The evolution of heteromorphic sex chromosomes in plants

AUTHORS

Sarah B. Carey[1], Philip C. Bentz[1], John T. Lovell[1], Laramie M. Akozbek[1,2], Joshua S. Havill[3], Lillian Padgitt-Cobb[4], Ryan C. Lynch[4], Nicholas Allsing[4], Austin Osmanski[1], Katherine A. Easterling[5], Leonardo R. Orozco[6], Haley Hale[1], Hannah McCoy[1], Zachary Meharg[1,3], John McKay[7], Jane Grimwood[1], Daniela Vergara[6], Rafael F. Guerrero[8], Nolan C. Kane[6], Todd P. Michael[4], Gary J. Muehlbauer[2], and Alex Harkess[1*]

AFFILIATIONS

[1] HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
[2]Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, AL, USA
[3] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN, USA
[4]The Plant Molecular and Cellular Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[5]United States Department of Agriculture, Agricultural Research Service, Wapato, WA, USA
[6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA
[7]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA
[8]Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA
[*]Author for correspondence: A.H., aharkess@hudsonalpha.org

SUMMARY

Ancient heteromorphic sex chromosomes are common in mammals, but not in plants. Sex chromosomes in the plant family Cannabaceae, which includes species like hops and hemp, were identified a century ago because of their obviously heteromorphic XYs. However, we know very little about their structure, nor their role in the development of the economically, medicinally, and culturally important flowers. Here we assembled genomes, with phased sex chromosomes, for three XY male hops (*Humulus* sp.), two XY male *Cannabis*, and two monoecious *Cannabis* individuals. Most of the non-recombining regions evolved prior to the divergence between *Cannabis* and *Humulus* around 47 MYA supporting this sex chromosome origin is among the oldest known in flowering plants. The Y chromosomes are highly-degenerate, with very few genes remaining, and show substantial rearrangements, while the X also shows faster rates of molecular evolution. The Cannabaceae sex chromosomes contain critical genes for flower development and the production of key compounds like bitter acids and cannabinoids, which are important traits for the plants. The sex chromosomes thus have played a vital role in the domestication and breeding of these species, and more generally, the Cannabaceae family.

MAIN TEXT

Since their discovery [1], sex chromosomes have been of keen interest to study because of their unique inheritance patterns that can drive differences in their size, structure, and gene content. A century ago, the sex chromosomes of *Humulus* and *Cannabis* were among the first to be identified in flowering plants [2,3]. *Humulus lupulus* var. *lupulus*, the European hop, has one of the only known heteromorphic XY pairs in plants, where the Y chromosome is smaller than the X. *Cannabis*, instead, has a Y chromosome bigger than the X, but this form of heteromorphy is also apparently rare across angiosperms; most examined plant sex chromosomes are homomorphic [4]. Curiously, the XY pairs across other members of *Humulus* differ greatly in their cytology with homomorphic (American hops) and multiple sex chromosomes ($X_1X_2Y_1Y_2$ in *H. lupulus* var. *cordifolius* and $X_1Y_1Y_2$ in *H. scandens* [5]). The shared origin of sex chromosomes [6] in Cannabaceae, in addition to their extreme structural diversity (heteromorphic, homomorphic, multiple, polyploid), makes it an ideal clade for examining the genomic causes and consequences of sex chromosome evolution in angiosperms and more generally across eukaryotes.

The inflorescences in the Cannabaceae family have a deep history of human use, of which the females (XX) are typically the economically important sex. *Cannabis sativa* (hereafter, *Cannabis)* was domesticated by humans ~12 thousand years ago [7] and is still cultivated globally for bast fiber in stems, multi-purpose oil in 'grain' (achenes/seeds), and medicinal compounds (cannabinoids) concentrated in glandular trichomes on female inflorescences [8]. Records from over 1,200 years ago support the use of female hops in brewing beer [9], where they provide the characteristic bitterness and aroma. Like *Cannabis*, secondary compounds in hops are concentrated in glandular trichomes on female inflorescences, thus females in both genera have been strongly selected upon over millennia of human domestication. Despite their importance to the domestication and productivity of hops and *Cannabis*, very little is known about the origin and sequence structure of Cannabaceae sex chromosomes. Such an obvious knowledge gap is largely due to basic challenges of building and exploring highly repetitive and large X and Y sex chromosomes [10]. Despite these difficulties, here we assemble fully-phased, chromosome-scale reference genomes with both X and Y chromosomes for *Humulus lupulus* and *Cannabis sativa*, which represent easily the most complete genomic resources that include both sex chromosomes for the important crops of the Cannabaceae. We use these genomes to examine patterns of sex chromosome evolution across the Cannabaceae that contribute to these sex-specific dynamics.

PHASED CANNABACEAE SEX CHROMOSOME ASSEMBLIES

Until recently, assembling Cannabaceae genomes was a challenge, because its species exhibit highly-heterozygous and repeat-rich genomes [11,12]. Further, sex chromosomes are particularly arduous to contiguously assemble [13]; for example, the Y chromosome was the last to be included in the human 'T2T' genome [10], and sex chromosomes are incomplete in many mammalian genomes on NCBI. While progress has been made to identify sex-linked sequences (e.g., [14,15]), to our knowledge no approach has been designed to assemble and validate perfectly-phased sex

chromosomes. Therefore, we developed a custom pipeline to phase the sex chromosomes that relies on the generation of sex-linked *k*-mers to identify X- versus Y-linked contigs (Supplementary Discussion).

To examine the structure and gene content of the Cannabaceae sex chromosomes, we built genomes that span their cytological range. This included males of three botanical varieties of *H. lupulus* found across the breeding pedigree [16]: a line of *H. lupulus* var. *lupulus* (USDA 21110M) and two wild, American relatives, *H. lupulus* vars. *lupuloides* and *neomexicanus*. We also built genomes for two male cultivars of *Cannabis* (Carmagnola and Otto II) and two monoecious *Cannabis* cultivars (Uso 31 and Futura 75). We used hifiasm with HiFi long reads to build contig assemblies and Omni-C incorporation to phase haplotypes [17] (Table S1). Prior to scaffolding, we used male-specific *k*-mers (henceforth, Y-mers [18]) and a haplotype-combined scaffolding approach to manually place the X/Y contigs into chromosomes (Table S2; Supplementary Discussion). After sex chromosome phase correction, in *Cannabis*, the Haplotype 1 (HAP1) and 2 (HAP2) assemblies are ~760.6-828.8 Mb (initial contig N50s: 22-43.8 Mb). The hop genomes are ~3 times larger at ~2.4-2.7 Gb (initial contig N50s: 56.1-167.6 Mb; Table S3). The final assemblies are consistent with estimated genome sizes [19–21] and we recovered the expected 10 chromosomes per haplotype in all individuals (Fig. S1-2). These seven new, haplotype-resolved genomes contain very few gaps (Fig. S3-4) and exhibit Merqury *k*-mer completeness values [22] of 92.95-98.31% (Table S3). We annotated Transposable Elements (TEs) using EDTA [23] and found the *Cannabis* genomes to be on average 74.3% repetitive (Fig. 1), consistent with recent assemblies [11]. Hops were higher for TEs at an average of 86.8%, an increase from previous assemblies [12], likely due to greater contiguity of repetitive regions (Table S4). Using a combination of RNAseq and protein homology, we annotated on average 34.9k gene models using BRAKER [24,25], with BUSCOs ranging between 93-97.2% (Table S3-4). The Y is the largest chromosome in the *Cannabis* genome at ~110.6-114 Mb, while the X is ~76% smaller at 84.4-87.4 Mb (Fig. 1). Similar to the increased overall genome size, the Xs of the three varieties of hops are ~3 times larger than in *Cannabis*, at ~267.4-272 Mb. The hop Y chromosomes show the most drastic difference. The wild-collected American isolates are cytologically homomorphic, where they are 89.3-92% the size of the X. The Y in *H. lupulus* var. *lupulus* is the smallest chromosome in its genome and the smallest hop Y at 168.7 Mb, making it ~62.9% the size of the X and fitting the approximate 1.5:1 ratio observed in the cytotype [5]. Thus, these sex chromosome assemblies meet the expected diverse range of cytological patterns (Fig. 1-2).

Sex chromosomes have unique inheritance patterns relative to autosomes. In an XY system, the non-recombining sex-determination region (SDR) of the Y chromosome is only inherited by males, while the remaining pseudoautosomal region (PAR) recombines relatively freely with homologous regions of the X (HXR) [26]. To delineate the boundary between the SDR and the PAR, we used a rigorous Y-mer mapping and gene tree approach. The hop and *Cannabis* Y chromosomes contain a single PAR that represents between 20-30% of the Y sequence, also matching cytological observations (Fig. 1) [27]. These near-perfect, phased sex chromosomes mark

a methodological advancement in the field of comparative genomics, enabling the generation of chromosome-scale assemblies from a single heterogametic sex individual without trio-binning (Supplemental Discussion).

TIMING AND EVOLUTION OF CANNABACEAE SEX CHROMOSOMES

Reconstructions of sexual systems suggest that dioecy (i.e., female and male flowers develop on separate individuals) is the ancestral state in Cannabaceae and its sister families Moraceae and Urticaceae, that contain figs, mulberries, and nettles [28]. Example species within these families have been shown to switch sex when ethylene is modified (e.g., ethephon treatment causes karyotypic males to produce female flowers [29,30]) suggesting the possibility of a shared sex-determining mechanism. Currently it is unknown if the two families share an ancestral XY pair; however, recent assemblies in Moraceae [31,32] and our phased XYs, provide an opportunity to test this and refine our understanding of the timing of the evolution of the Cannabaceae XYs.

