bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A happy accident: a novel turfgrass reference genome

Alyssa R. Phillips[*,†,1], **Arun S. Seetharam**[*,2], **Taylor AuBuchon-Elder**[3], **Edward S. Buckler**[4,5,6], **Lynn J. Gillespie**[7], **Matthew B. Hufford**[2], **Victor Llaca**[8], **M. Cinta Romay**[5], **Robert J. Soreng**[9], **Elizabeth A. Kellogg**[3] **and Jeffrey Ross-Ibarra**[†,1,10]

[*]Co-first authors, [†]Corresponding authors, [1]Dept. of Evolution and Ecology and Center for Population Biology, University of California, Davis, CA, USA, [2]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, IA, USA, [3]Donald Danforth Plant Science Center, Olivette, MO, USA, [4]School of Integrative Plant Sciences, Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA, [5]Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA, [6]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, USA, [7]Botany Section, Research and Collections, Canadian Museum of Nature, Ottawa, ON, Canada, [8]Corteva Agriscience, Johnston, IA, USA, [9]Deptartment of Botany, Smithsonian Institution, Washington, DC, USA, [10]Genome Center, University of California, Davis, CA, USA

ABSTRACT *Poa pratensis*, commonly known as Kentucky bluegrass, is a popular cool-season grass species used as turf in lawns and recreation areas globally. Despite its substantial economic value, a reference genome had not previously been assembled due to the genome's relatively large size and biological complexity that includes apomixis, polyploidy, and interspecific hybridization. We report here a fortuitous *de novo* assembly and annotation of a *P. pratensis* genome. The draft assembly consists of 6.09 Gbp with an N50 scaffold length of 65.1 Mbp, and a total of 118 scaffolds, generated using PacBio long reads and Bionano optical map technology. We annotated 256K gene models and found 58% of the genome to be composed of transposable elements. To demonstrate the applicability of the reference genome, we evaluated population structure and estimated genetic diversity in three North American wild *P. pratensis* populations. Our results support previous studies that found high genetic diversity and population structure within the species. The reference genome and annotation will be an important resource for turfgrass breeding and biologists interested in this complex species.

KEYWORDS Poaceae, genome assembly, aneuploidy, polyploidy, genetic diversity, population structure, Kentucky bluegrass, turfgrass

Introduction

Poa pratensis L., commonly known as Kentucky bluegrass, is an economically valuable horticultural crop grown globally on lawns and recreational areas as turf (Haydu *et al.* 2006). Native to Europe and Asia, it was introduced to North America in the seventeenth century by European colonizers as a forage crop (Carrier and Bort 1916; Raggi *et al.* 2015). Today, Kentucky bluegrass is the most popular cool-season grass used for turf due to it's vigorous growth and quick establishment that creates a dense, strong sod with a long lifespan (Casler and Duncan 2003).

Today, there are 40 million acres of managed turf in the United States (U.S.), an area approximately the size of the state of Florida (Milesi *et al.* 2005). While this massive area has the potential to serve as an important carbon sink, the large water and fertilization resources required currently outweigh the benefits (Milesi *et al.* 2005). Breeding efforts are underway to improve environmental-stress tolerances, disease and insect resistance, seed quality and yield, as well as uniformity and stability of traits (reviewed in Bonos and Huff 2013). While the economic value of *P. pratensis* is high, it is highly invasive, and in the last 30 years has aggressively invaded the North American Northern Great Plains, altering ecosystem function by reducing pollinator and plant diversity and altering nutrient dynamics (Kral-O'Brien *et al.* 2019; DeKeyser *et al.* 2015; Hendrickson *et al.* 2021). Continued research into the genetic diversity of wild *P. pratensis* is needed to understand how invasive populations are rapidly adapting, and the study of wild populations may enable identification of disease or environmentally tolerant ecotypes for use in turfgrass breeding.

Previous studies using RAPD, ISSR, and SRR markers demonstrated high genetic diversity in both developed cultivars and wild populations but limited population structure between groups (Bushman *et al.* 2013; Raggi *et al.* 2015; Honig *et al.* 2012, 2018, but see Dennhardt *et al.* 2016). Population divergence has been detected amongst some wild populations (Dennhardt *et al.* 2016) but the extent of population structure is unclear. There are a number of potential reasons for these findings, including gene flow, the independent development of cultivated lines from locally adapted ecotypes (Raggi *et al.* 2015; Bonos and Huff 2013), and geographic heterogeneity in patterns of genetic diversity. Repeated reversion of cultivars to wild forms has also been suggested, but may be unlikely (Dennhardt *et al.* 2016). Alternatively, previous studies may simply not have had sufficient marker resolution to detect population structure in a highly heterozygous polyploid species like *P. pratensis*.

Genetic analysis and improvement of turfgrass are challenging because of apomixis and polyploidy (Bushman and Warnke 2013). *Poa pratensis* is a facultative apomict, meaning it can reproduce sexually or asexually by aposporous apomixis, and it is a polyploid with frequent aneuploidy (Brown 1939). Although apomixis is a highly valued trait for seed production, high rates of apomixis stymie the recombination needed to genetically analyze traits or recombine beneficial traits into one cultivar (Bonos and Huff 2013). Polyploidy and aneuploidy fur-

[1] Dept. of Evolution and Ecology, University of California, Davis, CA, USA E-mail: arphillips@ucdavis.edu; rossibarra@ucdavis.edu

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ther these difficulties due to copy number variation of regions of interest and non-Mendelian inheritance resulting from double reduction. While some progress has been made in managing apomixis (Funk *et al.* 1967; Pepin and Funk 1971; Matzk 1991), including the discovery of its genetic basis (Albertini *et al.* 2004; **?**), the development of additional molecular and genomic tools in *P. pratensis* are needed to move genetic analysis and efforts forward in the face of its complex biology.

Here, we report the first *P. pratensis* genome. While attempting to assemble the genome for a C4 prairie grass, *Andropogon gerardi*, we unknowingly sequenced and assembled a wild *Poa* growing in the same pot. Fortunately, this resulted in a highly contiguous, near complete genome assembly for *Poa*. The long reads used in the assembly not only provided increased resolution of repetitive regions, but also captured haplotypic variations present within the genome. We utilized the reference genome and wild *Poa* from three populations to investigate the genetic diversity and population structure of North American *Poa*. The reference genome and annotation presented here are an important advance for Kentucky bluegrass breeding and conservation. Additionally, this reference genome provides an important resource for the study of closely related bluegrasses including *P. trivialis* L., *P. annua* L., and *P. arachnifera* Torr.

Materials and Methods

Sample collection

Rhizomes of *Poa* spp. were collected fortuitously as part of a different project aimed at collecting major C4 prairie grasses (*Andropogon gerardi* Vitman, *Sorghastrum nutans* (L.) Nash, and *Schizachyrium scoparium* (Michx.) Nash) in moist prairies in Manitoba, Canada and Colorado, USA (Supplement 2). Necessary permissions and permits were obtained prior to collecting. Plants were brought back to the United States from Canada under phytosanitary certificate 3193417.

The C4 focal plants were dug up with a shovel late in the growing season in 2018 (when the *Poa* was dormant and thus invisible), soil was washed off, rhizomes were wrapped in wet paper towels, and leaves were cut back to about 4 inches height to reduce transpiration. The focal C4 plant was placed in a 1-gallon Ziploc bag and returned to the plant growth facility at the Donald Danforth Plant Science Center in St. Louis, MO, USA. Plants were potted in 2:1 BRK20 promix soil to turface. The previously dormant *Poa* plants produced fresh green leaves in this setting and grew faster than the C4 plant with which it was entwined. Once it was discovered that *Poa* had interpolated itself into the rhizome and root area of the C4 plants, the *Poa* plants were extricated and placed in separate pots.

