Title

Genomic changes and stabilisation following homoploid hybrid speciation of the Oxford ragwort *Senecio squalidus*

Authors

Bruno Nevado[1,2,3,10,*], Mark A. Chapman[4], Adrian C. Brennan[5], James W. Clark[1,6], Edgar L.Y. Wong[1], Tom Batstone[6], Shane A. McCarthy[7], Alan Tracey[7], James Torrance[7], Ying Sims[7], Richard J. Abbott[8,†], Dmitry Filatov[1], Simon J. Hiscock[1,9]

[1] Department of Biology, University of Oxford, OX1 3RB Oxford, UK. [2] cE3c, Centre for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal. [3] Department of Animal Biology, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal. [4] School of Biological Sciences, University of Southampton, Southampton, SO17 1BJ, UK. [5] Biosciences Department, University of Durham, DH1 3LE Durham, UK. [6] Milner Centre for Evolution, Department of Life Sciences, University of Bath, BA2 7AY Bath, UK. [7] Wellcome Sanger Institute, CB10 1SA Cambridge, UK. [8] School of Biology, University of St. Andrews, KY16 9ST St. Andrews, UK. [9] University of Oxford Botanic Garden and Arboretum, Rose Lane, Oxford, OX1 4AZ Oxford, UK. [†] deceased. [10] Lead contact.

*** Correspondence:** Bruno Nevado, bnevado@fc.ul.pt

Summary

Oxford ragwort (*Senecio squalidus*) is one of only two homoploid hybrid species known to have originated very recently, so is a unique model for determining genomic changes and stabilisation following homoploid hybrid speciation. Here we provide a chromosome-level genome assembly of *S. squalidus* with 95% of the assembly contained in the 10 longest scaffolds, corresponding to its haploid chromosome number. We annotated 30,249 protein-coding genes and estimated that ca. 62% of the genome consists of repetitive elements. We then characterised genome-wide patterns of linkage disequilibrium, polymorphism and divergence in *S. squalidus* and its two parental species, finding that (i) linkage disequilibrium is highly heterogeneous, with a region on chromosome 4 showing increased values across all three species but especially in *S. squalidus*; (ii) regions harbouring genetic incompatibilities between the two parental species tend to be large, show reduced recombination, and have lower polymorphism in *S. squalidus*; (iii) the two parental species have an unequal contribution (70:30) to the genome of *S. squalidus*, with long blocks of parent-specific ancestry supporting a very rapid stabilisation of the hybrid lineage after hybrid formation; and (iv) genomic regions with major parent ancestry exhibit an overrepresentation of loci with evidence for divergent selection occurring between the two parental species on Mount Etna. Our results show that both genetic incompatibilities and natural selection play a role in determining genome wide reorganisation following hybrid speciation, and that patterns associated with homoploid hybrid speciation – typically seen in much older systems – can evolve very quickly following hybridisation.

Introduction

Hybridisation can be a creative force in organic evolution [1-4], enabling the transfer of genes between species (introgression) and the origin of new hybrid species involving either no change in chromosome number (homoploid hybrid speciation) or whole genome duplication (allopolyploidy). Homoploid hybrid speciation is considered rare [5], although this might be partly due to difficulties in recognising it [6] and the stringent criteria required to demonstrate its occurrence, particularly proving that reproductive barriers between the hybrid and its parents arose via hybridisation [7], though see [8] for other perspectives. However, these difficulties are beginning to be overcome by genomic and genetic analyses, and there is now good evidence for homoploid hybrid speciation having occurred in plants, including: sunflowers [9], *Ostryopsis* [10] and *Senecio* [11,12]; and animals, including: butterflies [13,14], finches [15], bears [16], and monkeys [17]. In most cases, the origins of known homoploid hybrid species are relatively ancient and, consequently, it is difficult to distinguish changes that occurred in the hybrid during its origin from those happening at a later stage. Only two homoploid hybrid species are known to be of very recent origin, a finch species that originated in the Galapágos Islands between 1981 and 2012 [15] and the Oxford ragwort (*Senecio squalidus*), a plant species that originated in the UK at the end of the 17th century [12,18]. These two species are particularly valuable for determining genomic and genetic changes during the initial stages of homoploid hybrid speciation. Here, we focus on such changes in the Oxford ragwort.

Senecio squalidus L. (Asteraceae) holds a unique place in the natural history of the UK and Ireland. This short-lived perennial herb, now a common sight along railways, road verges and wasteland in urban areas across the UK, originated from hybridisation between *S. aethnensis* Jan ex DC. and *S. chrysanthemifolius* Poir. [12,18-21] (Figure 1). The two parental species occur naturally at high (*S. aethnensis,* >2000m) and low (*S. chrysanthemifolius,* <1000m) elevations on Mount Etna, Sicily, and form a hybrid zone at intermediate elevations

[18,22-24]. During the late 17th century, both of these species were introduced to Britain, and hybridisation between them gave rise to a new hybrid lineage in the garden of the Duchess of Beaufort at Badminton, Gloucestershire and at the Oxford Botanic Garden [12,20]. The new hybrid lineage was subsequently cultivated extensively at the Oxford Botanic Garden, from where it escaped and naturalised in Oxford during the late 18th and early 19th centuries. During the industrial revolution of the 19th century *S. squalidus* spread from Oxford via the clinker beds of the expanding railway network and went on to colonize much of the British Isles over a period of ca. 150 years [20,25-28]. It's range now extends as far north as central Scotland, and west into Cornwall, Wales and Northern and Southern Ireland. More recently, it may have been introduced elsewhere in Europe and North America [29,30]. The spread of *S. squalidus* across the UK has triggered a burst of evolution in UK *Senecio* following its hybridisation with the native tetraploid species *S. vulgaris* L. ($2n = 40$). This has resulted in the origin of a new allohexaploid species, *Senecio cambrensis* Rosser ($2n = 60$), and two tetraploid introgressant taxa, *S. eboracensis* Abbott & Lowe and *S. vulgaris* var. *hibernicus* Syme [24,28,31], which have themselves become models for studying introgression and allopolyploidy in plants [32,33]. In addition, there is evidence of another tetraploid species, *S. viscosus* L., having been introgressed by *S. squalidus* [34].

The speed of colonization of the UK is intriguing in the context of the population history of *S. squalidus* and the fact that, like its parental species, it is strongly self-incompatible [28,35-38]. According to Baker's Rule, self-incompatible species tend to be poor colonizers compared to self-compatible species [39-41], especially if their founding population contains very few *S*-haplotypes that limit mate availability, as has been shown for *S. squalidus* [37,38,42-45]. This has made *S. squalidus* an especially interesting study system in terms of the inheritance and evolution of its sporophytic self-incompatibility (SSI) system and its origin and spread in the UK [24,28,46].

Previous studies estimated that the two parental species of *S. squalidus* diverged on Mt. Etna in the last 150,000 years [47,48], and remain distinct despite ongoing gene flow [48,49,50]. This is likely due to strong ecological selection as identified by clinal patterns of variation [49] and analysis of genomic differentiation across the Mt. Etna hybrid zone [51]. Crosses between the two species have also identified numerous loci showing transmission ratio distortion in their progeny [11,52,53] and in some instances hybrid breakdown [53], suggesting rapid establishment of intrinsic reproductive isolation mechanisms (incompatibilities) that may be (partly) responsible for the maintenance of the two species in the face of ongoing gene flow. Segregation of these incompatibilities in *S. squalidus* has, in turn, likely contributed to this species' reproductive isolation from its parents [11]. In addition, crosses between the two parental species have shown significant changes in gene expression in the hybrids [54-56], including transgressive expression patterns, which may explain how *S. squalidus* managed to colonise Britain, an environment where its parental species were never reported outside cultivation [57], and where both parental species perform poorly [58].