To investigate the timing of stepwise changes in sex chromosome structure across the phylogeny, we used a phylogeny based on whole-plastome sequences from across Rosales and fossil records to calibrate and estimate a divergence time tree. Resulting age estimates were transposed onto a species tree, inferred using a multi-species coalescent summary approach [33] with 775 nuclear gene trees at concordant nodes (Fig. 3, S5). The ancestors of *Cannabis* and *Humulus* began diverging between 33.9-41.4 (95% CI; mean=36.38) MYA, whereas Cannabaceae diverged from Moraceae around 66-84.4 (95% CI; mean=74.2) MYA (Fig. 3, S5). We next used evidence from synonymous protein changes (Ks) between one-to-one XY orthologs (hereafter, gametologs) to estimate the SDR age. The maximum Ks in hops is 0.4 (mean Ks, top 10% = 0.325) and we estimate the age to be 39.1-48.8 MYA, firmly placing the SDR origin on the stem branch leading to the split between *Cannabis* and *Humulus*. Consistent with our age estimates and previous comparative analyses [6], our results support independent evolutions of sex chromosomes from different autosomes in Moraceae and Cannabaceae and using gene trees no topology suggests gene capture prior to their divergence. These analyses support that the Cannabaceae XYs are among the oldest identified in plants.

Since the homologous SDRs do not recombine, genes captured in the same event are expected to have similar levels of Ks, where the older captures will have higher Ks between the X and Y compared to younger captures (i.e., evolutionary strata [34]). Genes with shared captures prior to the divergence of *Cannabis* and *Humulus* will show a tree topology where all Y-linked genes form a clade, sister to the X-linked clade (Fig. S6). Attributable to the difference in generation times (1 year in *Cannabis*, 2-3 years hops), in *Cannabis*, the max Ks is twice as high as hops (Ks = 0.81, mean=0.34). Importantly, Ks in both genera is largely consistent across the entirety of the SDR or HXR (Fig. 1), suggesting most of the recombination suppression occurred in one event. We additionally found 41 OrthoGroups shared across all Y SDRs (Table S5) and for the majority we were able to construct a gene tree with a topology that supports these shared captures (Fig. S6). In addition to the shared stratum, there is evidence for independent gene captures in *Cannabis* and *Humulus*. In *Cannabis*, Ks along the SDR shows a clear, continuous

pattern of gene addition from the PAR boundary at 29.3 Mb to 35.7 Mb, suggesting that these 27 gametologs have been added into the SDR over time, rather than in a large SV (Fig. 1). This gradual expansion of the SDR is likely related to centromere proximity (Fig. S7) as the existing reduced recombination may facilitate sex-linkage [35,36]. In hops, we also found evidence for ~50 Y-linked genes that were captured in the American hops, that are clearly in the PAR in *H. lupulus* var. *lupulus* (Fig. S6). Thus we found the Cannabaceae SDRs have two regions: a shared region that was captured prior to the divergence of the two genera encompassing the majority of the SDR, and a region of independent captures. After their initial evolution, the XYs have continued to progress dynamically, resulting in the massive cytological variation extant today.

STRUCTURAL EVOLUTION OF CANNABACEAE SEX CHROMOSOMES

These phased sex chromosomes allow for tracking the complex structural variant landscape within and among species in the Cannabaceae, revealing PAR and SDR dynamics as they relate to the evolution of novel reproductive forms. Sex chromosomes often show increased SVs across all classes including inversions [37,38] and fusions that create neo-sex chromosomes [39]. Fusions between an autosome and a sex chromosome are theorized to be favored when the autosome contains sexually antagonistic loci, where the fusion increases linkage with the sex-determining locus [40]. Models of sex chromosome:autosome fusions suggest that they more often occur in the SDR or HXR than the PAR [41], but testing these models has been challenging due to the lack of fully-phased sex chromosome assemblies in many plants.

Most genes on the Cannabaceae sex chromosomes reside on the PARs and maintain a linear order between XY pairs within and across species, in stark contrast with the SDRs that have very little synteny remaining (Fig. 1). Some large structural variations (SV) have occurred following divergence of the two species; in hops, the X chromosome haplotypes have a large inverted translocation (Fig. 1), which likely caused the previously observed inconsistency where subtelomeric probes are interstitial compared to Xs on *H. scandens* and *Cannabis* [42,43]. This SV is especially apparent in the gene landscape on the hop Xs; most chromosomes contain gene dense arms towards the telomeres, but the SV has altered the location of this genic region inwards. Within hops, we also found the predicted centromere locations differ across varieties. The most striking difference in centromere location is between the X and Y of the *H. lupulus* var. *lupulus* isolate. For the X, the centromere is ~142 Mb from the PAR, supporting its submetacentric location, but the Y is acrocentric and ~20 Mb away (Fig. 4). This result is consistent with observations that the XY bivalent only associates by a terminal chiasma between the short arm of the Y and the long arm of the X during meiosis [5], unlike any other bivalent in the karyotype. Interestingly, the centromere location differs among hop varieties; in *lupuloides* and *neomexicanus* the X centromere changed locations, and was driven by an inversion (Fig. S7).

Sex chromosome evolution, and particularly large-scale incorporation of autosomal sequence, has co-occurred with the evolution of novelty in other systems [44]. We found a translocation of part of an autosome to the *Cannabis* XYs, increasing their size. This

chromosome appears to have undergone a centric fission (based on the putative centromere location in hops Chr01; Fig. 1, S7), then fused to the PAR of the *Cannabis* X and Y, forming a neo-XY region (Fig. 1). In some lineages of hops, a second set of autosomes have also become incorporated into the ancestral pair, forming neo-XYs. In *H. scandens*, cytological evidence supports a fusion to the X chromosome of an autosome, but the two Ys remain unfused and both show signatures of suppressed recombination [27]. Using F_{st} between the sexes and mapping of Y-mers in *H. scandens,* we identified an ~160 Mb region of Chr03 that supports being the second SDR of the Y_2 (Fig. S8) and differs from the chromosome fused in the *Cannabis* lineage. The fusion between Chr03 and the HXR in *H. scandens* could involve a recessive, female beneficial allele, in contrast with the *Cannabis* fusion, which occurred on the PAR end, and is expected to be rarer based on models and observations across known neo-XY systems [41]. Given that the PARs are a relatively small proportion (20-30%) of the Cannabaceae XYs (Fig. 1), both events could also be explained by the Fragile Y hypothesis [45], which describes that recombination within the PARs is expected to occur for the XY to properly pair at meiosis, and fusions of an autosome to either the PAR or SDR can extend the size of the recombining region, reducing the probability of aneuploidy.

In the closely-related and sympatric American hop varieties, *H. lupulus* vars. *lupuloides* and *neomexicanus* [46], the Ys have evolved substantial structural variation since they diverged ~0.24 MYA (Fig. 3). The Ys between these varieties have evolved ~4.65 times more translocations and ~4.34 times more inverted translocations than the autosomes, while the Xs are more similar to the autosomes (1.27 times more translocations and 0.8 times more inverted translocations; Fig. 1, Table S6). The Ys have ~⅓ fewer inversions compared to autosomes, but they are large, representing ~35% of the assembled Y sequence. Extending this to comparisons between the *H. lupulus* var. *lupulus* Y shows they are substantially less alignable than other regions of the genome (Fig. 1), consistent with the greater divergence time between var. *lupulus* and American hops (1.61 MYA; Fig. 3). In contrast to hops, the *Cannabis* Ys also contain SVs, but fewer than the autosomes. The most likely explanation for this difference is that the *Cannabis* isolates are more-closely related than the hops, consistent with the neutral evolution of SVs on Ys, though it could reflect differences in *Cannabis*, such as with long-term breeding practices.

Like SDRs, PARs are distinct regions of the genome that are often mistakenly thought to evolve and recombine freely like autosomes, but their partial linkage to the SDR can drive different evolutionary dynamics [26]. The PARs are also highly-structurally variable, especially in the var. *lupulus* isolate, where there are two large inversions (each ~9Mb) and a translocation (~4.8Mb) (Fig. 1); curiously these SVs co-locate at the PAR/SDR boundary, which may lead to the expansions and contractions of sex chromosomes over time. This suggests that the PAR is ~34.2 Mb of the var. *lupulus* XYs, but only ~⅓ can freely recombine. Importantly, these results highlight the massive variation within and between sex chromosome pairs and their associated PAR and SDR regions. One sex chromosome reference is likely not sufficient to describe inter- and intra-specific variation [47].

Y DEGENERATION AND FASTER-X OF ANCIENT PLANT SEX CHROMOSOMES

The suppressed recombination of an SDR reduces the efficacy of natural selection and can drive the accumulation of slightly deleterious mutations [48]. As a consequence, SDRs tend to accumulate TEs, which can quickly increase the size of the Y and over greater time can lead to gene loss and reduce its size [49,50]. If this linear trajectory holds true, then this process of Y degeneration should increase with time, but the Cannabaceae sex chromosomes defy these expectations as they have a shared origin with extreme cytological variation (Fig. 1, 2), supporting the trend that time alone does not drive these changes [4].

The sex chromosomes show clear signs of degeneration, where the SDR is gene poor and TE rich compared to the autosomes and X chromosome (Fig. 1). In *H. lupulus* var. *lupulus*, the SDR contains only 382 gene models, ~17% the number in the HXR, with similarly low proportions in the Ys of vars. *lupuloides* and *neomexicanus* (both ~20%; Table S7). In *Cannabis*, which has the largest Y relative to its X counterpart, there is also evidence for substantial degeneration (SDR models ~46% of the HXR), though to a lesser extent than in the hops. Given the variation in the physical sizes of the Ys (Fig. 1), one potential cause is through the differential loss of genes within the SDRs. However, in hops, which show the most extreme variation in size within a species, the number of XY gametologs are nearly the same, suggesting that the larger Ys in the wild hops have lost nearly as many genes as the smaller Y in *H. lupulus* var. *lupulus*, and that physical size difference is likely driven by transposon accumulation. Indeed, TE content of the SDR is 16% and 7% higher compared with the genome-wide averages for *Cannabis* and *Humulus*, respectively, and mostly LTR elements (Table S8). Recent bursts in TE activity are evident on the Y chromosomes compared to their X counterparts, consistent with the increased overall percentages (Fig. 4). Importantly, the *H. lupulus* var. *lupulus* Y shows reduced TE expansion relative to the other Ys since the SDRs evolved, which could explain its smaller size (Fig. 4). Together these results support that the Cannabaceae Y chromosomes are highly-degenerated and that the cytological variations in Y chromosome sizes are a complex, ongoing interplay between gene loss and repeat expansions.

