Evolution of the spinach sex-linked region within a rarely recombining pericentro-meric region

Evolution of the spinach sex-linked region within a rarely recombining pericentromeric region

Hongbing She[1†], Zhiyuan Liu[1†], Shufen Li[2†], Zhaosheng Xu[1], Helong Zhang[1], Feng Cheng[1], Jian Wu[1], Xiaowu Wang[1], Chuanliang Deng[2], Deborah Charlesworth[3*], Wujun Gao[2*], Wei Qian[1*]

[1]State Key Laboratory of Vegetable Biobreeding, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2] College of Life Sciences, Henan Normal University, Xinxiang 453007, China

[3]Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK.

† These authors contributed equally to this work.

Running title: Decipher the sex-linked region of *S. oleracea*

*** Correspondence:**

Wei Qian, qianwei@caas.cn, Tel: +86-010-62194559

Wujun Gao, gaowujun@htu.cn

Deborah Charlesworth, Deborah.Charlesworth@ed.ac.uk

Abstract

Sex chromosomes have evolved independently in many different plant lineages. Here, we describe new reference genomes for spinach (*Spinacia oleracea*) X and Y haplotypes by sequencing homozygous XX females and YY males. The long arm of the 185 Mb chromosome 4 carries a 13 Mb X-linked region (XLR) and 24.1 Mb Y-linked region (YLR), of which 10 Mb is Y-specific. We describe evidence that this reflects insertions of autosomal sequences creating a "Y duplication region" or "YDR" whose presence probably directly reduces genetic recombination in the immediately flanking regions, although both the X and Y SLRs are within a large pericentromeric region of chromosome 4 that recombines rarely in meiosis of both sexes. Sequence divergence estimates using synonymous sites indicate that YDR genes started diverging from their likely autosomal progenitors at roughly the same time as the flanking YLR stopped recombining with the XLR, about 3 MYA. These flanking regions have a higher density of repetitive sequences in the YY than the XX assembly, and include slightly more pseudogenes, compared with the XLR, and the YLR has lost about 11% of the ancestral genes, suggesting some degeneration. Insertion of a male-determining factor would have caused Y-linkage across the entire pericentromeric region, creating physically small, highly recombining, terminal pseudo-autosomal regions.

Key words: inversion, recombination, sex-linked region, gene duplication, pericentromeric region, sex chromosome turnover

Introduction

Genome regions controlling individuals' genders have originated separately multiple times in different groups of organisms, including flowering plants (Charlesworth, 1996), and sometimes these have evolved into sex chromosomes (Westergaard, 1958). The evolution of extensive fully sex-linked regions (such as the familiar mammalian XY chromosome pair (Lahn and Page, 1999) is not yet well understood, and the large non-recombining regions in plants, including *Silene latifolia* and *Cannabis sativa*, are especially puzzling because the main hypothesis to explain such regions, involving sexual selection, or other situations creating conflicts between the sexes, is less likely to apply to plants than animals.

Some sex-linked regions in flowering plants probably evolved *de novo* from functionally hermaphroditic species, which requires at least two mutations, and generates selection for their closer linkage (Charlesworth and Charlesworth, 1978). Two-gene systems must exist in species in which Y-linked mutations (or deletions) can convert males into functional hermaphrodites, as has been observed in species from five different angiosperm families, *Silene latifolia*, *Vitis vinifera*, *Carica papaya*, *Asparagus officinalis* and *Actinidia chinensis* (Westergaard, 1958; Liu et al., 2004; Picq et al., 2014; Kazama et al., 2016; Akagi et al., 2019). In *Diospyros* species (the persimmon, family Ebenaceae), however, the Y-linked factor is a duplicated copy of a gene with a female-promoting allele, whose activity the duplication suppresses, so that presence/absence of the Y-linked factor controls male versus female development; the Y-linked region recombines, except close to the duplication (Akagi et al., 2014). Y-linked duplications are also found in the genus *Populus*, in the family Salicaceae

(Müller et al., 2020; Xue et al., 2020). However, these may involve "turnovers" like those in animals, in which an autosomal sequence duplicates, creating a new Y-linked region (Pan et al., 2021), or in *Fragaria* species (family Rosaceae), in which small female-determining regions have moved between different genomic locations (reviewed by Cauret et al. 2022). Duplications in any of the scenarios just outlined can prevent chromosome pairing, creating small non-recombining regions (Charlesworth, 2019).

Here, we describe new results from spinach, a diploid plant with 2n = 12 chromosomes, has genetic sex-determination, and a homomorphic X and Y chromosome pair (Lizuka and Janick, 1962; Deng et al., 2013). Self-fertilization of occasional monoecious XY plants (males with some female flowers), can produce viable homozygous YY progeny. Therefore spinach has no extensive completely Y-linked region that has undergone genetic degeneration leading to loss of gene functions and deletions of genes, making males hemizygous for X-linked genes (as reviewed by Bachtrog 2008). Interestingly, YY males are sterile (Wadlington and Ming, 2018), so the Y must lack at least one essential male function gene carried on the X. Sequencing YY and XY plants allows reliable assembly of both the Y and X chromosomes, whereas relying on a reference genome from XX plants risks errors. Importantly, reads from Y-linked regions that are missing from the female genome will map erroneously to homologous sequences elsewhere in the genome, and duplications will often be missed. PacBio sequencing and assemblies of the spinach Y and X sex-linked regions (SLRs) were recently reported, indeed revealing a duplication, which they concluded reflects an inserted region (Ma et al., 2022).

We obtained further new information using a different long-read sequencing technology (Oxford Nanopore, ONT) and different spinach material. Our results confirm the duplication in the spinach sex-linked region, but show that it is larger than that previously detected (Ma et al., 2022), and our analyses clarify understanding of the evolution of the extensive region showing sex-linkage, supporting the recent suggestion that a duplication can instantaneously create such a region (Charlesworth, 2019), initiating evolution of the unusual properties of Y chromosomes, including accumulation of repetitive sequences and genetic degeneration.

Results

Female (XX) and male (YY) genome assembly and annotation

As described in our Supplementary Methods section, we generated high-quality assemblies for the XX and YY genomes, named Sp_XX_v1 and Sp_YY_v1 (Supplementary Table 1, Table 2, Table 3 and Table 4), with estimated assembly sizes 978 and 952 Mb, respectively (Supplementary Fig. 1, Fig. 2, Fig. 3, Fig. 4, Fig. 5, Table 5 and Table 6). Integrated assembly, including Hi-C analysis, anchored 142 contigs to the expected six chromosomes (Supplementary Fig. 6). As expected, the two genomes had high sequence identity and aligned well, apart from several inversions on the largest chromosome in both the Sp_XX_v1 and Sp_YY_v1 assemblies, which we term chromosome 4 (Supplementary Table 6, Supplementary Fig. 7). As the sex chromosome pair is the largest in the karyotype (Lizuka and Janick, 1962; Deng et al., 2013), chromosome 4 is probably the sex chromosome pair. In the XX and YY assemblies of the strains sequenced by Ma et al. (2022), this is called chromosome 1. Rearrangements between the genomes affect about 14% of the genes across all chromosome assemblies of Ma et al. (2022); such differences are unexpected for autosomes, and suggest

lower accuracy of the PacBio than ONT assemblies (see Supplementary Fig. 8B), consistent with our higher BUSCO value (97%, versus the previous value of 94.8%; Supplementary Table 2). Both our XX and YY assemblies include similar numbers of predicted protein-coding genes, 28,359 and 26,573 in our female and male genomes, respectively, approximately 90% of which were annotated (Supplementary Table 5, Table 7 and Table 8), as well as tRNAs, rRNAs, and miRNAs in similar numbers in our XX and YY assemblies. Repetitive sequence densities average about 73%, in all assemblies, including many long terminal repeat (LTR) retrotransposons (Fig. 1, Supplementary Table 9).

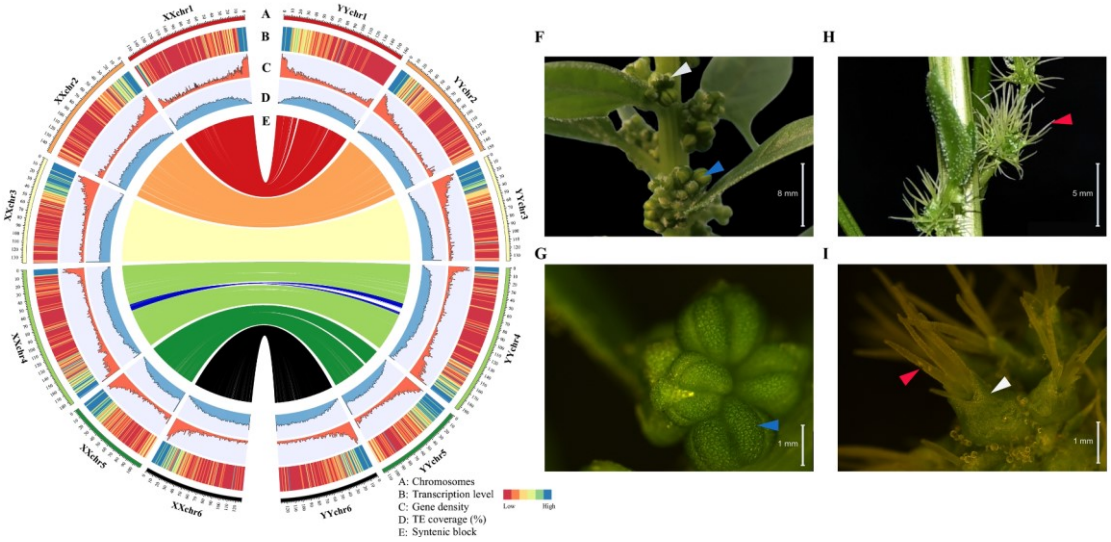

Figure 1. Genomic landscape and flowers between female and male spinach. (A) Ideogram of the chromosomes from the female (left) and male genomes (Mb scale). **(B)** Transcription levels estimated from read counts per million mapped reads in 1 Mb windows. **(C)** Gene density in 1 Mb windows. **(D)** Transposable element densities (TEs per 1 Mb). **(E)** Genes found in both the female and male assemblies. Blue lines represent the sex determining region. **(F)** and **(G)** show male (YY) flowers, and **(H)** and **(I)** show female flowers. White arrows indicate sepals, blue arrows stamens, and red arrows stigmas.

Identification of the spinach sex-determining region and evidence that it is within a large rarely recombining pericentromeric region

Aligning the X and Y chromosome sequences revealed three large inversions between the two assemblies (Supplementary Fig. 9, Fig. 10 and Fig. 11), in agreement with Ma et al. (2022), who number this chromosome 1 (Supplementary Fig. 8B, C). Both assemblies detected a Y-specific region interrupting the second inverted region (IV2), which corresponds to a contiguous X chromosome region from 85.8–98.8 Mb. As this seems likely to have been created by duplication events (as explained below), we termed it the Y-duplication region, or "YDR". It separates two Y segments, IV2-1 from 86.5 to 95.8 Mb (9.3 Mb), and IV2-2, from 105.9 to 110.6 Mb (4.7 Mb). In our assembly, the YDR occupies 10-Mb (95,894,781–105.947,428 Mb), considerably larger than the total of 2.2 Mb inserted sequences detected by Ma et al. 2022. The large amount of sequence missing from the previous assembly, including many genes compromises efforts to infer the male-determining factor(s), and, as shown in Supplementary Fig. 8C, some genes assigned to the inversion region are within the YDR in our assembly, making them better candidates than appears from the assembly of Ma et al. (2022). 49 inferred YDR genes are not present on the X. Deletion of such a large X-linked region is implausible, because it would reduce fitness in homozygotes (Manna et al., 2012); duplications into regions that rarely recombine are, however, very common, including in sex-linked regions, for example in *Silene latifolia* (Kejnovsky et al., 2006).

