Article type : Genomic Resources Article

Marx et al.—RNA-seq of vascular plants at Harvard Forest

Genomic Resources Article

Pilot RNA-seq data from 24 species of vascular plants at Harvard Forest

Hannah E. Marx[1,2], Stacy A. Jorgensen[1], Eldridge Wisely[3], Zheng Li[1], Katrina M. Dlugosch[1], and Michael S. Barker[1,4]

Manuscript received 11 April 2020; revision accepted 7 December 2020.

[1] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

[2] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109-1048, USA

[3] Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona 85721, USA

[4] Author for correspondence: msbarker@arizona.edu

Citation: Marx, H. E., S. A. Jorgensen, E. Wisely, Z. Li, K. M. Dlugosch, and M. S. Barker. 2021. Pilot RNA-Seq data from 24 species of vascular plants at Harvard Forest. *Applications in Plant Sciences* 9(2):

PREMISE: Large-scale projects such as the National Ecological Observatory Network (NEON) collect ecological data on entire biomes to track climate change. NEON provides an opportunity to launch community transcriptomic projects that ask integrative questions in ecology and evolution. We conducted a pilot study to investigate the challenges of collecting RNA-seq data from diverse plant communities.

METHODS: We generated >650 Gbp of RNA-seq for 24 vascular plant species representing 12 genera and nine families at the Harvard Forest NEON site. Each species was sampled twice in 2016 (July and August). We assessed transcriptome quality and content with TransRate, BUSCO, and Gene Ontology annotations.

RESULTS: Only modest differences in assembly quality were observed across multiple *k*-mers. On average, transcriptomes contained hits to >70% of loci in the BUSCO database. We found no significant difference in the number of assembled and annotated transcripts between diploid and polyploid transcriptomes.

DISCUSSION: We provide new RNA-seq data sets for 24 species of vascular plants in Harvard Forest. Challenges associated with this type of study included recovery of high-quality RNA from diverse species and access to NEON sites for genomic sampling. Overcoming these challenges offers opportunities for large-scale studies at the intersection of ecology and genomics.

KEY WORDS community transcriptomics; NEON; polyploidy; RNA-seq; transcriptome assembly.

Many questions in ecology and evolutionary biology increasingly require combining data from these fields at large scales. In particular, integrated, large-scale analyses of multispecies ecological and phylogenetic data sets have become critical to understanding plant distributions and responses to climate change (Zanne et al., 2014; Swenson and Jones, 2017; Maitner et al., 2018; Enquist et al., 2019; Gallagher et al., 2020; McFadden et al., 2019; Rice et al., 2019; Baniaga et al., 2020; Román-Palacios and Wiens, 2020). Recognizing this need, the National Science Foundation (NSF) recently launched the National Ecological Observatory Network (NEON) to generate large-scale data in areas including species occurrence, phenology, and climate, for ecological communities across the United States (Collinge, 2018; Knapp and Collins, 2019). Metagenomic and genomic sampling are also being used to identify and estimate changes in abundance and composition of some taxa, especially microbial communities (https://www.neonscience.org/data). Although these data and analyses will be crucial for understanding ecosystem-scale processes, the collection of genomic data from a broader array of species across NEON sites would allow researchers to further integrate ecological and evolutionary processes in the analyses of communities.

Genomic analyses of single species, although important, do not capture the larger patterns occurring within an interacting community of plants. Transcriptome profiling or genome sequencing of multiple species and individuals within a community will open new, integrative avenues of analysis and allow us to address existing questions that require sampling of floras and communities (Bragg et al., 2015; Fitzpatrick and Keller, 2015; Bowsher et al., 2017; Han et al., 2017; Swenson and Jones, 2017; Zambrano et al., 2017; Matthews et al., 2018; Subrahmaniam et al., 2018; Breed et al., 2019). This is especially true for understanding responses to climate change where community-level analyses are needed to capture the interacting dynamics of different species responses (Liu et al., 2018; Komatsu et al., 2019; Snell et al., 2019). The integration of community-level genomic data from non-model species with ecological and trait data will improve our understanding of plant responses to climate change. Collecting genomic data at the community level with repeated sampling that mirrors other trait data collection will permit assessments of the genetic diversity of entire plant communities and how they change over time, estimates of gene flow and hybridization, measurement of in situ gene expression variation across species in response to shared climate events, and a genomic perspective on functional diversity within and between plant communities. Metagenomics analyses of microbiomes have transformed our understanding of and approaches for studying microbial biology (Fierer et al., 2012a, b; Turner et al., 2013; Delgado-Baquerizo et al., 2018; Jansson and Hofmockel, 2020). Similar plant community transcriptomics and genomics studies could open new avenues of research and provide the crucial data to understand plant responses to climate change.

To explore the potential and challenges of plant community transcriptomics, we conducted a pilot RNA-seq study at the Harvard Forest NEON site. Whereas many RNA-seq studies are focused on collections of related species, an approach that simplifies collection and RNA extraction, a major challenge of community-level transcriptomics is that a diverse range of plant species need to be sampled for RNA extraction in the field. In this pilot study, we evaluated RNA-seq results generated following a protocol that we developed (Field Setup 2 of Yang et al., 2017) for collecting material at distant field sites and returning samples by shipping. Harvard

Forest was selected for this pilot study because of access to a field station that simplified the logistics of working with liquid nitrogen. At Harvard Forest, we sampled 24 species of vascular plants from sites adjacent to the NEON plot. Each species was sampled on two different dates one month apart (in July and August 2016), as close to the same time of day as possible. Species were selected from a phylogenetically diverse range of plants that included ferns, trees, and herbaceous annuals. These plants were selected because they represented the diversity of form and habit that is present in the deciduous forest community at Harvard Forest. Another potential challenge for plant transcriptomics is the abundance of polyploid species and cytotypes (Barker et al., 2016a). With potentially twice as many (or more) genes in a polyploid genome, these species could require more sequencing reads than related diploids to obtain reference transcriptomes of similar quality. To explore the impacts of polyploidy on transcriptome surveys, we made an effort to select sets of related polyploid and diploid species. Here, we give an overview of our data collection, present new reference transcriptomes and translated protein collections for each species, and evaluate the quality of these assemblies using multiple approaches.

<h1>METHODS

<h2>Taxon selection and sampling

The *Harvard Forest Flora* (Jenkins et al., 2008) was used to guide our taxonomic selections and find species to represent each category (diploid/polyploid). Putative diploids and neo-polyploid species were identified from chromosome counts obtained from the Chromosome Counts Database (Rice et al., 2015). Congeneric species pairs were selected based on their phylogenetic relatedness. Our sampling included nine polyploid and 11 diploid species (Table 1). We could not determine the ploidal level of four species. The Harvard Forest Flora Database (Jenkins et al., 2008) was used to locate sampling sites.

Field collection for plant RNA-seq followed the approach described in Yang et al. (2017). The only difference was here we sampled tissue from mature leaves of an apparently healthy individual (e.g., lacking herbivore or pathogen damage) rather than young flower or leaf buds to maintain developmental consistency as much as possible over time. Each target species was sampled from the same population on two different dates about one month apart (July and August) during the 2016 growing season (Fig. 1). We attempted to sample as close to the same time of day as possible on both dates by sampling species in the same order on both trips, but this was not always achievable due to challenges of fieldwork, such as weather and time to relocate sample populations. Leaf tissues were flash-frozen in liquid nitrogen in the field and shipped on dry ice to the University of Arizona for RNA extraction. After leaf tissue collection, additional leaf tissue was preserved on silica for DNA backup, and the remaining plant material was pressed for a herbarium specimen (see Appendix 1 for voucher information and collection details).

<h2>RNA extraction and RNA-seq

Total RNA was extracted from leaf tissue collected on each sampling date for all species using the Spectrum Plant Total RNA Kit (Sigma-Aldrich Co., St. Louis, Missouri, USA) following the manufacturer's Protocol A. RNA was used to prepare cDNA using the Ovation RNA-Seq System (catalogue no. 7102-A01; NuGEN, Redwood, California, USA) via single primer isothermal amplification and automated on the Apollo 324 liquid handler (TaKaRa Bio, Kusatsu, Shiga, Japan). cDNA was quantified on the NanoDrop (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and was sheared to approximately 300-bp fragments using the Covaris M220 ultrasonicator (Covaris, Woburn, Massachusetts, USA). Libraries were generated using Kapa Biosystem's library preparation kit for Illumina (KK8201; Roche, Wilmington, Massachusetts, USA). Fragments were end repaired and A-tailed, and individual indexes and adapters (catalogue no. 520999; Bioo Scientific, Austin, Texas, USA) were ligated on each separate sample. The adapter-ligated molecules were cleaned using AMPure beads (A63883; Agencourt Bioscience/Beckman Coulter, Indianapolis, Indiana, USA) and amplified with the KAPA HiFi enzyme (KK2502; Roche). Each library was then analyzed for fragment size on an Agilent TapeStation 4200 (Agilent Technologies, Santa Clara, California, USA) and quantified by qPCR (KAPA Library Quantification Kit, KK4835) on the QuantStudio 5 Real-Time PCR System (Thermo Fisher Scientific). Multiplex pooling (13–16 samples per lane) and paired-end sequencing at 2 × 150 bp were then performed on the Illumina NextSeq500 platform at Arizona State University's CLAS Genomics Core facility. Raw read quality was assessed using FastQC (Andrews, 2010).