One *Poa* was found inside the pot for an *Andropogon gerardi* genotype which was used to attempt assembly of a reference genome. Instead of collecting tissue from the *A. gerardi*, tissue was accidentally sampled from the *Poa* plant. This *Poa* individual is referred to as the *Poa* reference individual (Supplement 2). Eight additional *Poa*, referred to here as the *Poa* population panel, were discovered in various pots for C4 grasses whose genomes we attempted to sequence.

As *Poa* species generally require vernalization to flower, several plants were over-wintered outside under mulch and flowered in spring 2020 and/or 2021; voucher specimens were taken from these plants to verify species identity and have been deposited at the Smithsonian Institution (Washington, District of Columbia, U.S.A) and the Missouri Botanical Garden (St. Louis, MO, U.S.A.) (Heide 1994). Not all *Poa* individuals survived, so some specimens lack vouchers. Additionally, not all surviving *Poa* flowered so vegetative vouchers were submitted (Supplement 2).

PacBio sequencing

Approximately 4.1g fresh tissue from the reference individual was extracted for PacBio sequencing using a igh Molecular Weight (HMW) DNA approach based on the Circulomics Big DNA Kit (Circulomics, USA). This method yields DNA with a center of mass at 200 Kb, which is sufficient to construct PacBio CLR 20kb+ libraries. Sequencing was completed on the Sequel II across four SMRTCells. DNA extraction and sequencing was completed by Corteva Agriscience^{TM}.

Bionano optical map generation

DNA was extracted from 0.7 g of fresh leaf tissue from the reference individual using agarose embedded nuclei using the Bionano Prep^{TM} Plant Tissue DNA Isolation kit. DNA extraction, labeling, imaging, and optical map assembly followed the methods previously described in Hufford *et al.* (2021) and was completed by Corteva Agriscience^{TM}.

Genome size estimation

Genome size was estimated for the *Poa* reference individual and 5 of the population panel individuals (Supplement 2). Not all population panel individuals were sampled as some plants died prior to estimation. Genome size estimation methods using an internal standard are modified from Doležel *et al.* (2007). The maize B73 inbred line was used as the internal standard (5.16 2C/pg). Approximately 10x1cm of fresh leaf tissue for the target and sample standard were placed in a plastic square petri dish. A chopping solution composed of 1mL LB01 buffer solution, 250µL PI stock (2mg/mL), and 25µL RNase (1 mg/mL) added to the dish (1.25 mL; (Doležel *et al.* 2007)). The tissue was then chopped into 2-4 mm lengths and the chopping solution was mixed through the leaves by pipetting. The solution was then pipetted through a 30µm sterile single-pack CellTrics® filter into a 2mL Rohren tube on ice. Three replicates were chopped separately and analyzed for each *Poa* individual. The samples were left to chill for 20 minutes before analysis with a BD Accuri^{TM} C6 flow cytometer. Samples were run in Auto Collect mode with a 5-minute run limit, slow fluidics option, a FSC-H threshold with less than 200,000 events, and a 1-cycle wash. The cell count, coefficient of variation of FL2-A, and mean FL2-A were recorded for the target and reference sample with no gating. Results were analyzed separately for each replicate and manually annotated to designate the set of events. The three replicates for each *Poa* individual were averaged to calculate genome size.

Illumina sequencing of the Poa population panel

DNA was extracted from the *Poa* population panel using approximately 100 mg of lyophilized leaf tissue and a DNeasy® Plant Kit (Qiagen Inc., Germantown, MD). High throughput Illumina Nextera ® libraries were constructed and samples were sequenced with other plant samples in pools of 96 individuals in one lane of an S4 flowcell in an Illumina Novaseq 6000 System with paired-end 150bp reads, providing approximately 0.80X coverage for each sample.

Species identification

Species identification was completed using both morphological and DNA sequence data. Morphological assessment was completed for the *Poa* reference genome and three of the population

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

panel samples using flowering and vegetative vouchers. Phylogenetic inference was completed for species identification of all samples using one plastid and two nuclear ribosomal DNA loci: *trn*T-*trn*L-*trn*F (TLF), external transcribed spacer (ETS), and internal transcribed spacer (ITS), respectively. Trees for *mat*K and *rpo*B-*trn*C were also evaluated but the sequences showed little variation across sampled species.

Sequences for these loci were extracted from the *Poa* population panel whole genome sequence data by aligning reads to a *P. pratensis* sequence for each locus downloaded from Genbank (Supplement 3) using the default options of bwa mem (v0.7.17; Li 2013). The alignment files were sorted using SAMtools (v1.7; Danecek *et al.* 2021), read groups were added using Picard AddOrReplaceReadGroups, and duplicates removed with Picard MarkDuplicates using default settings (http://broadinstitute.github.io/picard). We identified variable sites for each sample separately using GATK (v4.1) HaplotypeCaller with default options (Van der Auwera and O'Connor 2020). SNPs were filtered to remove sites with low mapping quality (< 40) and low sequencing quality (< 40) (`gatk VariantFiltration -filter "QUAL < 40.0" -filter "MQ < 40.0"` and default `gatk SelectVariants`). A consensus sequence for each locus and sample was generated using GATK FastaAlternateReferenceMaker, which replaces the gene reference bases at variable sites with the alternate allele.

Sequences were extracted from the reference genome by aligning the *P. pratensis* sequences downloaded from Genbank to the reference genome with bwa mem using default options (v0.7.17; Li 2013). This allowed us to identify the position of each locus in the reference. Each locus only mapped to a single region in the reference genome, which was extracted using bioawk (https://github.com/lh3/bioawk).

Sequences from the reference genome and the population panel were included in a dataset with 119 *Poa* samples from previous work (Supplement 4; Cabi *et al.* 2016, 2017; Gillespie *et al.* 2007, 2008, 2009, 2018; Giussani *et al.* 2016; Refulio-Rodriguez *et al.* 2012; Soreng and Gillespie 2018; Soreng *et al.* 2015, 2017, 2020; Sylvester *et al.* 2021). These samples were chosen to represent the phylogenetic diversity of the genus *Poa*, and include all seven currently recognized subgenera as well as 29 of 38 sections and several unclassified species groups (classification according to Gillespie *et al.* (2007), with updates by Cabi *et al.* (2017); Gillespie *et al.* (2008, 2018); Soreng and Gillespie (2018); Soreng *et al.* (2020)). Since formal infrageneric taxonomic delimitations are often imperfect, and the genus *Poa* is large and highly complex, genotype codes are used in Supplement 4 as shorthand for the plastid and nrDNA clades found in a sample or species (see Soreng *et al.* (2020) for the most recent iterations).

Sequences were aligned using the auto-select algorithm and default parameters in the MAFFT plugin (v7.017; Katoh and Standley 2013) in Geneious (v8.1.9; http://www.geneious.com) followed by manual adjustment. *Poa* sect. *Sylvestres* was used as the outgroup to root trees based on its strongly supported position as sister to all other *Poa* species in previous plastid analyses (Gillespie *et al.* 2007, 2009, 2018). Bayesian Markov chain Monte Carlo analyses were conducted in MrBayes (v3.2.6; Ronquist *et al.* 2012). Optimal models of molecular evolution were determined using the Akaike Information Criterion (AIC; Akaike 1974) conducted through likelihood searches in jModeltest (Darriba *et al.* 2012) with default settings. Models were set at `GTR` + Γ for ETS and `GTR + I +` Γ for ITS and TLF based on the AIC scores and the models allowed in MrBayes. Two independent runs of four chained searches were performed for three or four million generations, sampling every 500 generations, with default parameters. Analyses were stopped when a split frequency of 0.005 was closely approached. A 25% burn-in was implemented prior to summarizing a 50% majority rule consensus tree and calculating Bayesian posterior probabilities. Trees were visualized and annotated in R using ggtree (v2.0.4) with ape (v5.4) and treeio (v1.10) (Yu 2020; R Core Team 2017; Wang *et al.* 2020; Paradis and Schliep 2019).

