Progress Towards Plant Community Transcriptomics: 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]*

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

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

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

*Corresponding author: msbarker@arizona.edu

ABSTRACT

- *Premise of the study:* Large scale projects such as NEON are collecting ecological data on entire biomes to track and understand plant responses to climate change. NEON provides an opportunity for researchers 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 phylogenetically diverse NEON plant communities, including species with diploid and polyploid genomes.
- *Methods:* We used Illumina NextSeq to generate >20 Gb of RNA-seq for each of 24 vascular plant species representing 12 genera and 9 families at the Harvard Forest NEON site. Each species was sampled twice, in July and August 2016. We used Transrate, BUSCO, and GO analyses to assess transcriptome quality and content.
- *Results:* We obtained nearly 650 Gb of RNA-seq data that assembled into more than 755,000 translated protein sequences across the 24 species. We observed only modest differences in assembly quality scores across a range of k-mer values. 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 genes between diploid and polyploid transcriptomes.

- *Discussion:* Our resource provides new RNA-seq datasets 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 clear opportunities for large scale studies at the intersection of ecology and genomics.

INTRODUCTION

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., 2019; McFadden et al., 2019; Rice et al., 2019; Baniaga et al., 2020; Román-Palacios and Wiens, 2020). Recognizing this need, NSF recently launched the National Ecological Observatory Network (NEON) to generate large-scale data on species occurrence, phenology, climate, and more, for ecological communities across the United States (Collinge, 2018; Knapp and Collins, 2019). Metagenomic and genomic sampling is 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, 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. Trancriptome profiling or genome sequencing of multiple species and individuals within a community will open new, integrative avenues of analyses 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, Lauber, et al., 2012; Fierer, Leff, et al., 2012; 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 (HARV). We sampled 24 species of vascular plants from sites adjacent to the NEON plot. Species were selected from a phylogenetically diverse range of plants that included ferns, trees, and herbaceous annuals. For plants in particular, the abundance of polyploid species in communities has the potential to pose challenges for genomic studies. To explore the impacts of polyploidy on transcriptome surveys, we made an effort to select sets of related polyploid and diploid species. Each species was sampled at two different time points in July and August 2016. 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.

METHODS

Taxon selection and sampling

The Harvard Forest Flora (Jenkins et al., 2008) was used to select taxa to represent each category (native/invasive, diploid/polyploid). Invasive species status was determined from the Harvard Forest Flora Database (Jenkins and Motzkin, 2009). 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 eleven diploid species (Table 1). We could not determine the ploidal level of four species. The Harvard Forest Flora Database was used to locate sampling sites.

Tissue from mature leaves was collected from an individual representing each target species at two time points (July and August) during the 2016 growing season. The same individual was sampled at both time points for perennial individuals, and the same population was sampled for annuals. Field sampling for plant RNA-seq followed the protocol described in Yang *et al.* 2017 (Yang et al., 2017). Leaf tissues were flash frozen in liquid nitrogen in the field, and shipped on dry ice to the University of Arizona for RNA extraction.

RNA extraction and RNA-seq

Total RNA was extracted from leaf tissue collected at each time point for all species using the Spectrum Plant Total RNA Kit (Sigma-Aldrich Co., St. Louis, MO, USA) following Protocol A. RNA was used to prepare cDNA using Nugen's Ovation RNA-Seq System via single primer isothermal amplification (Catalogue # 7102-A01) and automated on the Apollo 324 liquid handler (Wafergen). cDNA was quantified on the Nanodrop (Thermo Fisher Scientific) and was sheared to approximately 300 bp fragments using the Covaris M220 ultrasonicator. Libraries were generated using Kapa Biosystem's library preparation kit (KK8201). Fragments were end repaired and A-tailed, and individual indexes and adapters (Bioo, catalogue #520999) were ligated on each separate sample. The adapter ligated molecules were cleaned using AMPure beads (Agencourt Bioscience/Beckman Coulter, A63883), and amplified with Kapa's HIFI enzyme (KK2502). Each library was then analyzed for fragment size on an Agilent's Tapestation, and quantified by qPCR (KAPA Library Quantification Kit, KK4835) on Thermo Fisher Scientific's Quantstudio 5 before multiplex pooling (13-16 samples per lane) and paired-end sequencing at 2x150 bp on the Illumina NextSeq500 platform at Arizona State University's CLAS Genomics Core facility. Raw read quality was assessed using fastQC (Andrews, 2010).