Dense genetic maps show that all spinach chromosomes have large rarely recombining regions at one end, probably representing the centromeres, except for the sex chromosome pair, in which it is in the middle (in both male and female meiosis, Fig. 4A, B below; Supplementary Fig. 4 and Fig. 12). Repeat density analysis (Fig. 1, see Supplementary Methods, Fig. 13, Fig. 14, and Table 14) supports these candidate centromeric regions on all autosomes,

as high densities are expected in low recombination regions, including centromeres (Charlesworth et al., 1994); this probably accounts for the uncertain assemblies at the chromosome ends (Supplementary Fig. 4). The extremely large gene-poor/TE-rich region at one end of each autosome in both the female and male assemblies (Figure 1) are probably pericentromeric regions in which recombination is rare. We further confirmed the autosomes' inferred acrocentric or submetacentric morphology by using FISH analysis (Supplementary Methods, Fig. 15, Fig. 16 and Table 14).

Our X and Y chromosome assemblies both also have very extensive gene-poor/TE-rich regions, again suggesting that these recombine rarely and have accumulated repetitive sequences. These are in the middle regions (Fig. 1A–D), consistent with FISH experiments (Fig. 2) supporting this chromosome pair's metacentric morphology. FISH analysis using seven single-copy sequences from the YDR, four from IV2-1, and three from the IV2-2 region as probes further confirmed our assembly, with the IV regions flanking the YDR and the counterpart X chromosome position. The YDR does not coincide with the centromere, but is within the pericentromeric rarely recombining region of chromosome 4's long arm (Supplementary Fig. 15 and Fig. 16).

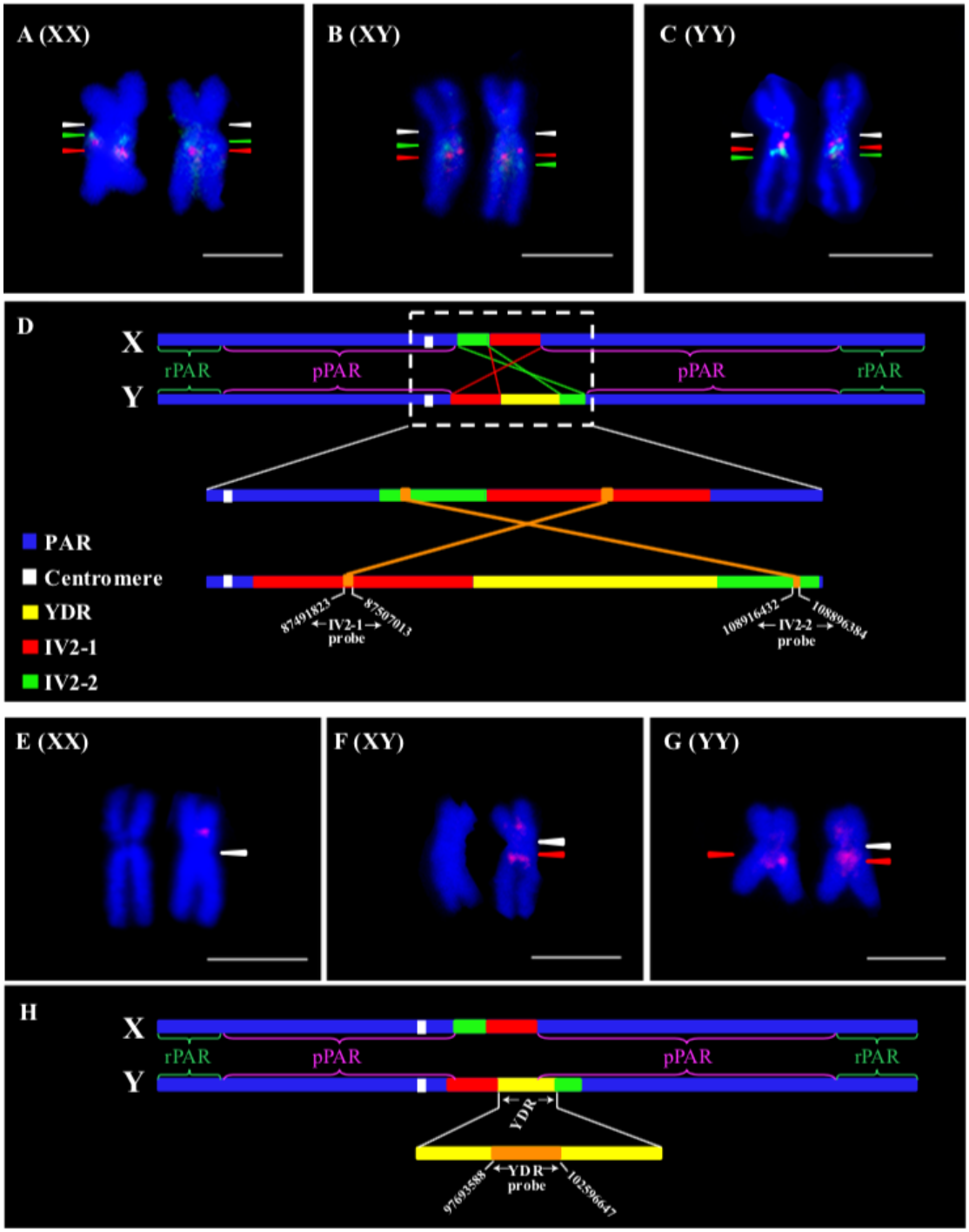

Figure 2. FISH mapping of IV2 and YDR on the metacentric sex chromosomes in spinach metaphase preparations. (A–D) show FISH mapping of IV2 in **(A)** XX female, **(B)** XY male, and **(C)** YY male individuals. White arrows represent the centromere position . The single-copy sequences of IV2-1 and IV2-2 are labeled with red and green signal, respectively. **(D)** Ideogram showing the physical position of the IV2 probes. Based on repeat densities, we identified a candidate centromeric repeat sequence, following Su et al. (2021), which identified that rnd-4_family-324 is enriched near all the centromeres but absent elsewhere in the genome, see Supplementary Fig. 13 and Fig. 14. **(E–H)** FISH mapping of the YDR in indi-

viduals of known sexes, as follows: **(E)** XX female, **(F)** XY male, and **(G)** YY male individuals. White and red arrows represent the centromere and the YDR signal, respectively. Unexpected red signal at the top of the chromosome is due to the fact that the probe sequences within the YDR are not all single-copy sequences. (**H)** Ideogram showing the physical position of the YDR probes. Bars = 5 μm. PAR: pseudoautosomal region; rPAR: recombining PAR; pPAR: pericentromeric PAR. rPAR and pPAR were inferred based on relationship between genetic map and the sex chromosomes (see Supplementary Fig. 12)

The rarity of recombination makes it difficult to define the spinach completely sex-linked region, or SLR, precisely, and the inference of its location by Ma et al. (2022) is imprecise (Supplementary Fig. 8). We therefore took advantage of four markers used in spinach breeding, T11A, V20A, S5.7, and S9.5, that have been reliably male-specific for many years in more than 10,000 individuals genotyped annually. We mapped these and 9 previously reported sex-linked markers or variants in male-associated BAC sequences (Akamatus et al., 1998; Liu et al., 2015; Kudoh et al., 2017) to our new XX female and YY male assemblies (Supplementary Table 10). All four male-specific markers, and all five male-associated BAC markers, map in the YDR, and the other markers in the IV2-1 and IV2-2 regions, except for SO4 to the right of IV2 (Fig. 3A and an expanded version in Supplementary Fig. 17).

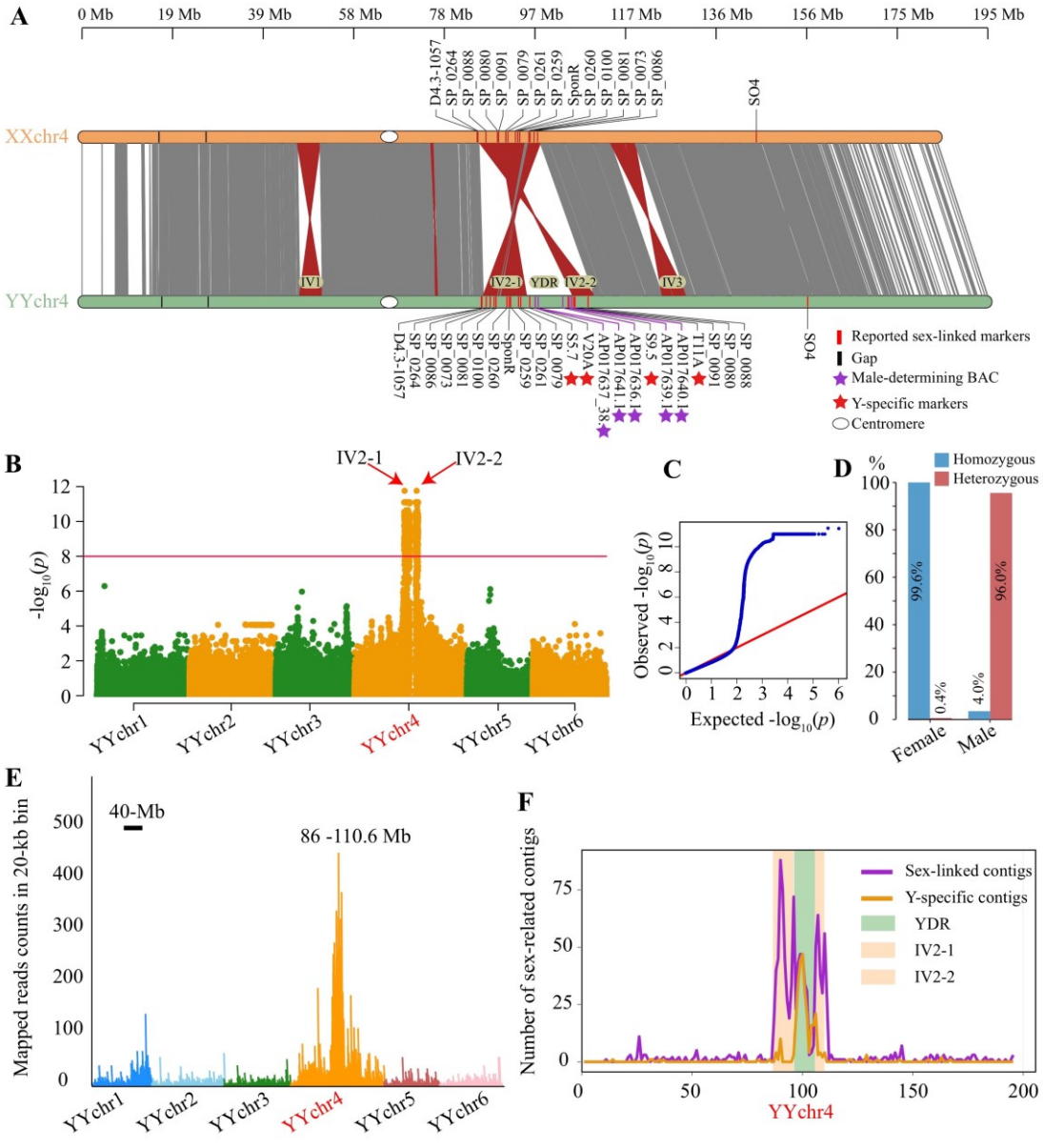

Figure 3. Identification and validation of the fully sex-linked region. (A) Sequence alignment of the X and Y chromosomes. The grey and red lines between the two chromosome assemblies indicate non-inverted and inverted alignments, respectively. Previously reported markers were mapped to the sex chromosome assemblies, as shown above the X chromosome and below the Y. Four genes whose sequences include Y-specific markers are present only in the Y assembly, and are indicated with asterisks. IV: inversion; YDR: Y-duplication region. **(B)** Manhattan plot based on the GWAS results using 20 female and 41 male individuals. The

vertical axis represents the -$\log_{10}$(*P* value) from the test of association with sex for each SNP. The red line indicates the $\alpha < 0.05$ value. **(C)** Quantile-quantile (QQ) plot from the GWAS analysis. **(D)** The percentages of homozygous and heterozygous SNPs that are significantly associated with sex **(E)** Mapping of male-specific reads to the Sp_YY_v1 genome. **(F)** Enrichment of contigs containing sex-associated markers on the Y chromosome.