Genome assembly

PacBio subreads obtained as BAM files were converted to FASTA format using SAMtools (v1.10; Danecek *et al.* 2021) and error-correction was performed using overlap detection and error correction module (first stage) of Falcon (v1.8.0; Chin *et al.* 2016). For running Falcon, the following options were used: the expected genome size was set to 6.4 Gbp (`-genome_size = 6400000000`), a minimum of two reads, maximum of 200 reads, and minimum identity of 70% for error corrections (`-min_cov 2 -max_n_read 200, -min_idt 0.70`), using the 40x seed coverage for auto-calculated cutoff. The average read correction rate was set to 75% (`-e 0.75`) with local alignments at a minimum of 3000 bp (`-l 3000`) as suggested by the Falcon manual. For the DAligner step, the exact matching length of k-mers between two reads was set to 18 bp (`-k 18`) with a read correction rate of 80% (`-e 0.80`) and local alignments of at least 1000 bp (`-l 1000`). Genome assembly was performed with Canu (v1.9: Koren *et al.* 2017) using the error-corrected reads from Falcon. For sequence assembly, the corrected reads had over 70x coverage for the expected genome size of *Poa* and were characterized by N50 of 25.6 Kbp and average length of 16.3 Kbp. These reads were trimmed and assembled with Canu using the default options except for `ovlMerThreshold=500`.

The Canu generated contig assembly was further scaffolded utilizing the Bionano optical map with Bionano Solve (v3.4) and Bionano Access (v1.3.0), as described previously by Hufford *et al.* 2021. The default config file (hybridScaffold_DLE1_config.xml) and the default parameters file (optArguments_nonhaplotype_noES_noCut_DLE1_saphyr.xml) were used for the hybrid assembly. The scaffolding step of Bionano Solve incorporates three types of gaps: 1) gaps of estimated size (varying N-size, but not 100bp or 13bp), using calibrated distance conversion of optical map to basepair (cases when contiguous optical map connects two contigs); 2) gaps of unknown sizes (100-N gaps), when distance could not be estimated (cases when large repeat regions like rDNA or centromeres interrupt the optical map but evidence to connect the map is present); and 3) 13-N gaps, in regions where two or more independently assembled contigs align to the same optical map, overlapping at the ends. The 13-N gaps are usually caused by sequence similarity sufficient for aligning to the optical map, but less than required to merge contigs. This could be caused by either high heterozygosity in that region, highly repetitive sequence, paralogous regions of the sub-genomes, or assembly errors. The contig overlaps, regardless of the size, are connected end-to-end by adding 13-N gaps when processed using Bionano Solve. Due to the polyploid nature of Poa as well as its high heterozygosity, these 13-N gaps had to be manually curated. Using Bionano Access (v1.3.0) we inspected the contig alignments to the optical map, either to trim the overlapping sequence or to remove exact duplicates to generate error-free assembly.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genome annotation

Gene prediction was carried out using a comprehensive method combining *ab initio* predictions (from BRAKER v2.1.6; Brůna *et al.* 2021) with direct evidence (inferred from transcript assemblies) using the BIND strategy (Li *et al.* 2021). Briefly, 58 RNA-seq libraries were downloaded from NCBI (Supplement 5) and mapped to the genome using a STAR (v2.5.3a; Dobin *et al.* 2013)-indexed genome and an iterative two-pass approach under default options to generate mapped BAM files. BAM files were used as input for multiple transcript assembly programs to assemble transcripts: Class2 (v2.1.7; Song *et al.* 2016), Cufflinks (v2.2.1; Trapnell *et al.* 2012), Stringtie (v2.1.4; Pertea *et al.* 2015) and Strawberry (v1.1.2; Liu and Dickerson 2017). Redundant assemblies were collapsed and the best transcript for each locus was picked using Mikado (v2.3.3; Venturini *et al.* 2018) by filling in the missing portions of the ORF using TransDecoder (v5.5.0; Haas *et al.* 2013) and homology as informed by the NCBI BLASTX (v2.10.1+; Altschul *et al.* 1990) results to the SwissProtDB Duvaud *et al.* 2021. Splice junctions were also refined using Portcullis (v1.2.1; Mapleson *et al.* 2018) to identify isoforms and to correct misassembled transcripts. Both *ab initio* and direct evidence predictions were analyzed with TESorter (v1.3.0; Zhang *et al.* 2019) to identify and remove any TE-containing genes before merging them. Merging was done using the GeMoMa (v1.8) Annotation Filter tool, to combine and filter gene predictions from BRAKER, Mikado and additional homology-based gene predictions generated by the GeMoMa pipeline using *Hordeum vulgare* annotations (Mascher *et al.* 2021). The predictions were prioritized using weights, with highest for homology (1.0), followed by direct evidence (0.9) and lowest for gene predictions from *ab initio* methods (0.1). Homology is defined by GeMoMa as protein sequence similarity and and intron position conservation relative to *Hordeum vulgare*. The Annotation Filter was run with settings to enforce the completeness of the prediction (`start=='M' stop=='*'`), external evidence support (`score/aa>=0.75`), and RNAseq support (`evidence>1` or `tpc==1.0`). The final predictions were subjected to phylostratiography analyses using phylostratr (v0.20; Arendsee *et al.* 2019) species specific genes (orphan genes) as well as genes belonging to various strata. Final gene-level annotations were saved in GFF3 format and the predicted peptides/CDS sequences were extracted using `gffread` of the Cufflinks package (v2.2.1; Trapnell *et al.* 2012).

Assessment of the assembly

Genome contiguity statistics were computed using the Assemblathon script (Bradnam *et al.* 2013). Gene space completeness was measured using BUSCO (v4.0; Manni *et al.* 2021) using the liliopsida_odb10 profile (n = 3278) and poales_odb10 profile (n = 4896) with default options. The contiguity of TE assembly was then assessed using the LTR Assembly Index (LAI; Ou *et al.* 2018). To compute LAI, we first annotated repeats using the Extensive *de-novo* TE Annotator (EDTA v1.9.6; Ou *et al.* 2019), and intact LTR retrotransposons (LTR-RT) were identified using LTRharvest (v1.6.1; Manchanda *et al.* 2020), and LTR_FINDER_parallel (v1.1; Ellinghaus *et al.* 2008). LTR_retriever (v2.9.0; Ou *et al.* 2018) was then used to filter the intact LTRs and computed the LAI score for the genome.

Population genetics of Poa

The population panel was mapped to the scaffold assembly, excluding the alternate scaffolds, using bwa mem (v0.7.17) (Li 2013). Reads were sorted using SAMtools (v1.7; Danecek *et al.* 2021), read groups were added using Picard AddOrReplaceReadGroups, and duplicates removed with Picard MarkDuplicates (http://broadinstitute.github.io/picard) using default settings. Genotype likelihoods (GLs) were utilized for the population genetic analyses to account for uncertainty in genotyping resulting from low sequence coverage. To evaluate the relationship between the sampled *Poa* populations, GL were called in beagle format for all individuals using the SAMtools GL method implemented in ANGSD (v0.934; Korneliussen *et al.* 2014) (`angsd -GL 1 -doGlf 2`) (Li 2011). Reads were filtered prior to GL calculation, retaining unique reads, reads with a flag below 255, and proper pairs (`-uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0`), as well as a minimum mapping and base quality of 30 (`-minMapQ 30 -minQ 30`). ANGSD assumes sites are diploid when estimating GLs. Although *P. pratensis* is a polyploid, we can identify sites that are diploid-like by excluding sites where paralogs may be mapping using a strict maximum depth cutoff. Assuming read depth follows a Poisson distribution with a mean of 0.8, we expect 99% of reads to have a depth of 4 or less. We included sites with a minimum depth of 1 and a maximum depth of 4 and required all genotypes to have data at a site (`-doCounts 1 -setMinDepthInd 1 -setMaxDepthInd 4 -minInd 8`). These GL were then used to evaluate population structure using a principal component analysis (PCA) implemented in PCAngsd (v1.02; Meisner and Albrechtsen 2018).

