Title: The *Ceratodon purpureus* genome uncovers structurally complex, gene rich sex chromosomes

Authors: Sarah B. Carey[1,§,‡], Jerry Jenkins[2], John T. Lovell[2], Florian Maumus[3], Avinash Sreedasyam[2], Adam C. Payton[1,4], Shenqiang Shu[5], George P. Tiley[6], Noe Fernandez-Pozo[7], 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], Jane Grimwood[2,5], Jeremy Schmutz[2,5], and Stuart F. McDaniel[1,*]

Affiliations:

[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]US 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 & Biodiversity Unit, University of Turku, Finland

[12]School of Integrative Plant Science, Section of Plant Biology and the L.H. Bailey Hortorium, 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

[§]Current address: Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, AL, USA

[‡]Current address: HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

[*]Corresponding author: stuartmcdaniel@ufl.edu

Abstract: Non-recombining sex chromosomes, like the mammalian Y, often lose genes and accumulate transposable elements, a process termed degeneration[1,2]. The correlation between suppressed recombination and degeneration is clear in animal XY systems[1,2], 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[3,4]. Here we test for degeneration in the bryophyte UV sex chromosome system through genomic comparisons with new female and male chromosome-scale reference genomes of the moss *Ceratodon purpureus*. We show that the moss sex chromosomes evolved over 300 million years ago and expanded via two chromosomal fusions. Although the sex chromosomes show signs of weaker purifying selection than autosomes, we find suppressed recombination alone is insufficient to drive gene loss on sex-specific chromosomes. Instead, the U and V sex chromosomes harbor thousands of broadly-expressed genes, including numerous key regulators of sexual development across land plants.

Main text: Sex chromosomes arise when an ordinary pair of autosomes gains the capacity to determine sex[5]. A defining characteristic of sex chromosomes is suppressed recombination in the heterogametic sex. It is widely believed this lack of meiotic recombination makes natural selection less effective, predisposing non-recombining chromosomes, like the mammalian Y, to degeneration and gene-loss[1,2]. However, although some non-recombining chromosomes rapidly degenerate, or are completely lost, the sex chromosomes in other groups remain homomorphic or expand[1]. This diversity of form and gene content suggests 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 non-recombining U and males inherit a non-recombining V[3,4]. The sex-specific transmission pattern of both chromosomes means factors that are confounded in XY or ZW systems, such as suppressed recombination, hemizygosity, and sex-limited inheritance, are independent on UV chromosomes. Many UV sex chromosome systems may be ancient[4], 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 new chromosome-scale female and male genomes of the moss, *Ceratodon purpureus*.

Results: Ancestral-state reconstructions of dioecy suggest that sex chromosomes evolved early in the history of the extant mosses[6]. To reconstruct the evolutionary history of the bryophyte UV sex chromosomes, we assembled and annotated chromosome-scale genomes of the 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 [7], 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 (Supplementary Tables 1-4; Extended Data Fig. 1). The version 1.0 genome assembly of R40 comprises 358 Megabases (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[8]. 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 over 1.5 billion RNA-seq reads for each of the genome lines and additional *de novo* assemblies of other *C. purpureus* isolates (Supplementary Table 5), 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*[9]).

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)[10], and we found 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 (Fig. 1, Extended Data Fig. 1), consistent with previous transcriptomic[10,11] and our own *Ks*-based analyses (Extended Data Fig. 1; Supplementary Table 6). We also identified abundant synteny between the *C. purpureus* and *P. patens* chromosomes, which diverged over 200 million years ago (MYA)[12] (Fig. 1). This result demonstrates the ancestral karyotype of most extant mosses consisted of seven chromosomes[12], which we refer to as Ancestral Elements A-G (Fig. 1), and suggests major parts of the gene content of moss chromosomes are stable over hundreds of millions of years, similar to the “Muller Elements" in *Drosophila*[13]. Curiously, we could not detect the homeologs of the *C. purpureus* chromosomes 5 and 9 using synteny, an observation we return to below.

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 non-recombining Y or W sex chromosomes in other systems[14], 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 $P<2x10^{-16}$; 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 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 $P<1.5x10^{-14}$) 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 $P<1x10^{-4}$; Fig. 1; Supplementary Fig. 2). The distribution of repeats in *C. purpureus* and the physical proximity of the autosomes inferred from the Hi-C contact map (Extended Data Fig. 1) together highlight the enigmatic isolation of the sex chromosomes in the nucleus[15].

Unlike other non-recombining sex chromosomes, neither the U nor V show signs of major degeneration beyond the increased TE density. Sex-linked genes used on average one more codon than autosomes (ENC), less frequently use optimal codons (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 (Fig. 2; *MWU,* autosomes to U or V $P<6x10^{-6}$ for all metrics). Although, notably, the U and V-linked genes were not significantly different from one another (*MWU*, ENC $P=0.8$; fop $P=0.22$; GC3s $P=0.18$; dN/dS $P=0.73$), suggesting transmission through one sex or the other has no detectable effect on purifying selection. Consistent with this observation, the U and the V possess 3,450 and 3,411 transcripts, respectively, representing ~12% of the *C. purpureus* gene content. This stands in stark contrast to the non-recombining mammalian Y chromosome, or even other UV systems, which typically contain an order of magnitude fewer genes, at most[16–18]. These observations indicate that although suppressed recombination decreases the efficacy of natural selection, alone it is insufficient to drive gene loss on non-recombining sex chromosomes[19].

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 allows us to estimate a minimum age for the sex chromosome system. To accomplish this, we

used a phylogenomic approach with stringent inclusion criteria. We built 744 gene trees, 402 with U and V-linked homologs. We found most genes became sex-linked in the *C. purpureus* lineage, after the divergence from *Syntrichia princeps* (μ *Ks*=0.16; Fig. 3; Supplementary Table 7). However, 13 U-V orthologous pairs diverged at the base of the Dicranidae (μ *Ks*=0.85), and three pairs diverged prior to 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 prior to the split between *Buxbaumia aphylla* and the remaining Bryopsida, ~300 MYA (based on previous fossil-calibrated, relaxed-clock analyses[20]; *Ks*=2.8; Extended Data Fig. 2), making the moss sex chromosomes among the oldest known to date across Eukarya.

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*[21]. However, on the *C. purpureus* sex chromosomes we found *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 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 most homologs are from Ancestral Elements D and B (Supplementary Table 7) indicating the missing homeologous chromosomes to *C. purpureus* 5 and 9, respectively, had fused to the sex chromosomes (Fig. 3), but the scrambling of gene order had rendered them undetectable using synteny alone.

To extend this ancestral reconstruction to liverwort sex chromosomes, we examined gene trees of sex-linked genes in *M. polymorpha*[17,22]. Like in *C. purpureus*, we found no evidence of syntenic strata when we compared *Ks* between the U and V-linked orthologs (Supplementary Table 8). We also found evidence of four liverwort-specific capture events, with

the oldest diverging ~400 MYA, prior to the split of Marchantiidae and Pelliidae[20] (Fig. 3; Extended Data Fig. 2). Most sex-linked genes in *M. polymorpha* (Supplementary Table 8), like two of the oldest genes in *C. purpureus* (Supplementary Table 7) have homologs from moss Ancestral Element A, leading to the remarkable suggestion that this element played a key role in sex determination early in the history of both lineages, ~500 MYA.

A key factor explaining the retention of transcripts on non-recombining sex chromosomes is broad expression[23,24], which in plants include the haploid phase. In transcriptomic data from multiple tissues we found more than 1,700 U and V-linked genes expressed (mean count ≥1) (Supplementary Tables 9-11), including essential components of the cytoskeleton (e.g., Tubulin) and DNA repair complexes (e.g., *RAD51*). In fact, based on an analysis of sex-biased gene expression, we found far more sex-specific sex-linked genes (i.e., only on the U or V) than significant autosomal genes (mean count ≥1, fold change ≥2, adjusted $P \leq 0.05$), suggesting sex-linked loci contribute more to expression differences between the sexes than do autosomes (Extended Data Table 1). Furthermore, in contrast to data from gene-poor sex chromosome systems, we found nearly all the genes in the male and female co-expression modules, including the hubs, are sex-linked (Extended Data Fig. 3; Supplementary Table 12).

