An ancient X chromosomal region harbours three genes potentially controlling sex determination in *Cannabis sativa*

Matteo Toscani, Afsheen Malik, Ainhoa Riera-Begue, Caroline A. Dowling, Susanne Schilling and Rainer Melzer

UCD School of Biology and Environmental Science and Earth Institute, University College Dublin, Ireland

Author emails and ORCID:

matteo.toscani@ucdconnect.ie 0009-0006-8216-6814

ainhoa.rierabegue@ucdconnect.ie 0009-0005-1519-5626

afsheen.malik@ucd.ie 0000-0001-8686-9740

caroline.a.dowling.1@gmail.com 0000-0002-1184-7798

susanne.schilling@ucd.ie 0000-0002-1142-6520

rainer.melzer@ucd.ie 0000-0003-0631-9023 20

Abstract

The development of different sexes has fascinated humankind for millennia. In plants unisexual flowers evolved multiple times independently, yet the molecular mechanisms governing sex determination remain largely elusive.

Here, we analysed sex determination in *Cannabis sativa*, one of the world's oldest and most versatile crops. *C. sativa* is primarily dioecious, with separate male and female individuals, but some cultivars are monoecious, bearing both male and female flowers on the same plant.

Using QTL mapping we identified *Monoecy1,* a locus on the X chromosome controlling monoecy vs dioecy. Strikingly, transcriptomic and genomic analyses revealed that *Monoecy1* contains three genes within just 60,000 bp which have distinct male-, female- and monoecy-specific expression patterns and encode for putative developmental regulators. We propose that the combinatorial interaction of those genes can explain not only the molecular control of monoecy vs. dioecy but also of male vs. female individuals.

30 **Main**

31 *Cannabis sativa* is mainly a dioecious crop with male and female individuals possessing XY
32 and XX sex chromosomes[1], respectively. The sex chromosomes are among the oldest in
33 flowering plants[2] and contain a non-recombining region spanning approximately 70 Mb,
34 encompassing thousands of genes[2,3]. Numerous candidate sex determination genes have been
35 proposed[4,5], but the large size and complex structure of the sex chromosomes complicates the
36 analysis of the sex determination system[4–6].

37 Beyond dioecious *C. sativa* plants, also monoecious cultivars exist which possess XX sex
38 chromosomes[1,7]. Sex determination in *C. sativa* is thus multilayered and pertains not only to
39 the genetic control of male vs. female but also to the genetics of monoecious vs. dioecious
40 plants.

41 However, monoecious and dioecious cultivars are interfertile, which offers the opportunity to
42 analyse the monoecy-dioecy sex determination system genetically using QTL analyses.
43 Towards this goal, we analyzed sex ratios in an F_2 population derived from a cross between the
44 monoecious cultivar 'Felina 32' and the dioecious cultivar 'FINOLA'[8]. The F_1 generation was
45 entirely dioecious with no monoecious individuals observed (Fig. 1a, Supplementary Table
46 S1), suggesting dominance of the dioecious over the monoecious trait. Monoecy reappeared in
47 the F_2 generation with sex frequencies depending on the F_1 mother plant, suggesting maternal
48 influence on sex inheritance (Fig. 1b). The sex expression of F_3 individuals depended
49 significantly on the sex of the F_2 mother, where monoecious mothers had approximately three
50 times more monoecious offspring than female mother plants (Fig. 1c, Supplementary Fig. S1).
51 This observation supports the notion that dioecy is dominant over monoecy.

52 To further characterize the genetics of monoecy vs. dioecy, we sequenced the entire mapping
53 population using SLAF-seq. After filtering, 1,383 high-quality markers were obtained
54 (Supplementary Figs. S2, S3). A sex determination QTL analysis using the 189 monoecious

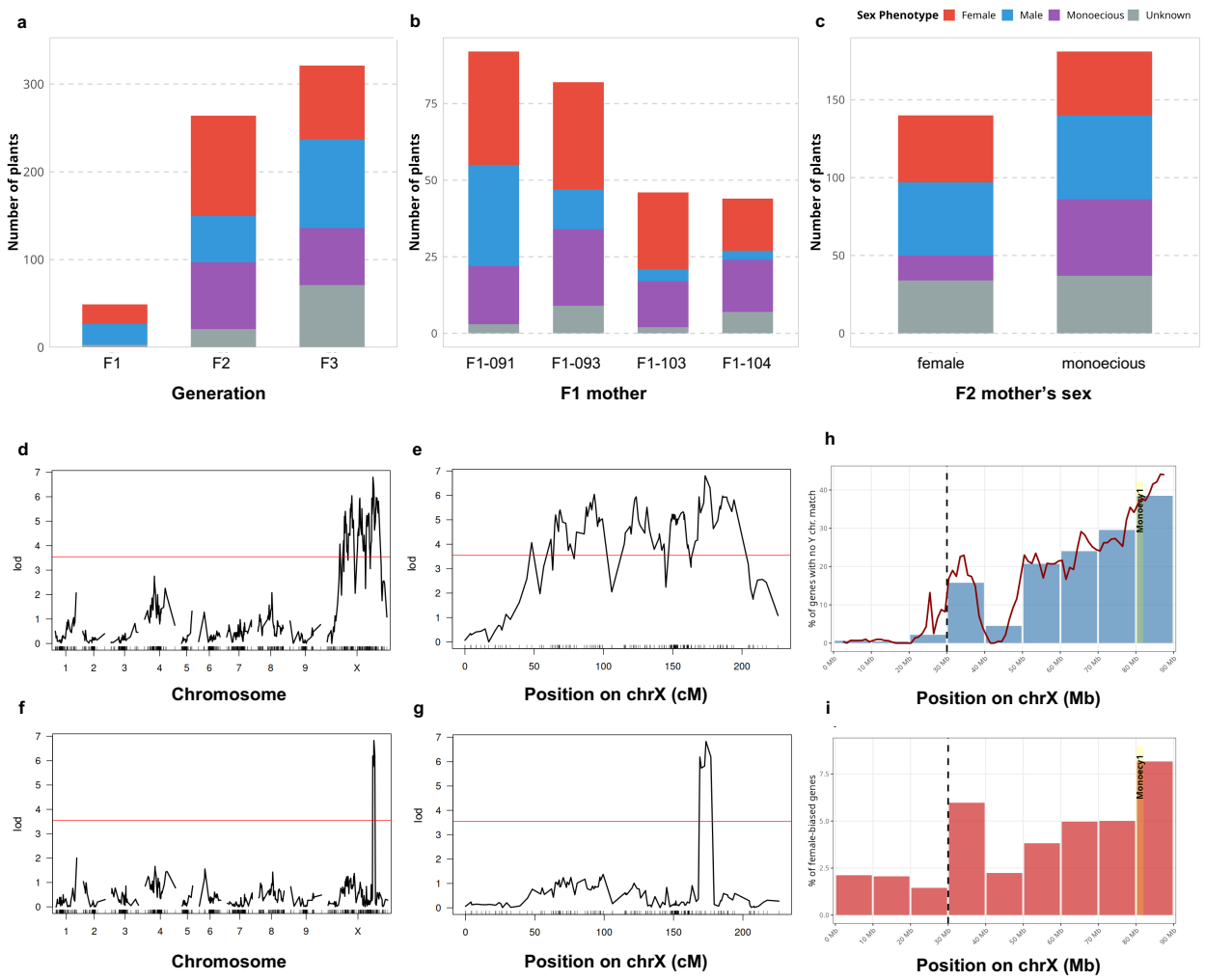

Figure 1. Sex inheritance patterns across generations, mapping and sex chromosomes genomic comparison analysis in *Cannabis sativa*.

