Sequencing and De Novo Transcriptome Assembly of Brachypodium sylvaticum (Poaceae)

Samuel E. Fox  
Oregon State University

Justin Preece  
Oregon State University

Jeffrey A. Kimbrel  
Joint BioEnergy Institute

Gina L. Marchini

Abigail Sage  
Oregon State University

See next page for additional authors

Follow this and additional works at: https://pdxscholar.library.pdx.edu/bio_fac

Part of the Biology Commons, and the Genomics Commons

Let us know how access to this document benefits you.

Citation Details
Fox, Samuel E., et al. "Sequencing and De Novo Transcriptome Assembly of Brachypodium sylvaticum (Poaceae)." Applications in Plant Sciences

This Article is brought to you for free and open access. It has been accepted for inclusion in Biology Faculty Publications and Presentations by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Authors
Samuel E. Fox, Justin Preece, Jeffrey A. Kimbrel, Gina L. Marchini, Abigail Sage, Ken Youens-Clark, Mitchell B. Cruzan, and Pankaj Jaiswal

This article is available at PDXScholar: https://pdxscholar.library.pdx.edu/bio_fac/29
Sequencing and De Novo Transcriptome Assembly of *Brachypodium sylvaticum* (Poaceae)

Author(s): Samuel E. Fox, Justin Preece, Jeffrey A. Kimbrel, Gina L. Marchini, Abigail Sage, Ken Youens-Clark, Mitchell B. Cruzan, and Pankaj Jaiswal

Source: Applications in Plant Sciences, 1(3).

Published By: Botanical Society of America

DOI: [http://dx.doi.org/10.3732/apps.1200011](http://dx.doi.org/10.3732/apps.1200011)

Genomic Resources Note

Sequencing and de novo transcriptome assembly of *Brachypodium sylvaticum* (Poaceae)\(^1\)

Samuel E. Fox\(^2\), Justin Preece\(^2\), Jeffrey A. Kimbrel\(^3\), Gina L. Marchini\(^4\), Abigail Sage\(^2\), Ken Youens-Clark\(^5\), Mitchell B. Cruzan\(^4\), and Pankaj Jaiswal\(^2,6\)

\(^2\)Department of Botany and Plant Pathology and Center for Genome Research and Biocomputing, 2082 Cordley Hall, Oregon State University, Corvallis, Oregon 97331 USA; \(^3\)Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, 5885 Hollis Street, Fourth Floor, Emeryville, California 94608 USA; \(^4\)Department of Biology, Portland State University, Portland, Oregon 97201 USA; and \(^5\)Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 USA

\(\cdot\) \textit{Premise of the study:} We report the de novo assembly and characterization of the transcriptomes of *Brachypodium sylvaticum* (slender false-brome) accessions from native populations of Spain and Greece, and an invasive population west of Corvallis, Oregon, USA.

\(\cdot\) \textit{Methods and Results:} More than 350 million sequence reads from the mRNA libraries prepared from three *B. sylvaticum* genotypes were assembled into 120,091 (Corvallis), 104,950 (Spain), and 177,682 (Greece) transcript contigs. In comparison with the *B. distachyon* Bd21 reference genome and GenBank protein sequences, we estimate \(>90\%\) exome coverage for *B. sylvaticum*. The transcripts were assigned Gene Ontology and InterPro annotations. *Brachypodium sylvaticum* sequence reads aligned against the Bd21 genome revealed 394,654 single-nucleotide polymorphisms (SNPs) and \(>20,000\) simple sequence repeat (SSR) DNA sites.

\(\cdot\) \textit{Conclusions:} To our knowledge, this is the first report of transcriptome sequencing of invasive plant species with a closely related sequenced reference genome. The sequences and identified SNP variant and SSR sites will provide tools for developing novel genetic markers for use in genotyping and characterization of invasive behavior of *B. sylvaticum*.

**Key words:** *Brachypodium sylvaticum*; comparative genomics; de novo transcriptome; invasive species; simple sequence repeat (SSR); single-nucleotide polymorphism (SNP).