<h2>De novo transcriptome assembly, protein translation, and quality assessment

Raw sequence reads were processed using the SnoWhite pipeline (Barker et al., 2010; Dlugosch et al., 2013), which included trimming adapter sequences and bases with a quality score below 20 from the 3′ ends of all reads, removing reads that are entirely primer and/or adapter fragments using TagDust (Lassmann et al., 2009), and removing poly(A/T) tails with SeqClean (https://sourceforge.net/projects/seqclean/). The cleaned reads from each sampling date were merged and cleaned to synchronize read pairs using fastq-pair (Edwards and Edwards, 2019) and pooled to assemble a reference de novo transcriptome for each species.

Due to the significant time involved in running and evaluating multiple assemblies for each species, we chose five species that represent the phylogenetic diversity of our samples (*Dryopteris intermedia* (Muhl. ex Willd.) A. Gray, *Galium mollugo* L., *Juglans cinerea* L., *Plantago major* L., and *Persicaria sagittata* (L.) H. Gross) to identify the optimal *k*-mer to use for assembling all 24 species. For these five exemplar taxa, we examined the quality of assemblies generated by SOAPdenovo-Trans version 1.03 (Xie et al., 2014) across a range of *k*-mers (37, 47, 57, 67, 77, 87, 97, 107, 117, and 127). Assembly quality across the different *k*-mers was assessed by mapping the raw reads to each assembly with TransRate version 1.0.3 (Smith-Unna et al., 2016) and evaluating the optimal assembly scores. TransRate calculates assembly scores by remapping the reads back to the assembly and combining a variety of metrics for each contig, including estimates of whether a base pair was called correctly, whether a base should be a part of the final transcript, the probability that a contig was derived from a single transcript, and the probability that a contig is structurally complete. We selected a *k*-mer that produced the average highest optimal assembly score across the five species. This *k*-mer (57, see Results) was used to assemble reference transcriptomes for the entire collection of species.

We used TransPipe (Barker et al., 2010) to identify plant proteins from the assembled transcripts for each reference transcriptome and provide protein and in-frame nucleic acid sequences for each species. The reading frame and protein translation for each sequence was identified by comparison to protein sequences from 25 sequenced and annotated plant genomes from Phytozome (Goodstein et al., 2012). Using BLASTX (Wheeler et al., 2008), best-hit proteins were paired with each transcript at a minimum cutoff of 30% sequence similarity over at least 150 sites. Transcripts that did not have a best-hit protein at this level were removed. To determine the reading frame and generate estimated amino acid sequences, each transcript was aligned against its best-hit protein by GeneWise 2.2.2 (Birney et al., 2004). Based on the highestscoring GeneWise DNA–protein alignments, stop and "N"-containing codons were removed to produce estimated amino acid sequences for each transcript. Output included translated protein sequences and their corresponding nucleic acid sequences.

To assess the quality of the assembled transcriptomes for the full set of 24 species, we analyzed each with TransRate and BUSCO. Summary statistics, including the number of scaffolds, mean scaffold lengths, and N50, were calculated by TransRate version 1.0.3 for all scaffolds as well as for the subset of sequences that were identified as plant proteins and translated. We evaluated the completeness of our transcriptome coverage with BUSCO version 4.0.5 (Seppey et al., 2019). BUSCO compares sequences to a collection of universal single-copy orthologs for the Viridiplantae (Viridiplantae Odb10) and the eukaryotes (Eukaryote Odb10). We also used the TransRate and BUSCO statistics to compare differences in the assemblies of diploid and polyploid species.

<h2>Gene Ontology annotation and comparison

Gene Ontology (GO) annotations of all transcriptomes were obtained through translated BLAST (BLASTX) searches against the annotated *Arabidopsis thaliana* (L.) Heynh. protein database from TAIR (Lamesch et al., 2012) to find the best hit with a length of at least 100 bp and an *E*-value of at least 1e-10. GO-slim annotations based on the plant GO-slims from TAIR were obtained for the whole transcriptome for each species and presented as a heatmap. The heatmap columns were clustered by hierarchical clustering with default parameters in R with the order of GO categories set arbitrarily by the ranking in *Lysimachia ciliata* L. Rankings of the GO slim categories were determined by the relative frequency of the GO term among the transcripts in each transcriptome.

<h1>RESULTS

We found relatively little variation in the optimal TransRate scores across assemblies with different *k*-mers. The optimal TransRate scores ranged from ~0.1–0.15, with each of the five exemplar species peaking at different *k*-mers (Fig. 2). Scores trended downward for all species at higher *k*-mers, with no sharp peaks in the score apparent in most taxa. The mean *k*-mer of the top-scoring assemblies for each species was 61, and the closest *k*-mer to this value (57) was used to assemble reference transcriptomes for all 24 species.

Assemblies for most of the 24 species appeared to be of relatively high quality. By combining RNA-seq libraries representing two different time points, we obtained an average of 27 Gbp of reads for each reference transcriptome (Table 1). With a *k*-mer = 57, the assemblies contained an average of 483,084 scaffolds with a mean length of 281 bp and N50 of 960 bp. The translated nucleic acids for each assembly had an average of 31,470 sequences with a mean length of 652 bp and N50 of 789 bp. We observed no significant relationship between the number of scaffolds or number of translated proteins and sequencing depth (Fig. 3). The mean complete plus fragmented BUSCO percentages were 73.2% against the Viridiplantae database and 76% against the eukaryote database (Table 2). We found that the number of hits to sequences in the BUSCO databases plateaued at around 20 Gbp of sequencing effort and around 20,000 proteins (Fig. 4).

Polyploid species did not have significantly more translated proteins than diploid species, with 31,152 average proteins translated compared to 30,804 (Fig. 5A; two-tailed *t*-test: $P = 0.95$). Similarly, polyploid species did not have a significantly higher proportion of duplicated BUSCO matches than diploids (Fig. 5B; two-tailed *t*-test: $P = 0.11$). In some cases, the number of proteins or duplicated BUSCO proportion was lower when comparing a polyploid species with its related diploids (e.g., *Dryopteris* Adans.). This may be due to variation in read and/or assembly quality rather than differences in the biology of these species. However, it is not clear that this is due to differences in data quality because in most cases, including *Dryopteris*, all of the species have similarly high read depth (>20 Gbp).

GO annotations of the transcriptomes of the 24 species were largely similar (Fig. 6). Categories such as *other cellular processes*, *other metabolic processes*, and *other intracellular components* were the largest fraction of all transcriptomes, whereas *receptor binding or activity* and *electron transport or energy pathways* were among the smallest. The rank order of each GO-slim category was largely consistent across most species. Species from the same genus were sometimes clustered together by the similarity of their GO-slim representations, such as in *Dryopteris* and *Lysimachia* L., but in most cases the species were not clustered with their congeners. *Polygonum cilinode* Michx. was unique in having many differences in GO category rank compared to the other taxa. It was also the lowest-scoring transcriptome assembly, with only 6088 translated proteins and nearly 80% of BUSCO genes missing (Table 2).