The hemizygous state of the X chromosome in males is predicted to cause elevated rates of molecular evolution, a phenomenon referred to as the faster-X effect, that has mostly been investigated in animal systems thus far [51] (but see [52]). There are two potential causes for faster-X, the first being that the heterogamety in males leads to stronger purifying selection against recessive deleterious mutations and positive selection for recessive beneficial mutations [53]. Another possibility is related to the effective population size (N_e) of the X [51], which controls the rate of genetic drift, and is expected to be ¾ N_e of the autosomes under neutral processes [54]. We used the McDonald Kreitman test (MKT), which uses within-species polymorphism and between-species divergence data, to estimate adaptive substitution rates (α, [55,56] across the X and autosomes. The HXR genes (n=370) show a significantly higher α relative to autosomal genes (n=6,786; Wilcoxon rank-sum test with Benjamini-Hochberg correction, p=0.031). In contrast, the PARs (n=212) were not significantly different from the HXR (p=0.13) nor from the

autosomes (p=0.82; Fig. 4). These results suggest that the hop X chromosomes show signatures of the faster-X effect, which may be due to a reduction in drift, as the X has nearly the same effective population size as the autosomes.

These molecular patterns are likely contributing to observations when crossing hops of distinct varieties, as there is ongoing evidence that sex chromosomes play a critical role in speciation [57]. The highly degenerated Y and faster-X, suggest the possibility of Haldane's Rule (HR), where the inviability of one sex will be the heterogametic sex when crossing hybrids [58]. Hops have a strong female-biased sex ratio, which becomes more extreme when crossing distinct varieties (e.g., 7:1 for within *lupulus*, 67:1 for *lupulus* to *cordifolius* [59]). The large-X effect, the second of the "two rules of speciation" along with HR, proposes that the X plays a disproportionate role in hybrid inviability, which may also play a role in the diversification of the *Humulus* genus and could drive difficulties with crossing distinct varieties.

SEX CHROMOSOMES AND HOP CONES

In hops, only XX females develop the prized "cones" that are used in brewing. Hop cones (technically the inflorescences) contain modified trichomes, called lupulin glands, that produce the characteristic bitter acids and flavors. The sex chromosomes contain the sex determination genes required for XX female versus XY male sexual differentiation, but also genes that contribute to secondary sex characteristics in both vegetative and reproductive tissues. Given the shared origin of a sex chromosome in Cannabaceae, we searched for highly-conserved, putative sex determination genes, which we expect to act early in flower development [5]. Five genes in particular are shared across all Cannabaceae Ys, none of which appear to be obvious sex-determination genes based on their developmental timing and orthologous functions, in the absence of functional validation in hops (Table S5). For instance, the maintenance of *SIGNAL PEPTIDE PEPTIDASE* (*AtSPP*) and *ROP ENHANCER 1* (*REN1*) suggests a conserved and critical role in pollen in Cannabaceae. In *Arabidopsis*, *AtSPP* is required for pollen development and pollen germination [60] and *REN1* plays an important role in pollen tube growth [61]. This line of evidence could support that sex-determination in Cannabaceae is through a dosage system of X chromosomes, rather than an active Y system, as has been suggested by crossing experiments [62].

Natural variation in sexual systems can be used to overcome limitations in functional genomics. Reversions to monoecy, where a plant can produce both functional female and male flowers, have been observed. One possible route to transitioning from dioecy to monoecy involves movement of SDR-linked sequence to the X chromosome. To test this hypothesis, we compared our dioecious to our monoecious assemblies in *Cannabis*. All monoecious karyotypes were XX, consistent with genome size estimates [63]. The monoecious Xs (X_m) are highly-syntenic to the dioecious Xs and show no clear synteny to the Y SDR (Fig. 3). There are no smaller-scale signatures of the Y chromosome on the X_m, including no dense peaks of Y-mers mapping to the X_m (Table S2) and assessing 1,581 gene trees have no evidence of a topology suggesting an SDR-linked gene translocated. These data suggest that in these individuals, monoecy evolved through a modification on the X chromosome and not movement of the Y chromosome SDR

sequence into the X. This also supports that sex-determination in Cannabaceae occurs through a dosage system of X chromosomes, or that a novel mechanism has evolved in monoecious lines that can influence sex-specific development.

Sex chromosomes can harbor genes that have functions beyond sex determination that require their contextualization in the Y-linked SDR, the HXR, and the PAR regions. Indeed, many critical floral development genes appear on the sex chromosomes (Table S9). Males flower earlier in Cannabaceae [5,64] and the sex-linkage of many of the regulators of flowering suggests that these genes may underlie these known phenological differences. For instance, homologs of *FLOWERING LOCUS D* (*FD*), *TERMINAL FLOWER 1* (*TFL1*), *EMBRYONIC FLOWER 1* (*EMF1*), and *EARLY FLOWERING IN SHORT DAYS 1* (*EFS1*) in the SDR or HXRs (Fig. S9). The PAR also contains homologs of several critical genes in the photoperiod-dependent pathways for flowering including *FLOWERING LOCUS T* (*FT*), *CONSTANS* (*CO*), and *GIGANTEA* (*GI*), and their conservation across plants could suggest similar functions in Cannabaceae [65,66]. Consistent with these findings, QTL analyses in *Cannabis* identified peaks for flowering time on the X chromosome, though not exclusively [67]. Other critical floral development genes, like *LEUNIG* (*LUG*) and *SEUSS* (*SEU*) are located on the HXR and PAR, respectively, which function together as regulators of the class C (carpel and stamen identity) homeotic gene *AGAMOUS* (*AG*) [68]. The abundance of floral genes on the Cannabaceae XYs could suggest this chromosome was predisposed to take on the role of sex chromosomes [69,70].

Female hops are strongly artificially selected for their production of bitter acids and metabolites used in the brewing process, but the impact of selection on the sex chromosome is unknown. The bitter acids are prenylated acylphloroglucinols, commonly referred to as alpha (e.g., humulone) and beta acid (e.g., lupulone). The bitter acid pathway has been well characterized [71,72] and three of the key enzymes are Valerophenone Synthase (*VPS*) and prenyltransferases (*PT1-2*). In hops that the *PT1-2* genes are tandem duplications located in the PAR of the XY chromosomes and curiously within an inversion with *FT* (Fig. S9). Leveraging the HLAGA7 sex-linked marker [73]), we traced one of the previously identified bitter acid QTLs to the sex chromosomes, consistent with the location of the *PT* genes [73,74]. Because bitter acids have been the target for breeding in hops, we tested for evidence of selection on these genes using the MKT. *PT1*, *PT2*, and *VPS* show signatures of adaptive variation (alpha = 0.76, 0.1, 0.58, respectively), of which *PT1* shows the strongest signal under Fisher's exact test (p=0.03, FDR= n.s.), consistent with selection on XX females for alpha and beta acid production, in addition to the most abundant prenylated flavinoid, xanthohumol [75]. Sex chromosomes therefore have been instrumental in the ongoing selection of genes in cultivated hop for the production of beer, and this presents a unique parallel with *Cannabis*, where the same *PT* gene family is also critical for the production of all cannabinoids [76] in XX females.

CONCLUSIONS

Our near-complete assemblies have shown the evolution of sex chromosomes in *Cannabis* and *Humulus* in unprecedented detail. Their shared origin, ~47 MYA, is among one of the oldest

known sex chromosome pairs in flowering plants. Substantial cytological variation exists, driven by differences in gene content, repeat variation, and structural variation. There is evidence of substantial degeneration of the Y chromosome, as well as faster-X. Many critical genes are found on the XY pairs, particularly involved with development of their economically important flowers. These massively diverse sex chromosomes, and their associated reproductive genes, played a pivotal role in the domestication, breeding, and ongoing production of hops to support the modern beer industry. The heteromorphic sex chromosomes thus have played an indispensable role in the domestication and ongoing breeding of these species, revealing that there is far more variability within and among sex chromosomes than can be gleaned from cytology alone.

REFERENCES

1. Stevens, N. M. *Studies in Spermatogenesis.* (1905).
2. Winge, O. On sex chromosomes, sex determination and preponderance of females in some dioecious plants. *Compt. Rend. Trav. Lab. Carlsberg* **15**, 1–26 (1923).
3. Hirata, K. Sex reversal in hemp. *Jour. Soc. Agr. For. Sapporo* (1924).
4. Renner, S. S. & Müller, N. A. Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. *Nat Plants* **7**, 392–402 (2021).
5. Shephard, H. L., Parker, J. S., Darby, P. & Ainsworth, C. C. Sexual development and sex chromosomes in hop. *New Phytol.* **148**, 397–411 (2000).
6. Prentout, D. *et al.* Plant genera Cannabis and Humulus share the same pair of well-differentiated sex chromosomes. *New Phytol.* **231**, 1599–1611 (2021).
7. Ren, G. *et al.* Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. *Sci. Adv.* **7**, (2021).
8. Small, E. & Marcus, D. Hemp: A new crop with new uses for North America. *Trends in new crops and new uses* **24**, 284–326 (2002).
9. Hornsey, I. S. *Brewing*. (Royal Society of Chemistry, 2013).
10. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
11. Gao, S. *et al.* A high-quality reference genome of wild Cannabis sativa. *Hortic Res* **7**, 73 (2020).
12. Padgitt-Cobb, L. K., Pitra, N. J., Matthews, P. D., Henning, J. A. & Hendrix, D. A. An improved assembly of the ‘Cascade’ hop (Humulus lupulus) genome uncovers signatures of molecular evolution and refines time of divergence estimates for the Cannabaceae family. *Horticulture Research* **10**, uhac281 (2023).
13. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
14. Muyle, A. *et al.* SEX-DETector: A probabilistic approach to study sex chromosomes in non-model organisms. *Genome Biol. Evol.* **8**, 2530–2543 (2016).

