PLANT SCIENCES

Gene-rich UV sex chromosomes harbor conserved regulators of sexual development

Sarah B. Carey[1†‡], Jerry Jenkins[2], John T. Lovell[2], Florian Maumus[3], Avinash Sreedasyam[2], Adam C. Payton[1,4], Shengqiang Shu[5], George P. Tiley[6], Noe Fernandez-Pozo[7], Adam Healey[2], Kerrie Barry[5], Cindy Chen[5], Mei Wang[5], Anna Lipzen[5], Chris Daum[5], Christopher A. Saski[8], Jordan C. McBreen[1], Roth E. Conrad[9], Leslie M. Kollar[1], Sanna Olsson[10], Sanna Huttunen[11], Jacob B. Landis[12], J. Gordon Burleigh[1], Norman J. Wickett[13], Matthew G. Johnson[14], Stefan A. Rensing[7,15,16], Jane Grimwood[2,5], Jeremy Schmutz[2,5], Stuart F. McDaniel[1]*

Nonrecombining sex chromosomes, like the mammalian Y, often lose genes and accumulate transposable elements, a process termed degeneration. The correlation between suppressed recombination and degeneration is clear in animal XY systems, but the absence of recombination is confounded with other asymmetries between the X and Y. In contrast, UV sex chromosomes, like those found in bryophytes, experience symmetrical population genetic conditions. Here, we generate nearly gapless female and male chromosome-scale reference genomes of the moss *Ceratodon purpureus* to test for degeneration in the bryophyte UV sex chromosomes. We show that the moss sex chromosomes evolved over 300 million years ago and expanded via two chromosomal fusions. Although the sex chromosomes exhibit weaker purifying selection than autosomes, we find that suppressed recombination alone is insufficient to drive degeneration. Instead, the U and V sex chromosomes harbor thousands of broadly expressed genes, including numerous key regulators of sexual development across land plants.

INTRODUCTION

Sex chromosomes arise when an ordinary pair of autosomes gains the capacity to determine sex (*1*). A defining characteristic of sex chromosomes is suppressed recombination in the heterogametic sex. It is widely believed that this lack of meiotic recombination makes natural selection less effective, predisposing nonrecombining chromosomes, like the mammalian Y, to degeneration and gene loss (*2*, *3*). However, although some nonrecombining chromosomes rapidly degenerate, or are completely lost, the sex chromosomes in other groups remain homomorphic or expand (*2*). This diversity of form and gene content suggests that the role of suppressed recombination in the long-term trajectory of sex chromosome evolution must be modulated by other processes related to the life history of the organism. Identifying these important processes requires comparative analyses across multiple eukaryotic lineages.

Many organisms, including bryophytes, algae, and some fungi, have a haploid UV sex chromosome system, in which females inherit a nonrecombining U and males inherit a nonrecombining V (*4*, *5*). The sex-specific transmission pattern of both chromosomes means that factors that are confounded in XY or ZW systems, such as suppressed recombination, hemizygosity, and sex-limited inheritance, are independent on UV chromosomes (*4*–*6*). Many UV sex chromosome systems may be ancient (*5*), providing ample time for degenerative processes to act. However, the structural complexity of sex chromosomes has precluded genomic analyses in UV systems. Here, we evaluate the relative roles of gene gain and degeneration in shaping the evolution of the bryophyte UV sex chromosomes using nearly-gapless, chromosome-scale female and male genomes of the moss *Ceratodon purpureus*.

RESULTS

Assembly of *C. purpureus* female and male genomes

Ancestral-state reconstructions of dioecy suggest that sex chromosomes evolved early in the history of the extant mosses (*7*). To reconstruct the evolutionary history of the bryophyte UV sex chromosomes, we assembled and annotated chromosome-scale genomes of GG1 (female) and R40 (male) *C. purpureus* isolates. Although the *C. purpureus* genome is relatively small, the sex chromosomes are large and have extensive repeat content, making them a challenge to assemble (*8*), particularly with short-read technologies, which often do not span a whole repeat. We therefore used a combination of Illumina, bacterial artificial chromosomes (BACs), PacBio, and Dovetail Hi-C (figs. S1 and S2 and tables S1 to S4). The version 1.0 genome assembly of R40 comprises 358 Mb in 601 contigs (N50 1.4 Mb), with 98.3% of the assembled sequence in the largest 13 pseudomolecules,

[1]Department of Biology, University of Florida, Gainesville, FL, USA. [2]Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [3]Université Paris-Saclay, INRAE, URGI, 78026 Versailles, France. [4]RAPiD Genomics, Gainesville, FL, USA. [5]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]Department of Biology, Duke University, Durham, NC, USA. [7]Plant Cell Biology, University of Marburg, Marburg, Germany. [8]Department of Plant and Environmental Sciences, Clemson University, Clemson, SC, USA. [9]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [10]Department of Forest Ecology and Genetics, INIA-CIFOR, Madrid, Spain. [11]Department of Biology and Biodiversity Unit, University of Turku, Turku, Finland. [12]L.H. Bailey Hortorium and Section of Plant Biology, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA. [13]Negaunee Institute for Plant Conservation Science and Action, Chicago Botanic Garden, Glencoe, IL, USA. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX, USA. [15]Center for Synthetic Microbiology (SYNMIKRO), University of Marburg, Hans-Meerwein-Straße 6, 35032 Marburg, Germany. [16]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schänzlestraße 18, 79104 Freiburg im Breisgau, Germany.
*Corresponding author. Email: stuartmcdaniel@ufl.edu
†Present address: Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, AL, USA.
‡Prensent address: HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA.

corresponding to the 13 chromosomes in its karyotype (*9*). The version 1.0 GG1 assembly is 349.5 Mb in 558 contigs (N50 1.4 Mb), with 97.9% of assembled sequence in the largest 13 pseudomolecules. Using more than 1.5 billion RNA sequencing (RNA-seq) reads for each of the genome lines (GG1 and R40) and additional de novo assemblies of other *C. purpureus* isolates (table S5), we annotated 31,482 genes on the R40 assembly and 30,425 on GG1 [BUSCO v3.0 of 69% using Embryophyte; 96.7 and 96.4%, respectively using Eukaryote; values similar to the moss *Physcomitrium patens* (*10*)].

Identifying the moss ancestral chromosome elements

To examine the conservation of genome architecture, we performed synteny analyses between the two *C. purpureus* genomes and the *P. patens* genome. GG1 and R40 were collected from distant localities (Gross Gerungs, Austria and Rensselaer, New York, USA, respectively) (*11*), and we found that the assemblies had numerous structural differences (Fig. 1). In the self-synteny analysis, we found clear homeologous chromosome pairs resulting from an ancient whole-genome duplication (WGD) (Fig. 1 and fig. S2), consistent with previous transcriptomic (*11*, *12*) and our own *Ks*-based analyses (fig. S2 and table S6). We also identified abundant synteny between the *C. purpureus* and *P. patens* chromosomes, which diverged over 200 million years (Ma) ago (Fig. 1) (*13*). This result demonstrates that the ancestral karyotype of most extant mosses consisted of seven chromosomes (*13*), which we refer to as ancestral elements A to G (Fig. 1), and suggests that major parts of the gene content of moss chromosomes are stable over hundreds of millions of years, similar to the "Muller Elements" in *Drosophila* (*14*). Curiously, we could not detect the homeologs of the *C. purpureus* chromosomes 5 and 9 using synteny, an observation we return to below.

Suppressed recombination causes weak degeneration but not gene loss

The major exception to the long-term genomic stability observed in *C. purpureus* was the sex chromosomes, which also share no obviously syntenic regions with each other or the autosomes (Fig. 1). The sex chromosomes are ~30% of each genome (110.5 Mb on the R40 V, 112.2 Mb on the GG1 U; Fig. 1), four times the size of the largest autosome. The size is largely attributable to an increase in transposable elements (TEs), which comprise 78.2 and 81.9% of the U and V, respectively, similar to the nonrecombining Y or W sex chromosomes in other systems (*15*), but far more than the *C. purpureus* autosomes [mean (μ), 46.4%; Mann-Whitney *U* with Benjamini and Hochberg correction (MWU), autosomes to U or V $P < 2 \times 10^{-16}$; Fig. 1]. While some TEs have a homogeneous distribution across all chromosomes (e.g., Copia; μ: autosomes = 0.8%, U = 1.3%, V = 1.2%; MWU, all pairwise comparisons $P > 0.09$), the U and V chromosomes are enriched for very different classes of repeats compared to

