

Transcriptome characterization and detection of gene expression differences in aspen (*Populus tremuloides*)

Hardeep S. Rai · Karen E. Mock · Bryce A. Richardson ·
Richard C. Cronn · Katherine J. Hayden ·
Jessica W. Wright · Brian J. Knaus · Paul G. Wolf

Received: 21 August 2012 / Revised: 11 February 2013 / Accepted: 27 February 2013 / Published online: 16 April 2013
© Springer-Verlag Berlin Heidelberg (outside the USA) 2013

Abstract Aspen (*Populus tremuloides*) is a temperate North American tree species with a geographical distribution more extensive than any other tree species on the continent. Because it is economically important for pulp and paper industries and ecologically important for its role as a foundation species in forest ecosystems, the decline of aspen in large portions of its range is of serious concern. The availability and annotation of the black cottonwood (*Populus trichocarpa*) genome enables a range of high throughput sequencing approaches that can be used to

understand rangewide patterns of genetic variation, adaptation, and responses to environmental challenges in other *Populus* species, including aspen. Gene expression studies are particularly useful for understanding the molecular basis of ecological responses, but are limited by the availability of transcriptome data. We explored the aspen transcriptome through the use of high-throughput sequencing with two main goals: (1) characterization of the expressed portion of the *P. tremuloides* genome in leaves and (2) assessment of variation in gene expression among genets collected from distinct latitudes but reared in a common garden. We also report a large single nucleotide polymorphism dataset that provides the groundwork for future studies of aspen evolution and ecology, and we identify a set of differentially expressed genes across individuals and population boundaries for the leaf transcriptome of *P. tremuloides*.

Communicated by P. Ingvarsson

H. S. Rai · K. E. Mock
Ecology Center and Department of Wildland Resources,
Utah State University, 5230 Old Main Hill,
Logan, UT 84322, USA

B. A. Richardson
Rocky Mountain Research Station, USDA Forest Service, Provo,
UT 84606, USA

R. C. Cronn · B. J. Knaus
Pacific Northwest Research Station, USDA Forest Service,
Corvallis, OR 97331, USA

K. J. Hayden
Department of Environmental Science, Policy, and Management,
University of California, Berkeley, CA 94720, USA

J. W. Wright
Pacific Southwest Research Station, USDA- Forest Service, 1731
Research Park Drive,
Davis, CA 95618, USA

P. G. Wolf (✉)
Ecology Center and Department of Biology, Utah State University,
5305 Old Main Hill,
Logan, UT 84322, USA
e-mail: paul.wolf@usu.edu

Keywords Trembling aspen · Quaking aspen · RNA-Seq · Differential expression · SNP (single nucleotide polymorphism) · *Populus trichocarpa*

Introduction

The poplars (*Populus* L., Salicaceae) are widely distributed northern hemisphere trees that comprise 22–85 species, usually in six sections, depending on taxonomic scheme (Eckenwalder 1977; Hamzeh and Dayanandan 2004). Here, we have sequenced the leaf tissue transcriptome of *Populus tremuloides* Michx. (trembling aspen), a close relative of the fully sequenced model organism black cottonwood, *Populus trichocarpa* Torr. & Gray. *P. tremuloides* has the broadest distribution, in terms of both latitude and longitude, of any North American tree species (Jones 1985) and has been

proposed as an ideal study system for both adaptation and responses to climate change (Hogg and Hurdle 1995; Isebrands et al. 2001). Its distribution across many climatic zones suggests that *P. tremuloides* is highly plastic, highly adaptable, or both (Mitton and Grant 1996). *P. tremuloides* has been shown to have extremely high levels of genetic variation at the population level, both in terms of trait heritability and molecular marker diversity (Cole 2005; Liu and Furnier 1993). Aspen is economically important for pulp and paper industries and ecologically important for its role as a foundation species in forest ecosystems (Mitton and Grant 1996). Dramatic mortality in certain portions of the species' range has recently been documented, raising concerns about the ability of *P. tremuloides* to respond to a changing climate (Frey et al. 2004; Iverson and Prasad 2002; Rehfeldt et al. 2009; Worrall et al. 2008; Hogg et al. 2008). A pervasive loss of this tree species at a continental scale would have significant ecological and economic impact and could result in dramatic carbon fluxes (Kurz and Apps 1999; Hogg and Hurdle 1995).

The availability of a high-quality genomic reference in the genus, *P. trichocarpa* (Tuskan et al. 2006), along with dramatic advances in sequencing technologies (reviewed by Fang and Cui 2011; Metzker 2009), provide new opportunities to understand the evolutionary history of *Populus* species and to characterize molecular responses to environmental challenges (e.g., Jiang et al. 2012; Qiu et al. 2011). In *P. tremuloides*, these technologies enable studies at the molecular level to examine how a single species can achieve such an enormous ecological amplitude (e.g., through plasticity, adaptation, migration) and what factors limit its persistence in changing climates. Answers to these questions are critical for effective management of natural populations as well as for silvicultural improvement aimed at specific applications (e.g., biofuels, carbon sequestration).

Here, we present an initial characterization of the *P. tremuloides* transcriptome. The high-throughput sequencing of cDNA libraries (RNA-Seq) enables the direct study of the expressed portion of the genome in sampled tissues (Marioni et al. 2008; Nagalakshmi et al. 2008; Perkins et al. 2009; Wang et al. 2009). Compared to array-based technologies, RNA-Seq has the advantage of higher sensitivity and the ability to capture a significantly larger component of gene expression (Severin et al. 2010), allowing for the assay of tens of thousands of transcripts, or even the entire transcriptome (e.g., Bajgain et al. 2011; Bräutigam et al. 2011; Coppe et al. 2010; Geraldine et al. 2011; Nagalakshmi et al. 2008; Severin et al. 2010; Wang et al. 2009). A distinct advantage of RNA-Seq is that the relative abundance of sequencing reads, aligned with or without the existence of a reference genome, can be used as an estimate of transcript abundance,

which can be compared among individuals or under different environmental conditions (Bashir et al. 2010; Wang et al. 2009).

Variation within the transcriptome of other *Populus* species has been investigated for responses to various sources of stress using microarrays (Grisel et al. 2010; Hamanishi et al. 2010), for comparative studies (using real-time PCR and microarray analysis: Quesada et al. 2008), and even as a tool for understanding ecosystem interactions (using RNA-Seq data: Larsen et al. 2011). However, little attention has been paid to the descriptive, functional, or comparative genomics of *P. tremuloides*. Here, we present an overview of the aspen transcriptome, focusing on expression in leaf tissues using an RNA-Seq approach. Our major objectives are (1) to provide a snapshot of the expressed portion of the aspen genome and (2) to characterize variation present within and among aspen from two latitudinally distinct regions.

