# QA/QC

Quality scores and potential adapter contaminants were screened using FastQC version 0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All libraries were of good quality for assembly and expression profiling. Reads were trimmed of low-quality bases and adapter contamination with Trimmomatic version 0.33 (Bolger et al. 2014). Reads that were shorter than 18bp after trimming were discarded, and both reads in a pair were discarded if one read was discarded. Trimmed reads were then cleaned of ribosomal RNA sequences. A nonredundant collection of ribosomal RNA sequences was derived from the SILVA database (Quast et al. 2013) using the “dedupe2” tool from the BBTools suite (cite). The reads were cleaned of ribosomal RNA sequences by filtering with BBDuk version 38.39 (cite); reads that shared a 31-mer with one of the nonredundant ribosomal sequences were removed.

# Assembly

The reads that survived the base quality, adapter content, sequence length, and ribosomal RNA filter were concatenated into single forward and reverse read files for assembly. The transcriptome was assembled with Trinity version 2.8.6 (Grabherr et al. 2011; Haas et al. 2013). Trinity’s “in silico read normalization” routine was used, and the target coverage was set to 200bp to reduce runtime and memory requirements. The minimum contig size was set to 250bp, and the K-mer size was set to 25bp.

# Expression Profiling

To measure the gene expression in each of the eight tissues, the reads from each of the libraries were separately re-processed. Individual samples were trimmed of low-quality bases and adapter contamination with Trimmomatic, using the same parameters as described in the QA/QC section above. Reads were not cleaned for ribosomal RNA content to allow for an estimate of the proportion of reads from protein coding genes in each tissue. Trimmed reads were mapped to the genome assembly with HISAT2 (Kim et al. 2015; Kim et al. 2019) using known splice sites as estimated from the Funannotate workflow (https://zenodo.org/record/2604804). Resulting BAM files were cleaned of records with a mapping quality of less than 60. The BAM files were then sorted and used as input for gene-level quantification with featureCounts version 1.6.2 from the subread package (Liao et al. 2013). The resulting counts matrix was used to calculate the 𝜏 index of tissue specificity (Yanai et al. 2005). Genes with 𝜏 > 0.8 were considered to be tissue-specific, and the tissue with the greatest normalized expression was identified for each tissue-specific gene.

# Orthogroup Identification

Orthologous gene groups between Northern Wild Rice and other grasses were identified with Orthofinder version 2.3.11 (Emms and Kelly 2019). Proteome sequences from a total of 14 other grass species from Phytozome and Ensembl Plants were used as input to the Orthofinder algorithm (**Table SX**). Orthofinder was run with the “BLAST” option for performing the sequence similarity searches, and the “MSA” option for generating the gene trees of orthologous gene groups.

Scripts to perform high-throughput sequence data QA/QC, assembly, expression analyses, and orthologous gene group identification are available at (**[Public GitHub Link TBD]**). A workflow document that details all analyses, including the exact commands that were executed is available in Appendix SX.