a) Sex distribution across three generations of a 'FINOLA' (dioecious) × 'Felina 32' (monoecious) cross. The F_1 generation consisted exclusively of dioecious plants, with 49% male and 44.9% female F_1 plants. In the F_2 generation 43% of individuals were female, 29% monoecious, 20% male and 8 % did not flower. In the the F_3 generation 26% of individuals were female, 20% monoecious, 31% male and 22% did not flower. **b)** Sex distribution of F_2 plants grouped by F_1 mother plant (F1-091, F1-093, F1-103 and F104; F1-095 was the father plant in all crosses). Sex frequencies varied significantly between F_1 mother plants, with male offspring ranging from 6.8% to 35.9% (χ^2 = 23.65, df = 6, p = 0.0006). **c)** Sex distribution of F_3 plants grouped by the sex phenotype of their F_2 maternal parent. Monoecious F_2 mothers produced significantly more monoecious offspring (27% monoecious, 23% female) compared to female F_2 mothers (11% monoecious, 31% female) (χ^2 = 11.783, df = 2, p = 0.0028). The F_3 was generated by random intercrossing of F_2 individuals. **d)** Genome-wide QTL scan for the monoecy trait across all chromosomes using binary model (r/QTL, EM algorithm). The X chromosome shows the only significant LOD scores. **e)** Detailed view of the X chromosome for the monoecy trait. The peak LOD score of 6.97 was detected at marker X_80375543. **f)** Composite Interval Mapping of the monoecy trait across all chromosomes, confirming the X chromosome as the only significant QTL location. **g)** Detailed view of Composite Interval Mapping of the X chromosome, refining the QTL to a 2 Mb interval between 80-82 Mb, designated as the *Monoecy1* locus, which explains 15% of the phenotypic variance (R/qtl fitqtl). Red horizontal lines indicate the 5% significance threshold determined by 10,000 permutations. **h)** Percentage of X-chromosomal genes with no detectable homologs on the Y chromosome, analyzed in 10 Mb bins. The dashed line indicates the pseudoautosomal region boundary at 30 Mb. A sliding window analysis (red line) reveals that the highest gene loss (~40%) occurs at the terminus of the X chromosome, which includes *Monoecy1*, indicating this region represents the oldest evolutionary strata of sex chromosome differentiation. The analysis was repeated using the cultivar 'Kompolti' instead of the reference Pink pepper genome, pinpointing to the same region with the same ratio of gene loss (Supplementary Fig. S7). **i)** Distribution of female-biased genes along the X chromosome shown as percentage of genes with higher expression in females compared to males. The highest concentration of female-biased genes coincides with the region containing *Monoecy1*, suggesting a functionally conserved role in sex determination. The *Monoecy1* locus is highlighted in yellow.

55

and female plants of the F_2 population (excluding the 53 male plants as maleness is always determined by the Y chromosome, Supplementary Table S2) yielded significant associations on the non-recombining region of the X chromosome (Fig. 1d, e, Supplementary Table S3), identifying the sex chromosomes as the main cause of not only male vs. female sex determination, but also of monoecy vs. dioecy. To refine the region of interest on the X chromosome, Composite Interval Mapping was employed, confirming the highest LOD score at marker X_80375543 (Fig. 1f, g, Supplementary Table S4). This analysis pinpointed a more specific QTL region, narrowing the peak to a region that corresponds to a 2 Mb interval between 80–82 Mb on the X chromosome. We termed this locus *Monoecy1*.

Further genotype-phenotype association (Supplementary Fig. S4) analysis by comparing the genotype at the identified marker X_80375543 with the monoecy/dioecy phenotype of the F_2 plants provided support for the *Monoecy1* locus. A more detailed breakdown of genotypes at this marker and their sex phenotype classifications revealed a clear association (Supplementary Table S5): Individuals homozygous for allele X_{D803} (X_{D803}, which is the parental dioecious FINOLA genotype), exhibited an almost exclusively dioecious female phenotype (88 %, although the $X_{D803}X_{D803}$ genotype was unexpected in F_2 plants, see Supplementary Text 1, Supplementary Fig. S5, S6). In contrast, genotype X_{M803}, the parental monoecious Felina 32 genotype, was predominantly associated with monoecy (68 %). Individuals heterozygous for the parental genotypes ($X_{D803}X_{M803}$) showed an intermediate distribution (62 % female, 38 % monoecious).

The incomplete penetrance observed in our segregation analysis supports a polygenic model of sex determination. *Monoecy1* explains approximately 15% of phenotypic variance for monoecy/dioecy, suggesting additional genetic factors influence sex expression. These may include multiple loci selected through breeding, environmental interactions, and epigenetic regulation.

81 To better understand the evolutionary dynamics of the sex chromosomes and *Monoecy1* we
82 next conducted a sequence analysis of the sex chromosomes. The X and Y chromosomes likely
83 originated from autosomes that diverged over time[9]. We hypothesized that genes on the Y
84 chromosome will have been lost over evolutionary timescales while they have been retained
85 on the X chromosome. We therefore estimated the proportion of protein coding genes that were
86 retained on the X chromosome but lost on the Y chromosome. Overall, 418 out of 3,446 (12.1
87 %) of protein coding genes on chromosome X did not have a counterpart on chromosome Y.
88 The position of those hemizygous genes was not evenly distributed along the chromosome.
89 The terminal region spanning from 80 Mb to the end of the X chromosome displays the highest
90 levels of gene loss, confirmed also on an alternative cultivar genome (Supplementary Fig. S7),
91 with a conservative estimate of 40 % of X-encoded proteins lacking a Y counterpart (Fig. 1h,
92 Supplementary Table S6). Interestingly, this region also contains *Monoecy1*, which thus resides
93 in the most divergent region between X and Y chromosomes, suggesting from this region sex
94 chromosome evolution in *C. sativa* was initiated. This conclusion is further supported by
95 synonymous substitution rate (ds) analysis between X and Y chromosome homologs[2].
96 Furthermore, there is a clear pattern of the gene loss ratio increasing from one end of the X
97 chromosome to the opposite end, being near zero in the section from 0-30 Mbp. This is in
98 agreement with the position of the previously described pseudo-autosomal region[3] which is
99 still actively recombining between X and Y chromosomes.
100 An inverse analysis was also performed measuring the genes on the Y chromosome that are
101 not present on the X chromosome. This showed a much lower ratio of Y-exclusive genes
102 (Supplementary Fig. S8), consistent with the observation that the X chromosome is not actively
103 degenerating as the Y chromosome[10].
104 In our previous transcriptomics experiments[4] male and female apical meristems were used to
105 analyse if genes on the X chromosome showed a female-male biased expression. We found a