The sex-specific gene expression networks are enriched for proteins with known reproductive functions across green plant lineages. For example, the male GO and KEGG terms are enriched for microtubule-based processes, which play a role in sperm production in other systems[25,26] (Extended Data Fig. 3; Supplementary Tables 13-14). We also found both male and female co-expression modules are enriched for genes involved in circadian rhythm, like phytochrome, which are involved in flower development in *Arabidopsis thaliana*[27]. The male co-expression module also contained a V-specific *ABC1* gene orthologous to a V-linked copy in *M. polymorpha* (Extended Data Fig. 4), and genes in this family are involved with pollen

development in angiosperms[28]. The female co-expression module contains a U-specific *RWP-RK* TF orthologous to *M. polymorpha MpRKD*, which is a component of the egg development pathway in *M. polymorpha* and *A. thaliana* and mating-type loci in green algae[16,29,30] (Extended Data Fig. 4). The cis-acting sexual dimorphism switch *MpFGMYB*[31], which promotes female development in *M. polymorpha* has orthologous U and V-linked copies in *C. purpureus* (Extended Data Fig. 4). Additionally, several other TFs or transcriptional regulators (TR) are found in the sex-specific co-expression modules (e.g., V-linked *R2R3-MYB*) or are only found on the U or V (e.g., *HD DDT*, *Med7*, and *SOH1*; Supplementary Table 15), together suggesting candidate regulators of sex-specific developmental processes are enriched on the *C. purpureus* UV sex chromosomes.

Conclusions: 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, resulting 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[23]. However, unlike flowering plants, where hermaphroditism is the norm[32], the antiquity of dioecy in bryophytes more closely mirrors the sexual systems in animals[33,34]. 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.

References:

1. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 1563–1572 (2000).
2. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).
3. Bachtrog, D. *et al.* Are all sex chromosomes created equal? *Trends Genet.* **27**, 350–357 (2011).
4. Carey, S. B., Kollar, L. M. & McDaniel, S. F. Does degeneration or genetic conflict shape gene content on UV sex chromosomes? *EcoEvoRxiv* (2020) doi:10.32942/osf.io/hs6w3.
5. Bull, J. J. *Evolution of sex determining mechanisms*. (The Benjamin/Cummings Publishing Company, Inc., 1983).
6. McDaniel, S. F., Atwood, J. & Burleigh, J. G. Recurrent evolution of dioecy in bryophytes. *Evolution* **67**, 567–572 (2013).
7. McDaniel, S. F., Willis, J. H. & Shaw, A. J. A linkage map reveals a complex basis for segregation distortion in an interpopulation cross in the moss Ceratodon purpureus. *Genetics* **176**, 2489–2500 (2007).
8. Fritsch, R. Index to bryophyte chromosome counts. (1991).
9. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
10. Szövényi, P. *et al.* De novo assembly and comparative analysis of the Ceratodon purpureus transcriptome. *Mol. Ecol. Resour.* **15**, 203–215 (2015).
11. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
12. Lang, D. *et al.* The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).
13. Schaeffer, S. W. 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).
14. Bergero, R. & Charlesworth, D. The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* **24**, 94–102 (2009).
15. Montgomery, S. A. *et al.* Chromatin Organization in Early Land Plants Reveals an Ancestral Association between H3K27me3, Transposons, and Constitutive Heterochromatin. *Curr. Biol.* **30**, 573–588.e7 (2020).
16. Ferris, P. *et al.* Evolution of an expanded sex-determining locus in Volvox. *Science* **328**, 351–354 (2010).
17. Bowman, J. L. *et al.* Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell* **171**, 287–304.e15 (2017).
18. Ahmed, S. *et al.* A haploid system of sex determination in the brown alga Ectocarpus sp. *Curr. Biol.* **24**, 1945–1957 (2014).
19. Immler, S. & Otto, S. P. The evolution of sex chromosomes in organisms with separate haploid sexes. *Evolution* **69**, 694–708 (2015).
20. Laenen, B. *et al.* Extant diversity of bryophytes emerged from successive post-Mesozoic diversification bursts. *Nat. Commun.* **5**, 5134 (2014).

21. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
22. Materials and methods are available as supporting material on *Science* Online.
23. Chibalina, M. V. & Filatov, D. A. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* **21**, 1475–1479 (2011).
24. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
25. Pazour, G. J., Dickert, B. L. & Witman, G. B. The DHC1b (DHC2) isoform of cytoplasmic dynein is required for flagellar assembly. *J. Cell Biol.* **144**, 473–481 (1999).
26. Koshimizu, S. *et al.* Physcomitrella MADS-box genes regulate water supply and sperm movement for fertilization. *Nat Plants* **4**, 36–45 (2018).
27. Endo, M., Nakamura, S., Araki, T., Mochizuki, N. & Nagatani, A. Phytochrome B in the mesophyll delays flowering by suppressing FLOWERING LOCUS T expression in Arabidopsis vascular bundles. *Plant Cell* **17**, 1941–1952 (2005).
28. Kuromori, T., Ito, T., Sugimoto, E. & Shinozaki, K. Arabidopsis mutant of AtABCG26, an ABC transporter gene, is defective in pollen maturation. *J. Plant Physiol.* **168**, 2001–2005 (2011).
29. Rövekamp, M., Bowman, J. L. & Grossniklaus, U. Marchantia MpRKD Regulates the Gametophyte-Sporophyte Transition by Keeping Egg Cells Quiescent in the Absence of Fertilization. *Curr. Biol.* **26**, 1782–1789 (2016).
30. Tedeschi, F., Rizzo, P., Rutten, T., Altschmied, L. & Bäumlein, H. RWP- RK domain-containing transcription factors control cell differentiation during female gametophyte development in Arabidopsis. *New Phytol.* **213**, 1909–1924 (2017).
31. Hisanaga, T. *et al.* A cis-acting bidirectional transcription switch controls sexual dimorphism in the liverwort. *EMBO J.* **38**, (2019).
32. Renner, S. S. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am. J. Bot.* **101**, 1588–1596 (2014).
33. Eppley, S. M. & Jesson, L. K. Moving to mate: the evolution of separate and combined sexes in multicellular organisms. *J. Evol. Biol.* **21**, 727–736 (2008).
34. Sasson, D. A. & Ryan, J. F. A reconstruction of sexual modes throughout animal evolution. *BMC Evol. Biol.* **17**, 242 (2017).
35. Norrell, T. E., Jones, K. S., Payton, A. C. & McDaniel, S. F. Meiotic sex ratio variation in natural populations of Ceratodon purpureus (Ditrichaceae). *Am. J. Bot.* **101**, 1572–1576 (2014).
36. McDaniel, S. F. & Shaw, A. J. Selective sweeps and intercontinental migration in the cosmopolitan moss Ceratodon purpureus (Hedw.) Brid. *Mol. Ecol.* **14**, 1121–1132 (2005).
37. Xiao, C.-L. *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
38. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
39. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
40. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
42. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
43. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods Mol. Biol.* **1418**, 283–334 (2016).
44. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
45. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
46. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–10 (2012).
47. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
48. Mamidi, S. *et al.* A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **38**, 1203–1210 (2020).
49. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
50. Banks, J. A. *et al.* The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960–963 (2011).
51. Merchant, S. S. *et al.* The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007).
52. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. (2015).
53. Smit, A. F. A., Hubley, R. & Green, P. RepeatModeler Open-1.0. 2008--2015. *Seattle, USA: Institute for Systems Biology. Available from: httpwww. repeatmasker. org, Last Accessed May* **1**, 2018 (2015).
54. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516–522 (2000).
55. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
56. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
57. Lovell, J. T. *et al.* The genomic landscape of molecular responses to natural drought stress in Panicum hallii. *Nat. Commun.* **9**, 5213 (2018).
58. Zwaenepoel, A. & Van de Peer, Y. wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).
59. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
60. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
61. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

62. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
63. Proost, S. *et al.* i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11–e11 (2012).
64. Tiley, G. P., Barker, M. S. & Burleigh, J. G. Assessing the Performance of Ks Plots for Detecting Ancient Whole Genome Duplications. *Genome Biol. Evol.* **10**, 2882–2898 (2018).
65. Rensing, S. A. *et al.* An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evol. Biol.* **7**, 130 (2007).
66. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
67. Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, (2005).
68. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
69. Hoede, C. *et al.* PASTEC: an automatic transposable element classification tool. *PLoS One* **9**, e91929 (2014).
70. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
72. Team, R. C. R: a language and environment for statistical computing (version 3.5. 3)[software]. (2019).
73. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
74. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
75. McKnight, P. E. & Najab, J. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* 1–1 (2010).
76. Lang, D. *et al.* 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).
77. Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae. *Genome Biol. Evol.* **9**, 3384–3397 (2017).
78. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
79. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
80. Olney, K. C., Brotman, S. M., Valverde-Vesling, V., Andrews, J. & Wilson, M. A. Aligning RNA-Seq reads to a sex chromosome complement informed reference genome increases ability to detect sex differences in gene expression. *BioRxiv* 668376 (2019).
81. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

82. Love, M., Anders, S. & Huber, W. Differential analysis of count data--the DESeq2 package. *Genome Biol.* **15**, 10–1186 (2014).
83. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
84. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
85. Csardi, G., Nepusz, T. & Others. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1–9 (2006).
86. Briatte, F. ggnetwork: Geometries to Plot Networks with'ggplot2'. *R package version 0. 5* **1**, (2016).
87. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version* **2**, 2010 (2010).
88. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
89. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
90. Alaba, S. *et al.* The liverwort Pellia endiviifolia shares microtranscriptomic traits that are common to green algae and land plants. *New Phytol.* **206**, 352–367 (2015).
91. Johnson, M. G., Malley, C., Goffinet, B., Shaw, A. J. & Wickett, N. J. 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).
92. Gao, S. *et al.* Cloning and functional characterization of a phenolic acid decarboxylase from the liverwort Conocephalum japonicum. *Biochem. Biophys. Res. Commun.* **481**, 239–244 (2016).
93. Li, F.-W. *et al.* Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat Plants* **4**, 460–472 (2018).
94. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
95. Haas, B., Papanicolaou, A. & Others. TransDecoder (find coding regions within transcripts). *Github, nd https://github. com/TransDecoder/TransDecoder (accessed May 17, 2018)* (2015).
96. Bateman, A. *et al.* Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27**, 260–262 (1999).
97. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
98. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
99. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
100. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
101. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

102. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–12 (2006).
103. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
104. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
105. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. 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).
106. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
107. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
108. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
109. Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).
110. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
111. Stenico, M., Lloyd, A. T. & Sharp, P. M. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Res.* **22**, 2437–2446 (1994).
112. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).
113. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).
114. Luo, M. & Wing, R. A. An improved method for plant BAC library construction. *Methods Mol. Biol.* **236**, 3–20 (2003).
115. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Res.* **8**, 175–185 (1998).
116. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
117. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
118. Perroud, P.-F. *et al.* The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J.* **95**, 168–182 (2018).
119. Cove, D. J. *et al.* Culturing the moss Physcomitrella patens. *Cold Spring Harb. Protoc.* **2009**, db.prot5136 (2009).
120. Leder, E. H. *et al.* 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).
121. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
122. Yuan, Y.-W. & Wessler, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7884–7889 (2011).
123. Tsubouchi, H. & Roeder, G. S. The Mnd1 protein forms a complex with hop2 to promote

439 homologous chromosome pairing and meiotic double-strand break repair. *Mol. Cell. Biol.*
440 **22**, 3078–3088 (2002).

441

Acknowledgements: The authors thank Ralph Quatrano, David Cove, and the Quatrano laboratory at Washington University in St. Louis for incubating the *C. purpureus* genome project; Bernie Hauser, Thomas Colquhoun, and Drake Garner for assisting with the hormone and light perturbations; Bernard Goffinet, Michelle Mack, Sharon Robinson, Todd Rosenstiel, Blanka Shaw, and the late Norton Miller for providing field collections. Susanne Renner, Sally Otto, Mark Kirkpatrick, and Matthew Hahn provided valuable feedback on a draft of the manuscript. The One Thousand Plant Transcriptome Initiative generously 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. This work was supported by NSF DEB-1541005 and start-up funds from UF to SFM; microMORPH Cross-Disciplinary Training Grant, Sigma-Xi Grant-In-Aid of Research, and Society for the Study of Evolution Rosemary Grant Award to SBC; NSF DEB-1239992 to NJW; Emil Aaltonen Foundation and the University of Turku to SO; NSF DEB-1541506 to JGB and SFM. 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: SFM; Data curation: SBC, KB; Formal analysis: SBC, JJ, SS, JTL, FM, AS, GPT, NFP, SAR; Funding acquisition: SBC, SH, NJW, JGB, SAR, SFM; Investigation: SBC, JJ, ACP, SS, JTL, FM, AS, GPT, NFP, CC, MW, AL, CD, CS, JCM, REC, LMK, SO, SH, JBL, JGB, SAR, SFM; Methodology: SBC, ACP, JJ, JS, SFM; Project administration: SBC, KB, JG, JS, SFM; Resources: NJW, MJG, JG, JS, SFM; Supervision: MGJ, SAR, JG, JS, SFM; Visualization: SBC, JJ, JTL, AS, GPT; Writing – original draft: SBC, SFM; Writing – review and editing: SBC, ACP, JJ, JTL, FM, AS, GPT, CAS, LMK, JGB NJW, MGJ, SAR, JG, JS, SFM. All authors approve of the final draft of this manuscript.

Competing interests: The authors declare no competing interests.

Data availability: DNA and RNA data for this manuscript can be found under NCBI BioProjects listed in Supplementary Table 5. 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 transposable element has been deposited in NCBI GenBank under MT647524. The published sequence data used in this study can be found in Supplementary Table 16. 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: Materials and correspondence should be addressed to stuartmcdaniel@ufl.edu.