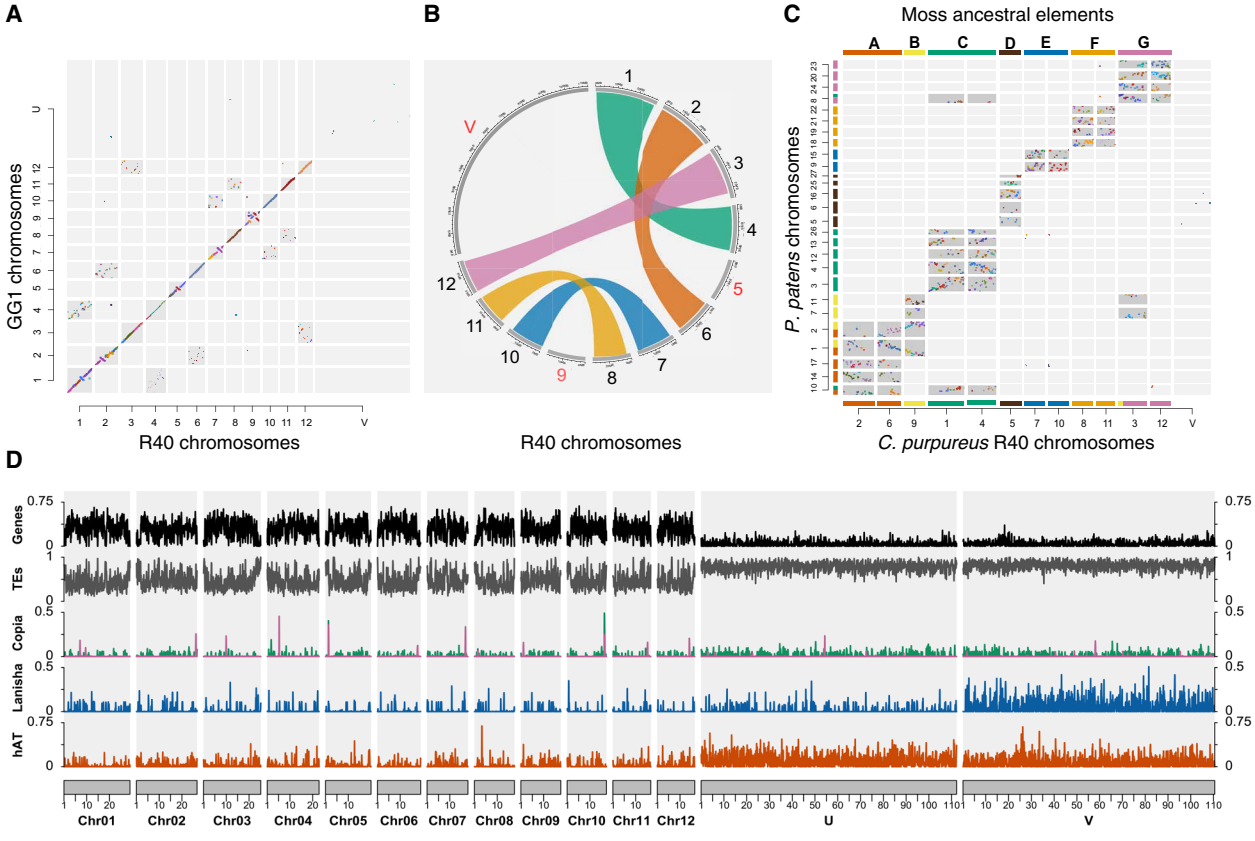

Fig. 1. Chromosome architecture in *C. purpureus*. (**A**) Dot plot of syntenic orthogroup blastp hits between *C. purpureus* GG1 and R40 isolates, showing structural variation on autosomes and a lack of synteny across the sex chromosomes. (**B**) Self-synteny plot of *C. purpureus* R40 isolate showing homeologous chromosomes from a WGD. (**C**) Dot plot of syntenic orthogroup blastp hits between *C. purpureus* R40 and *P. patens*, highlighting the seven ancestral chromosomes that we refer to as the moss ancestral elements A to G. (**D**) Density plots across *C. purpureus* chromosomes (in megabases). Densities show the proportion of a 100-kb window (90-kb jump) of each feature. Local density peaks of RLC5 Copia elements (purple Copia peaks) on each chromosome represent candidate centromeric regions, similar to *P. patens* (*13*).

each other and the autosomes. For example, the U was enriched for hAT (μ: autosomes = 2.3%, U = 10.1%, V = 7.4%; MWU, all pairwise comparisons $P < 1.5 \times 10^{-14}$) and the V was enriched in a previously undescribed superfamily of cut-and-paste DNA transposons, which we refer to as Lanisha elements (μ: autosomes = 1%, U = 1.2%, V = 5.8%; MWU, all pairwise comparisons $P < 1 \times 10^{-4}$; Fig. 1 and fig. S3). The distribution of repeats in *C. purpureus* and the physical proximity of the autosomes inferred from the Hi-C contact map (fig. S2) together highlight the enigmatic isolation of the sex chromosomes in the nucleus (*16*).

Unlike other nonrecombining sex chromosomes, neither the U nor the V shows signs of major degeneration beyond the increased TE density. Sex-linked genes used on average one more codon than autosomes [effective number of codons (ENC)], less frequently use optimal codons [frequency of optimal codon (fop)], have a loss of preferred GC bias in the third synonymous codon position (GC3s), and have a higher rate of protein evolution (dN/dS), all consistent with weaker selection (MWU, autosomes to U or V $P < 6 \times 10^{-6}$ for all metrics; Fig. 2). Although, notably, the U- and V-linked genes were not different from one another (MWU, ENC $P = 0.8$; fop $P = 0.22$; GC3s $P = 0.18$; dN/dS $P = 0.73$), suggesting that transmission through one sex or the other has no detectable effect on purifying selection. Consistent with this observation, the U and the V have 3450 and 3411 transcripts, respectively, representing ~12% of the *C. purpureus* gene content. This stands in stark contrast to the nonrecombining mammalian Y chromosome, or even other UV systems, which typically contain an order of magnitude fewer genes, at most (*17*–*19*). These observations indicate that although suppressed recombination decreases the efficacy of natural selection, alone, it is insufficient to drive gene loss on nonrecombining sex chromosomes (*20*).

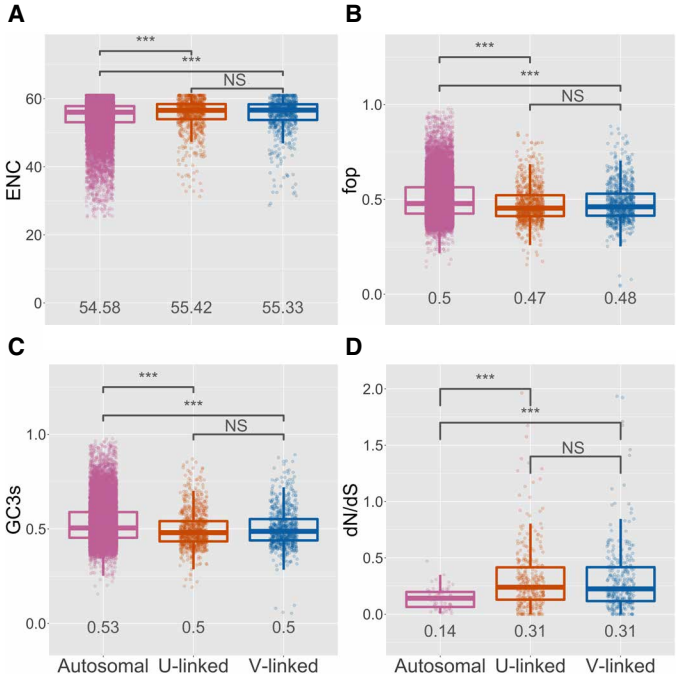

Fig. 2. Molecular evolution of autosomal and sex-linked genes in *C. purpureus*. (**A**) Autosomal genes are significantly different from U- or V-linked genes in the ENC. (**B**) fop. (**C**) GC content of the third, synonymous codon (GC3s) and (**D**) protein evolution (dN/dS) (MWU, autosomes to U or V $P < 6 \times 10^{-6}$ for all metrics, indicated by ***; numbers show means). However, U- and V-linked genes were not significantly different [MWU, ENC $P = 0.8$; fop $P = 0.22$; GC3s $P = 0.18$; dN/dS $P = 0.73$, indicated by NS (not significant)], suggesting weak but not significantly different degeneration on the U and V. For dN/dS, four U-linked genes and two V-linked genes fell above the given scale of the *y* axis.

Moss sex chromosomes are ancient but evolutionarily dynamic

The lack of degeneration means that thousands of genes can be used to reconstruct a detailed history of gene gain on the *C. purpureus* UV sex chromosomes. Critically, the times to the most recent common ancestor between orthologous genes on the U and V chromosomes allow us to estimate a minimum age for the sex chromosome system. In principle, the evolution of sex linkage should mimic the effects of a gene duplication, with identical U- and V-linked clades that coalesce at the node where recombination between them ceased (fig. S4). To identify these nodes, we used a phylogenomic approach with stringent inclusion criteria. We built 744 gene trees, 402 with U- and V-linked homologs. We found that most genes became sex-linked in the *C. purpureus* lineage, after the divergence from *Syntrichia princeps* (μ *Ks* = 0.16; Fig. 3 and table S7). However, 13 U-V orthologous pairs diverged at the base of the Dicranidae (μ *Ks* = 0.85), and three pairs diverged before the split between the two diverse clades Bryidae and Dicranidae (μ *Ks* = 1.64). The most ancient U-V divergence (a Zinc finger Ran binding protein of unknown function) was before the split between *Buxbaumia aphylla* and the remaining Bryopsida, ~300 Ma ago [based on previous fossil-calibrated, relaxed-clock analyses (*21*)] (*Ks* = 2.8; fig. S4).

It is possible that the Zinc finger gene duplicated before the split between *B. aphylla* and the remaining Bryopsida, and these duplicates each were independently captured by the sex-determining locus in these two lineages. However, a more parsimonious explanation is that the origin of the bryophyte sex chromosome system predated the divergence between *B. aphylla* and the remaining Bryopsida. This observation provides support for the maximum parsimony ancestral state reconstruction of sexual system of McDaniel *et al.* (*7*) but pushes back the origins of dioecy in the common ancestor of the Bryidae and Dicranidae to before the origin of the arthrodontous mosses (Fig. 3). These data also provide an independent means to evaluate the inferred transition bias toward dioecy suggested by that analysis.

A classic signature of gene capture on sex chromosomes is the presence of strata, where neighboring genes added in the same recombination suppression event have a similar *Ks* (*22*). However, on the *C. purpureus* sex chromosomes, we found that *Ks* was not associated with gene order (Fig. 3). Even genes with very low *Ks*, presumably from the most recent recombination suppression event, were found across the entirety of the U or V, meaning that gene order was shuffled soon after the evolution of sex linkage. To understand the mechanism by which the region of suppressed recombination acquires new genes, we combined inferences from phylogenomic analyses with the physical position of orthologs among the ancestral karyotypic elements. When we examined gene trees for the two most recent capture events, we found that the overwhelming majority are from ancestral elements D (~80% of *C. purpureus*–specific captures; table S7) and B (~92% of Dicranidae captures; table S7) indicating that the missing homeologous chromosomes to *C. purpureus* 5 and 9, respectively, had fused to the sex chromosomes

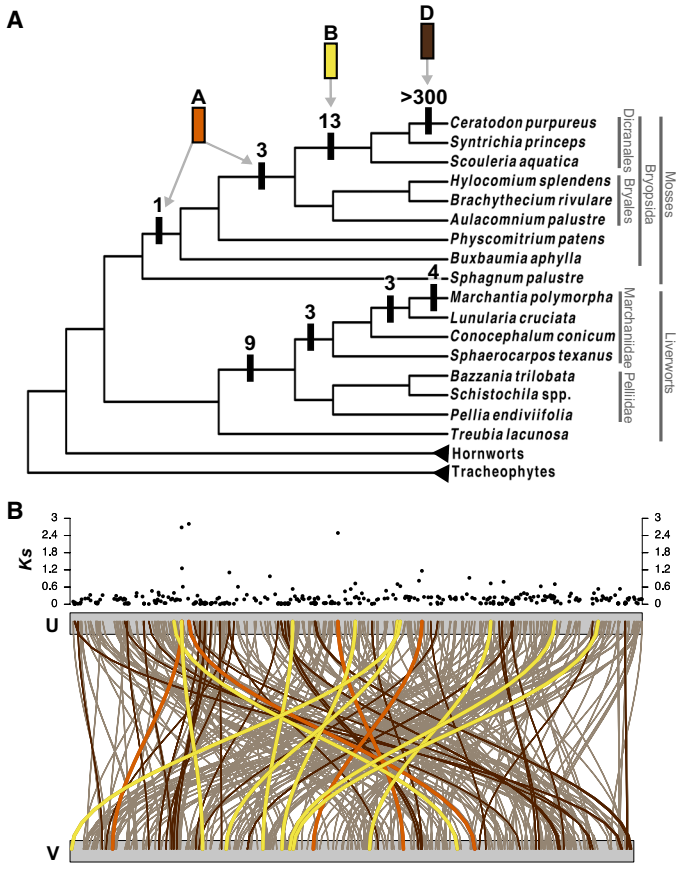

Fig. 3. Evolutionary history of moss and liverwort sex chromosomes. (**A**) Capture events of genes on moss and liverwort sex chromosomes. Numbers indicate how many extant genes were captured at the indicated branch based on the topology of the tree. The capture events in mosses can be traced back to three ancestral elements (A, B, and D), where the oldest sex-linked genes were from ancestral element A and homeologous chromosomes from B and D fused to the sex chromosomes. (**B**) *Ks* of one-to-one U-V orthologs plotted on U and V sex chromosomes of *C. purpureus*. Lines connect the U-V orthologs, where colors correspond to the ancestral elements in (A). The darker brown lines indicate genes with $Ks \leq 0.02$, presumably representing the most recently captured genes, which highlights the rapid rearrangement of genes on the sex chromosomes. These data, in addition to synteny (Fig. 1), also suggest a lack of a pseudo-autosomal region between the *C. purpureus* U and V.

