Supporting Information

Estep et al. 10.1073/pnas.1404177111

SI Materials and Methods

Taxon Sampling. We sampled 114 accessions representing two outgroup species (*Paspalum malacophyllum* and *Plagiantha tenella*), two species of *Arundinella*, and 100 species of Andropogoneae in 40 genera. Plant material came from our own collections, the US Department of Agriculture (USDA) Germplasm Resources Information Network (GRIN), the Kew Millenium Seed Bank, and material sent by colleagues. All plants acquired as seeds (e.g., those from USDA and from Kew) were grown to flowering in the greenhouse at the University of Missouri-St. Louis to verify identification. Vouchers are listed in Table S1.

Sequencing and Processing. Total genomic DNA was extracted using a modified cetyl triethylammonium bromide (CTAB) procedure (1) or Qiagen DNeasy kits, following the manufacturer's protocol (Qiagen) (2). Five regions of four loci were PCR amplified following Estep et al. (3). The loci are *Aberrant panicle organization1* (*apo1*), *Dwarf*8 (*d8*), two exons of *Erect panicle2* (*ep2*), and *Retarded palea1* (*rep1*). In sorghum (a diploid) *apo1* and *d8* are on chromosomes 1 and 10, respectively. *ep2* and *rep1* are on chromosome 2, 14.5 Mbp apart; based on estimated recombination frequency in euchromatic regions, this distance could be *ca*. 50 cM (4). Thus, the four markers are unlinked. Each marker has homologs on two chromosomes in maize, as expected in a tetraploid. Importantly for this study, we found no evidence that these loci are lost after polyploidization; in known polyploids, paralogues are found consistently as long as sequence depth is sufficient.

PCR products were gel purified, cloned using pGEM-T easy kits, and transformed into JM109 high-efficiency competent cells, following the manufacturers' protocols (Promega) or using a Topo-4 cloning vector and transformed into One Shot Top 10 cells (Invitrogen). At least eight positive clones for each PCR product were sequenced in both directions using universal primers (M13, Sp6, or T7) on an ABI3730 DNA sequencer at the Penn State Huck Institute of the Life Sciences or Beckman Coulter Genomics. Chromatogram files were trimmed of vector and low quality sequences manually or using Geneious Pro-5.5.6 (BioMatters), and reverse and forward sequences for each clone were assembled. Only clones with 80% or more double-stranded sequence were used for downstream analyses. Internal primers were designed as necessary for loci over 1,000 bp long (*D8* and *Ep2* exon7). All good quality contigs for each sample were then aligned using Geneious, and primer sequences were removed. Recombinants were initially identified by eye and then confirmed with networks in SplitsTree (5), and removed from the alignment. Use of computational methods alone missed many recombinant sequences that were readily identifiable by eye. Occasionally, it was difficult to determine which sequences were genomic and which were PCR recombinants; in these cases, we compared the problematic sequences with unambiguous sequences from other species to determine the nonrecombined sequences. Singletons were identified with MEGA (6) and were interpreted as PCR errors. Sequences were translated and aligned using MUSCLE, as implemented in Geneious Pro-5.5.6.

Data Matrix Assembly. The dataset for each locus consisted of numerous redundant clones. To reduce the number of sequences to one per paralogue per locus, preliminary phylogenetic analyses were conducted for each marker in RAxML (7) including all clones for all taxa. Clones that formed a clade in preliminary analyses and that differed by fewer than five nucleotides were inferred to represent a single locus and were combined into a majority-rule consensus sequence. Clones that did not meet these criteria were kept separate through another round of RAxML analyses. We identified clades with a bootstrap value ≥ 50 that comprised the same accession for each locus. Accessions in these clades were reduced to a single majority-rule consensus sequence using the perl script clone_reducer (github.com/mrmckain).