Figures

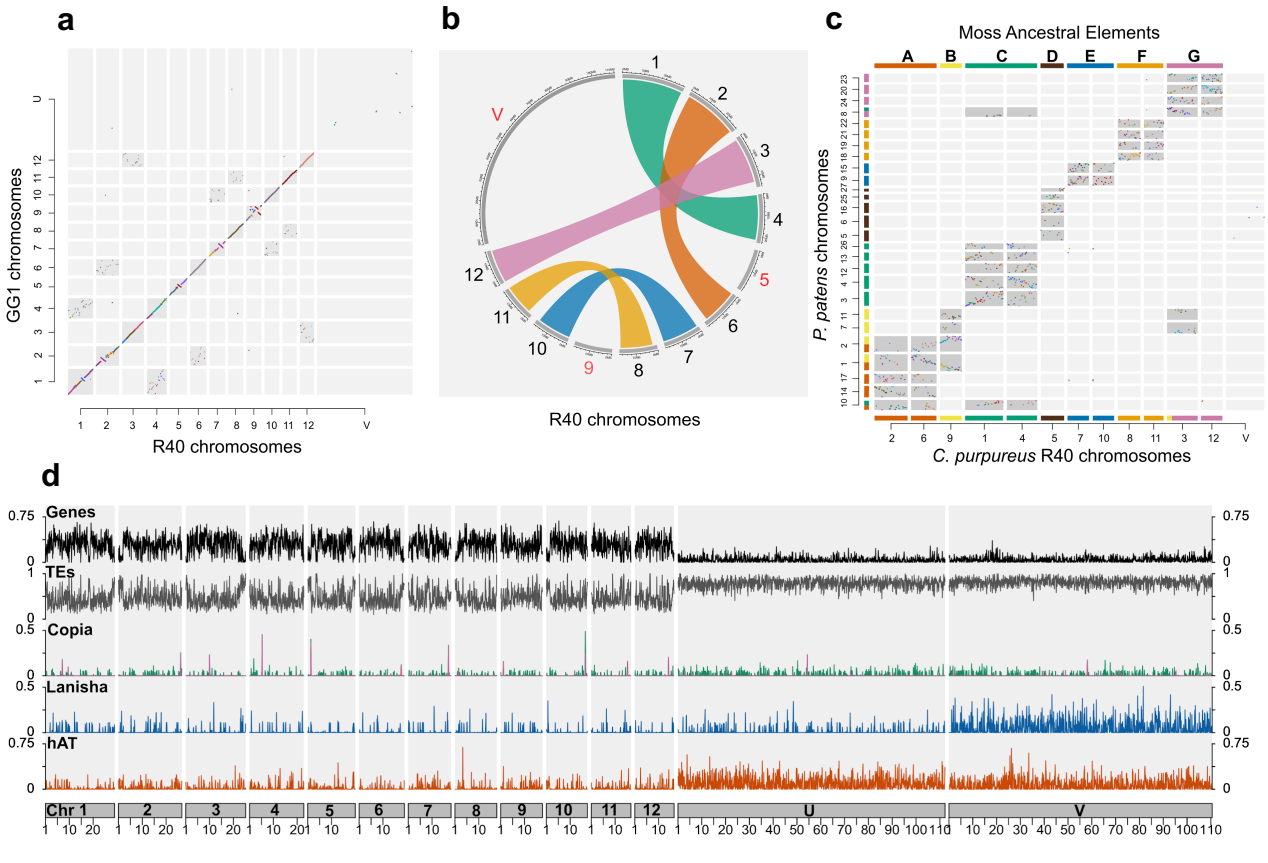


Fig. 1. Chromosome architecture in *C. purpureus*. a, Dot plot of syntenic orthogroup blastp hits between *C. purpureus* GG1 and R40 isolates, showing structural variation on autosomes and a lack of synteny across the sex chromosomes. **b,** Self-synteny plot of *C. purpureus* R40 isolate showing homeologous chromosomes from a whole-genome duplication. **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-G. **d,** Density plots across *C. purpureus* chromosomes (in Mb). Densities show the proportion of a 100 Kilobase (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*[12].

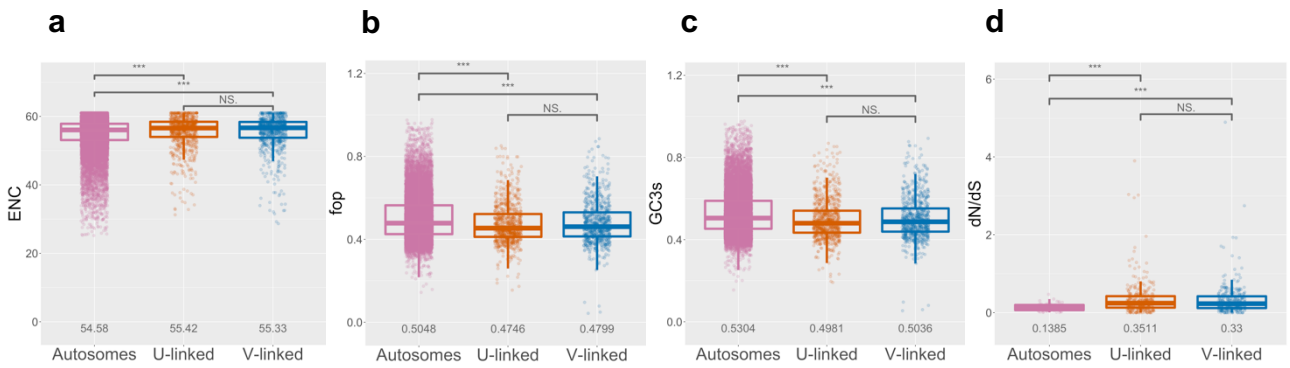


Fig. 2. Molecular evolution of autosomal and sex-linked genes in *C. purpureus*. a, Autosomal genes are significantly different from U or V-linked genes in the effective number of codons (ENC) **b,** Frequency of optimal codons (fop) **c,** GC content of the third, synonymous codon (GC3s) and **d,** Protein evolution (dN/dS) (*MWU,* autosomes to U or V $P<6x10^{-6}$ 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), suggesting weak but not significantly different degeneration on the U and V.

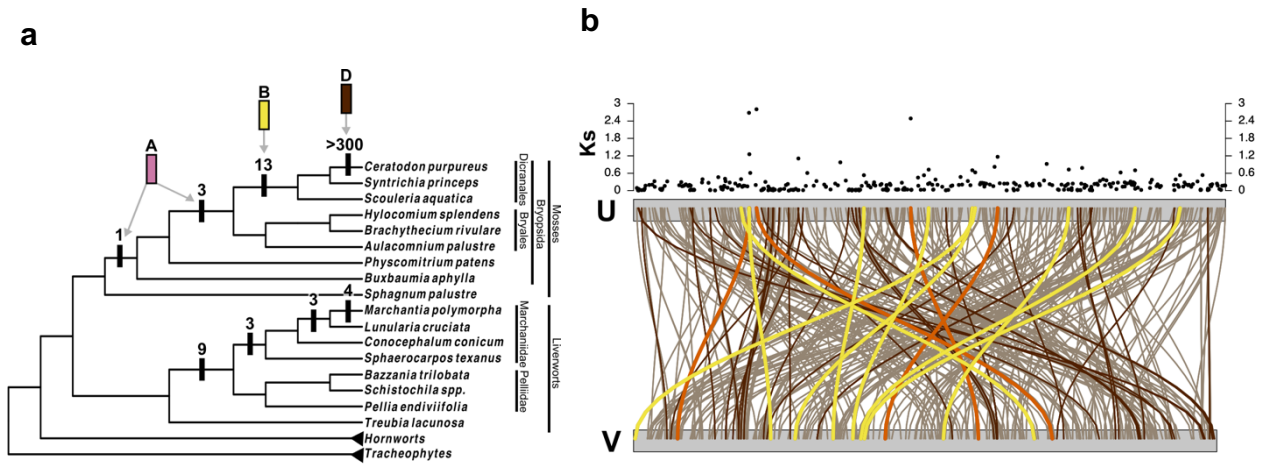


Fig. 3. Evolutionary history of moss and liverwort sex chromosomes. a, Capture events of genes on moss and liverwort sex chromosomes. Numbers indicate how many extant genes were captured at the indicated branch based on the topology of the tree. The capture events in mosses can be traced back to three 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*≤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 pseudoautosomal region between the *C. purpureus* U and V.

Methods

Isolate collection and tissue culture. All *C. purpureus* tissue used in this study was isolated from a single-spore[35], from field-collected sporophytes (Supplementary Table 5)[10,36]. In-depth methods for tissue generation for DNA and RNA, library preparation, and sequencing can be found in the Supplementary 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 (DOE) Joint Genome Institute (JGI) 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 400bp insert 2x150 Illumina fragment library (133.14x for GG1, 146.45x for R40) was sequenced along with one 2x150 Dovetail HiC library (252.86x GG1, 442.71x R40) (Supplementary Table 1). Prior to assembly, Illumina fragment reads were screened for phix contamination. Reads composed of >95% simple sequence were removed. Illumina reads <50bp after trimming for adapter and quality (q<20) were removed. For the PacBio sequencing, a total of 8 PB 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 113.77x (GG1) and 132.11x (R40) (Supplementary Table 2). Finally, a total of 1,032 BAC clones sequenced with Illumina indexed libraries were used for patching the final chromosome gaps.

566 *Genome assembly and construction of pseudomolecule chromosomes:* Improved versions 1.0
567 of the *C. purpureus* (var. GG1 and var. R40) assemblies were generated by assembling the
568 4,195,510 PacBio GG1 reads (113.77x sequence coverage) and 5,238,148 PacBio reads R40
569 (132.11x sequence coverage) separately using the MECAT assembler[37] and subsequently
570 polished using QUIVER[38]. For GG1, this produced 637 scaffolds (637 contigs), with a contig
571 N50 of 1.2 Mb, 475 scaffolds larger than 100 Kb, and a total genome size of 347.1 Mb
572 (Supplementary Table 3). For R40, this produced 731 scaffolds (731 contigs), with a contig N50
573 of 1.1 Mb, 497 scaffolds larger than 100 Kb, and a total genome size of 361.3 Mb
574 (Supplementary Table 3).
575 Hi-C scaffolding using the JUICER pipeline[39] was used to identify misjoins in the initial
576 MECAT assembly. Misjoins were characterized as a discontinuity in the GG1 or R40 linkage
577 group. A total of 73 misjoins were identified and resolved in GG1 and 64 in R40. The resulting
578 broken contigs were then oriented, ordered, and joined together into 13 chromosomes (12
579 autosomal and 1 sex chromosome designated as “U” in the GG1 release and 12 autosomal and
580 1 sex chromosome designated as “V” in the R40 release) using both the map, as well as the Hi-
581 C data. A total of 579 joins were made in GG1 and 625 in R40 during this process. Each
582 chromosome join is padded with 10,000 Ns. Significant telomeric sequence was identified using
583 the $(TTTAGGG)_n$ repeat, and care was taken to make sure that it was properly oriented in the
584 production assembly. The remaining scaffolds were screened against bacterial proteins,
585 organelle sequences, GenBank nr and removed if found to be a contaminant. For GG1, a set of
586 1,032 BAC clones (107.8 Mb total sequence) sequenced with Illumina indexed libraries were
587 used to patch remaining gaps in the chromosomes. Clones were aligned to the chromosomes
588 using BLAT[40], and clone contigs crossing gaps were used to form patches. A total of 35 gaps
589 were patched.
590 Finally, homozygous SNPs and INDELs were corrected in the release consensus
591 sequence using ~88x of Illumina reads (2x150, 400bp insert) by aligning the reads using bwa