Nucleotide diversity (θ_π) was estimated across *P. pratensis* and within the *P. pratensis* of the Boulder population in ANGSD. GLs were re-called following the parameters described above and the site allele frequency likelihood (SAF) was calculated (`-doSaf 1`) for the two groups: all Boulder *P. pratensis* and one genotype from each *P. pratensis* population (Boulder, Tolstoi, and Argyle). The SAF was used to estimate the global folded site frequency spectrum (SFS) (`realSFS -fold 1`) and θ_π was calculated for each site (`realSFS saf2theta`) (Nielsen *et al.* 2012). Then, θ_π was calculated in 10,000 bp sliding windows (`thetaStat do_stat -win 10000 -step 10000`). Windows of size 50 Kbp and 1 Kbp were also evaluated and θ_π was not greatly impacted by window size. Results are reported for the 10 Kbp window size. Windows with fewer than 10% of sites sequenced were dropped, and the window-wise θ_π of the remaining windows was normalized by the number of sites sequenced in the window. Average genome-wide θ_π per bp was calculated as the mean of these windows. PCA and nucleotide diversity results were visualized with ggplot2 in R (R Core Team 2017; Wickham 2016).

Data availability

The genome assembly and annotations are available from the European Nucleotide Archive (ENA) at X. The raw Illumina sequence data for the *Poa* population panel is available from NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA730042. The code for the entirety of assembly, annotation, and population genetic analyses is documented at https://github.com/phillipsar2/poa_genome.

Results and Discussion

Species identification and validation

Herbarium vouchers for the *Poa* reference genome and 2 of the population panel genotypes were identified as *P. pratensis* by their morphology (Supplement 2).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The *Poa* reference genotype can be further classified as subspecies *angustifolia*, characterized by narrower and involute leaf blades, usually with strigose hairs on the adaxial surface of blades. The blades of *P. pratensis* subspecies *angustifolia* are firmer and tend to be more consistently glaucous. The intravaginal shoots are often disposed in fascicles of more than one shoot, the inflorescences are generally narrower, and the spikelets are smaller than other *P. pratensis* subspecies (Soreng and Barrie 1999; Soreng 2007; Cope and Gray 2009). This subspecies is the most likely classification for the reference genotype, although the infraspecies structure is complex and the subspecies genetically and morphologically grade into one another (Soreng and Barrie 1999; Soreng 2007; Cope and Gray 2009).

The remaining *Poa* population genotypes did not survive long enough for detailed morphological identification. We identified the remaining genotypes, and confirmed the morphological IDs, using phylogenetic inference with three commonly used loci (ETS, ITS, TLF). The reference genome was identified as *P. pratensis* by all three loci and 7 of the 8 genotypes in the *Poa* population panel were identified as *P. pratensis* by two of the three loci (ITS and ETS; Figure 1;Supplementary figures S1-3; Supplement 2). The 7 *P. pratensis* genotypes in the population panel held an unresolved position within the subgenus *Poa* in the TLF tree (Figure S3). The eighth population panel genotype was identified as *P. compressa* L. by all three loci. Phylogenetic identification thus supports our morphological identification of the reference genome as *P. pratensis*.

Genome assembly

Error-corrected PacBio reads (100 Gb; 70X coverage) were assembled into 27,953 contigs. The contig assembly was oriented and further scaffolded using a Bionano optical map resulting in 118 primary scaffolds and 10 alternate scaffolds (Table 1).

The assembly is approximately 124% of the genome size estimated using flow cytometry (4900 pg/1C; Table 1). The flow cytometry estimate suggests the genotype is likely an octoploid (Stoneberg Holt *et al.* 2005).

Completeness of the assembly was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) and the LTR Assembly Index (LAI). The assembly contains 99% of the expected conserved genes (BUSCOs), 98% of which were duplicated reflecting the polyploid nature of the assembly. Additionally, the transposable element assembly is also complete as demonstrated by a LAI value of 25.8 (Ou *et al.* 2018).

Genome annotation

We identified 256,281 gene models, approximately 32K per subgenome assuming octoploidy, using a hybrid gene prediction pipeline that combined *ab initio* gene models with direct evidence annotations. Phylostrata demonstrated approximately 13% of the gene models are species-specific, which is higher than would be expected from orphan genes alone (Arendsee *et al.* 2014). The excess of species-specific genes likely results from a lack of closely related high-quality reference genomes available for comparison in the Phylostrata analysis. This demonstrates the important gap a *Poa* reference genome fills in the green tree of life.

Transposable elements (TEs) were comprehensively annotated using EDTA (Ou *et al.* 2019) and found to compose 58% of the genome. More specifically, Class I LTR retrotransposons and Class II DNA transposons comprise 36% and 15% of the genome, respectively. At the level of superfamily, the RLG (*Ty3*) LTR retrotransposon superfamily was the most common at 18% of the genome.

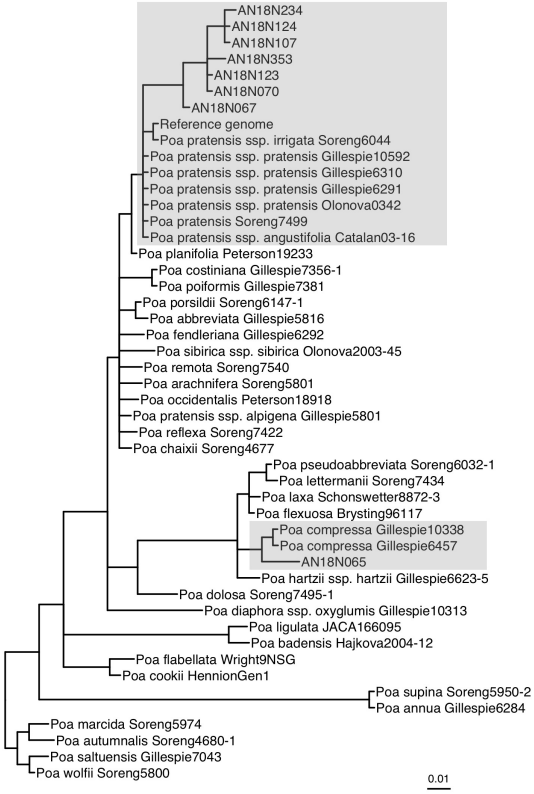

Figure 1 Phylogenetic inferences identifies the reference genome as *Poa pratensis*.
The tree is a subset of the full ITS tree in Figure S1. Reference the full trees in the supplement for clade support values. The unknown *Poa* population samples are labeled with their sample IDs (beginning with 'AN'). The shaded boxes indicate the two clades the reference genome and population panel fall within: *P. pratensis* and *P. compressa*. Branch length is the expected substitutions per site.