Within the transcribed regions reported here, we characterize sequence variation [single nucleotide polymorphisms (SNPs)] present in the expressed regions of the genome of aspen within and among individual genotypes. Gene expression patterns vary during development, between tissue types, and with external stimuli, and play a key role in the functioning of all organisms. We also highlight novel transcribed regions of the genome relative to the transcript assemblies of black cottonwood (*P. trichocarpa* v2; Phytozome v8.0). The availability of this type of data has utility in future studies of aspen, including the discovery and identification of novel genes, identification of genetic variants useful for mapping and association studies, assessment of plasticity and adaptive mechanisms, and identification of genetic markers for quantitative trait loci (QTL).

Materials and methods

Plant material and RNA isolation

We obtained root segments from five aspen genets in two US states (three individuals from a site in Arizona and two from a site in Montana; Table 1). Root segments were propagated following Schier (1978), and the shoots were grown in a controlled greenhouse (currently housed at the USDA Forest Sciences Laboratory in Logan, UT, USA). Each genet was represented by a single ramet. From each of the five ramets in the greenhouse, we extracted total RNA from a single undamaged, mature, mid-season leaf. We used 75 mg of leaf tissue from the proximal end of the leaf, avoiding the mid-vein. We isolated RNA using a Spectrum Plant Total RNA kit (Sigma-Aldrich), with a DNase digestion step, and the final product was eluted twice from each column.

Table 1 Geographic sources of root cuttings used for greenhouse common garden propagation

Root sample	Collection date	Location	Geographic coordinates
MT-5	17Nov09	Trout Creek, MT, USA	Latitude: 47°52'2.454" N Longitude: 115°36'42.078" W
MT-6	16Nov09	Trout Creek, MT, USA	Latitude: 47°53'24.630" N Longitude: 115°37' 33.050" W
MT-8	16Nov09	Trout Creek, MT, USA	Latitude: 47° 53' 28.169" N Longitude: 115° 37' 14.896" W
AZ-8	20Oct09	Mt. Lemon, AZ, USA	Latitude: 32°25'7.368" N Longitude: 110°43'55.673" W
AZ-9	20Oct09	Mt. Lemon, AZ, USA	Latitude: 32°27'7.721" N Longitude: 110°46'59.474" W

We examined quantity and quality of RNA using a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA).

cDNA library preparation

Sera-Mag Oligo(dT) beads (Thermo Scientific, Fremont, CA, USA) were used to enrich poly-adenylated RNA by hybridization, following the manufacturer's protocol, and the enriched RNA was resuspended in 50 µl of 10 mM Tris–HCl. Library preparation followed the Illumina GAI RNA-Seq protocol (San Diego, CA; part no. 1004898 Rev. A), except that custom barcoded adapters were used during ligation steps (Cronn et al. 2008). We used PCR to enrich libraries for fragments containing adapters ligated in the correct orientation. Following each enzymatic step, a cleaning step was performed using AMPure XP beads (Agencourt, Danvers, MA, USA). Libraries were quantified using a Qubit fluorometer and their quality validated using an Agilent 2100 Bioanalyser (Santa Clara, CA, USA). The five libraries were pooled in equimolar quantities for a 5 pM final concentration and run on three separate lanes of the Illumina sequencer (two of these lanes were part of the same flow cell).

High-throughput sequencing

Illumina high-throughput sequencing was performed at the Oregon State University Center for Genome Research and

Table 2 Summary of reads generated from Illumina sequencing for five *Populus tremuloides* individuals from two populations in Western North America

Individual	Reads	Average length	Total bases
Montana-5	9,754,551	72.25	704,806,138
Montana-6	8,682,585	72.4	628,655,398
Montana-8	5,418,515	72.04	390,372,241
Arizona-8	10,531,371	72.16	759,894,435
Arizona-9	9,355,761	71.86	672,266,046
Combined	43,742,783	72.15	3,155,994,258

Biocomputing, on an Illumina GAIx using 80 bp single-end indexing runs, resulting in 73-bp sequences and a 6-bp index. The samples were sequenced across three Illumina lanes. Base calling was performed in Casava v1.8 (Illumina, San Diego, CA, USA). We culled low-quality reads using the Illumina analysis pipeline. We then used custom perl scripts (<http://brianknaus.com>) for barcode (index) deconvolution. The final 43.7 M reads remaining after passing quality control steps are summarized in Table 2.

Transcriptome assembly

We trimmed adapter sequences and binned single-end reads by barcode (deconvolution) and aligned them to the *P. trichocarpa* v2 transcript assemblies [45,033 protein-coding transcripts (file name: Ptrichocarpa_156_transcript.fa.gz); Phytozome v8.0] using CLC Genomics Workbench (v4.9; CLC Bio, Cambridge, MA, USA). We mapped to transcript assemblies rather than to the genome to avoid problems with gapped alignments at splice junctions. The maximum gap and mismatch count were set to 2, and insertion and deletion costs were set to 3, with a minimum contig length of 200 bp. Length fraction and similarity parameters were set to 0.5 and 0.8, respectively.

Gene ontology (GO) annotations were assigned using Blast2GO (Conesa et al. 2005). GO terms were searched and retrieved to match the results from a blastx search

Table 3 Summary of blast results of the unmapped reads in *Populus tremuloides* from the reference-guided assembly. A *de novo* assembly of these reads resulted in 13,544 contigs

	Number of contigs	Percent of total
<i>Populus</i>	3394	25.1
Other plants	420	3.1
Bacterial origin	5984	44.2
Mammalian	117	0.9
Other	38	0.3
No significant match	3591	26.6

against the NCBI nr database (maximum of 10 hits for each contig with an e-value cutoff of $1 \times e^{-10}$), and the annotation function of Blast2GO was used to select final GO terms for each contig. Finally, the terms were analyzed through the GO Slim for plants function of Blast2GO to provide a broad overview of the gene product functions contained within the ontology content.