Two independent approaches using multiple individuals, GWAS (Klein et al., 2005) and a reference-blind approach using k-mers (Akagi et al., 2018) independently identified the same SLR. 1,100 SNPs associated with sex in our GWAS were all located within IV2-1 or IV2-2 (Fig. 3B, C; Supplementary Table 11). 99.6% were homozygous in females, while approximately 96.0% were heterozygous in males, consistent with spinach's known male heterogamety (Fig. 3D; Supplementary Table 11, Fig. S12). Furthermore, reads including male-specific k-mers were enriched in the 24.6 Mb YY chromosome 4 region from 86–110.6 Mb (Fig. 3E), similar to the 17 Mb chromosome 1 region in the PacBio assembly previously identified as sex-linked, from 145 to 162 Mb (Ma et al., 2022), although this lacks much of the YDR (see above). These reads assembled into 4,659 initial contigs, of which 1,362 showed sex-linkage and 304 were Y-specific (see Supplementary Methods); 70.9% of the sex-linked ones mapped to the IV2-1 and -2 regions, and 93.4% of those specific to the YY assembly mapped to the YDR (Fig. 3F; Supplementary Table 12 and Table 13). The boundaries defined by sequences containing these k-mers coincide with those of a region within which genes present in both the X and Y assemblies show higher divergence than the rest of the chromosome (Fig. 6A below).

Validation of the YDR

Together, these findings show that only the part of chromosome 4 including the YDR and IV2 region is a completely Y-linked region (YLR), which must include the male-determining factor. Coverage of the YDR confirmed its presence in all males, but at half the coverage values of the other regions (Supplementary Fig. 18 and Fig. 19). All the rest of chromosome 4 had equal coverage in 20 male and 20 female accessions, like the autosomes (Supplementary Fig. 18). Other than the IV2 region, chromosome 4 consists of vast genetically pseudo-autosomal regions (PARs) that recombine rarely, which we term "pericentromeric PARs" (denoted by pPAR in Fig. 2), and physically small terminal PARs with high recombination rates.

Analysis of linkage disequilibrium (LD) on the X and Y chromosomes using 20 female and 41 male accessions confirms that the region physically close to the SLR recombines even less than the rest of the pericentromeric region, and that LD extends across a wider region of the Y than the X, consistent with recombination being especially infrequent in males (Fig. 4C, D).

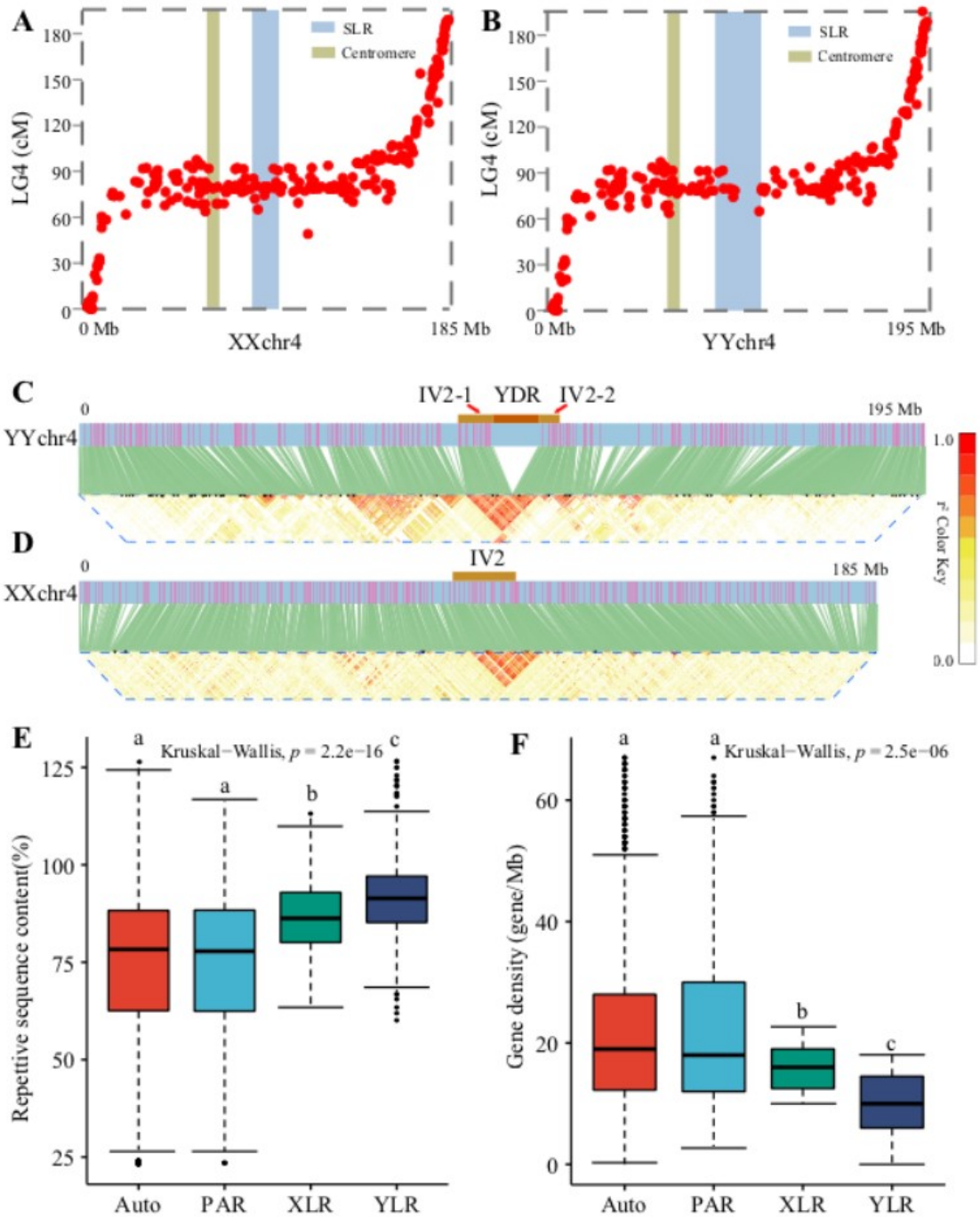

Figure 4. Sex chromosomes exhibited large pericentromeric region. Alignment of the **(A)** X and **(B)** Y chromosomes with the SLAF marker linkage genetic map (Qian et al., 2017). The horizontal axis shows physical positions in the sex chromosome assemblies, with the blue and brown bars indicating the SLR and centromere positions, respectively, and the estimated genetic map positions are shown on the vertical axis. LG: linkage group. **(C and D)** Linkage disequilibrium (LD, measured as r^2) analysis of one thousand random SNPs on the Y and X chromosomes using our 61 accessions. SNPs are depicted as pink and green lines. Boxplot of **(E)** repetitive sequences coverage in 200-Kb sliding windows with 20-Kb steps, and **(F)** gene

density. XLR: X-linked region; YLR: Y-linked region. Auto: autosomes. PAR: pseudoautosomal region. the letters above the boxplots indicate significance in Kruskal-Wallis tests.

Repetitive sequence accumulation in the YLR and XLR

Within the YLR, 14 Mb is syntenic with the X-linked region (XLR), but the YDR insertion adds 10 Mb. A higher repetitive sequence content (Fig. 4E) also increases the YLR size (yellow region in Fig. 2H), as 88.9% of its sequence consists of repetitive sequences (of many types), versus 85.0% of the XLR. Even though the autosomes and the PARs also include large pericentromeric repeat-rich portions (the large blue regions in Fig. 2D and H), only 73.2% of the autosomes and 74.1% of the PARs are repetitive (Supplementary Table 15 and Table 16). In the terminal recombining portions of the PARs the repeat density averages only 56.6%, much lower than the pericentromeric PARs (82.22%; Supplementary Table 16). The high XLR repeat density may therefore simply reflect its location entirely within the pericentromeric region whose recombination rate is low enough for repeats to have accumulated (Fig. 4A; Supplementary Fig. 4). The YLR's higher repeat content, and higher LD (see above), show that recombination is lower than in the pericentromeric PARs, allowing Y-specific variants to be maintained.

Genes in the SLRs and their expression

Having defined the completely Y-linked region, and shown that it is larger than previously inferred by Ma et al. (2022), we sought to identify candidate male-determining genes and their origins. We identified 211 protein-coding genes in the XLR (Supplementary Table 17, Fig. 20), and 245 in the YLR, 182 in the IV2 regions of synteny with the X, and 49 in the

YDR and absent from the X, of which 18 represent parts of TEs (red font in Supplementary Table 18).

The YDR could represent either a duplication, or a deletion from the X. Under the deletion hypothesis, Y-specific genes should mostly be single-copy genes whose only copy is in the YDR, whereas the duplication hypothesis predicts that they should often have autosomal copies (paralogs) that might include progenitors of the YDR genes. 16 of the 49 YDR genes have paralogs (half with multiple autosomal copies), while 33 are single-copy male-specific genes without candidate autosomal progenitors (Supplementary Fig. 21). Although this test appears to favour the deletion hypothesis, it does not do so unambiguously. The Discussion section evaluates the duplication hypothesis, using further results which we describe next.

A single event that duplicated all 49 genes, or even all 16 with likely autosomal progenitors, is unlikely, and indeed the paralogs are scattered on all spinach autosomes, with no autosome carrying multiple candidate progenitor genes located close to one another (Supplementary Table 20, which excludes TE sequences). It is more likely that multiple independent duplications occurred into this region. This probably did not involve individual genes duplicating into the region by retrotransposition of mRNAs, since 44 of the 49 have at least one intron (Supplementary Table 18); the autosomal putative progenitors of four of the five intronless YDR genes are also intronless.

Because the male determining factor is within the completely Y-linked region identified above, candidate genes cannot be evaluated by finer genetic mapping. We therefore compared gene expression in males and females, using RNA-seq (Fig. 5A). Excluding 27 pseudogenes, there are 218 YLR genes, of which 19 were differentially expressed in males,

including 10 YDR genes absent in females and thus completely male-specific (Fig. 5B, C). Four of the 16 YDR genes with autosomal paralogs (red font in Supplementary Table 21) are highly expressed in both sexes. Other YDR genes were not expressed, suggesting that they are non-functional, the commonest fate of duplicated genes (Walsh, 1995).

