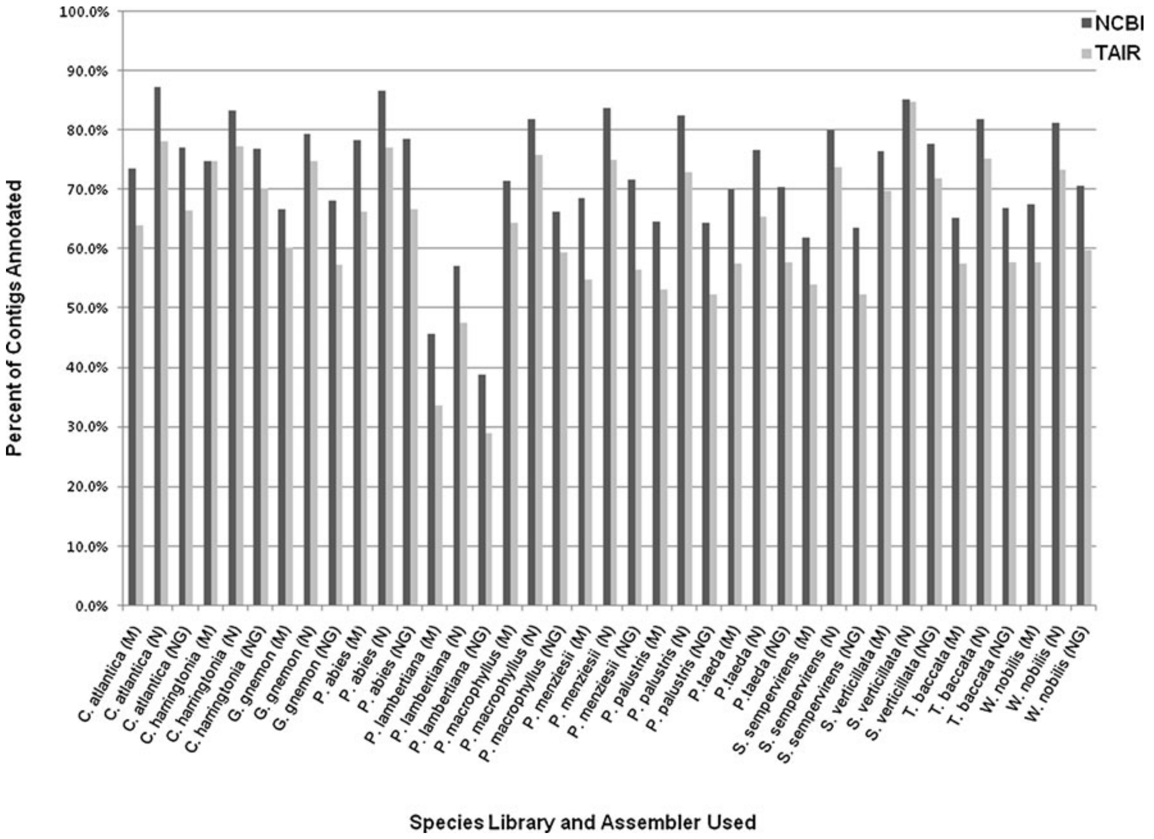

Fig. 2 BLAST hit metrics. Percent of contigs annotated in each species and for each assembler after BLASTX query against NCBI non-redundant and TAIR9 databases. *M* MiraEST, *N* Newbler, *NG* NGen assembler. *E* value cutoff for returned HSPs was $\leq 1 \times 10^{-15}$

species studied, had significant DNA sequence resources available in GenBank. These transcriptome assemblies are already providing a valuable resource for studies of genes and gene family evolution in the conifers (Bagal et al. 2012) and for improved interpretation of results from the PtGen2 microarray (Lorenz et al. 2011).