592 mem[41] and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool[42].
593 A total of 108 homozygous SNPs and 5,291 homozygous INDELs in GG1 and 19 homozygous
594 SNPs and 867 homozygous INDELs in R40 were corrected in the release. The final version 1.0
595 GG1 release contains 349.5 Mb of sequence (1.3% gap), consisting of 558 contigs with a contig
596 N50 of 1.4 Mb and a total of 97.9% of assembled bases in chromosomes. The final version 1.0
597 R40 release contains 358.0 Mb of sequence (1.2% gap), consisting of 601 contigs with a contig
598 N50 of 1.4 Mb and a total of 98.3% of assembled bases in chromosomes.
599 Completeness of the euchromatic portion of the version 1.0 GG1 and 1.0 R40
600 assemblies was assessed by aligning an rnaSEQ library (library code GNGZB for GG1, GNGZC
601 for R40). The aim of this analysis is to obtain a measure of completeness of the assembly,
602 rather than a comprehensive examination of gene space. The transcripts were aligned to the
603 assembly using GSNAP[43]. The alignments indicate that 96.88% of the GG1 RNAseq reads
604 aligned to the version 1.0 GG1 release and 97.01% of the R40 RNAseq reads aligned to the
605 version 1.0 R40 release.
606
607 *Construction of the scaffold assembly:* A total of 4,195,510 PacBio reads (113.77x) in GG1 and
608 5,238,148 PacBio reads (132.11x) in R40 were assembled using MECAT[37], and formed the
609 starting point of the version 1.0 release for each. The 310,662,272 Illumina sequence reads
610 (133.14x sequence coverage) in GG1 and 353,932,084 Illumina sequence reads (146.45x
611 sequence coverage) in R40 were used for fixing homozygous snp/indel errors in the consensus.
612 A total of 310,662,272 Hi-C reads (252.86x sequence coverage) in GG1 and 1,062,837,932 Hi-
613 C reads (442.71x sequence coverage) in R40 were used for chromosome construction.
614
615 *Screening and final assembly release:* Scaffolds that were not anchored in a chromosome were
616 classified into bins depending on sequence content. Contamination was identified using blastn
617 against the 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 24mers that occur more than 4 times in the genome) (16 scaffolds, 482.8 Kb), chloroplast (1 scaffold, 158.7 Kb), and low quality (>50% unpolished bases post polishing, 3 scaffolds, 48.3 Kb) In R40, additional scaffolds were classified as repetitive (>95% masked with 24mers that occur more than 4 times in the genome) (12 scaffolds, 489.6 Kb), chloroplast (1 scaffold, 50.2 Kb), and low quality (>50% unpolished bases post polishing, 6 scaffolds, 236.8 Kb). Resulting final statistics are shown in Supplementary Table 4.

GG1 assessment of assembly accuracy: A set of 17 finished contiguous Sanger BAC clones >100Kb were selected in order 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% bp error) with an example being given in Supplementary Fig. 1. All dot plots were generated using Gepard[44]. The remaining 3 BACs indicate a higher error rate due mainly to their placement in more regions containing tandem repeats (Supplementary Fig. 2). The overall bp error rate in the BAC clones is 0.016% (269 discrepant bp out of 1,599,605 bp).

Genome Annotations. Transcript assemblies were made from ~1.5 billion pairs of 2x150 stranded paired-end Illumina RNA-seq reads from *C. purpureus* GG1 and ~1.6 billion pairs from *C. purpureus* R40 using PERTRAN (Shu, unpublished), 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 (Chile, Dur, Ecud, Ren, and UCONN; Supplementary Table 5) 180,954 (GG1) and 194,414 (R40) transcript assemblies were constructed using PASA[45] from RNA-seq transcript assemblies above and a bit of *C. purpureus* ESTs. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from *Arabidopsis*

644 *thaliana*[46], soybean[47], *Setaria viridis*[48], grape[49], *Sphagnum magellanicum*, *Physcomitrium*
645 *patens*[12], *Selaginella moellendorffii*[50], *Chlamydomonas reinhardtii*[51], filtered Trinity assembly
646 ORFs described above, high-confidence gene models from the first round of *C. purpureus* R40
647 gene call, uniprot Bryopsida and Swiss-Prot proteomes to repeat-soft-masked *C. purpureus*
648 GG1 genome using RepeatMasker[52] with up to 2K BP extension on both ends unless extending
649 into another locus on the same strand. Repeat library consists of De Novo repeats by
650 RepeatModeler [53] on *C. purpureus* GG1 genome and repeats in RepBase. Gene models were
651 predicted by homology-based predictors, FGENESH+[54], FGENESH_EST (similar to
652 FGENESH+, EST as splice site and intron input instead of protein/translated ORF), and
653 EXONERATE[55], PASA assembly ORFs (in-house homology constrained ORF finder) and from
654 AUGUSTUS via BRAKER1[56]. The best scored predictions for each locus are selected using
655 multiple positive factors including EST and protein support, and one negative factor: overlap
656 with repeats. The selected gene predictions were improved by PASA. Improvement includes
657 adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene
658 model proteins were subject to protein homology analysis to above mentioned proteomes to
659 obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual
660 best hit) BLASTP score and protein coverage is the highest percentage of protein aligned to the
661 best of homologs. PASA-improved transcripts were selected based on Cscore, protein
662 coverage, EST coverage, and its CDS overlapping with repeats. The transcripts were selected if
663 its Cscore is larger than or equal to 0.5 and protein coverage larger than or equal to 0.5, or it
664 has EST coverage, but it's CDS overlapping with repeats is less than 20%. For gene models
665 whose CDS overlaps with repeats for more than 20%, its Cscore must be at least 0.9 and
666 homology coverage at least 70% to be selected. The selected gene models were subject to
667 Pfam analysis and gene models whose protein is more than 30% in Pfam TE domains were
668 removed and weak gene models. Incomplete gene models, low homology supported without

669 fully transcriptome supported gene models and short single exon (<300 bp CDS) without protein
670 domain nor good expression gene models were manually filtered out.
671
672 **Synteny analysis within *C. purpureus* and between *P. patens*.** We ran the default
673 GENESPACE pipeline[57] with a minimum block size of 5 genes and a maximum gap / search
674 radius of 15 genes. In short, GENESPACE runs orthofinder on synteny-constrained blastp hits.
675 This offers higher stringency when exploring highly diverged genomes (or ancient whole-
676 genome duplications) by removing high scoring, but randomly distributed, blast hits.
677
678 **Ks-plot analysis to identify the *C. purpureus* whole-genome duplication.** Whole-genome
679 duplications (WGD) were detected with conventional *Ks* plot analyses. We used the wgd
680 pipeline[58]. An all-by-all BLASTP search[59] was performed for the *C. purpureus* GG1 and R40
681 genomes as well as *P. patens* and *M. polymorpha*. Paralogs were clustered with MCL[60]. For
682 each cluster, all pairwise *Ks* estimates were obtained from PAM[61] with the GY94 model with
683 F3x4 equilibrium codon frequencies[62]. Hierarchical clustering was used to reduce redundant
684 comparisons and obtain node-averaged *Ks* estimates. This process was repeated for syntenic
685 paralogs too, which were obtained from I-ADHoRe v3.0 with default settings[63] based on all-by-
686 all BLASTP results. Orthologous gene divergences used reciprocal best BLASTP hits between
687 *C. purpureus* and *P. patens*.
688 Peaks in *Ks* plots can be identified visually, but we also applied mixture models that
689 were selected by the difference in BIC scores, such that a difference less than 3.2 is used as a
690 stopping criterion. Mixture models were implemented with the *bic.test.wgd* function available on
691 GitHub (https://github.com/gtiley/Ks_plots). Mixture models can be problematic in their
692 interpretation due to over-fitting, therefore we looked for peaks that were consistently detected
693 across models and the maximum *Ks* value allowed[64]. When analyzing all paralogs, a single
694 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 (Supplementary Table 6). The more consistent results in R40 imply more paralogs from this WGD event have survived on the V chromosome compared to the U chromosome. This WGD post-dates the divergence of *C. purpureus* and *P. patens* (Fig. S12). This is determined by visual inspection but agrees with previous analyses of WGD in both *C. purpureus* and *P. patens*[10–12]. 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 (Extended Fig. 1), which suggest slightly more recent WGD ages (Supplementary Table 6). However, analyses of syntenic paralogs from *P. patens* supported the presence of two WGDs following divergence from *C. purpureus* (Supplementary Table 6), similar to previous findings when using syntenic data[12] compared to all paralogs from genomic or transcriptomic data[11,65].