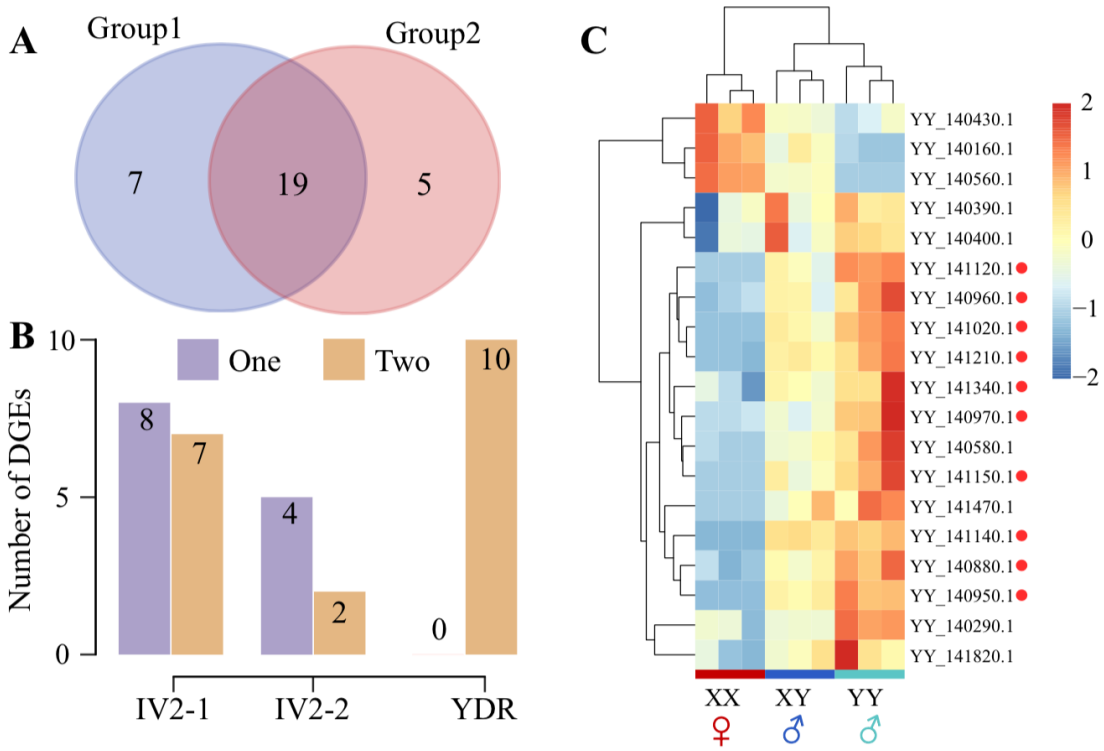

Figure 5. Expression patterns of genes within the sex-linked region on the Y chromosome. (A) Venn diagram of the DEGs in two groups, each comparing males and females to test for DEGs using mRNA-Seq: group 1 includes a set of three XX females and three XY males, while group 2 compared the same three females with three YY individuals. **(B)** Plot showing whether the YLR DEGs were detected in just one of the groups just defined, or in both groups. **(C)** Heat map of the DEGs found in both groups, shown separately for the three genotypes in the SLR. The colors denote gene the expression levels ($\log_2$(TPM+1)). Red dots indicate genes expressed only in males. Seven and five genes are DEGs in one comparison but not the other, probably reflecting the male sterility of the YY genotype, which implies different expression of some genes between XY and YY males.

Among 182 genes in the IV2 XLR and YLR regions with expression high enough for reliable conclusions about differences between individuals with and without a Y chromosome, 4 IV2-1 and 2 IV2-2 genes showed higher expression in the XY and YY males than in XX females (Supplementary Table 21) and are also candidate male-determining genes, or possible masculinized genes (genes that evolved higher expression in males after the region became Y-linked, or ones affecting male functions that duplicated into the region). Five genes with inconsistent results in the two comparisons are less promising male-determination candidates, as are the three IV2 genes with consistently lower expression in males in both comparisons, and 7 in only one of them.

We also tested whether the IV2 region has evolved changes predicted after a male-determining factor appears on a chromosome. We have described evidence that the 10-Mb YDR does reflects an insertion on the Y chromosome, rather than deletion of part of the X chromosome (see also the Discussion section). This created a completely Y-linked region, consistent with 14 non-TE genes being found only in the Y sequence. In non-recombining regions, insertions and deletions are predicted to occur much more often than in other genome regions. Consistent with this, 27 genes were found only in the X sequence, suggesting possible losses from the Y-linked region. Also as expected, the Y sequence included more pseudogenes (27 genes, or 11%) than the X (only eight pseudogenes, or 3%; Supplementary Fig. 20) (Fisher's exact test, p =0.006). However, the ratio of nonsynonymous to synonymous site divergence (K_a/K_s) between homologous sequences in the XX and YY assemblies is similar for the IV2 region and the PARs (Supplementary Fig. 22). Either recombination stopped

too recently, or Y-linkage in IV2 is incomplete, and rare recombination has prevented degeneration.

The YDR evolved at about the time the X- and Y-linked inversion 2 sequences started diverging

Synonymous site divergence (K_s) values between the YLR and XLR sequences in the IV2 region flanking the YDR reflects the number of generations since recombination between them stopped. The weighted mean for these 182 XY gene pairs is 0.035, much higher than the value of 0.003 between genes in the large XX and YY chromosome 4 pericentromeric region sequences (Fig. 2A, Fig. 6A, Supplementary Table 22 and Table 23). The higher diversity is therefore evidence that (despite their low recombination rate in our mapping families) the regions flanking the YDR must undergo recombination, at least occasionally, supporting the conclusion above.

Divergence between the YDR genes and their autosomal paralogs averages 0.043, suggesting recent appearance. However, several YDR sequences have much higher divergence, and some much lower, so that some duplication events are probably independent, and occurred at different times, and involve progenitors in different genome regions (Fig. 7b below). Four adjacent YDR genes may have duplicated more recently than the others (based on no synonymous differences and high identity values based on all site types in their coding regions, see Supplementary Table 20). The mean divergence is slightly higher than the Y-X Ks estimate, whose median is 0.034 (Fig. 6B). This is consistent with a duplication that created the YDR and caused a lack of pairing, preventing Y-X recombination (excluding the

four low values just mentioned, the YDR-paralog Ks value is very similar to the Y-X Ks). A molecular clock rate of 7.0×10^{-9} (Xu et al., 2017), suggests YDR and IV2 divergence between 2.56 and 3.10 MYA.

Despite this Y-linked region's lack of major degeneration, it has accumulated repetitive sequences. TE densities within IV2 are 83.5% for the X sequence, and 84.9% for the Y, versus 70.8% and 71.9% in the XX and YY chromosome PAR sequences, respectively. The YLR density of 10 genes per Mb is lower than the density of 16 in the XLR, or that in the autosomes or the rest of chromosome 4 (19 and 28 genes/Mb, respectively; Fig. 4F; Supplementary Table 19). When Y-X recombination stopped, heterozygous insertions could drift to high frequencies within the IV2 Y population, even if they reduce fitness (unlike recombining regions, which can become homozygous and eliminate TE insertions with deleterious effects), explaining these density differences. The insertion ages, estimated using LTR-retrotransposons within the YDR, IV2 region, and the rest of chromosome 4 (40 and 56 intact LTR elements, respectively, mostly LTR/Copia elements, especially in the YDR; Fig. 6C) are 3.15 and 2.97 MYA, respectively in the YDR and IV2 regions, supporting the time of the YDR creation estimated above. Insertions in the flanking regions are younger (2.28 MYA, based on 869 elements; Fig. 6D, E), as expected.

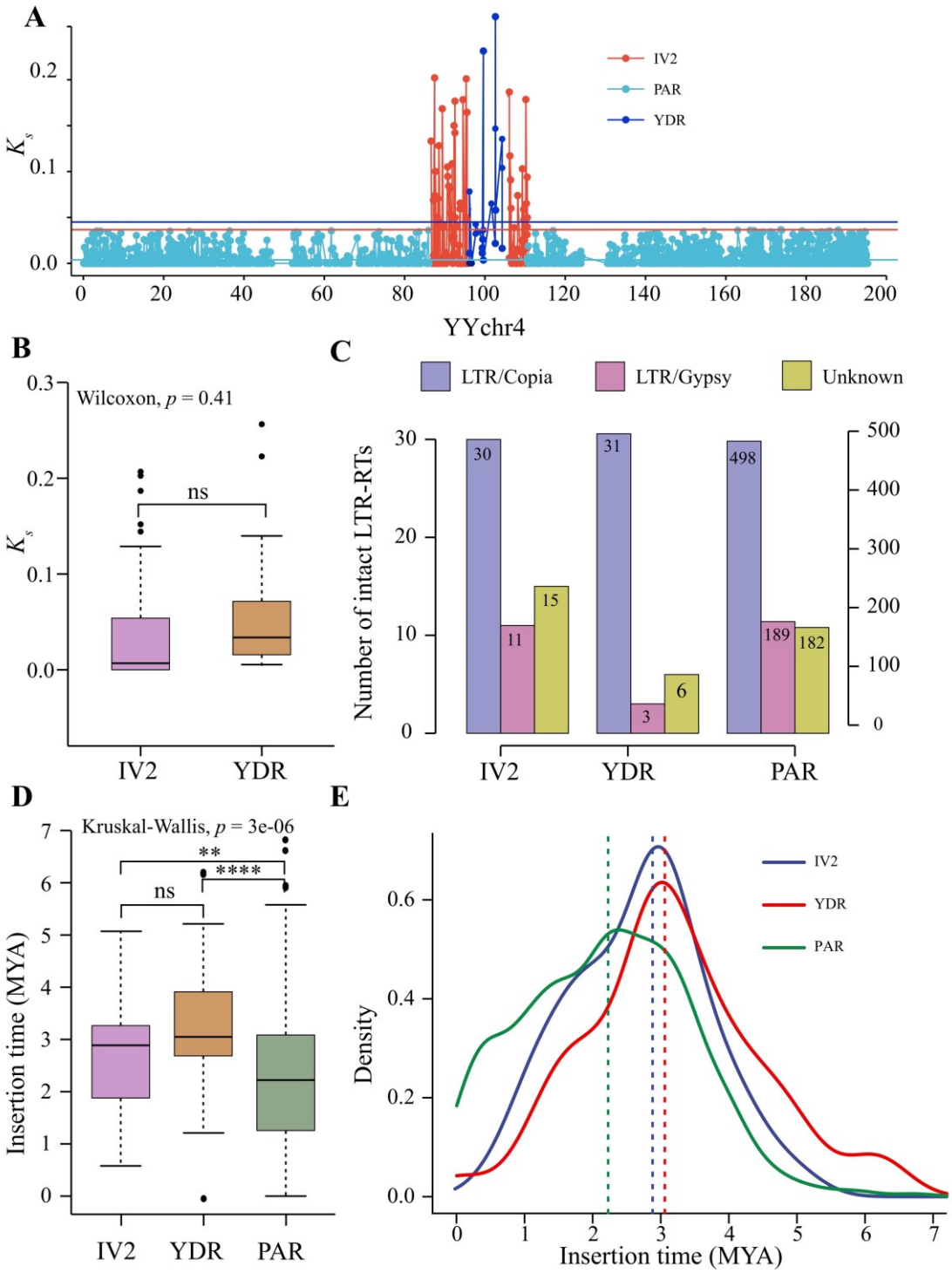

Figure 6. Divergence estimates of IV2 and YDR genes. (A) The K_s values between X/Y gene pairs in the XX and YY assemblies (red) and in the PARs (teal), and blue indicates divergence between YDR genes and their paralogs. Each dot represents one gene pair. The horizontal lines show mean values for the three regions, weighted by the sequence lengths. **(B)** Wilcoxon test comparing K_s values of genes in the IV2 and the YDR regions; ns indicates no significant difference. **(C)** Number of intact LTR-RTs in IV2, YDR and PAR. **(D)** Boxplot of insertion times in the same three regions, the two YLR portions of chromosome 4, IV2, YDR,

and the PAR, showing significant differences between the PAR and both YLR portions, by Kruskal-Wallis tests. **(E)** Distribution of estimated LTR insertion times in the three regions indicated in the figure. The dotted lines indicate the median insertion times MYA: million years ago.