<h1>DISCUSSION

Overall, the RNA sampling approach we developed and employed (Field Setup 2 of Yang et al., 2017) allowed us to sequence and assemble RNA-seq data from a diverse range of species at Harvard Forest. The transcriptomes we assembled for 24 species of vascular plants at Harvard Forest appear to be relatively high quality and consistent with our expectations for de novo plant transcriptome assemblies. Our assemblies were reasonably complete, with more than 70% of BUSCO genes present on average. This is a similar distribution of BUSCO scores to those in the recently published 1KP project (Carpenter et al., 2019; One Thousand Plant Transcriptomes Initiative, 2019) and other studies (Blande et al., 2017; Evkaikina et al., 2017; Pokorn et al., 2017; Weisberg et al., 2017). In our analyses, BUSCO scores and scaffold numbers appear to plateau after approximately 20 Gbp of sequencing effort for diploid and polyploid species, but previous studies indicate that reference assemblies of similar quality can be generated with substantially less sequencing effort for high-quality RNA samples. For example, data from the 1KP project suggest that as little as 2–3 Gbp of read depth is sufficient (Carpenter et al., 2019). Larger amounts of data were collected in this project to facilitate future gene expression analyses. Notably, the few samples that had low BUSCO scores or BUSCO scores that were relatively low for the sequencing effort, such as *Polygonum cilinode*, also had lower numbers of translated proteins but more scaffolds than most species. The relatively poor quality of these outlier assemblies is likely related to lower RNA quality rather than to sequencing effort or ploidal level. In contrast, assemblies with higher BUSCO scores yield translated protein numbers that are more consistent with the number of genes in sequenced plant genomes (Michael, 2014; Wendel et al., 2016). For example, our transcriptomes of *Prunus virginiana* L. and *P. serotina* Ehrh. contained 38,773 and 30,812 translated proteins each. Genomes of related *Prunus* L. species had similar numbers of annotated genes, including 27,852 in *P. persica* (L.) Batsch (International Peach Genome Initiative et al., 2013), 41,294 in *P. yedoensis* Matsum. (Baek et al., 2018), and 43,349 in *P. avium* (L.) L. (Shirasawa et al., 2017). However, these comparisons should be interpreted cautiously because transcriptome assemblies can contain multiple isoforms of protein-coding genes. Like many transcriptome assemblies (Johnson et al., 2012; Carpenter et al., 2019; Patterson et al., 2019), our assemblies also contain a large number of small scaffolds (<300 bp). Small scaffolds are likely artifacts of library amplification and sequencing, considering that most did not translate to a known plant protein sequence.

We found no significant difference in the number of translated transcripts between the diploid and polyploid transcriptome assemblies. Although this could be due to the modest sample size or variation in the age and fractionation level of polyploids, it may also reflect biological differences in expressed transcriptome size and diversity that impact the number of assembled transcripts. Under a simple null model of polyploid transcriptome size, one may expect to observe an approximate doubling of the diploid transcriptome size that may translate to doubling the number of assembled transcripts. However, recent research indicates polyploid transcriptomes may be smaller than expected. Research in *Glycine* Willd. has found that the expressed transcriptome size of polyploid species is less than 2× the diploid size (Coate and Doyle, 2015; Doyle and Coate, 2018; Visger et al., 2019). For example, the transcriptome of the allotetraploid *G. dolichocarpa* Tateishi & H. Ohashi was 1.4× the size of its diploid progenitors (Coate and Doyle, 2010). The apparent lower-than-expected level of the quantity of gene expression in polyploids may be an artifact of comparing diploids and polyploids without accounting for differences in cell numbers or biomass (Coate and Doyle, 2015; Doyle and Coate, 2018; Visger et al., 2019). However, smaller transcriptome sizes in polyploids may also be related to which genes are expressed at a given time or in a particular tissue. This is likely relevant when comparing the assembled gene space for diploid and polyploid transcriptomes, as we do here. Our non-model reference transcriptomes are built from the expressed genes in each sample rather than being based on a reference genome collection. Thus, only genes and alleles that are expressed will be captured in our assemblies and observed in our comparisons. Not all genes or alleles in a polyploid need to be expressed at one time and the overall diversity of the transcriptome at any given time may look more like a diploid, with other alleles being expressed at different times or tissues. Indeed, differential homoeolog silencing is well characterized in polyploid plants (Adams et al., 2003; Coate and Doyle, 2010) and may reduce the sampled transcript diversity of a polyploid genome. If this is the case, we would expect that sampling across more tissues, development times, and environments would lead to greater sampling of the polyploid gene space. Although RNA spike-ins and cell counting may improve differential expression analyses (Visger et al., 2019), capturing the full genome diversity of non-model polyploid species from RNA-seq assemblies remains an additional challenge.

Our pilot study of RNA-seq sampling of diverse species in the field demonstrated some familiar challenges. Building on our past experience with extracting RNA from diverse species

(Barker et al., 2008; Dempewolf et al., 2010; Der et al., 2011; Lai et al., 2012; Dlugosch et al., 2013; Hodgins et al., 2014; Barker et al., 2016b; Mandáková et al., 2017; Qi et al., 2017; Yang et al., 2017; An et al., 2019; Carpenter et al., 2019), we developed an approach for this study to obtain high-quality RNA from field samples (Field Setup 2 of Yang et al., 2017). We found that flash-freezing leaves in liquid nitrogen in situ for later RNA extraction worked well for our diverse samples. A few samples, especially *Polygonum cilinode*, yielded lower-quality RNA, which could potentially be related to leaf age at the time of sampling. Different RNA extraction methods will be needed to deal with the secondary compounds (e.g., polyphenolics) that are present in mature and senescing tissues. Recovering high-quality RNA in the field, across a range of time points and from leaves of different ages, will be a challenge for future studies.

Other challenges that will need to be overcome are associated with sampling at NEON sites. Sampling within NEON permanent plots is generally not allowed for collections outside of NEON's own standard protocol, and therefore our sampling was limited to sites adjacent to NEON plots. This limitation raises some significant issues for researchers who wish to leverage data being collected within NEON sites (https://data.neonscience.org/). First, many NEON sites are located in areas where there is no similar adjacent field site available for sampling, due to land restrictions or ecological variation. We ultimately selected Harvard Forest because we could sample at sites outside of the NEON plot itself. The second major issue is that sampling outside of the NEON plot means that there is no guarantee of continued access to plant populations in the future. There is a great opportunity for ecologists and evolutionary biologists to leverage the wealth of data that NEON is generating for our community. However, access for researchers that wish to conduct RNA and DNA sampling of plants (and other organisms) within NEON sites is an essential issue that requires further development across the network. Sequencing costs will continue to decline over the planned 30-year life span of NEON, and strategies to accommodate sequencing for plants and other eukaryotes will offer opportunities to greatly expand large-scale studies at the intersection of ecology and evolution.

<h1>ACKNOWLEDGMENTS

The authors thank Manisha V. Patel at Harvard Forest for sampling logistics, George M. Ferguson at the University of Arizona Herbarium (ARIZ) for help with identifications and voucher deposits, and Jason Steel at the Biodesign Institute at Arizona State University for RNA-seq preparation and sequencing. This project was supported by the National Science Foundation (NSF #1550838 to M.S.B. and K.M.D and NSF #1750280 to K.M.D.).

<h1>AUTHOR CONTRIBUTIONS

H.E.M., K.M.D., and M.S.B. conceived and designed the experiments. H.E.M. and S.A.J collected the samples, extracted the RNA, and collected the vouchers. H.E.M., E.W., Z.L., K.M.D., and M.S.B. analyzed the data. H.E.M., Z.L., K.M.D., and M.S.B. drafted the manuscript, and all authors approved the final manuscript.

<h1>DATA ACCESSIBILITY

Raw reads for all samples for 24 species are deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRP127805: https://www.ncbi.nlm.nih.gov/sra/SRP127805; BioProject: PRJNA422719). Assembled transcriptomes for each species are archived on Zenodo and available at https://doi.org/10.5281/zenodo.3727312 (Marx et al., 2020).

<h1>LITERATURE CITED

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences, USA* 100: 4649–4654.

An, H., X. Qi, M. L. Gaynor, Y. Hao, S. C. Gebken, M. E. Mabry, A. C. McAlvay, et al. 2019. Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid *Brassica napus*. *Nature Communications* 10: 2878.

Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data [Online]. Website http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [accessed 22 December 2020].

Baek, S., K. Choi, G.-B. Kim, H.-J. Yu, A. Cho, H. Jang, C. Kim, et al. 2018. Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biology* 19: 127.

Baniaga, A. E., H. E. Marx, N. Arrigo, and M. S. Barker. 2020. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecology Letters* 23: 68–

78.

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, R. W. Michelmore, S. J. Knapp, and L. H. Rieseberg. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

Barker, M. S., K. M. Dlugosch, L. Dinh, R. S. Challa, N. C. Kane, M. G. King, and L. H. Rieseberg. 2010. EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. *Evolutionary Bioinformatics Online* 6: 143–149.

Barker, M. S., N. Arrigo, A. E. Baniaga, Z. Li, and D. A. Levin. 2016a. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210: 391–398.

Barker, M. S., Z. Li, T. I. Kidder, and C. R. Reardon. 2016b. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *American Journal of Botany* 103: 1203–1211.

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and Genomewise. *Genome Research* 14: 988–995.

Blande, D., P. Halimaa, A. I. Tervahauta, M. G. M. Aarts, and S. O. Kärenlampi. 2017. *De novo* transcriptome assemblies of four accessions of the metal hyperaccumulator plant *Noccaea caerulescens*. *Scientific Data* 4: 160131.