De novo transcriptome assembly from short read length sequence datasets can be difficult in the absence of a reference genome due to the many factors that contribute to misassembly of contigs. Cloning and sequencing errors, sequence polymorphisms, sequence repeats including homopolymers, contaminating sequences (e.g., transposons, mitochondrial, and rDNA sequences), variations in transcript abundance, splicing variants, allelic variation, and paralogous genes can all lead to erroneous joining (overassembly) or splitting (underassembly) of transcripts (Wall et al. 2009; Papanicolaou et al. 2009; Surget-Groba and Montoya-Burgos 2010). A variety of assembly tools are available for de novo transcriptome assembly (e.g., Papanicolaou et al. 2009; Weber et al. 2007), and all have strengths and weaknesses that reflect algorithmic tendencies to lump or split sequences. One recent study compared six of the most popular assemblers for their respective strengths and weaknesses and found that Newbler 2.5 generally returned longer contigs and better alignments, while SeqMan NGen did a better job of recapitulating known transcripts and identified more novel contigs (Kumar and Blaxtaer 2010). Our findings support their conclusions, and we echo their recommendation to use multiple assemblers because integration of complementary output from different programs provided the most informative final product. For example, we have used the combined MiraEST and Newbler assembly results to identify full-length members of the RecQ helicase gene family (F. Hartung, personal communication), as well as five full-length PAL genes in *P. taeda* (Bagal et al. 2012).

In general, the more collapsed Newbler assemblies were a good place to start for associating specific read sequences with broad functional annotations. The more finely divided contigs produced by NGen and MiraEST could then be used to better separate gene family members. None of the assemblers, however, was particularly good with transcripts over 10 kb as many large contigs were misassembled. No doubt the default assembler settings used in this study could be fine-tuned to return improved transcriptome assembly, as was demonstrated by Kumar and Blaxter (2010). New assembly builds using later versions of Newbler and MiraEST, along with modified assembly parameters, are in progress and will be uploaded to Conifer DBMagic as they become available.

The Conifer DBMagic resource greatly expands the list of expressed genes in gymnosperms and should be of immediate use in development of new tools for functional genomics studies in conifers. The new sequence information made available for previously unstudied conifer species will greatly facilitate phylogenetic analyses and improve our understanding of higher plant evolution.

Acknowledgments The authors wish to acknowledge Yuan-Sheng Yu and Ujwal Bagal for their technical assistance. We also wish to thank Dr. Trevor Fenning for supplying the *P. abie*s samples and Dr. Richard Cronn for early access to *P. taeda* chloroplast genome sequence information. These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. Particular thanks go to Christa Pennacchio, Kerrie Barry, Erika Lindquist, and their associates at JGI for directing cDNA library construction, 454 pyrosequencing, and some of the data filtering reported for this project. Partial funding for this work was provided by USDA/NRI CSREES CAP Award #2007-55300-18603.

References

Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M, Giuliano G, Chiusano ML, Baldoni L, Perrotta G (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 10:399

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bagal UR, Leebens-Mack JH, Lorenz WW, Dean JFD (2012) The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage. BMC Genomics 13 (Suppl. 3):S1. doi:10.1186/1471-2164-13-S3-S1

Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA (2011) Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). BMC Genomics 12:370

Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, Wheeler N, Sederoff R, Carlson JE (2009) Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. BMC Plant Biol 9:51

Bordeaux JM (2008) Characterization of growth conditions for production of a laccase-like phenoloxidase by *Amylostereum areolatum*, a fungal pathogen of pines and other conifers. M.S. Thesis, University of Georgia, Athens, GA, 110 pg.

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 USA 97:4092–4097

Bowyer JL, Shmulsky R, Haygreen JG (2007) Forest products and wood science: an introduction, 5th edn. Blackwell, Ames, p 576

Burleigh JG, Mathews S (2004) Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am J Bot 91:1599–1613

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD (2000) Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci USA 97:4086–4091

Chen YA, Lin CC, Wang CD, Wu HB, Hwang PI (2007) An optimized procedure greatly improves EST vector contamination removal. BMC Genomics 8:416

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159

Delano-Frier JP, Aviles-Arnaut H, Casarrubias-Castillo K, Casique-Arroyo G, Castrillón-Arbeláez PA, Herrera-Estrella L, Massange-Sánchez J, Martínez-Gallardo NA, Parra-Cota FI, Vargas-Ortiz E, Estrada-Hernández MG (2011) Transcriptomic analysis of grain amaranth (*Amaranthus hypochondriacus*) using 454 pyrosequencing: comparison with *A. tuberculatus*, expression profiling in stems and in response to biotic and abiotic stress. BMC Genomics 12:363