Gene trees were estimated using RAxML v.7.3.0 with the GTR + Γ model and 500 bootstrap replicates for each locus. We used individual gene-tree topologies as a guide to identify and concatenate paralogues from the same genome for each accession. If a polyploid had two paralogues in each gene tree, and one paralogue was always sister to a particular diploid or other polyploid, then we inferred that those paralogues represented the same genome and used them to create a concatenated sequence. If a locus did not have a sequence congruent to the topology, then the locus was marked as missing data. Five datasets were assembled. One included only accessions with full sampling of all loci for all genomes, a second included accessions that had four out of five markers, a third, three out of five, and so forth. Preliminary trees for all datasets were congruent, but those trees in which some genomes of some taxa were represented by only one or two sequences were less well supported. The results presented here are based on the dataset with a minimum of three out of five loci for each taxon, which maximized species inclusion while still providing enough information for robust results. All five datasets are deposited at Dryad (datadryad.org).

Phylogenetic Analysis, Divergence Time, and Diversification Estimates. Concatenated trees were reconstructed using both maximum likelihood (ML) and Bayesian approaches and rooted at *Paspalum*. The ML tree was estimated with RAxML using the GTR + Γ model and 500 bootstrap replicates. The Bayesian tree was estimated using MrBayes v.3.2.1 (8) using a gamma model with six discrete categories. Two independent runs with 50 million generations each were sampled every 1,000 generations. Convergence of the separate runs was verified using AWTY (9).

Divergence times were estimated using BEAST 1.7.5 (10) on the CIPRES Science Gateway (11). The concatenated analysis was run for 100 million generations sampling every 1,000 under the GTR + Γ model with six gamma categories. The tree prior used the birth-death with incomplete sampling model (12), with the starting tree being estimated using unweighted pair group method with arithmetic mean (UPGMA). The site model followed an uncorrelated lognormal relaxed clock (13). The analysis was rooted to *Paspalum*, with the age of the root estimated as a normal distribution describing an age of 25.5 ± 5 million y (14). Convergence statistics were estimated using Tracer v.1.5 (15) after a burn-in of 50,000 sampled generations. Chain convergence was estimated to have been met when the effective sample size was greater than 200 for all statistics. Ultimately, 10,000 trees were used in TreeAnnotator v.1.7.5 to produce the maximum clade credibility tree and to determine the 95% highest posterior density (HPD) for each node. The final tree was drawn using FigTree v.1.4.0 (tree.bio.ed.ac.uk/software/figtree/).

Tests for shifts in the underlying model of diversification were conducted using Bayesian Analysis of Macroevolutionary Mixtures (BAMM) (16); priors were chosen as recommended using the function setBAMMpriors. Analyses were conducted on a tree constructed by pruning the BEAST tree to leave only a single

paralogue (genome) per species, as well as on the unpruned trees. Analyses were run for 10 million and 100 million generations; 10 million provided an adequate effective sample size for both rate shifts and log-likelihoods (>>500), and additional generations did not change the results. Incomplete taxon sampling was accommodated by assigning a backbone sampling fraction of 0.5, corresponding to our sample of 50% of the genera, and assigning specific fractions for the sample for each clade. Preliminary analyses using a backbone sample of 0.08 or using a global estimate of sampling intensity produced largely similar results.

Differences in speciation rate for polyploids versus diploids were estimated using the Binary State Speciation and Extinction (BiSSE) model (17) as implemented in Mesquite (18) and diversitree (19). Species were coded as allopolyploid or diploid, and the characters were mapped on the pruned BEAST tree. Adjustment for incomplete sampling was implemented in diversitree, using a sample of 0.08 and also 0.5 to correspond to the analyses in BAMM. Analyses were run for 100,000 generations. Values from an unconstrained analysis were compared with those of an analysis in which speciation rates were constrained to be equal.