106 pattern similar to the gene loss analysis: the greatest percentage of female biased genes is at
107 the end of the X chromosome, in the region where *Monoecy1* is located (Fig. 1i).
108 We next reasoned that candidate genes controlling monoecy might be differentially expressed
109 between monoecious and dioecious cultivars. Initial comparisons between dioecious female
110 'FINOLA' and monoecious 'Felina 32' plants revealed 3,420 differentially expressed genes
111 (Fig. 2a, Supplementary Table S7, S8). To refine our analysis, we compared gene expression
112 profiles between monoecious and dioecious female F_2 plants, which reduced the number of
113 differentially expressed genes to 90 (Fig. 2b, Supplementary Table S7, S8). Only 10 of those
114 genes were located on the X chromosome with only one differentially expressed gene,
115 LOC115699949, being located in *Monoecy1*. LOC115699949 bears homologies to the gene
116 encoding the *Arabidopsis thaliana* transcription factor gene *KANADI4* (*KAN4*/*ATS*) and was
117 therefore termed *CsKAN4* (Supplementary Fig. S9). Based on publicly available RNAseq
118 data[11] we confirmed the low expression of *CsKAN4* in the monoecious cultivar 'Santhica 27'
119 (Supplementary Table S9).
120 We also extended our transcriptomic analysis to F_4 monoecious and dioecious samples of our
121 mapping population. Here, the number of genes differentially expressed between monoecious
122 and female dioecious plants was just 17 (Fig. 2c, Supplementary Table S7, S8). We further
123 identified nine genes consistently downregulated across all generations of our mapping
124 population in monoecious vs. female dioecious plants (Supplementary Table S10), one of them
125 being likely involved in the gibberellic acid (GA) pathway: LOC115706670 (encoding a
126 putative gibberellin 2-beta-dioxygenase 1), located on chromosome 1 (Fig. 2d).
127 We did not detect differential expression of *CsKAN4* in the F_4 generation (Fig. 2e), presumably
128 because F_4 samples were isolated from flowering tissue and subsequent analyses showed that
129 *CsKAN4* has a tissue- and stage-dependent expression pattern, showing relevant expression in
130 dioecious plants mainly during vegetative stages (Fig. 2f).

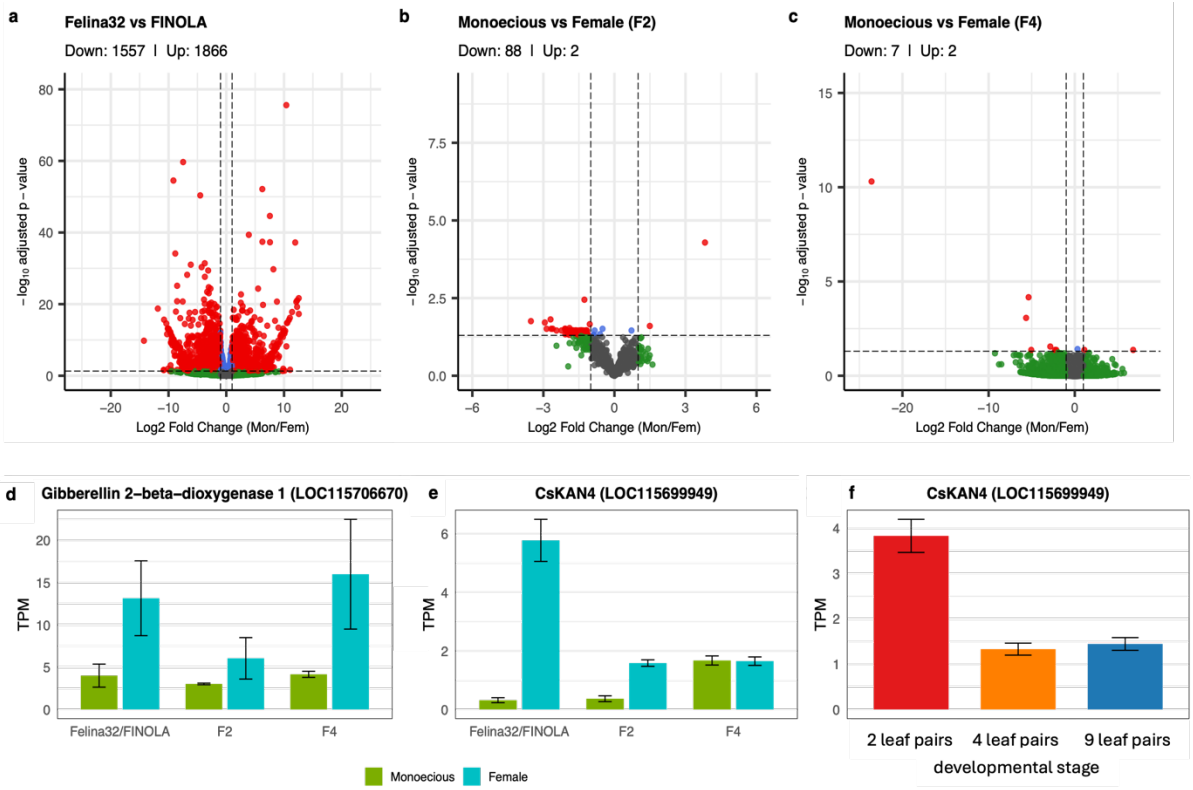

Figure 2. Transcriptomic analysis of monoecy across *Cannabis sativa* generations.
a-c) Volcano plots depicting differential gene expression between monoecious and female *C. sativa* plants across three generations. Log2 fold change (x-axis) represents expression in monoecious relative to female plants; -log10 adjusted p-value (y-axis) indicates statistical significance. Red dots represent significantly differentially expressed genes (DEGs; adjusted $p < 0.05$, $|\log2FC| > 1$). **a)** Comparison between parental lines 'Felina32' (monoecious) and 'FINOLA' (dioecious females) reveals extensive transcriptional differences. **b)** F_2 segregating population shows a marked reduction in DEGs as genetic recombination has randomised cultivar-specific traits unrelated to sex expression. **c)** F_4 generation demonstrates further reduction of DEGs. **d-f)** TPM-normalized expression of key monoecy related candidate genes across generations and developmental stages. **d)** *Gibberellin 2-beta-dioxygenase 1* (LOC115706670), thought to be degrading GA, displays consistently downregulation in monoecious plants across generations. Error bars represent standard error of the mean. **e)** *CsKAN4* (LOC115699949) shows downregulation in leaf samples of monoecious parental line 'Felina 32' and in F_2 monoecious plants compared to females, but not in flower meristem samples in F_4 plants. **f)** *CsKAN4* (LOC115699949) is expressed in early stages of plant development (2 leaf pairs) but expression is lower in later stages (4 and 9 leaf pairs), data from[4].

Strikingly, a detailed analysis of the *CsKAN4* gene locus revealed that it is just 60,000 bp away from *CsREM16*. We previously found that *CsREM16* was strongly upregulated in males as compared to females and proposed that it constitutes a molecular switch between male and female plants[4].