Origin of sequences within the spinach SLR

To understand the evolution of the SLR, we searched for regions of synteny in three outgroup species in the Amaranthaceae with genome assemblies. Spinach is more closely related to *Chenopodium quinoa* than *Beta vulgaris* (see Supplementary Fig. 23A and Supplementary Table 24). These three species differ by inter- and intra-chromosome rearrangements (Supplementary Fig. 23B). A diploid outgroup species, *Amaranthus hypochondriacus*, had regions corresponding to the entire IV2-1 region (but not the smaller IV2-2), scattered on multiple chromosomes (Fig. 7A), but parts of both were found on a single *C. quinoa* chromosome, as expected if YDR sequences inserted into an ancestral genome region like the spinach X with adjacent IV2-1 and -2 regions (Fig. 7A). 12 out of 16 spinach YDR genes have putative paralogs with sequences much closer than any possible *C. quinoa* orthologs (Supplementary Fig. 24), consistent with duplications since these lineages split. Duplications appear to have brought three non-TE genes into the YDR (Supplementary Table S25). However, the matches cover only 565 kb, in 73 sequences ranging from 5 to 16.9 kb, scattered across the spinach genome, while 45 YDR sequences match transposable elements, and the origins of the other genes are currently unknown (Fig. 7B; Supplementary Table 25).

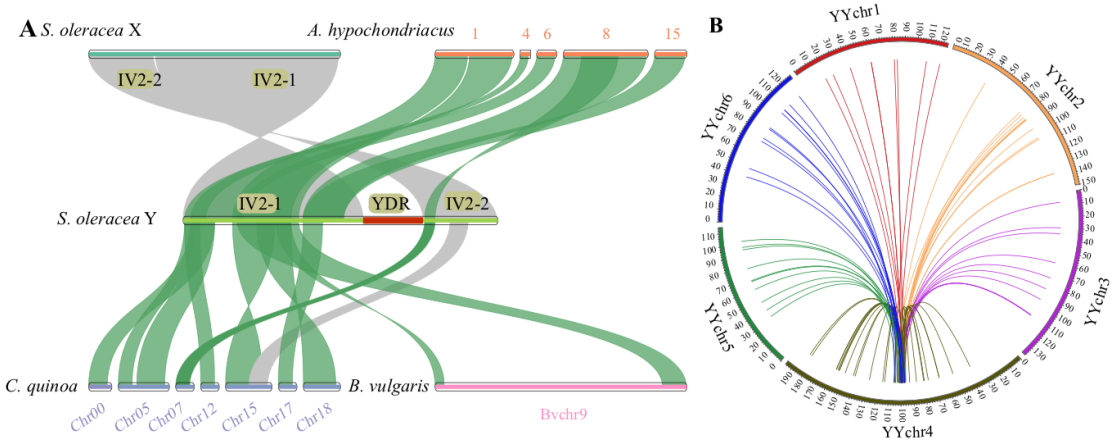

Figure 7. Origin of the SD region in spinach. (A) Syntenic relationships of the sex-determining region between *S. oleracea* and the outgroup species, *A. hypochondriacus*, *B. vulgaris* and *C. quinoa*. Corresponding syntenic regions are drawn. **(B)** Segmental duplications into the YDR originating from progenitor sequences scattered throughout the whole genome. Each line within the circle connects sequences of lengths ≥5 kb and showing ≥97% identity.

Discussion

Possible origins of the spinach male-determining region: duplication or X deletion

The YDR probably arose by duplication of a male-determining gene into the Y chromosome (or a gene that evolved and became a male-determining gene). The spinach YDR's small size is consistent with a lack of major sex chromosome heteromorphism, as the pericentromeric PARs have accumulated repetitive sequences and become very large, so that the extra size contributed by the YDR is cytologically undetectable. The spinach sex-determining system may have evolved from an ancestor that was already dioecious, with a single male-determining factor and femaleness the default state in its absence. Then, as described in the Introduction section, a turnover event could have occurred, in which the maleness factor moved or duplicated to its present location, or a different gene took over control of sex-determination. Alternatively, a femaleness factor elsewhere in the genome (analogous to

401 the persimmon *MeGI* gene) could have duplicated to form the maleness factor, like the per-
402 simmon *OGI* gene (Akagi et al., 2014).

403 A deletion of X-linked genes is less plausible. If the ancestral population was cosexual
404 (hermaphroditic or monoecious), a deletion could abolish an essential male function, creating
405 females, but the non-deleted chromosome would determine cosexuality, and another closely
406 linked mutation would be required to produce males. This conclusion is supported by the ob-
407 servation that the YDR of the spinach relative, *S. tetrandra* is smaller than those of *S. turke-*
408 *stanica* and *S. oleracea* (She et al., 2021). The single copy YDR gene, *YY_141140.1*, is
409 male-specific in all three species (She et al., 2021), and is therefore a candidate for the
410 male-determining factor, while another YDR gene, *YY_140950.1*, was not detected in *S.*
411 *tetrandra*, eliminating it as a candidate. The *YY_141140.1* is a bZIP transcription factor re-
412 quired for positive regulation of flowering in *Arabidopsis thaliana* (Li et al., 2019), though its
413 function in spinach has not yet been tested.t

414 The YDR could have evolved by movement of a pre-existing male-determining region
415 with several genes, as in *Fragaria* (Cauret et al., 2022) and *Actinidia* (Akagi et al., 2023).
416 However, the results from *S. tetrandra* suggest that the current spinach YDR has probably
417 expanded recently, consistent with being larger than the largest *Fragaria* translocation (about
418 31 kb in *F. chiloensis*, including sequences flanking the pre-translocation progenitor regions).
419 Indeed, there is no unique autosomal progenitor for the spinach YDR, whose genes either di-
420 verged from their paralogs recently, or have no evident paralog. K_s of the *S. oleracea*/*S. tur-*
421 *kestanica YY_141140.1* from its *S. tetrandra* ortholog is 0.079, suggesting an origin of the
422 male-determining gene before the split of most YDR genes and their paralogs (Supplementary

Fig. S25). This suggests that the male-determining factor may pre-date the split of the different spinach lineages, as inferred in *Mercurialis* (Gerchen et al., 2021). An initial small insertion that moved or created a male-determining factor/gene could have been followed by duplications of further sequences from other autosomes into the region (as is known to occur in completely non-recombining plant genome regions, Gisby and Catoni, 2022).

Effects of a turnover on recombination

Because the YDR is hemizygous, it cannot recombine with the X, and the insertion probably also prevents pairing in the flanking regions, stopping recombination within the adjacent IV-2 regions, with nearly 200 X-linked counterpart genes in the wider SLR. This SLR is within the very large pericentromeric region, occupying most of each spinach chromosome, in which genetic mapping does not detect crossovers (Qian et al., 2017). Interestingly, given the lack of available data on recombination in such pericentromeric regions, these regions of spinach chromosome 4 must sometimes recombine, as we detect sharp changes in Y-X divergence between the SLR and its flanking pericentromeric regions, defining the latter as genetically pseudo-autosomal (Fig. 2, Fig. 3, Fig. 4 and Fig. 6). Thus the spinach XY pair has two kinds of pseudo-autosomal regions, the very large rarely recombining pericentromeric region with high repetitive sequence densities, and physically small terminal regions with very high recombination rates in males, as in mammalian PARs. Although a low recombination rate probably pre-dates the presence of the male-determining factor that created Y linkage on spinach Chr4, the IV regions have evolved higher repetitive sequence densities, as expected after recombination with their X-linked counterparts stopped, creating an isolated and permanently heterozygous YLR.

SLRs within regions with ancestrally low recombination have been reported in other plant species (Wang et al., 2012; Pilkington et al., 2019; He et al., 2021; Rifkin et al., 2021; Xue et al., 2021) and fish (Bergero et al., 2019). Unlike papaya, whose SLR has undergone subsequent recombination suppression (Wang et al., 2012), the spinach SLR is about 20 Mb from the chromosome 4 centromere, and there is no definitive evidence for subsequent recombination suppression. The duplication may have directly reduced recombination in the region. The inversion may also contribute, and may already have been present in the ancestral rarely recombining region into which the YDR sequences inserted; high repeat densities and some linkage disequilibrium are also seen in females' XLR (Figs. 4D and E), and the inversion might even still be segregating among spinach X chromosomes as the selective disadvantage of rearrangement heterozygosity is small in such regions, and rearrangements are common. The female studied in our FISH experiment was not heterozygous for this inversion, but larger samples are needed. It is often speculated that inversions are the cause of recombination suppression between Y and X chromosomes (e.g. (Lahn and Page, 1999), as is theoretically possible (Charlesworth and Charlesworth, 1978). However, inversions have only rarely been shown to have contributed to suppressing recombination (Lemaitre et al., 2009; Peichel et al., 2020), and many observed inversions evolved after recombination became suppressed.

Candidates for the spinach male-determining factor

We detected many genes in the spinach pericentromeric regions, which are often viewed as gene deserts. However, it is now recognized that genes exist in such regions, in organisms including *Drosophila melanogaster* (Corradini et al., 2007) and plants, including the

well-studied case of barley (Baker et al., 2014). Sex differences in expression would be predicted to evolve at some of the many genes in the rarely recombining PARs, but this has not yet been tested. However, if the duplication on spinach chromosome 4 created a Y-linked region by bringing in a male-determining factor, the YDR should contain it. As this region includes few genes, it may be possible to identify this factor in spinach. The YDR genes here named *YY_140960.1* (*EIF3A*) and *YY_141020.1* (FCF1) are potential candidates, based on expression in male but not female flowers (Ma et al., 2022), and this is confirmed by our expression results (Supplementary Table 21, Fig. S8C); their third candidate was not found in our study, but *YY_141140.1* was identified above as a new candidate. *YY_141020.1* has no autosomal or X-linked copies, but *YY_140960.1* has autosomal copies and an X copy (Supplementary Table 20). Synonymous site divergence of *YY_140960.1* from its closest autosomal copy is 3.1%, based on a 665 bp sequence (Supplementary Table 18 and Table 20), is similar to the divergence of Y-linked non-YDR sequences from their X-linked alleles, but considerably lower than the new candidate male-determiner, *YY_141140.1*, which is male-specific in the outgroup species, *S. tetrandra*. As mentioned above, the less diverged sequences could reflect translocations after a male-determining factor inserted.