15. Käfer, J., Lartillot, N., Marais, G. A. B. & Picard, F. Detecting sex-linked genes using genotyped individuals sampled in natural populations. *Genetics* **218**, (2021).
16. Haunold, A. Hop Production, Breeding, and Variety Development in Various Countries. *J. Am. Soc. Brew. Chem.* **39**, 27–34 (1981).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
18. Carey, S. B. *et al.* ZW sex chromosome structure in Amborella trichopoda. *Nat. Plants* (2024) doi:10.1038/s41477-024-01858-x.
19. Zonneveld, B. J. M., Leitch, I. J. & Bennett, M. D. First nuclear DNA amounts in more than 300 angiosperms. *Ann. Bot.* **96**, 229–244 (2005).
20. Natsume, S. *et al.* The Draft Genome of Hop (Humulus lupulus), an Essence for Brewing. *Plant Cell Physiol.* **56**, 428–441 (2014).
21. Balant, M. *et al.* Novel Insights into the Nature of Intraspecific Genome Size Diversity in Cannabis sativa L. *Plants* **11**, (2022).
22. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
23. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
24. Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: incorporating protein homology information into gene prediction with GeneMark-EP and AUGUSTUS. *Plant and Animal Genomes XXVI* (2018).
25. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* (2024) doi:10.1101/gr.278090.123.
26. Otto, S. P. *et al.* About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **27**, 358–367 (2011).
27. Razumova, O. V., Divashuk, M. G., Alexandrov, O. S. & Karlov, G. I. GISH painting of the Y chromosomes suggests advanced phases of sex chromosome evolution in three dioecious Cannabaceae species (Humulus lupulus, H. japonicus, and Cannabis sativa). *Protoplasma* **260**, 249–256 (2023).
28. Zhang, Q., Onstein, R. E., Little, S. A. & Sauquet, H. Estimating divergence times and ancestral breeding systems in Ficus and Moraceae. *Ann. Bot.* **123**, 191–204 (2019).
29. Dennis Thomas, T. In Vitro Modification of Sex Expression in Mulberry (Morus Alba) by Ethrel and Silver Nitrate. *Plant Cell Tissue Organ Cult.* **77**, 277–281 (2004).
30. Mohan Ram, H. Y. & Sett, R. Modification of growth and sex expression in cannabis sativa by aminoethoxyvinylglycine and ethephon. *Z. Pflanzenphysiol.* **105**, 165–172 (1982).
31. Zhang, X. *et al.* Genomes of the Banyan Tree and Pollinator Wasp Provide Insights into Fig-Wasp Coevolution. *Cell* **183**, 875–889.e17 (2020).
32. Xia, Z. *et al.* Chromosome-level Genomes Reveal the Genetic Basis of Descending

Dysploidy and Sex Determination in Morus Plants. *Genomics Proteomics Bioinformatics* (2022) doi:10.1016/j.gpb.2022.08.005.

33. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
34. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
35. Yu, Q. *et al.* Chromosomal location and gene paucity of the male specific region on papaya Y chromosome. *Mol. Genet. Genomics* **278**, 177–185 (2007).
36. She, H. *et al.* Evolution of the spinach sex-linked region within a rarely recombining pericentromeric region. *Plant Physiol.* **193**, 1263–1280 (2023).
37. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
38. Makova, K. D. *et al.* The complete sequence and comparative analysis of ape sex chromosomes. *Nature* **630**, 401–411 (2024).
39. Pennell, M. W. *et al.* Y fuse? Sex chromosome fusions in fishes and reptiles. *PLoS Genet.* **11**, e1005237 (2015).
40. Charlesworth, D. & Charlesworth, B. Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genet. Res.* **35**, 205–214 (1980).
41. Wilhoit, K. T., Alexander, E. P. & Blackmon, H. Worse than nothing at all: the inequality of fusions joining autosomes to the PAR and non-PAR portions of sex chromosomes. *PeerJ* **12**, e17740 (2024).
42. Divashuk, M. G., Alexandrov, O. S., Kroupin, P. Y. & Karlov, G. I. Molecular cytogenetic mapping of Humulus lupulus sex chromosomes. *Cytogenet. Genome Res.* **134**, 213–219 (2011).
43. Divashuk, M. G., Alexandrov, O. S., Razumova, O. V., Kirov, I. V. & Karlov, G. I. Molecular cytogenetic characterization of the dioecious Cannabis sativa with an XY chromosome sex determination system. *PLoS One* **9**, e85118 (2014).
44. Waters, P. D., Duffy, B., Frost, C. J., Delbridge, M. L. & Graves, J. A. The human Y chromosome derives largely from a single autosomal region added to the sex chromosomes 80-130 million years ago. *Cytogenet. Cell Genet.* **92**, 74–79 (2001).
45. Blackmon, H. & Demuth, J. P. Estimating tempo and mode of Y chromosome turnover: explaining Y chromosome loss with the fragile Y hypothesis. *Genetics* **197**, 561–572 (2014).
46. Reeves, P. A. & Richards, C. M. Species delimitation under the general lineage concept: an empirical example using wild North American hops (Cannabaceae: Humulus lupulus). *Syst. Biol.* **60**, 45–59 (2011).
47. van der Bijl, W. & Mank, J. E. Hiding in plain sight: the Y chromosome and its reinvigorated role in evolutionary processes. *Evol. Lett.* qrae065 (2024) doi:10.1093/evlett/qrae065.
48. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic

sex chromosomes. *Heredity* **95**, 118–128 (2005).
49. Papadopulos, A. S. T., Chester, M., Ridout, K. & Filatov, D. A. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13021–13026 (2015).
50. Sacchi, B. *et al.* Phased assembly of Neo-sex chromosomes reveals extensive Y degeneration and rapid genome evolution in Rumex hastatulus. *Mol. Biol. Evol.* **41**, msae074 (2024).
51. Mank, J. E., Vicoso, B., Berlin, S. & Charlesworth, B. Effective population size and the Faster-X effect: empirical results and their interpretation. *Evolution* **64**, 663–674 (2010).
52. Krasovec, M., Nevado, B. & Filatov, D. A. A comparison of selective pressures in plant X-linked and autosomal genes. *Genes (Basel)* **9**, 234 (2018).
53. Charlesworth, B., Coyne, J. A. & Barton, N. H. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *Am. Nat.* **130**, 113–146 (1987).
54. Sayres, M. A. W. Genetic diversity on the sex chromosomes. *Genome Biol. Evol.* **10**, 1064 (2018).
55. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–654 (1991).
56. Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in Drosophila. *Nature* **415**, 1022–1024 (2002).
57. Payseur, B. A., Presgraves, D. C. & Filatov, D. A. Introduction: Sex chromosomes and speciation. *Mol. Ecol.* **27**, 3745–3748 (2018).
58. Cowell, F. 100 years of Haldane's rule. *J. Evol. Biol.* **36**, 337–346 (2023).
59. Jakse, J., Stajner, N., Kozjak, P., Cerenak, A. & Javornik, B. Trinucleotide microsatellite repeat is tightly linked to male sex in hop (Humulus lupulus L.). *Mol. Breed.* **21**, 139–148 (2008).
60. Han, S., Green, L. & Schnell, D. J. The signal peptide peptidase is required for pollen function in Arabidopsis. *Plant Physiol.* **149**, 1289–1301 (2009).
61. Hwang, J.-U., Vernoud, V., Szumlanski, A., Nielsen, E. & Yang, Z. A tip-localized RhoGAP controls cell polarity by globally inhibiting Rho GTPase at the cell apex. *Curr. Biol.* **18**, 1907–1916 (2008).
62. Parker, J. S. & Clark, M. S. Dosage sex-chromosome systems in plants. *Plant Sci.* **80**, 79–92 (1991).
63. Razumova, O. V., Alexandrov, O. S., Divashuk, M. G., Sukhorada, T. I. & Karlov, G. I. Molecular cytogenetic analysis of monoecious hemp (Cannabis sativa L.) cultivars reveals its karyotype variations and sex chromosomes constitution. *Protoplasma* **253**, 895–901 (2016).
64. Mediavilla, V., Jonquera, M. & Schmid-Slembrouck, I. Decimal code for growth stages of hemp (Cannabis sativa L.). (1998).
65. Brandoli, C., Petri, C., Egea-Cortines, M. & Weiss, J. Gigantea: Uncovering new functions in flower development. *Genes (Basel)* **11**, 1142 (2020).