*Brachypodium sylvaticum* (Huds.) P. Beauv. (slender false-brome; Poaceae), with an estimated genome size of 470 Mb and 17 chromosomes (Foote et al., 2004), is a perennial bunchgrass native to Europe, Asia, and North Africa and is closely related to the bioenergy feedstock model grass *B. distachyon* (L.) P. Beauv. (Wolny et al., 2011), which has a sequenced genome of 272 Mb and five chromosomes. In its native range, *B. sylvaticum* occurs in habitats ranging from forest understory to open meadows and tolerates conditions from full shade to full sun (Holten, 1980; Long, 1989; Aarrestad, 2000; Kirby and Thomas, 2000). In the United States, *B. sylvaticum* is invasive and listed as a noxious weed covering the west coast of California, Oregon, and Washington (Oregon Department of Agriculture, 2009; Washington State Department of Agriculture, 2009; Lionakis Meyer and Effenberger, 2010). It is also expanding into the eastern United States where it has been reported in Missouri and Virginia (Roy, 2010). In Oregon, *B. sylvaticum* forms thick monocultures in open forests at elevations from nearly sea level to approximately 1200 m. It threatens the endangered Oregon Willamette Valley oak savanna ecosystem by replacing the native flora and reduces habitat for rare butterflies (Kaye and Blakeley-Smith, 2006; Severns and Warren, 2008). False brome is shade tolerant (Murchie and Horton, 1997; Holmes et al., 2010), which makes it a particularly dangerous invasive threat to undisturbed habitats (Martin et al., 2009).

False brome was first introduced into Oregon via plant introduction studies in the early part of the twentieth century, and later identified and collected from the wild in 1939 (Chambers, 1966; Kaye and Blakeley-Smith, 2006). Field trials for exotic grasses were performed by the United States Department of Agriculture (USDA) in Corvallis at the Oregon State University facilities, and *B. sylvaticum* was widely planted to “improve range” throughout the western United States (Hull, 1974). Oregon State University herbarium records indicate that two separate experimental gardens were established in Eugene and Corvallis, Oregon. Genetic profiling with microsatellite markers confirm these introductions were independent and that they probably consisted of the same set of multiple accessions from the native range in Europe that had been collected by the USDA Division of Plant Introduction (Rosenthal et al., 2008). Accessions in each of these two plantings have crossed, and the invasive plants that are now spreading across Oregon forests are recombinant products of hybridization. *Brachypodium sylvaticum*

---

\(^{1}\) Manuscript received 21 December 2012; revision accepted 7 February 2013.

The authors thank members of the Center for Genome Research and Biocomputing at Oregon State University for sequencing and computational support, Tanya Cheeke for suggestions on the manuscript, and Cathy Gresham and Fiona McCarthy for InterProScan analysis. Thanks to Sharon Wei and Doreen Ware from Gramene database, and Henry Priest and Todd Mockler from BrachyBase for help with the integration of data sets in the respective databases. Funded by laboratory startup funds provided to P.J.

\(^{6}\) Author for correspondence: jaiswalp@science.oregonstate.edu

doi:10.3732/apps.1200011

Applications in Plant Sciences 2013 1(3): 1200011; http://www.bioone.org/loi/apps © 2013 Fox et al. Published by the Botanical Society of America. This work is licensed under a Creative Commons Attribution License (CC-BY-NC-SA).
has thus become increasingly common in the past 15 yr (Chambers, 1966; Kaye and Blakeley-Smith, 2006; Rosenthal et al., 2008). The introductions in Corvallis and Eugene retain unique marker signatures, but Bayesian cluster analyses indicate that similar sets of native accessions from Western Europe contributed to the hybrid genotypes that are spreading from each introduction location (Rosenthal et al., 2008).

Despite the economic and environmental impact of *B. sylvaticum*, there remains a lack of adequate genomic resources for the study of the invasive populations. To thoroughly investigate genetic differences between invasive and native populations, a reference transcriptome specific for *B. sylvaticum* is needed to precisely align, map, and interrogate gene sequences. The major aim of this study was to assemble, annotate, and characterize a high-quality reference transcriptome that will enable researchers to assess gene expression levels, conduct comparative analyses, and identify putative single-nucleotide polymorphism (SNP) and simple sequence repeat (SSR) sequence sites in the genomes of *B. sylvaticum* populations for developing genetic markers to be used in future genotyping, identification, and genetic tracking studies.

Over the past several years, next-generation sequencing (NGS) has emerged as a low-cost, large-scale, fast, and accurate approach for de novo transcriptome sequencing (reviewed in Ward et al., 2012). Moreover, the tremendous depth of coverage generated by the Illumina sequencing platform in particular enables marker and gene discovery, comparative genomics, and gene expression analysis in nonmodel organisms (Wang et al., 2010; Huang et al., 2012; Nicolai et al., 2012; Varshney et al., 2012; Wang et al., 2012; Zhao et al., 2012). We conducted RNA-Seq transcriptome assemblies on *B. sylvaticum* plants originating west of Corvallis, Oregon (hereafter referred to as Corvallis, or Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre). We generated >36 Gb of *B. sylvaticum* transcriptome and assembled the sequences into 120,091, 104,095, and 177,095 transcript contigs with an average length of 1652, 1728, and 1566 bp for samples from Corvallis, Spain, and Greece, respectively. The cDNA sequence files are provided in fasta format for each population in Appendices S1 (Brasy-Cor), S2 (Brasy-Esp), and S3 (Brasy-Gre). We estimate that these transcriptomes represent >90% of the *B. sylvaticum* gene space. Furthermore, we identified SNP and SSR sequences that could be used to design genetic markers for use in future population studies. Along with the substantial genomic resources available in a close congener (*B. distachyon*; Mur et al., 2011) providing reference and comparative material, this transcriptome assembly will be useful on a broad scale as a greatly needed resource for ecologists and geneticists conducting research on native and invasive populations of *B. sylvaticum*, and on the role of climate change and adaptation toward successful invasiveness.