To gain a better understanding of the processes underpinning homoploid hybrid speciation and how they affect rapid adaptation to a novel environment, we generated and analysed a chromosome-level genome assembly of *S. squalidus*. The availability of this contiguous genome assembly together with a re-analysis of RNAseq data from this species (28 specimens covering most of the species' distribution range in Great Britain; Figure 1, Table S1) and its two parental species (16 specimens each, Table S1) [12,47] has allowed us to shed light on how pre-existing hybrid incompatibilities between the parental species, and selection acting on different parental alleles, together contributed to shaping the genome of *S. squalidus* and to fuelling its rapid spread across the UK following homoploid hybrid speciation.

Results

A chromosome-level genome assembly of S. squalidus

The chromosome-level assembly of *S. squalidus* consisted of 592 scaffolds, with an N50 of 66.7 Mb and a total length of 662.2 Mb (Table 1). We estimated the haploid genome size of the same individual as ca. 775 Mb using flow cytometry. This estimate is slightly lower than the ca. 880 Mb estimate for this species obtained previously [62] and implies that ca. 85% of the genome of *S. squalidus* is represented in our new assembly. The ten longest scaffolds accounted for over 95% of the assembly (631.8 Mb) and corresponded to the haploid chromosome number in *S. squalidus* [63]. Detailed statistics of the newly assembled genome are available in the BlobToolKit browser [64], and the Hi-C contact map on the genome-note server [65].

We annotated 30,249 protein-coding genes in the *S. squalidus* genome (Table S2) and 97.2% of the single copy plant orthologs (BUSCOs) were present and complete in the annotation (Table 1). Approximately 62% of the genome consists of repetitive elements (REs), with LTR elements being the most frequent (Figure 2A). Repetitive elements were not homogeneously distributed along chromosomes, with fewer repetitive regions found in terminal regions of each chromosome (Figure 2C). We identified SSRs in *S. squalidus* and four other members of the Asteraceae family, and found ca. 71,000 SSRs in the *S. squalidus* nuclear genome, which was less than in lettuce (ca. 265,000), sunflower (ca. 252,000) and globe artichoke (ca. 134,000) and more than in *Erigeron* (ca. 59,000). The distribution of repeat types and density in these genomes was similar, with dinucleotide repeats predominating (Figure 2B).

Chloroplast genome assembly

The chloroplast genome of *S. squalidus* (Figure S1) was 150,803 bp in length, with a large single copy region (LSC, 82,949 bp), a small single copy region (SSC, 18,213 bp) and a pair

of inverted repeats (IR, 24,821bp). The cpDNA genome is therefore slightly smaller than that from lettuce, sunflower and globe artichoke (151,104 to 152,765 bp).

Annotation of the chloroplast genome identified 116 genes including 80 protein coding, 5 rRNA genes and 31 tRNA genes as well as 126 SSRs (all of which were mononucleotide repeat SSRs) (Table S2). Comparison of the chloroplast genomes of *S. squalidus* and its two parental species (using a single individual of each species) identified 3 indels and 3 SNPs across all three species: 2 indels and 1 SNP supported a closer relationship between *S. squalidus* and *S. aethnensis*, while the remaining polymorphisms supported a closer relationship between the two parental species (Figure S1). However, follow-up work analysing more individuals of all three species, and ideally including *S. aethnensis* from higher elevations, is required to confirm this because the *S. aethnensis* individual used for cpDNA assembly was collected at 2,036m elevation, where admixed individuals may still be found.

Large-scale synteny across Asteraceae

To place the observed synteny changes on an evolutionary timescale, we estimated phylogenetic relationships and divergence times across representative Asteraceae genomes using 440 low-copy orthologs (Figure 3A). We estimate that the crown group of Asteraceae originated in the Palaeocene (65.8 – 55.3 Mya; Ypresian-Danian) and that the divergence between the lineages leading to *Lactuca* (Cichorieae) and the Asteroideae occurred during the Eocene (48.3 – 40.5 Mya, Bartonian-Lutetian). The divergence between the lineages leading to *Helianthus* and *Senecio* occurred later during the Eocene (42.5-35.6 Mya, Priabonian-Bartonian). The divergence among the *vulgaris*-clade of *Senecio* [66] was characterised by a rapid radiation during the Miocene (4.9-3.9 Ma). We note that the crown age for Asteraceae estimated here is somewhat younger than previous studies [67]. Given that the two studies use an almost identical set of node calibrations, the difference most likely results from the different

taxon sampling: a stronger emphasis on the origin of the Senecioneae in this study, vs a focus on the backbone of the Asteraceae phylogeny in previous work.

We investigated synteny between *S. squalidus*, the common sunflower and lettuce, and found many syntenic blocks between *S. squalidus* and each of the other species that spanned over 10 Mb (Figure 3C, D). The whole genome duplication specific to the sunflower clade [68] is evident from patterns of synteny across several pairs of chromosomes (Figure 3D). Of note, chromosome 4 of *S. squalidus* showed synteny over its entire length with chromosome 2 of *L. sativa* and with chromosome 14 of *H. annus*, as well as with several other Asteraceae species analysed (Figure 3B). This large scale synteny pattern is remarkable given that it involves species that diverged up to 48 Mya, and that *Helianthus* has since experienced an independent genome duplication and expansion [69,70]. The mechanisms promoting maintenance of synteny over large divergence times remain unclear, but the new genome assembly provided in this study can be used to leverage additional information and gain insight into how collinearity of such a large region is maintained across Asteraceae species.

Linkage disequilibrium is highly heterogeneous along the genome of S. squalidus

To estimate recombination rates along the genome of *S. squalidus* we mapped markers from the most extensive genetic linkage map available [53], which was obtained from crosses between the two parental species, to the new assembly. We found that 87% of markers from this genetic map were in the same order along the new assembly and we used these to estimate chromosome-wide average recombination rates; these ranged from 1.6 to 4.4 cM/Mb (Table S3). Regions with low local recombination rates putatively indicate the centromeres of chromosomes, often found towards the centre of chromosomes (Figure S2). To estimate effective recombination rate along the genome, we used pairwise LD estimates between pairs of SNPs within each chromosome with population-level RNAseq data collected for each

192 species. The results show overall higher LD in *S. squalidus* compared to both parental species,
193 which is in line with its recent origin.

194 Analysis of genome-wide LD identified a region on chromosome 4 (the first ca. 15 Mb)
195 with reduced recombination in *S. squalidus* compared to the rest of chromosome 4 (Figure
196 S3A). This region exhibits significantly lower recombination within all three species, with the
197 effect being larger in *S. squalidus* (Figure S3A,B). Furthermore, F_{ST} between all pairs of
198 species are significantly higher in this region (Figure S3C), and *S. squalidus* exhibits relatively
199 high polymorphism, high Tajima's D, and an excess of *S. aethnensis* diagnostic alleles (Figure
200 4).