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 polyA/T tails with SeqClean (https://sourceforge.net/projects/seqclean/). The cleaned reads from each sample time point 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 (*Dryopteris intermedia*, *Galium mollugo*, *Juglans cinerea*, *Plantago major*, and *Persicaria sagittata*) 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 v1.03 (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 v1.0.3 (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 structural 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 within 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 gene at a minimum cutoff of 30% sequence similarity over at least 150 sites. Genes 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 gene was aligned against its best hit protein by Genewise 2.2.2 (Birney et al., 2004). Based on the highest scoring Genewise DNA-protein alignments, stop and 'N' containing codons were removed to produce estimated amino acid sequences for each gene. Output included paired DNA and protein sequences with the DNA sequence reading frame corresponding to each protein sequence.

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 v1.0.3 for all scaffolds as well as the subset of sequences that were identified as plant proteins and translated. We evaluated the completeness of our transcriptome coverage with BUSCO v4.0.5 (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.

Gene Ontology (GO) Annotation and Comparison

Gene Ontology annotations of all transcriptomes were obtained through Translated BLAST (Blastx) searches against annotated *A. thaliana* protein database from TAIR (Lamesch et al., 2012) to find the best hit with length of at least 100 bp and an e-value of at least 1e-10. GO annotations 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. The overall ranking of GO category rows was determined by the ranking of GO annotations among the transcriptome of *Lysimachia ciliata*, as an arbitrary standard. The colors of the heatmap represent the percentage of the transcriptome represented by a particular GO category, with red being highest and purple lowest.

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 (Figure 1). 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 relatively high quality. We achieved a mean read depth of 27 Gb 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 (Figure 2), suggesting that our sequencing effort on these libraries was sufficient. The mean complete + fragmented BUSCO percentages were 73.2% against the viridiplantae database and 76% against the eukaryote database (Table 2). We found more hits to sequences in the BUSCO databases with more translated proteins, but the hits plateaued around 20,000 proteins (Figure 3).

Polyploid species did not have significantly more translated proteins than diploids with 31,152 average proteins translated compared to 30,804 (Figure 4A; two-tailed t-test p = 0.95). Similarly, polyploid species did not have a significantly higher proportion of duplicated BUSCO matches than diploids (Figure 4B; 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*). 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 Gb).

Gene ontology (GO) annotations of the transcriptomes of the 24 species were largely similar (Figure 5). 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, such as in *Dryopteris* and *Lysimachia*, but in most cases the species were not clustered with their congeners. *Polygonum cilinode* 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 6,088 translated proteins and nearly 80% of BUSCO genes missing (Table 2).

DISCUSSION

Overall, 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 plant transcriptomes. In particular, the number of scaffolds that matched known plant proteins was consistent with the number of genes in sequenced plant genomes (Michael, 2014; Wendel et al., 2016). For example, our transcriptomes of *Prunus virginiana* and *P. serotina* contained 38,773 and 30,812 translated proteins each. Genomes of related *Prunus* species had similar numbers of annotated genes, including 27,852 in *P. persica* (International Peach Genome Initiative et al., 2013), 41,294 in *P. yedoensis* (Baek et al., 2018), and 43,349 in *P. avium* (Shirasawa et al., 2017). The assemblies are also 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). Like many transcriptome assemblies (Johnson et al., 2012; Carpenter et al., 2019; Patterson et al., 2019), these 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 genes between the diploid and polyploid transcriptome assemblies. Although this could be due to the modest sample size, it may also reflect biological differences in expressed transcriptome size and diversity that impact the number of assembled genes. 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 genes. However, recent research indicates polyploid transcriptomes may be smaller than expected. Research in *Glycine* has found that the expressed transcriptome size of polyploid species are less than 2X the diploid size (Coate and Doyle, 2015; Doyle and Coate, 2018). For example, the transcriptome of the allotetraploid *Glycine dolichocarpa* was 1.4X 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 (Visger et al., 2019). However, smaller transcriptome sizes in polyploids may also be related to which genes are expressed at a given time or 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 comparison to 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 and unique challenges. Building on our past experience with extracting RNA from diverse species (Barker et al., 2008, 2016; Dempewolf et al., 2010; Der et al., 2011; Lai et al., 2012; Dlugosch et al., 2013; Hodgins et al., 2014; 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. 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 RNAwhich could potentially be related to leaf age at the time of sampling. Different RNA extractions methods will be needed to deal with the secondary compounds that are present in mature and senescing tissues. Recovering high quality RNA across a range of time points, in the field, 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 that wish to leverage data being collected within NEON sites. 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 other than NEON the 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.

ACKNOWLEDGEMENTS

We 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 NSF #1550838 to M.S.B. and K.M.D and NSF #1750280 to K.M.D.

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.

DATA AVAILABILITY

Raw reads for all samples for 24 species are deposited on the 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.

REFERENCES

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 of the United States of America* 100: 4649–4654.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Website http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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.