(Fig. 3), but the scrambling of gene order had rendered them undetectable using synteny alone.

The mechanism by which the chromosome fusions occur in UV systems is not entirely clear. In other systems, the sex chromosomes have a pseudo-autosomal region (PAR) that is linked to the nonrecombining portion of the sex chromosome. Unlike the sex-limited region, the PAR pairs normally at meiosis, presumably to assure 1:1 segregation. The addition of new genes to both sex chromosome partners can proceed through the expansion of suppressed recombination into the PAR or the translocation of other genomic regions to the PAR. Curiously, we could find no trace of a PAR in the *C. purpureus* assemblies, nor in any of the unassembled scaffolds (Figs. 1 and 3). We suspect that the PAR reported in *C. purpureus* genetic maps (*7*, *8*) reflects artifactual linkage generated by wide genetic crosses (*8*, *23*). Nevertheless, abundant genotyping of haploid spores from single diploid sporophytes strongly suggests that some PAR-independent mechanism enforces 1:1 segregation of the U and V sex chromosomes (*24*). Thus, while small genomic regions could become independently incorporated into the U or the V, the translocation of whole chromosomes would likely result in homologous pairing between the neo-sex chromosome arms and ultimately the incorporation of the region into both the U and V.

To extend the ancestral reconstruction to liverwort sex chromosomes, we generated gene trees using transcriptome data combined with previously identified sex-linked genes in *Marchantia polymorpha* (*18*). Like in *C. purpureus*, we found no evidence of syntenic strata when we compared *Ks* between the U- and V-linked orthologs (table S8). We also found evidence of four liverwort-specific capture events, with the oldest diverging ~400 Ma ago, before the split of Marchantiidae and Pelliidae (Fig. 3 and fig. S4) (*21*). Our analyses show that most sex-linked genes in *M. polymorpha* (table S8), like two of the oldest genes in *C. purpureus* (table S7), have homologs from moss ancestral element A. This new insight leads to the remarkable suggestion that this element played a key role in sex determination early in the history of both lineages, ~500 Ma ago, making the Setaphyte sex chromosomes among the oldest known to date across Eukarya.

The *C. purpureus* sex chromosomes harbor broadly expressed, conserved regulators of sexual development

A key factor explaining the retention of transcripts on nonrecombining sex chromosomes is broad gene expression (*25*, *26*), which in plants includes the haploid phase. In transcriptomic data from multiple tissues, we found more than 1700 U- and V-linked genes expressed (mean count ≥ 1) (fig. S5 and tables S9 to S11), including essential components of the cytoskeleton (e.g., tubulin) and DNA repair complexes (e.g., *RAD51*). We found that the number of sex-biased autosomal genes (mean count ≥ 1, fold change ≥ 2, adjusted $P \leq 0.05$) was far eclipsed by expressed sex-specific genes (i.e., those only on the U or V), suggesting that sex-linked loci contribute more to expression differences between the sexes than do autosomes (fig. S5). Furthermore, in contrast to data from gene-poor sex chromosome systems, we found that nearly all the genes in the female- and male-specific coexpression modules, including the hubs, were sex-linked (fig. S6 and table S12).

The sex-specific gene expression networks are enriched for proteins with known reproductive functions across green plant lineages. For example, the male Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms show enrichment for microtubule-based processes, which play a role in sperm production in other systems (fig. S6 and tables S13 and S14) (*27*, *28*). We also found that both female and male coexpression modules are enriched for genes involved in circadian rhythm, like phytochrome, which are involved in flower development in *Arabidopsis thaliana* (*29*). The male coexpression module also contained a V-specific *ABC1* gene orthologous to a V-linked copy in *M. polymorpha* (fig. S7), and genes in this family are involved with pollen development in angiosperms (*30*). The female coexpression module contains a U-specific *RWP-RK* transcription factor (TF) orthologous to *M. polymorpha MpRKD*, which is a conserved component of the egg development pathway across land plants and are mating-type loci in green algae (fig. S7) (*17*, *31*, *32*). Moreover, the cis-acting sexual dimorphism switch *MpFGMYB* (*33*), which promotes female development in *M. polymorpha*, as well as the *P. patens* CRINKLY4 (*PpCR4*) gene,

which functions in archegonial neck development (fig. S7) (*34*), have orthologous U- and V-linked copies in *C. purpureus* (fig. S7).

Several other TFs or transcriptional regulators (TRs) are found in the sex-specific coexpression modules (e.g., V-linked *R2R3-MYB*) or are only found on the U or V (e.g., *HD DDT*, *Med7*, and *SOH1*; table S15), together suggesting that candidate regulators of sex-specific developmental processes are enriched on the *C. purpureus* UV sex chromosomes. In addition, 187 U- and/or V-linked genes are homologous to over 250 *A. thaliana* genes with reproductive roles (fig. S5 and table S16). It is, of course, difficult to prove shared gene function across these diverse taxa, particularly for genes found in large gene families, and neo-functionalization can lead to altered functions in specific lineages. Nevertheless, complementing mutants of these genes in hermaphroditic species, like *A. thaliana* or *P. patens*, with sex-linked homologs from *C. purpureus* is likely to provide a powerful means to interrogate the evolution and function of sex-limited gene regulatory networks.

DISCUSSION

Our analyses challenge the idea that suppressed recombination and sex-limited inheritance are sufficient to drive sex chromosome degeneration. Clearly, the lack of meiotic recombination both weakens purifying selection, which results in decreased codon bias and increased protein evolution, and facilitates massive structural variation and highly differentiated TE accumulation between the U and V. Like in other plants, haploid gene expression in *C. purpureus* apparently slows sex chromosome degeneration, even over millions of years of suppressed recombination (*25*). However, unlike flowering plants, where hermaphroditism is the norm (*35*), the antiquity of dioecy in bryophytes more closely mirrors the sexual systems in animals (*36*, *37*). Thus, the gene-rich *C. purpureus* sex chromosomes provide a powerful comparative tool for studying the long-term evolution of sex-limited gene regulatory networks that govern sexual differentiation.

MATERIALS AND METHODS

Isolate collection and tissue culture

All *C. purpureus* tissue used in this study was isolated from a single spore (*24*), from field-collected sporophytes (table S5) (*11*, *38*). In-depth methods for tissue generation for DNA and RNA, library preparation, and sequencing can be found in Supplementary Materials and Methods.

Genome assemblies

We sequenced *C. purpureus* (var. GG1 and var. R40) using a whole-genome shotgun sequencing strategy and standard sequencing protocols. Sequencing reads were collected using Illumina, PacBio, and Sanger platforms. Illumina, PacBio, and Sanger reads were sequenced at the Department of Energy Joint Genome Institute in Walnut Creek, California and the HudsonAlpha Institute in Huntsville, Alabama. Illumina reads were sequenced using the Illumina HiSeq 2000 and X10 platform, and the PacBio reads were sequenced using the SEQUEL I platform. Sanger BACs were sequenced using an ABI 3730XL capillary sequencer. For both GG1 and R40, one 400–base pair (bp) insert 2 × 150 Illumina fragment library (133.14× for GG1, 146.45× for R40) was sequenced along with one 2 × 150 Dovetail Hi-C library (252.86× GG1, 442.71× R40) (table S1). Before assembly, Illumina fragment reads were screened for phix contamination. Reads composed of >95% simple sequence were removed. Illumina reads <50 bp after trimming for adapter and quality (q < 20) were removed. For the PacBio sequencing, a total of eight chemistry 2.1 cells (10-hour movie time) were sequenced each for GG1 and R40 on Sequel 1 with a raw sequence yield of 39.82 Gb (GG1) and 46.24 Gb (R40) with a total coverage of 113.77× (GG1) and 132.11× (R40) (table S2). Last, a total of 1032 BAC clones sequenced with Illumina indexed libraries were used for patching the final chromosome gaps.

Genome assembly and construction of pseudomolecule chromosomes

Improved versions 1.0 of the *C. purpureus* (var. GG1 and var. R40) assemblies were generated by separately assembling the 4,195,510 PacBio GG1 reads (113.77× sequence coverage) and 5,238,148 PacBio reads R40 (132.11× sequence coverage) using the MECAT assembler (*39*) and subsequently polished using QUIVER (*40*). For GG1, this produced 637 scaffolds (637 contigs), with a contig N50 of 1.2 Mb, 475 scaffolds larger than 100 kb, and a total genome size of 347.1 Mb (table S3). For R40, this produced 731 scaffolds (731 contigs), with a contig N50 of 1.1 Mb, 497 scaffolds larger than 100 kb, and a total genome size of 361.3 Mb (table S3).

Hi-C scaffolding using the JUICER pipeline (*41*) was used to identify misjoins in the initial MECAT assembly. Misjoins were characterized as a discontinuity in the GG1 or R40 linkage group. A total of 73 misjoins were identified and resolved in GG1 and 64 in R40. The resulting broken contigs were then oriented, ordered, and joined together into 13 chromosomes (12 autosomal and 1 sex chromosome designated as "U" in the GG1 release and 12 autosomal and 1 sex chromosome designated as "V" in the R40 release) using both the map and the Hi-C data. A total of 579 joins were made in GG1 and 625 in R40 during this process. Each chromosome join is padded with 10,000 Ns. Significant telomeric sequence was identified using the $(TTTAGGG)_n$ repeat, and care was taken to make sure that it was properly oriented in the production assembly. The remaining scaffolds were screened against bacterial proteins, organelle sequences, and GenBank non-redundant database and removed if found to be a contaminant. For GG1, a set of 1032 BAC clones (107.8-Mb total sequence) sequenced with Illumina indexed libraries were used to patch remaining gaps in the chromosomes. Clones were aligned to the chromosomes using BLAT (*42*), and clone contigs crossing gaps were used to form patches. A total of 35 gaps were patched.