66. Romero, J. M. *et al.* CONSTANS, a HUB for all seasons: How photoperiod pervades plant physiology regulatory circuits. *Plant Cell* **36**, 2086–2102 (2024).
67. Steel, L., Welling, M., Ristevski, N., Johnson, K. & Gendall, A. Comparative genomics of flowering behavior in Cannabis sativa. *Front. Plant Sci.* **14**, (2023).
68. Franks, R. G., Wang, C., Levin, J. Z. & Liu, Z. SEUSS, a member of a novel family of plant regulatory proteins, represses floral homeotic gene expression with LEUNIG. *Development* **129**, 253–263 (2002).
69. Marshall Graves, J. A. & Peichel, C. L. Are homologies in vertebrate sex determination due to shared ancestry or to limited options? *Genome Biol.* **11**, 205 (2010).
70. Kratochvíl, L., Gamble, T. & Rovatsos, M. Sex chromosome evolution among amniotes: is the origin of sex chromosomes non-random? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **376**, 20200108 (2021).
71. Clark, S. M., Vaitheeswaran, V., Ambrose, S. J., Purves, R. W. & Page, J. E. Transcriptome analysis of bitter acid biosynthesis and precursor pathways in hop (Humulus lupulus). *BMC Plant Biol.* **13**, 12 (2013).
72. Champagne, A. & Boutry, M. A comprehensive proteome map of glandular trichomes of hop (Humulus lupulus L.) female cones: Identification of biosynthetic pathways of the major terpenoid-related compounds and possible transport proteins. *Proteomics* **17**, 1600411 (2017).
73. Cerenak, A. *et al.* Identification of QTLs for alpha acid content and yield in hop (Humulus Lupulus L.). *Euphytica* **170**, 141–154 (2009).
74. McAdam, E. L. *et al.* Quantitative trait loci in hop (Humulus lupulus L.) reveal complex genetic architecture underlying variation in sex, yield and cone chemistry. *BMC Genomics* **14**, 360 (2013).
75. Yang, S., Chen, R., Cao, X., Wang, G. & Zhou, Y. J. De novo biosynthesis of the hops bioactive flavonoid xanthohumol in yeast. *Nat. Commun.* **15**, 253 (2024).
76. Luo, X. *et al.* Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **567**, 123–126 (2019).
77. Gasic, K., Hernandez, A. & Korban, S. S. RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. *Plant Mol. Biol. Rep.* **22**, 437–438 (2004).
78. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
79. Astashyn, A. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 60 (2024).
80. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
81. Dudchenko, O. *et al.* The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *BioRxiv* (2018) doi:10.1101/254797.

82. Lovell, J. T. *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, (2022).
83. Grassa, C. J. *et al.* A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytol.* **230**, 1665–1679 (2021).
84. Lynch, R. C. *et al.* Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *bioRxiv* (2024) doi:10.1101/2024.05.21.595196.
85. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
86. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
87. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. Preprint at (2015).
88. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
89. van Velzen, R. *et al.* Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4700–E4709 (2018).
90. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, (2023).
91. Ou, S. *et al.* Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome Res.* **34**, 1140–1153 (2024).
92. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
93. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
94. Madden, T. L. The BLAST Sequence Analysis Tool. *The NCBI handbook* (2013).
95. Neumann, P. *et al.* Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* **2**, 4 (2011).
96. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
97. Marcais, G. & Kingsford, C. Jellyfish: A fast k-mer counter. *Tutorialis e Manuais* 1–8 (2012).
98. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
99. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
100. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

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. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
103. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
104. Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).
105. Khan, A. *et al.* A phased, chromosome-scale genome of 'Honeycrisp' apple (Malus domestica). *GigaByte* **2022**, gigabyte69 (2022).
106. Xia, Z. *et al.* Haplotype-resolved chromosomal-level genome assembly reveals regulatory variations in mulberry fruit anthocyanin content. *Hortic. Res.* **11**, uhae120 (2024).
107. Yocca, A. *et al.* A chromosome-scale assembly for 'd'Anjou'pear. *G3: Genes, Genomes, Genetics* **14**, jkae003 (2024).
108. Yang, M. *et al.* Insights into the evolution and spatial chromosome architecture of jujube from an updated gapless genome assembly. *Plant Commun.* **4**, 100662 (2023).
109. Kubatko, L. S. & Degnan, J. H. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* **56**, 17–24 (2007).
110. Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94**, 447–462 (2016).
111. Johnson, M. G. *et al.* HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**, 1600016 (2016).
112. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
113. 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).
114. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
115. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
116. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
117. Zhou, C. *et al.* Oatk: a de novo assembly tool for complex plant organelle genomes. *bioRxiv* (2024) doi:10.1101/2024.10.23.619857.
118. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18 (2017).

119. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
120. Stadler, T. Sampling-through-time in birth-death trees. *J. Theor. Biol.* **267**, 396–404 (2010).
121. Heath, T. A., Huelsenbeck, J. P. & Stadler, T. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2957–66 (2014).
122. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
123. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
124. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437–453 (2015).
125. E, K. & H. *Monographie Der Früchte Und Samen in Der Kreide von Mitteleuropa*. (Ústřední ústav geologický v Academii, nakl. Československé akademie věd, 1986).
126. Friis, E. M., Crane, P. R. & Pedersen, K. R. Early Flowers and Angiosperm Evolution.,(Cambridge University Press: Cambridge, UK). (2011).
127. Collinson, M. E. The fossil history of the Moraceae, Urticaceae (including Cecropiaceae), and Cannabaceae. *Evolution, systematics, and fossil history of the Hamamelidae* **2**, 319–339 (1989).
128. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
129. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
130. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
131. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
132. Korunes, K. L. & Samuk, K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* **21**, 1359–1368 (2021).
133. Murga-Moreno, J., Coronado-Zamora, M., Hervas, S., Casillas, S. & Barbadilla, A. iMKT: the integrative McDonald and Kreitman test. *Nucleic Acids Res.* **47**, W283–W288 (2019).
134. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
135. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.* **5**, e310 (2007).

FIGURES

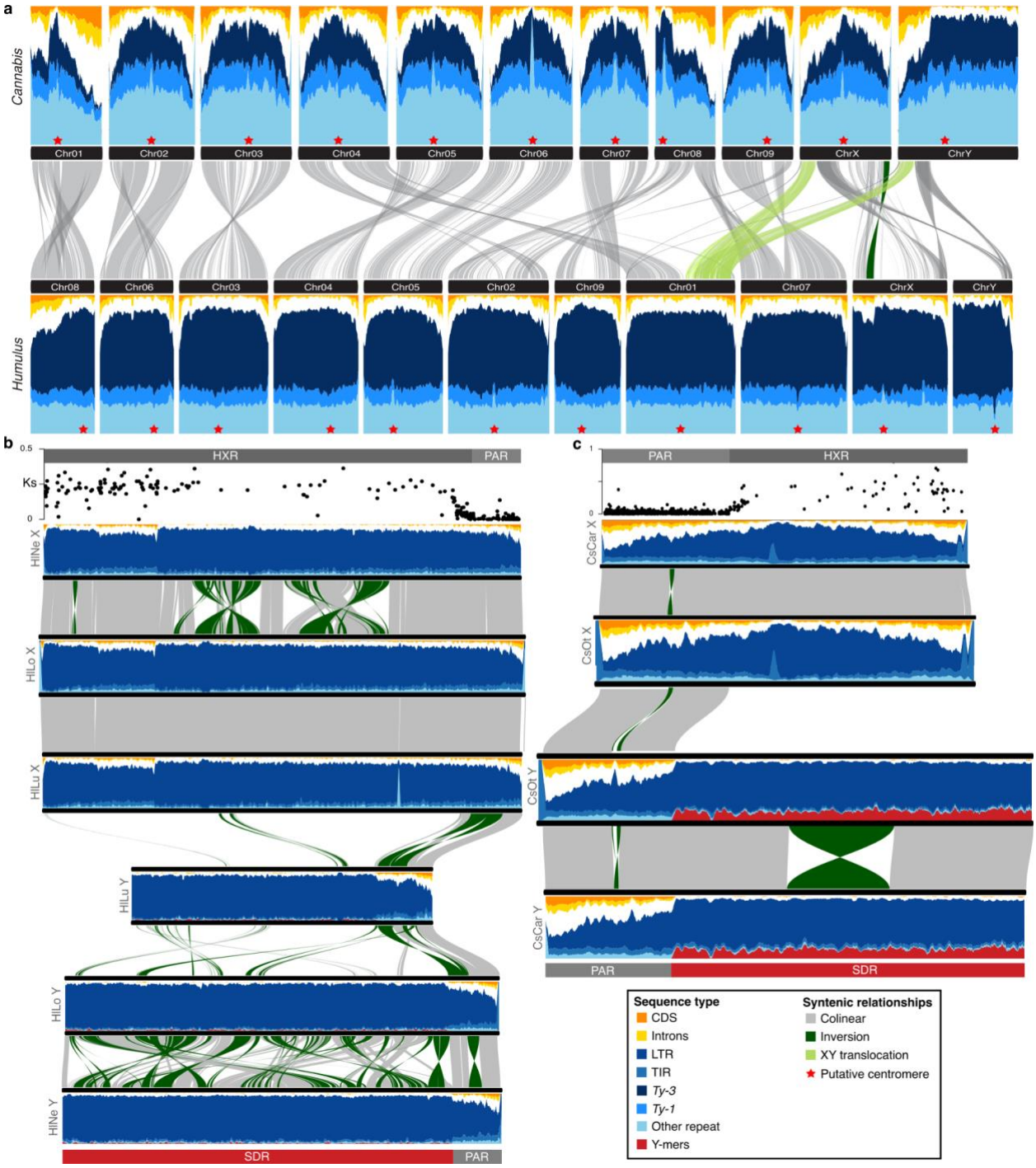

Figure 1. The genome architecture of *Cannabis* and *Humulus*. a) gene and repeat landscapes of *Cannabis* and *Humulus* and their syntenic relationships. These species exhibit a typical angiosperm architecture, with repeats increasing at the centromeres and gene dense arms. One exception is the X chromosome in hops, where an inverted translocation has altered the gene-dense region towards the centromere. The genera are also largely syntenic, though some

structural changes have occurred. The autosomes are represented with a single haplotype but both sex chromosomes. Due to the genome size differences, *Humulus* was scaled to *Cannabis*. **b**) gene and repeat content, and structural relationships between the X and Y chromosomes within hops and *Cannabis* (**c**). Y-mer density highlights the sex-determination regions (SDR), while synonymous protein changes (Ks) for one-to-one XY gametologs highlight the homologous region of the Xs (HXR; values above one were removed). On the Ys, repeats sharply increase in the SDRs relative to the pseudoautosomal regions (PAR), and become gene poor. The Y chromosomes are highly structurally variable in hops, though more conserved in these *Cannabis* isolates, and in both genera show very little synteny with the X chromosomes outside of the PARs.