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., 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. 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., Z. Li, T. I. Kidder, and C. R. Reardon. 2016. 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. *The 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. *The 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, E. Bekele, J. M. M. Engels, Q. C. B. Cronk, and L. H. Rieseberg. 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* 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.

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. 2012. 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. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America* 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-Gomez, V. Vandvik, W. D. Pearse, F. D. Schneider, et al. 2019. The open traits network: Using open science principles to accelerate trait-based science across the tree of life. *EcoEvoRxiv* https://doi.org/10.32942/osf.io/kac45.

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: D1178–86.

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. *The 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., and G. Motzkin. 2009. Harvard Forest Flora Database. *Harvard Forest Data Archive: HF116.*

Jenkins, J., G. Motzkin, and K. Ward. 2008. The Harvard Forest Flora: An Inventory, Analysis, & Ecological History. Harvard Forest Paper 28: 266.

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 of the United States of America* 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: D1202–10.

Lassmann, T., Y. Hayashizaki, and C. O. Daub. 2009. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* .

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 of the United States of America* 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 S. McMahon [ed.],. *Methods in Ecology and Evolution / British Ecological Society* 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. *The Plant Journal* 91: 3–21.

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. Chris 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. *The 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 of the United States of America* 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, NY.

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. *The Plant Journal* 93: 747–770.

Swenson, N. G., and F. A. Jones. 2017. Community transcriptomics, genomics and the problem of species co-occurrence. *The 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.

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. *The Journal of Ecology*: 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.

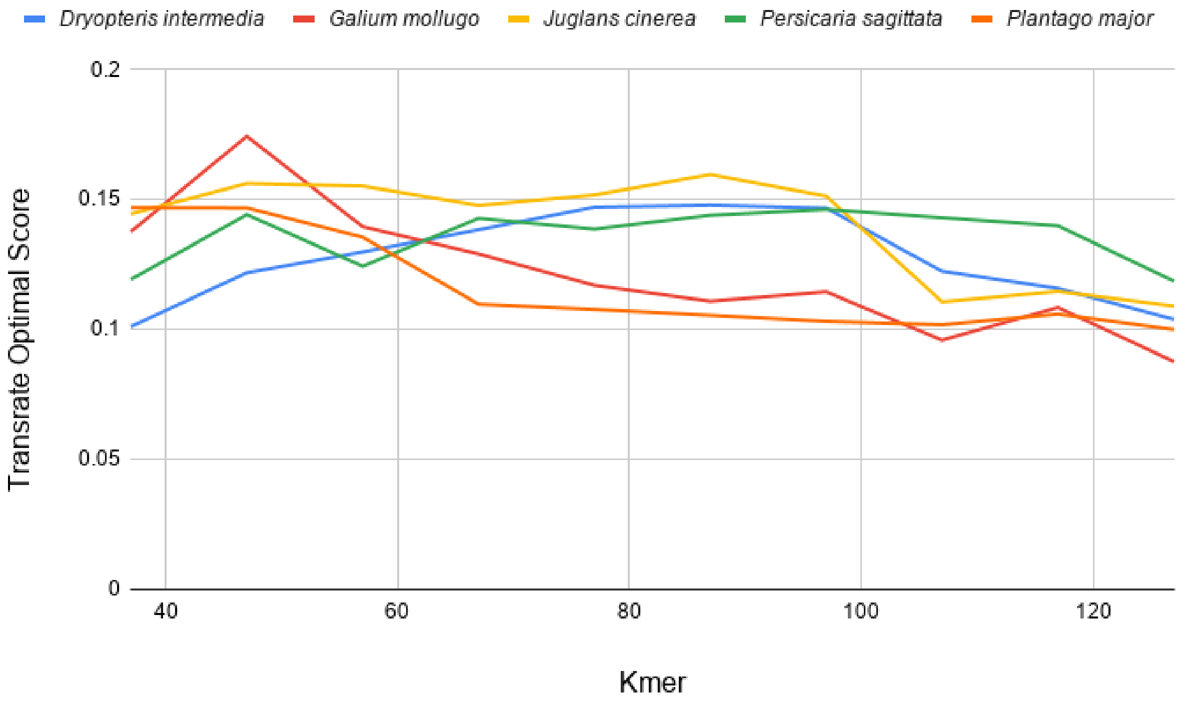

Figure 1. Transrate Optimal Scores for assemblies of five exemplar species from Harvard Forest. A reference transcriptome for each species was assembled with different K-mers starting at K = 37 and increasing in increments of 10 to K = 127.