Last, homozygous single-nucleotide polymorphisms (SNPs) and insertions or deletions (INDELs) were corrected in the release consensus sequence using ~88× of Illumina reads (2 × 150, 400-bp insert) by aligning the reads using bwa mem (*43*) and identifying homozygous SNPs and INDELs with the Genome Analysis Toolkit UnifiedGenotyper tool (*44*). A total of 108 homozygous SNPs and 5291 homozygous INDELs in GG1 and 19 homozygous SNPs and 867 homozygous INDELs in R40 were corrected in the release. The final version 1.0 GG1 release contains 349.5 Mb of sequence (1.3% gap), consisting of 558 contigs with a contig N50 of 1.4 Mb and a total of 97.9% of assembled bases in chromosomes. The final version 1.0 R40 release contains 358.0 Mb of sequence (1.2% gap), consisting of 601 contigs with a contig N50 of 1.4 Mb and a total of 98.3% of assembled bases in chromosomes.

Completeness of the euchromatic portion of the version 1.0 GG1 and 1.0 R40 assemblies was assessed by aligning an RNA-seq library

(library code GNGZB for GG1 and GNGZC for R40). The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The transcripts were aligned to the assembly using GSNAP (*45*). The alignments indicate that 96.88% of the GG1 RNA-seq reads aligned to the version 1.0 GG1 release and 97.01% of the R40 RNA-seq reads aligned to the version 1.0 R40 release.

Construction of the scaffold assembly

A total of 4,195,510 PacBio reads (113.77×) in GG1 and 5,238,148 PacBio reads (132.11×) in R40 were assembled using MECAT (*39*) and formed the starting point of the version 1.0 release for each. The 310,662,272 Illumina sequence reads (133.14× sequence coverage) in GG1 and 353,932,084 Illumina sequence reads (146.45× sequence coverage) in R40 were used for fixing homozygous SNP/INDEL errors in the consensus. A total of 310,662,272 Hi-C reads (252.86× sequence coverage) in GG1 and 1,062,837,932 Hi-C reads (442.71× sequence coverage) in R40 were used for chromosome construction.

Screening and final assembly release

Scaffolds that were not anchored in a chromosome were classified into bins depending on sequence content. Contamination was identified using blastn against the National Center for Biotechnology Information (NCBI) nucleotide collection (NR/NT) and blastx using a set of known microbial proteins. In GG1, additional scaffolds were classified as repetitive (>95% masked with 24-mers that occur more than four times in the genome) (16 scaffolds, 482.8 kb), chloroplast (1 scaffold, 158.7 kb), and low quality (>50% unpolished bases after polishing, 3 scaffolds, 48.3 kb). In R40, additional scaffolds were classified as repetitive (>95% masked with 24-mers that occur more than four times in the genome) (12 scaffolds, 489.6 kb), chloroplast (1 scaffold, 50.2 kb), and low quality (>50% unpolished bases after polishing, 6 scaffolds, 236.8 kb). Resulting final statistics are shown in table S4.

GG1 assessment of assembly accuracy

A set of 17 finished contiguous Sanger BAC clones >100 kb were selected to assess the accuracy of the assembly. A range of variants were detected in the comparison of the BAC clones and the assembly. In 14 of the BAC clones, the alignments were of high quality (<0.05% base pair error) with an example being given in fig. S1. All dot plots were generated using Gepard (*46*). The remaining three BACs indicate a higher error rate due mainly to their placement in more regions containing tandem repeats (fig. S1). The overall base pair error rate in the BAC clones is 0.016% (269 discrepant bp of 1,599,605 bp).

Genome annotations

Transcript assemblies were made from ~1.5 billion pairs of 2 × 150 stranded paired-end Illumina RNA-seq reads from *C. purpureus* GG1 and ~1.6 billion pairs from *C. purpureus* R40 using PERTRAN, which conducts genome-guided transcriptome short-read assembly via GSNAP (*47*) and builds splice alignment graphs after alignment validation, realignment, and correction (*48*), PERTRAN assemblies from G100m_X_G150f_Sporo reads on the *C. purpureus* GG1 or R40 genome, and filtered open reading frames (ORFs) from Trinity assemblies from stranded paired-end Illumina reads from additional *C. purpureus* cultivars (is Navarino, Magellanes, Chile; Durham, NC, USA; Otavalo, Ecuador; Renesselaer, NY, USA; and Storrs, CT, USA; table S5). 180,954 (GG1) and 194,414 (R40) transcript assemblies were constructed using the Program to Assemble Spliced Alignments (PASA) (*49*) from RNA-seq transcript assemblies above and a bit of *C. purpureus* expressed sequence tags (ESTs). Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from *A. thaliana* (*50*), soybean (*51*), *Setaria viridis* (*52*), grape (*53*), *Sphagnum magellanicum*, and *P. patens* (*13*), *Selaginella moellendorffii* (*54*), and *Chlamydomonas reinhardtii* (*55*), filtered Trinity assembly ORFs described above, high-confidence gene models from the first round of *C. purpureus* R40 gene call, UniProt Bryopsida, and Swiss-Prot proteomes to repeat-soft-masked *C. purpureus* GG1 genome using RepeatMasker (*56*) with up to 2000-bp extension on both ends unless extending into another locus on the same strand. Repeat library consists of de novo repeats by RepeatModeler (*57*) on *C. purpureus* GG1 genome and repeats in RepBase. Gene models were predicted by homology-based predictors, FGENESH+ (*58*), FGENESH_EST (similar to FGENESH+, EST as splice site and intron input instead of protein/translated ORF), and EXONERATE (*59*) and PASA assembly ORFs (in-house homology constrained ORF finder) and from AUGUSTUS via BRAKER1 (*60*). The best-scored predictions for each locus are selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding untranslated regions, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subject to protein homology analysis to abovementioned proteomes to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual best hit) BLASTP score, and protein coverage is the highest percentage of protein aligned to the best of homologs. PASA-improved transcripts were selected on the basis of Cscore, protein coverage, EST coverage, and its coding sequence (CDS) overlapping with repeats. The transcripts were selected if its Cscore is larger than or equal to 0.5 and protein coverage is larger than or equal to 0.5, or it has EST coverage, but its CDS overlapping with repeats is less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cscore must be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis, and gene models whose protein is more than 30% in Pfam TE domains were removed and considered weak gene models. Incomplete gene models, low homology supported without fully transcriptome-supported gene models of short single exons (<300-bp CDS) with neither protein domain nor good expression gene models were manually filtered out.

Synteny analysis within *C. purpureus* and between *P. patens*

We ran the default GENESPACE pipeline (*48*) with a minimum block size of 5 genes and a maximum gap/search radius of 15 genes. In short, GENESPACE runs Orthofinder (*61*, *62*) on synteny-constrained blastp hits. This offers higher stringency when exploring highly diverged genomes (or ancient WGDs) by removing high-scoring, but randomly distributed, blast hits.

Ks plot analysis to identify the *C. purpureus* WGD

WGDs were detected with conventional *Ks* plot analyses. We used the wgd pipeline (*63*). An all-by-all BLASTP search (*64*) was performed for the *C. purpureus* GG1 and R40 genomes as well as *P. patens* and *M. polymorpha*. Paralogs were clustered with MCL (*65*). For each cluster, all pairwise *Ks* estimates were obtained from PAM (*66*) with the GY94 model with F3x4 equilibrium codon frequencies (*67*). Hierarchical clustering was used to reduce redundant comparisons and obtain node-averaged *Ks* estimates. This

process was repeated for syntenic paralogs too, which were obtained from I-ADHoRe v3.0 with default settings (*68*) based on all-by-all BLASTP results. Orthologous gene divergences used reciprocal best BLASTP hits between *C. purpureus* and *P. patens*.

Peaks in *Ks* plots can be identified visually, but we also applied mixture models that were selected by the difference in BIC scores, such that a difference less than 3.2 is used as a stopping criterion. Mixture models were implemented with the *bic.test.wgd* function available on GitHub (https://github.com/gtiley/Ks_plots). Mixture models can be problematic in their interpretation because of overfitting; therefore, we looked for peaks that were consistently detected across models and the maximum *Ks* value allowed (*69*). When analyzing all paralogs, a single prominent peak was observable in *C. purpureus* with a mean between a *Ks* of 0.65 and 0.97 in GG1 and a *Ks* between 0.68 and 0.74 in R40 (table S6). The more consistent results in R40 imply that more paralogs from this WGD event have survived on the V chromosome compared to the U chromosome. This WGD postdates the divergence of *C. purpureus* and *P. patens* (fig. S2). This is determined by visual inspection but agrees with previous analyses of WGD in both *C. purpureus* and *P. patens* (*11*–*13*). The presence of a single WGD that occurred in *C. purpureus* following divergence from *P. patens* is supported by analyses of syntenic paralogs as well (fig. S2), which suggests slightly more recent WGD ages (table S6). However, analyses of syntenic paralogs from *P. patens* supported the presence of two WGDs following divergence from *C. purpureus* (table S6), similar to previous findings when using syntenic data (*13*) compared to all paralogs from genomic or transcriptomic data (*12*, *70*).

Ks plot analyses are provocative of older WGD events that predate the divergence of *C. purpureus* and *P. patens*. Notably, low numbers of syntenic paralogs are evident between *Ks* of 3.0 and 4.0; although, the same is true for *M. polymorpha* that putatively has no history of ancient WGD. Any identifiable peaks in *Ks* plot analyses are too speculative given the lack of evidence from mixture models and nor does their existence affect our proposed model of karyotype evolution. It should be noted though that analyses of gene trees that reconcile duplication and loss events onto a species tree have implied a shared large-scale duplication event shared by *C. purpureus* and *P. patens* ["B3" (*12*)] and an even older event shared by all mosses ["B2" (*12*)]. Testing these ancient hypotheses is beyond the scope of *Ks* plot analyses, even with syntenic data. Rather, macrosyntenic evidence from more moss species, such as *Sphagnum fallax*, will be needed to identify the presence of expected syntenic ratios among genes, similar to the identifiable 1:4 ratios between *C. purpureus* and *P. patens* investigated here.