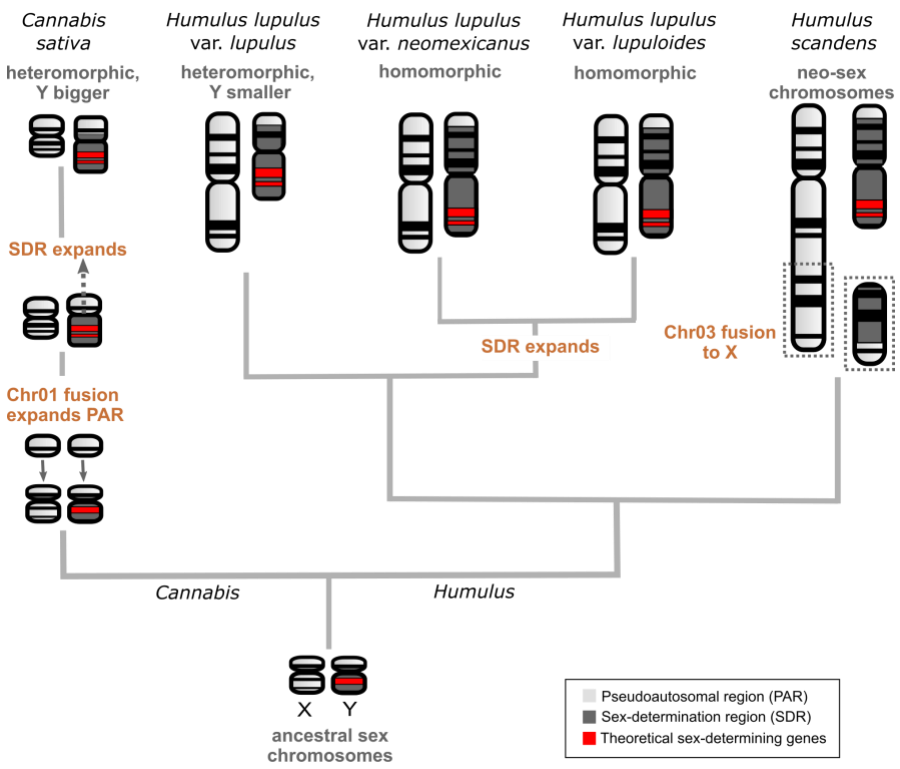

Figure 2. Model of sex chromosome evolution in Cannabaceae. The *Cannabis* and *Humulus* sex chromosomes have extreme structural diversity with heteromorphic, homomorphic, and multiple sex chromosome pairs. The sex chromosomes evolved from an ancestral autosomal pair and have been on independent trajectories once each lineage diverged. A fusion of part of an autosome to the pseudoautosomal region (PAR) expanded the size of *Cannabis* XYs and the sex-determination region (SDR) has expanded, and over time capturing more gene content into the non-recombining region. The American hops, *H. lupulus* vars. *neomexicanus* and *lupuloides* also had independent expansions of the SDRs. In *H. scandens*, a different chromosome independently fused to the X chromosome than in *Cannabis*, while the two Y chromosomes remain unfused. For clarity, chromosome numbers provided here are based on *Humulus*.

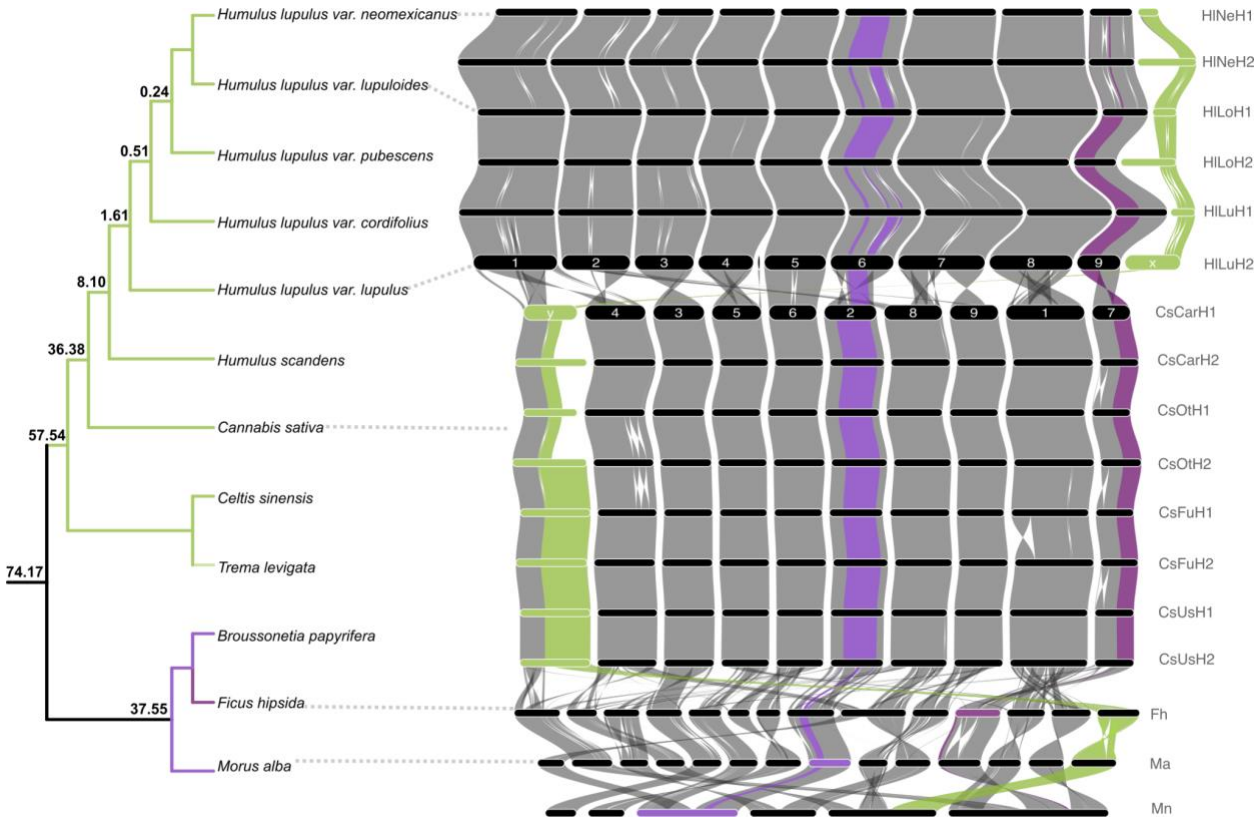

Figure 3. Phylogenetic relationships and synteny in Cannabaceae and Moraceae. Dates on the cladogram are the mean age in millions of years. Green branches highlight the Cannabaceae and purple the Moraceae families. The gray dashed lines connect species in the cladogram that are represented in the GENESPACE synteny plot. Highlighted in green are the XY chromosomes in Cannabaceae (the pseudoautosomal region only, as the sex-determination regions show very little synteny, Fig. 1). In purple are the sex-determination regions in Moraceae, highlighting they are on different chromosomes than in Cannabaceae, consistent with their independent evolutions.