Bowsher, A. W., P. Shetty, B. L. Anacker, A. Siefert, S. Y. Strauss, and M. L. Friesen. 2017. Transcriptomic responses to conspecific and congeneric competition in co-occurring Trifolium. *Journal of Ecology* 105: 602–615.

Bragg, J. G., M. A. Supple, R. L. Andrew, and J. O. Borevitz. 2015. Genomic variation across landscapes: Insights and applications. *New Phytologist* 207: 953–967.

Breed, M. F., P. A. Harrison, C. Blyth, M. Byrne, V. Gaget, N. J. C. Gellie, S. V. C. Groom, et al. 2019. The potential of genomics for restoring ecosystems and biodiversity. *Nature Reviews Genetics* 20: 615–628.

Carpenter, E. J., N. Matasci, S. Ayyampalayam, S. Wu, J. Sun, J. Yu, F. R. Jimenez Vieira, et al. 2019. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *GigaScience* 8: giz126.

Coate, J. E., and J. J. Doyle. 2010. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: An example from a plant allopolyploid. *Genome Biology and Evolution* 2: 534–546.

Coate, J. E., and J. J. Doyle. 2015. Variation in transcriptome size: Are we getting the message? *Chromosoma* 124: 27–43.

Collinge, S. K. 2018. NEON is your observatory. *Frontiers in Ecology and the Environment* 16: 371.

Delgado-Baquerizo, M., A. M. Oliverio, T. E. Brewer, A. Benavent-González, D. J. Eldridge, R. D. Bardgett, F. T. Maestre, et al. 2018. A global atlas of the dominant bacteria found in soil. *Science* 359: 320–325.

Dempewolf, H., N. C. Kane, K. L. Ostevik, M. Geleta, M. S. Barker, Z. Lai, M. L. Stewart, et al. 2010. Establishing genomic tools and resources for *Guizotia abyssinica* (L.f.) Cass.—The development of a library of expressed sequence tags, microsatellite loci and the sequencing of its chloroplast genome. *Molecular Ecology Resources* 10: 1048–1058.

Der, J. P., M. S. Barker, N. J. Wickett, C. W. dePamphilis, and P. G. Wolf. 2011. *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics* 12: 99.

Dlugosch, K. M., Z. Lai, A. Bonin, J. Hierro, and L. H. Rieseberg. 2013. Allele identification for transcriptome-based population genomics in the invasive plant *Centaurea solstitialis*. *G3:* Genes, Genomes, Genetics 3: 359-367.

Doyle, J. J., and J. E. Coate. 2018. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* 180: 1–52.

Edwards, J. A., and R. A. Edwards. 2019. Fastq-pair: Efficient synchronization of paired-end fastq files. bioRxiv 552885 [Preprint] [published 19 February 2019]. Available at https://doi.org/10.1101/552885 [accessed 22 December 2020].

Enquist, B. J., X. Feng, B. Boyle, B. Maitner, E. A. Newman, P. M. Jørgensen, P. R. Roehrdanz, et al. 2019. The commonness of rarity: Global and future distribution of rarity across land plants. *Science Advances* 5: eaaz0414.

Evkaikina, A. I., L. Berke, M. A. Romanova, E. Proux-Wéra, A. N. Ivanova, C. Rydin, K. Pawlowski, and O. V. Voitsekhovskaja. 2017. The *Huperzia selago* shoot tip transcriptome sheds new light on the evolution of leaves. *Genome Biology and Evolution* 9: 2444–2460.

Fierer, N., C. L. Lauber, K. S. Ramirez, J. Zaneveld, M. A. Bradford, and R. Knight. 2012a. Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME Journal* 6: 1007–1017.

Fierer, N., J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, et al. 2012b. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences, USA* 109: 21390–21395.

Fitzpatrick, M. C., and S. R. Keller. 2015. Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters* 18: 1–16.

Gallagher, R. V., D. S. Falster, B. S. Maitner, R. Salguero-Gómez, V. Vandvik, W. D. Pearse, F. D. Schneider, et al. 2020. Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution* 4: 294-303.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, et al. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* 40: D1178–D1186.

Han, B., M. N. Umaña, X. Mi, X. Liu, L. Chen, Y. Wang, Y. Liang, et al. 2017. The role of transcriptomes linked with responses to light environment on seedling mortality in a subtropical forest, China. *Journal of Ecology* 105: 592–601.

Hodgins, K. A., Z. Lai, L. O. Oliveira, D. W. Still, M. Scascitelli, M. S. Barker, N. C. Kane, et al. 2014. Genomics of Compositae crops: Reference transcriptome assemblies and evidence of hybridization with wild relatives. *Molecular Ecology Resources* 14: 166–177.

International Peach Genome Initiative, I. Verde, A. G. Abbott, S. Scalabrin, S. Jung, S. Shu, F. Marroni, et al. 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics* 45: 487–494.

Jansson, J. K., and K. S. Hofmockel. 2020. Soil microbiomes and climate change. *Nature Reviews Microbiology* 18: 35–46.

Jenkins, J., G. Motzkin, and K. Ward. 2008. The Harvard Forest Flora: An inventory, analysis, and ecological history. Harvard Forest Paper 28. Harvard Forest, Harvard University, Petersham, Massachusetts, USA.

Johnson, M. T. J., E. J. Carpenter, Z. Tian, R. Bruskiewich, J. N. Burris, C. T. Carrigan, M. W. Chase, et al. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE* 7: e50226.

Knapp, A. K., and S. L. Collins. 2019. Reimagining NEON operations: We can do better. *Bioscience* 69: 956–959.

Komatsu, K. J., M. L. Avolio, N. P. Lemoine, F. Isbell, E. Grman, G. R. Houseman, S. E. Koerner, et al. 2019. Global change effects on plant communities are magnified by time and the number of global change factors imposed. *Proceedings of the National Academy of Sciences, USA* 116: 17867–17873.

Lai, Z., N. C. Kane, A. Kozik, K. A. Hodgins, K. M. Dlugosch, M. S. Barker, M. Matvienko, et al. 2012. Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany* 99: 209–218.

Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, et al. 2012. The *Arabidopsis* Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research* 40: D1202–D1210.

Lassmann, T., Y. Hayashizaki, and C. O. Daub. 2009. TagDust—A program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25: 2839-2840.

Liu, H., Z. Mi, L. Lin, Y. Wang, Z. Zhang, F. Zhang, H. Wang, et al. 2018. Shifting plant species composition in response to climate change stabilizes grassland primary production. *Proceedings of the National Academy of Sciences, USA* 115: 4051–4056.

Maitner, B. S., B. Boyle, N. Casler, R. Condit, J. Donoghue II, S. M. Durán, D. Guaderrama, et al. 2018. The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods in Ecology and Evolution* 9: 373–379.

Mandáková, T., Z. Li, M. S. Barker, and M. A. Lysak. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *Plant Journal* 91: 3–21.

Marx, H. E., S. A. Jorgensen, E. Wisely, Z. Li, K. M. Dlugosch, and M. S. Barker. 2020. Progress toward plant community transcriptomics: Pilot RNA-seq data from 24 species of vascular plants at Harvard Forest [Data set] [published 25 March 2020]. Available at Zenodo repository https://zenodo.org/record/3727313#.X-Jg5OBMHX8 [accessed 22 December 2020].

Matthews, B., R. J. Best, P. G. D. Feulner, A. Narwani, and R. Limberger. 2018. Evolution as an ecosystem process: Insights from genomics. *Genome* 61: 298–309.

McFadden, I. R., B. Sandel, C. Tsirogiannis, N. Morueta-Holme, J.-C. Svenning, B. J. Enquist, and N. J. B. Kraft. 2019. Temperature shapes opposing latitudinal gradients of plant taxonomic and phylogenetic β diversity. *Ecology Letters* 22: 1126–1135.

Michael, T. P. 2014. Plant genome size variation: Bloating and purging DNA. *Briefings in Functional Genomics* 13: 308–317.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

Patterson, J., E. J. Carpenter, Z. Zhu, D. An, X. Liang, C. Geng, R. Drmanac, and G. K.-S. Wong. 2019. Impact of sequencing depth and technology on *de novo* RNA-Seq assembly. *BMC Genomics* 20: 604.

Pokorn, T., S. Radišek, B. Javornik, N. Štajner, and J. Jakše. 2017. Development of hop transcriptome to support research into host-viroid interactions. *PLoS ONE* 12: e0184528.

Qi, X., H. An, A. P. Ragsdale, T. E. Hall, R. N. Gutenkunst, J. C. Pires, and M. S. Barker. 2017. Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Molecular Ecology* 26: 3373–3388.