Genome-Size Estimation. Genome size was measured by flow cytometry at the Flow Cytometry and Imaging Core laboratory at the Benaroya Research Institute, Virginia Mason Research Center. Genome sizes were measured a minimum of four times using standard methods, and the values were averaged (20). Picograms of DNA/2C cell were converted to megabasepairs (Mbp)/1C by multiplying the average experimental value by 980 Mbp per 1 picogram of DNA and dividing by two.

Estimating the Number of Allopolyploidization Events. To obtain a minimum estimate of the number of allopolyploidy events, we looked for a clear phylogenetic signal of allopolyploidy, in the form of a multiple-labeled gene tree (Fig. S1). This method provides unambiguous evidence of allopolyploidy, but will overlook genetic and taxonomic autopolyploidy, allopolyploidy between two very similar species or populations of the same species, and allopolyploidy for which our sequencing was not sufficiently deep to retrieve paralogues. Thus, the number of events estimated here is a robust minimum estimate. We also report genome size for many of the sequenced specimens (Figs. S2 and S3). Although genome size gives a rough approximation of ploidy, it cannot distinguish between polyploids and plants whose genomes have expanded via transposon amplification (21). We did not use published data on chromosome numbers because the literature on these numbers is unreliable; many counts lack photodocumentation or voucher specimens and thus cannot be verified as to number or species identity. In addition, many species have several numbers reported; these may represent real variation or errors.

1. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11–15.
2. Teerawatananon A, Jacobs SWL, Hodkinson TR (2011) Phylogenetics of Panicoideae (Poaceae) based on chloroplast and nuclear DNA sequences. *Telopea (Syd)* 13: 115–142.
3. Estep MC, Vela Diaz DM, Zhong J, Kellogg EA (2012) Eleven diverse nuclear-encoded phylogenetic markers for the subfamily Panicoideae (Poaceae). *Am J Bot* 99(11): e443–e446.
4. Mace ES, Jordan DR (2011) Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and of gene-rich regions with significant implications for crop improvement. *Theor Appl Genet* 123(1):169–191.
5. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2):254–267.
6. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.
7. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
8. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.
9. Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): A system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* 24(4):581–583.
10. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29(8):1969–1973.
11. Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES science gateway for inference of large phylogenetic trees. *Proceedings of the 2010 Gateway Computing Environments Workshop (GCE)* (Institute of Electrical and Electronics Engineers, New York), pp 1–8.
12. Stadler T (2009) On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J Theor Biol* 261(1):58–66.
13. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4(5):e88.
14. Vicentini A, Barber JC, Giussani LM, Aliscioni SS, Kellogg EA (2008) Multiple coincident origins of C_4 photosynthesis in the Mid- to Late Miocene. *Glob Change Biol* 14: 2963–2977.
15. Rambaut A, Suchard MA, Xie D, Drummond AJ (2013) Tracer, Version 1.5. Available at tree.bio.ed.ac.uk/software/tracer/.
16. Rabosky DL (2014) Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* 9(2):e89543.
17. Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. *Syst Biol* 56(5):701–710.
18. Maddison WP, Maddison DR (2011) Mesquite: A Modular System for Evolutionary Analysis, Version 2.75. Available at mesquiteproject.org.
19. FitzJohn RG (2012) Diversitree: Comparative phylogenetic analysis of diversification in R. *Methods Ecol. Evol.* 3:1084–1092.
20. Arumuganathan K, Earle ED (1991) Estimation of nuclear DNA amounts of plants by flow cytometry. *Plant Mol Biol Rep* 9:229–241.
21. Kellogg EA, Bennetzen JL (2004) The evolution of nuclear genome structure in seed plants. *Am J Bot* 91(10):1709–1725.