Eberhardt TL, Bernards MA, He L, Davin LB, Wooten JB, Lewis NG (1993) Lignification in cell-suspension cultures of *Pinus taeda*. In situ characterization of a gymnosperm lignin. J Biol Chem 268:21088–21096

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

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

Fernandez-Pozo N, Canales J, Guerrero-Fernández D, Villalobos DP, Díaz-Moreno SM, Bautista R, Flores-Monterroso A, Guevara MÁ, Perdiguero P, Collada C, Cervera MT, Soto A, Ordás R, Cantón FR, Avila C, Cánovas FM, Claros MG (2011) EuroPineDB: a high-coverage web database for maritime pine transcriptome. BMC Genomics 12:366

Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M, Birol I, Jones SJ, Cronk QC, Douglas CJ (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. Mol Ecol Resour 11 (Suppl 1):81–92

Hilton J, Bateman RM (2006) Pteridosperms are the backbone of seed-plant phylogeny. J Torrey Bot Soc 133:119–168

Hsiao YY, Chen YW, Huang SC, Pan ZJ, Fu CH, Chen WH, Tsai WC, Chen HH (2011) Gene discovery using next-generation pyrosequencing to develop ESTs for *Phalaenopsis* orchids. BMC Genomics 12:360

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 Genom 7:457–468

Keeling CI, Madilao LL, Zerbe P, Dullat HK, Bohlmann J (2011) The primary diterpene synthase products of *Picea abies* levopimaradiene/abietadiene synthase (PaLAS) are epimers of a thermally unstable diterpenol. J Biol Chem 286:21145–21153

Kumar S, Blaxter ML (2010) Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11:571

Liang C, Sun F, Wang H, Qu J, Freeman RM Jr, Pratt LH, Cordonnier-Pratt MM (2006) MAGIC-SPP: a database-driven DNA sequence processing package with associated management tools. BMC Bioinforma 7:115

Logacheva MD, Kasianov AS, Vinogradov DV, Samigullin TH, Gelfand MS, Makeev VJ, Penin AA (2011) De novo sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12:30

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, Sun F, Liang C, Kolychev D, Wang H, Zhao X, Cordonnier-Pratt MM, 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

Lorenz WW, Yu YS, Dean JFD (2010) An improved method of RNA isolation from loblolly pine (*P. taeda* L.) and other conifer species. J Vis Exp (36).pii:1751

Lorenz WW, Alba R, Yu YS, 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

McKain MR, Wickett N, Zhang Y, Ayyampalayam S, McCombie WR, Chase MW, Pires JC, Depamphilis CW, Leebens-Mack J (2012) Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). Am J Bot 99:397–406

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet-next generation sequence assembly visualization. Bioinformatics 26:401–402

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312

Papanicolaou A, Stierli R, Ffrench-Constant RH, Heckel DG (2009) Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinforma 10:447

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11:180

Plomion C, Bouquet J, Kole C (2012) Genetics, genomics and breeding of conifers. CRC Press, Boca Raton, p 456

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

Schultz RP (1999) Loblolly: the pine for the twenty-first century. New Forests 17:71–88

Sun H, Paulin L, Alatalo E, Asiegbu FO (2011) Response of living tissues of *Pinus sylvestris* to the saprotrophic biocontrol fungus *Phlebiopsis gigantea*. Tree Physiol 31:438–451

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 20:1432–1440

Varshney RK, Hiremath PJ, Lekha P, Kashiwagi J, Balaji J, Deokar AA, Vadez V, Xiao Y, Srinivasan R, Gaur PM, Siddique KH, Town CD, Hoisington DA (2009) A comprehensive resource of drought and salinity-responsive ESTs for gene discovery and marker development in chickpea (*Cicer arietinum* L.). BMC Genomics 10:523

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang SP, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, Jackson SA (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Wall PK, Leebens-Mack JH, Chanderbali A, Barakat A, Wolcott E, Liang H, 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

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol 144:32–42