### METHODS AND RESULTS

#### Transcriptome sequencing and de novo assembly

Populations from Spain and Greece were selected to investigate correlations between the Oregon, USA samples and European progenitors used in the USDA field trials. Our samples were drawn from populations in Corvallis, Oregon (population OR-C1; Rosenthal et al., 2008; GPS coordinates: 44°59'25"N, 124°45'41"W), Avila, Spain (population SASA, USDA accession PI 318962; GPS coordinates: 40°39'27"N, 5°18'38"W), and Thessaloniki, Greece (population GRE, USDA accession PI 206546; GPS coordinates: 40°37'48"N, 22°57'36"E). The OR-C1 plants were selected from field-collected seed, and the GRE and SASA seed samples were collected and maintained by the USDA Plant Germplasm division in Pullman, Washington (USA). All plants were grown in a common greenhouse garden at the Portland State University campus in Portland, Oregon, under 12 hours light at 25°C and 12 hours dark at 15°C. At 60 wk, leaf tissue was collected from two individuals per population and ground in liquid nitrogen. Total cellular RNA was extracted using a modified protocol described elsewhere (Fox et al., 2009). In brief, RNA was extracted using RNA Plant Reagent (Life Technologies, Grand Island, New York, USA) and treated with RNase-free Turbo DNase (Life Technologies). Concentration, integrity, and extent of contamination by ribosomal RNA were assessed using an ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and Bioanalyzer 2100 (Agilent Technologies, Santa Clara, California, USA).

Samples were prepared using the TruSeq RNA Sample Preparation Kit (v2) and sequenced on the Illumina HiSeq 2000 instrument (Illumina, San Diego, California, USA). We generated >358 million 101-bp paired-end reads with a fragment size of ~325 bp. This represents an overall total of 36.19 Gb of 101-bp paired-end *B. sylvaticum* transcriptome sequence. The reads from each sample were indexed for use in preliminary expression analyses and represent between 47 and 79 million reads per sample (Table 1). A total of 12.1 Gb paired-end sequences were used to assemble the Corvallis reference transcriptome, and 10.9 and 12.8 Gb were used for the Greece and Spain references, respectively.

The raw Illumina reads were processed for quality and parsed for index sequences and pairs using custom Perl scripts prior to assembly. The metrics used to assess transcriptome assembly quality included the overall number (coverage) of contigs, the average length of contigs, and the diversity of contigs (the estimated number of discrete loci assembled). We compared de novo assemblies using Velvet/Oases and Trinity algorithms, two published software programs built specifically to assemble de novo transcriptomes from short-read sequence data (Grabherr et al., 2011; Schulz et al., 2012). After evaluating the performance of the Velvet/Oases and Trinity algorithms, we conducted our final assembly using Velvet/Oases (Appendix S4). Our assembly generated 120,091, 104,095, and 177,682 contigs for the Corvallis, Spain, and Greece assemblies (Table 1; Appendices S1, S2, and S3). In all three *B. sylvaticum* assemblies, ~96% of assembled contigs were longer than 250 bp, with the longest contig for each of the three assemblies being greater than 15,000 bp. The average contig length was 1652, 1728, and 1566 bp for samples from Corvallis, Spain, and Greece, respectively, while the average length of the reference *B. distachyon* Bd21 gene models (v1.2) is 1086 bp. The median for each assembly, while lower, was near the average, indicating that the assemblies were not composed of an over-representative set of small contigs. Furthermore, the frequency distribution of contig scaffold sizes shows the majority of the lengths near the median and the overall frequency distribution similar to that observed in *B. distachyon* (Fig. 1A). However, when we compared the number of assembled contigs with the average length, we observed many more contigs in our de novo assemblies than the number of gene models in *B. distachyon*, and the average contig lengths in *B. sylvaticum* were greater than that of *B. distachyon* as well (Fig. 1B). The greater number of contigs in the *B. sylvaticum* assemblies could result from several factors, including gene splicing isoforms,

### Table 1. Statistics of data sets from the sequencing and de novo transcriptome assemblies of the *Brachypodium sylvaticum* samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre) and their comparison with the transcriptome of *B. distachyon* Bd21 sequenced genome.