We therefore analysed the region containing *CsREM16* and *CsKAN4* in more detail. Between *CsREM16* and *CsKAN4* only two other genes are located, a putative long noncoding RNA (LOC133032448) and a putative pseudogene (LOC115699946) (Fig. 3a). Sequence analysis

140 revealed that the lncRNA contains a region of ~300 nt with high similarity to *CsREM16*
141 (Supplementary Fig. S10, Supplementary Text S2). We therefore termed LOC133032448
142 *lncREM16.*
143 To establish further evidence of the sex related expression pattern of the genes in this region,
144 we integrated 64 RNA-seq samples from previous experiments, encompassing monoecious and
145 dioecious samples from different developmental stages and growth conditions[4,8,12] (Fig. 3b,
146 Supplementary Table S11). This confirmed the dioecious-biased expression of *CsKAN4*, the
147 expression of *CsREM16* in female and monoecious individuals, and showed that *lncREM16* is
148 largely silenced in monoecious and female dioecious individuals. Remarkably, despite being
149 located on the X chromosome, *lncREM16* is expressed in male samples (Fig. 3b,
150 Supplementary Table S11). Furthermore, genome sequencing data from the monoecious
151 'Felina 32' cultivar revealed an insertion containing a transposable element domain ca. 8,000
152 bp upstream the start of the transcription site of *CsKAN4*, absent in the dioecious 'FINOLA'
153 cultivar (Supplementary Fig. S11).
154
155 Taken together, our data indicate that one of the oldest parts of the Cannabis X chromosome
156 contains the *Monoecy1* QTL which contributes to the development of monoecious vs. dioecious
157 individuals. *Monoecy1* contains three genes in the space of just 60,000 bp, the combinatorial
158 expression of which predicts the sex of *C. sativa* plants: in monoecious individuals *CsREM16*
159 is expressed, in female individuals *CsREM16* and *CsKAN4* are expressed and in male
160 individuals *lncREM16* and *CsKAN4* are expressed (Fig. 3c).
161 We propose that the *CsREM16, CsKAN4 and lncREM16* are the primary molecular
162 determinants of both monoecy/dioecy and male/female sex expression in *Cannabis sativa*.
163 CsREM16 and CsKAN4 are transcription factors belonging to the B3 and GARP family
164 respectively, and homologs of those proteins have been demonstrated to be involved in

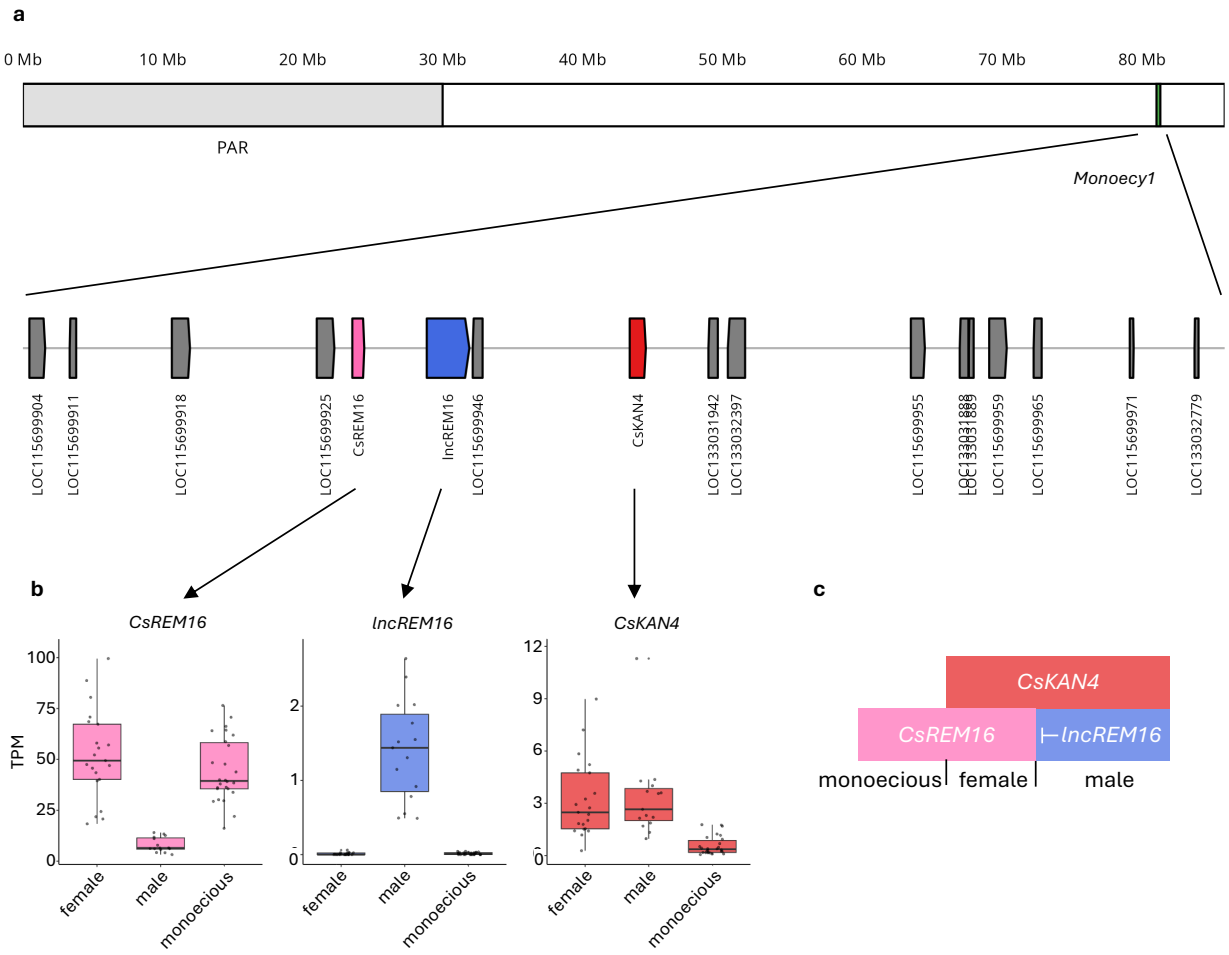

Figure 3. Genomic organization and expression patterns of the potential sex-determining genes in the *Monoecy1* locus.
a) Chromosome X ideogram and detailed gene organization within the *Monoecy1* region (81.025-81.300 Mb). The upper panel shows the X chromosome with the pseudoautosomal region (PAR, gray) and the highlighted region containing potential sex-determining genes contained in the *Monoecy1* locus (green). The lower panel displays gene arrangements with directional arrows indicating strand orientation. The putative sex-determining genes are colored according to their sex-bias classification: *CsREM16* (pink, XX-biased), *lncREM16* (blue, male-biased), and *CsKAN4* (red, dioecious-biased). **b)** Expression analysis of the three key genes using RNA-seq data from 64 'Felina 32', 'FINOLA' and F_2 samples across different sex phenotypes. Box plots show transcript abundance (TPM) for *CsREM16* (LOC115699937), *lncREM16* (LOC133032448), and *CsKAN4* (LOC115699949) in female, male, and monoecious individuals. *CsREM16* shows high expression in females and monoecious plants but low expression in males, supporting a female-determining role. *lncREM16* exhibits male-specific expression despite being X-linked. *CsKAN4* displays reduced expression in monoecious plants compared to dioecious individuals, consistent with its proposed role in monoecy determination. Individual data points are overlaid as jitter plots. Box plots show median (center line), interquartile range (box), and 1.5× interquartile range (whiskers). **c)** Combinatorial model of sex determination in *C. sativa* governed by three key regulators: *CsREM16*, *CsKAN4* and *lncREM16*. XX females express *CsREM16* and *CsKAN4* but not *lncREM16*, resulting in female flower development. XY males express *lncREM16* and *CsKAN4* but *CsREM16* is silenced (potentially by *lncREM16*), leading to male flower development. XX monoecious plants express *CsREM16* but not *lncREM16* and *CsKAN4*, enabling development of both male and female flowers on the same individual.