201 At least three scenarios could explain the peculiar pattern of recombination,
202 polymorphism, and divergence on this region of chromosome 4. First, this region could harbour
203 chromosomal rearrangements between the two parental species, thus being involved in
204 homoploid hybrid speciation according to the recombinational model [71,72]. This hypothesis is
205 supported by high F_{ST} between the parental species in this region (F_{ST} = 0.44±0.15 vs genome-
206 wide F_{ST} = 0.37±0.16; Figure S3C), which could indicate reduced introgression caused by
207 rearrangements between the two species. However, this hypothesis would not explain the
208 increased LD found on this region in each of the parental species, given that a fixed
209 rearrangement is not predicted to cause reduced recombination within species. Furthermore,
210 we did not find any genetic incompatibility between the parental species on this region (Figure
211 4, top). Finally, the high polymorphism and Tajima's D observed in this region in *S. squalidus*
212 suggests that this region still harbours alleles from both species, which contrasts with the
213 recombinational speciation model that predicts regions with rearrangements would quickly fix
214 for alternative parental alleles.

215 A second possible explanation for the patterns observed would involve introgression of
216 a haplotype from a fourth species into *S. squalidus* during its colonisation of the UK. This

hypothesis could explain the high LD found in *S. squalidus* and the increased F_{ST} between *S. squalidus* and its parental species on this region (Figure S3C) as well as the high polymorphism and Tajima's D in this region in *S. squalidus* (Figure 4). However, this hypothesis would not explain the high F_{ST} and LD observed in both parental species in this region (Figure S3).

A third possible explanation for the patterns observed is that this region of chromosome 4 harbours the self-incompatibility (*S*)locus controlling SSI in *Senecio*. We favour this hypothesis as it could better explain all the patterns observed. First, *S*-loci are typically under strong balancing selection: as *S*-haplotypes become rarer, their fitness increases because individuals that carry them can mate with a larger pool of mates; conversely, any *S*-haplotype reaching high frequency will see its fitness decrease as fewer mates will be available for breeding. The elevated Tajima's D in this region could thus be explained by the action of frequency-dependent balancing selection on the *S*-locus. Second, because *S. aethnensis* was the minor contributor to the gene pool of *S. squalidus*, *S*-haplotypes of this species might be expected to be rarer in the hybrid lineage giving rise to *S. squalidus*. Selection favouring these rare *S*-haplotypes could thus explain the higher frequency of *S. aethnensis* diagnostic alleles in this region of chromosome 4. Third, *S*-loci are typically located in regions of reduced recombination [73], which ensures that both male and female SI-determining genes are inherited together. Given the recent hybrid origin of *S. squalidus*, the region of reduced recombination typical of the *S*-locus is expected to extend further than in either parental species, and this is what we observed (Figure S3).

To further narrow down the potential location of the *S*-locus within this region of chromosome 4, we estimated average LD between pairs of SNPs within 1 Mb windows across chromosome 4 in the three *Senecio* species. We reasoned that because the three species share the same *S*-locus, the window harbouring the *S*-locus should exhibit high LD in all species. We found only one window within this region that exhibited high LD in all species (Figure S4).

Analysis of tissue-specific gene expression in all three *Senecio* species identified putative *S* genes, however none of these candidate genes showed a clear functional similarity to any *S*-genes identified in other species (Table S4, Figure S4). Overall, while we lack strong evidence for specific genes involved in self-incompatibility in *S. squalidus*, the region identified in chromosome 4 represents an interesting target for future studies. These could leverage additional data (including tissue-specific gene expression of multiple *S. squalidus* individuals of known S-genotype) to test whether this region on chromosome 4 harbours the *S*-locus of *Senecio*.

The genome of S. squalidus is a mosaic of parental alleles

Genome-wide analysis of polymorphism and divergence within *S. squalidus* and between *S. squalidus* and its parental species revealed a general loss of polymorphism and an increase in Tajima's D (indicative of a reduction in low frequency variants) in *S. squalidus*. Genome-wide diversity measured as average Watterson's θ across non-overlapping sliding windows of 500 kb, was 0.0030±0.0013, 0.0035±0.0014 and 0.0041±0.0017; and genome-wide average Tajima's D was 1.06±0.92, 0.08±0.73 and 0.13±0.60; in *S. squalidus*, *S. chrysanthemifolius* and *S. aethnensis*, respectively. Watterson's θ was significantly lower and Tajima's D significantly higher in *S. squalidus* compared to either of the two parental species (Welch two sample t-test, $P < 0.0001$).

Genome-wide average differentiation measured as mean F_{ST} between *S. squalidus* and the two parental species confirmed *S. chrysanthemifolius* as the more genetically similar parental species (Welch two sample t-test, $P < 0.0001$), even though genome-wide estimates of F_{ST} were similar across all pairs (*S. squalidus* vs *S. chrysanthemifolius* F_{ST} =0.36±0.14; *S. squalidus* vs *S. aethnensis* F_{ST} =0.40±0.14; *S. aethnensis* vs *S. chrysanthemifolius* F_{ST} = 0.37±0.16). In line with our previous study [12] and in agreement with the recent hybrid origin

of *S. squalidus*, we find that the majority of diagnostic SNPs, i.e. those where the two parental species are nearly fixed for different alleles, are found in a polymorphic state in *S. squalidus* (64% of all diagnostic SNPs). However, diagnostic SNPs that were nearly fixed in *S. squalidus* more often carried the *S. chrysanthemifolius* allele (1377 SNPs) rather than the *S. aethnensis* allele (539 SNPs) (binomial test, $P < 0.0001$). Diagnostic SNPs fixed in *S. squalidus* are not randomly distributed along the genome (Pearson's Chi-squared test, $P < 0.0001$; Figure 4). Instead, *S. aethnensis*-like alleles are preferentially found on chromosomes 4 and 5 (and to a lesser extent on chromosome 2), while elsewhere in the genome *S. chrysanthemifolius*-like alleles are more common (Figure S5).

Despite lower genome-wide average polymorphism, many regions of the genome show levels of polymorphism that are higher in *S. squalidus* than in either parental species (Figure 4, top). Given the extremely young age of *S. squalidus*, the higher polymorphism in these regions most likely reflects the retention of alleles from both parental species, rather than *de novo* mutations accumulating since the origin of *S. squalidus*. Also evident from the genome-wide analysis is that the large regions harbouring incompatibilities in chromosomes 2, 3 and 4 (discussed below) are almost completely devoid of high-polymorphism windows (average Watterson's theta within these regions = 0.0028 vs outside = 0.0031; Welch two-sample t-test, $P = 0.027$), which is in line with sorting of parental haplotypes in *S. squalidus* in these regions.

Genetic Incompatibilities are located in regions of reduced recombination and polymorphism

To gain insight into the role of pre-existing genetic incompatibilities (between the two parental species) in the evolution of *S. squalidus*, we analysed genome-wide patterns of LD and the distribution of genetic incompatibilities identified in previous studies[11,52,53]. Here we define incompatibilities broadly as regions showing evidence for significant transmission ratio

291 distortion (TRD) in experimental crosses, as these imply a fitness cost on hybridisation between
292 the two species.