Methods

Plant materials

A male from an inbred spinach line bearing some hermaphrodite flowers (10S15) was self-pollinated to produce XX females, XY males, and YY male individuals (Fig. 1F–I), to lower the frequency of heterozygous sites in the genome, to aid assembly. Low frequencies of

heterozygous sites were indeed achieved, as estimated by Illumina data (0.076% for female, and 0.119% for male individuals, see Supplementary Fig. 1). A single XX female and YY male identified in our previous study (She et al., 2021) were used for ONT Nanopore sequencing and *de novo* assembly. Sixty-one spinach accessions from different inbred lines (Supplementary Table 26) cultivated by our team were used for new resequencing and GWAS analyses to test which variants in the long-read sequences are consistent Y-X differences, as described below. Subsets of 15 female and 16 male individuals were used to identify male-specific k-mers (MSKs). All these plants were grown in the field at the Institute of Vegetables and Flowers (IVF) of the Chinese Academy of Agricultural Sciences (CAAS) in spring 2018. Each plant's floral morphology was inspected visually to determine its sex phenotype.

Library construction and sequencing

Fresh leaves from each individual were collected and frozen in liquid nitrogen prior to high-quality genomic DNA extraction following (Murray and Thompson, 1980). The DNA quality and concentration were assessed using electrophoresis on 1.0% agarose gels and an ND-2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). The genomic DNA was used for Illumina and ONT library construction and sequencing.

ONT libraries were constructed with the Ligation Sequencing Kit 1D (SQK-LSK109) protocol. First, DNA fragments were repaired using NEBNext FFPE Repair Mix. Then, after end-repair and 3'-adenylation with the NEBNext End repair/dA-tailing Module reagents, the adapters were ligated using the NEBNext Quick Ligation Module (E6056). Finally, the ONT

509 library was sequenced using a Nanopore PromethION P48 instrument by BioMarker (Beijing,
510 China).

511 To correct possible errors in the ONT long reads, Illumina genomic libraries were se-
512 quenced from the same XX and YY individuals. These libraries, with insert sizes of 300 bp,
513 were constructed using the Illumina Genomic DNA Sample Preparation kit according to the
514 manufacturer's instructions (Illumina, San Diego, CA, USA) and were sequenced using a
515 HiSeq 2500 instrument (Illumina) to generate 150-bp paired-end reads by BioMarker (Beijing,
516 China).

517 In order to use Hi-C sequencing data to validate the ordering and orientation of the con-
518 tigs, we created a male (XY) individual Hi-C library from the inbred line 12S4 (different from
519 the 10S15 line used for ONT and Illumina sequencing) using the method of Xie et al. (2015),
520 which was sequenced on the Illumina HiSeq X Ten platform.

521 For RNA sequencing, flower buds from female (XX), XY male, and YY male individu-
522 als from the inbred line 10S15 were sampled at the early inflorescence stage (60 days after
523 sowing) on June 5, 2018. Three biological replicates were sequenced from each individual.
524 Total RNA was extracted using TRIzol reagent (Thermo Fisher Scientific Inc.) and purified
525 by phenol/chloroform extraction. cDNA was synthesized using a TranScript One-Step gDNA
526 Removal and cDNA Synthesis Kit (TransGen Biotech, Beijing, China). The mRNA libraries
527 were constructed using the protocol of Zhong et al. (Zhong et al., 2011) and sequenced on a
528 HiSeq 2500 (150-bp paired-end reads) by BerryGenomics (Beijing, China).

529 **Genome assembly of both female (XX) and male (YY) individuals**

First, a genome survey was performed based on ~50× Illumina reads from the chosen YY male and XX female individuals, using Genome Characteristics Estimation (GCE v1.0.0) (Liu et al., 2013). Then, Nextdenovo (v2.2.0; https://github.com/Nextomics/NextDenovo) was used to assemble these genomes. Specifically, the raw ONT long reads of the female (~58×) and male (~42×) were assembled using Nextdenovo (parameters "read_cutoff = 1k, minimap2_options_raw = -x ava-ont –minlen 1000, random_round = 20"); three rounds of polishing were applied using Racon (v1.3.3) (Vaser et al., 2002) with the XX or YY ONT long reads, and two rounds of polishing using NextPolish (v1.2.2) (Hu et al., 2019) with the Illumina reads (~50×). Additionally, NECAT (v0.01) (Chen et al., 2021) was used with default parameters to assemble the two individuals and correct and extend the contigs. The assembled contigs were further polished using Medaka (v0.11.5; https://github.com/nanoporetech/medaka) and NextPolish.

A high-density genetic linkage map (Qian et al., 2017) was used to orient and order the contigs, and adjacent contigs were separated by 100 Ns. The female (XX) and YY male assemblies are termed Sp_XX_v1 and Sp_YY_v1, and those of each chromosome are designated with prefixes XX and YY, respectively, followed by the chromosome number used in the linkage map.

The Benchmarking Universal Single-Copy Orthologs (BUSCO) program was used to perform a preliminary assessment of the assembly results, using the embryophyta_odb10 database (Waterhouse et al., 2017). Analysis of synteny between the female (Sp_XX_v1) and male (Sp_YY_v1) assemblies was performed using NUCmer (v3.1, parameters "-c 100") (Kurtz et al., 2004). The alignment blocks were then filtered to remove mapping noise, and

552 one-to-one alignments were identified by delta-filter with parameter settings "-r -q". A
553 dot-plot of aligned regions >2000 bp was generated by D-GENIES (Cabanettes and Klopp,
554 2018). To assess the female and male genome quality, we filtered the Hi-C reads using fastp
555 (v0.20.0) (Chen et al., 2018), and then aligned the clean reads to the two assemblies using
556 BWA (v0.7.17, (Li, 2013) with the parameters "-SP5M –F 256". We then filtered the align-
557 ment file as described in ALLHIC (v0.9.13, (Zhang et al., 2019) with default parameters. Fi-
558 nally, the ALLHIC_plot was used to show the contact map for the entire genome in 500 kb
559 windows.

560 **Annotation of repetitive sequences**

561 A *de novo* repetitive-element library was constructed using RepeatModeler (v.1.0.11;
562 http://www.repeatmasker.org/RepeatModeler.html). The repetitive elements in the female
563 (XX) and male (YY) genomes were annotated using RepeatMasker (v.4.0.7, (Zhang et al.,
564 2012). We identified LTR-RTs using LTR_Finder (v1.07, Zhao and Hao, 2007) with the pa-
565 rameters "-D 15000 –d 1000 –L 7000 –l 100 –p 20 –C –M 0.85." LAT assembly index (LAI)
566 scores were then calculated using the LAI program (Ou and Jiang, 2018) of the LTR retriever
567 package, with 3-Mb sliding windows and 300-Kb steps (Ou et al., 2018). The insertion times
568 of elements with intact LTR sequences were estimated using LTR_retriever, using the substi-
569 tution rate per base of 7.0×10^{-9} estimated by Xu et al. (2017). We predicted whether the
570 centromere regions were at the ends of chromosomes, or in the middle (for the sex chromo-
571 some pair) based on the distribution of the transposable element densities on each chromo-
572 some, and narrowed down the region using the top 10 transposable elements subfamilies as
573 described by Su et al. (2021).

Protein-coding gene prediction

The female (XX) and YY male genome sequences were used for gene prediction using MAKER v2.31.6 (Cantarel et al., 2008), which combines evidence from *ab initio*, transcript mapping, and protein homology-based predictions. *Ab initio* gene predictions were also produced using SNAP v2006-07-28 (Korf, 2004), AUGUSTUS v3.3.2 (Stanke et al., 2006), and GeneMark v4.57_lic (Besemer and Borodovsky, 2005). To map transcripts, 18 spinach RNA-Seq data sets from NCBI (accession number SRP076521), and three from each of our XX female and YY male sequenced individuals, plus an XY male, also from the 10S15 progeny (see above) were combined, and then filtered using fastp v0.20.0 (Chen et al., 2018) (Supplementary Table 7). The RNA-Seq data were then aligned to the XX and YY genomes using STAR v1.5.2 (Dobin et al., 2012) and assembled using Stringtie v2.1.2 (Kovaka et al., 2019) and gffread (Pertea and Pertea, 2020). Protein sequences from sugar beet (also in the family Amaranthaceae), and the more distantly related *Arabidopsis thaliana* (Brassicaceae), and from the Swiss-Prot database, were used for protein homology-based prediction. Predicted genes without start codons, stop codons, or with a length less than 50 bp were filtered out from further analyses.

Analysis of the sex chromosomes

We performed an initial comparison of our reference X and Y chromosome sequences using minimap2 (v2.15-r915-dirty; https://github.com/lh3/minimap2) with default parameters (Li, 2018). A total of 24 previously inferred spinach sex-linked sequences (Arumuganathan and Earle, 1991; Groben and Wricke, 1998; Kudoh et al., 2017; Wadlington and Ming, 2018; Okazaki et al., 2019) were used in this study. Specifically, 15 sex-linked sequences that are

present on both chromosomes but exhibit SNP or simple sequences repeats (SSR) differences, and four sequences found only in the Y chromosome sequence or in five bacterial artificial chromosome (BAC) sequences believed to be in the male-determining (as described in the Results section, these suggest the presence of a duplication into the Y-linked region, named "YDR", see below). To identify the sex-linked region (SLR), we used BLASTN ("-evalue 1e-10") to align these sequences to the female and male assemblies. To validate the presence of a male-specific region, the XX female ONT long reads were also mapped to the Y chromosome, using minimap2 (v2.15-r915-dirty).

Identification of sex-related contigs using reference-free k-mer analysis

A total of 15 female and 16 male individuals from the different inbred lines (Supplementary Table 26) were used to identify sex-related sequences through a comparison of the specific lengths of sequence numbers between the female and male sequences as described by Akagi et al. (Akagi et al., 2018). Details are in the Supplementary Methods file.

Genome-wide association studies

All 61 cultivated spinach accessions mentioned above (20 females and 41 males) were used for resequencing and GWAS analysis based on the Sp_YY_v1 assembly. High-quality SNPs were identified using PopSeq2Geno (Cheng et al., 2016), and GWAS was performed using the compressed mixed linear model within the GAPIT package of R. The details are in the Supplementary Methods. To analyze linkage disequilibrium (LD) on the Y chromosome, we randomly selected 1,000 high-quality SNPs from those that the GWAS analysis inferred to be in sequences on the chromosome that carries the Y-linked region. LD was estimated for the Y using LDBlockShow (v1.40) (Dong et al., 2020). Similarly, we identified high-quality SNPs

on the Sp_XX_v1 assembly, and randomly selected 1,000 SNPs to analyze X chromosome LD.

Inversion analysis

Based on the alignment of the reference X and Y chromosome assemblies, three large inversions were detected between the X and Y (named "IV" regions below). To validate the inversions, the XX and YY corrected ONT reads generated from NECAT were aligned to the Sp_XX_v1 and Sp_YY_v1 genome assemblies using minimap2 (v2.15-r915-dirty) with parameter "-x asm5". The mapped reads with the highest mapping quality (60) and coverage in the aligned sequences on both sides of the inferred inversions >= 98% were used to confirm the inversion borders. Specifically, to infer an inversion on the Y chromosome, three criteria were required (i) at least 10 YY ONT corrected reads were required to map to the candidate inversion border, (ii) the border outside the inversion was required to share a homologous region on the X, while, (iii) inside the border, the X counterpart sequence was inverted.