<table>
<thead>
<tr>
<th>Transcriptomes</th>
<th>Raw sequences</th>
<th>Assembled contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of reads</td>
<td>Gb</td>
</tr>
<tr>
<td>Brasy-Cor</td>
<td>120,443,086</td>
<td>12.16</td>
</tr>
<tr>
<td>Brasy-Esp</td>
<td>109,942,266</td>
<td>11.1</td>
</tr>
<tr>
<td>Brasy-Gre</td>
<td>127,927,820</td>
<td>12.92</td>
</tr>
<tr>
<td><em>B. distachyon</em> (Bd21)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

http://www.bioone.org/loi/apps
Fig. 1. Sequence comparisons. (A) Histogram of frequency distribution of contig lengths of *Brachypodium sylvaticum* samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre) de novo transcriptome assemblies compared with *B. distachyon* cDNA lengths (X-axis has been truncated at 6 kb). The frequency distribution of the *B. sylvaticum* transcriptomes closely mirrors that of the *B. distachyon* transcriptome. (B) Comparisons of number of contigs and the average length of *B. sylvaticum* assembled contigs and *B. distachyon* transcripts. Although the overall number of contigs is far greater than the number of gene loci in *B. distachyon*, the average length observed in *B. sylvaticum* transcriptomes is larger than *B. distachyon*.
gene paralogs, and mis-assemblies. As discussed below, the vast majority of assembled contigs are well-annotated, indicating that the inflated number of contigs is not due to mis-assemblies but more likely due to transcript isoforms and/or paralogous loci.

**Homology-based transcriptome annotation**—We examined the number of homologous loci identified through BLAST to approximate the diversity and coverage of the assembled gene loci (Mount, 2007). We annotated genes based on BLAST similarity (E-value threshold cutoff of 1e-10) to sequences available in the GenBank protein database and directly to transcripts identified in the sequenced monocot genomes. We used our transcript scaffolds as queries in BLASTX to search against the National Center for Biotechnology Information (NCBI) nonredundant protein database (Table 2), which resulted in >80% correlation between the contigs of each *B. sylvaticum* assembly and GenBank peptide sequences.

Similar trends were observed when we compared the *B. sylvaticum* nucleotide sequences directly against other monocot nucleotide sequences using BLASTn (E-value threshold cutoff of 1e-10 and percent identity >90%). The *B. distachyon*, *Oryza sativa* (Japonica, MSU6), and *Sorghum bicolor* v1.4 transcriptome databases were obtained from Gramene BioMart (Spoon et al., 2012). In the Corvallis assembly, 103,752 (86.4%) contigs hit *B. distachyon* genes, 96,375 (80.3%) contigs hit rice genes, and 93,751 (78.1%) contigs hit *S. bicolor* genes, with similar results for the Spain and Greece assemblies (Table 2). We further examined the number of homologous loci identified through direct comparisons between *B. sylvaticum* and *B. distachyon* (Fig. 2). We found that the Corvallis assembly uncovered 28,791 (92.8%) of the gene paralogs, and mis-assemblies. As discussed below, the vast majority of assembled contigs are well-annotated, indicating that the inflated number of contigs is not due to mis-assemblies but more likely due to transcript isoforms and/or paralogous loci.

**Table 2.** BLAST comparisons of *Brachypodium sylvaticum* transcriptomes against various databases. Shown are the number and percentage of contigs from *B. sylvaticum* samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre) that hit a gene from the respective database. BLASTx comparisons were made to the GenBank nonredundant peptide database (nr), while BLASTn nucleotide comparisons were made against the *B. distachyon*, *Oryza sativa*, and *Sorghum bicolor* cDNA databases (E-value threshold cutoff of 1e-10).

<table>
<thead>
<tr>
<th>Database compared</th>
<th>Brasy-Cor (120,091)</th>
<th>Brasy-Esp (104,950)</th>
<th>Brasy-Gre (177,682)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of hits</td>
<td>% hits</td>
<td>No. of hits</td>
</tr>
<tr>
<td>GenBank peptide (nr)</td>
<td>96,140</td>
<td>80.1</td>
<td>86,791</td>
</tr>
<tr>
<td><em>B. distachyon</em> v1.2</td>
<td>103,752</td>
<td>86.4</td>
<td>93,291</td>
</tr>
<tr>
<td><em>O. sativa</em> (japonica) vMSU6</td>
<td>96,375</td>
<td>80.3</td>
<td>86,577</td>
</tr>
<tr>
<td><em>S. bicolor</em> v1.4</td>
<td>93,751</td>
<td>78.1</td>
<td>84,023</td>
</tr>
</tbody>
</table>