Poa population structure

Genotype likelihoods were estimated for 1,722,320 sites. Principal Component Analysis (PCA) and nucleotide diversity (θ_π) were utilized to evaluate population structure and genetic diversity of the *Poa* population panel. In the PCA, most genetic variation was explained by species (40.5%) followed by population (12.7%) (Figure 2A). *P. compressa* is quite distantly related to *P. pratensis* (Figure 1), and the first principal component separates the *P. compressa* genotype from the *P. pratensis* genotypes. The second principal component separates the *P. pratensis* genotypes in the Boulder population from the Canadian *P. pratensis* populations Tolstoi and Argyle (Figure 2A). Genotypes from the Boulder population remain tightly clustered. These results suggest North American *P. pratensis* exhibit population structure and support previous findings of population divergence in Northern Great Plains populations (Dennhardt *et al.* 2016).

To further understand the structure of genetic diversity across

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1 Scaffold assembly statistics

Variable	Description
Scaffolds	118
Estimated genome size	4.90 Gbp
Assembled genome size	6.09 Gbp
N50	65,127,037 bp
L50	31
Longest scaffold	177,118,352 bp
Scaffolds > 1 Mb	110
Scaffolds > 10 Mb	98
Average scaffold length	51,622,171 bp
Average length of gaps	44,233 bp
Complete BUSCOs	99.2%
LAI	25.8

P. pratensis populations, we estimated nucleotide diversity (π) within the Boulder population (nSites = 46,951,318) and across the three populations excluding the *P. compressa* genotype (nSites = 37,263,868). Mean diversity across *P. pratensis* populations is high (π = 0.0098, SD = .0038; Figure 2B), which is consistent with previous studies of *P. pratensis* (Bonos and Huff 2013; Raggi *et al.* 2015; Bushman *et al.* 2013; Honig *et al.* 2018, 2012). The Boulder *P. pratensis* has lower diversity (π = 0.0061, SD = 0.0037) compared to the across-population diversity. This difference in nucleotide diversity further demonstrates population structure exists amongst our samples.

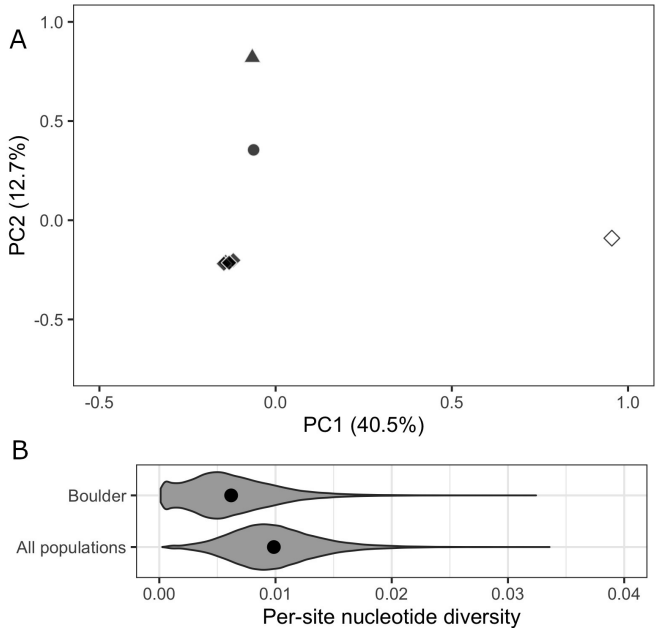

Figure 2 Population structure of *Poa* and nucleotide diversity in *P. pratensis*.
(A) The first two principal components (PCs) of a PCA of the entire *Poa* population panel. The percent of genetic variation explained by each PC is reported in parenthesis on each axis. Samples are indicated by shape (circle = Argyle, triangle = Tolstoi, diamond = Boulder) and species are colors (white = *P. compressa*, black = *P. pratensis*. Points are slightly jittered for visualization (`geom_point(position = position_jitter(w = 0.02, h = 0.02))`). (B) Per-site nucleotide diversity calculated in 10K bp windows for only *P. pratensis* in Boulder and across the three populations. Mean diversity in each group is marked with a black circle.

Conclusions

Poa pratensis is a globally popular turfgrass species used in lawns and recreation areas. Despite its economic value, progression of molecular tools to aid breeding has been slow compared to other turfgrasses as a result of polyploidy and apomixis (Bushman and Warnke 2013). Utilizing long read technology and a Bionano optical map, we have assembled and annotated the first high quality *P. pratensis* reference genome. We demonstrated the utility of the reference by evaluating the genetic diversity and population structure of wild North American *Poa* and provided the first estimate of nucleotide diversity in *P. pratensis*. The reference genome and annotation will serve as an important resource in the study of bluegrasses.

Acknowledgments

This project was supported by the National Science Foundation grant number 1822330. Thank you to Dr. Chrissy McAllister and Bess Bookout for sharing samples collected with permission from Nature Conservancy Canada properties. Additionally, thank you to the City of Boulder Open Space and Mountain Parks for permission to collect on their properties and Lynn Riedel for assisting these collections. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin, HPC@ISU equipment at Iowa State University (partially funded by NSF under MRI grant number 1726447), for providing HPC resources that have contributed to the research results reported within this paper. We thank Dr. Kevin Fengler (for providing assembly instructions) and Dr. Gina Zastrow-Hayes (for establishing sequencing contracts), of Corteva Agriscience for their help in this project. Additionally, the authors would like to thank our *Andropogon gerardi* reference plant for being contaminated with *Poa* and Felix Andrews for his alleged role in the happy accident that led to this work. Finally, thank you to Bob Ross for inspiring a generation of scientists to persevere.

References

Akaike, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**: 716–723.

Albertini, E., G. Marconi, G. Barcaccia, L. Raggi, and M. Falcinelli, 2004 Isolation of candidate genes for apomixis in *Poa pratensis* L. Plant molecular biology **56**: 879–894.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. Journal of molecular biology **215**: 403–410.

Arendsee, Z., J. Li, U. Singh, A. Seetharam, K. Dorman, *et al.*, 2019 phylostratr: A framework for phylostratigraphy. Bioinformatics **35**: 3617–3627.

Arendsee, Z. W., L. Li, and E. S. Wurtele, 2014 Coming of age: orphan genes in plants. Trends in plant science **19**: 698–708.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Bonos, S. A. and D. R. Huff, 2013 Cool-season grasses: Biology and breeding. Turfgrass: Biology, use, and management **56**: 591–660.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, *et al.*, 2013 Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience **2**: 2047–217X.

Brown, W. L., 1939 Chromosome complements of five species of *Poa* with an analysis of variation in *Poa pratensis*. American Journal of Botany **26**: 717–723.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics **3**, lqaa108.

Bushman, B. S. and S. E. Warnke, 2013 Genetic and genomic approaches for improving turfgrass. Turfgrass: biology, use, and management **56**: 683–711.

Bushman, B. S., S. E. Warnke, K. L. Amundsen, K. M. Combs, and P. G. Johnson, 2013 Molecular markers highlight variation within and among Kentucky bluegrass varieties and accessions. Crop Science **53**: 2245–2254.

Cabi, E., R. J. Soreng, and L. Gillespie, 2017 Taxonomy of *Poa jubata* and a new section of the genus (Poaceae). Turkish Journal of Botany **41**: 404–415.

Cabi, E., R. J. Soreng, L. Gillespie, and N. Amiri, 2016 *Poa densa* (Poaceae), an overlooked Turkish steppe grass, and the evolution of bulbs in *Poa*. Willdenowia **46**: 201 – 211.

Carrier, L. and K. S. Bort, 1916 The history of Kentucky bluegrass and white clover in the united states. Agronomy Journal **8**: 256–267.

Casler, M. D. and R. R. Duncan, 2003 Turfgrass biology, genetics, and breeding .

Chin, C.-S., P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, *et al.*, 2016 Phased diploid genome assembly with single-molecule real-time sequencing. Nature methods **13**: 1050–1054.