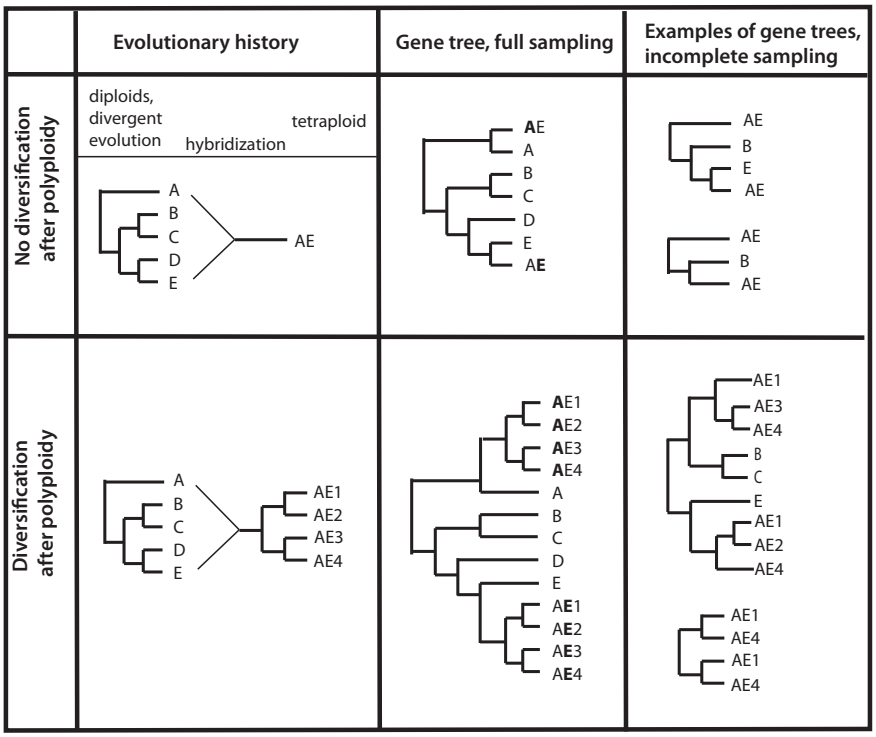

Fig. S1. Gene-tree patterns resulting from hybridization and polyploidy (genetic allopolyploidy). (*Left*) The actual evolutionary history of divergent evolution at the diploid level, followed by hybridization between taxon A and taxon E (reticulation) and formation of a tetraploid. Subsequent speciation may (*Lower*) or may not (*Upper*) occur. (*Center*) The double-labeled gene trees that will be produced by each pattern of evolution. Tetraploids appear twice in the tree; the A genome (**AE**) will appear with its A genome progenitor and the E genome (A**E**) with its E genome progenitor. (*Right*) Some of the many possible gene trees that will result from incomplete taxon sampling. In this study, any of these patterns were taken as indication of allopolyploidy.