Fig. 2. Sequence comparisons between contigs from *Brachypodium sylvaticum* samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre) and *B. distachyon* transcripts and genome. (A) *B. distachyon* genes hit by *B. sylvaticum* contigs. Greater than 87% of *B. distachyon* loci were hit by all three *B. sylvaticum* transcriptomes, and more than 96% were hit by a minimum of one contig from at least one *B. sylvaticum* transcriptome. (B) RPKM values averaged over 0.5 megabase intervals across the *B. distachyon* genome. This image shows the uniform coverage of *B. sylvaticum* reads aligned to the five *B. distachyon* chromosomes. Blue histograms indicate RPKM values averaged along the positive strand while green histograms indicate RPKM values of minus strand alignments.

http://www.bioone.org/loi/apps
reading frame (ORF) using the ORFPredictor tool (Min et al., 2005) and were functionally characterized using InterProScan version 4.8 (Quevillon et al., 2005; Hunter et al., 2012). Nearly half of the translated ORFs were assigned InterPro and Gene Ontology (GO) annotations (Appendix S5), which is consistent with other published annotated genomes. The functional annotations were enriched by performing Blast2GO analysis by adopting a stringent BLASTx search (E-value ≤1e-20 and percent identity ≥90%) against the NCBI GenBank nonredundant protein database (Conesa and Gotz, 2008), and the resulting best hits with GO annotations were used to project GO assignments to B. sylvaticum contigs (Gotz et al., 2008; Barrell et al., 2009). Enrichment resulted in assigning GO annotations to 69,628 Corvallis, 61,015 Spain, and 107,700 Greece contig assemblies (Appendix S6).

**Gene expression analysis**—We conducted relative gene expression analyses to assess the utility of these reference transcriptomes for future differential gene expression studies. We mapped B. sylvaticum Illumina reads to B. distachyon loci on all five chromosomes and graphically depict the RPKM values averaged over 31,029 B. distachyon v1.2 transcripts, while the Spain and Greece assemblies hit 91.3% and 94.1% of the B. distachyon transcripts, respectively (Fig. 2A). When we compared the B. distachyon genes hit by all three B. sylvaticum assemblies, we found that ~96.8% (30,054) of the B. distachyon genes were hit by a contig from at least one of the three B. sylvaticum transcriptomes. In addition, 27,187 (87.6%) B. distachyon genes were commonly hit by all three B. sylvaticum transcriptome assemblies.

Further, when we extended our comparisons with other sequenced grass genomes, we found that the Corvallis contigs hit 44,538 (~67%) of 66,338 rice transcripts and 24,999 (~84%) of 29,448 sorghum transcripts with similar results for the other two B. sylvaticum assemblies. Overall, ~90% of all the B. sylvaticum contigs were assigned a homology-based annotation. While the properties of any transcriptome are uniquely associated with the spatial, temporal, and environmental factors present at the precise time of tissue sampling, these results indicate that we have sequenced the great majority of B. sylvaticum gene loci and have assembled three quality reference transcriptomes for B. sylvaticum.

**Functional characterization of the transcriptome**—All the B. sylvaticum contigs were translated into peptides by querying the longest predicted open reading frame (ORF) using the ORFPredictor tool (Min et al., 2005) and were functionally characterized using InterProScan version 4.8 (Quevillon et al., 2005; Hunter et al., 2012). Nearly half of the translated ORFs were assigned InterPro and Gene Ontology (GO) annotations (Appendix S5), which is consistent with other published annotated genomes. The functional annotations were enriched by performing Blast2GO analysis by adopting a stringent BLASTx search (E-value ≤1e-20 and percent identity ≥90%) against the NCBI GenBank nonredundant protein database (Conesa and Gotz, 2008), and the resulting best hits with GO annotations were used to project GO assignments to B. sylvaticum contigs (Gotz et al., 2008; Barrell et al., 2009). Enrichment resulted in assigning GO annotations to 69,628 Corvallis, 61,015 Spain, and 107,700 Greece contig assemblies (Appendix S6).