Ks-plot analyses are provocative of older WGD events that pre-date 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 lack of evidence from mixture models, and nor do 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"[11]) and an even older event shared by all mosses ("B2"[11]). Testing such 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.

Transposable element annotation. We combined 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)[66]. Parameters were set to consider repeats with at least 5 copies. We obtained a library of 4,699 consensus sequences that was filtered to keep only those that are found at least once as full length copy in the combined assembly and we retained 2,523 of them. This library of consensus sequences was then used as digital probe for whole genome annotation by the TEannot[67] 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[68]). The library of consensus sequences was classified using PASTEC followed by manual curation[69]. To improve classification, remote homology detection was performed using HH-suite3[70]. 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.)[71]. These results were plotted in R (v3.5.3)[72] using the package karyoploteR (v1.8.8)[73] and edited in Inkscape (v0.92.2) (https://inkscape.org/en/) (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 Mann-Whitney U test with a Benjamini and Hochberg correction for multiple tests[74,75] using the sliding window densities (n_{Auto}=2736, n_U=1247, n_V=1229).

Transcription factor and regulator annotation. Transcription associated proteins (TAPs) comprise transcription factors (TFs, acting in sequence-specific manner, typically by binding to cis-regulatory elements) and transcriptional regulators (TRs, acting on chromatin or via protein-protein interaction). We classified all *C. purpureus* proteins into 122 families and sub-families of TAPs by a domain-based rule set[76,77]. We compared this genome-wide classification with

746 relevant organisms. All proteins in which a domain was found are listed with their family
747 assignment. In cases when the domain composition does not allow an unambiguous
748 assignment they are assigned no_family_found.
749
750 **Gene expression and co-expression.** Gene expression and co-expression analyses were
751 done using three male-female sibling pairs ($n_{isolates}$=6, 3 of each sex) at gametophore and
752 protonemal stages (n_{stages}=2) in triplicate ($n_{replicates}$=3) (Supplementary Table 5; see
753 Supplementary Methods for details on tissue conditions). Raw reads were filtered for
754 contaminants and adapters removed using BBDuk (v38.00) (Bushnell,
755 http:\\bbtools.jgi.doe.gov). This included removing reads with 93% identity to human, mouse,
756 dog, or cat or align to common microbial references. Further filtering removed reads with any
757 'N's, an average quality of 10, or a length <50 or 33% of the full read length. Adapters were
758 trimmed and reads were right-quality-trimmed if quality was below 6. Paired-end reads were
759 split into forward and reverse reads (novaseq_FASTQ_de_interlacer.pl,
760 https://doi.org/10.5061/dryad.v41ns1rsm). Reads were further filtered for quality using
761 Trimmomatic (v0.36)[78] using leading and trailing values of 3, window size of 10, quality score of
762 30, and minimum length of 40. We assessed the quality of the remaining reads using fastqc
763 (v0.11.4) (Andrews, 2010).
764 Filtered reads were mapped using HISAT2 (v2.1.0)[79] to the *C. purpureus* R40 genome
765 (autosomes and V sex chromosome) concatenated with the GG1 U sex chromosome. We hard
766 masked the U chromosome for males, and the V for females, using Bedtools (v2.27.1)[71]
767 maskfasta[80]. Genes greater than 300bp were assembled using StringTie (v1.3.3)[81], gene counts
768 were extracted using StringTie's prepDE.py script
769 (http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual#deseq), and gene IDs renamed
770 (using mstrg_prep.pl, https://gist.github.com/gpertea/b83f1b32435e166afa92a2d388527f4b).
771 Only genes matching the original genome annotation file were used for co-expression analyses

below. To identify differentially-expressed genes (DEGs), we used DESeq2 (v1.22.2[82]) 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<1, and an adjusted p-value <0.05. For sex-linked genes, we calculated the mean normalized count across only males or females for protonema and gametophore separately. To identify which sex-linked genes were sex-specific we used the output from Orthofinder below.

Co-expression network construction and module detection: Weighted gene co-expression networks were constructed using the WGCNA R package (v1.69)[83] with genes expression data normalized using variance stabilizing transformation from the DESeq2 R package (v1.26.0)[84]. The data retained after filtering genes showing low expression levels (minimum read count = 6 and minimum total read count = 10) were used to construct co-expression network modules using the block-wise 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). And 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 module eigengene distance threshold of 0.25 was used to merge highly similar modules. Top 10 hub genes in each module were identified based on module membership. These co-expression 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)[85] and ggnetwork (v0.5.8)[86] R packages and in order to focus on the relevant gene pair relationships,

798 network depictions were limited to an adjacency threshold of 0.2 and the top 3000
799 edges/interactions between nodes/gene models.
800
801 *GO and KEGG pathway enrichment analysis:* Gene Ontology (GO) enrichment analysis was
802 carried out using topGO, an R Bioconductor package (v2.38.1)[87] with Fisher's exact test; only
803 GO terms with a P <0.05 were considered significant. To identify redundant GO terms, semantic
804 similarity among GO terms were measured using Wang's method implemented in the
805 GOSemSim, an R package (v2.12.1)[88]. KEGG [89] pathway enrichment analysis was performed
806 based on hypergeometric distribution test and pathways with P <0.05 were considered
807 enriched.
808
809 **Phylogenomic analyses of moss and liverwort sex chromosomes.** The genome and
810 transcriptome lines used for phylogenomic analyses can be found in Supplementary Table
811 16[11,12,17,50,90–93]. For all RNA seq data, we filtered for quality using Trimmomatic (v0.36)[78] using
812 leading and trailing values of 3, a window size of 10, a quality score of 30, and a minimum
813 length of 40. We assessed the quality of the remaining reads using fastqc (v0.11.4) (Andrews,
814 2010). To *de novo* assemble genes, we used Trinity (vr20170205-2.4.0)[94] following default
815 parameters (the exception being with C*. purpureus*, for which used –SS_lib_type RF). We next
816 determined the single best open reading frame using TransDecoder (v5.0.2)[95]. Our reading
817 frames were checked first against pFam (v32.0)[96] and if no hit was found the frame was
818 determined by Transdecoder. To reduce protein redundancy, we next ran our open reading
819 frames through CD-HIT (4.6.3)[97,98] using a 0.99 threshold.
820 We first found orthogroups for the in-frame genes using Orthofinder (v2.2.0)[99,100]. We
821 built gene trees for genes annotated on the *M. polymorpha* and *C. purpureus* sex chromosomes
822 by first filtering clusters for at least eight species present in the tree (orthogroup_filter.pl,
823 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)[101]. We back translated our alignments to DNA using pal2nal (v14)[102]. Alignments were filtered for column occupancy of 0.5 using trimal (v1.2)[103] and filtered to remove any sequences less than 300bp (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)[104] 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 which contained annotated U and V-linked genes and determined the most-distantly related species found in the same clade (e.g., Extended Data Fig. 2). 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)[105,106] in R (v3.5.3)[72] 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)[107] 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 a custom python script (physco_outgroup.py, https://doi.org/10.5061/dryad.v41ns1rsm), which used ETE3[108] to first intentify the sex-linked genes then finding the closest *P. patens* gene based on brach length. For these genes, we also determined if a paralogous *C. purpureus* chromosome 5 paralog was present and reported only those that clearly showed gene duplication, presumably from the whole-genome duplication event.