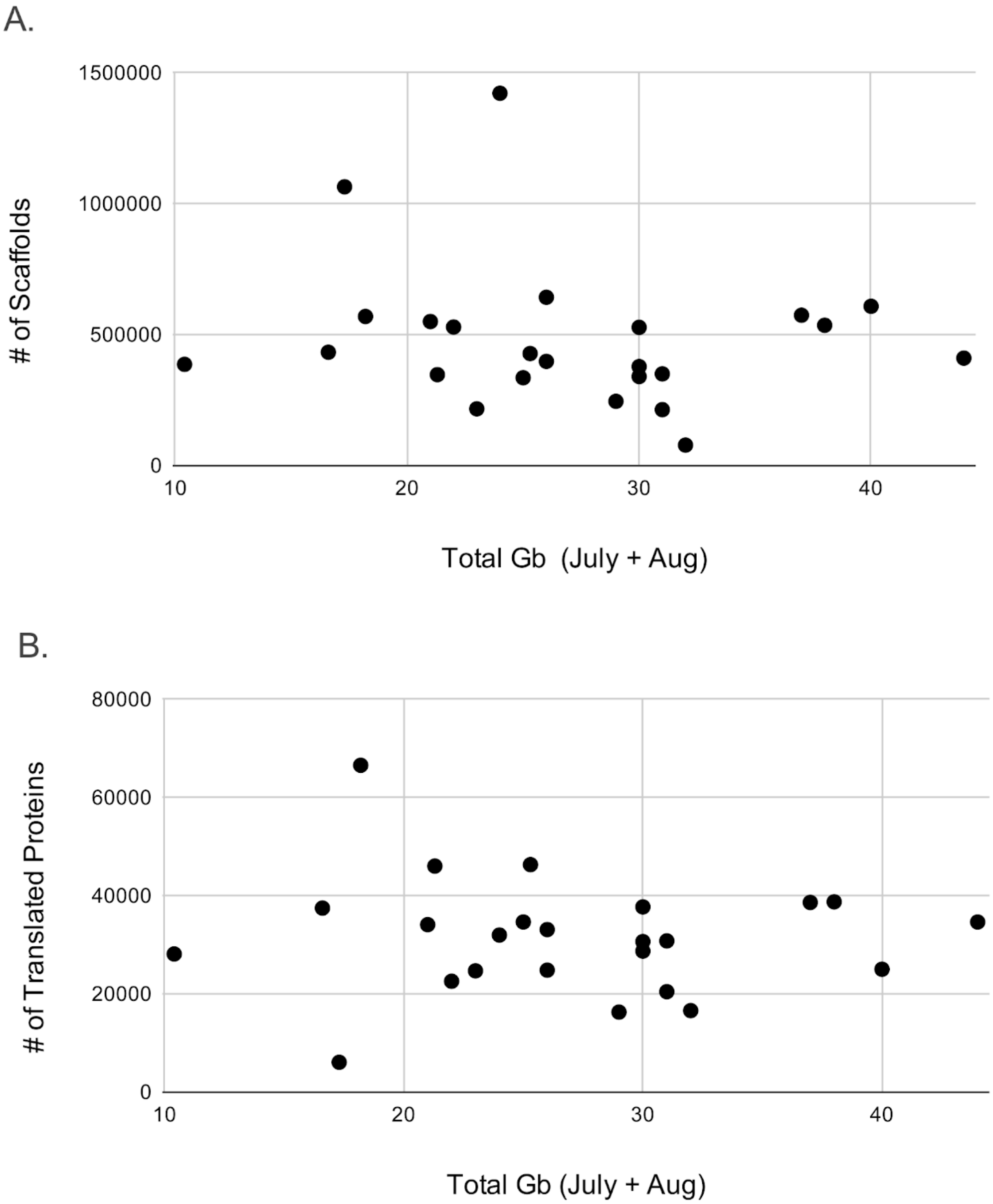

Figure 2. Comparison of the number of A) scaffolds and B) translated proteins produced by each assembly with the total gigabases sequenced for each species.