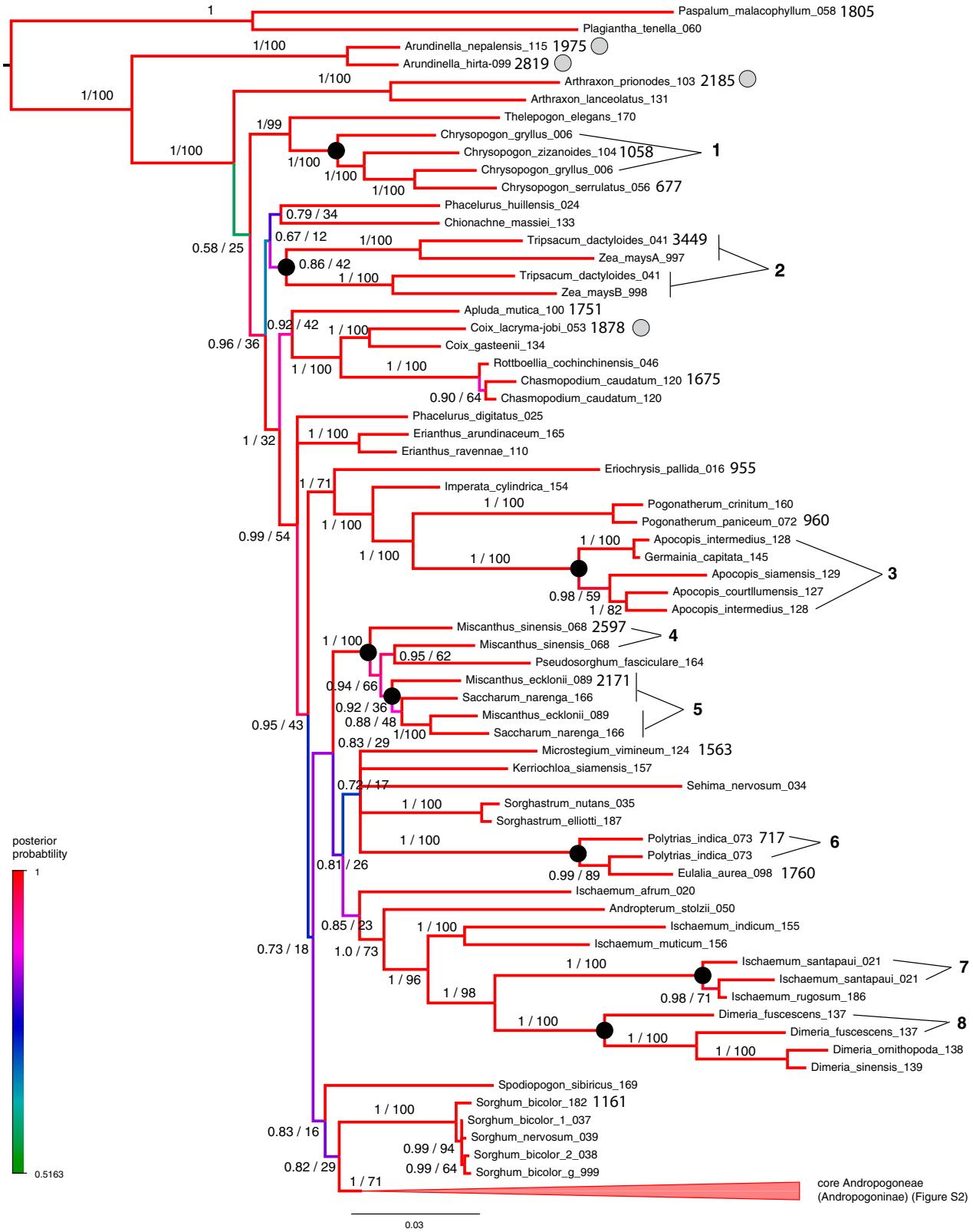

Fig. S2. Bayesian phylogeny of Andropogoneae, species not included in the "core Andropogoneae." Numbers on branches are posterior probability (pp)/ML. Branch color reflects pp values. Genome-size estimates as Mbp/1C nucleus are indicated after species names, where available. Numbered polyploidization events correspond to those in Fig. 1, Figs. S2 and S3, and Table 1. Accession numbers follow the species name for all accessions.