TE annotation

We combined the R40 assembly (autosomes and V) with the U sex chromosome assembled from GG1 to run de novo repeat detection using the TEdenovo pipeline from the REPET package (v2.4) (*71*). Parameters were set to consider repeats with at least five copies. We obtained a library of 4699 consensus sequences that was filtered to keep only those that are found at least once as a full-length copy in the combined assembly, and we retained 2523 of them. This library of consensus sequences was then used as digital probe for whole-genome annotation by the TEannot (*72*) pipeline from the REPET package v2.4. Threshold annotation scores were determined for each consensus as the 99th percentile of the scores obtained against a randomized sequence [reversed input, not complemented and masked using Tandem Repeats Finder with parameters 2 7 7 80 10 70 10 (*73*)]. The library of consensus sequences was classified using PASTEC followed by manual curation (*74*). To improve classification, remote homology detection was performed using HH-suite3 (*75*). For the density plot of genes and TEs (Fig. 1), we calculated the proportion of coverage of each feature in a 100-kb window with a 90-kb jump using Bedtools (v2.27.) (*76*). These results were plotted in R (v3.5.3) (*77*) using the package karyoploteR (v1.8.8) (*78*) (ceratodon_genome_plots.R, https://doi.org/10.5061/dryad.v41ns1rsm). To examine differences in enrichment between the autosomes, U, and V, we ran a pairwise MWU for multiple tests (*79*, *80*) using the sliding window densities (n_{Auto} = 2736, n_U = 1247, n_V = 1229).

TF and regulator annotation

Transcription-associated proteins (TAPs) comprise TFs (acting in a sequence-specific manner, typically by binding to cis-regulatory elements) and TRs (acting on chromatin or via protein-protein interaction). We classified all *C. purpureus* proteins into 122 families and subfamilies of TAPs by a domain-based rule set (*81*, *82*). We compared this genome-wide classification with relevant organisms. All proteins in which a domain was found are listed with their family assignment. In cases when the domain composition does not allow an unambiguous assignment, they are assigned no_family_found.

Gene expression and coexpression

Gene expression and coexpression analyses were done using three male-female sibling pairs ($n_{isolates}$ = 6, three of each sex) at gametophore and protonemal stages (n_{stages} = 2) in triplicate ($n_{replicates}$ = 3) (table S5; see Supplementary Materials and Methods for details on tissue conditions). Raw reads were filtered for contaminants and adapters removed using BBDuk (v38.00) (Bushnell, http:\\bbtools.jgi.doe.gov). This included removing reads with 93% identity to human, mouse, dog, or cat or aligning to common microbial references. Further filtering removed reads with any "N's," an average quality of 10, or a length <50 or 33% of the full read length. Adapters were trimmed and reads were right quality–trimmed if quality was below 6. Paired-end reads were split into forward and reverse reads (novaseq_FASTQ_de_interlacer.pl, https://doi.org/10.5061/dryad.v41ns1rsm). Reads were further filtered for quality using Trimmomatic (v0.36) (*83*) using leading and trailing values of 3, a window size of 10, a quality score of 30, and a minimum length of 40. We assessed the quality of the remaining reads using fastqc (v0.11.4) Andrews (2010), https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Filtered reads were mapped using HISAT2 (v2.1.0) (*84*) to the *C. purpureus* R40 genome (autosomes and V sex chromosome) concatenated with the GG1 U sex chromosome. We hard masked the U chromosome for males, and the V for females, using Bedtools (v2.27.1) (*76*) maskfasta (*85*). Genes greater than 300 bp were assembled using StringTie (v1.3.3) (*86*), gene counts were extracted using StringTie's prepDE.py script (http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual#deseq), and gene IDs renamed (using mstrg_prep.pl, https://gist.github.com/gpertea/b83f1b32435e166afa92a2d388527f4b). Only genes matching the original genome annotation file were used for coexpression analyses below.

To identify differentially expressed genes, we used DESeq2 (v1.22.2) (*87*), where we contrasted males and females at both the protonemal and gametophore stages. For autosomal genes, we removed those with baseMean < 1, a $\log_2$ fold change < 2, and an adjusted $P > 0.05$. For sex-linked genes, we calculated the mean normalized

count across only females or males for protonema and gametophore stages separately. To identify which sex-linked genes were sex specific versus homologous, we used the output from Orthofinder below. Heatmaps of gene expression were made using variance stabilized counts using DESeq2 (*88*). We converted the transformed counts to long format using reshape2 (v1.4.3) (*89*), clustered the samples using hierarchical clustering, and generated a dendrogram using ggdendro (v0.1.22) (*90*). The plots were modified using gtable (v0.3.0) (*91*). The final heatmaps were generated with ggplot2 (v3.3.1) (*92*) and gridExtra (v2.3) (*93*) using the color palate viridis (v0.5.1) (ceratodon_genome_plots.R, https://doi.org/10.5061/dryad.v41ns1rsm) (*94*).

Coexpression network construction and module detection

Weighted gene coexpression networks were constructed using the WGCNA R package (v1.69) (*95*) with gene expression data normalized using variance stabilizing transformation from the DESeq2 R package (v1.26.0) (*88*). The data retained after filtering genes showing low expression levels (minimum read count = 6 and minimum total read count = 10) were used to construct coexpression network modules using the blockwise network construction procedures. Briefly, pairwise Pearson correlations between each gene pair were weighted by raising them to power (β). To select a proper soft-thresholding power, the network topology for a range of powers was evaluated and appropriate power was chosen that ensured an approximate scale-free topology of the resulting network. The pairwise weighted matrix was transformed into topological overlap measure (TOM). In addition, the TOM-based dissimilarity measure (1 − TOM) was used for hierarchical clustering, and initial module assignments were determined using a dynamic tree-cutting algorithm. Pearson correlations between each gene and each module eigengene, referred to as a gene's module membership, were calculated, and a module eigengene distance threshold of 0.25 was used to merge highly similar modules. Top 10 hub genes in each module were identified on the basis of module membership. These coexpression modules were assessed to determine their correlation with expression patterns distinct to conditions. Interesting modules having significant relationships with conditions, such as sex, were visualized using the igraph (v1.2.5) (*96*) and ggnetwork (v0.5.8) (*97*) R packages and to focus on the relevant gene pair relationships; network depictions were limited to an adjacency threshold of 0.2 and the top 3000 edges/interactions between nodes/gene models.

GO and KEGG pathway enrichment analysis

GO enrichment analysis was carried out using topGO, an R Bioconductor package (v2.38.1) (*98*) with Fisher's exact test; only GO terms with $P < 0.05$ were considered significant. To identify redundant GO terms, semantic similarity among GO terms was measured using Wang's method implemented in the GOSemSim, an R package (v2.12.1) (*99*). KEGG (*100*) pathway enrichment analysis was performed on the basis of hypergeometric distribution test, and pathways with $P < 0.05$ were considered enriched.

Phylogenomic analyses of moss and liverwort sex chromosomes

The genome and transcriptome lines used for phylogenomic analyses can be found in table S17 (*12*, *13*, *18*, *54*, *101*–*104*). For all RNA-seq data, we filtered for quality using Trimmomatic (v0.36) (*83*) using leading and trailing values of 3, a window size of 10, a quality score of 30, and a minimum length of 40. We assessed the quality of the remaining reads using fastqc (v0.11.4) (Andrews, 2010). To de novo assemble genes, we used Trinity (vr20170205-2.4.0) (*105*) following default parameters (the exception being with *C. purpureus*, for which used –SS_lib_type RF). We next determined the single best ORF using TransDecoder (v5.0.2) (*106*). Our reading frames were checked first against pFam (v32.0) (*107*), and if no hit was found, the frame was determined by Transdecoder. To reduce protein redundancy, we next ran our ORFs through CD-HIT (4.6.3) (*108*, *109*) using a 0.99 threshold.

We first found orthogroups for the in-frame genes using Orthofinder (v2.2.0) (*61*, *62*). We built trees for genes annotated on the *M. polymorpha* and *C. purpureus* sex chromosomes by first filtering clusters for at least eight species present in the tree (orthogroup_filter.pl, https://doi.org/10.5061/dryad.v41ns1rsm). For these clusters, we wrote FASTA files for both amino acid and cds files of genes clustered within an orthogroup (fasta_from_OrthoFinder.pl, https://doi.org/10.5061/dryad.v41ns1rsm). We next aligned our amino acid fasta files using MAFFT (v7.407) (*110*). We back-translated our alignments to DNA using pal2nal (v14) (*111*). Alignments were filtered for column occupancy of 0.5 using trimal (v1.2) (*112*) and filtered to remove any sequences less than 300 bp (alignment_length_filter.pl, https://doi.org/10.5061/dryad.v41ns1rsm). These final alignments were used to build bootstrapped trees using RAxML (v8.2.8) (*113*) using the GTRGAMMA model and 100 bootstrap replicates. We visually analyzed trees to determine when genes became sex-linked. To accomplish this, we identified the clades that contained annotated U- and V-linked genes and determined the most distantly related species found in the same clade (e.g., fig. S4). All trees and alignments can be found on Dryad under https://doi.org/10.5061/dryad.v41ns1rsm. All tree plots were made using ggtree (v1.14.6) (*114*, *115*) in R (v3.5.3) (*77*) and edited in Inkscape (v0.92.2) (https://inkscape.org/en/) (ceratodon_genome_plots.R, https://doi.org/10.5061/dryad.v41ns1rsm).

To identify the ancestral element from which sex-linked genes descended, trees with one-to-one U-V orthologs were rooted using *Azolla*, *Salvinia*, *Selaginella*, *Takakia*, or *Sphagnum* as an outgroup (in this order of preference) using newick utils (v1.6) (*116*), and only the longest isoform within a clade for the same sample was retained (edlwtre2.pl, https://doi.org/10.5061/dryad.v41ns1rsm). To determine the closest *P. patens* gene, we used an in-house script (physco_outgroup.py, https://doi.org/10.5061/dryad.v41ns1rsm), which used ETE3 (*117*) to first identify the sex-linked genes and then find the closest *P. patens* gene based on branch length. For these genes, we also determined whether a *C. purpureus* chromosome 5 paralog was present and, of these reported, only those that clearly showed gene duplication, presumably from the WGD event.

To identify homologs between the *C. purpureus* sex chromosomes and *A. thaliana*, we used Orthofinder (v2.3.12) (*61*, *62*). We used gene annotations for *A. thaliana* TAIR10 (*50*), and reproductive genes were identified using functions and GO terms (from www.arabidopsis.org/).