We also used CLC to perform a de novo assembly of those unmapped sequences that did not produce significant alignments to the reference in the *P. trichocarpa* reference-guided assembly above. CLC's de novo assembler performs a two-step assembly by first assembling simple contigs using De Bruijn graphs and then mapping back all reads using the simple contigs as references. The mismatch count was set to 2, and insertion and deletion costs were set to 3, with a minimum contig length of 200 bp. Length fraction and similarity parameters were set to 0.5 and 0.8, respectively. Nonspecific matches were discarded because we chose to be conservative in our treatment of reads that mapped to more than one position. We then performed a blast search of the contigs resulting from this de novo assembly against the

NCBI "nr" database, to identify potential contaminants (see Table 3).

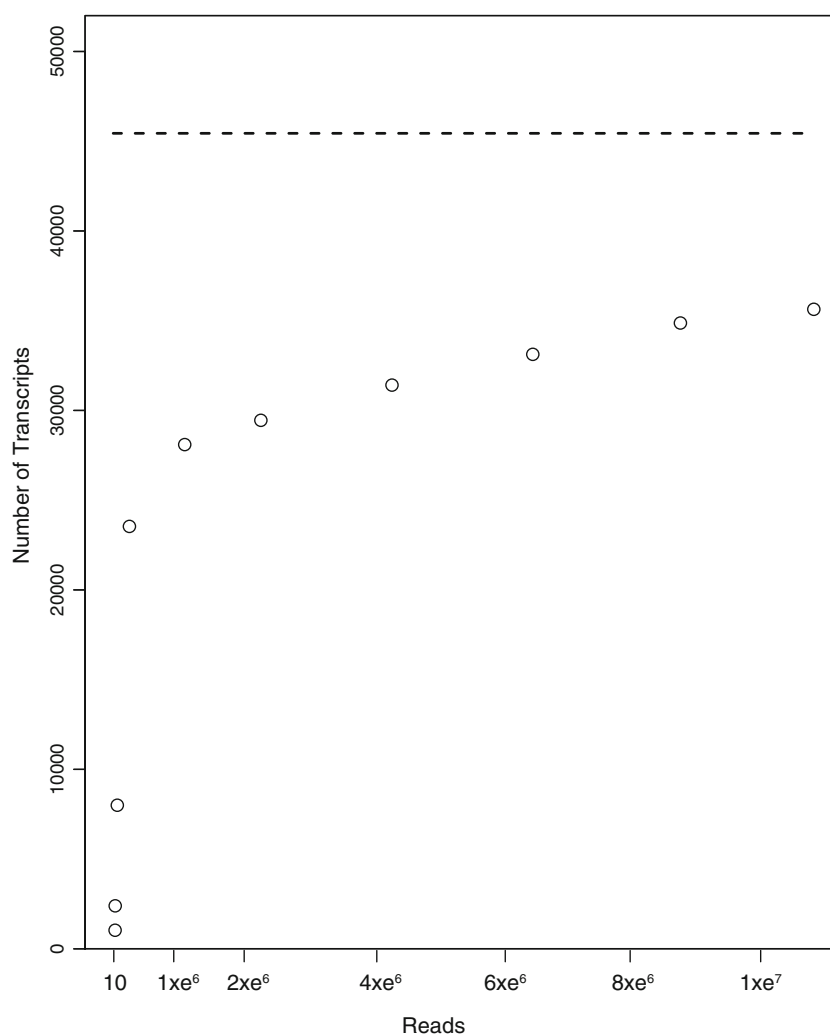
SNP searching

We used our reference-guided assembly (above) to detect SNPs using CLC Genomics Workbench to detect SNPs occurring between *P. tremuloides* and *P. trichocarpa* as well as SNPs between the two sampled populations of *P. tremuloides*. Only variants with a minimum coverage of $8\times$ and a minimum variant count (same as minimum allele depth) of 4 were called. Maximum coverage for the SNP search was set at $500\times$ in order to avoid highly repetitive regions of the genome. The resulting search for variable sites was filtered for SNPs specific to aspen and also filtered for SNPs within open reading frames (ORFs) of more than 200 bp.

Differential expression

We used our raw reads to explore differential expression among *P. tremuloides* genets within and among sampling sites

Fig. 1 Example of a rarefaction plot for one *Populus tremuloides* clone (from Arizona). The raw reads were plotted against the *P. trichocarpa* transcript assembly (v2). Line indicates total number of whole-plant transcripts in reference (*P. trichocarpa*)



representing different latitudes. In order to quantify transcript abundance, the 73 bp reads were mapped to the *P. trichocarpa*