FISH analysis

The YDR and IV2 regions just mentioned (and described in detail in the Results section below) were confirmed using FISH analysis in female, and male (XY and YY) individuals from the inbred line 10S15. Seven YDR fragments, and seven IV2 sub-regions, four in IV2-1, and three in the IV2-2 region (see Results) were tested using PCR amplifications with the primer sequences summarized in Supplementary Table 27. Mitotic metaphase spreads from the root tips were prepared as described by Li et al. (Li et al., 2019), see detailed information in the Supplementary Methods.

Synteny and divergence analysis

Synteny blocks between the female and YY male assemblies were identified and visualized using the Python version of MCscan (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) (Tang et al., 2008) with the parameter "--minspan=30". Images of repetitive sequences, gene density, and synteny blocks of the two genomes (from the female and YY male) were drawn using Circos (Krzywinski et al., 2009). Paralogous genes within the YDR were detected by mapping genes within the YDR to the Sp_YY_v1 and Sp_XX_v1 assemblies using BLASTN (cutoff identity <92%, and coverage <75%). We then identified homologs in *C. quinoa* for each of the 16 genes, using TBtools (Chen et al., 2020). Phylogenetic trees of homologs and paralogs were estimated using IQ-TREE (v2.0.3, Lam-Tung et al., 2015).

Synteny analysis between Sp_YY_v1 and Monoe_Viroflay (Cai et al., 2021), Cornell-No.9 (Ma et al., 2022) was performed using NUCmer (v3.1, Kurtz *et al.*, 2004) with parameters "-c 100". A dot-plot of aligned regions >2000 bp was generated by D-GENIES (Cabanettes and Klopp, 2018). Gene synteny analysis was performed using the Python version of MCscan (Tang et al., 2008).

The yn00 program of PAML (v4.9j, Yang, 2007) was used to estimate synonymous site divergence (K_s) and non-synonymous site divergence (K_a) between X/Y gene pairs in the IV2 region (see Results section). K_s estimates were also used to relate the time of Y-X divergence to that between 16 genes in the YDR region and their putative autosomal paralogs. The divergence times in years were calculated based on the substitution rate of 7.0×10^{-9} for *Arabidopsis thaliana*, following Xu et al. (2017).

Identification of pseudogenes in the sex-linked region

We identified pseudogenes within the sex-liked region using liftoff (v1.6.3, Shumate and Salzberg, 2020). Specifically, we anchored genes on the X (Y) to the Y (X) chromosomes, and then classified coding sequences with premature stop codons or frame shift as pseudogenes.

RNA-Seq Analysis

mRNA-Seq reads from three XX females, three XY males and three YY males from inbred line 10S15 were used to identify genes expressed at high levels in flowers, and find genes with expression differences between the sexes (termed "DEGs"). The experiment included two comparisons: group1 compared three XX females and three XY males, and group2 compared three XX females and three YY males. These mRNA-Seq reads were aligned to the Sp_YY_v1 sequence using HISAT2 (v4.8.2) with default parameters (Kim et al., 2015). Read counts per gene were generated using featureCounts (v2.0.1, Yang et al., 2014), and converted to transcripts per million (TPM) using a custom Python script (https://github.com/Spinach-lab/Sp_YY_v1-Sp_XX_v1). Expression levels were compared between male (XY)/YY and female individuals using the R package DESeq (v1.14, Anders and Huber, 2010). $P < 0.05$ and fold change (FC) > 2 were used to identify DEGs. Venn diagrams of the DEGs in the three groups were generated using BMKCloud (http://www.biocloud.net/).

Data and code availability

The genome assemblies of male and female, resequencing reads, and transcriptome sequencing reads used in the study have been deposited in the Genome Warehouse in the BIG Data Center (BIGDataCenterMembers, 2017), Beijing Institute of Genomics (BIG), Chinese

Academy of Sciences, under accession numbers GWHBOUO00000000, GWHBOUV00000000, and CRA004067 that are publicly accessible at http://bigd.big.ac.cn. All scripts used in the study were deposited in https://github.com/Spinach-lab/Sp_YY_v1-Sp_XX_v1.

Acknowledgements

This work was performed in the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, Beijing, China, and was supported by the Chinese Academy of Agricultural Sciences Innovation Project (CAAS-ASTIP-IVFCAAS, CAAS-ZDRW202103), China Agricultural Research System (CARS-23-A-17), Beijing Joint Research Program for Germplasm Innovation and New Variety Breeding (G20220628003), the Youth innovation Program of Chinese Academy of Agricultural Sciences(Y2023QC07).

Author Contributions

WQ designed the study. HS conducted the experiments. HS and ZL analyzed the data. SL performed FISH analysis. HS wrote the manuscript. WQ, DC, JW, WG, SL, ZL, CD and XW revised the manuscript. DC reformulated the manuscript. WQ, ZL, HZ, and ZX prepared the samples. FC helped analyze the data.

Competing Interests

No conflict of interest declared.

References

Akagi T, Henry IM, Ohtani H, Morimoto T, Beppu K, Kataoka I, Tao R (2018) A Y-encoded suppressor of feminization arose via lineage-specific duplication of a cytokinin response regulator in kiwifruit. Plant Cell **30:** 780-795

706 **Akagi T, Henry IM, Tao R, Comai L** (2014) A Y-chromosome–encoded small RNA
707 acts as a sex determinant in persimmons. Science **346:** 646-650
708 **Akagi T, Pilkington SM, Varkonyi-Gasic E, Henry IM, Sugano SS, Sonoda M,**
709 **Firl A, McNeilage MA, Douglas MJ, Wang T, Rebstock R, Voogd C,**
710 **Datson P, Allan AC, Beppu K, Kataoka I, Tao R** (2019) Two
711 Y-chromosome-encoded genes determine sex in kiwifruit. Nat Plants **5:**
712 801-809
713 **Akagi T, Varkonyi-Gasic E, Shirasawa K, Catanach A, Henry IM, Mertten D,**
714 **Datson P, Masuda K, Fujita N, Kuwada E, Ushijima K, Beppu K, Allan**
715 **AC, Charlesworth D, Kataoka I** (2023) Recurrent neo-sex chromosome
716 evolution in kiwifruit. Nature Plants
717 **Akamatus T, Suzuki T, Uchimiya H** (1998) Determination of male or female of
718 spinach by using DNA marker.**:** Japan: Sakata no tane KK
719 **Anders S, Huber W** (2010) Differential expression analysis for sequence count data.
720 Genome Biology **11:** R106
721 **Arumuganathan K, Earle ED** (1991) Nuclear DNA content of some important plant
722 species. Plant Molecular Biology Reporter **9:** 208-218
723 **Bachtrog D** (2008) The temporal dynamics of processes underlying Y chromosome
724 degeneration. Genetics **179:** 1513-1525
725 **Baker K, Bayer M, Cook N, Dressig S, Dhillon T, Russell J, Hedley PE, Morris J,**
726 **Ramsay L, Colas I, Waugh R, Steffenson B, Milne I, Stephen G, Marshall**
727 **D, Flavell AJ** (2014) The low-recombining pericentromeric region of barley

restricts gene diversity and evolution but not gene expression. Plant Journal **79:** 981-992

Bergero R, Gardner J, Bader B, Yong L, Charlesworth D (2019) Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. Proceedings of the National Academy of Sciences of the United States of America **116:** 6924-6931

Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research **33:** W451-W454

BIGDataCenterMembers (2017) The BIG Data Center: from deposition to integration to translation. Nucleic Acids Research **45:** D18-D24

Cabanettes F, Klopp C (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ **6:** e4958

Cai XF, Sun XP, Xu CX, Sun HH, Wang XL, Ge CH, Zhang ZH, Wang QX, Fei ZJ, Jiao C, Wang QH (2021) Genomic analyses provide insights into spinach domestication and the genetic basis of agronomic traits. Nature Communications **12:** 7246

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research **18:** 188-196

Cauret CMS, Mortimer SME, Roberti MC, Ashman TL, Liston A (2022) Chromosome-scale assembly with a phased sex-determining region resolves

features of early Z and W chromosome differentiation in a wild octoploid strawberry. Genes Genomes Genetics **12:** jkac139

Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. Current Biology **6:** 149-162

Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. The American Naturalist **112:** 975-997

Charlesworth B, Charlesworth D (1978) Selection of new inversions in multi-locus genetic systems. Genetical Research **21**

Charlesworth B, Sniegowski P, Stephan WJN (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371:** 215-220

Charlesworth D (2019) Young sex chromosomes in plants and animals. New Phytol **224:** 1095-1107

Chen CJ, Chen H, Zhang Y, Thomas HR, Frank MH, He YH, Xia R (2020) TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Molecular Plant **13:** 1194-1202

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34:** 884-890

Chen Y, Nie F, Xie SQ, Zheng YF, Xiao CL (2021) Efficient assembly of nanopore reads via highly accurate and intact error correction. Nature Communications **12:** 60

Cheng F, Sun R, Hou X, Zheng H, Zhang F, Zhang Y, Liu B, Liang J, Zhuang M, Liu Y, Liu D, Wang X, Li P, Liu Y, Lin K, Bucher J, Zhang N, Wang Y,

Wang H, Deng J, Liao Y, Wei K, Zhang X, Fu L, Hu Y, Liu J, Cai C, Zhang S, Zhang S, Li F, Zhang H, Zhang J, Guo N, Liu Z, Liu J, Sun C, Ma Y, Zhang H, Cui Y, Freeling MR, Borm T, Bonnema G, Wu J, Wang X (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. Nature Genetics **48:** 1218-1224

Corradini N, Rossi F, Giordano E, Caizzi R, Verni F, Dimitri P (2007) Drosophila melanogaster as a model for studying protein-encoding genes that are resident in constitutive heterochromatin. Heredity **98:** 3-12

Deng C, Qin R, Gao J, Cao Y, Li S, Gao W, Lu L (2013) Identification of sex chromosome of spinach by physical mapping of 45s rDNAs by FISH. Caryologia **65:** 322-327

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2012) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29:** 15-21

Dong SS, He WM, Ji JJ, Zhang C, Yang TL (2020) LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Briefings in Bioinformatics **22:** 1-6

Gerchen JF, Veltsos P, Pannell JR (2021) Recurrent allopolyploidization, Y-chromosome introgression and the evolution of sexual systems in the plant genus *Mercurialis*. Phil. Trans. R. Soc. B **377:** 20210224

Gisby JS, Catoni M (2022) The widespread nature of Pack-TYPE transposons

794 reveals their importance for plant genome evolution. Plos Genetics **18:** e1010078
795 **Groben R, Wricke G** (1998) Occurrence of microsatellites in spinach sequences
796 from computer databases and development of polymorphic SSR markers.
797 Plant Breeding **117:** 271-274
798 **He L, Jia KH, Zhang RG, Wang Y, Shi TL, Li ZC, Zeng SW, Cai XJ, Wagner ND,**
799 **Horandl E, Muyle A, Yang K, Charlesworth D, Mao JF** (2021)
800 Chromosome-scale assembly of the genome of Salix dunnii reveals a
801 male-heterogametic sex determination system on chromosome 7. Molecular
802 Ecology Resources **21:** 1966-1982
803 **Hu J, Fan J, Sun Z, Liu S** (2019) NextPolish: a fast and efficient genome polishing
804 tool for long-read assembly. Bioinformatics **36:** 2253-2255
805 **Kazama Y, Ishii K, Aonuma W, Ikeda T, Kawamoto H, Koizumi A, Filatov DA,**
806 **Chibalina M, Bergero R, Charlesworth D, Abe T, Kawano S** (2016) A new
807 physical mapping approach refines the sex-determining gene positions on the
808 Silene latifolia Y-chromosome. Sci Rep **6:** 18917
809 **Kejnovsky E, Kubat Z, Hobza R, Lengerova M, Sato S, Tabata S, Fukui K,**
810 **Matsunaga S, Vyskot B** (2006) Accumulation of chloroplast DNA sequences
811 on the Y chromosome of Silene latifolia. Genetica **128:** 167-175
812 **Kim D, Langmead B, Salzberg SL** (2015) HISAT: a fast spliced aligner with low
813 memory requirements. Nature Methods **12:** 357-360
814 **Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK,**
815 **SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J,**