**Fig. 3.** Comparison of SNPs identified by mapping sequence reads from Brachypodium sylvaticum genotypes from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre) to the B. distachyon Bd21 reference genome. (A) Venn diagram showing the overlap of SNPs within the three B. sylvaticum genotypes. The numbers inside the Venn diagram exclude the 628 common sites exhibiting variation among the B. sylvaticum genotypes. We identified more than 157,835 nucleotide variants common among the three B. sylvaticum genotypes when compared to B. distachyon. Brasy-Gre had the greatest number of SNPs identified. (B) Genotype-specific SNPs were mapped to show the SNP density and chromosome-wide distribution against the B. distachyon genome. (C) Phylogenetic construction of a maximum likelihood tree based on the 628 common sites exhibiting variation among the B. sylvaticum genotypes indicates the proposed evolutionary relationship among the B. sylvaticum genotypes and the B. distachyon Bd21 reference genome.
Applications in Plant Sciences 1(3): 1200011

Fox et al.—Brachypodium sylvaticum transcriptome

doI:10.3732/apps.1200011

0.5 megabase intervals (Fig. 2B). When we compared the RPKM values among the three B. sylvaticum plants, we found very high Pearson’s correlation coefficients between gene expression data sets (Appendix S7). Although the three B. sylvaticum plants do not exhibit significant expression differences over our mapped intervals, we do observe a uniform distribution of contigs mapping to the B. distachyon genome (Fig. 2B; B. distachyon v1.0 genome sequence from http://mips.helmholtz-muenchen.de/plant/brachypodium/). To further demonstrate the utility of pairing the B. sylvaticum transcriptomes and their close congener to investigate questions regarding gene expression, we used the BrachyCyc pathway tool (http://pathway.granome.org/granome/brachycyc.shtml), which contains biochemical pathways consisting of over 7000 B. distachyon genes coding for enzymes. Using the BrachyCyc pathway tool, we mapped RPKM values to BrachyCyc metabolic pathways to examine gene expression profiles of homologous genes (Appendix S7). These results demonstrate the potential of the B. sylvaticum transcriptomes for use in future studies investigating differential gene expression and metabolomics, as well as for making comparisons with B. distachyon.

Genetic variation—To quantify the number of SNP sites across the three transcriptomes, we mapped the reads from each B. sylvaticum sample to the B. distachyon genome v1.0 using Bowtie version 0.12.8 (Langmead et al., 2009). We then used custom Perl scripts to identify nucleotide differences in positions with at least eight aligned reads and 75% of those aligned reads confirming the SNP (Kimbrle, unpublished). Using these criteria, we identified 394,654 putative SNPs among the three B. sylvaticum genomes (Fig. 3A; Appendix S8). Of these, 157,835 SNPs were in sites common to all three genotypes. Although the total number of SNPs was similar among the genotypes, we observed more SNPs unique to the Greece sample (66,963) when compared to Corvallis (39,936) and Spain (40,027). To address the biological relevance of these SNPs and their potential role to be studied in the future for relevance to invasive phenotypes, we predicted the potential effects of the variants and identified a diverse set of consequences (McLaren et al., 2010). Notably, we identified more than 230,000 downstream variants, more than 92,000 missense variants, and 234 stop codons introduced (Tables 3, 4). We observed only slight variation when we mapped the SNP densities of each B. sylvaticum genotype onto B. distachyon chromosomes. Generally, the B. sylvaticum SNP density mirrors the B. distachyon gene density and centromeric region (Fig. 3B). We also constructed a maximum parsimony tree based on concatenated variant sites where at least five reads from each B. sylvaticum transcriptome aligned with Bowtie (Fig. 3C). Of the 157,835 SNPs, only 628 loci were polymorphic within at least one of the three B. sylvaticum samples. These 628 positions were concatenated to generate the maximum likelihood tree depicting relative relationships among the three B. sylvaticum genotypes. While this SNP analysis shows the utility of the B. sylvaticum transcriptomes for genotyping studies, much work needs to be done to fully elucidate the relationships of the various native and invasive populations.

We mined the assembled B. sylvaticum contigs for SSRs using Perl code from the Simple Sequence Repeat Identification Tool (SSRIT; Ternykht et al., 2001; http://www.granome.org/db/markers/ssritool), looking for di-, tri-, tetra-, penta-, and hexanucleotide SSRs with a minimum of nine, six, five, and five repeat units, respectively (Table 5; Appendix S9). In total, we identified 23,535 SSRs in Corvallis, 20,303 in Spain, and 32,847 in Greece transcriptome sequences (Table 5). These SSRs were identified in 18,281 contigs from Corvalgis, 15,975 from Spain, and 25,567 from Greece. To test whether SSRs showing polymorphism in our computational analysis can be used to develop potential genetic markers, we conducted a test PCR amplification of a sample SSR site from the three sequenced genotypes as well as additional B. sylvaticum plants from Oregon, USA, and Europe (Appendix S10).