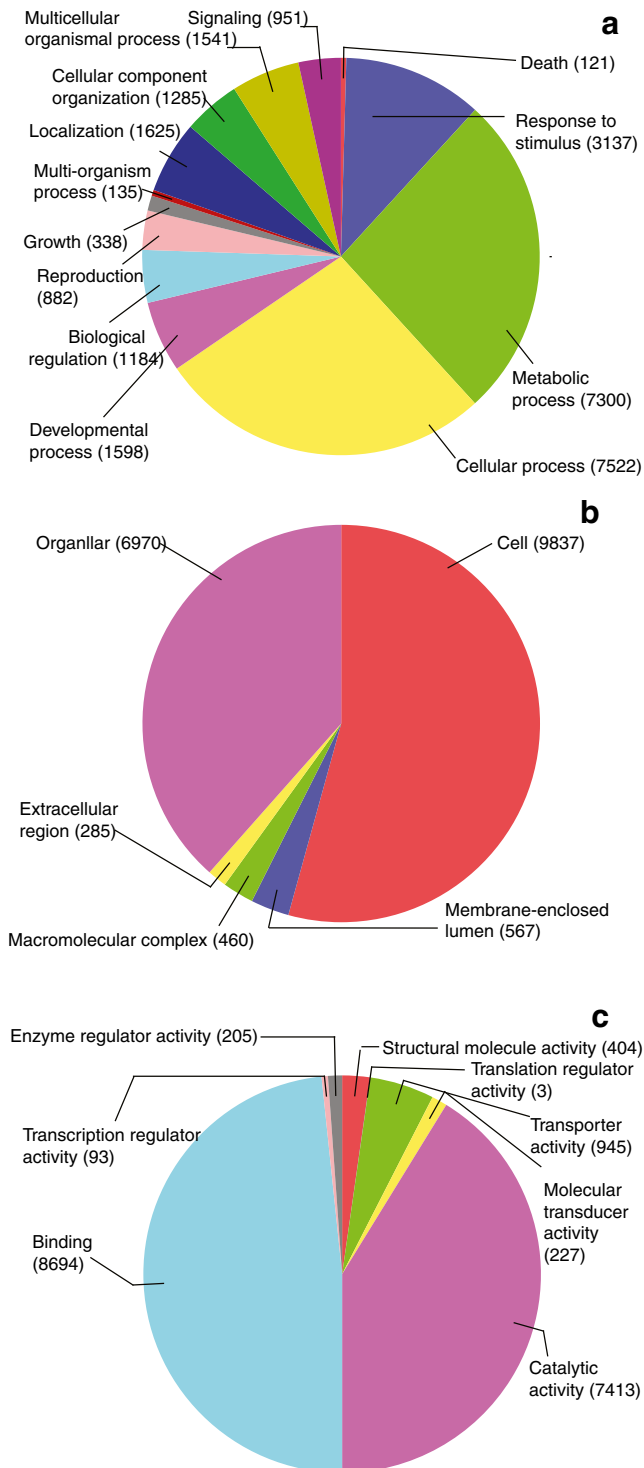


Fig. 2 Distribution of assembled sequences of *Populus tremuloides* in three main GO categories. The annotated sequences were run through GO-Slim to provide a high level summary of functions. **a** Biological processes (10,556 contigs annotated). **b** Cellular components (12,593 contigs annotated). **c** Molecular function (10,503 contigs annotated). Note that a contig can be assigned to more than one category

reference using the reference-based aligner Bowtie version 0.12.7 (<http://bowtie-bio.sourceforge.net/index.shtml>), resulting in SAM format files. Bowtie deals with multimapping reads by assigning them randomly among transcript sites (Langmead et al. 2009). The maximum number of mismatches was set to 2 in the $-n$ alignment mode. The number of reads mapped to each transcript was tallied using a custom perl script. Unbalanced library sizes were down-sampled using the method implemented in NBPSeg (Di et al. 2011); the smallest library size was identified, and the other libraries were down-sampled to this number of reads, based on a probabilistic step, resulting in approximately equal library sizes. Once libraries were normalized to balanced sizes, comparisons of differential expression between genets were made using Fisher's exact test (Fisher 1922), which is appropriate for small sample sizes (1 or 2). We corrected for multiple comparisons by setting the false discovery rate to 0.05 (Benjamini and Hochberg 1995). For comparison across the two sampling sites, the dispersion parameter for the negative binomial distribution was estimated for each transcript using the Bioconductor package edgeR (Robinson et al. 2010) using the “estimateTagwiseDisp” function, and negative-binomial exact tests were fit for each gene using the dispersion parameters estimated above. Results from the statistical analysis of differential expression are presented with MA plots (log ratio versus log abundance) using the plotSmear function in EdgeR (Robinson et al. 2010).

Results and discussion

In this study, we characterized the leaf transcriptome of *P. tremuloides*, an ecologically important tree species in western North America. Although we lacked biological and experimental replication, our study design included five individual aspen genotypes from two distinct and latitudinally separated sampling sites. This enabled us to make the first examination of variation in gene expression in aspen based on RNA-Seq data. We mapped the trimmed 73-bp sequence reads to the *P. trichocarpa* reference genome (Tuskan et al. 2006), from which we generated a functional characterization of the transcriptome and a set of putative SNPs. We also characterized the differences in gene expression patterns between sampled regions and among individual genets.

Transcriptome assembly and analysis

We generated 43,742,783 sequence reads from three independent runs of pooled and barcoded RNA-Seq libraries (Table 2). All reads are deposited at the NCBI Sequence Read Archive (accession number SRA057223). Assuming a high degree of sequence and transcript similarity (Cervera et al. 2005; Chen et al. 2010; Chen et al. 2007; Cole 2005;

Table 4 Frequency of SNPs in *Populus tremuloides* by mutation type

	SNP type	Count	Percent count	Percent total
Transitions	C/T	54,477	30.8	58.2
	A/G	48,435	27.4	
Transversions	A/T	19,171	10.8	36.1
	C/G	12,021	6.8	
	A/C	16,500	9.3	
	G/T	16,261	9.2	
Complex SNP		9,832	5.6	5.6
	TOTAL	176,697		

Hamzeh and Dayanandan 2004; Hamzeh et al. 2006; Unneberg et al. 2005), sequences from all *P. tremuloides* individuals were combined into a single reference-mapped assembly against the 45,033 annotated, protein-coding transcripts of *P. trichocarpa*. A large proportion of the total sequencing reads, 84 % (~37M), were successfully mapped to the reference library with an average depth of ~66X. The 6,813,822 sequences of the ~44M total reads were not mapped to the reference transcripts in this assembly. We detected 7.8M nonspecific reads, mapping to more than one *P. trichocarpa* transcript. In these cases, we assigned the reads randomly. Thus, we may have slightly

overestimated the total coverage and underestimated the sequencing depth.

The above assembly resulted in the pooled *P. tremuloides* sequencing reads mapping to 38,177 of the 45,033 *P. trichocarpa* reference transcripts with at least a single read in each contig (of these, 20,906 had an average read depth of 4×). Even when this was examined for each of the five genets individually, a large proportion of the *P. trichocarpa* transcripts were recovered: MT5, 32,963; MT6, 32,049; MT8, 31,272; AZ8, 33,109; AZ9, 34,007 total transcripts for each genet, respectively. This demonstrates our ability to detect low abundance transcripts using RNA-Seq, with which we have captured a large proportion of expressed genes in the *P. tremuloides* leaf transcriptome, and that our aspen leaf transcriptome, contains a substantial proportion of the total annotated *P. trichocarpa* transcripts.