165

166 reproductive development[13,14]. Previous studies have indicated that sex determination in *C.*
167 *sativa* is under X chromosome gene dosage control[5], which is consistent with the X
168 chromosomal location and male or female biased expression of the genes identified here.
169 Though the exact molecular mechanisms of how the three genes control sex determination
170 remains elusive, it is interesting to note that *kan4* mutants in *A. thaliana* show an increased
171 expression of GA biosynthesis genes and an increased level of GA[15]. Likewise, our
172 transcriptomic data indicate that in monoecious individuals a gene encoding a putative
173 gibberellin 2-beta-dioxygenase 1 (*GA2ox1*) is downregulated in conjunction with *CsKAN4*.
174 GA2ox1 enzymes are involved in GA catabolism. Intriguingly, exogenous GA applications
175 induce the formation of male flowers on female plants in *C. sativa*[16] and a gene involved in the
176 GA pathway has previously been implicated in sex determination in *C. sativa*[17]. A reduction of
177 *CsKAN4* expression in monoecious plants may thus lead to an increase in GA, which in turn
178 fosters the development of male flowers.
179 The molecular role of *CsREM16* in sex determination is more difficult to pinpoint, but our data
180 indicate that its expression is strongly associated with XX as opposed to XY individuals across
181 different developmental stages and cultivars[4]. Recent data from *A. thaliana* indicate that REM
182 transcription factors recruit RNA polymerase IV to activate expression of small interfering
183 RNAs that in turn initiate DNA methylation of specific loci[18]. Also considering that some *A.*
184 *thaliana REM* mutants display strong female gametophytic defects[19,20], *CsREM16* may have
185 been co-opted to control sex determination of the sporophyte via changing the DNA
186 methylation landscape.
187 The question arises as to why *CsREM16* and *CsKAN4* are located so closely together on the X
188 chromosome. The expression of both genes may be required for the development of a plant
189 with exclusively female flowers; they therefore became linked over evolutionary timescales.
190 In this model, *CsREM16* expression is absolutely necessary for female flowers. The

requirement for *CsKAN4* however is not as strict, with monoecious plants that possess male and female flowers emerging if the gene is downregulated (Figure 3c).

A number of different (not necessarily mutually exclusive) models have been proposed recently on how sex expression and sex determination in the Cannabaceae is controlled[4,5,21,22]. Each of them makes testable predictions about the role of different X and Y chromosomal genes. It will be fascinating to see how future experiments shape our understanding of sex determination in *C. sativa* and beyond.

208 **Methods**

209 **Plant Material and Growth Conditions**

210 The 'Felina 32'x'FINOLA' mapping population was described previously[8]. Briefly, an
211 individual of the dioecious 'FINOLA' cultivar was used as pollen donor and the monoecious
212 'Felina 32' cultivar as pollen aceptor . The F_2 mapping population was generated by crossing
213 a single male F_1 as a pollinator with four F_1 female plants. The F_2 population was grown in a
214 glasshouse under natural light conditions in Dublin, Ireland from June to September 2020. Leaf
215 tissue was harvested from the F_2 population and stored at -80°C until further processed. The
216 DNA from all F_2 individuals, the F_1 individuals, and the parents was extracted using DNeasy®
217 Plant Mini Kit (Qiagen GmbH, Germany) as per manufacturer's instructions. The quality and
218 quantity of DNA was checked by agarose gel electrophoresis with a NanoDrop
219 spectrophotometer (ND-1000). The samples were prepared to a final concentration of ≥ 35
220 ng/µl and sent to BMKGene (Beijing, China) for SLAF-Seq (Specific Locus Amplified
221 Fragment Sequencing).

222

223 **SLAF-seq**

224 SALF-seq was carried out by BMKgene. Based on the pilot experiments the restriction
225 enzymes RsaI and HaeIII were selected for fragmenting the DNA into SLAF-tags of about 364
226 - 444 bp sizes.

227 On average 12640 SLAF tags were produced per sample with a mean sequencing depth of 20X.
228 Libraries generated from SLAF-tags after a series of steps were processed on the Illumina
229 platform and about 51.60 Mb reads were generated.

230

231

232

233 **Quality control**

234 The quality of SLAF-seq fastq files was checked using fastQC and summarised using MultiQC
235 softwares[23].

236 All reads exceeded quality 20 and consequently no filtering was applied at this stage. The
237 adapter sequences were trimmed by the BMKgene sequencing company and consequently no
238 further trimming was performed.

239

240 **Mapping and genotyping**

241 After quality control, the paired-end reads from female and monoecious samples were mapped
242 against the Pink pepper reference genome (GCF_029168945.1_ASM2916894v1) using BWA-
243 mem2[24]. The reads from male samples were mapped against the reference genome to which
244 the non recombining region of the Y chromosome from the Kompolti assembly[25] (*KOMPb*)
245 was appended, to avoid reads from the Y chromosome getting mapped to similar regions in the
246 X chromosome and falsely genotyping heterozygous SNPs in male X chromosome positions.
247 The resulting BAM files were filtered to keep reads of mapping quality MAPQ > 20. Due to
248 SLAF-seq reads being repetitive by design, duplicate reads were not removed. The genotyping
249 was then done using GATK 1.6.0 functions Haptlotypecaller and GenotypeGVCFs.

250

251 **QTL analysis**

252 The resulting VCF file from GenotypeGVCFs, containing SNPs from the mapping population,
253 was filtered using VCFtools 0.1.17[26] to keep biallelic markers that were genotyped in at least
254 80 % of the samples, with quality greater than 30 and with depth of at least 10, with the
255 following parameters: --min-alleles 2 --max-alleles 2 --max-missing 0.8 --minQ 30 --minGQ
256 30—minDP 10.

257 The genotypes from the filtered VCF file were then imported in R using the package onemap,
258 which allowed also to convert the VCF file into a genotype table coded by the parental origin:
259 markers from F_2 samples which were homozygous for the 'FINOLA' parent were coded as
260 AA, markers homozygous for the monoecious 'Felina 32' parent were coded as BB,
261 heterozygous markers from both parents were coded as AB. To this genotype table, the
262 phenotype data was then manually added, with the plant sex encoded as 1 for female and 0 for
263 monoecious, in columns as required by r/QTL cross format, and the cross object was imported
264 with the read.cross() function. For the QTL analysis, female and monoecious plants were used,
265 as male plants inherit the whole Y chromosome. Markers from the autosomes which show
266 severely distorted segregation (P.value < 0.001) from the expected 1:2:1 ratio were removed.
267 For each chromosome, markers were reordered based on their recombination frequency using
268 the r/QTL orderMarkers() function. The genetic map in centimorgans is then estimated with
269 the est.map() function, with parameters map.function = "kosambi",error.prob = 0.01, with
270 average and median chromosomal lengths of 105.5 cM and 97.67 cM, respectively. The
271 average marker spacing across chromosomes was <1 cM. Low recombination fractions for
272 markers on the same chromosomes (Supplementary Fig. S3), together with high LOD scores,
273 were used as indicators to confirm that markers assigned to the same linkage groups were
274 closely linked, as expected for markers within the same chromosome, to indicate the accuracy
275 of our genetic map and its concordance with the reference genome.
276 For each phenotypic trait the QTL LOD is calculated with the scanone function and method
277 "mr-argmax", and the 5 % significance threshold is generated by executing 10000
278 permutations. To refine the significant QTL to a more precise genomic interval, we additionally
279 applied the composite interval mapping method using the cim() function in the R/qtl package,
280 using the same genetic map and phenotypic dataframe, improving resolution by fitting multiple
281 marker covariates outside the testing interval to control for genetic background noise, thereby