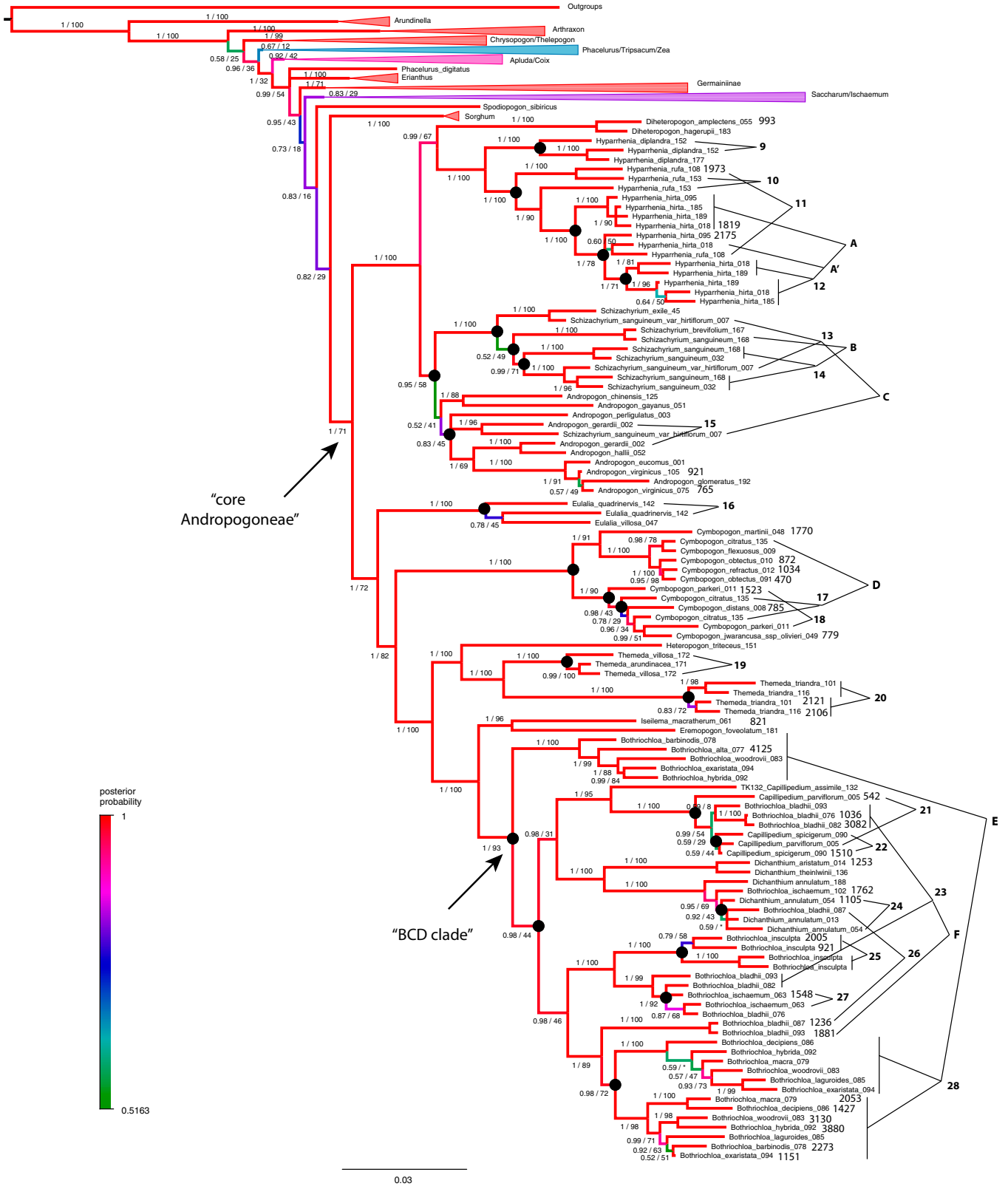

Fig. S3. Bayesian phylogeny of core Andropogoneae. Branch numbers, colors, and genome sizes are as in Fig. S2.