Brachypodium sylvaticum resources and data download—All sequence, annotations, and data files are available from the project website (http://jaiswalab.cgrb.oregonstate.edu/genomics/brasy). SNPs and contig alignments to B. distachyon are also available from the B. distachyon Bd21 genome browses available from BrachyBase (http://www.brachypodium.org/) and Gramene (http://www.granome.org/Brachypodium_distachyon/) databases. The raw sequence data are available from the NCBI Sequence Read Archive (SRA; accession SRA062655).

CONCLUSIONS

The list of noxious invasive plants identified by the Oregon Department of Agriculture (2009) includes species such as kudzu, goatgrass, knapweed, and others that exact high economic and ecological costs in regions where they have been introduced. Ecologists make distinctions among species that are introduced (able to persist), naturalized (establishes self-sustaining populations, but is not a dominant component of the vegetation), and invasive (is able to dominate habitats to the exclusion of native species). False brone meets the “invasive” criteria but has not yet become notorious because its distribution is still restricted compared to other invasive plants. Our ultimate goal is to use B. sylvaticum as a model for studying adaptation and invasiveness and for the general study of grasses. Therefore, we consider it a necessary first step to establish baseline resources for B. sylvaticum by generating de novo transcriptomes from multiple genotypes, use them to study gene expression and regulation, and identify functional nucleotide polymorphisms to develop new sets of genetic markers for future population-wide

<table>
<thead>
<tr>
<th>Predicted variant effect</th>
<th>No. of SNP sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>3′ UTR variant</td>
<td>46,871</td>
</tr>
<tr>
<td>3′ UTR variant, splice region variant</td>
<td>48</td>
</tr>
<tr>
<td>5′ UTR variant</td>
<td>6636</td>
</tr>
<tr>
<td>5′ UTR variant, splice region variant</td>
<td>34</td>
</tr>
<tr>
<td>Downstream gene variant</td>
<td>231,969</td>
</tr>
<tr>
<td>Initiator codon variant</td>
<td>38</td>
</tr>
<tr>
<td>Intron variant</td>
<td>36,334</td>
</tr>
<tr>
<td>Intron variant, splice region variant</td>
<td>13,336</td>
</tr>
<tr>
<td>Missense variant</td>
<td>92,532</td>
</tr>
<tr>
<td>Missense variant, splice region variant</td>
<td>310</td>
</tr>
<tr>
<td>Splice acceptor variant</td>
<td>3222</td>
</tr>
<tr>
<td>Splice donor variant</td>
<td>5502</td>
</tr>
<tr>
<td>Splice region variant, initiator codon variant</td>
<td>1</td>
</tr>
<tr>
<td>Stop gained</td>
<td>234</td>
</tr>
<tr>
<td>Stop gained, splice region variant</td>
<td>3</td>
</tr>
<tr>
<td>Stop lost</td>
<td>117</td>
</tr>
<tr>
<td>Stop lost, splice region variant</td>
<td>2</td>
</tr>
<tr>
<td>Stop retained variant</td>
<td>180</td>
</tr>
<tr>
<td>Synonymous variant</td>
<td>185,986</td>
</tr>
<tr>
<td>Synonymous variant, splice region variant</td>
<td>958</td>
</tr>
<tr>
<td>Upstream gene variant</td>
<td>174,506</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNP</th>
<th>Substitution type</th>
<th>Brasy-Cor</th>
<th>Brasy-Esp</th>
<th>Brasy-Gre</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → C</td>
<td>Transversion</td>
<td>14,208</td>
<td>13,836</td>
<td>14,695</td>
</tr>
<tr>
<td>A → G</td>
<td>Transversion</td>
<td>43,635</td>
<td>42,985</td>
<td>43,135</td>
</tr>
<tr>
<td>A → T</td>
<td>Transversion</td>
<td>9,695</td>
<td>9,834</td>
<td>10,118</td>
</tr>
<tr>
<td>C → A</td>
<td>Transversion</td>
<td>11,998</td>
<td>12,241</td>
<td>11,459</td>
</tr>
<tr>
<td>C → G</td>
<td>Transversion</td>
<td>18,834</td>
<td>18,072</td>
<td>17,862</td>
</tr>
<tr>
<td>C → T</td>
<td>Transversion</td>
<td>36,201</td>
<td>37,250</td>
<td>34,717</td>
</tr>
<tr>
<td>G → A</td>
<td>Transversion</td>
<td>35,918</td>
<td>36,980</td>
<td>34,383</td>
</tr>
<tr>
<td>G → C</td>
<td>Transversion</td>
<td>18,962</td>
<td>18,103</td>
<td>17,813</td>
</tr>
<tr>
<td>G → T</td>
<td>Transversion</td>
<td>11,892</td>
<td>12,143</td>
<td>11,393</td>
</tr>
<tr>
<td>T → A</td>
<td>Transversion</td>
<td>9,647</td>
<td>9,970</td>
<td>10,039</td>
</tr>
<tr>
<td>T → C</td>
<td>Transversion</td>
<td>44,121</td>
<td>43,302</td>
<td>43,790</td>
</tr>
<tr>
<td>T → G</td>
<td>Transversion</td>
<td>14,077</td>
<td>13,912</td>
<td>14,371</td>
</tr>
</tbody>
</table>