In order to test whether the depth of our individual RNA-Seq libraries was sufficient to maximize recovery of the expressed portion of the aspen leaf genome, we randomly sampled our *P. tremuloides* reads and counted the number that corresponded to the reference (*P. trichocarpa*) transcripts. We plotted the number of reads for each individual separately. All individuals show the same asymptote for transcripts recovered after the initial 2 million reads (Fig. 1). The results of these rarefaction plots indicate that

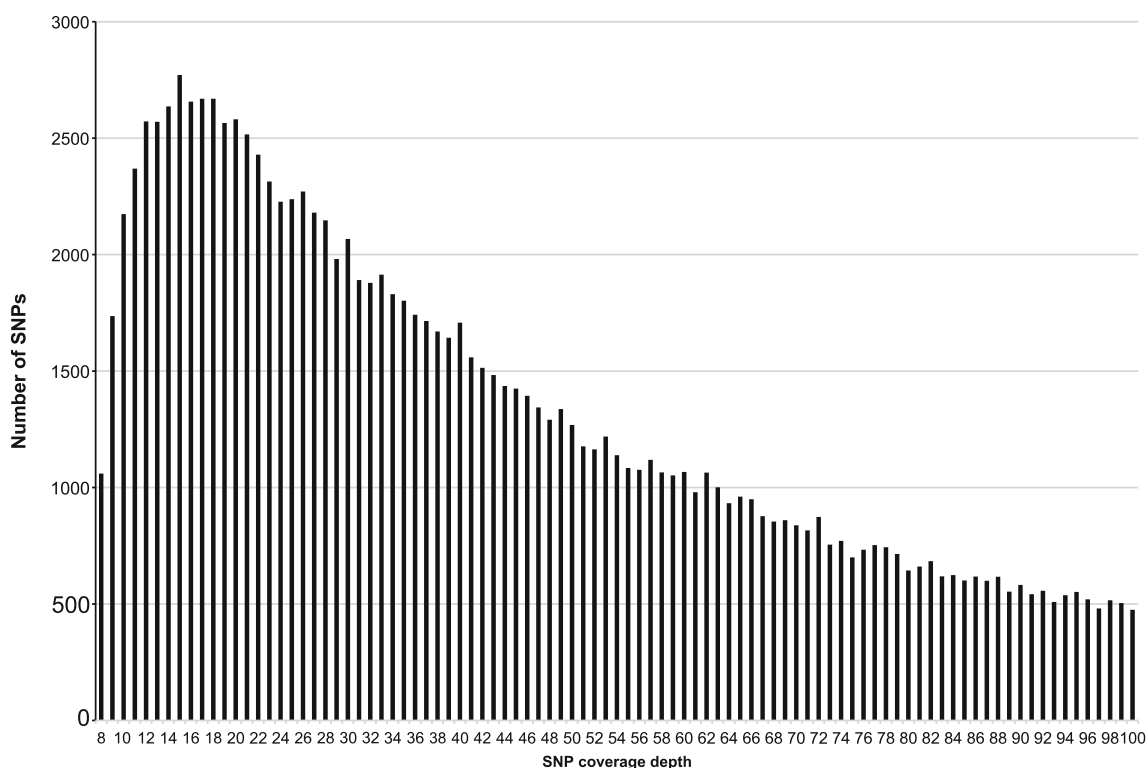


Fig. 3 Distribution of SNPs by read depth in *Populus tremuloides* (176,697 total SNPs). SNPs with read coverage >100× are not shown. The coverage depth for each SNP is the total number of reads that includes representatives for both alleles present in each diploid individual

our current level of sequencing for each of the genets is sufficient to recover a reasonable sampling of the entire leaf transcriptome and that additional sequencing is likely to yield only a small number of additional expressed transcripts (likely very low-expression level genes in this portion of the transcriptome).

We assigned GO annotations to the assembled contigs with a significant match within the Viridiplantae using Blast2GO (Conesa et al. 2005). Of the 20,906 contigs that we analyzed, 15,421 were annotated with a GO term. The distribution of GO terms assigned to the contigs is presented in Fig. 2.

The de novo assembly of the unmapped reads resulted in 13,544 assembled contigs >200 bp with an N50 of 493 bp (4,042,046 unmapped reads assembled). An initial search using blastx against the NCBI “nr” database was performed (with a minimum *e* value cutoff of 1×10^{-10} ; Table 3). Approximately 27 % of the de novo assembled contigs had no significant match in the NCBI “nr” database, whereas 28 % had significant homology to at least one angiosperm sequence (25 % of these were reported from the genus *Populus*). Not surprisingly, almost half (44 %) of the contigs assembled from the unmapped reads showed sequence similarity (based on the above blast criteria) with bacteria, suggesting low-level contamination (fewer than 1 % of all high-throughput sequencing reads).

SNP discovery and genotyping

We found 396,909 total SNPs, of which 220,212 (55.5 %) were fixed differences between *P. tremuloides* and *P. trichocarpa*. Of these differences, 177,880 (80.8 %) were within ORFs and 83,201 (46.8 %) of these were synonymous differences. Within *P. tremuloides*, we found 176,697 SNPs, equivalent to 1 SNP/224 bp. These putative SNPs include heterozygous sites within aspen individuals as well as nucleotide substitutions among the individuals. Of the 146,970 aspen SNPs detected in ORFs, 69,888 (47.6 %) were synonymous substitutions. We detected 77,082 nonsynonymous substitutions, of which 59,866 were found in transcripts with only a single nonoverlapping ORF. Thus, the proportion of nonsynonymous SNPs within *P. tremuloides* was similar to that between *P. tremuloides* and *P. trichocarpa*. Within the aspen leaf transcriptome, we found levels of SNP polymorphism similar to those found in studies of other plant species (e.g., Bajgain et al. 2011; Novaes et al. 2008; Coppe et al. 2010; Gerald et al. 2011). A summary of SNP frequency and mutation type is provided in Table 4 and the distribution of read coverage depth in Fig. 3. We have made the full list of putative SNPs within aspen publically available at the AspenDB website (<http://aspen-db.uga.edu/>), which is part of the Dendrome Project (<http://dendrome.ucdavis.edu/>). The data reported for each putative

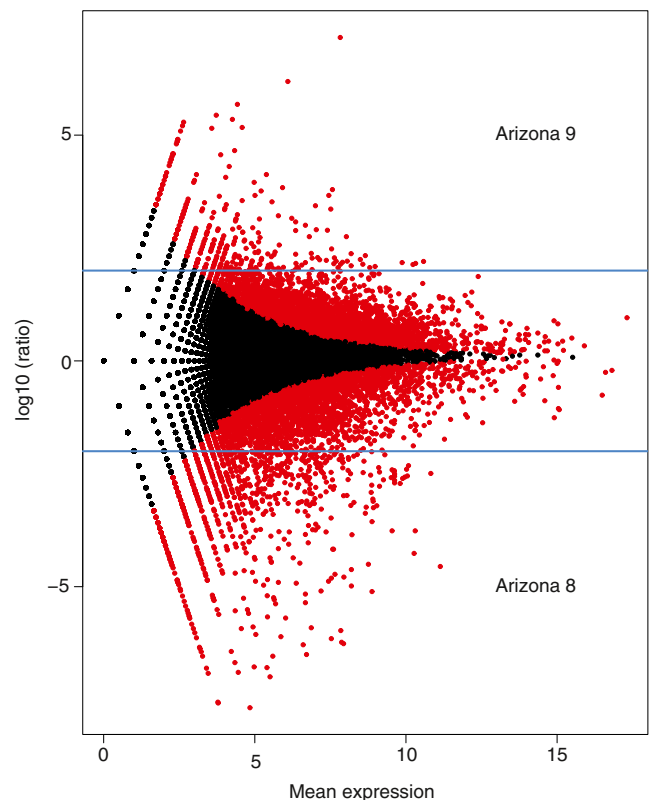


Fig. 4 Example MA plot of differentially expressed genes in comparisons of two *Populus tremuloides* individuals from the same population using Fisher's exact test to examine significance in expression difference at a false discovery rate value 0.05. (7,886 genes; see Table 4 for complete pairwise results)