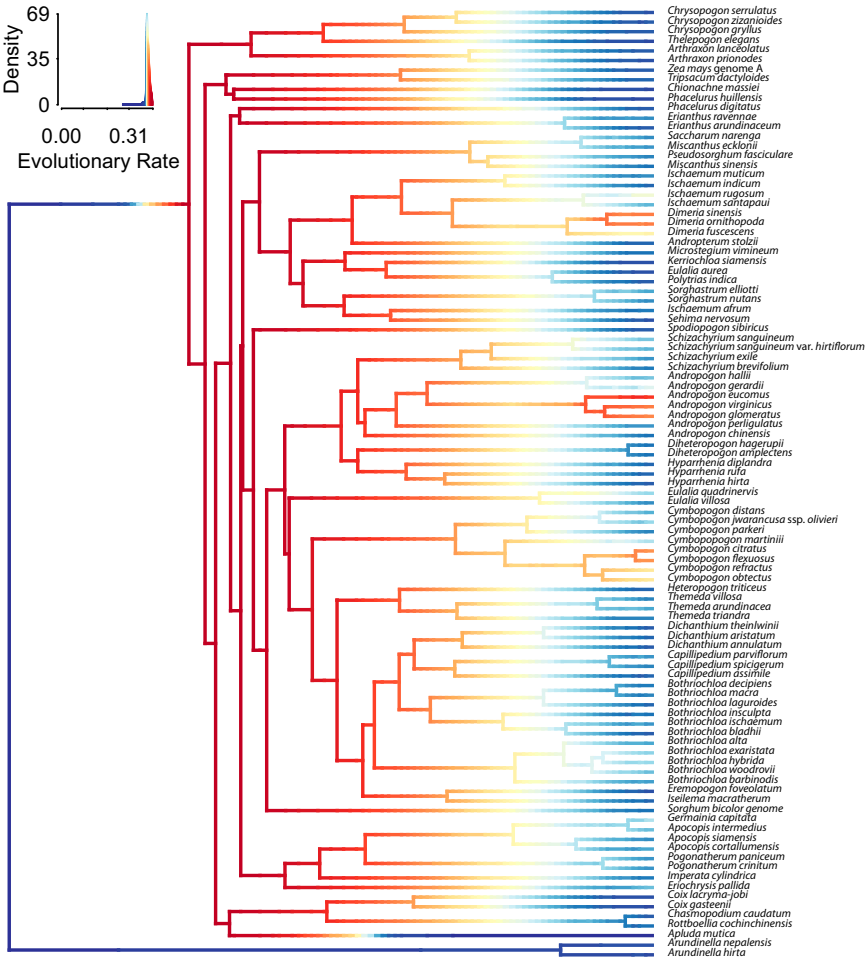

Fig. S4. Phylorate tree from BAMM (16), with species sample frequency adjusted for each clade, and backbone sampling of 0.5. Models of rate shifts with the highest posterior probability indicate either no shifts or one. The marginal probabilities of a shift are similar across much of the tree; in other words, the position of the shift cannot be assigned with certainty.

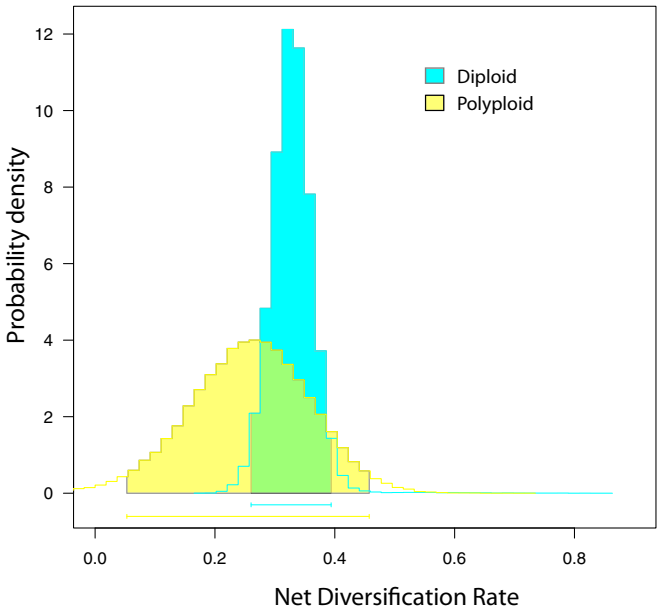

Fig. S5. Net diversification rates (speciation minus extinction) for diploids (blue) and polyploids (yellow), calculated using BiSSE, as implemented in diversitree (19). ML estimates of the rates are significantly different ($P < 0.02$) compared with a model in which speciation rates are constrained to be equal. Analysis assumes a species sample of 8%.

Other Supporting Information Files

Table S1 (DOCX)