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

Sarah B. Carey1‡*, Jerry Jenkins2, John T. Lovell3, Florian Maumus7, Avinash Sreedasyam2, Adam C. Payton1,4, Shengqiang Shu5, George P. Tiley6, Noe Fernández-Pozo7, Adam Healey2, Kerrie Barry5, Cindy Chen5, Mei Wang5, Anna Lipzen5, Chris Daum5, Christopher A. Sasaki8, Jordan C. McBreen4, Roth E. Conrad9, Leslie M. Kollar1, Sanna Olsson10, Sanna Huttunen11, Jacob B. Landis12, J. Gordon Burleigh1, Norman J. Wickett13, Matthew G. Johnson14, Stefan A. Rensing7,15,16, Jane Grimwood2,5, Jeremy Schmutz2,5, Stuart F. McDaniel1*

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 nonrecombinating U and males inherit a nonrecombinating 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.
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 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 homologous 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 autoosome. 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 × 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 × 10⁻⁴) 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 × 10⁻³; 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 × 10⁻⁶ for all metrics; Fig. 2). Although, notably, the U- and V-linked genes were not different from one another (MWU, autosomes to U or V 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).

**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 homologous chromosomes to C. purpureus 5 and 9, respectively, had fused to the sex chromosomes.

**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 × 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.
Carey et al., Sci. Adv. 2021; 7 : eabh2488     30 June 2021

SCIENCE ADVANCES | RESEARCH ARTICLE

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 MrRKD, 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 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...
Transcript assemblies were made from ~1.5 billion pairs of 2 × 150 bp genome annotations (1,599,605 bp). More regions containing tandem repeats (fig. S1). The overall base bly. In 14 of the BAC clones, the alignments were of high quality selected to assess the accuracy of the assembly. A range of variants were shown in table S4.

Resulting final statistics are 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 constructs 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 Bryopsis, 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 Csore 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 1-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/tjile/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, macro-syntenic 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 TEadenovo 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 ($n_{\text{Auto}} = 2736, n_{U} = 1247, n_{V} = 1229$).

**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 ($n_{\text{isolates}} = 6$, three of each sex) at gametophore and protonemal stages ($n_{\text{stages}} = 2$) in triplicate ($n_{\text{replicates}} = 3$) (table S5; see Supplementary Materials and Methods for details on tissue conditions). Raw reads were filtered for contaminants and adapters removed using BBADuk (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?l=manual#deseq), and gene IDs renamed (using mstrg_prep.pl, https://gist.github.com/gpertea/b83f1b323435e166afa92ad23852f7b). 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 protone-mal and gametophore stages. For autosomal genes, we removed those with baseMean < 1, a log 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 gg dendro (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 (B). 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 Trimomatic (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 (v8.2.16) (117) 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 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 homolog 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) ($n_{Autosomes} = 61$, $n_U = 314$, $n_V = 315$).

Ka and Ks analyses
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 karyplotR (v1.8.8) (78) and edited in Inkscape (v0.92.2) (https://inkscape.org/en/).

Codon analyses
To analyze codon-usage biases, we used CodonW (v1.4.2; J. Peden, 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 top 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 U, ENC, and GC3s between autosomes, U, and V, we ran a pairwise MWU for multiple tests (79, 80) in R (v3.5.3) (77) ($n_{Autosomes} = 15,677$, $n_U = 797$, $n_V = 736$) 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/eabf2488/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 degenaration 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? Ecology 92, 2092–2099 (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, 2211–2222 (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 interspecific cross in the moss Ceratodon purpureus. Genetics 176, 2489–2500 (2007).
9. R. Fritsch, Index to Byphphyte Chromosome Counts (1991); available at www.schweizerbart.de/publications/detail/isbn/9783443620127/.
10. One Thousand Plant Transcriptomes Initiative, One thousand plant transcriptomes of moss Ceratodon purpureus was 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) ($n_{Autosomes} = 61$, $n_U = 314$, $n_V = 315$).
31. M. S. Kabeya, Y. Palacio-Mejía, C. R. Renner, The relative and molecular responses to sexual modes throughout animal evolution. *EMBO J.* **38**, e100240 (2019).

32. J. A. Banks, T. Nishiyama, M. Hasebe, J. L. Bowman, M. Grivko, S. dePamphilis, V. A. Albert, N. Anoos, T. Ayamizu, A. Ambrose, N. W. Ashton, M. J. Antell, E. Barker, M. R. Barker, J. L. Bennezet, N. D. Bonowitz, C. Chaplin, C. Cheng, L. G. G. Correa, D. Dreze, J. Debray, I. E. M. Cass, M. Elias, E. M. Engstrom, M. Estrelle, L. Feng, C. Finet, S. K. Floyd, W. D. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwastshi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Koliskaugolü, M. Kubo, T. Kurata, S. Landonde, K. L. Li, L. L. Li, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S. I. Morinaga, T. Murata, B. Mueller-Roever, D. R. Nelson, M. Obara, Y. Ogino, R. G. Olmstead, N. Onodera, B. L. Petersen, B. Pili, M. Prigg, A. S. Rensing, D. M. Riaño-Pachón, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shikakov, N. Shikagushi, N. Shinohara, D. E. Shippen, I. Sorensen, R. Sotovka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurzid, G. Theissen, P. Uhlovsk, S. Wakazu, J. K. Wang, W. G. W. T. W. Delhanty, D. W. Wolf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, L. Roquet, R. Olilar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lukas, D. Rojkars, I. V. Girgoriev, The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *The Selaginella genome* **332**, 960–966 (2013).