282 increasing the power to detect QTL, and potentially distinguish closely linked loci. In our case,
283 we used the default model settings without specifying additional covariates, relying on the
284 automatic forward selection procedure from the r/qtl package to select cofactors from the
285 genome scan.
286
287 **Gene loss ratio estimate**
288 The degree of degeneration of different segments of the Y chromosome was estimated using
289 the gene loss ratio as a proxy for divergence with the X chromosome. The X chromosome was
290 divided into 10 bins of maximum 10 million nucleotides. From each bin all the proteins
291 encoded by genes in that chromosomal section were BLASTed against the genomic sequences
292 of the Y chromosome assemblies of 3 different cultivars: Kompolti (KOMPa), BooneCounty
293 (BCMa), and a landrace cultivar from Germany (GRMa), using the tblastn algorithm. If no
294 match with e-value < 0.05 was found in any of the assemblies, a gene was considered to have
295 been lost on the Y chromosome. The percentage of X chromosomal genes with no counterpart
296 on the Y chromosome was calculated for each bin and plotted.
297 We consider this approach a conservative estimate of gene loss due to the fact that a gene
298 encoded on the X chromosome belonging to a gene family can have a match on the Y
299 chromosome if a close member of this family is encoded on the Y chromosome, hence the
300 "true" gene loss ratio segment is likely to be higher than the 40 % measured.
301 For each X chromosome section bin, additionally, the percentage of female biased genes was
302 calculated based on previously published transcriptomics analyses[4]. Genes that were
303 overexpressed in females at either the vegetative or flowering stage were counted as female-
304 biased.
305
306

307 **Transcriptomics analysis**

308 RNA-seq reads from monoecious 'Felina 32' leaves samples and from dioecious FINOLA
309 females were mapped against the transcriptome of reference Pink pepper using Salmon, gene
310 expression was quantified and a comparative analysis of differential gene expression was
311 performed using the R Deseq2 package. The same comparison was performed between 3 F_2
312 female and 3 F_2 monoecious individuals with leaf samples that were obtained from the mapping
313 population[12]. An F_3 and F_4 population was generated from the F_2 population by random
314 intercrossing. From the F_4 population 8 monoecious and 8 female dioecious samples from floral
315 meristem and axillary inflorescence meristem were taken and analyzed with the same
316 aforementioned pipeline. In DESeq2, adjusted p-values (padj) are set to NA for genes with
317 extremely low expression, zero variance across samples, or in cases where the statistical model
318 fails to converge, often caused by a small limited sample size caused high dispersion. Our F_2
319 transcriptome dataset, comprising six biological replicates, may contribute to increased
320 variability and dispersion, leading to a substantial number of genes with NA padj, including
321 some with extremely low unadjusted p-values. To identify genes consistently downregulated
322 in monoecious plants across generations, we applied a filtering threshold of $p < 0.05$ and $|\log_2$
323 fold change$| > 1$, to retain differentially expressed genes regardless of padj availability.

324 We analyzed the expression of three candidate sex-determination genes (*CsKAN4*, *CsREM16*,
325 *lncREM16*) in 64 male, female, and monoecious samples from the cultivars 'Felina32' and
326 'FINOLA'. These samples were collected as part of a previously published experiment[12] and
327 are publicly available under BioProject PRJNA956491.

328

329

330

References

1. 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).
2. Prentout, D. *et al.* Plant genera Cannabis and Humulus share the same pair of well-differentiated sex chromosomes. *New Phytol.* **231**, 1599–1611 (2021).
3. Prentout, D. *et al.* An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of Cannabis sativa. *Genome Res.* **30**, 164–172 (2020).
4. Shi, J., Toscani, M., Dowling, C. A., Schilling, S. & Melzer, R. Identification of genes associated with sex expression and sex determination in hemp (*Cannabis sativa* L.). *J. Exp. Bot.* erae429 (2024) doi:10.1093/jxb/erae429.
5. Akagi, T. *et al.* Evolution and functioning of an X–A balance sex-determining system in hops. *Nat. Plants* 1–14 (2025) doi:10.1038/s41477-025-02017-6.
6. Schilling, S. *et al.* The Cream of the Crop: Biology, Breeding, and Applications of Cannabis Sativa. in *Annual Plant Reviews online* 471–528 (John Wiley & Sons, Ltd, 2021). doi:10.1002/9781119312994.apr0740.
7. Faux, A.-M., Berhin, A., Dauguet, N. & Bertin, P. Sex chromosomes and quantitative sex expression in monoecious hemp (Cannabis sativa L.). *Euphytica* **196**, 183–197 (2014).
8. Dowling, C. A. *et al.* A FLOWERING LOCUS T ortholog is associated with photoperiod-insensitive flowering in hemp (Cannabis sativa L.). *Plant J.* **119**, 383–403 (2024).
9. Charlesworth, D. Plant sex chromosome evolution. *J. Exp. Bot.* **64**, 405–420 (2013).
10. Charlesworth, D. Young sex chromosomes in plants and animals. *New Phytol.* **224**, 1095–1107 (2019).
11. Guerriero, G. *et al.* Transcriptomic profiling of hemp bast fibres at different developmental stages. *Sci. Rep.* **7**, 4961 (2017).

12. Dowling, C. A., Michael, T. P., McCabe, P. F., Schilling, S. & Melzer, R. FT-like genes in Cannabis and hops: sex specific expression and copy-number variation may explain flowering time variation. *bioRxiv* 2024.10.04.616617 (2024) doi:10.1101/2024.10.04.616617.
13. Safi, A. *et al.* The world according to GARP transcription factors. *Curr. Opin. Plant Biol.* **39**, 159–167 (2017).
14. Mantegazza, O. *et al.* Analysis of the arabidopsis REM gene family predicts functions during flower development. *Ann. Bot.* **114**, 1507–1515 (2014).
15. Gomez, M. D., Ventimilla, D., Sacristan, R. & Perez-Amador, M. A. Gibberellins Regulate Ovule Integument Development by Interfering with the Transcription Factor ATS. *Plant Physiol.* **172**, 2403 (2016).
16. Ram, H. Y. M. & Jaiswal, V. S. Induction of male flowers on female plants of Cannabis sativa by gibberellins and its inhibition by abscisic acid. *Planta* **105**, 263–266 (1972).
17. Petit, J., Salentijn, E. M. J., Paulo, M.-J., Denneboom, C. & Trindade, L. M. Genetic Architecture of Flowering Time and Sex Determination in Hemp (Cannabis sativa L.): A Genome-Wide Association Study. *Front. Plant Sci.* **11**, 569958 (2020).
18. Wu, Z. *et al.* Arabidopsis REM transcription factors and GDE1 shape the DNA methylation landscape through the recruitment of RNA Polymerase IV transcription complexes. *bioRxiv* 2025.02.21.639493 (2025) doi:10.1101/2025.02.21.639493.
19. Mendes, M. A. *et al.* Live and let die: a REM complex promotes fertilization through synergid cell death in Arabidopsis. *Development* **143**, 2780–2790 (2016).
20. VERDANDI Is a Direct Target of the MADS Domain Ovule Identity Complex and Affects Embryo Sac Differentiation in Arabidopsis | The Plant Cell | Oxford Academic. https://academic.oup.com/plcell/article/22/6/1702/6095927.
21. Carey, S. B. *et al.* The evolution of heteromorphic sex chromosomes in plants. Preprint at https://doi.org/10.1101/2024.12.09.627636 (2024).