293 The three regions harbouring incompatibilities described using RNA-based genetic
294 mapping [53] spanned ca. 5.9 Mb of chromosome 2, 33 Mb of chromosome 3 and 23 Mb of
295 chromosome 4 (Figure 4). Watterson's θ was significantly lower (Welch two sample t-test, P
296 = 0.027) and Tajima's D significantly higher (Welch two sample t-test, P = 0.001) in genomic
297 windows harbouring incompatibilities compared to the rest of the genome (Figure S6A, B).
298 We also found that markers overlapping the regions of genetic incompatibilities were
299 significantly more likely to be located in regions of low recombination compared to the rest of
300 the genome (mean local recombination rate in incompatibility markers = 1.24 cM/Mb; non-
301 incompatibilities = 2.75 cM/Mb; t-test P = 0.002; Figure S6C). This pattern does not seem to
302 be driven solely by putative location of incompatibilities near centromeres, as incompatibilities
303 span very large areas (Figure 4). Mapping of incompatibilities identified using a different cross
304 [52] revealed that incompatibilities are present on chromosomes 4, 5, 7 and 8 but their sparse
305 occurrence precluded more detailed analysis (Figure 4).

306 Genomic windows harbouring genetic incompatibilities between the parental species
307 thus exhibit lower polymorphism and higher Tajima's D in *S. squalidus* compared to the rest
308 of the genome (Figure S6) and are less likely to exhibit higher polymorphism than found in
309 either parental species (Figure 4, top). This suggests that, in regions harbouring genetic
310 incompatibilities, alleles from one of the parental species have been fixed in *S. squalidus*, while
311 in other regions of the genome alleles from both species might still be segregating. Importantly,
312 across the genome different genetic incompatibility regions have been fixed for different
313 parental haplotypes [11], which agrees with the hypothesis that sorting of incompatibilities is
314 important in generating reproductive barriers between homoploid hybrids and their parental
315 species [74] and could play a role in speciation [71,72].

A role for natural selection on the sorting of parental alleles following hybrid speciation

As expected for a homoploid hybrid species, and demonstrated in hybrid sunflowers [75], the distribution of parental alleles across the *S. squalidus* genome is not random. Instead, parent-specific alleles occur in blocks, i.e. tracts of the genome where all diagnostic SNPs are inherited from the same parent. After hybridisation, the size of these blocks is reduced due to recombination until fixation of haplotypes from either parental species occurs, at which point recombination can no longer reduce block size [76]. The distribution of block sizes can be used to infer the time taken between generation of a hybrid population and establishment of a hybrid species [76,77]. In *S. squalidus* we find that two chromosomes (4 and 5) carry long blocks of *S. aethnensis*-specific alleles, with *S. chrysanthemifolius*-specific blocks more common elsewhere (Figure 4). These long ancestry blocks in *S. squalidus* point to a very rapid establishment of the newly formed hybrid lineage and is in line with the historical records and demographic reconstructions that imply a very strong genetic bottleneck in the origin of *S. squalidus* and the establishment of a new, stabilised hybrid lineage within 30-100 generations [12].

We recover two additional genomic patterns in *S. squalidus* that are in-line with other, much older cases of homoploid hybrid speciation [10,14,17,78,79]. First, the *S. squalidus* genome exhibits an unequal contribution of genetic material from the two parental species: of 1916 diagnostic SNPs that were fixed or nearly fixed in *S. squalidus*, 71.9% carry the *S. chrysanthemifolius* allele and 28.1% the *S. aethnensis* one. Second, regions with minor parent ancestry exhibit higher recombination rates (Figure S7A), although this trend is non-significant (Welch two sample t-test, $P = 0.087$). Whether these genome-wide patterns are due to neutral or selective processes is central to our understanding of homoploid hybrid speciation, and its role in generating novel phenotypes.

The unequal contribution of the two parental species could be due to preferential backcross of *S. squalidus* with *S. chrysanthemifolius*. Indeed, historical records suggest that *S. aethnensis* x *S. chrysanthemifolius* hybrid material was grown alongside *S. chrysanthemifolius* (but not *S. aethnensis*) in Oxford, allowing for backcrossing and introgression of *S. chrysanthemifolius* alleles into *S. squalidus*. However, the period during which this occurred was relatively short, as all *S. chrysanthemifolius*-like herbarium specimens from Oxford Botanic Garden pre-date 1720 (*pers. comm.* Stephen A. Harris). Furthermore, it is unclear whether these *S. chrysanthemifolius*-like specimens were 'pure' *S. chrysanthemifolius* plants or admixed individuals, and only in the former case would this scenario explain the preferential introgression of *S. chrysanthemifolius* alleles into *S. squalidus*.

An alternative explanation for preferential fixation of *S. chrysanthemifolius* alleles in *S. squalidus* is that such fixation was driven by natural selection. This could be due to purifying selection removing deleterious alleles from the minor parent, as shown for hybrid swordtail populations [79] and hybrid chestnut trees species [78]; or positive selection driving fixation of advantageous alleles from the major parent in the hybrid genomic background, as inferred in other systems [10,14,16,17]. Our results do not support a role for purifying selection removing deleterious alleles from the minor parent, because genetic diversity of *S. aethnensis* is greater than in *S. chrysanthemifolius* (Watterson's $\theta = 0.0041 \pm 0.0017$ vs 0.0035 ± 0.0014), which suggests that deleterious mutations would be more common in the major parent. As for positive natural selection driving fixation of major parent alleles, it is worth noting that *S. chrysanthemifolius* in Sicily grows in disturbed habitats (roadsides, derelict buildings and abandoned orchards and vineyards) akin to those favoured by *S. squalidus* in the UK, such that natural selection could favour alleles from this species in *S. squalidus*. Given the very recent origin of *S. squalidus* and the strong bottleneck associated with its origin in the UK, we are unable to apply tests for selection that rely on fixation of alleles in *S. squalidus* since its origin.

However, we can test whether alleles that experience divergent selection between the two parental species on Mount Etna are preferentially fixed for either parental species in *S. squalidus*.

Previous studies found evidence for divergent selection acting between the two parental species on Mount Etna using different datasets and approaches [49,51]. Here we make use of the results of the latest analysis, which identified 76 outlier loci using a nextRADseq dataset representing multiple populations (192 individuals in total) of the two species [51]. We identified the genomic location of 44 of these loci on the newly assembled genome and mapped them onto 39 genomic windows (Figure 4, bottom). Of these 39 genomic windows, *S. squalidus* carries exclusively *S. chrysanthemifolius* alleles at all fixed diagnostic SNPs in 17 windows; exclusively *S. aethnensis* alleles at all fixed diagnostic SNPs in 5 windows; is polymorphic at diagnostic SNPs in 6 windows; and at the remaining 11 windows no diagnostic SNPs were identified. A permutation test shows that the observed number of windows with *S. chrysanthemifolius* ancestry is significantly higher than expected by chance: out of 1000 permutations, only six resulted in 17 or more genomic windows carrying outlier loci and with only *S. chrysanthemifolius* diagnostic SNPs (Figure S7B).