Rice, A., L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, A. Salman-Minkov, J. Mayzel, et al. 2015. The Chromosome Counts Database (CCDB): A community resource of plant chromosome numbers. *New Phytologist* 206: 19–26.

Rice, A., P. Šmarda, M. Novosolov, M. Drori, L. Glick, N. Sabath, S. Meiri, et al. 2019. The global biogeography of polyploid plants. *Nature Ecology & Evolution* 3: 265–273.

Román-Palacios, C., and J. J. Wiens. 2020. Recent responses to climate change reveal the drivers of species extinction and survival. *Proceedings of the National Academy of Sciences, USA* 117: 4211–4217.

Seppey, M., M. Manni, and E. M. Zdobnov. 2019. BUSCO: Assessing genome assembly and annotation completeness. *In* M. Kollmar [ed.], Gene prediction: Methods and protocols, 227–245. Springer, New York, New York, USA.

Shirasawa, K., K. Isuzugawa, M. Ikenaga, Y. Saito, T. Yamamoto, H. Hirakawa, and S. Isobe. 2017. The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. *DNA Research* 24: 499–508.

Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. 2016. TransRate: Reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Research* 26: 1134–1144.

Snell, R. S., N. G. Beckman, E. Fricke, B. A. Loiselle, C. S. Carvalho, L. R. Jones, N. I. Lichti, et al. 2019. Consequences of intraspecific variation in seed dispersal for plant demography, communities, evolution and global change. *AoB Plants* 11: lz016.

Subrahmaniam, H. J., C. Libourel, and E. P. Journet. 2018. The genetics underlying natural variation of plant–plant interactions, a beloved but forgotten member of the family of biotic interactions. *Plant Journal* 93: 747–770.

Swenson, N. G., and F. A. Jones. 2017. Community transcriptomics, genomics and the problem of species co-occurrence. *Journal of Ecology* 105: 563–568.

Turner, T. R., E. K. James, and P. S. Poole. 2013. The plant microbiome. *Genome Biology* 14: 209.

Visger, C. J., G. K.-S. Wong, Y. Zhang, P. S. Soltis, and D. E. Soltis. 2019. Divergent gene expression levels between diploid and autotetraploid *Tolmiea* relative to the total transcriptome, the cell, and biomass. *American Journal of Botany* 106: 280–291.

Weisberg, A. J., G. Kim, J. H. Westwood, and J. G. Jelesko. 2017. Sequencing and *de novo* assembly of the *Toxicodendron radicans* (poison ivy) transcriptome. *Genes* 8: 317.

Wendel, J. F., S. A. Jackson, B. C. Meyers, and R. A. Wing. 2016. Evolution of plant genome architecture. *Genome Biology* 17: 37.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 36: D13–21.

Xie, Y., G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, et al. 2014. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30: 1660-1666.

Yang, Y., M. J. Moore, S. F. Brockington, A. Timoneda, T. Feng, H. E. Marx, J. F. Walker, and S. A. Smith. 2017. An efficient field and laboratory workflow for plant phylotranscriptomic projects. *Applications in Plant Sciences* 5: 1600128.

Zambrano, J., Y. Iida, R. Howe, L. Lin, M. N. Umana, A. Wolf, S. J. Worthy, and N. G. Swenson. 2017. Neighbourhood defence gene similarity effects on tree performance: A community transcriptomic approach. *Journal of Ecology* 105: 616–626.

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.

FIGURE TITLES

FIGURE 1. Workflow of our RNA-seq collection and assembly.

FIGURE 2. TransRate optimal scores for assemblies of five exemplar species from Harvard Forest. A reference transcriptome for each species was assembled with different *k*-mers starting at $k = 37$ and increasing in increments of 10 to $k = 127$.

FIGURE 3. Comparison of the number of (A) scaffolds and (B) translated proteins produced by each assembly with the total number of giga base pairs (Gbp) sequenced for each species.

FIGURE 4. The percentage of BUSCO complete (C) plus fragmented (F) matches compared to the (A) total giga base pairs (Gbp) sequenced and (B) number of translated proteins in each assembly. Green diamonds represent BUSCO matches to the Viridiplantae database, whereas purple circles represent matches to the eukaryote database.

FIGURE 5. Comparison of (A) the number of translated proteins and (B) BUSCO duplicated/single copy (D/S) ratio for assemblies of diploid and polyploid species. In neither case were diploids significantly different from polyploids.

FIGURE 6. Heat map of Gene Ontology (GO) slim categories present in the entire transcriptome of each species. Each column represents the annotated GO categories from each analyzed transcriptome, whereas the rows represent a particular GO category. The colors of the heat map represent the percentage of the transcriptome represented by a particular GO category, with red being highest and purple lowest. The overall ranking of GO category rows was determined by the ranking of GO annotations in the transcriptome of *Lysimachia ciliata*. Hierarchical clustering was used to organize the heatmap columns.

TABLE 1. Summary statistics for RNA-seq data sets, assemblies with a *k*-mer = 57, and translations.

Species	Ploidy[a]	Chromosome no.	July SRA	Aug SRA	Total Gbp (July + Aug)	No. of scaffolds	Mean scaffold length (bp)	Scaffold N50 (bp)	No. of translated proteins
Dryopteris carthusiana	P	82	SAMN08277176	SAMN08277204	26	643,129	264.5	710	24,851
Dryopteris intermedia	D	41	SAMN08277187	SAMN08277216	22	529,510	267.1	822	22,595
Dryopteris marginalis	D	41	SAMN08277188	SAMN08277217	21	550,548	260.1	917	34,121
Galium mollugo	D	11	SAMN08277173	SAMN08277201	40	608,764	179.2	1028	25,040
Galium tinctorium	D	12	SAMN08277181	SAMN08277210	32	78,487	400.6	1091	16,610
Galium triflorum	P	33	SAMN08277189	SAMN08277219	37	574,562	246	899	38,650
Hypericum perforatum	P	16	SAMN08277171	SAMN08277199	25	335,837	233.4	867	34,670
Juglans cinerea	D	16	SAMN08277174	SAMN08277202	18.2	569,859	359.5	1151	66,595
Lonicera tatarica var. *morrowii*	NA	9	SAMN08277167	SAMN08277195	10.4	386,927	324.9	1216	28,147
Lysimachia ciliata	P	48	SAMN08277190	SAMN08277220	24	1,422,451	207	828	32,005
Lysimachia nummularia	D	17	SAMN08277194	SAMN08277223	25.3	428,232	320.9	1102	46,343
Lysimachia quadrifolia	P	42	SAMN08277183	SAMN08277212	30	340,491	290.4	944	37,737
Persicaria arifolia	NA	NA	SAMN08277180	SAMN08277209	30	528,292	245	787	28,741

Persicaria hydropiperoides	NA	NA	SAMN08277172	SAMN08277200	21.3	347,558	344	913	46,058
Persicaria sagittata	NA	20	SAMN08277179	SAMN08277208	26	398,304	319.6	1148	33,118
Plantago lanceolata	D	6	SAMN08277178	SAMN08277206	31	213,834	293.1	1073	20,470
Plantago major	D	6	SAMN08277169	SAMN08277197	23	217,041	308.6	1032	24,715
Plantago rugelii	P	12	SAMN08277170	SAMN08277198	30	378,418	276.1	1161	30,673
Polygonum cilinode	D	11	SAMN08277184	SAMN08277213	17.3	1,065,186	186.8	415	6088
Potentilla argentea	P	21	SAMN08277177	SAMN08277207	29	245,734	425.6	1268	16,306
Potentilla canadensis	D	14	SAMN08277192	SAMN08277222	16.6	433,249	216.5	667	37,503
Prunus serotina	P	16	SAMN08277186	SAMN08277215	31	350,572	267.5	1017	30,812
Prunus virginiana	D	8	SAMN08277185	SAMN08277214	38	536,216	235	1110	38,773
Reynoutria japonica	P	33	SAMN08277193	SAMN08277224	44	410,810	279.6	870	34,662

Note: NA = not available; SRA = Sequence Read Archive.

[a]P and D are polyploid and diploid species, respectively.

TABLE 2. BUSCO results for comparisons to the Viridiplantae and eukaryote databases.