22. Monthony, A. S. *et al.* Sex-Specific Ethylene Responses and Convergent Plasticity: Multi-Omic Insights into *Cannabis* Sexual Plasticity. Preprint at https://doi.org/10.1101/2025.06.03.657661 (2025).
23. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
24. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, Rio de Janeiro, Brazil, 2019). doi:10.1109/IPDPS.2019.00041.
25. Lynch, R. C. *et al.* Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *Nature* (2025) doi:10.1038/s41586-025-09065-0.
26. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann - Research Ireland under grant number IRCLA/2022/3294. CAD was supported by an Irish Research Council–Environmental Protection Agency Government of Ireland Postgraduate Scholarship (grant no. GOIPG/2019/1987).

405 **Figures**

406 **Figure 1. Sex inheritance patterns across generations, mapping and sex chromosomes**
407 **genomic comparison analysis in *Cannabis sativa*.**

408 **a)** Sex distribution across three generations of a 'FINOLA' (dioecious) × 'Felina 32'
409 (monoecious) cross. The F_1 generation consisted exclusively of dioecious plants, with 49%
410 male and 44.9% female F_1 plants. In the F_2 generation 43% of individuals were female, 29%
411 monoecious, 20% male and 8 % did not flower. In the the F_3 generation 26% of individuals
412 were female, 20% monoecious, 31% male and 22% did not flower. **b)** Sex distribution of F_2
413 plants grouped by F_1 mother plant (F1-091, F1-093, F1-103 and F104; F1-095 was the father
414 plant in all crosses). Sex frequencies varied significantly between F_1 mother plants, with male
415 offspring ranging from 6.8% to 35.9% ($\chi^2 = 23.65$, df = 6, p = 0.0006). **c)** Sex distribution of
416 F_3 plants grouped by the sex phenotype of their F_2 maternal parent. Monoecious F_2 mothers
417 produced significantly more monoecious offspring (27% monoecious, 23% female) compared
418 to female F_2 mothers (11% monoecious, 31% female) ($\chi^2 = 11.783$, df = 2, p = 0.0028). The F_3
419 was generated by random intercrossing of F_2 individuals. **d)** Genome-wide QTL scan for the
420 monoecy trait across all chromosomes using binary model (r/QTL, EM algorithm). The X
421 chromosome shows the only significant LOD scores. **e)** Detailed view of the X chromosome
422 for the monoecy trait. The peak LOD score of 6.97 was detected at marker X_80375543. **f)**
423 Composite Interval Mapping of the monoecy trait across all chromosomes, confirming the X
424 chromosome as the only significant QTL location. **g)** Detailed view of Composite Interval
425 Mapping of the X chromosome, refining the QTL to a 2 Mb interval between 80-82 Mb,
426 designated as the *Monoecy1* locus, which explains 15% of the phenotypic variance (R/qtl
427 fitqtl). Red horizontal lines indicate the 5% significance threshold determined by 10,000
428 permutations. **h)** Percentage of X-chromosomal genes with no detectable homologs on the Y
429 chromosome, analyzed in 10 Mb bins. The dashed line indicates the pseudoautosomal region

430 boundary at 30 Mb. A sliding window analysis (red line) reveals that the highest gene loss
431 (~40%) occurs at the terminus of the X chromosome, which includes *Monoecy1*, indicating this
432 region represents the oldest evolutionary strata of sex chromosome differentiation. The analysis
433 was repeated using the cultivar 'Kompolti' instead of the reference Pink pepper genome,
434 pinpointing to the same region with the same ratio of gene loss (Supplementary Fig. S7). **i)**
435 Distribution of female-biased genes along the X chromosome shown as percentage of genes
436 with higher expression in females compared to males. The highest concentration of female-
437 biased genes coincides with the region containing *Monoecy1*, suggesting a functionally
438 conserved role in sex determination. The *Monoecy1* locus is highlighted in yellow.
439
440 **Figure 2. Transcriptomic analysis of monoecy across *Cannabis sativa* generations.**
441 **a-c)** Volcano plots depicting differential gene expression between monoecious and female *C.*
442 *sativa* plants across three generations. Log2 fold change (x-axis) represents expression in
443 monoecious relative to female plants; -log10 adjusted p-value (y-axis) indicates statistical
444 significance. Red dots represent significantly differentially expressed genes (DEGs; adjusted p
445 < 0.05, $|\log2FC| > 1$). **a)** Comparison between parental lines 'Felina32' (monoecious) and
446 'FINOLA' (dioecious females) reveals extensive transcriptional differences. **b)** F_2 segregating
447 population shows a marked reduction in DEGs as genetic recombination has randomised
448 cultivar-specific traits unrelated to sex expression. **c**) F_4 generation demonstrates further
449 reduction of DEGs. **d-f)** TPM-normalized expression of key monoecy related candidate genes
450 across generations and developmental stages. **d)** *Gibberellin 2-beta-dioxygenase 1*
451 (LOC115706670), thought to be degrading GA, displays consistently downregulation in
452 monoecious plants across generations. Error bars represent standard error of the mean. **e**)
453 *CsKAN4* (LOC115699949) shows downregulation in leaf samples of monoecious parental line
454 'Felina 32' and in F_2 monoecious plants compared to females, but not in flower meristem

455 samples in F_4 plants. **f)** *CsKAN4* (LOC115699949) is expressed in early stages of plant
456 development (2 leaf pairs) but expression is lower in later stages (4 and 9 leaf pairs)[4].
457
458 **Figure 3. Genomic organization and expression patterns of the potential sex-determining**
459 **genes in the *Monoecy1* locus.**
460 **a)** Chromosome X ideogram and detailed gene organization within the *Monoecy1* region
461 (81.025-81.300 Mb). The upper panel shows the X chromosome with the pseudoautosomal
462 region (PAR, gray) and the highlighted region containing potential sex-determining genes
463 contained in the *Monoecy1* locus (green). The lower panel displays gene arrangements with
464 directional arrows indicating strand orientation. The putative sex-determining genes are
465 colored according to their sex-bias classification: *CsREM16* (pink, XX-biased), *lncREM16*
466 (blue, male-biased), and *CsKAN4* (red, dioecious-biased). **b)** Expression analysis of the three
467 key genes using RNA-seq data from 64 'Felina 32', 'FINOLA' and F_2 samples across different
468 sex phenotypes. Box plots show transcript abundance (TPM) for *CsREM16* (LOC115699937),
469 *lncREM16* (LOC133032448), and *CsKAN4* (LOC115699949) in female, male, and
470 monoecious individuals. *CsREM16* shows high expression in females and monoecious plants
471 but low expression in males, supporting a female-determining role. *lncREM16* exhibits male-
472 specific expression despite being X-linked. *CsKAN4* displays reduced expression in
473 monoecious plants compared to dioecious individuals, consistent with its proposed role in
474 monoecy determination. Individual data points are overlaid as jitter plots. Box plots show
475 median (center line), interquartile range (box), and 1.5× interquartile range (whiskers). **c)**
476 Combinatorial model of sex determination in *C. sativa* governed by three key regulators:
477 *CsREM16*, *CsKAN4* and *lncREM16*. XX females express *CsREM16* and *CsKAN4* but not
478 *lncREM16*, resulting in female flower development. XY males express *lncREM16* and *CsKAN4*
479 but *CsREM16* is silenced (potentially by *lncREM16*), leading to male flower development. XX

480 monoecious plants express *CsREM16* but not *lncREM16* and *CsKAN4*, enabling development

481 of both male and female flowers on the same individual.