Our analysis shows that genomic windows of major parent ancestry are more likely to be under divergent selection between the two parental species on Mount Etna than expected by chance. A limitation of this analysis is that selective regimes are likely different in the UK compared to Mount Etna, such that genes that experience divergent selection on Mount Etna might be evolving neutrally in the UK. The reverse is also true: genes that evolved neutrally between the parental species on Mount Etna might have been important in adaptation of *S. squalidus* to its new environment in the UK. Regardless of this limitation, our results lend support to a role for natural selection in driving preferential fixation of *S. chrysanthemifolius* alleles in *S. squalidus*.

Discussion

The generation of a chromosome-level genome assembly for *Senecio squalidus*, combined with re-analysis of transcriptome data for this species and its two parental species [12,47] have provided a greater understanding of the genomic and genetic changes that occurred in *S. squalidus* following its hybrid origin. We have determined how this species' hybrid genome is structured, how its genome compares with those of other Asteraceae species in terms of synteny, where genomic incompatibilities are located, and how natural selection may have determined the genomic contribution of the two parental species of *S. squalidus*. The availability of a high-quality genome for *S. squalidus* also sets the stage for future population genomic studies, particularly to pinpoint the combinations of alleles important in adapting the species to conditions in the UK, and to clarify the evolution of other hybrid taxa that have originated very recently in the UK following hybridisation with *S. squalidus* [24,28,31].

How then did hybridisation fuel the adaptation and rapid spread of *S. squalidus* across the UK? It is remarkable that neither of the parental species of *S. squalidus* is established in the UK outside cultivation [57], and that common garden experiments have shown that both parental species, their naturally occurring hybrids from Mount Etna, and newly synthetised hybrids between the two species perform poorly in the UK [58]. Our results suggest that natural selection preferentially drove *S. chrysanthemifolius* alleles to fixation in the new hybrid species, possibly as this species occupies ecologically similar habitats to those *S. squalidus* would find in the UK. This has two important implications. First, while selection may have preferentially favoured *S. chrysanthemifolius* alleles in *S. squalidus*, whatever *S. aethnensis* alleles remain in the hybrid species must be central to the success of the hybrid lineage in its new environment – without these minor alleles *S. squalidus* would become genetically identical to *S. chrysanthemifolius*, and would thus perform poorly in the UK. The identification of these minor

parent genes, how they underpin adaptation to the UK, and in particular how this effect might depend on epistatic interactions with *S. chrysanthemifolius* alleles in other genes, are promising avenues for future research. Second, sorting of parental alleles in *S. squalidus* preferentially occurred at genomic regions that harbour highly differentiated alleles between the two parental species, suggesting that natural selection acted on alleles that were already under divergent selection between the parental species. Thus natural selection may have acted in a novel combination of alleles that were themselves previously subjected to natural selection on a different environment, which could help explain how adaptation to a novel environment proceeded so quickly, and further supports the hypothesis that hybridisation plays a creative role in generating novel phenotypes that are able to colonise new niches.

Acknowledgements

This paper is dedicated to the memory of our co-author Richard Abbott who sadly died while the paper was in revision. Richard was a pioneer in the field of homoploid hybrid speciation, with many of his insights stemming from his research of *Senecio*. This work was supported by Natural Environment Research Council grants NE/G018448/1 and NE/P002145/1 to SJH and NE/G017646/1 to DAF. BN is supported by funds from Fundação para a Ciência e a Tecnologia (https://doi.org/10.54499/CEECIND/00229/2018/CP1553/CT0002 and 2022.15825.CPCA). Work at the Tree of Life programme at the Wellcome Sanger Institute was performed under the umbrella of the 25 genomes for 25 years project, and funded through the Institute's core award from the Wellcome Trust (award number 206194). We thank Mike Stratton and Julia Wilson for their support for the 25 genomes for 25 years project, and the Sanger Institute Scientific Operations Long Read team for expert support in DNA extraction and sequence generation. We thank Mark Blaxter, Varvara Fazalova, Ivo Chelo, Vítor Sousa, Alex Blanckaert and Stephen Harris for comments on earlier versions of this manuscript.

441

442 ***Author contributions***

443 Study design: BN, SJH, TB, DAF. Data collection: BN, EW, TB, JC, ACB, SAM, AT, JT, YS.
444 Genome assembly: BN, SAM, AT, JT, YS. Genome annotation: BN. Data analysis: BN, MAC,
445 JC. Writing: BN, MAC, JC, RJA, SJH, ACB. All authors read and approved the manuscript.

446

447 ***Declaration of interests***

448 The authors declare no competing interests.

449

450 ***Main Figures and Tables legends***

451

452 **Figure 1: *Senecio squalidus* and its two parental species.** Photos show typical plants of *S.*
453 *squalidus* in the UK (top) and its two parental species from Mount Etna: *S. aethnensis* is found
454 in volcanic soils at high elevation (above 2,000m) and *S. chrysanthemifolius* in disturbed
455 habitats at lower elevation (below 1,000m). Inset to the top right notes key events, inferred
456 from historical records, detailing the origin of *S. squalidus* and its expansion across the British
457 Isles. Map to the bottom left details the sampling of *S. squalidus* used in this study (size of red
458 circles denotes number of individuals sequenced from each location). See also Table S1.

459

460 **Figure 2: Analyses of repetitive regions in the newly assembled genome of *S. squalidus*.**
461 (A) Overall percentage of the genome consisting of different repeat elements. (B) Number of
462 Simple Sequence Repeats (SSRs) per kb identified in *S. squalidus* and other Asteraceae species
463 with published genomes. (C) The distribution of all repeat elements calculated as percentage
464 of sequence over 1 Mb windows across each chromosome. See also Figures S1-S2 and Table
465 S3.

Figure 3: Synteny and molecular dating analysis across representative Asteraceae species. (A) Dated phylogeny of the representative Asteraceae species analysed in this study (excluding outgroups outside Asteraceae used for calibration), showing the divergence times of *S. squalidus*, *H. annus* and *L. sativa*. (B) Synteny plots showing in detail the large scale synteny between chromosome 4 of *S. squalidus* and remaining Asteraceae species. (C,D) Synteny analysis between *S. squalidus* and *L. sativa* (C) and between *S. squalidus* and *H. annus* (D) genomes, shown as a dot plot and covering only the 10 chromosomes of *S. squalidus*. See also Tables S6–S7.

Figure 4: **Analysis of polymorphism using non-overlapping windows (500kb size) along the genome of *S. squalidus.*** Middle panel denotes polymorphism (Watterson's estimator) in *S. squalidus* along the 10 chromosomes (alternating black and grey line). Top bars denote location of incompatibilities described in the two studies mentioned in the main text (green from [53], orange from [52]), genomic windows with high Tajima's D in *S. squalidus* (Zscore > 1.65, or approx. 1-tailed $P < 0.05$) and genomic windows where polymorphism in *S. squalidus* is higher than in both parental species. Bottom panel denotes the proportion of fixed diagnostic SNPs (PFDS) from each parental species that are fixed or nearly fixed in *S. squalidus*. Genomic windows without diagnostic SNPs, or only with diagnostic SNPs that are polymorphic in *S. squalidus* are not shown. In the bottom panel, asterisks atop the bars denote genomic windows with evidence for divergent natural selection acting on the two parental species on Mount Etna (from [51]). See also Figures S3–S7 and Table S4.

Table 1: Statistics of the genome assembly of *S. squalidus* compared to other Asteraceae species. See also Table S2, Table S5.