Protein evolution

To examine protein evolution of sex-linked and autosomal genes, we first pruned the trees described above at the closest *P. patens* homolog (prune_tree.py, https://doi.org/10.5061/dryad.v41ns1rsm) (*117*). For genes that had a *C. purpureus* chromosome 5 homolog, the R40 and GG1 leaves were identified instead and pruned at the closest homolog in *P. patens*. The chromosome 5 homologs were used to assess dN/dS on autosomal genes in *C. purpureus* and were

specifically targeted given the recent fusion of the chromosome 5 homeolog to the sex chromosomes. All other copies of a *C. purpureus* gene were removed, and which copy of a gene to keep for all other species was chosen at random. To get dN/dS ratios, we used PAML (v4.9a) (*66*) (additional scripts for this analysis in https://doi.org/10.5061/dryad.v41ns1rsm). For the sex-linked gene trees, we allowed the U and V to evolve at different rates than the rest of the tree. For the chromosome 5 homologs, the GG1 and R40 branches could evolve at a different rate than the rest of the tree. dN/dS values >5 were removed from further analyses. To determine whether there is a significant difference in dN/dS on autosomal, U-, and V-linked genes, we ran a pairwise MWU for multiple tests (*79*, *80*) ($n_{\text{Autosomes}} = 61$, $n_{\text{U}} = 314$, $n_{\text{V}} = 315$).

Ka and Ks analysis

FASTA files of in-frame *C. purpureus* and *M. polymorpha* sex-linked genes were aligned (see above) and converted to axt format (array_hash_extractor_fasta_unlock_ks.pl and aln_to_axt.pl, https://doi.org/10.5061/dryad.v41ns1rsm). *Ka*, *Ks*, and *Ka*/*Ks* were calculated using KaKs_Calculator (v2.0) (*118*) using the Goldman and Yang model (*67*) on only one-to-one UV orthologs. *Ks* was plotted on the U and V sex chromosomes (Fig. 3) in R (v3.5.3) (*77*) using karyoploteR (v1.8.8) (*78*) and edited in Inkscape (v0.92.2) (https://inkscape.org/en/). One gene with *Ks* > 3, but coalescence in *C. purpureus* was removed from the plot (ceratodon_genome_plots.R, https://doi.org/10.5061/dryad.v41ns1rsm).

Codon analyses

To analyze codon-usage biases, we used CodonW (v1.4.2; J. Penden, https://sourceforge.net/projects/codonw). We first removed any gene that had no expression to remove potential pseudogenes. We also removed genes with less than 200 codons to reduce the variance around calculated codon values (alignment_length_filter.pl, https://doi.org/10.5061/dryad.v41ns1rsm) (*119*). We ran a correspondence analysis on autosomal, U-, and V-linked genes together to determine the optimal codons in *C. purpureus*. We next separately determined the fop usage (*120*), ENC, and GC content of the third synonymous position of a codon (GC3s) on autosomes, U-, and V-linked genes. To determine whether there is a significant difference between fop, ENC, and GC3s between autosomes, U, and V, we ran a pairwise MWU for multiple tests (*79*, *80*) in R (v3.5.3) (*77*) ($n_{\text{Autosomes}} = 15{,}677$, $n_{\text{U}} = 797$, $n_{\text{V}} = 736$) and plotted the results (Fig. 2) using ggplot2 (v3.2.1) (*92*) using default box-plot elements (ceratodon_genome_plots.R, https://doi.org/10.5061/dryad.v41ns1rsm).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/27/eabh2488/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. J. J. Bull, *Evolution of Sex Determining Mechanisms* (The Benjamin/Cummings Publishing Company Inc., 1983).
2. B. Charlesworth, D. Charlesworth, The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 1563–1572 (2000).
3. D. Bachtrog, Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).
4. D. Bachtrog, M. Kirkpatrick, J. E. Mank, S. F. McDaniel, J. C. Pires, W. Rice, N. Valenzuela, Are all sex chromosomes created equal? *Trends Genet.* **27**, 350–357 (2011).
5. S. Carey, L. Kollar, S. McDaniel, Does degeneration or genetic conflict shape gene content on UV sex chromosomes? *EcoEvoRxiv* 10.32942/osf.io/hs6w3 (2020).
6. S. F. McDaniel, K. M. Neubig, A. C. Payton, R. S. Quatrano, D. J. Cove, RECENT gene-capture on the UV sex chromosomes of the moss ceratodon purpureus : Sex chromosome evolution in Ceratodon purpureus. *Evolution* **67**, 2811–2222 (2013).
7. S. F. McDaniel, J. Atwood, J. G. Burleigh, Recurrent evolution of dioecy in bryophytes. *Evolution* **67**, 567–572 (2013).
8. S. F. McDaniel, J. H. Willis, A. J. Shaw, A linkage map reveals a complex basis for segregation distortion in an interpopulation cross in the moss Ceratodon purpureus. *Genetics* **176**, 2489–2500 (2007).
9. R. Fritsch, *Index to Bryophyte Chromosome Counts* (1991); available at www.schweizerbart.de/publications/detail/isbn/9783443620127/%23).
10. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
11. P. Szövényi, P.-F. Perroud, A. Symeonidi, S. Stevenson, R. S. Quatrano, S. A. Rensing, A. C. Cuming, S. F. McDaniel, De novo assembly and comparative analysis of the Ceratodon purpureus transcriptome. *Mol. Ecol. Resour.* **15**, 203–215 (2015).
12. One Thousand Plant Transcriptomes Initiative, One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
13. D. Lang, K. K. Ullrich, F. Murat, J. Fuchs, J. Jenkins, F. B. Haas, M. Piednoel, H. Gundlach, M. Van Bel, R. Meyberg, The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).
14. S. W. Schaeffer, Muller "Elements" in *Drosophila*: How the search for the genetic basis for speciation led to the birth of comparative genomics. *Genetics* **210**, 3–13 (2018).
15. R. Bergero, D. Charlesworth, The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* **24**, 94–102 (2009).
16. S. A. Montgomery, Y. Tanizawa, B. Galik, N. Wang, T. Ito, T. Mochizuki, S. Akimcheva, J. L. Bowman, V. Cognat, L. Maréchal-Drouard, H. Ekker, S.-F. Hong, T. Kohchi, S.-S. Lin, L.-Y. D. Liu, Y. Nakamura, L. R. Valeeva, E. V. Shakirov, D. E. Shippen, W.-L. Wei, M. Yagura, S. Yamaoka, K. T. Yamato, C. Liu, F. Berger, Chromatin organization in early land plants reveals an ancestral association between H3K27me3, transposons, and constitutive heterochromatin. *Curr. Biol.* **30**, 573–588.e7 (2020).
17. P. Ferris, B. J. S. C. Olson, P. L. De Hoff, S. Douglass, D. Casero, S. Prochnik, S. Geng, R. Rai, J. Grimwood, J. Schmutz, I. Nishii, T. Hamaji, H. Nozaki, M. Pellegrini, J. G. Umen, Evolution of an expanded sex-determining locus in Volvox. *Science* **328**, 351–354 (2010).
18. J. L. Bowman, T. Kohchi, K. T. Yamato, J. Jenkins, S. Shu, K. Ishizaki, S. Yamaoka, R. Nishihama, Y. Nakamura, F. Berger, C. Adam, S. S. Aki, F. Althoff, T. Araki, M. A. Arteaga-Vazquez, S. Balasubrmanian, K. Barry, D. Bauer, C. R. Boehm, L. Briginshaw, J. Caballero-Perez, B. Catarino, F. Chen, S. Chiyoda, M. Chovatia, K. M. Davies, M. Delmans, T. Demura, T. Dierschke, L. Dolan, A. E. Dorantes-Acosta, D. M. Eklund, S. N. Florent, E. Flores-Sandoval, A. Fujiyama, H. Fukuzawa, B. Galik, D. Grimanelli, J. Grimwood, U. Grossniklaus, T. Hamada, J. Haseloff, A. J. Hetherington, A. Higo, Y. Hirakawa, H. N. Hundley, Y. Ikeda, K. Inoue, S.-I. Inoue, S. Ishida, Q. Jia, M. Kakita, T. Kanazawa, Y. Kawai, T. Kawashima, M. Kennedy, K. Kinose, T. Kinoshita, Y. Kohara, E. Koide, K. Komatsu, S. Kopischke, M. Kubo, J. Kyozuka, U. Lagercrantz, S.-S. Lin, E. Lindquist, A. M. Lipzen, C.-W. Lu, E. De Luna, R. A. Martienssen, N. Minamino, M. Mizutani, M. Mizutani, N. Mochizuki, I. Monte, R. Mosher, H. Nagasaki, H. Nakagami, S. Naramoto, K. Nishitani, M. Ohtani, T. Okamoto, M. Okumura, J. Phillips, B. Pollak, A. Reinders, M. Rövekamp, R. Sano, S. Sawa, M. W. Schmid, M. Shirakawa, R. Solano, A. Spunde, N. Suetsugu, S. Sugano, A. Sugiyama, R. Sun, Y. Suzuki, M. Takenaka, D. Takezawa, H. Tomogane, M. Tsuzuki, T. Ueda, M. Umeda, J. M. Ward, Y. Watanabe, K. Yazaki, R. Yokoyama, Y. Yoshitake, I. Yotsui, S. Zachgo, J. Schmutz, Insights into land plant evolution garnered from the marchantia polymorpha genome. *Cell* **171**, 287–304.e15 (2017).
19. S. Ahmed, J. M. Cock, E. Pessia, R. Luthringer, A. Cormier, M. Robuchon, L. Sterck, A. F. Peters, S. M. Dittami, E. Corre, M. Valero, J.-M. Aury, D. Roze, Y. Van de Peer, J. Bothwell, G. A. B. Marais, S. M. Coelho, A haploid system of sex determination in the brown alga Ectocarpus sp. *Curr. Biol.* **24**, 1945–1957 (2014).
20. S. Immler, S. P. Otto, The evolution of sex chromosomes in organisms with separate haploid sexes. *Evolution* **69**, 694–708 (2015).
21. B. Laenen, B. Shaw, H. Schneider, B. Goffinet, E. Paradis, A. Désamoré, J. Heinrichs, J. C. Villarreal, S. R. Gradstein, S. F. McDaniel, D. G. Long, L. L. Forrest, M. L. Hollingsworth, B. Crandall-Stotler, E. C. Davis, J. Engel, M. Von Konrat, E. D. Cooper, J. Patiño, C. J. Cox, A. Vanderpoorten, A. J. Shaw, Extant diversity of bryophytes emerged from successive post-Mesozoic diversification bursts. *Nat. Commun.* **5**, 5134 (2014).
22. B. T. Lahn, D. C. Page, Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
23. S. F. McDaniel, J. H. Willis, A. J. Shaw, The genetic basis of developmental abnormalities in interpopulation hybrids of the moss Ceratodon purpureus. *Genetics* **179**, 1425–1435 (2008).