SNP include the position of the SNP relative to the *P. trichocarpa* genome assembly, the reference base, the nucleotide variants and their frequencies present in the reads mapped here, and the coverage depth for each nucleotide site. These putative SNPs should help guide future population studies, comparative analyses, or even possibly development of QTLs within aspen, and add to the growing SNP resources for aspen (Kelleher et al. 2012) and other *Populus* species (Fladung and Buschbom 2009). Further characterization and validation of these data will be invaluable for research related to aspen.

Table 5 Pairwise comparison results of the numbers of significantly differentially expressed genes among *Populus tremuloides* individuals: significance was based on Fisher's exact test with a false discovery rate of 0.05

	MT5	MT6	MT8	AZ8	AZ9
MT5	–				
MT6	6,244	–			
MT8	8,192	84,18	–		
AZ8	10,234	10,223	3,723	–	
AZ9	7,134	7,402	5,679	7,886	–

Differential expression

To characterize the variation present among *P. tremuloides* samples, we analyzed gene expression differences between the RNA-Seq libraries and determined whether there were statistically significant differences in expression levels of transcripts across samples. We made two separate comparisons: one among the five genets and a second comparison between the two geographic regions. All of our RNA-Seq libraries were carefully selected from leaves at the same developmental stages. Collected at the same time on the same date from trees grown in a greenhouse common garden.

One characteristic of the individual pairwise comparisons for differential expression was the large number of genes that differed statistically in expression among sampled genets, whether or not they originated from geographically distant regions (Fig. 4 and Table 5). In other studies of differential gene expression, the high sensitivity of the RNA-Seq method has allowed for a new level of detail and detection when considering closely related genotypes (Zenoni et al. 2010; Kawahara-Miki et al. 2011). The amount of expression variation between individuals of

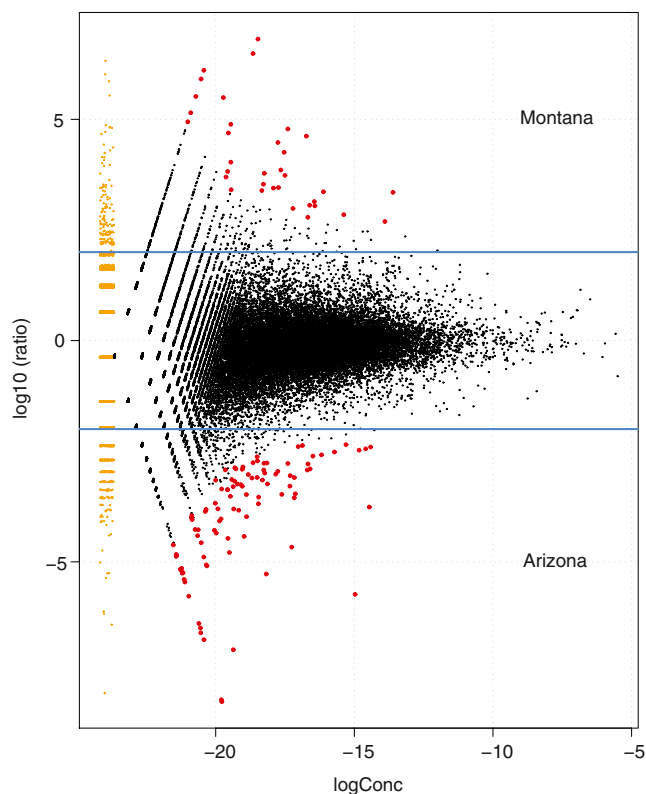


Fig. 5 Differential expression between *Populus tremuloides* populations (two and three individuals from Arizona and Montana, respectively). The MA (log ratio versus log abundance) plot highlights 154 statistically significant transcripts at a false discovery rate value of 0.05

aspen is higher than similar comparisons of gene expression in other plants at this taxonomic scale (Cohen et al. 2010; Matzkin 2012, 2008; Zhou et al. 2012; Beritognolo et al. 2011; Kawahara-Miki et al. 2011), but it is consistent with results from studies that used other measures of genotypic diversity in *P. tremuloides* (Chong et al. 1994; Cole 2005; Lund et al. 1992). We suspect that the high interindividual expression variability may be due in part to our lack of replication and suggest that future transcriptome studies include replication of populations at contrasting latitudes, increased representation of individuals within populations, as well as within-individual (technical) replication. Otherwise, it is difficult to determine whether the large numbers