Species	*Senecio squalidus*	*Helianthus annuus*	*Lactuca sativa*	*Erigeron canadensis*	*Cynara cardunculus*
Source	This study	GCA_002127325.2	GCF_002870075.2	GCF_010389155.1	GCF_001531365.1
Number of scaffolds	592	332	8325	357	13,588
Scaffold N50 (Mb)	66.7	176	1.77	45.5	0.125
Contig N50 (Kb)	157	2000	28	1600	19
Total length (Mb)	652	3010	2380	426	725
Number of genes	30,249	83,308	38,919	44,592	26,889
Complete BUSCOs (%)	97.2	98.1	99.6	99.5	98.1
Complete single-copy BUSCOs (%)	87.1	87.3	99.1	97.6	96.0

STAR Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Bruno Nevado (bnevado@fc.ul.pt).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The genome assembly of *S. squalidus* is available from NCBI under accession number GCA_910822075.1. The raw sequence data used in the genome assembly are available from NCBI SRA repository under various accession numbers (see Table S5). The population-level raw sequence data for each species is available from NCBI SRA repository under BioProject ID PRJNA549571.
- This paper does not report original code.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

For genome assembly we selected a single, healthy, *S. squalidus* individual (accession name: Ox6) collected from an Oxford population and previously confirmed as heterozygous for self-incompatibility (*S*) haplotypes *S1* and *S4* [38]. This genotype was maintained clonally via cuttings in glasshouses at the Universities of Bristol and Oxford. DNA extraction was carried out using fresh material (young leaves) at the Sanger institute using the BioNano PlantTissue DNA Isolation protocol (https://bionanogenomics.com/wp-content/uploads/2017/01/30068-Bionano-Prep-Plant-Tissue-DNA-Isolation-Protocol.pdf). Genome size of Ox6 was estimated using flow cytometry (https://www.plantcytometry.nl/).

METHOD DETAILS

Genome sequencing

DNA was prepared by shearing for sequencing on the Pacific Biosciences SEQUEL I platform. Two library fragments sizes were prepared (~6 kb and ~12 kb) and sequenced over 16 SMRT cells (14 for 6kb, 2 for 12 kb), generating 84 Gb of raw data, ~160-fold coverage, with an overall read N50 of 7 kb. We generated ~300-fold base coverage in 10X Chromium Genome long fragment read clouds. We commissioned ~100-fold coverage each in Dovetail Chicago and Hi-C data. Details of sequencing are given in Table S5).

Nuclear genome assembly

Assembly was carried out following the Vertebrate Genome Project pipeline v1.0 [80] with FALCON-UNZIP [81]. Haplotypic duplication was identified and removed with PURGE_DUPS [82]. A first round of scaffolding was carried out with 10X Genomics read clouds using SCAFF10X

(available from https://github.com/wtsi-hpag/Scaff10X). Scaffolding with Hi-C data [83] was performed using SALSA2 [84]. The Hi-C scaffolded assembly was polished with arrow using the PacBio data, then polished with 10X Illumina data by aligning to the assembly with LONGRANGER align (available from https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines), calling variants with FREEBAYES [85] and applying homozygous non-reference edits using BCFTOOLS consensus. Two rounds of the Illumina polishing were applied. The assembly was checked for contamination using BLOBTOOLKIT [86] and contaminating scaffolds were removed.

Chloroplast genome assembly

To assemble the chloroplast genome of *S. squalidus* we used the *de novo* assembler NOVOPLASTY v2.5 [87] on the adaptor-trimmed Illumina PE reads. We used the first 500bp of the *Jacobea vulgaris* complete cpDNA sequence as seed (GenBank accession number HQ234669) and set the following options: insert size automatic, genome range 120-200k bp, K-mer 39, insert range 1.6, insert range strict 1.2 and coverage cut off 1000. The chloroplast genome was annotated using GESEQ [88] and displayed using OGDRAW [89] available from https://chlorobox.mpimp-golm.mpg.de. The resulting complete cpDNA genomes were aligned by hand using SEAVIEW v 4.0 [90]. This was repeated for one individual of each of *S. aethnensis* and *S. chrysanthemifolius* – DNA extracts for these individuals were obtained from fresh leaves of individuals grown in the greenhouse, using Qiagen DNeasy Plant kit. Material was sequenced on an Illumina HiSeq 2000.

Annotation

Prior to annotation we identified repeat regions within the genome using both REPEATMASKER v4.0 [91] and a custom repeat library generated for our species using REPEATMODELER v1.0 [92].

To annotate the genome assembly of *S. squalidus* we used the MAKER pipeline v2.31 [93]. In the first annotation pass we used both *ab initio* and transcriptome-based gene prediction evidence obtained from the transcriptome reference of *S. squalidus* assembled in our previous work [12] and the proteomes of globe artichoke (C*ynara cardunculus* var. *scolymus*, GenBank accession number: GCA_001531365.1) and common sunflower (*Helianthus annuus*, GCA_002127325). The obtained gene models were then improved with SNAP v. 2006-07-28 [94] and a second and final annotation pass with MAKER was performed using these improved gene models. To evaluate the completeness of the genome assembly and the performance of the annotation pipeline we used the homology-based approach implemented in BUSCO v4.1 [95]. Simple sequence repeat markers (SSRs; aka microsatellites) were identified in the *S. squalidus* genome using MISA [96] and a minimum of 8 repeats for dinucleotide repeats, 6 repeats for trinucleotide repeats, and 4 repeats for tetra-, penta- and hexanucleotide repeats. The same settings were used to mine the genomes of lettuce (NCBI accession number GCF_002870075.2), sunflower (GCA_002127325.2), globe artichoke (GCF_001531365.1) and *Erigeron canadensis* (GCF_010389155.1) for comparison. Gene sequences annotated across the 10 chromosomes of *S. squalidus* were translated to proteins and blasted against the UniProt sequence database [97] using BLASTP [98]. Blast hits were loaded into BLAST2GO and INTERPROSCAN [99] was used to add InterPro terms to each annotated gene. These multiple sources of information were used to annotate each gene with its most likely gene ontology term.

Tissue-specific RNAseq of S. squalidus

To infer tissue-specific gene expression values we used RNAseq expression data. We sampled different tissues (roots, young leaves, fully developed leaves, capitulum buds, flower buds and whole open flowers) from a single *S. squalidus* individual grown in the greenhouse under a 16:8h light cycle, planted in a mixture of soil and perlite. Tissues were flash-frozen in liquid