24. T. E. Norrell, K. S. Jones, A. C. Payton, S. F. McDaniel, Meiotic sex ratio variation in natural populations of Ceratodon purpureus (Ditrichaceae). *Am. J. Bot.* **101**, 1572–1576 (2014).
25. M. V. Chibalina, D. A. Filatov, Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* **21**, 1475–1479 (2011).
26. D. W. Bellott, J. F. Hughes, H. Skaletsky, L. G. Brown, T. Pyntikova, T.-J. Cho, N. Koutseva, S. Zaghlul, T. Graves, S. Rock, C. Kremitzki, R. S. Fulton, S. Dugan, Y. Ding, D. Morton, Z. Khan, L. Lewis, C. Buhay, Q. Wang, J. Watt, M. Holder, S. Lee, L. Nazareth, J. Alföldi, S. Rozen, D. M. Muzny, W. C. Warren, R. A. Gibbs, R. K. Wilson, D. C. Page, Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
27. G. J. Pazour, B. L. Dickert, G. B. Witman, The DHC1b (DHC2) isoform of cytoplasmic dynein is required for flagellar assembly. *J. Cell Biol.* **144**, 473–481 (1999).
28. S. Koshimizu, R. Kofuji, Y. Sasaki-Sekimoto, M. Kikkawa, M. Shimojima, H. Ohta, S. Shigenobu, Y. Kabeya, Y. Hiwatashi, Y. Tamada, T. Murata, M. Hasebe, Physcomitrella MADS-box genes regulate water supply and sperm movement for fertilization. *Nat. Plants.* **4**, 36–45 (2018).
29. M. Endo, S. Nakamura, T. Araki, N. Mochizuki, A. Nagatani, Phytochrome B in the mesophyll delays flowering by suppressing FLOWERING LOCUS T expression in Arabidopsis vascular bundles. *Plant Cell* **17**, 1941–1952 (2005).
30. T. Kuromori, T. Ito, E. Sugimoto, K. Shinozaki, Arabidopsis mutant of AtABCG26, an ABC transporter gene, is defective in pollen maturation. *J. Plant Physiol.* **168**, 2001–2005 (2011).
31. M. Rövekamp, J. L. Bowman, U. Grossniklaus, Marchantia MpRKD regulates the gametophyte-sporophyte transition by keeping egg cells quiescent in the absence of fertilization. *Curr. Biol.* **26**, 1782–1789 (2016).
32. F. Tedeschi, P. Rizzo, T. Rutten, L. Altschmied, H. Bäumlein, RWP-RK domain-containing transcription factors control cell differentiation during female gametophyte development in Arabidopsis. *New Phytol.* **213**, 1909–1924 (2017).
33. T. Hisanaga, K. Okahashi, S. Yamaoka, T. Kajiwara, R. Nishihama, M. Shimamura, K. T. Yamato, J. L. Bowman, T. Kohchi, K. Nakajima, A *cis*-acting bidirectional transcription switch controls sexual dimorphism in the liverwort. *EMBO J.* **38**, e100240 (2019).
34. V. Demko, E. Ako, P.-F. Perroud, R. Quatrano, O.-A. Olsen, The phenotype of the CRINKLY4 deletion mutant of Physcomitrella patens suggests a broad role in developmental regulation in early land plants. *Planta* **244**, 275–284 (2016).
35. S. S. Renner, The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database. *Am. J. Bot.* **101**, 1588–1596 (2014).
36. S. M. Eppley, L. K. Jesson, Moving to mate: The evolution of separate and combined sexes in multicellular organisms. *J. Evol. Biol.* **21**, 727–736 (2008).
37. D. A. Sasson, J. F. Ryan, A reconstruction of sexual modes throughout animal evolution. *BMC Evol. Biol.* **17**, 242 (2017).
38. S. F. McDaniel, A. J. Shaw, Selective sweeps and intercontinental migration in the cosmopolitan moss Ceratodon purpureus (Hedw.) Brid. *Mol. Ecol.* **14**, 1121–1132 (2005).
39. C.-L. Xiao, Y. Chen, S.-Q. Xie, K.-N. Chen, Y. Wang, Y. Han, F. Luo, Z. Xie, MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
40. C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
41. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer provides a one-click system for analyzing loop-resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
42. W. J. Kent, BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
43. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (16 March 2013).
44. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
45. T. D. Wu, J. Reeder, M. Lawrence, G. Becker, M. J. Brauer, GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. *Methods Mol. Biol.* **1418**, 283–334 (2016).
46. J. Krumsiek, R. Arnold, T. Rattei, Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
47. T. D. Wu, S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
48. J. T. Lovell, J. Jenkins, D. B. Lowry, S. Mamidi, A. Sreedasyam, X. Weng, K. Barry, J. Bonnette, B. Campitelli, C. Daum, S. P. Gordon, B. A. Gould, A. Khasanova, A. Lipzen, A. MacQueen, J. D. Palacio-Mejía, C. Plott, E. V. Shakirov, S. Shu, Y. Yoshinaga, M. Zane, D. Kudrna, J. D. Talag, D. Rokhsar, J. Grimwood, J. Schmutz, T. E. Juenger, The genomic landscape of molecular responses to natural drought stress in Panicum hallii. *Nat. Commun.* **9**, 5213 (2018).
49. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr., L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, O. White, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
50. P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, E. Huala, The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
51. J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, S. A. Jackson, Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
52. S. Mamidi, A. Healey, P. Huang, J. Grimwood, J. Jenkins, K. Barry, A. Sreedasyam, S. Shu, J. T. Lovell, M. Feldman, J. Wu, Y. Yu, C. Chen, J. Johnson, H. Sakakibara, T. Kiba, T. Sakurai, R. Tavares, D. A. Nusinow, I. Baxter, J. Schmutz, T. P. Brutnell, E. A. Kellogg, A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **38**, 1203–1210 (2020).
53. O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, P. Wincker; French-Italian Public Consortium for Grapevine Genome Characterization, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
54. J. A. Banks, T. Nishiyama, M. Hasebe, J. L. Bowman, M. Gribskov, C. dePamphilis, V. A. Albert, N. Aono, T. Aoyama, B. A. Ambrose, N. W. Ashton, M. J. Axtell, E. Barker, M. S. Barker, J. L. Bennetzen, N. D. Bonawitz, C. Chapple, C. Cheng, L. G. G. Correa, M. Dacre, J. DeBarry, I. Dreyer, M. Elias, E. M. Engstrom, M. Estelle, L. Feng, C. Finet, S. K. Floyd, W. B. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwatashi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Kolukisaoglu, M. Kubo, T. Kurata, S. Lalonde, K. Li, Y. Li, A. Litt, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S.-I. Morinaga, T. Murata, B. Mueller-Roeber, D. R. Nelson, M. Obara, Y. Oguri, R. G. Olmstead, N. Onodera, B. L. Petersen, B. Pils, M. Prigge, S. A. Rensing, D. M. Riaño-Pachón, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shakirov, N. Shibagaki, N. Shinohara, D. E. Shippen, I. Sørensen, R. Sotooka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurdzic, G. Theissen, P. Ulvskov, S. Wakazuki, J.-K. Weng, W. W. G. T. Willats, D. Wipf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, D. Loqué, R. Otillar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lucas, D. Rokhsar, I. V. Grigoriev, The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960–963 (2011).
55. S. S. Merchant, S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Maréchal-Drouard, W. F. Marshall, L.-H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C.-L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernández, H. Fukuzawa, D. González-Ballester, D. González-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J.-P. Ral, D. M. Riaño-Pachón, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C.-J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y. W. Huang, J. Jhaveri, Y. Luo, D. Martínez, W. C. A. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Grigoriev, D. S. Rokhsar, A. R. Grossman, The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007).
56. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. 2013–2015 (2015).
57. A. F. A. Smit, R. Hubley, P. Green, RepeatModeler Open-1.0. 2008–2015 (Institute for Systems Biology, 2015); www. repeatmasker.org [last accessed 1 May 2018].
58. A. A. Salamov, V. V. Solovyev, Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516–522 (2000).
59. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