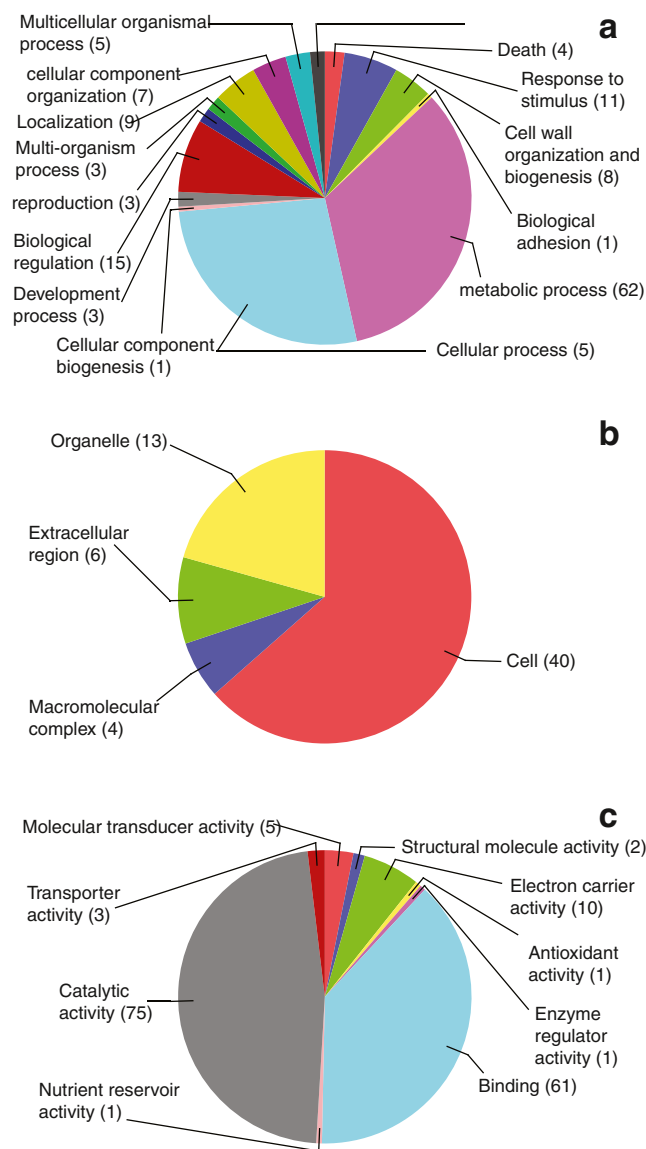


Fig. 6 Gene ontology (GO) graphs of differentially expressed genes from pooled individuals of *Populus tremuloides* within two collection sites: Trout Creek, MT, USA and Mount Lemon, AZ, USA. **a** Biological processes. **b** Cellular component. **c** Molecular function

of differentially expressed genes is a genus-wide phenomenon resulting from different sources of unaccounted-for variation such as temporal or mechanical effects, stochastic variation or whether it is simply false positives resulting from lack of replication.

We discovered a small set of genes that were expressed at significantly different levels between two latitudinally separated sampling regions, despite the enormous variation present between individuals (Fig. 5). We further characterized these differentially expressed transcripts by querying against NCBI's "nr" database and assigning GO terms in Blast2GO (Fig. 6). Overall, the GO term contributions were similar to those for total transcripts (Fig. 2), although it appears that the organellar component of differentially expressed genes represents a smaller overall contribution compared to the proportion in the totals.

Conclusions

Aspen has the potential to be an important study system for responses to climate change and its effects across a landscape that spans a large majority of the North American continent. The characterization of the leaf transcriptome signals an important step in the direction of high-throughput comparative genomics in this ecologically significant tree species. We have identified a large number of SNPs in transcribed genes that should prove useful for species-wide population genetic studies of *P. tremuloides*, enabling a range of high-throughput sequencing and genotyping methodologies (e.g., microfluidic SNP genotyping across large number of individuals). Our study of differential gene expression in aspen is a part of a small but growing body of literature that documents differential gene expression between latitudinally separate populations (Gerald et al. 2011; Kawahara-Miki et al. 2011; Beritognolo et al. 2011; Eckert et al. 2009; Ellison et al. 2011) and provides a first look into the differentiation and gene expression changes within the transcriptome of aspen.

Acknowledgments Thanks to Tim Benedict, Mary Lou Fairweather, and E. Pfalzer for sample collections. Thanks to Tara Jennings for laboratory assistance and Chris Sullivan for biocomputing assistance. This research was funded by the USDA Forest Service Western Forest Transcriptome Survey and National Fire Plan (2012-NFP-GSD-1).

References

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

- Bashir A, Bansal V, Bafna V (2010) Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. BMC Genomics 11:385
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Series B (Methodological) 57:289–300
- Beritognolo I, Harfouche A, Brilli F, Prosperini G, Gaudet M, Brosché M, Salani F, Kuzminsky E, Auvinen P, Paulin L (2011) Comparative study of transcriptional and physiological responses to salinity stress in two contrasting *Populus alba* L. genotypes. Tree Physiol 31(12):1335–1355
- Bräutigam A, Mullick T, Schliesky S, Weber APM (2011) Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C(3) and C(4) species. J Exp Bot 62:3093–3102
- Cervera MT, Storme V, Soto A, Ivens B, Van Montagu M, Rajora OP, Boerjan W (2005) Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers. TAG Theoretical Appl Genet 111(7):1440–1456
- Chen JH, Sun H, Wen J, Yang YP (2010) Molecular phylogeny of *Salix* L. (Salicaceae) inferred from three chloroplast datasets and its systematic implications. Taxon 59(1):29–37
- Chen K, Peng Y, Wang Y, Korpelainen H, Li C (2007) Genetic relationships among poplar species in section *Tacamahaca* (*Populus* L.) from western Sichuan, China. Plant Sci 172(2):196–203
- Chong DKX, Yang RC, Yeh FC (1994) Nucleotide divergence between populations of trembling aspen (*Populus tremuloides*) estimated with RAPDS. Curr Genet 26(4):374–376
- Cohen D, Bogeat-Triboulot MB, Tisserant E, Balzergue S, Martin-Magniette ML, Lelandais G, Ningre N, Renou JP, Tamby JP, Le Thiec D (2010) Comparative transcriptomics of drought responses in *Populus*: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes. BMC Genomics 11(1):630
- Cole CT (2005) Allelic and population variation of microsatellite loci in aspen (*Populus tremuloides*). New Phytol 167:155–164
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676
- Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, Bernatchez L, Zane L, Bortoluzzi S (2010) Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EelBase opens new perspectives for the study of the critically endangered European eel. BMC Genomics 11:635
- 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. Nucl Acids Res 36:e122
- Di Y, Schafer DW, Cumbie JS, Chang JH (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol 10:art24
- Eckenwalder J (1977) North American cottonwoods (*Populus*, Salicaceae) of sections *Abaso* and *Aigeros*. J Arnold Arboret 58:193–208
- Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, Clair JBS, Neale DB (2009) Association genetics of coastal Douglas Fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. cold-hardiness related traits. Genetics 182(4):1289–1302
- Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW (2011) Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Nat Acad Sci (USA) 108:2831–2836