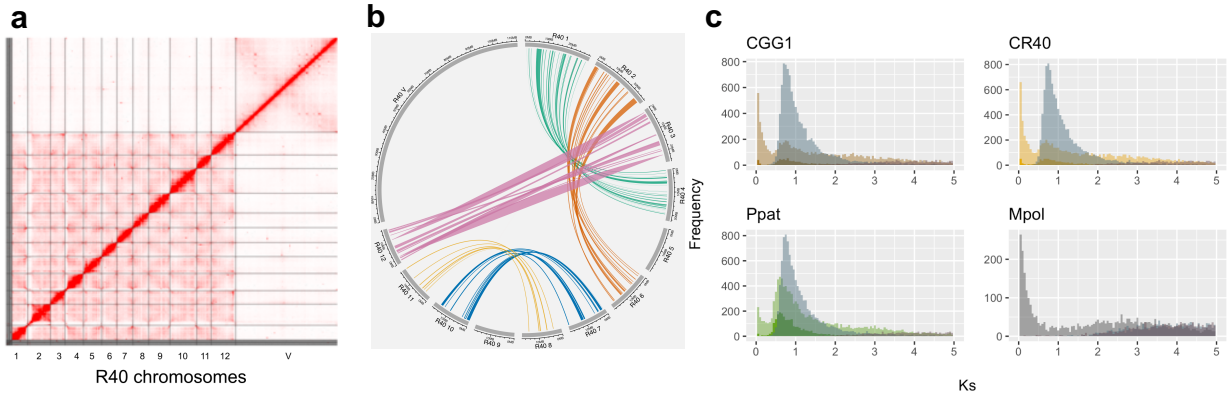
Protein evolution. To examine protein evolution of sex-linked and autosomal genes, we first pruned the trees described above at the closest *P. patens* homolog (prune_tree.py, https://doi.org/10.5061/dryad.v41ns1rsm)[108]. For genes that had a *C. purpureus* chromosome 5 homolog, the R40 and GG1 leaves 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[61]; 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 if there is a significant difference in dN/dS on autosomal, U, and V-linked genes we ran a pairwise Mann-Whitney U test with a Benjamini and Hochberg correction for multiple tests[74,75] (n_{auto}=61, n_U=314, n_V=315).

Ka 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)[109] using the Goldman and Yang model[62] on only one-to-one UV orthologs. *Ks* was plottred on the U and V sex chromosomes (Fig. 3) in R (v3.5.3)[72] using karyoploteR (v1.8.8)[73] 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).

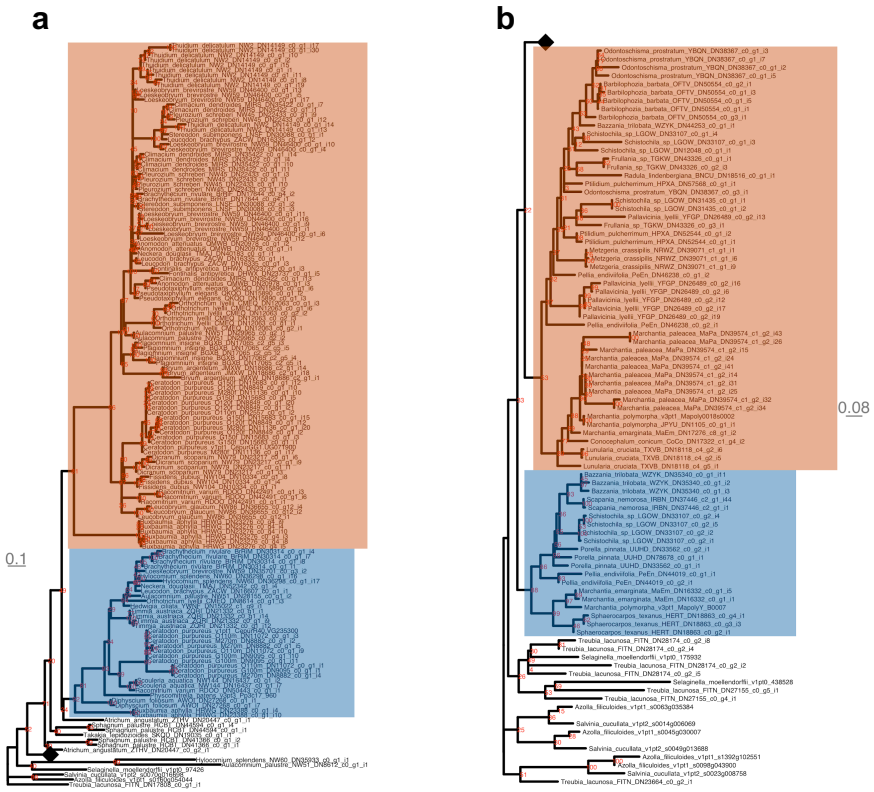
873

874 **Codon analyses.** To analyze codon-usage biases we used CodonW (v1.4.2; J. Penden
875 https://sourceforge.net/projects/codonw). We first removed any gene that had no expression to
876 remove potential pseudogenes (gene expression methods below). We also removed genes with
877 less than 200 codons to reduce the variance around calculated codon values using a custom
878 PERL script (alignment_length_filter.pl, https://doi.org/10.5061/dryad.v41ns1rsm[110]). We ran a
879 correspondence analysis on autosomes, U, and V-linked genes together to determine the
880 optimal codons in *C. purpureus.* We next determined the frequency of optimal codon usage
881 (fop[111]), effective number of codons (ENC), and GC content of the third synonymous position of
882 a codon (GC3s) on autosomes, U, and V-linked genes separately. To determine if there is a
883 significant difference between fop, ENC, and GC3s between autosomes, U, and V we ran a
884 pairwise Mann-Whitney U test with a Benjamini and Hochberg correction for multiple tests[74,75] in
885 R (v3.5.3[112]) (n_{Auto}=15,677, n_U=797, n_V=736) and plotted the results (Fig. 2) using ggplot2
886 (v3.2.1[113]) using default box-plot elements (ceratodon_genome_plots.R,
887 https://doi.org/10.5061/dryad.v41ns1rsm).
888