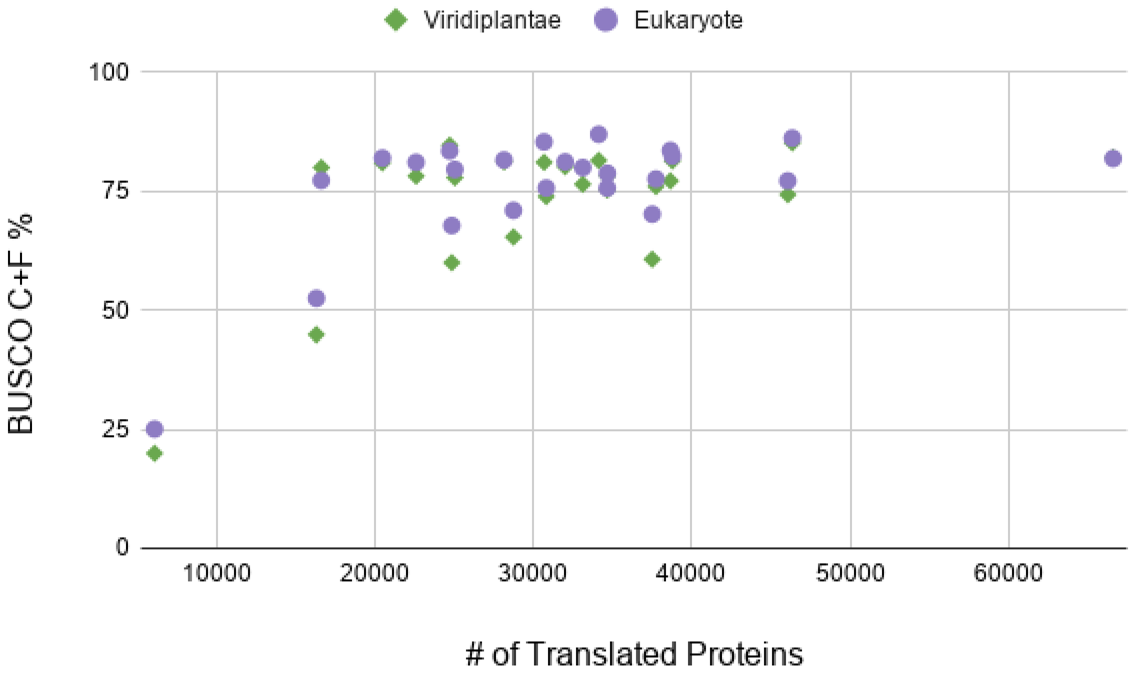

Figure 3. The percentage of BUSCO complete (C) plus fragmented (F) matches compared to the number of translated proteins in each assembly. Green diamonds represent BUSCO matches to the viridiplantae database, whereas purple circles represent matches to the eukaryote database.