- Fang Z, Cui X (2011) Design and validation issues in RNA-seq experiments. *Brief Bioinform* 12:280–287
- Fisher RA (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P. *J Roy Stat Soc* 85(1):87–94
- Fladung M, Buschbom J (2009) Identification of single nucleotide polymorphisms in different *Populus* species. *Trees-Struct Funct* 23(6):1199–1212
- Frey BR, Lieffers VJ, Landhausser SM (2004) Predicting landscape patterns of aspen dieback: mechanisms and knowledge gaps. *Canad J Forest Res* 34(7):1379–1390
- Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M, Birol I, Jones SJM, Cronk QCB, Douglas CJ (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Mol Ecol Resour* 11:81–92
- Grisel N, Zoller S, Künzli-Gontarczyk M, Lampart T, Münsterkötter M, Brunner I, Bovet L, Métraux J-P, Sperisen C (2010) Transcriptome responses to aluminum stress in roots of aspen (*Populus tremula*). *BMC Plant Biol* 10:15
- Hamanishi ET, Raj S, Wilkins O, Thomas BR, Mansfield SD, Plant AL, Campbell MM (2010) Intraspecific variation in the *Populus balsamifera* drought transcriptome. *Plant Cell Env* 33:1742–1755
- Hamzeh M, Dayanandan S (2004) Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast *trnT-trnF* region and nuclear rDNA. *Amer J Bot* 91(9):1398–1408
- Hamzeh M, Périnet P, Dayanandan S (2006) Genetic relationships among species of *Populus* (Salicaceae) based on nuclear genomic data. *J Torrey Bot Soc* 133(4):519–527
- Hogg EH, Brandt JP, Michaelian M (2008) Impacts of a regional drought on the productivity, dieback, and biomass of western Canadian aspen forests. *Canad J Forest Res* 38:1373–1384
- Hogg EH, Hurdle PA (1995) The aspen parkland in western Canada: a dry-climate analogue for the future boreal forest? *Water Air Soil Pollut* 82(1):391–400
- Isebrands JG, McDonal EP, Kruger E, Hendrey G, Percy K, Pregitzer K, Sober J, Karnosky DF (2001) Growth responses of *Populus tremuloides* clones to interacting elevated carbon dioxide and tropospheric ozone. *Environ Pollut* 115:359–371
- Iverson LR, Prasad AM (2002) Potential redistribution of tree species habitat under five climate change scenarios in the eastern US. *For Ecol Manag* 155:205–222
- Jiang H, Peng S, Zhang S, Li X, Korpelainen H, Li C (2012) Transcriptional profiling analysis in *Populus yunnanensis* provides insights into molecular mechanisms of sexual differences in salinity tolerance. *J Exp Bot* 63:3709–3726
- Jones JR (1985) Distribution. In: DeByle NV, Winokur RP (eds) *Aspen: ecology and management in the western United States*. General Technical Report RM-1 19. US Department of Agriculture, Forest Service, Fort Collins, pp 9–10
- Kawahara-Miki R, Wada K, Azuma N, Chiba S (2011) Expression profiling without genome sequence information in a non-model species, pandalid shrimp (*Pandalus latirostris*), by next-generation sequencing. *PLoS One* 6(10):e26043
- Kelleher CT, Wilkin J, Zhuang J, Cortés AJ, Quintero ÁLP, Gallagher TF, Bohlmann J, Douglas CJ, Ellis BE, Ritland K (2012) SNP discovery, gene diversity, and linkage disequilibrium in wild populations of *Populus tremuloides*. *Tree Genet Genomes* 8:821–829
- Kurz WA, Apps MJ (1999) A 70-year retrospective analysis of carbon fluxes in the Canadian forest sector. *Ecol Appl* 9(2):526–547
- Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25
- Larsen PE, Sreedasyam A, Trivedi G, Podila GK, Cseke LJ, Collart FR (2011) Using next generation transcriptome sequencing to predict an ectomycorrhizal metabolome. *BMC Systems Biol* 5:14
- Liu Z, Furnier GR (1993) Comparison of allozyme, RFLP, and RAPD markers for revealing genetic variation within and between trembling aspen and bigtooth aspen. *Theoret Appl Genet* 87(1):97–105
- Lund ST, Furnier GR, Mohn CA (1992) Isozyme variation in quaking aspen in Minnesota. *Canad J Forest Res* 22(4):521–524
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509–1517
- Matzkin LM (2008) The molecular basis of host adaptation in cactophilic *Drosophila*: molecular evolution of a glutathione S-transferase gene (*GstD1*) in *Drosophila mojavensis*. *Genetics* 178(2):1073–1083
- Matzkin LM (2012) Population transcriptomics of cactus host shifts in *Drosophila mojavensis*. *Mol Ecol* 21:2428–2439
- Metzker ML (2009) Sequencing technologies: the next generation. *Nature Rev Genet* 11(1):31–46
- Mitton J, Grant M (1996) Genetic variation and the natural history of quaking aspen. *Bioscience* 46:25–31
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, NY)* 320:1344–1349
- Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9(1):312
- Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella Typhi*. *PLoS Genetics* 5(7):e1000569
- Qiu Q, Ma T, Hu Q, Liu B, Wu Y, Zhou H, Wang Q, Wang J, Liu J (2011) Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. *Tree Physiol* 31(4):452–461
- Quesada T, Li Z, Dervinis C, Li Y, Bock P, Tuskan GA, Casella G, Davis JM, Kirst M (2008) Comparative analysis of the transcriptomes of *Populus trichocarpa* and *Arabidopsis thaliana* suggests extensive evolution of gene expression regulation in angiosperms. *New Phytol* 180:408–420
- Rehfeldt GE, Ferguson DE, Crookston NL (2009) Aspen, climate, and sudden decline in western USA. *For Ecol Manag* 258:2353–2364
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140
- Schier GA (1978) Vegetative propagation of Rocky Mountain aspen. USDA Forest Service General Technical Report INT-44
- Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC (2010) RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol* 10:16
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313(5793):1596
- Unneberg P, Strömberg M, Lundeberg J, Jansson S, Sterky F (2005) Analysis of 70,000 EST sequences to study divergence between two closely related *Populus* species. *Tree Genet Genomes* 1(3):109–115
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev Genet* 10:57–63

- Worrall JJ, Egeland L, Eager T, Mask RA, Johnson EW, Kemp PA, Shepperd WD (2008) Rapid mortality of *Populus tremuloides* in southwestern Colorado, USA. For Ecol Manag 255(3):686–696
- Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. Plant Physiol 152(4):1787–1795
- Zhou S, Campbell TG, Stone EA, Mackay TFC, Anholt RRH (2012) Phenotypic plasticity of the *Drosophila* transcriptome. PLoS Genetics 8(3):e1002593