482 **Supplementary Material**

483 **Supplementary Tables**

484 **Table S1. Sex phenotype segregation across three generations of Cannabis sativa crosses**
485 **and maternal inheritance patterns**

486 **Table S2. Genetic linkage map and quantitative trait locus (QTL) analysis for sex**
487 **determination using standard interval mapping**

488 **Table S3. Refined QTL mapping for sex determination using composite interval mapping**
489 **analysis**

490 **Table S4. Association between marker X_80375543 genotypes and sex phenotype**
491 **distribution in F_2 female and monoecious plants**

492 **Table S5. Validation of male sex phenotype through Y chromosome read mapping**
493 **analysis**

494 **Table S6. Gene content and sex-biased expression patterns in the oldest evolutionary**
495 **stratum of the X chromosome (80-86 Mb)**

496 **Table S7. Differential gene expression analysis between monoecious and dioecious**
497 **Cannabis plants across parental and segregating generations**

498 **Table S8. Significantly differentially expressed genes ($p < 0.05$, $|\log2FC| > 1$) between**
499 **monoecious and dioecious plants across generations**

500 **Table S9. Expression profiles of sex determination candidate genes in the monoecious**
501 **cultivar Santhica 27 (BioProject PRJNA435671)**

502

503 **Table S10. Core set of genes consistently downregulated in monoecious plants across**
504 **multiple generations**

505 **Table S11. Expression validation of sex determination candidate genes across 64 RNA-**
506 **seq samples from parental cultivars (BioProject PRJNA956491)**

507

508 **Supplementary Figures**

509 **Figure S1. Progeny sex distribution by maternal phenotype.**
510 Sex ratios of F_3 progeny grouped by F_2 maternal genotype (monoecious or female). Each bar
511 represents offspring from a single F_2 mother plant (identified on y-axis). Random pollination
512 occurred within the F_2 population. Monoecious F_2 mothers produced higher proportions of
513 monoecious offspring compared to female F_2 mothers, which predominantly yielded female
514 progeny.

515

516 **Figure S2. F_2 population genotype matrix.**
517 Visualization of genotypes across all F_2 individuals (rows) and markers (columns) after marker
518 reordering based on recombination frequencies. Color coding: red (homozygous FINOLA),
519 blue (heterozygous), green (homozygous Felina 32), and white (missing data). Marker
520 clustering within chromosomes demonstrates genetic linkage.

521

522 **Figure S3. Genetic map and marker linkage analysis.**
523 a) Genetic map constructed from 1,383 high-quality SLAF-seq markers spanning 1,055
524 centiMorgans (cM) across 10 chromosomes, with an average marker spacing of < 1 cM.
525 Chromosomes 1-9 represent autosomes, and chromosome X represents the sex chromosome.
526 b) Physical map showing marker distribution along the chromosomes in megabases (Mb). The

527 X chromosome spans approximately 86 Mb on the reference genome. c) Pairwise marker
528 relationships displayed as a heat map. Upper-left triangle: recombination fractions between
529 markers (brighter shades indicate lower recombination frequencies, suggesting physical
530 proximity). Lower-right triangle: LOD scores for marker linkage (brighter colors indicate
531 stronger statistical evidence for linkage). The distinct diagonal blocks confirm accurate marker
532 assignment to chromosomes.
533
534 **Figure S4. Genotype distribution by sex phenotype at X chromosome markers.**
535 Frequency distribution of genotypes (X_DX_D: homozygous FINOLA , X_DX_M: heterozygous,
536 X_MX_M: homozygous Felina 32) across X chromosome markers in F_2 individuals, separated by
537 sex phenotype. Female plants show enrichment for X_DX_D genotypes while monoecious plants
538 display higher prevalence of X_MX_M genotypes, supporting X-linked inheritance of the monoecy
539 trait.
540
541 **Figure S5. Genetic map after removal of unexpected genotypes.**
542 a) Genetic map and linkage analysis after setting unexpected X chromosome genotypes
543 (homozygous FINOLA in female plants) to missing values. Marker order and genetic distances
544 remain largely unchanged compared to the complete dataset, demonstrating robustness of the
545 genetic map. b) Recombination fractions and LOD scores show genetic map coherence after
546 removing the X_DX_D genotypes from the X chromosome.
547
548 **Figure S6. QTL analysis with filtered genotype data.**
549 Monoecy trait QTL scan after removal of unexpected AA X chromosome genotypes. While
550 overall LOD significance is reduced, the peak location remains consistent with the original
551 analysis, confirming the robustness of the *Monoecy1* locus identification.

552 **Figure S7. Gene loss analysis using alternative reference genome.**

553 Percentage of X-chromosome hemizygous genes (genes without Y chromosome homologs)
554 across X chromosome segments using the 'Kompolti' cultivar reference genome. The pattern
555 is similar to findings from the Pink Pepper reference, with increasing divergence from the
556 pseudoautosomal region toward the chromosome terminus.

557

558 **Figure S8. Y-specific gene content distribution.**

559 Percentage of Y-chromosome hemizygous genes (genes without X chromosome homologs)
560 across Y chromosome segments. Lower percentages compared to X-hemizygous genes reflect
561 the reduced gene content of the degenerating Y chromosome.

562

563 **Figure S9. Phylogenetic reconstruction of KAN4/ATS homologs.**

564 Maximum-likelihood phylogeny of KANADI family proteins from shoot.bio with CsKAN4
565 (LOC115699949) highlighted. The tree demonstrates orthology between CsKAN4 and
566 Arabidopsis KAN4/ATS (AT5G42630), suggesting functional conservation.

567

568 **Figure S10. Sequence similarity between *lncREM16* and *CsREM16***

569 a) Alignment from blast showing that a section of *lncREM16* aligns with the genomic region
570 of *CsREM16* spanning part of the third and forth exons and the entire intron between the two
571 exons b) dotplot and c) alignment from mafft of the matching regions.

572

573 **Figure S11. Transposable element presence in monoecious cultivars 'Santhica 27' and**
574 **'Felina 32'.**

575 A transposable element was identified upstream of the transcription factor *CsKAN4* in the
576 monoecious cultivar 'Santhica27'. The presence of this element was confirmed in the

monoecious cultivar 'Felina32' by mapping sequencing reads from the 'Felina32' parent to the 'Santhica27' assembly and verifying coverage in this region. The absence of mapped reads from the dioecious cultivar 'FINOLA' in this region indicates the transposable element is absent in ‘FINOLA’.