582 nitrogen and RNA extracted using the method described in [56]. Expression values for each gene
583 and each tissue were estimated with Trinity v 2.12 [100] using the bowtie alignment option and
584 the RSEM abundance estimation method.
585
586 ***Synteny across the Asteraceae***
587 Macrosyteny was inferred using chromosome-scale assemblies across a diversity of Asteraceae
588 species. Gene models and genome annotation files were obtained from publicly available
589 databases and patterns of synteny were analysed in pairwise comparisons of all species using
590 the MCSCAN pipeline [101] and visualised as dotplots and synteny plots using the JCVI pipeline
591 [102].
592 To gain a better understanding on how synteny changes through time within Asteraceae,
593 we used phylogenetic and molecular clock methods to date split events between representative
594 species of this family. We obtained proteomes for other *Senecio* species, sunflower and lettuce,
595 as well as for outgroup species (Table S6). Orthologous genes among proteomes were
596 identified using ORTHOFINDER v2 [103], with a Diamond similarity search and default parameters.
597 Low copy number gene families from ORTHOFINDER were aligned using MAFFT with the
598 localpair option [104] and trimmed using TRIMAL with the automated1 option [105]. Phylogenetic
599 trees for individual gene families were reconstructed under the best fitting model (-MFP) in
600 IQTREE [106], with 10,000 ultrafast bootstrap replicates [107]. The bootstrap consensus trees were
601 provided as input to ASTRAL-III to reconstruct the species relationships [108]. Gene families were
602 clustered by rate, approximated using the root to tip distance, into five clusters. Each cluster
603 was concatenated and formed a single partition in a partitioned molecular clock analysis. Eight
604 relevant fossil calibrations were selected (Table S7) and modelled as a uniform distribution
605 between a hard minimum and a soft maximum age, with a 1% probability tail that the maximum
606 could be exceeded. The selection of the fossil *Tubulifloridites lillei* has proven contentious in

607 the past [67], and as such a 2.5% probability tail was attached to the minimum age here. Clock
608 analyses were run using the normal approximation method in [109], where branch lengths and the
609 Hessian matrix are first estimated prior to the clock analysis [110]. A relaxed clock was selected,
610 where the clock rate for each branch is independently drawn from a lognormal distribution.
611 The prior on the mean was modelled as a gamma distribution with a shape parameter of 2 and
612 scale parameter of 20. Four independent chains were run for 5 million generations, with
613 effective sample sizes measured using Tracer [111] to determine convergence.
614
615 ***Population genomics statistics***
616 To evaluate the effect of hybridisation on genome-wide patterns of polymorphism in *S.*
617 *squalidus*, we re-analysed RNAseq data from this species (n = 26) and its two parental species
618 (n = 16 each) from previous studies [12,47] (Table S1). Raw sequence reads were trimmed for low
619 quality bases and adaptors using TRIMMOMATIC v0.35 [112], and mapped to the genome with the
620 splice-aware aligner STAR v2.7 [113] using default settings and including the annotation
621 information generated for the new genome. Duplicate reads were marked with PICARDTOOLS
622 v2 (available from http://broadinstitute.github.io/picard/) using the markduplicates function
623 and SNP calling performed with SAMTOOLS v1.3 [114] bcftools command. We used the
624 multiallelic SNP caller, disregarded reads with mapping quality below 20 and bases with base
625 quality below 20 and included in the output homozygous-reference blocks with a minimum
626 depth of 8 reads (-g8). Inclusion of homozygous-reference blocks is essential to distinguish
627 regions of missing data (i.e., that have not been sequenced to high enough depth to perform
628 confident genotype calling) from regions that were adequately covered but where no SNPs are
629 present (i.e., truly invariant positions). We further filtered the resulting SNP set by excluding
630 SNPs covered by fewer than 8 reads (the same depth threshold as used for the homozygous
631 blocks); SNPs within 3bp of an indel (-g3); SNPs with quality below 15; and heterozygous

SNPs with fewer than two reads supporting each allele. We converted the resulting vcf file into fasta format using VCF2FAS, which reads vcf files with reference-homozygous blocks and correctly assigns missing data and homozygous-reference genotypes (available from https://github.com/brunonevado/vcf2fas).

We obtained genome-wide estimates of polymorphism and divergence, namely Watterson's θ, F_{ST} and Tajima's D, using MSTATSPOP v0.1 (available from https://github.com/CRAGENOMICA/mstatspop) and applying a non-overlapping sliding window approach with 500 kb size and 500 kb steps. We used this software as it implements the algorithms described in [115] to provide unbiased estimates even in the presence of high levels of missing data between individuals. Smaller window sizes were explored as well, but these resulted in too many windows with too much missing data (not shown). To determine the parental contributions to the genome of *S. squalidus* we identified SNPs where the two parental species were fixed or nearly fixed for alternative alleles (>90% in one parent and <10% in the second parent and at least 10 alleles of each species sequenced), and where *S. squalidus* carried almost exclusively one of these alleles (frequency of most common allele > 90% in *S. squalidus*). We further identified windows with significantly higher Tajima's D than the genomic background using a Z-test (Zscore > 1.65, roughly equivalent to a 1-tailed test at $P < 0.05$). Resulting data was plotted in the R statistical package (available from https://www.r-project.org).

Testing for the role of selection in sorting of parental alleles in S. squalidus

To test for a role for natural selection in the sorting of parental alleles in *S. squalidus*, we identified the genomic location of outlier loci identified in a previous study as being under divergent selection between the two parentals species across an elevation gradient on Mount Etna [51]. We blasted the sequences of the 76 nextRAD outlier loci identified in that study against

657 the newly assembled genome of *S. squalidus*, retaining hits with more than 95% sequence
658 identity and at least 140 bp long. Markers with hits on multiple scaffolds and identical match
659 statistics were disregarded. For markers with multiple hits on the same scaffold, the hit with no
660 gaps and higher sequence identity was kept.

661 To test whether regions containing loci under selection were more likely to have already
662 sorted for either parental alleles, we identified the ancestry of each genomic window based on
663 the presence of diagnostic SNPs: windows where all diagnostic SNPs that are fixed or nearly
664 fixed in *S. squalidus* carry the same parental allele were labelled as having sorted for that
665 parental species; windows where diagnostic SNPs are either polymorphic in *S. squalidus* or
666 nearly fixed for different parental alleles were labelled as polymorphic windows; and regions
667 without diagnostic SNPs were labelled as having unknown ancestry. We then calculated how
668 many windows that sorted for the two parental alleles carried also loci under divergent selection
669 on Mount Etna, based on the presence of at least one outlier nextRAD locus. To test whether
670 this number is higher than expected by chance, we performed a permutation test in R: we
671 randomly selected the same number of windows and assigned them to "outlier" status and
672 calculated how many of these windows have also sorted for each parental species. For this
673 permutation test we ignored windows with unknown ancestry and performed 1000 replicates
674 to assess whether the values observed are likely to have occurred by chance alone.

675

676 ***Genetic linkage maps and mapping genetic incompatibilities***

677 To estimate recombination rates along the genome of *S. squalidus* we mapped markers from
678 the most extensive genetic linkage map available [53], which was obtained from crosses between
679 the two parental species, to the new assembly. We used BLASTN v2.2 [98], retaining only the top-
680 hit with an E-value below 1e-30 and an identity above 95%. After removal of markers with
681 different order in the linkage map and the genomic assembly, we used MAREYMAP v1.3 [116] to

682 estimate local recombination rate along each chromosome. Additionally, we located markers
683 flanking the genomic position of the genetic incompatibilities identified in crosses between the
684 two parental species [52,53] on the *S. squalidus* genome using BLASTN with the same settings as
685 above.

686 As an additional measure of the recombination rate along the genomes of *S. squalidus*,
687 *S. aethnensis* and *S. chrysanthemifolius*, we used the population-level data described above to
688 estimate effective recombination based on observed patterns of LD between SNPs. For each
689 species, we obtained a new VCF file with SAMTOOLS as described above but performed
690 genotype calling jointly for all conspecific individuals. We filtered these joint VCF files with
691 VCFTOOLS v0.1 [117], retaining only biallelic SNPs with less than 80% missing data and excluding
692 SNPs within 1,000 bp of each other. For each chromosome we phased the filtered SNP subsets
693 with Beagle v 5.2 [118] using default values, and plotted resulting patterns of LD between pairs
694 of SNPs using the R package LDheatmap [119].