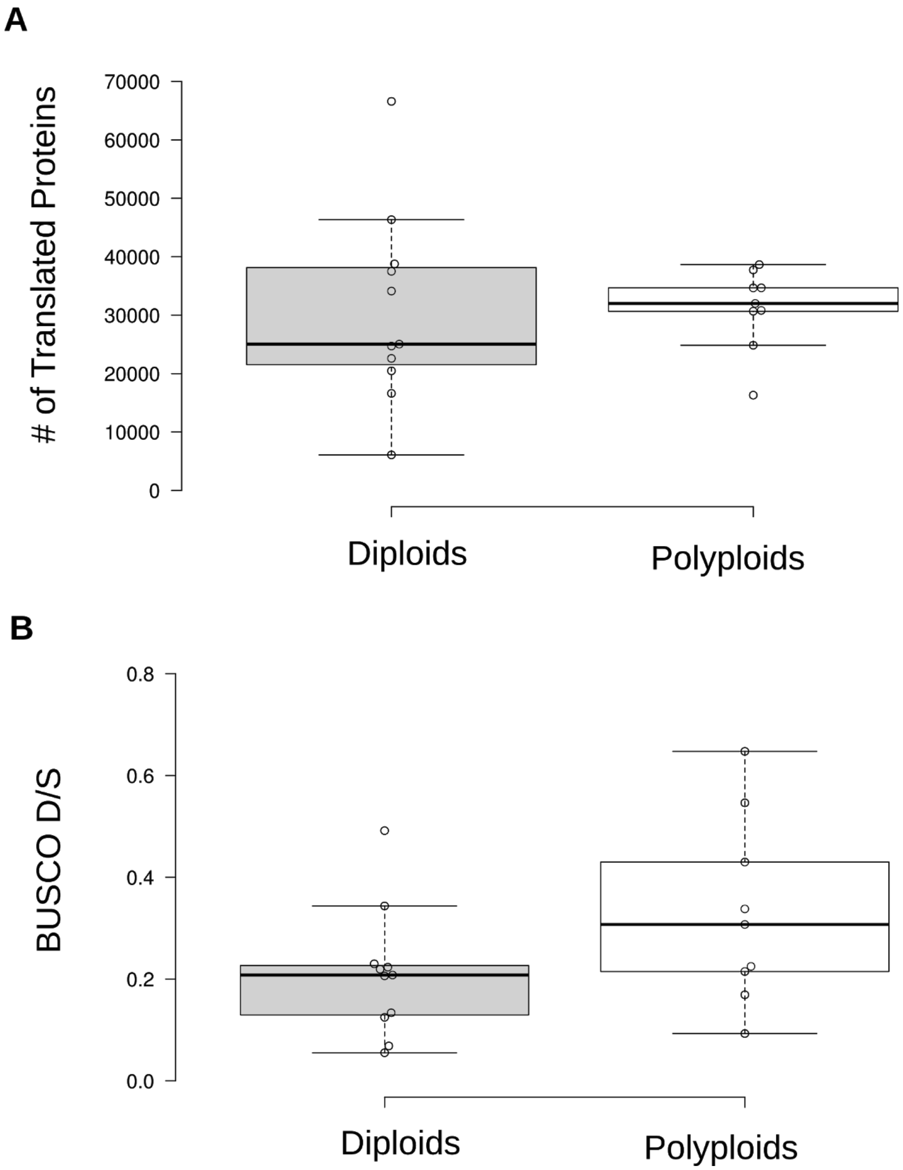

Figure 4. Comparison of A) the number of translated proteins and B) BUSCO duplicated/single copy ratio for assemblies of diploid and polyploid species. In neither case were diploids significantly different from polyploids.

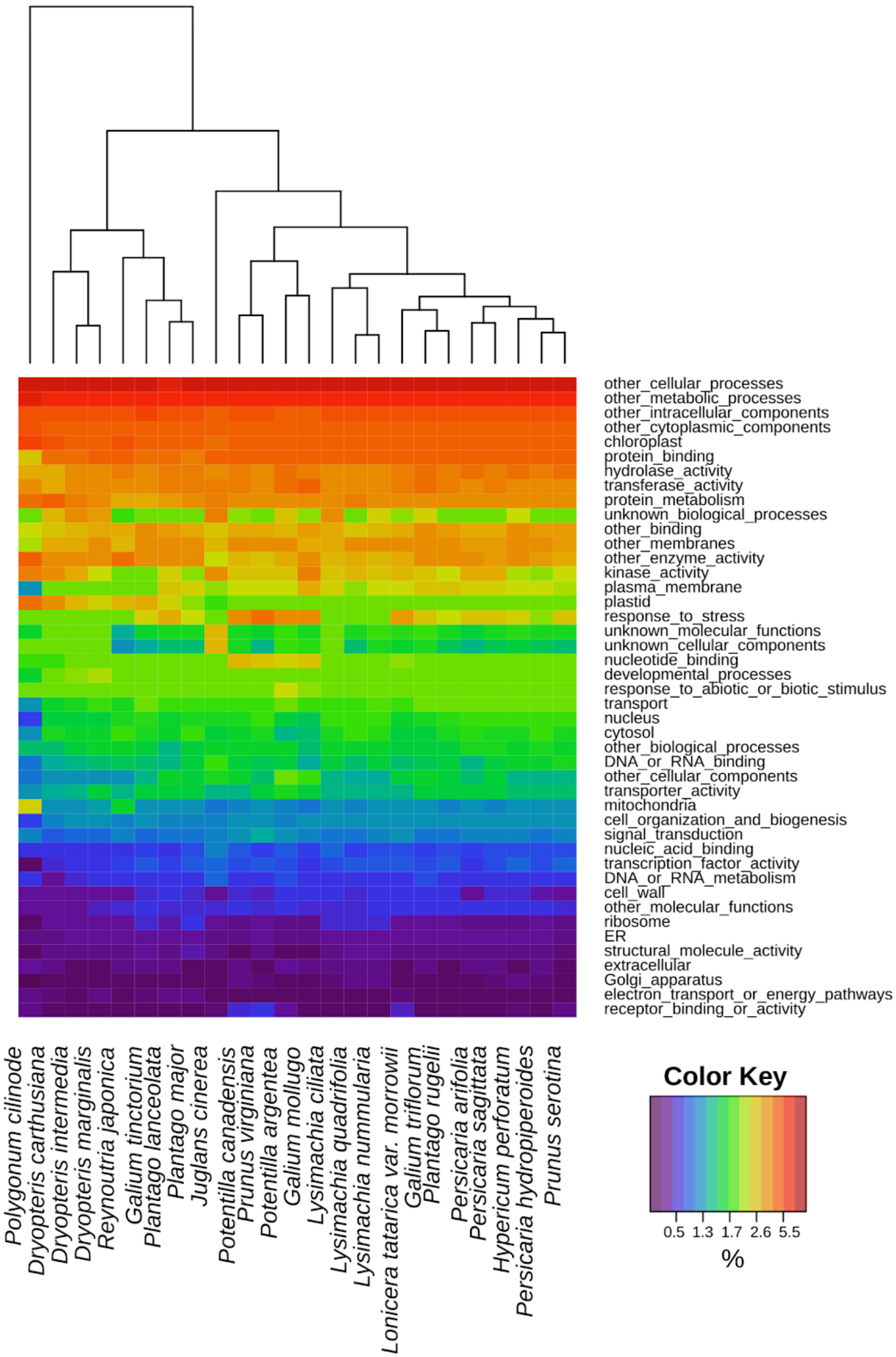
other_cellular_processes
other_metabolic_processes
other_intracellular_components
other_cytoplasmic_components
chloroplast
protein_binding
hydrolase_activity
transferase_activity
protein_metabolism
unknown_biological_processes
other_binding
other_membranes
other_enzyme_activity
kinase_activity
plasma_membrane
plastid
response_to_stress
unknown_molecular_functions
unknown_cellular_components
nucleotide_binding
developmental_processes
response_to_abiotic_or_biotic_stimulus
transport
nucleus
cytosol
other_biological_processes
DNA_or_RNA_binding
other_cellular_components
transporter_activity
mitochondria
cell_organization_and_biogenesis
signal_transduction
nucleic_acid_binding
transcription_factor_activity
DNA_or_RNA_metabolism
cell_wall
other_molecular_functions
ribosome
ER
structural_molecule_activity
extracellular
Golgi_apparatus
electron_transport_or_energy_pathways
receptor_binding_or_activity
Polygonum cilinode
Dryopteris carthusiana
Dryopteris intermedia
Dryopteris marginalis
Reynoutria japonica
Galium tinctorium
Plantago lanceolata
Plantago major
Juglans cinerea
Potentilla canadensis
Prunus virginiana
Potentilla argentea
Galium mollugo
Lysimachia ciliata
Lysimachia quadrifolia
Lysimachia nummularia
Lonicera tatarica var. morrowii
Galium triflorum
Plantago rugelii
Persicaria arifolia
Persicaria sagittata
Hypericum perforatum
Persicaria hydropiperoides
Prunus serotina
Color Key
0.5
1.3
1.7
2.6
5.5
%