816 **Barnstable C, Hoh J** (2005) Complement factor H polymorphism in
817 age-related macular degeneration. Science **308:** 385-389
818 **Korf I** (2004) Gene finding in novel genomes. BMC Bioinformatics **5:** 59
819 **Kovaka S, Zimin AV, Pertea GM, Razaghi R, Pertea M** (2019) Transcriptome
820 assembly from long-read RNA-seq alignments with StringTie2. Genome
821 Biology **20:** 289
822 **Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ,**
823 **Marra MA** (2009) Circos: an information aesthetic for comparative genomics.
824 Genome Research **19:** 1639-1645
825 **Kudoh T, Takahashi M, Osabe T, Toyoda A, Hirakawa H, Suzuki Y, Ohmido N,**
826 **Onodera Y** (2017) Molecular insights into the non-recombining nature of the
827 spinach male-determining region. Molecular Genetics and Genomics **293:**
828 557-568
829 **Kurtz S, Phillippy A, Delcher AL, Smoot M** (2004) Versatile and open software for
830 comparing large genomes. Genome Biology **5:** R12
831 **Lahn BT, Page DC** (1999) Four evolutionary strata on the human X chromosome.
832 Science **286:** 964-967
833 **Lam-Tung N, Schmidt HA, Arndt VH, Quang MBJMB, Evolution** (2015)
834 IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating
835 Maximum-Likelihood Phylogenies. Molecular Biology Evolution **32:** 268-274
836 **Lemaitre C, Braga MDV, Gautier C, Sagot MF, Tannier E, Marais GAB** (2009)
837 Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in

838 Human Sex Chromosomes. Genome Biology and Evolution **1:** 56-66
839 Li DB, Zhang HY, Mou MH, Chen YL, Xiang SY, Chen LG, Yu DQ (2019)
840 Arabidopsis Class II TCP Transcription Factors Integrate with the FT-FD
841 Module to Control Flowering. Plant Physiology 181: 97-111
842 **Li H** (2013) Aligning sequence reads, clone sequences and assembly contigs with
843 BWA-MEM. arXiv**:** 1303-3997
844 **Li H** (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
845 **34:** 3094-3100
846 **Li SF, Guo YJ, Li JR, Zhang DX, Gao WJ** (2019) The landscape of transposable
847 elements and satellite DNAs in the genome of a dioecious plant spinach
848 (Spinacia oleracea L.). Mobile DNA **10:** 3
849 **Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W** (2013)
850 Estimation of genomic characteristics by analyzing k-mer frequency in de
851 novo genome projects. Quantitative Biology **35:** 62-67
852 **Liu D, Qian W, Zhang H, Fan G, Xu Z** (2015) Development and application of
853 molecular markers linked with sex gene X/Y in spinach. Horticultural Plant
854 Journal **42:** 1583-1590
855 **Liu ZY, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu QY, Pearl HM, Kim**
856 **MS, Charloton JW, Stiles JI, Zee FT, Paterson AH, Ming R** (2004) A
857 primitive Y chromosome in papaya marks incipient sex chromosome evolution.
858 Nature **427:** 348-352
859 **Lizuka M, Janick J** (1962) Cytogenetic analysis of sex determination in Spinacia

oleracea. Genetics **47:** 1225-1241

Ma XK, Yu LA, Fatima M, Wadlington WH, Hulse-Kemp AM, Zhang XT, Zhang SC, Xu XD, Wang JJ, Huang HX, Lin J, Deng B, Liao ZY, Yang ZH, Ma YH, Tang HB, Van Deynze A, Ming R (2022) The spinach YY genome reveals sex chromosome evolution, domestication, and introgression history of the species. Genome Biology **23:** 23-75

Manna F, Gallet R, Martin G, Lenormand TJJoEB (2012) The high-throughput yeast deletion fitness data and the theories of dominance. Journal of Evolutionary Biology **25:** 892-903

Müller N, A., Kersten B, Leite Montalvao AP, Mähler N, Bernhardsson C, Bräutigam K, Carracedo Lorenzo Z, Hoenicka H, Kumar V, Mader M, Pakull B, Robinson KM, Sabatti M, Vettori C, Ingvarsson PK, Cronk Q, Street NR, Fladung M (2020) A single gene underlies the dynamic evolution of poplar sex determination. Nature Plants **6:** 630-637

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant ONA. Nucleic Acids Research **8:** 4321-4326

Okazaki Y, Takahata S, Hirakawa H, Suzuki Y, Onodera Y (2019) Molecular evidence for recent divergence of X- and Y-linked gene pairs in Spinacia oleracea L. Plos One **14:** e0214949

Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research **46:** e126

Ou S, Jiang N (2018) LTR_retriever: a highly accurate and sensitive program for

identification of long terminal repeat retrotransposons. Plant Physiology **176:** 1410-1422

Pan QW, Kay T, Depince A, Adolfi M, Schartl M, Guiguen Y, Herpin A (2021) Evolution of master sex determiners: TGF-beta signalling pathways at regulatory crossroads. Philosophical Transactions of the Royal Society B-Biological Sciences **376**

Peichel CL, McCann SR, Ross JA, Naftaly AFS, Urton JR, Ceah JN, Grimwood J, Schmutz J, Myers RM, Kingsley DM, White MA (2020) Assembly of the threespine stickleback Y chromosome reveals convergent signatures of sex chromosome evolution. Genome Biology **21:** 177

Pertea G, Pertea M (2020) GFF Utilities: GffRead and GffCompare. F1000 Research **9:** 304

Picq S, Santoni S, Lacombe T, Latreille M, Weber A, Ardisson M, Ivorra S, Maghradze D, Arroyo-Garcia R, Chatelet P, This P, Terral JF, Bacilieri R (2014) A small XY chromosomal region explains sex determination in wild dioecious V. vinifera and the reversal to hermaphroditism in domesticated grapevines. BMC Plant Biology **14**

Pilkington SM, Tahir J, Hilario E, Gardiner SE, Chagne D, Catanach A, McCallum J, Jesson L, Fraser LG, McNeilage MA, Deng C, Crowhurst RN, Datson PM, Zhang Q (2019) Genetic and cytological analyses reveal the recombination landscape of a partially differentiated plant sex chromosome in kiwifruit. Bmc Plant Biology **19:** 172

Qian W, Fan G, Liu D, Zhang H, Wang X, Wu J, Xu Z (2017) Construction of a high-density genetic map and the X/Y sex-determining gene mapping in spinach based on large-scale markers developed by specific-locus amplified fragment sequencing (SLAF-seq). BMC Genomics **18:** 276

Rifkin JL, Beaudry FE, Humphries Z, Choudhury BI, Barrett SC, Wright SI (2021) Widespread recombination suppression facilitates plant sex chromosome evolution. Molecular Biology and Evolution **38:** 1018-1030

She HB, Xu ZS, Zhang HL, Li GL, Wu J, Wang XW, Li Y, Liu ZY, Qian W (2021) Identification of a male-specific region (MSR) in Spinacia oleracea. Horticultural Plant Journal **7:** 341-346

Shumate A, Salzberg SL (2020) Liftoff: accurate mapping of gene annotations. Bioinformatics **37:** 1639-1643

Stanke M, Tzvetkova A, Morgenstern B (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology **7:** S11-S11

Su X, Wang B, Geng X, Du Y, Lin T (2021) A High-continuity and annotated tomato reference genome. BMC Genomics **22:** 898

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science **320:** 486-488

Vaser R, Sović I, Nagarajan N, Šikić M (2002) Racon-Rapid consensus module for raw de novo genome assembly of long uncorrected reads. Bioinformatics **18:** 452-464

926 **Wadlington WH, Ming R** (2018) Development of an X-specific marker and
927 identification of YY individuals in spinach. Theoretical and Applied Genetics
928 **131:** 1987-1994
929 **Walsh JB** (1995) How often do duplicated genes evolve new functions? Genetics **139:**
930 421-428
931 **Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, Aryal R, VanBuren R,**
932 **Murray JE, Zhang W, Navajas-Perez R, Feltus FA, Lemke C, Tong EJ,**
933 **Chen C, Wai CM, Singh R, Wang ML, Min XJ, Alam M, Charlesworth D,**
934 **Moore PH, Jiang J, Paterson AH, Ming R** (2012) Sequencing papaya X and
935 Yh chromosomes reveals molecular basis of incipient sex chromosome
936 evolution. Proceedings of the National Academy of Sciences of the U S A **109:**
937 13710-13715
938 **Waterhouse RM, Mathieu S, A SF, Mosè M, Panagiotis I, Guennadi K,**
939 **Kriventseva EV, Zdobnov EM** (2017) BUSCO applications from quality
940 assessments to gene prediction and phylogenomics. Molecular Biology and
941 Evolution **3:** 3
942 **Westergaard M** (1958) The mechanism of sex determination in dioecious plants.
943 Advances in Genetics **9:** 217–281
944 **Xie T, Zheng JF, Liu S, Peng C, Zhou YM, Yang QY, Zhang HY** (2015) De novo
945 plant genome assembly based on chromatin interactions: a case study of
946 Arabidopsis thaliana. Molecular Plant **8:** 489-492
947 **Xu C, Jiao C, Sun H, Cai X, Wang X, Ge C, Zheng Y, Liu W, Sun X, Xu Y, Deng**

J, Zhang Z, Huang S, Dai S, Mou B, Wang Q, Fei Z, Wang Q (2017) Draft genome of spinach and transcriptome diversity of 120 Spinacia accessions. Nature Communications **8:** 15275

Xue LJ, Wu HT, Chen YN, Li XP, Hou J, Lu J, Wei SY, Dai XG, Olson MS, Liu JQ, Wang MX, Charlesworth D, Yin TM (2020) Evidences for a role of two Y-specific genes in sex determination in Populus deltoides. Nature Communications **11:** 5893

Xue LZ, Gao Y, Wu MY, Tian T, Fan HP, Huang YJ, Huang Z, Li DP, Xu LH (2021) Telomere-to-telomere assembly of a fish Y chromosome reveals the origin of a young sex chromosome pair. Genome Biology **22:** 203

Yang L, Smyth GK, Wei S (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923-930

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution **24:** 1586-1591

Zhang N, Zeng L, Shan H, Ma H (2012) Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytologist **195:** 923-937

Zhang X, Zhang S, Zhao Q, Ming R, Tang H (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nature Plants **5:** 833-845

Zhao X, Hao W (2007) LTR_FINDER: an efficient tool for the prediction of

970 full-length LTR retrotransposons. Nucleic Acids Research **35:** W265-W268

971 **Zhong S, Joung JG, Zheng Y, Chen YR, Liu B, Shao Y, Xiang JZ, Fei Z,**

972 **Giovannoni JJ** (2011) High-throughput illumina strand-specific RNA

973 sequencing library preparation. Cold Spring Harbor Protocols **2011:** 940

974