	BUSCO Viridiplantae database							BUSCO Eukaryote dat			
Species	**C**	**(S)**	**(D)**	**F**	**M**	**D/S**	**C+F**	**C**	**(S)**	**(D)**	**F**
Dryopteris carthusiana	19.3	16.9	2.4	40.7	40	0.142	60	24.3	20	4.3	43.5
Dryopteris intermedia	39.8	35.1	4.7	38.4	21.8	0.134	78.2	48.2	44.3	3.9	32.9
Dryopteris marginalis	41.5	34.4	7.1	40	18.5	0.206	81.5	48.6	43.5	5.1	38.4
Galium mollugo	27.8	22.6	5.2	50.1	22.1	0.230	77.9	38	32.5	5.5	41.6
Galium tinctorium	40	37.9	2.1	40	20	0.055	80	50.2	46.3	3.9	27.1
Galium triflorum	30.8	21.9	8.9	46.4	22.8	0.406	77.2	33.4	21.6	11.8	50.2
Hypericum perforatum	29.5	22.4	7.1	48.7	21.8	0.317	78.2	38.8	29	9.8	40
Juglans cinerea	33.9	27.8	6.1	48.2	17.9	0.219	82.1	47.8	39.2	8.6	34.1
Lonicera tatarica var. *morrowii*	34.8	29.4	5.4	46.4	18.8	0.184	81.2	47.1	36.9	10.2	34.5
Lysimachia ciliata	34.1	28.7	5.4	46.1	19.8	0.188	80.2	49	40	9	32.2
Lysimachia nummularia	42.4	35.1	7.3	42.8	14.8	0.208	85.2	48.6	34.9	13.7	37.6
Lysimachia quadrifolia	31.8	26.4	5.4	44.2	24	0.205	76	38	32.5	5.5	39.6
Persicaria arifolia	13.9	10.6	3.3	51.5	34.6	0.311	65.4	24.7	16.9	7.8	46.3
Persicaria hydropiperoides	25.4	16.5	8.9	48.9	25.7	0.539	74.3	32.9	18.4	14.5	44.3
Persicaria sagittata	27.8	16.5	11.3	48.7	23.5	0.685	76.5	39.2	20.4	18.8	40.8
Plantago	40.5	36	4.5	40.5	19	0.125	81	45.9	39.2	6.7	36.1

lanceolata											
Plantago major	51.3	48	3.3	33.4	15.3	0.069	84.7	54.1	49.8	4.3	29.4
Plantago rugelii	34.5	20.9	13.6	46.6	18.9	0.651	81.1	45.8	27.8	18	39.6
Polygonum cilinode	11.5	9.4	2.1	8.5	80	0.223	20	5.5	4.7	0.8	19.6
Potentilla argentea	11.7	10.8	0.9	33.2	55.1	0.083	44.9	14.1	12.9	1.2	38.4
Potentilla canadensis	12.9	9.6	3.3	47.8	39.3	0.344	60.7	17.3	12.2	5.1	52.9
Prunus serotina	23.5	18.1	5.4	50.4	26.1	0.298	73.9	28.6	20	8.6	47.1
Prunus virginiana	27	18.1	8.9	54.4	18.6	0.492	81.4	37.7	25.5	12.2	44.7
Reynoutria japonica	30.1	22.8	7.3	45.2	24.7	0.320	75.3	30.2	23.1	7.1	45.5

Note: C = percentage of all complete BUSCO matches in the respective database; C+F = percentage of complete and fragmented BUSCO matches in the respective database; D = percentage of complete and duplicated BUSCO matches in the respective database; D/S = ratio of duplicated to single-copy complete sequences BUSCO matches in the respective database; F = percentage of fragmented BUSCO matches in the respective database; S = percentage of complete and single copy BUSCO matches in the respective database; M = percentage of missing BUSCO matches in the respective database.

Appendices

APPENDIX 1. Sample collection number, sampling date, time of day, and locality information, with catalog numbers for vouchers deposited at the University of Arizona Herbarium (ARIZ). RIN (RNA integrity number) scores for each sample are also included.

Family	Species	Collection no.	Sampling date	Sampling time	Sampling locality information	Latitude
Dryopteridaceae	*Dryopteris carthusiana*	2016-015	16-Jul-16	19:51	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Rd. to Lyford House, ~50 m east of State Route 32. Along bank on south side (south facing), in *Acer saccharum* dominated forest. Scattered in leaf litter.	42.5279
Dryopteridaceae	*Dryopteris carthusiana*	2016-050	16-Aug-16	12:19	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. South-facing slope on the roadway to Lyford House.	42.5278
Dryopteridaceae	*Dryopteris intermedia*	2016-028	17-Jul-16	17:30	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. Along path north of gate #29 (north of Dugway Rd.).	42.4674
Dryopteridaceae	*Dryopteris intermedia*	2016-066	17-Aug-16	13:01	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. Along path north of gate #29 (north of Dugway Rd.).	42.4676

Dryopteridaceae	*Dryopteris marginalis*	2016-029	17-Jul-16	17:30	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. Along path north of gate #29 (north of Dugway Rd.).	42.4674
Dryopteridaceae	*Dryopteris marginalis*	2016-067	17-Aug-16	13:06	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. Along path north of gate #29 (north of Dugway Rd.).	42.4676
Rubiaceae	*Galium mollugo*	2016-008	16-Jul-16	10:23	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Shed SE of Torrey Lab, along forest edge.	42.531
Rubiaceae	*Galium mollugo*	2016-044	15-Aug-16	17:31	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Around wood shed and generator, along forest margin east of Torrey Hall. South of dumpster and parking.	42.5318
Rubiaceae	*Galium tinctorium*	2016-022	17-Jul-16	9:40	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. Along trail SE of gate #26, to Connor Pond marsh margin. Followed margin back north ~10 m to opening in forest (mixed *Thuja*, *Acer*, *Pinus*). Collected at forest/marsh margin.	42.4663

Rubiaceae	*Galium tinctorium*	2016-059	16-Aug-16	17:37	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. Along trail SE of gate #26, to Connor Pond marsh margin. Followed margin back north ~10 m to opening in forest (mixed *Thuja*, *Acer*, *Pinus*). Collected at forest/marsh margin.	42.4663
Rubiaceae	*Galium triflorum*	2016-030	17-Jul-16	17:57	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. At gate #26.	42.4661
Rubiaceae	*Galium triflorum*	2016-069	17-Aug-16	13:42	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. At gate #26.	42.4660
Hypericaceae	*Hypericum perforatum*	2016-005	16-Jul-16	9:47	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. West of gas shed that is SE of Torrey Lab. Sandy parking lot near forest margin.	42.5320
Hypericaceae	*Hypericum perforatum*	2016-041	15-Aug-16	17:20	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. West of gas shed that is SE of Torrey Lab. Sandy parking lot near forest margin.	42.5319
Juglandaceae	*Juglans cinerea*	2016-009	16-Jul-16	11:36	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Along south side of trail (Locust Opening Rd.) that heads east from Torrey Lab.	42.5324

					~100 m past gate, next to phone pole (#135-5).	
Juglandaceae	*Juglans cinerea*	2016-045	16-Aug-16	10:09	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Along south side of trail (Locust Opening Rd.) that heads east from Torrey Lab. ~100 m past gate, next to phone pole (#135-5).	42.5326
Caprifoliaceae	*Lonicera tatarica* var. *morrowii*	2016-001	16-Jul-16	9:34	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. John H. Torrey Laboratory. North parking lot, along fence margin.	42.5417
Caprifoliaceae	*Lonicera tatarica* var. *morrowii*	2016-037	15-Aug-16	15:12	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. John H. Torrey Laboratory. North parking lot, along fence margin.	42.5324
Primulaceae	*Lysimachia ciliata*	2016-031	17-Jul-16	18:01	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. At gate #26.	42.4661
Primulaceae	*Lysimachia ciliata*	2016-070	17-Aug-16	13:47	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Simes Tract, Compartment Simes. At gate #26.	42.4660

Primulaceae	*Lysimachia nummularia*	2016-035	18-Jul-16	10:00	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P8. Harvard Farm. East of State Route 32 on Pierce Rd., just west of gate.	42.5253
Primulaceae	*Lysimachia nummularia*	2016-074	17-Aug-16	15:15	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P8. Harvard Farm. East of State Route 32 on Pierce Rd., just west of gate.	42.5234
Primulaceae	*Lysimachia quadrifolia*	2016-024	17-Jul-16	10:50	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. East of State Route 122, along trail leading south to gate #26. Forest margin / roadside clearing along fence. Weedy.	42.4671
Primulaceae	*Lysimachia quadrifolia*	2016-062	17-Aug-16	10:20	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. East of State Route 122, along trail leading south to gate #26. Forest margin / roadside clearing along fence. Weedy.	42.2671
Polygonaceae	*Persicaria arifolia*	2016-021	17-Jul-16	9:33	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. Along trail SE of gate #26, to Connor Pond marsh margin. Followed margin back north ~10 m to opening in forest (mixed *Thuja*, *Acer*, *Pinus*). Collected at forest/marsh margin.	42.4663