33. S. S. Merchant, S. E. Prochnik, O. Volland, E. H. Harris, S. J. Karpowicz, G. B. Wittman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Maréchal-Drouard, W. F. Marshall, L. H. Qi, D. R. Nelson, A. A. Sanderson, M. H. Spalding, V. V. Kapitonov, O. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grummond, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C.-L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, F. Enerle, H. Fuzukaçu, D. Gonzalez-Ballester, D. Gonzalez-Halphen, A. Halman, M. Hänikenne, M. Hippler, M. Inwood, J. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lamire, A. V. Lobanov, M. Lohn, A. Manuell, I. Meier, M. L. Mettig, T. Mittelmeier, J. V. Moroney, J. Moseley, G. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Puzot, S. Purton, J.-P. Ralph, D. M. Riaño-Pachón, W. Riekhof, L. Rymarquis, M. Schroda, D. Serr, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Alm, J. Ackerman, K. Bisova, C.-C. Chen, M. Elias, K. Gendler, C. Haverkamp, M. R. Lamb, H. Leedam, J. C. Long, J. Minagawa, M. D. Page, J. Fan, W. Posthakam, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Cowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roever, S. Rajaman, R. T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y. W. Huang, J. Javey, Y. Luo, D. Martinez, W. C. A. Ngau, B. Otillar, A. Poliaik, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Girgoriev, D. S. Rokhsar, A. R. Grossman, The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007).

34. A. F. A. Smir, H. Hubley, P. Green, RepeatMasker Open-4.0. 2013–2015 (2015).

35. A. F. A. Smir, H. Hubley, P. Green, RepeatModeler-Open-1.0. 2008–2015 (Institute for Systems Biology, 2015); www.repeatmasker.org [last accessed 1 May 2018].

36. A. A. Salamov, V. V. Solovyev, Ab initio gene finding in Drosoforma genomic DNA. *Genome Res.* **10**, 516–522 (2000).

37. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *JBC* **30**, 1 (2005).
RNA-Seq-based genome annotation with D. K. i-ADHoRe 3.0—Fast and detection of Z. data sets. genome duplication contributed to, 473 (2019).
M. (2014).
Res.
27
H.
displaying arbitrary data.
B.
R Core Team, R: A language and environment for statistical computing (2013); available features.
approach to classification reveals that many plant-specific transcription factors evolved H. B. and analysis of
H.
B. Gel, R. A language and environment for statistical computing (2013); available at www.R-project.org/).
B. Gel, E. Serra, karyoploter: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33, 3088–3090 (2017).
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).
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).
D. Lang, B. Weiche, G. Timmerhaus, S. Richardt, D. M. Riao-Pachón, L. G. G. Corrêa, R. Reski, B. Mueller-Roessler, S. A. Rensing, Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: A timeline of loss, gain, expansion, and correlation with complexity. Genome Biol. 2, 488–503 (2010).
P. K. i. Wilhelmsson, C. Mühlich, K. K. Ulrich, S. A. Rensing, Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biol. 9, 3384–3397 (2017).
A. M. Bolger, M. Lohse, B. Usadel, Trimmmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360 (2015).
K. C. Olney, S. M. Bottman, J. P. Andrews, V. A. Valverde-Vesling, M. A. Wilson, Reference genome and transcriptome informed by the sex chromosome complement of the sample data.
M. K. C. Olney, S. M. Brotman, J. P. Andrews, V. A. Valverde-Vesling, M. A. Wilson, Reference
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., S.S., and J.T.L. Project administration: S.B.C., K.B., J.G., J.S., and S.F.M. Supervision: M.G.J., S.J., 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 approved 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.v41ns1rs. Materials and correspondence should be addressed to stuartmcaldaniel@ufl.edu.

Submitted 24 February 2021
Accepted 14 May 2021
Published 30 June 2021
10.1126/sciadv.abb2488

**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. Sasaki, J. C. McBreen, R. E. Conrad, L. M. Kollar, S. Olsson, S. Huttunen, B. L. tandis, 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).
Gene-rich UV sex chromosomes harbor conserved regulators of sexual development

Sarah B. Carey, Jerry Jenkins, John T. Lovell, Florian Maumus, Avinash Sreedasyam, Adam C. Payton, Shengqiang Shu, George P. Tiley, Noe Fernandez-Pozo, Adam Healey, Kerrie Barry, Cindy Chen, Mei Wang, Anna Lipzen, Chris Daum, Christopher A. Sasaki, Jordan C. McBreen, Roth E. Conrad, Leslie M. Kollar, Sanna Olsson, Sanna Huttunen, Jacob B. Landis, J. Gordon Burleigh, Norman J. Wickett, Matthew G. Johnson, Stefan A. Rensing, Jane Grimwood, Jeremy Schmutz and Stuart F. McDaniel

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