Cope, T. A. and A. J. Gray, 2009 Grasses of the british isles. Botanical Society of the British Isles.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, *et al.*, 2021 Twelve years of SAMtools and BCFtools. GigaScience **10**, giab008.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada, 2012 jmodeltest 2: more models, new heuristics and parallel computing. Nature methods **9**: 772–772.

DeKeyser, E. S., L. A. Dennhardt, and J. Hendrickson, 2015 Kentucky bluegrass (*Poa pratensis*) invasion in the northern great plains: a story of rapid dominance in an endangered ecosystem. Invasive Plant Science and Management **8**: 255–261.

Dennhardt, L. A., E. S. DeKeyser, S. A. Tennefos, and S. E. Travers, 2016 There is no evidence of geographical patterning among invasive Kentucky bluegrass (*Poa pratensis*) populations in the northern great plains. Weed Science **64**: 409–420.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, 2013 Star: ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21.

Doležel, J., J. Greilhuber, and J. Suda, 2007 Estimation of nuclear DNA content in plants using flow cytometry. Nature protocols **2**: 2233–2244.

Duvaud, S., C. Gabella, F. Lisacek, H. Stockinger, V. Ioannidis, *et al.*, 2021 Expasy, the swiss bioinformatics resource portal, as designed by its users. Nucleic Acids Research **49**: W216–W227.

Ellinghaus, D., S. Kurtz, and U. Willhoeft, 2008 LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC bioinformatics **9**: 1–14.

Funk, C. R., J. H. Sang, *et al.*, 1967 Recurrent intraspecific hybridization – A proposed method of breeding Kentucky bluegrass (*Poa pratensis* l.). New Jersey Agricultural Experiment Station Bulletin .

Gillespie, L. J., A. Archambault, and R. J. Soreng, 2007 Phylogeny of *Poa* (Poaceae) based on trnT–trnF sequence data: major clades and basal relationships. Aliso: A Journal of Systematic and Evolutionary Botany **23**: 420–434.

Gillespie, L. J., R. J. Soreng, R. D. Bull, S. W. Jacobs, and N. F. Refulio-Rodriguez, 2008 Phylogenetic relationships in subtribe Poinae (Poaceae, Poeae) based on nuclear ITS and plastid trnT-trnL-trnF sequences. Botany (Ottawa) **86**: 938–967.

Gillespie, L. J., R. J. Soreng, E. Cabi, and N. Amiri, 2018 Phylogeny and taxonomic synopsis of *Poa* subgenus *Pseudopoa* (including *Eremopoa* and *Lindbergella*)(Poaceae, Poeae, Poinae). PhytoKeys **111**: 69–101.

Gillespie, L. J., R. J. Soreng, and S. W. Jacobs, 2009 Phylogenetic relationships of Australian *Poa* (Poaceae: Poinae), including molecular evidence for two new genera, *Saxipoa* and *Sylvipoa*. Australian Systematic Botany **22**: 413–436.

Giussani, L. M., L. J. Gillespie, M. A. Scataglini, M. A. Negritto, A. M. Anton, *et al.*, 2016 Breeding system diversification and evolution in American *Poa* supersect. *Homalopoa* (Poaceae: Poeae: Poinae). Annals of Botany **118**: 281–303.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nature protocols **8**: 1494–1512.

Haydu, J. J., A. W. Hodges, and C. R. Hall, 2006 Economic impacts of the turfgrass and lawncare industry in the united states. EDIS **2006**.

Heide, O., 1994 Control of flowering and reproduction in temperate grasses. New Phytologist **128**: 347–362.

Hendrickson, J., M. Liebig, J. Printz, D. Toledo, J. Halvorson, *et al.*, 2021 Kentucky bluegrass impacts diversity and carbon and nitrogen dynamics in a Northern Great Plains rangeland. Rangeland Ecology & Management **79**: 36–42.

Honig, J. A., V. Averello, S. A. Bonos, and W. A. Meyer, 2012 Classification of Kentucky bluegrass (*Poa pratensis* L.) cultivars and accessions based on microsatellite (simple sequence repeat) markers. HortScience **47**: 1356–1366.

Honig, J. A., V. Averello, C. Kubik, J. Vaiciunas, B. S. Bushman, *et al.*, 2018 An update on the classification of Kentucky bluegrass cultivars and accessions based on microsatellite (SSR) markers. Crop Science **58**: 1776–1787.

Hufford, M. B., A. S. Seetharam, M. R. Woodhouse, K. M. Chougule, S. Ou, *et al.*, 2021 De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. bioRxiv .

Katoh, K. and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30**: 772–780.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research **27**: 722–736.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of next generation sequencing data. BMC Bioinfor-

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

matics **15**: 356.

Kral-O'Brien, K. C., R. F. Limb, T. J. Hovick, and J. P. Harmon, 2019 Compositional shifts in forb and butterfly communities associated with Kentucky bluegrass invasions. Rangeland Ecology & Management **72**: 301–309.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**: 2987–2993.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 .

Li, J., U. Singh, P. Bhandary, J. Campbell, Z. Arendsee, *et al.*, 2021 Foster thy young: enhanced prediction of orphan genes in assembled genomes. Nucleic Acids Research gkab1238.

Liu, R. and J. Dickerson, 2017 Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-seq. PLoS computational biology **13**: e1005851.

Manchanda, N., J. L. Portwood, M. R. Woodhouse, A. S. Seetharam, C. J. Lawrence-Dill, *et al.*, 2020 GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. BMC genomics **21**: 1–9.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simao, and E. M. Zdobnov, 2021 Busco update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv preprint arXiv:2106.11799 .

Mapleson, D., L. Venturini, G. Kaithakottil, and D. Swarbreck, 2018 Efficient and accurate detection of splice junctions from RNA-seq with portcullis. GigaScience **7**: giy131.

Mascher, M., T. Wicker, J. Jenkins, C. Plott, T. Lux, *et al.*, 2021 Long-read sequence assembly: a technical evaluation in barley. The Plant cell **33**: 1888—1906.

Matzk, F., 1991 New efforts to overcome apomixis in *Poa pratensis* L. Euphytica **55**: 65–72.

Meisner, J. and A. Albrechtsen, 2018 Inferring population structure and admixture proportions in low-depth NGS data. Genetics **210**: 719–731.

Milesi, C., S. W. Running, C. D. Elvidge, J. B. Dietz, B. T. Tuttle, *et al.*, 2005 Mapping and modeling the biogeochemical cycling of turf grasses in the united states. Environmental management **36**: 426–438.

Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang, 2012 SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS ONE **7**: e37558.

Ou, S., J. Chen, and N. Jiang, 2018 Assessing genome assembly quality using the LTR assembly index (LAI). Nucleic acids research **46**: e126–e126.

Ou, S., W. Su, Y. Liao, K. Chougule, J. R. Agda, *et al.*, 2019 Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome biology **20**: 1–18.

Paradis, E. and K. Schliep, 2019 ape. Bioinformatics **35**: 526–528.

Pepin, G. W. and C. R. Funk, 1971 Intraspecific hybridization as a method of breeding Kentucky bluegrass (*Poa pratenis* L.) for turf. Crop science **11**: 445–448.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology **33**: 290–295.

R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raggi, L., E. Bitocchi, L. Russi, G. Marconi, T. F. Sharbel, *et al.*, 2015 Understanding genetic diversity and population structure of a *Poa pratensis* worldwide collection through morphological, nuclear and chloroplast diversity analysis. PLoS One **10**: e0124709.

Refulio-Rodriguez, N. F., J. T. Columbus, L. J. Gillespie, P. M. Peterson, and R. J. Soreng, 2012 Molecular phylogeny of *Dissanthelium* (Poaceae: Pooideae) and its taxonomic implications. Systematic Botany **37**: 122–133.

Ronquist, F., M. Teslenko, P. Van Der Mark, D. L. Ayres, A. Darling, *et al.*, 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology **61**: 539–542.