Figure 5. 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. P and D are polyploid and diploid species, respectively.

Species	Ploidy	Chrom. #	July SRA	Aug SRA	Total Gb (July + Aug)	# of Scaffolds	Mean Scaffold Length (bp)	Scaffold N50 (bp)	# of Translated Proteins	Mean Translated Length (nucleic acids, bp)	N50 (trans. nucleic acids, bp)
Dryopteris carthusiana	P	82	SAMN08277176	SAMN08277204	26	643129	264.5	710	24851	473.8	543
Dryopteris intermedia	D	41	SAMN08277187	SAMN08277216	22	529510	267.1	822	22595	587.1	702
Dryopteris marginalis	D	41	SAMN08277188	SAMN08277217	21	550548	260.1	917	34121	633.9	789
Galium mollugo	D	11	SAMN08277173	SAMN08277201	40	608764	179.2	1028	25040	692.9	822
Galium tinctorium	D	12	SAMN08277181	SAMN08277210	32	78487	400.6	1091	16610	811.2	1023
Galium triflorum	P	33	SAMN08277189	SAMN08277219	37	574562	246	899	38650	664.9	798
Hypericum perforatum	P	16	SAMN08277171	SAMN08277199	25	335837	233.4	867	34670	698.6	858
Juglans cinerea	D	16	SAMN08277174	SAMN08277202	18.2	569859	359.5	1151	66595	712.4	918
Lonicera tatarica var. morrowii	NA	9	SAMN08277167	SAMN08277195	10.4	386927	324.9	1216	28147	710.9	864
Lysimachia ciliata	P	48	SAMN08277190	SAMN08277220	24	1422451	207	828	32005	631.6	777
Lysimachia nummularia	D	17	SAMN08277194	SAMN08277223	25.3	428232	320.9	1102	46343	716.7	894

Lysimachia quadrifolia	P	42	SAMN08277183	SAMN08277212	30	340491	290.4	944	37737	674.9	813
Persicaria arifolia	NA	NA	SAMN08277180	SAMN08277209	30	528292	245	787	28741	558.8	639
Persicaria hydropiperoides	NA	NA	SAMN08277172	SAMN08277200	21.3	347558	344	913	46058	658.8	795
Persicaria sagittata	NA	20	SAMN08277179	SAMN08277208	26	398304	319.6	1148	33118	725.1	885
Plantago lanceolata	D	6	SAMN08277178	SAMN08277206	31	213834	293.1	1073	20470	742.7	906
Plantago major	D	6	SAMN08277169	SAMN08277197	23	217041	308.6	1032	24715	782.6	993
Plantago rugelii	P	12	SAMN08277170	SAMN08277198	30	378418	276.1	1161	30673	760.5	930
Polygonum cilinode	D	11	SAMN08277184	SAMN08277213	17.3	1065186	186.8	415	6088	471.6	498
Potentilla argentea	P	21	SAMN08277177	SAMN08277207	29	245734	425.6	1268	16306	525.5	687
Potentilla canadensis	D	14	SAMN08277192	SAMN08277222	16.6	433249	216.5	667	37503	526.5	594
Prunus serotina	P	16	SAMN08277186	SAMN08277215	31	350572	267.5	1017	30812	658.3	774
Prunus virginiana	D	8	SAMN08277185	SAMN08277214	38	536216	235	1110	38773	678.8	813
Reynoutria japonica	P	33	SAMN08277193	SAMN08277224	44	410810	279.6	870	34662	549.3	609