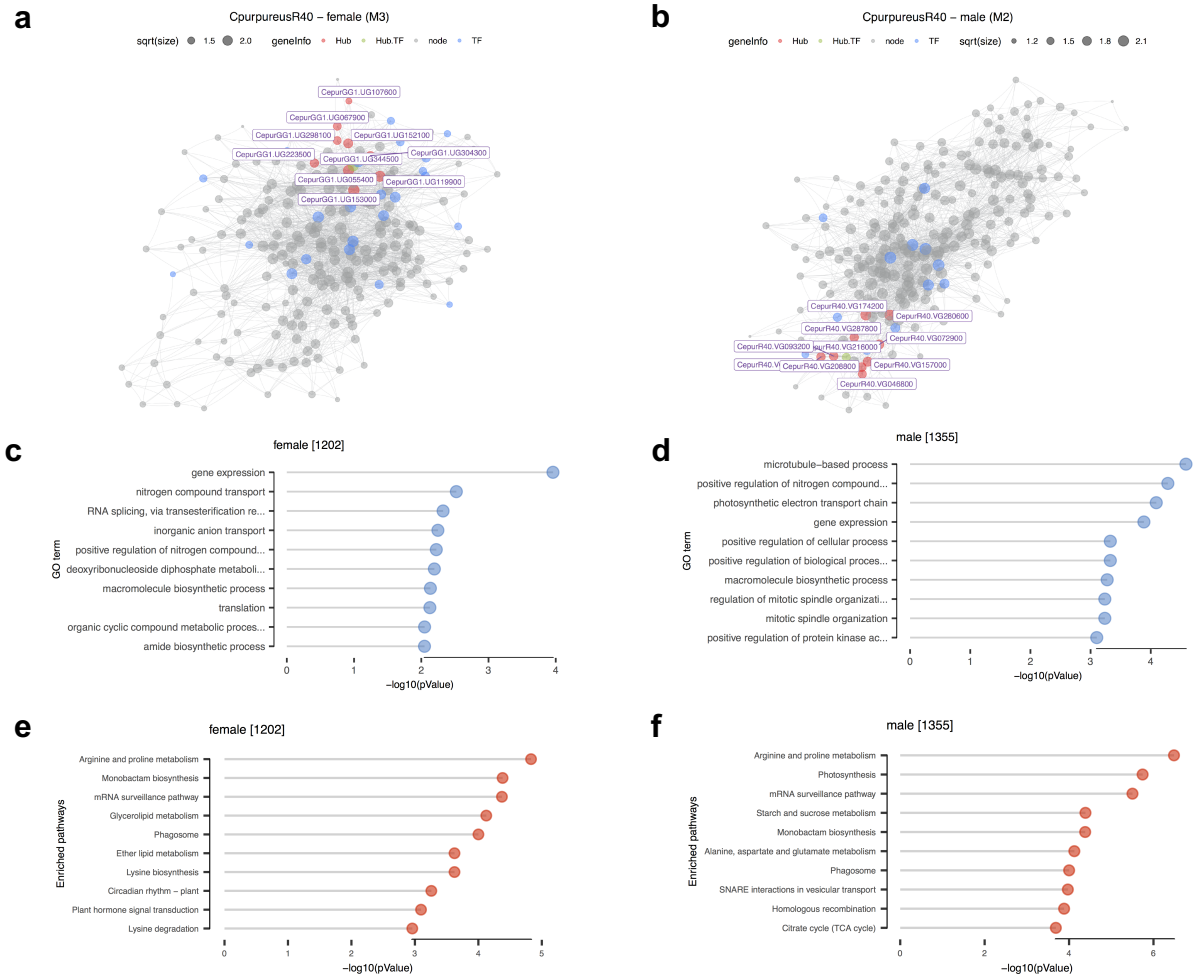
Extended Data



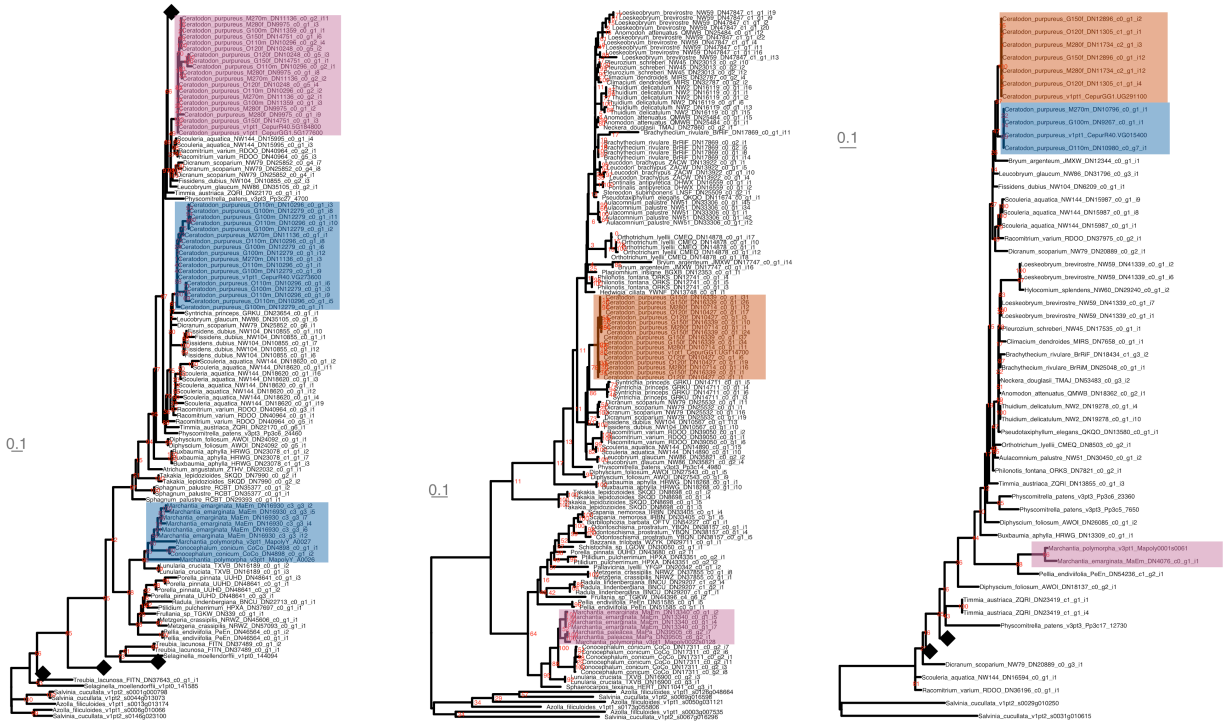
Extended Data Fig. 1. The *C. purpureus* genome assembly and whole-genome duplication analyses. a, Hi-C contact map for the *C. purpureus* R40 isolate, highlighting the chromosome-scale genome assembly. **b,** Collinear regions identified in R40, highlighting the homeologous chromosomes from a whole-genome duplication; synteny plots in GG1 show the same patterns (data not shown). **c,** *Ks* plots for *C. purpureus GG1* and R40, *P. patens*, and *M. polymorpha.* Node-averaged *Ks* values are shown for all paralogs (in the lighter color) as well as syntenic paralogs (the darker color). The third, blue-grey distribution are orthologs between *C. purpureus and P. patens*. The *M. polymorpha Ks* plot has no syntenic paralogs and shows *M. polymorpha – C. purpureus* orthologs in blue-grey as well *M. polymorpha – P. patens* orthologs in purple, which overlap and become indistinguishable.



Extended Data Fig. 2. Phylogeny of oldest sex-linked gene identified in mosses and liverworts. Trees were built using Maximum-likelihood and bootstrap support is shown in red. Trees were rooted using ferns (*Azolla* or *Salvinia*) as the outgroup. Black diamonds represent collapsed clades and the branch-length scale is in gray. The clades highlighted in blue are V-linked (male), orange are U-linked (female). **a,** Ancient sex-linked gene in mosses. The tree includes sex-specific data for *C. purpureus* and distantly-related *Brachythecium rivulare*. Evidence for the ancient sex-linkage includes isolates of the same sex being more closely related than the other sex of the same species (e.g., *C. purpureus* GG1 is more closely related to BrRiF and R40 more-closely related to BrRiM). **b,** One of the oldest sex-linked genes in liverworts. Eight other trees in liverworts show a similar topology (see Supplementary Table 8).



Extended Data Fig. 3. Co-expression, GO, and KEGG enrichments for female and male modules. a, Co-expression modules for females. **b,** Co-expression modules for males. **c,** female module GO enrichment. **d,** male module GO enrichment. **e,** female module KEGG enrichment. **f,** male module KEGG enrichment. HUBs for each module are identified in red and with the gene name shown. Transcription factors (TF) are in blue and nodes in gray.



Extended Data Fig. 4. Notable homologous genes in *C. purpureus* and *M. polymorpha*. Trees were built using Maximum-likelihood and bootstrap support is shown in red. Trees were rooted using ferns (*Azolla* or *Salvinia*) as the outgroup. Black diamonds represent collapsed clades and the branch-length scale is in gray. The clades highlighted in blue are V-linked (male), orange are U-linked (female) and purple are noteworthy autosomal homologs. **a,** An *ABC1* gene that is sex-linked in males of *C. purpureus* and *M. polymorpha.* The topology suggests this gene was independently captured between the two lineages. **b,**An *RWP-RK* gene that is sex-linked in females of *C. purpureus* and orthologous to *Marchantia MpRKD.* **c,** Gene tree of sex-linked genes in *C. purpureus* showing they are orthologous to *Marchantia MpFGMYB.*

977

978 **Extended Data Table 1. Sex-biased gene expression patterns in *C. purpureus* for**

979 **protonemal and gametophore stages.** Autosomal genes presented are significantly

980 differentially expressed at an adjusted P≤0.05. U and V-linked genes given are sex-specific (i.e.,

981 no evidence of a paralog on the other sex chromosome). Many homologous sex-linked genes

982 are also expressed, however, it is not possible to statistically test for differences in expression

983 between these because they are mapped to different chromosomes.

	Protonema		Gametophore	
	MC≥1	FC≥2; MC≥1	MC≥1	FC≥2; MC≥1
Male autosomal	2048	317	1307	256
Female autosomal	1900	366	1420	264
U	1006	-	1061	-
V	978	-	1054	-

984 Mean count (MC)

985 Fold change (FC)

986