Song, L., S. Sabunciyan, and L. Florea, 2016 CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. Nucleic acids research **44**: e98–e98.

Soreng, R., L. Gillespie, and L. Consaul, 2017 Taxonomy of the *Poa laxa* group, including two new taxa from Arctic Canada and Greenland, and Oregon, and a re-examination of *P.* sect. *Oreinos* (Poaceae). Nordic Journal of Botany **35**: 513–538.

Soreng, R. J., 2007 *Poa* L. Flora of North America, Poaceae, part 1, vol. 24. .

Soreng, R. J. and F. R. Barrie, 1999 (1391) proposal to conserve the name *Poa pratensis* (Gramineae) with a conserved type. Taxon **48**: 157–159.

Soreng, R. J. and L. J. Gillespie, 2018 *Poa secunda* J. Presl (Poaceae): a modern summary of infraspecific taxonomy, chromosome numbers, related species and infrageneric placement based on DNA. PhytoKeys p. 101.

Soreng, R. J., L. J. Gillespie, H. Koba, E. Boudko, and R. D. Bull, 2015 Molecular and morphological evidence for a new grass genus, *Dupontiopsis* (Poaceae tribe Poeae subtribe Poinae sl), endemic to alpine Japan, and implications for the reticulate origin of *Dupontia* and *Arctophila* within Poinae sl. Journal of Systematics and Evolution **53**: 138–162.

Soreng, R. J., M. V. Olonova, N. S. Probatova, and L. J. Gillespie, 2020 Breeding systems and phylogeny in *Poa*, with special attention to Northeast Asia: The problem of *Poa shumushuensis* and sect. *Nivicolae* (Poaceae). Journal of Systematics and Evolution **58**: 1031–1058.

Stoneberg Holt, S., L. Horová, P. Bureš, J. Janeček, and V. Černoch, 2005 The trnL-F plastid DNA characters of three *Poa pratensis* (Kentucky bluegrass) varieties. Plant, Soil and Environment **51**: 94–99.

Sylvester, S. P., R. J. Soreng, and L. J. Gilliespie, 2021 Resolving páramo *Poa* (Poaceae): morphometric and phylogenetic analysis of the 'Cucullata complex'of north-west South America. Botanical Journal of the Linnean Society **197**: 104–146.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols **7**: 562–578.

Van der Auwera, G. A. and B. D. O'Connor, 2020 *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Incorporated.

Venturini, L., S. Caim, G. G. Kaithakottil, D. L. Mapleson, and D. Swarbreck, 2018 Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience **7**: giy093.

Wang, L.-G., T. T.-Y. Lam, S. Xu, Z. Dai, L. Zhou, *et al.*, 2020 Treeio: an R package for phylogenetic tree input and output

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with richly annotated and associated data. Molecular biology and evolution **37**: 599–603.

Wickham, H., 2016 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yu, G., 2020 Using ggtree to visualize data on tree-like structures. Current protocols in bioinformatics **69**: e96.

Zhang, R.-G., Z.-X. Wang, S. Ou, and G.-Y. Li, 2019 TEsorter: lineage-level classification of transposable elements using conserved protein domains. bioRxiv .

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplement

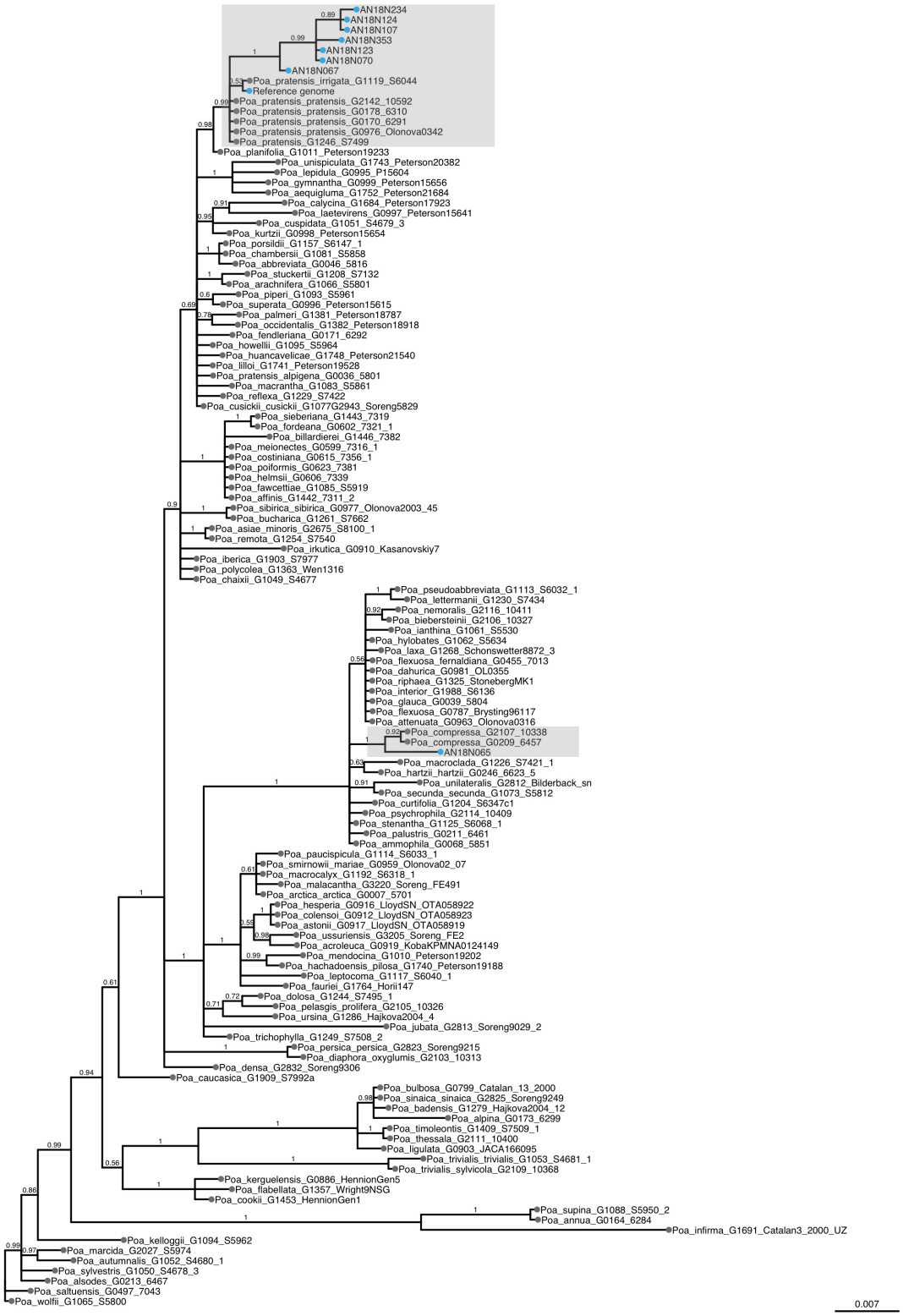

Figure S1 Bayesian 50% majority rule Consensus tree of ITS data. The *Poa* population panel and reference genome are indicated on the tree with blue dots. The unknown *Poa* population samples are labeled with their sample IDs (beginning with 'AN'). The shaded boxes indicate the two clades the reference genome and population panel group within: *P. pratensis* and *P. compressa*. Bayesian posterior probabilities shown above the branches and branch length is the expected substitutions per site.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