Table 2. BUSCO results for comparisons to the viridiplantae and eukaryote databases. C = % all complete, S = % complete and single copy, D = % complete and duplicated, F = % fragmented, M = % missing, D/S = ratio of duplicated to single copy complete sequences, and C+F = % of complete and fragmented BUSCO matches in the respective database.

	BUSCO Viridiplantae DB							**BUSCO Eukaryote DB**						
Species	**C**	**(S)**	**(D)**	**F**	**M**	**D/S**	**C+F**	**C**	**(S)**	**(D)**	**F**	**M**	**D/S**	**C+F**
Dryopteris carthusiana	19.3	16.9	2.4	40.7	40	0.142	60	24.3	20	4.3	43.5	32.2	0.215	67.8
Dryopteris intermedia	39.8	35.1	4.7	38.4	21.8	0.134	78.2	48.2	44.3	3.9	32.9	18.9	0.088	81.1
Dryopteris marginalis	41.5	34.4	7.1	40	18.5	0.206	81.5	48.6	43.5	5.1	38.4	13	0.117	87
Galium mollugo	27.8	22.6	5.2	50.1	22.1	0.230	77.9	38	32.5	5.5	41.6	20.4	0.169	79.6
Galium tinctorium	40	37.9	2.1	40	20	0.055	80	50.2	46.3	3.9	27.1	22.7	0.084	77.3
Galium triflorum	30.8	21.9	8.9	46.4	22.8	0.406	77.2	33.4	21.6	11.8	50.2	16.4	0.546	83.6
Hypericum perforatum	29.5	22.4	7.1	48.7	21.8	0.317	78.2	38.8	29	9.8	40	21.2	0.338	78.8
Juglans cinerea	33.9	27.8	6.1	48.2	17.9	0.219	82.1	47.8	39.2	8.6	34.1	18.1	0.219	81.9
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	18.4	0.276	81.6
Lysimachia ciliata	34.1	28.7	5.4	46.1	19.8	0.188	80.2	49	40	9	32.2	18.8	0.225	81.2
Lysimachia nummularia	42.4	35.1	7.3	42.8	14.8	0.208	85.2	48.6	34.9	13.7	37.6	13.8	0.393	86.2
Lysimachia quadrifolia	31.8	26.4	5.4	44.2	24	0.205	76	38	32.5	5.5	39.6	22.4	0.169	77.6

Persicaria arifolia	13.9	10.6	3.3	51.5	34.6	0.311	65.4	24.7	16.9	7.8	46.3	29	0.462	71
Persicaria hydropiperoides	25.4	16.5	8.9	48.9	25.7	0.539	74.3	32.9	18.4	14.5	44.3	22.8	0.788	77.2
Persicaria sagittata	27.8	16.5	11.3	48.7	23.5	0.685	76.5	39.2	20.4	18.8	40.8	20	0.922	80
Plantago lanceolata	40.5	36	4.5	40.5	19	0.125	81	45.9	39.2	6.7	36.1	18	0.171	82
Plantago major	51.3	48	3.3	33.4	15.3	0.069	84.7	54.1	49.8	4.3	29.4	16.5	0.086	83.5
Plantago rugelii	34.5	20.9	13.6	46.6	18.9	0.651	81.1	45.8	27.8	18	39.6	14.6	0.647	85.4
Polygonum cilinode	11.5	9.4	2.1	8.5	80	0.223	20	5.5	4.7	0.8	19.6	74.9	0.170	25.1
Potentilla argentea	11.7	10.8	0.9	33.2	55.1	0.083	44.9	14.1	12.9	1.2	38.4	47.5	0.093	52.5
Potentilla canadensis	12.9	9.6	3.3	47.8	39.3	0.344	60.7	17.3	12.2	5.1	52.9	29.8	0.418	70.2
Prunus serotina	23.5	18.1	5.4	50.4	26.1	0.298	73.9	28.6	20	8.6	47.1	24.3	0.430	75.7
Prunus virginiana	27	18.1	8.9	54.4	18.6	0.492	81.4	37.7	25.5	12.2	44.7	17.6	0.478	82.4
Reynoutria japonica	30.1	22.8	7.3	45.2	24.7	0.320	75.3	30.2	23.1	7.1	45.5	24.3	0.307	75.7