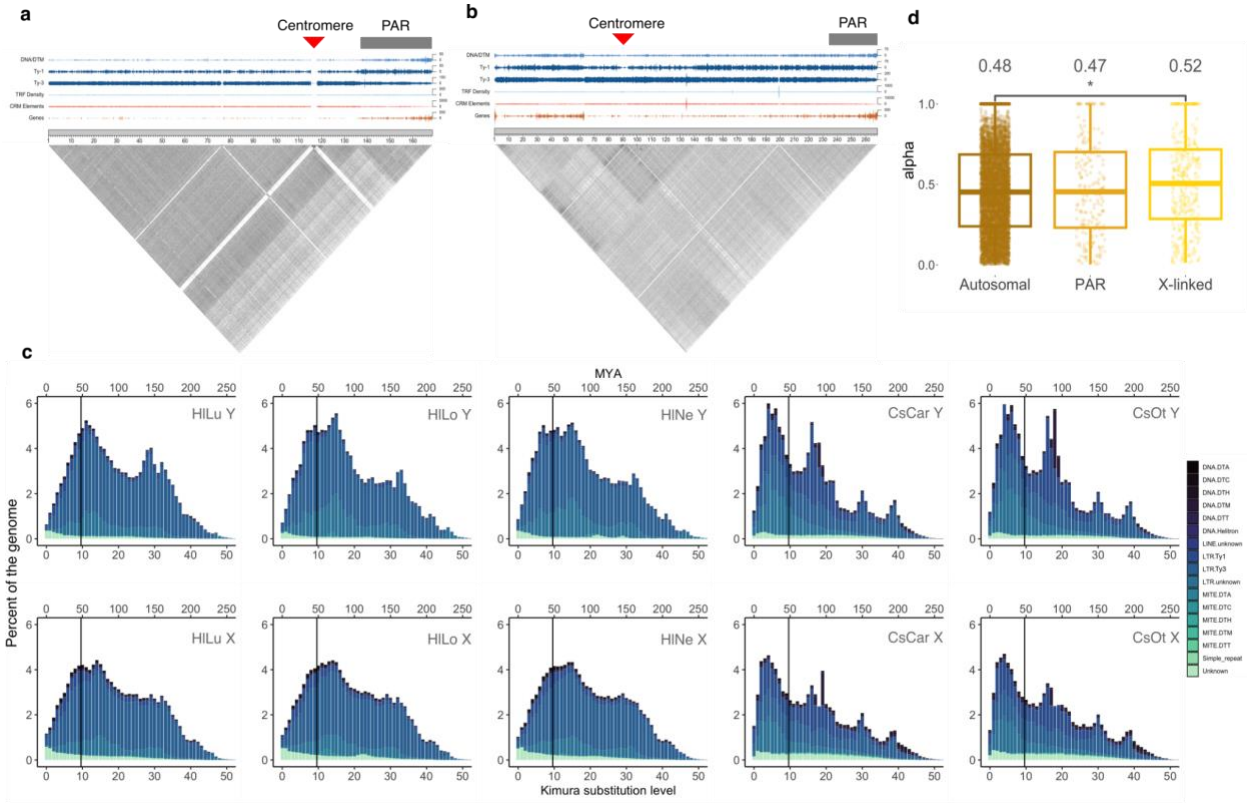

Figure 4. Molecular evolution of Cannabaceae XYs. a) putative centromere location for the *Humulus lupulus* var. *lupulus* Y and (**b**) X chromosomes showing their different proximities from the pseudoautosomal regions (PAR). The X pericentromere features low-complexity tandem repeats (gray) flanking a higher-complexity centromeric array (red arrow), a slight increase in *Ty-3* elements and CRM retrotransposons, and a reduction in *Ty-1* elements and other DNA transposons (DNA/DTM). The pericentromeric region is not observable on the Y, due to the high repetitive content; however, the centromeric array and reduction in *Ty-1* indicate the putative centromere location. (**b**) X chromosomes showing their different proximities from the pseudoautosomal regions (PAR). **c**) Kimura substitution plots, with lower values closer to 0 representing more recent events and higher values approaching 50 representing older events. Time was approximated by dividing the Kimura Distance by twice the mutation rate (we assumed a value of $1x10^{-9}$), and the black lines indicate the initial sex chromosome evolution. **d**) boxplot for alpha values from the McDonald Kreitman test. Values above each box are the mean and the asterisk indicates significance.

METHODS

Sequence generation

All sequencing was performed at HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA), unless otherwise noted. To obtain high molecular weight (HMW) DNA for the genome isolates (Table S1), we flash froze young leaf tissue in liquid nitrogen then extracted HMW genomic DNA using a Takara NucleoBond HMW kit (San Jose, CA) according to the manufacturer's protocol. Long-read sequencing was performed using circular consensus sequencing (CCS) mode on a PacBio Sequel II (Pacific Biosciences of California). Libraries were constructed using a SMRTbell Template Prep Kit 3.0, sheared using a Megaruptor (Diagenode), and sized on a Blue Pippin instrument (Sage Science, Beverly, MA, USA). Sequencing was performed using a 30 hour movie time with two hour pre-extension, and the resulting raw data were processed using the CCS6.2 algorithm.

To scaffold the genome assemblies, we generated Omni-C data. Libraries were constructed using standard protocols (Cantata Bio, Dovetail Omni-C Kit Catalog #21005). One gram of flash, frozen leaf tissue was ground and after nuclei isolation, the pellet was resuspended in 1x PBS, aliquoted into three tubes and centrifuged at 6000 rpm for five minutes. The supernatant was then discarded and the nuclei pellets flash frozen for storage at -80 degrees. One aliquoted nuclei pellet was then processed using the Dovetail Omni-C protocol V2. Libraries were run on an Illumina NovaSeq 6000 PE150.

To develop the male-specific k-mer lists (i.e., Y-mers), and data for the population genomic analyses, we generated Illumina whole-genome sequencing (WGS) for 22 *H. lupulus* individuals, spanning all five botanical varieties (Table S10). Isolates were extracted using a CTAB approach and sequencing libraries were constructed using an Illumina TruSeq DNA PCR-free library kit (Catalog #20015962) using standard protocols. Libraries were sequenced on an Illumina NovaSeq 6000 instrument using paired ends and a read length of 150 base pairs. For two isolates (64035M and Bullion) DNA was instead extracted using the BioEcho EchoLUTION Plant DNA 96 Kit (Product No. 010-103-002) and sequencing libraries were constructed using a seqWell purePlex DNA Library Preparation Kit. Sequencing was performed on an Illumina NovaSeq X Plus instrument using paired-end reads and a read length of 150 base pairs at NovoGene (Beijing, China).

To provide mRNA support for the gene annotations, we generated additional RNAseq data. We flash froze leaf (with some stem remnants likely) tissues and inflorescences for 21110M and Comet "male", respectively. We extracted RNA using a CTAB approach [77].

Sequencing libraries were constructed using an Illumina TruSeq Stranded mRNA Library Prep Kit (Catalog #20020595) using standard protocols. Libraries were sequenced on an Illumina NovaSeq 6000 instrument using paired ends and a read length of 150 base pairs.

Genome assemblies

For *Cannabis*, HiFi reads were assembled into contigs using hifiasm v0.16.0, with the Hi-C integration mode that incorporated Omni-C reads for phasing. For *Humulus*, we used hifiasm

v0.19.5. For 21110M, we also changed the '-s' parameter from 0.55 (default) to 0.2 to better balance haplotypes out of the assembly. Using our HiFi data, we polished our contigs using Racon v1.5.0 [78], removed contigs below 50 Kb using BBTools v38.18-0 (https://sourceforge.net/projects/bbmap/), and removed contigs identified as contamination using FCS-GX v0.4.0 [79].

Prior to scaffolding, we manually placed the sex chromosomes into a single haplotype (described in detail in the Supplementary Discussion). The sex chromosome-corrected assembly was then scaffolded by first mapping the Omni-C reads using BWA-MEM, with parameters -5SP and samtools version 1.10 with parameters view -S -h -b -F 2316. Next, we scaffolded the assembly using Yet Another Hi-C Scaffolding tool (YaHS) v1.1 with default parameters [80]. We edited contact maps manually with Juicebox Assembly Tools (JBAT) v2.15 [81] to produce the expected 10 chromosomes per haplotype. Scaffolds containing assembled telomeres were checked for correct orientation to the terminal ends by searching for the $(TTTAGGG)_n$ repeat using GENESPACE v1.3.1 [82]. The *Cannabis* assemblies were reordered and oriented to the existing cs10 reference [83] and hops were reordered and oriented to Cascade [12]. Final plots showing the number of contigs and telomeric regions (density of the window for the telomeric repeat set to 0.75) were generated using GENESPACE.

Genome annotations

To generate gene model annotations, for *Cannabis*, we annotated repeats and genes following previously described methods [84]. In brief, we used a repeat library built from existing assemblies and repeats were masked with RepeatMasker v4.1.2. We annotated gene models using Braker v2.1.6 and TSEBRA [85]. For hops, we annotated repeats by building a *de novo* library for each haplotype separately using RepeatModeler v2.0.5 [86], invoking the -LTRStruct parameter. We then masked repeats using RepeatMasker v4.1.5 [87]. We annotated gene models by using BRAKER3 v3.0.6 [25], using both protein and RNAseq as extrinsic evidence. For the proteins, we used the OrthoDB v11 [88] for Viridiplantae, combined with proteins from genome references for *Parasponia* [89], *Trema* [89], Cascade, and cs10. For RNAseq support, we used existing RNAseq data for *Humulus*, as well as RNAseq generated here (Table S10). We assessed the gene annotations using compleasm v0.2.6 [90] using the Eudicots database.

Repeat analyses

To examine the landscape of TEs, we additionally annotated repeats using EDTA v2.0.0 [23] for each haplotype separately. To compare repeat landscapes between *Cannabis* and hops, we generated a panEDTA library using EDTA v2.2.0 [91]. To identify putative centromere locations, we used StainedGlass v0.5[92] to visualize the massive tandem repeat arrays. To identify tandem repeats, we used Tandem Repeats Finder v4.09.1 [93] (parameters 2 7 7 80 10 50 500 -f -d -m -h) and to identify CRM elements we used BLAST v2.15.0 [94] using previously identified elements [95]. To plot the landscape of repeats and genes in the context of the putative centromere locations, we use sliding windows of 100kb and 200kb for the Y and X, respectively.

Identification of the sex chromosomes using Y-mers

We used whole-genome sequencing data, and HiFi long-reads (*H. lupulus* var. *lupulus* 21110M and *H. lupulus* var. *lupuloides* MN-1421 isolates only), to identify the sex-determining region (SDR) of the Y. All paired-end Illumina data had adapters removed and were quality filtered using TRIMMOMATIC v0.39 [96] with leading and trailing values of three, sliding window of 30, jump of 10, and a minimum remaining read length of 40. We next found all canonical 21-mers in each isolate using Jellyfish v2.3.0 [97] and used the bash *comm* command to identify all *k*-mers shared in all male isolates and not found in any female isolate (Y-mers). For *H. scandens* we used meryl v1.3 [22] to identify Y-mers. We used to meryl *count* to identify all 21-mers in each isolate and the *intersect* and *difference* commands for identifying Y-mers. We mapped the Y-mers to both haplotype assemblies using BWA-MEM, with parameters -k 21 -T 21 -a -c 10 and generated a table of coverage per contig using SAMtools v1.10 *coverage* [98].

Gene tree analyses

To compare the boundary of the SDR with the PAR when using Y-mer mapping, and examine the timing of gene capture into the SDR, we ran OrthoFinder v2.5.2, with the *Cannabis* and hop proteins and additional angiosperm genomes (Table S11), [99,100] in ultra-sensitive mode. For OrthoGroups that contained sex-linked genes, we aligned the genes using MAFFT v7.471 [101] with the parameter maxiterate set to 1000 and using genafpair. We built gene trees using RAxML v8.2.12 [102] with 100 bootstrap replicates and the model PROTGAMMAWAG. In total, we built 1,096 gene trees containing Y-linked gene models and visually examined the topology. To examine phylogenetic evidence of monoecy evolving from a Y-linked gene translocating to the X, we used ETE3 v3.1.1 [103] *check_monophyly* function.