Table 3. SNP consequence predictions. After aligning all of the 394,654 SNPs from the Brachypodium sylvaticum samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre) on the B. distachyon (Bd21) genome v1.0, we predicted the potential effects of SNP variant loci on the Bd21 genes. We identified a diverse set of consequences for the various gene loci.

Table 4. Characterization of the number of transitions and transversions that were predicted from our SNP analysis, showing a much higher prevalence of transitions over transversions as expected. The SNPs were identified in the Brachypodium sylvaticum samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre).
screening. The results of our de novo assemblies produced a relatively large number of long, reconstructed transcripts, as demonstrated by average contig lengths. Overall, we were able to assign homology-based annotations to ~90% of B. sylvi-

ticum contigs, and more than 50% of the translated sequences were functionally annotated by assigning InterPro and Gene Ontology annotations. More than 96% of B. distachyon Bd21 gene loci were associated with B. sylvi-
ticum contigs, thereby demonstrating diversity and broad coverage in our transcriptomic data. When compared to B. distachyon, we discovered ~390,000 SNPs, of which ~158,000 SNPs were common to the three B. sylvi-
ticum samples. Based on the SNP calls, we identified the number of SNPs with consequences to the gene and transcripts including those altering the potential intron splicing sites and translated protein sequences. These resources, when paired with well-established B. distachyon genomic data, will be useful in the future characterization of B. sylvi-
ticum invasiveness.

LITERATURE CITED


FOOTL, T. N., S. GRIFFITHS, S. ALLIUS, and G. MOORE. 2004. Construction and analysis of a BAC library in the grass Brachypodium sylvaticum: Its use as a tool to bridge the gap between rice and wheat in elucidat-


HOLTEN, J. I. 1980. Distribution and ecology of Brachypodium sylvati-


HUANG, L., X. YANG, P. SUN, W. TONG, and S. HU. 2012. The first Illumina-based de novo transcriptome sequencing and analysis of saf-


HULL, A. J. C. 1974. Species for seeding mountain rangelands in south-


LIONAKIS MEYER, D. J., and J. EFFENBERGER. 2010. California noxious weed disseminules identification manual. California Department of Food and Agriculture, Sacramento, California, USA.


MARTIN, P. H., C. D. CASHAM, and P. L. MARKS. 2009. Why forests appear resistant to exotic plant invasions: Intentional introductions, stand dy-

MCLAREN, W., B. Pritchard, D. RIOS, Y. CHEN, F. FLICK, and F. CUNNINGHAM. 2010. Deriving the consequences of genomic vari-


ologist 191(2): 334–347.


QUEVILLON, E., V. SILVTOENEN, S. PILLAI, N. HART, N. MULDER, R. APWIELER, and R. LOPEZ. 2005. InterProScan: Protein domains iden-


ROY, B. A. 2010. Brachypodium sylvaticum. Invasive species compendi-


http://www.bioone.org/loi/apps

Table 5. Summary of SSR sites identified in the transcriptomes of Brachypodium sylvaticum samples from Corvallis (Brasy-Cor), Spain (Brasy-Esp), and Greece (Brasy-Gre). We identified di-, tri-, tetra-

penta-, and hexanucleotide SSRs with a minimum of nine, six, five, and five repeat units, respectively. More than 20,000 SSRs were identified in the B. sylvaticum transcriptomes; trinucleotide repeats were the largest class of SSRs.

<table>
<thead>
<tr>
<th>Population</th>
<th>Total</th>
<th>Dimer</th>
<th>Trimer</th>
<th>Tetramer</th>
<th>Pentamer</th>
<th>Hexamer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brasy-Cor</td>
<td>23,535</td>
<td>6011</td>
<td>16,346</td>
<td>750</td>
<td>183</td>
<td>245</td>
</tr>
<tr>
<td>Brasy-Esp</td>
<td>20,304</td>
<td>5449</td>
<td>14,093</td>
<td>513</td>
<td>141</td>
<td>107</td>
</tr>
<tr>
<td>Brasy-Gre</td>
<td>32,847</td>
<td>8194</td>
<td>23,280</td>
<td>858</td>
<td>276</td>
<td>239</td>
</tr>
</tbody>
</table>