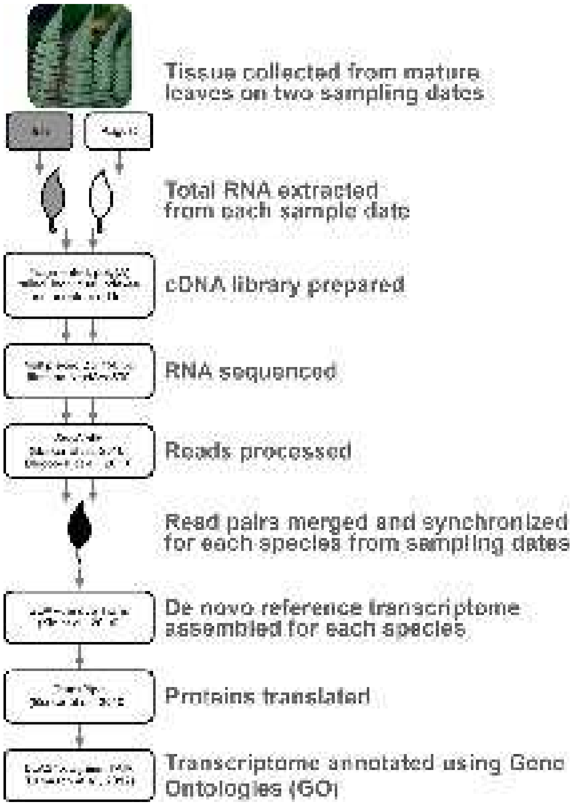
Polygonaceae	*Persicaria arifolia*	2016-058	16-Aug-16	17:30	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. Along trail SE of gate #26, to Connor Pond marsh margin. Followed margin back north ~10 m to opening in forest (mixed *Thuja*, *Acer*, *Pinus*). Collected at forest/marsh margin.	42.4663
Polygonaceae	*Persicaria hydropiperoides*	2016-006	16-Jul-16	9:50	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. West of dumpster SE of Torrey Lab, in sandy lot at forest edge, weedy patch.	42.5319
Polygonaceae	*Persicaria hydropiperoides*	2016-042	15-Aug-16	17:23	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Around dumpster east of Torrey Hall and south of shed.	42.5319
Polygonaceae	*Persicaria sagittata*	2016-020	17-Jul-16	9:31	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. Along trail SE of gate #26, to Connor Pond marsh margin. Followed margin back north ~10 m to opening in forest (mixed *Thuja*, *Acer*, *Pinus*).	42.4663
Polygonaceae	*Persicaria sagittata*	2016-057	16-Aug-16	17:24	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. Along trail SE of gate #26, to Connor Pond marsh margin. Followed margin back north ~10 m to opening in forest.	42.4663

Plantaginaceae	*Plantago lanceolata*	2016-018	16-Jul-16	20:06	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Lawn west of Lyford House.	42.528
Plantaginaceae	*Plantago lanceolata*	2016-053	16-Aug-16	13:43	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Lawn west of Lyford House (before fence).	42.5280
Plantaginaceae	*Plantago major*	2016-003	16-Jul-16	9:27	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. East of Torrey Lab, in sandy parking lot along forest margin.	42.5321
Plantaginaceae	*Plantago major*	2016-039	15-Aug-16	16:09	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. East of Torrey Lab, in sandy parking lot along forest margin.	42.5321
Plantaginaceae	*Plantago rugelii*	2016-004	16-Jul-16	9:30	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. East of Torrey Lab, in sandy parking lot along forest margin.	42.5321
Plantaginaceae	*Plantago rugelii*	2016-040	15-Aug-16	16:11	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. East of Torrey Lab, in sandy parking lot along forest margin.	42.5321

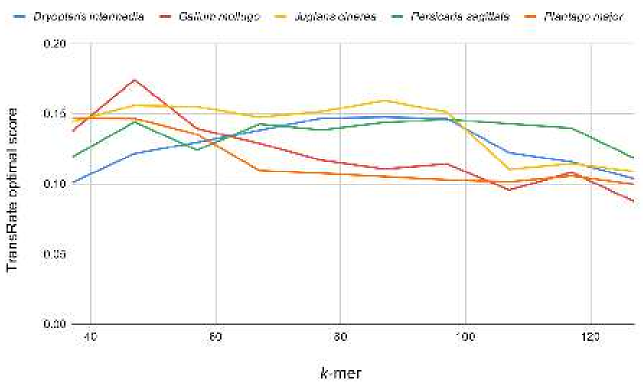
Polygonaceae	*Polygonum cilinode*	2016-025	17-Jul-16	11:13	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. West side of State Route 122, just north of gate #27 along guard rail. Disturbed soils.	42.4553
Polygonaceae	*Polygonum cilinode*	2016-063	17-Aug-16	10:47	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Slab City Tract, Compartment S6. West side of State Route 122, just north of gate #27 along guard rail. Disturbed soils.	42.4556
Rosaceae	*Potentilla argentea*	2016-017	16-Jul-16	19:53	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. Bordering garage south of Lyford House.	42.5281
Rosaceae	*Potentilla argentea*	2016-055	16-Aug-16	1:51	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P1. In sandy soil of parking area south of Lyford House.	42.5282
Rosaceae	*Potentilla canadensis*	2016-033	17-Jul-16	16:55	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P10. French Rd., east of gate #1. On north side of trail at forest edge, just east of old foundation. With *Acer*, *Fraxinus*, relatively open clearing.	42.5390

Rosaceae	*Potentilla canadensis*	2016-072	17-Aug-16	14:36	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P10. French Road, east of gate #1. On north side of trail at forest edge, just east of old foundation. With *Acer*, *Fraxinus*, relatively open clearing.	42.5390
Rosaceae	*Prunus serotina*	2016-027	17-Jul-16	16:38	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Tom Swam Tract, Compartment T1. Sunset Lane, west of State Route 32. About 50 m west of Riley Gate, along road-turned-path on left (south) side.	42.4930
Rosaceae	*Prunus serotina*	2016-065	17-Aug-16	12:16	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Tom Swam Tract, Compartment T1. Sunset Lane, west of State Route 32. About 50 m west of Riley Gate, along road-turned-path on left (south) side.	42.4927
Rosaceae	*Prunus virginiana*	2016-026	17-Jul-16	15:33	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Tom Swam Tract, Compartment T4. Road heading south from Tom Swamp Road at gate #18 (before gate #20). West side of road.	42.5094
Rosaceae	*Prunus virginiana*	2016-064	17-Aug-16	11:26	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Tom Swam Tract, Compartment T4. Road heading south from Tom Swamp Road at gate #18 (before gate #20). West side of road.	42.509

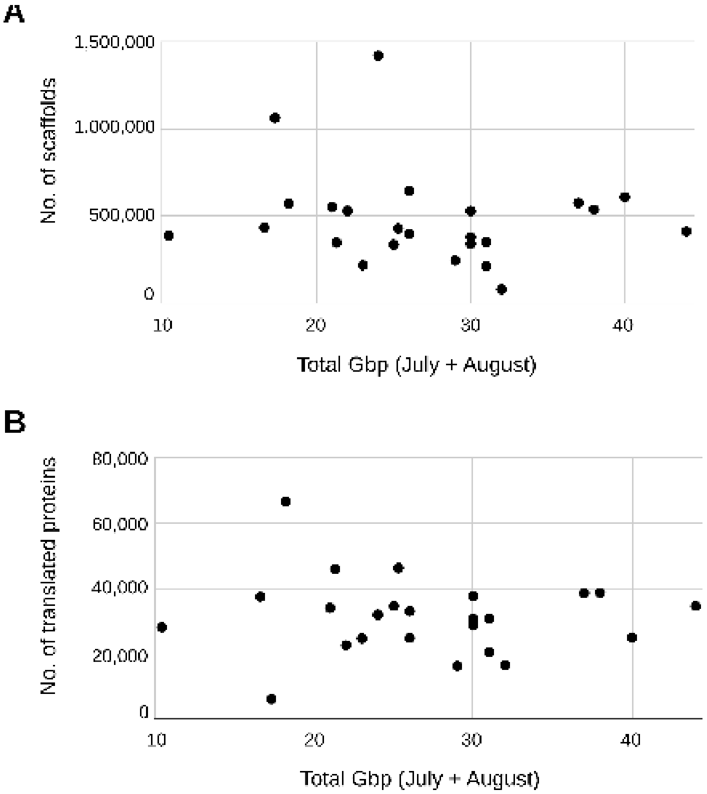
Polygonaceae	*Reynoutria japonica*	2016-034	17-Jul-16	19:40	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P10. East side of State Route 32, north of Prospect Hill Rd. Along freeway.	42.5330
Polygonaceae	*Reynoutria japonica*	2016-075	17-Aug-16	16:03	Harvard Forest, 26 Prospect Hill Rd., Petersham, MA, 0136, USA. Worcester County. Prospect Hill Tract, Compartment P10. East side of State Route 32, north of Prospect Hill Rd. Along freeway.	42.5329