60. K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, M. Stanke, BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
61. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
62. D. M. Emms, S. Kelly, OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
63. A. Zwaenepoel, Y. Van de Peer, wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).
64. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics.* **10**, 421 (2009).
65. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
66. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
67. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
68. S. Proost, J. Fostier, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer, K. Vandepoele, i-ADHoRe 3.0—Fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).
69. G. P. Tiley, M. S. Barker, J. G. Burleigh, Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol. Evol.* **10**, 2882–2898 (2018).
70. S. A. Rensing, J. Ick, J. A. Fawcett, D. Lang, A. Zimmer, Y. Van de Peer, R. Reski, An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evol. Biol.* **7**, 130 (2007).
71. T. Flutre, E. Duprat, C. Feuillet, H. Quesneville, Considering transposable element diversification in de novo annotation approaches. *PLOS ONE* **6**, e16526 (2011).
72. H. Quesneville, C. M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner, D. Anxolabehere, Combined evidence annotation of transposable elements in genome sequences. *PLOS Comput. Biol.* **1**, e22 (2005).
73. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
74. C. Hoede, S. Arnoux, M. Moisset, T. Chaumier, O. Inizan, V. Jamilloux, H. Quesneville, PASTEC: An automatic transposable element classification tool. *PLOS ONE* **9**, e91929 (2014).
75. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics.* **20**, 473 (2019).
76. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
77. R Core Team, R: A language and environment for statistical computing (2013); available at www.R-project.org/).
78. B. Gel, E. Serra, karyoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
79. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* **57**, 289–300 (1995).
80. H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
81. D. Lang, B. Weiche, G. Timmerhaus, S. Richardt, D. M. Riaño-Pachón, L. G. G. Corrêa, R. Reski, B. Mueller-Roeber, S. A. Rensing, Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: A timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503 (2010).
82. P. K. I. Wilhelmsson, C. Mühlich, K. K. Ullrich, S. A. Rensing, Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. *Genome Biol. Evol.* **9**, 3384–3397 (2017).
83. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
84. D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
85. K. C. Olney, S. M. Brotman, J. P. Andrews, V. A. Valverde-Vesling, M. A. Wilson, Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data. *Biol. Sex Differ.* **11**, 42 (2020).
86. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
87. M. Love, S. Anders, W. Huber, Differential analysis of count data–the DESeq2 package. *Genome Biol.* **15**, 10–1186 (2014).
88. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
89. H. Wickham, reshape2: Flexibly reshape data: A reboot of the reshape package. *R package version*. **1** (2012); available at http://cran.ms.unimelb.edu.au/web/packages/reshape2/.
90. A. de Vries, B. D. Ripley, ggdendro: Create Dendrograms and Tree Diagrams Using'ggplot2'. *R package version 0. 1–20* (2016).
91. H. Wickham, T. L. Pedersen, gtable: Arrange grobs in tables. *R package version 0. 2. 0* (2016).
92. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
93. B. Auguie, A. Antonov, M. B. Auguie, Package 'gridExtra.' *Miscellaneous Functions for "Grid" Graphics* (2017); available at https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf.
94. S. Garnier, N. Ross, B. Rudis, M. Sciaini, C. Scherer, viridis: Default Color Maps from 'matplotlib.' *R package version 0. 5. 1* (2018).
95. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
96. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**, 1–9 (2006).
97. F. Briatte, ggnetwork: Geometries to Plot Networks with'ggplot2'. *R package version 0. 5. 1* (2016).
98. A. Alexa, J. Rahnenfuhrer, topGO: Enrichment analysis for gene ontology. *R package version*, (2010).
99. G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, S. Wang, GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
100. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
101. S. Alaba, P. Piszczalka, H. Pietrykowska, A. M. Pacak, I. Sierocka, P. W. Nuc, K. Singh, P. Plewka, A. Sulkowska, A. Jarmolowski, W. M. Karlowski, Z. Szweykowska-Kulinska, The liverwort Pellia endiviifolia shares microtranscriptomic traits that are common to green algae and land plants. *New Phytol.* **206**, 352–367 (2015).
102. M. G. Johnson, C. Malley, B. Goffinet, A. J. Shaw, N. J. Wickett, A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta). *Mol. Phylogenet. Evol.* **98**, 29–40 (2016).
103. S. Gao, H.-N. Yu, Y.-F. Wu, X.-Y. Liu, A.-X. Cheng, H.-X. Lou, Cloning and functional characterization of a phenolic acid decarboxylase from the liverwort Conocephalum japonicum. *Biochem. Biophys. Res. Commun.* **481**, 239–244 (2016).
104. F.-W. Li, P. Brouwer, L. Carretero-Paulet, S. Cheng, J. de Vries, P.-M. Delaux, A. Eily, N. Koppers, L.-Y. Kuo, Z. Li, M. Simenc, I. Small, E. Wafula, S. Angarita, M. S. Barker, A. Bräutigam, C. dePamphilis, S. Gould, P. S. Hosmani, Y.-M. Huang, B. Huettel, Y. Kato, X. Liu, S. Maere, R. McDowell, L. A. Mueller, K. G. J. Nierop, S. A. Rensing, T. Robison, C. J. Rothfels, E. M. Sigel, Y. Song, P. R. Timilsena, Y. Van de Peer, H. Wang, P. K. I. Wilhelmsson, P. G. Wolf, X. Xu, J. P. Der, H. Schluepmann, G. K.-S. Wong, K. M. Pryer, Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4**, 460–472 (2018).
105. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
106. B. Haas, A. Papanicolaou, TransDecoder (find coding regions within transcripts). Github, nd (2015); https://github. com/TransDecoder/TransDecoder [accessed 17 May 2018].
107. A. Bateman, E. Birney, R. Durbin, S. R. Eddy, R. D. Finn, E. L. Sonnhammer, Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27**, 260–262 (1999).
108. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
109. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
110. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
111. M. Suyama, D. Torrents, P. Bork, PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W012 (2006).
112. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
113. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
114. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T. Lam, ggtree : An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

115. G. Yu, T. T.-Y. Lam, H. Zhu, Y. Guan, Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
116. T. Junier, E. M. Zdobnov, The Newick utilities: High-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
117. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
118. Z. Zhang, J. Li, X.-Q. Zhao, J. Wang, G. K.-S. Wong, J. Yu, KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).
119. F. Wright, The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
120. M. Stenico, A. T. Lloyd, P. M. Sharp, Codon usage in Caenorhabditis elegans: Delineation of translational selection and mutational biases. *Nucleic Acids Res.* **22**, 2437–2446 (1994).
121. M. Luo, R. A. Wing, An improved method for plant BAC library construction. *Methods Mol. Biol.* **236**, 3–20 (2003).
122. B. Ewing, L. Hillier, M. C. Wendl, P. Green, Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
123. B. Ewing, P. Green, Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
124. D. Gordon, C. Abajian, P. Green, Consed: A graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
125. P.-F. Perroud, F. B. Haas, M. Hiss, K. K. Ullrich, A. Alboresi, M. Amirebrahimi, K. Barry, R. Bassi, S. Bonhomme, H. Chen, J. C. Coates, T. Fujita, A. Guyon-Debast, D. Lang, J. Lin, A. Lipzen, F. Nogué, M. J. Oliver, I. Ponce de León, R. S. Quatrano, C. Rameau, B. Reiss, R. Reski, M. Ricca, Y. Saidi, N. Sun, P. Szövényi, A. Sreedasyam, J. Grimwood, G. Stacey, J. Schmutz, S. A. Rensing, The Physcomitrella patens gene atlas project: Large-scale RNA-seq based expression data. *Plant J.* **95**, 168–182 (2018).
126. D. J. Cove, P.-F. Perroud, A. J. Charron, S. F. McDaniel, A. Khandelwal, R. S. Quatrano, Culturing the moss *Physcomitrella patens*. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5136 (2009).
127. E. H. Leder, J. M. Cano, T. Leinonen, R. B. O'Hara, M. Nikinmaa, C. R. Primmer, J. Merilä, Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks. *Mol. Biol. Evol.* **27**, 1495–1503 (2010).
128. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
129. Y.-W. Yuan, S. R. Wessler, The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7884–7889 (2011).
130. H. Tsubouchi, G. S. Roeder, The Mnd1 protein forms a complex with hop2 to promote homologous chromosome pairing and meiotic double-strand break repair. *Mol. Cell. Biol.* **22**, 3078–3088 (2002).

Acknowledgments: We thank R. Quatrano, D. Cove, and the Quatrano laboratory at Washington University in St. Louis for incubating the *C. purpureus* genome project; B. Hauser, T. Colquhoun, and D. Garner for assisting with the hormone and light perturbations; and B. Goffinet, M. Mack, S. Robinson, T. Rosenstiel, B. Shaw, and the late N. Miller for providing field collections. S. Renner, S. Otto, M. Kirkpatrick, and M. Hahn provided valuable feedback on a draft of the manuscript, and we thank two anonymous reviewers for helpful feedback on the current version. The One Thousand Plant Transcriptome Initiative provided early access to moss and liverwort data, and the University of Florida Interdisciplinary Center for Biotechnology Research and HiPerGator provided vital technical support throughout the project. **Funding:** This work was supported by NSF DEB-1541005 and 1542609 and start-up funds from UF to S.F.M.; microMORPH Cross-Disciplinary Training Grant, Sigma-Xi Grant-In-Aid of Research, and Society for the Study of Evolution Rosemary Grant Award to S.B.C.; NSF DEB-1239992 to N.J.W.; the Emil Aaltonen Foundation and the University of Turku to S.O.; and NSF DEB-1541506 to J.G.B. and S.F.M. The work conducted by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. **Author contributions:** Conceptualization: S.F.M. Data curation: S.B.C. and K.B. Formal analysis: S.B.C., J.J., S.S., J.T.L., F.M., A.S., G.P.T., N.F.-P., A.H., and S.A.R. Funding acquisition: S.B.C., S.H., N.J.W., J.G.B., S.A.R., and S.F.M. Investigation: S.B.C., J.J., A.C.P., S.S., J.T.L., F.M., A.S., G.P.T., N.F.-P., A.H., C.C., M.W., A.L., C.D., C.A.S., J.C.M., R.E.C., L.M.K., S.O., S.H., J.B.L., J.G.B., S.A.R., and S.F.M. Methodology: S.B.C., A.C.P., J.J., J.S., and S.F.M. Project administration: S.B.C., K.B., J.G., J.S., and S.F.M. Resources: N.J.W., M.G.J., J.G., J.S., and S.F.M. Supervision: M.G.J., S.A.R., J.G., J.S., and S.F.M. Visualization: S.B.C., J.J., J.T.L., A.S., and G.P.T. Writing—original draft: S.B.C. and S.F.M. Writing—review and editing: S.B.C., A.C.P., J.J., J.T.L., F.M., A.S., G.P.T., A.H., C.A.S., L.M.K., J.G.B., N.J.W., M.G.J., S.A.R., J.G., J.S., and S.F.M. All authors approve of the final draft of this manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. DNA and RNA data for this manuscript can be found under NCBI BioProjects listed in table S5. The R40 and GG1 v1.0 genome assemblies can be found on NCBI GenBank under the accessions JACMSA000000000 and JACMSB000000000, respectively. The genome assemblies and annotations can also be found on Phytozome (https://phytozome-next.jgi.doe.gov/). The Lanisha TE has been deposited in NCBI GenBank under MT647524. The published sequence data used in this study can be found in table S17. Supporting documents for the phylogenomic analysis of sex-linked genes can be found on Dryad under https://doi.org/10.5061/dryad.v41ns1rsm. Materials and correspondence should be addressed to stuartmcdaniel@ufl.edu.

Submitted 24 February 2021
Accepted 14 May 2021
Published 30 June 2021
10.1126/sciadv.abh2488

Citation: S. B. Carey, J. Jenkins, J. T. Lovell, F. Maumus, A. Sreedasyam, A. C. Payton, S. Shu, G. P. Tiley, N. Fernandez-Pozo, A. Healey, K. Barry, C. Chen, M. Wang, A. Lipzen, C. Daum, C. A. Saski, J. C. McBreen, R. E. Conrad, L. M. Kollar, S. Olsson, S. Huttunen, J. B. Landis, J. G. Burleigh, N. J. Wickett, M. G. Johnson, S. A. Rensing, J. Grimwood, J. Schmutz, S. F. McDaniel, Gene-rich UV sex chromosomes harbor conserved regulators of sexual development. *Sci. Adv.* **7**, eabh2488 (2021).

ScienceAdvances

Gene-rich UV sex chromosomes harbor conserved regulators of sexual development

Sarah B. CareyJerry JenkinsJohn T. LovellFlorian MaumusAvinash SreedasyamAdam C. PaytonShengqiang ShuGeorge P. TileyNoe Fernandez-PozoAdam HealeyKerrie BarryCindy ChenMei WangAnna LipzenChris DaumChristopher A. SaskiJordan C. McBreenRoth E. ConradLeslie M. KollarSanna OlssonSanna HuttunenJacob B. LandisJ. Gordon BurleighNorman J. WickettMatthew G. JohnsonStefan A. RensingJane GrimwoodJeremy SchmutzStuart F. McDaniel

Sci. Adv., 7 (27), eabh2488. • DOI: 10.1126/sciadv.abh2488

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.