Ka/Ks

To identify one-to-one orthologs on the XY and examine protein evolution, we ran OrthoFinder v2.5.2 [99,100] within a reference's haplotypes. We calculated synonymous (Ks) and nonsynonymous (Ka) changes in codons using Ka/Ks Calculator v2.0 [104]. To use Ks to estimate the age of the SDR, we used the following the formula: time (T)= $(Ks/\boldsymbol{\pi}-1)x(2N_e t_g)$, where we used a μ of $1x10^{-9}$, $\boldsymbol{\pi}$=0.0158 (described below), and a generation time (t_g)=2.

Synteny analyses

To examine syntenic relationships between Cannabaceae and Moraceae, we use GENESPACE using default parameters. Visualization and analysis of local synteny was conducted with DEEPSPACE (https://github.com/jtlovell/DEEPSPACE), which aligns genomic windows between assemblies and uses GENESPACE machinery to infer syntenic blocks. Source code for the riparian plots was modified where needed to visually compare two genomes with such different total sizes.

Species tree inference

We used OrthoFinder v2.5.4 [99,100] to identify conserved, single-copy genes for species tree inference with ASTRAL v5.7.8 [33], which employs a coalescent-based summary method that is statistically consistent under the multi-species coalescent model. To identify orthologs conserved as single copy across the Rosales order, we used gene annotations from the following Rosales taxa: *H. lupulus* var. *lupuloides*, *H. lupulus* var. *lupulus*, *H. lupulus* var. *neomexicanus*, *C. sativa*, *Ficus hispida* [31], *Malus domestica* [105], *Morus alba* [106], *Pyrus commune* [107], and *Ziziphus jujuba* [108]. All single-copy genes conserved across those samples were used for gene and species tree inference, after removing all orthologs of sex chromosome linked genes (found in any species) since any of those may exhibit unique evolutionary histories in different lineages. Coalescent-based, rather than concatenation, methods were used to analyze nuclear genes because concatenation implicitly assumes an absence of recombination and ILS, or incomplete lineage sorting (i.e., a single history is shared amongst all genes), and can result in statistical inconsistencies in multi-locus datasets with high levels of gene tree discordance [109,110]; whereas coalescent approaches assume free recombination between genes, accounting for gene tree–species tree discordance due to ILS.

Along with our *Humulus* dataset, publicly available Illumina short reads were downloaded from NCBI SRA for additional lineages spanning the Rosales (Table S12), to produce gene assemblies for phylogenomic analysis. We employed HybPiper v2.1.6 [111] to i) map reads to target nucleotide sequences with BWA-MEM, ii) assemble mapped reads into contiguous sequences with SPAdes v3.15.2 [112], iii) extend gene assemblies into flanking intronic regions, and iv) test for paralogs with default HybPiper parameters. The target sequence file used for read mapping with HybPiper was manually curated with nucleotide coding sequences for 775 one-to-one orthologs from genome annotations for: 1) *H. lupulus* var. *neomexicanus*, 2) *M. domestica*, 3) *C. sativa*, 4) *M. alba*, and 5) *Z. jujuba*.

Multiple sequence alignments for each target gene were produced using MAFFT v7.520 with the *--auto* option, which automatically selects the best alignment strategy based on the data [101]. Poorly-aligned sequences were trimmed using trimAl v1.4.1 with the flag *-automated1* [113]. Gene tree inference was performed using maximum likelihood (ML) with IQ-TREE v1.6.2 [114], 1000 ultrafast bootstraps (BS) [115], and model selection based on Bayesian information criterion with ModelFinder [116]. Prior to the ASTRAL analysis, gene tree branches with <10% BS support were collapsed into polytomies to improve the accuracy of species tree inference [33]. The resulting species tree was rooted with Rosaceae taxa (i.e., *P. communis*, *M. domestica*).

Plastome divergence time estimations

Whole (circularized) plastome assemblies were used to estimate relationships and divergence times among the same, or similar, Rosales taxa included in the species tree analysis (Table S12). To assemble chloroplast genomes, we used oatk v0.2 [117] when HiFi data was available. For all others where we used Novoplasty v4.3.5 [118]. Limited taxon sampling was used to optimize time calibrations across the Rosales tree, and obtain robust divergence time estimates for ingroup branches (i.e., Cannabaceae and *Humulus*). Whole plastomes were aligned using MAFFT v7.520 with the *--auto* option and used as input for ML tree inference with IQ-TREE v1.6.2 (as performed for individual nuclear gene trees) and bayesian divergence time estimation with BEAST v2.7.7 [119]. BEAST was executed using a GTR+G+I substitution and site heterogeneity model with four rate categories; a fossilized birth-death (FBD) tree prior to model speciation and extinction [120,121]; and optimized relaxed clock model with a lognormal distribution [122]. The plastome tree topology was constrained according to the ML tree from IQ-TREE and rooted with

Rosaceae taxa. Two independent BEAST runs, totalling 48,858,000 MCMC generations, were combined to obtain >200 effective sample size (ESS) for all priors, with MCMC sampling every 1000 generations and 10% burn-in. ESS was assessed with Tracer v1.7.2 [123]. To calibrate the plastome tree to time (i.e., million years ago=Ma) and estimate the tree root height, we estimated the FBD origin under a normal distribution with a mean of 107 Ma and a liberal standard deviation of 4.0. Those parameters virtually constrained the total tree height to approximately 100-114 Ma, which is in accordance with crown group age estimates for Rosales in previous studies [28,124]. Under the FBD model, two fossil calibrations were applied to further calibrate the plastome tree to time: the first applied a 66 Ma minimum age for the Cannabaceae stem group [28,125,126] and the second applied a 33.9 Ma minimum age for the stem group of *Humulus* [28,127]**.**

Structural variant analyses

To identify structural variants between the genome assemblies we first generated alignments using minimap2 v2.24 [128] using --eqx. We called SVs using SyRI v1.7.0 [129], ignoring SNPs (--nosnp) and setting --invgaplen 10000000.

Population genetic analyses

BWA-MEM v0.7.17 [130] was used to map reads to the *H. lupulus* var. *lupulus* reference, with some modifications based on the analysis. We used bcftools v1.9 *mpileup* and *call* [131] functions to call variants, where we applied the -G parameter for *call*. We filtered the vcf file using 'QUAL>20 & DP>5 & MQ>30', minor allele frequency of 0.05, and dropped sites with > 25% missing data. To calculate **π**, for the genome reference, we used the autosomes for HAP1, removed the Y chromosome, but included the X chromosome, and restricted the analysis to females of the wild-collected hops. We used pixy v1.2.7.beta1 [132] using 100,000-bp windows with a 10,000-bp jump. To estimate N_e, we used the relationship $\theta = 4N_e\,\mu$, where we used **π** as our metric for θ, we and found $\theta \sim= 0.015$ for both the X and autosomes. To examine F_{st} for *H. scandens*, we mapped previously existing samples (Table S10) to the *H. lupulus* var. *lupulus* HAP1 reference that includes the Y, so we could compare with the Y-mer mapping. In pixy, we calculated F_{st} for gene models only.

McDonald Kreitman Test

To estimate adaptive substitutions, or alpha (α) [56], for the faster-X test, we used the McDonald Kreitman Test (MKT). We used the autosomes for HAP1, removed the Y chromosome, but included the X chromosome for the *H. lupulus* var. *lupulus* reference, and restricted the analysis to only female hops, including one *H. scandens* female as the outgroup species. We used the R package iMKT v0.1.1 [133], where we applied a Fay, Wyckoff, and Wu correction [134] of 0.15. Because the MKT can be underpowered when there are low counts in the 2-by-2 contingency table, we only compared results for which the row and column totals were six or greater [135]. Because negative α values indicate that the assumptions of the MKT were likely not met, and difficult to meaningfully interpret, we also excluded α values below zero. Thus we restrict our analysis to testing for differences in the proportion of substitution driven by positive selection, only for genes that meet the null hypothesis (α=0) or show evidence for adaptive evolution

(α>0). To test for selection on the bitter acid pathway genes, we followed the same approach with some exceptions. We included male isolates in the analysis and mapped to the *H. lupulus* var. *lupulus* HAP1 reference that includes the Y. After calculating the MKT, we used the Fisher exact test to test for significance on each gene and applied a false discovery rate adjustment to the p-values.

Data availability

The genome assemblies have been deposited on NCBI under BioProjects (TBD). Sequencing libraries for the genome assembly and annotation are publicly available on NCBI under BioProject (TBD) and the whole-genome sequencing of additional isolates are under (TBD). Individual accession numbers are provided in Supplementary Tables S1 and S10.

Acknowledgments

Support for this work was provided by the US Department of Agriculture National Institute of Food and Agriculture Postdoctoral Fellowship (USDA NIFA) no. 2022-67012-38987 (S.B.C.), USDA NIFA no. 2023-67013-39620 (A.H.), National Science Foundation (NSF) IOS-PGRP CAREER no. 2239530 (A.H.), NIH R35 GM147107 (R.F.G.), and Institute of *Cannabis* Research at Colorado State University Pueblo ICR-FY22-Kane (R.F.G., D.V., N.C.K.).

Author contributions

Concept and research design: S.B.C., J.S.H, G.J.M., A.H.
Sample collection, data collection, sequencing: S.B.C., J.S.H., K.E., L.R.O., H.H., H.M., Z.M., J.M., D.V., R.F.G., N.C.K., G.J.M., A.H.
Genome assembly and annotation: S.B.C., L.P.-C., R.L., N.A.., T.P.M., A.H.
Computational and statistical analysis: S.B.C., P.C.B., L.A.M., J.T.L., A.O.
Wrote the paper (with contributions from all authors): S.B.C., P.C.B., J.T.L., L.M.A., N.C.K., A.H.

Competing interests

D.V. is a board member of the 501(C)3 non-profit Agricultural Genomics Foundation and the sole owner of the company CGRI, LLC