Tissue collected from mature leaves on two sampling dates
Total RNA extracted from each sample date
cDNA library prepared
RNA sequenced
Reads processed
Read pairs merged and synchronized for each species from sampling dates
De novo reference transcriptome assembled for each species
Proteins translated
Transcriptome annotated using Gene Ontologies (GO)

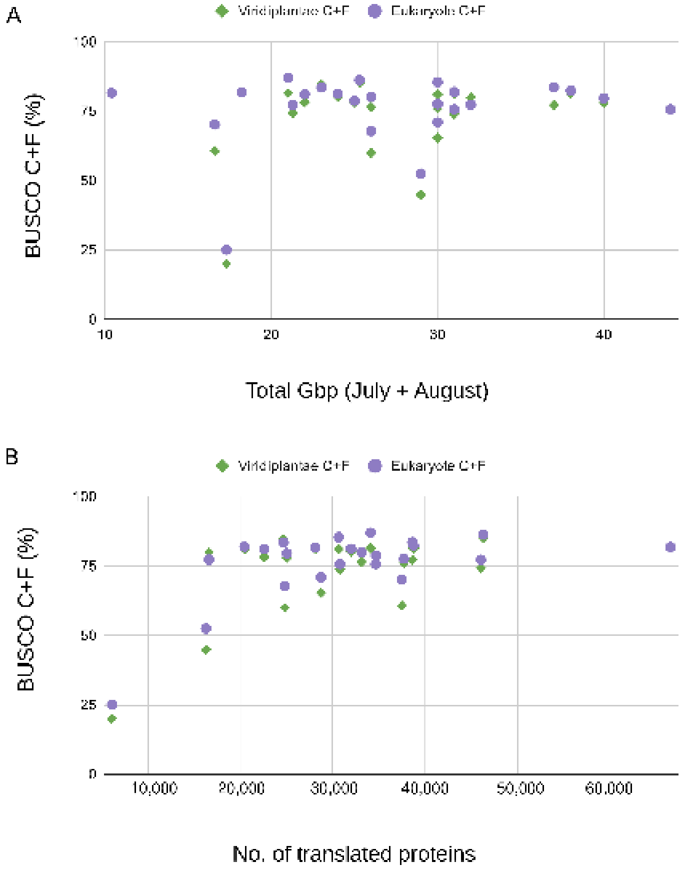
aps3_11409_f1.png



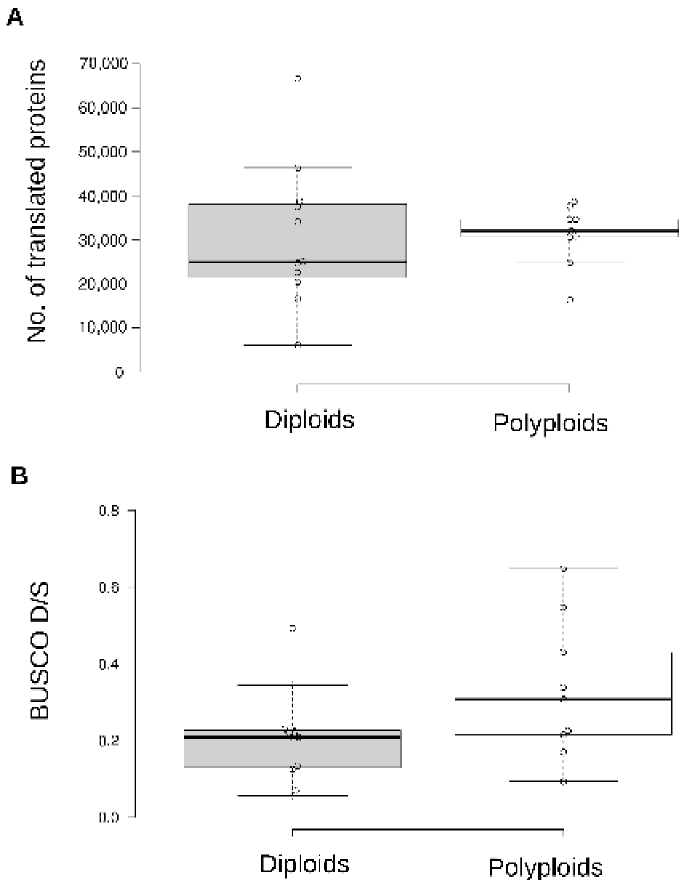
aps3_11409_f2.png



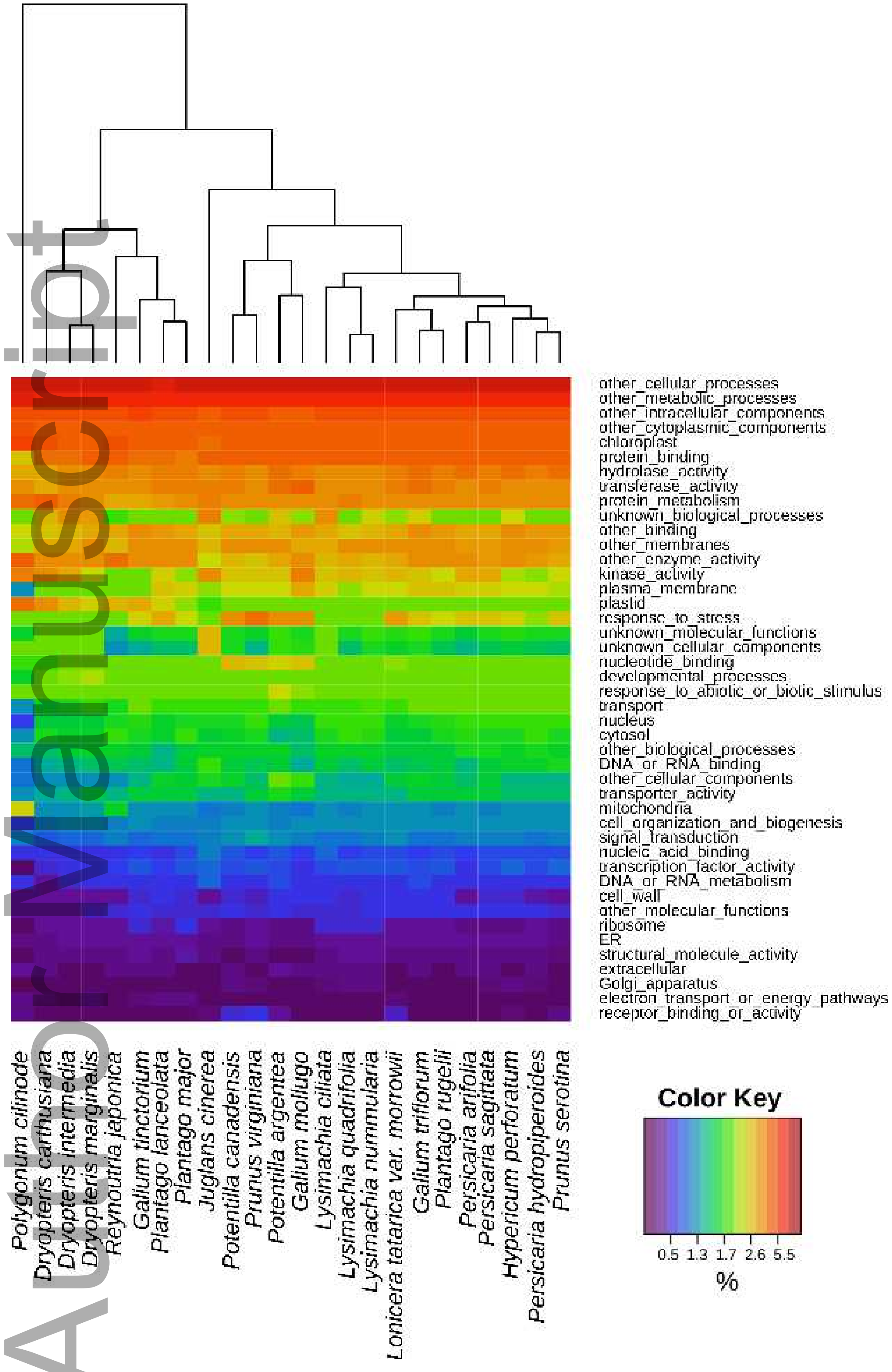
aps3_11409_f3.png



aps3_11409_f4.png



aps3_11409_f5.png



aps3_11409_f6.png