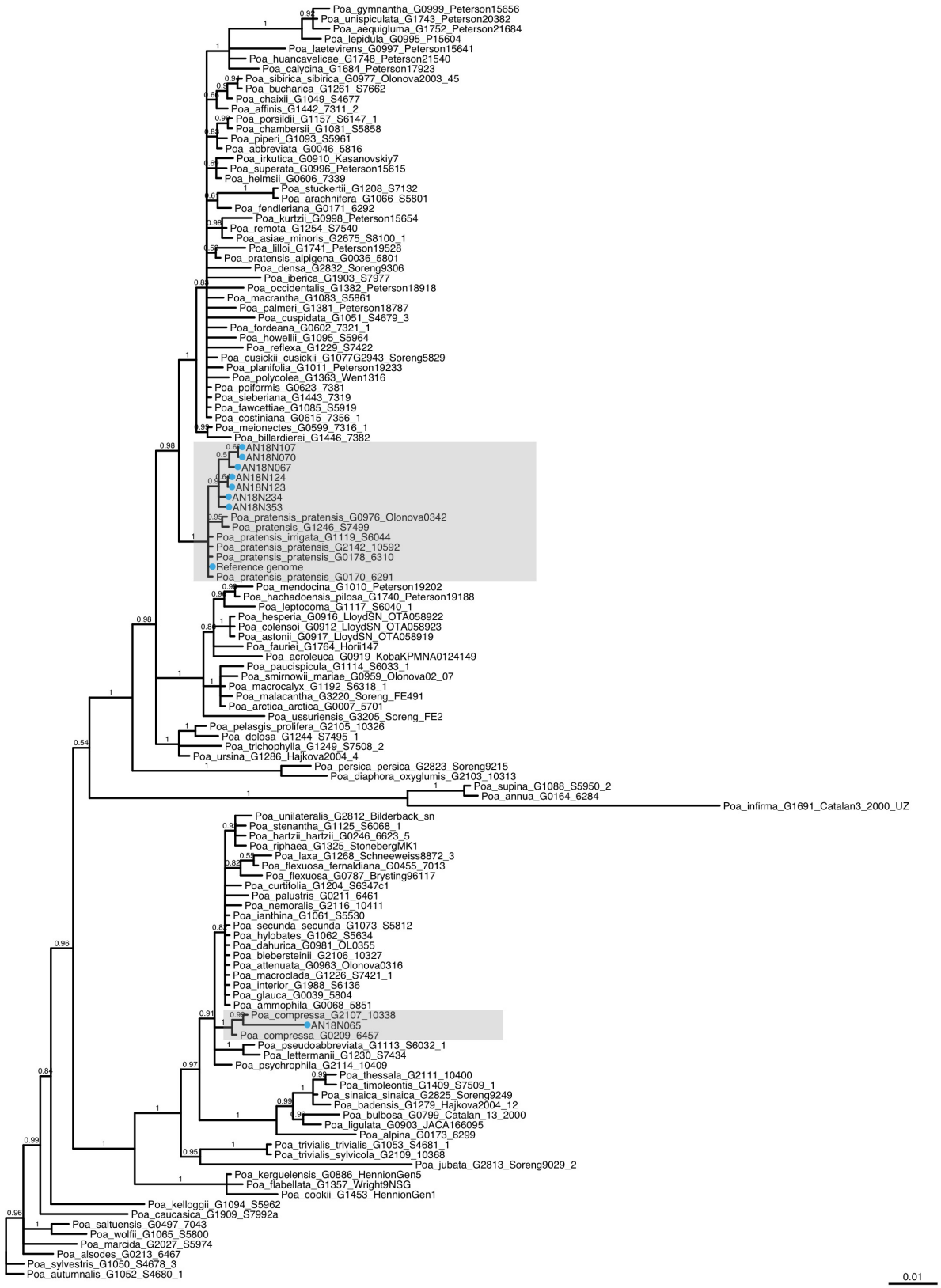

Figure S2 Bayesian 50% majority rule consensus tree of ETS data. See Supplement S1 for description of the figure components.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483531; this version posted March 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

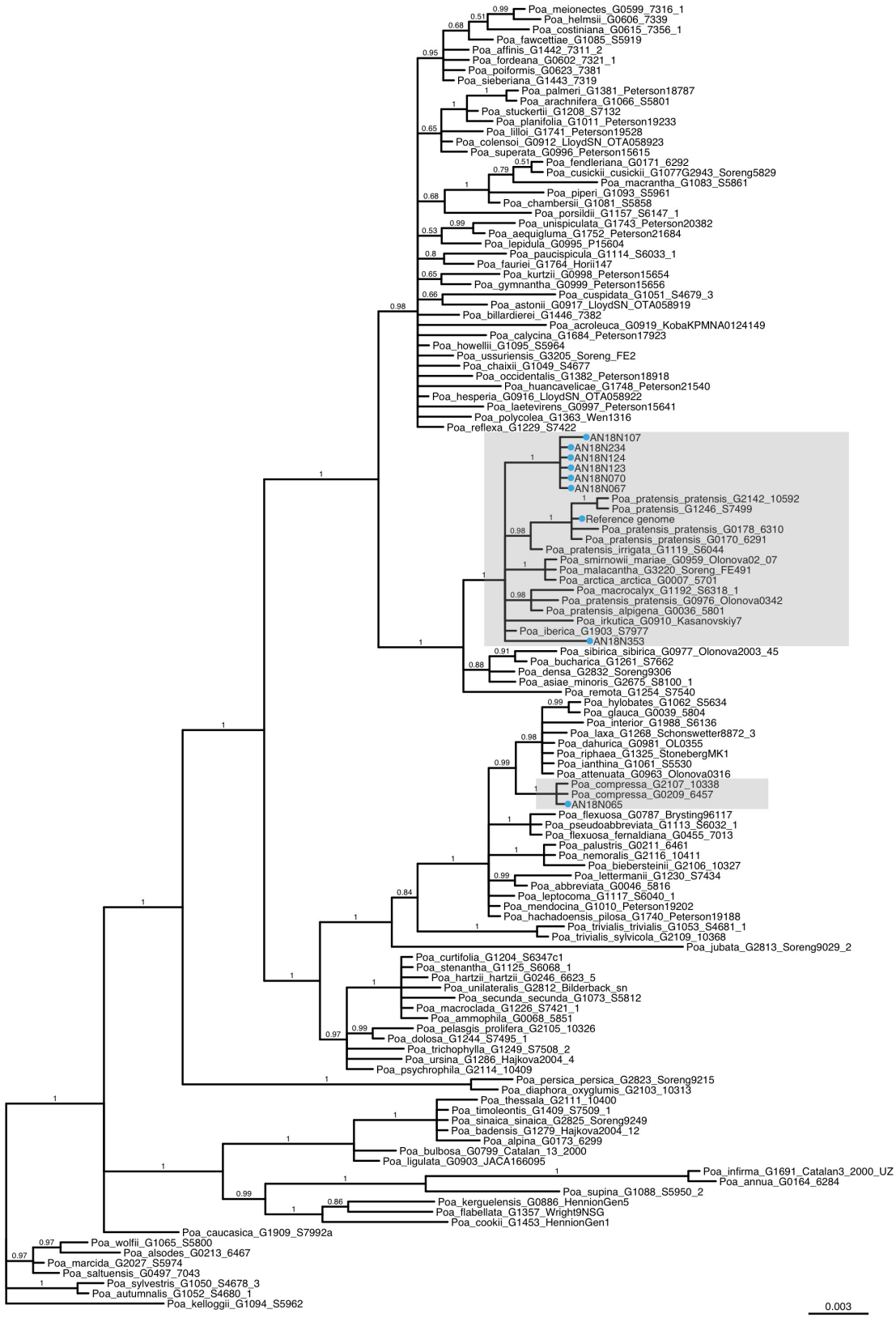

Figure S3 Bayesian 50% majority rule Consensus tree of TLF data. See Supplement S1 for description of the figure components.