695

696 ***QUANTIFICATION AND STATISTICAL ANALYSIS***

697 Statistical analyses were carried out in R v. 3.8 (https://www.r-project.org). For identification
698 of genomic regions with elevated Tajima's D, raw values were transformed into Z-scores (with
699 the *scale* function in R), and high Tajima D regions identified as those with a Z-score above
700 1.65 (approx. 1-tailed $P < 0.05$). For comparison of statistics related to polymorphism,
701 divergence and recombination in different genomic regions, Welch two sample t-tests were used
702 (with the *t.test* function in R).

703

704 ***Supplementary Tables not in SI file***

705 Table S1: Identification of samples of the three *Senecio* species used in this study (Excel spreadsheet).
706 **Related to Figure 1.**
707
708 Table S2: Annotation of the newly assembled genome of *S. squalidus* (gzipped gff file). **Related to**
709 **Table 1.**

References

1. Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C.A., Buggs, R., et al. (2013). Hybridization and speciation. J. Evol. Biol. *26*, 229-246.

2. Nieto Feliner, G., Casacuberta, J., and Wendel, J.F. (2020). Genomics of evolutionary novelty in hybrids and polyploids. Front. Genet. *11*, 792.

3. Moran, B.M., Payne, C., Langdon, Q., Powell, D.L., Brandvain, Y., and Schumer, M. (2021). The genomic consequences of hybridization. Elife *10*, e69016.

4. Bock, D.G., Cai, Z., Elphinstone, C., Gonzalez-Segovia, E., Hirabayashi, K., Huang, K., Keais, G.L., Kim, A., Owens, G.L., and Rieseberg, L.H. (2023). Genomics of plant speciation. Plant Commun. *4*, 100599.

5. Yakimowski, S.B., and Rieseberg, L.H. (2014). The role of homoploid hybridization in evolution: a century of studies synthesizing genetics and ecology. Am. J. Bot. *101*, 1247-1258.

6. Jiggins, C.D., Salazar, C., Linares, M., and Mavarez, J. (2008). Hybrid trait speciation and *Heliconius* butterflies. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 3047-3054.

7. Schumer, M., Rosenthal, G.G., and Andolfatto, P. (2014). How common is homoploid hybrid speciation? Evolution *68*, 1553-1560.

8. Nieto Feliner, G., Álvarez, I., Fuertes Aguilar, J., Heuertz, M., Marques, I., Moharrek, F., Piñeiro, R., Riina, R., Rosselló, J.A., Soltis, P.S., et al. (2017). Is homoploid hybrid speciation that rare? An empiricist's view. Heredity *118*, 513-516.

9. Owens, G.L., Huang, K., Todesco, M., and Rieseberg, L.H. (2023). Re-evaluating homoploid reticulate evolution in *Helianthus* sunflowers. Mol. Biol. Evol. *40*, msad013.

10. Wang, Z., Jiang, Y., Bi, H., Lu, Z., Ma, Y., Yang, X., Chen, N., Tian, B., Liu, B., Mao, X., et al. (2021). Hybrid speciation via inheritance of alternate alleles of parental isolating genes. Mol. Plant *14*, 208-222.

11. Brennan, A.C., Hiscock, S.J., and Abbott, R.J. (2019). Completing the hybridization triangle: the inheritance of genetic incompatibilities during homoploid hybrid speciation in ragworts (*Senecio*). AoB Plants *11*, ply078.

12. Nevado, B., Harris, S.A., Beaumont, M.A., and Hiscock, S.J. (2020). Rapid homoploid hybrid speciation in British gardens: the origin of Oxford ragwort (Senecio squalidus). Mol. Ecol. *29*, 4221-4233.

13. *Heliconius* Genome, C. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature *487*, 94-98.

14. Rosser, N., Seixas, F., Queste, L.M., Cama, B., Mori-Pezo, R., Kryvokhyzha, D., Nelson, M., Waite-Hudson, R., Goringe, M., Costa, M., et al. (2024). Hybrid speciation driven by multilocus introgression of ecological traits. Nature *628*, 811-817.

15. Lamichhaney, S., Han, F., Webster, M.T., Andersson, L., Grant, B.R., and Grant, P.R. (2018). Rapid hybrid speciation in Darwin's finches. Science *359*, 224-228.

16. Zou, T., Kuang, W., Yin, T., Frantz, L., Zhang, C., Liu, J., Wu, H., and Yu, L. (2022). Uncovering the enigmatic evolution of bears in greater depth: the hybrid origin of the Asiatic black bear. Proc. Natl. Acad. Sci. USA *119*, e2120307119.

17. Wu, H., Wang, Z., Zhang, Y., Frantz, L., Roos, C., Irwin, D.M., Zhang, C., Liu, X., Wu, D., Huang, S., et al. (2023). Hybrid origin of a primate, the gray snub-nosed monkey. Science *380*, eabl4997.

18. James, J.K., and Abbott, R.J. (2005). Recent, allopatric, homoploid hybrid speciation: the origin of *Senecio squalidus* (Asteraceae) in the British Isles from a hybrid zone on Mount Etna, Sicily. Evolution *59*, 2533-2547.

19. Abbott, R., James, J.K., Irwin, J., and Comes, H. (2000). Hybrid origin of the Oxford Ragwort, *Senecio squalidus* L. Watsonia *23*, 123-138.

20. Harris, S. (2002). Introduction of Oxford ragwort, *Senecio squalidus* L. (Asteraceae), to the United Kingdom. Watsonia *24*, 31-43.

21. Brennan, A.C., Barker, D., Hiscock, S.J., and Abbott, R.J. (2012). Molecular genetic and quantitative trait divergence associated with recent homoploid hybrid speciation: a study of *Senecio squalidus* (Asteraceae). Heredity *108*, 87-95.

22. Abbott, R.J., Hegarty, M.J., Hiscock, S.J., and Brennan, A.C. (2010). Homoploid hybrid speciation in action. Taxon *59*, 1375-1386.

23. Abbott, R.J., and Brennan, A.C. (2014). Altitudinal gradients, plant hybrid zones and evolutionary novelty. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130346.

24. Vallejo-Marin, M., and Hiscock, S.J. (2016). Hybridization and hybrid speciation under global change. New Phytol. *211*, 1170-1187.

773 25. Druce, G.C. (1927). The Flora of Oxfordshire, 2nd ed. (Oxford: Clarendon Press).

774 26. Kent, D.H. (1956). *Senecio squalidus* L. in the British Isles. 1. Early records (to
775 1877). Proceedings of the Botanical Society of the British Isles *2*, 115-118.

776 27. Kent, D.H. (1960). *Senecio squalidus* L. in the British Isles. 2. The spread from
777 Oxford (1879 - 1939). Proceedings of the Botanical Society of the British Isles *3*, 375-
778 379.

779 28. Abbott, R.J., Brennan, A.C., James, J.K., Forbes, D.G., Hegarty, M.J., and Hiscock,
780 S.J. (2009). Recent hybrid origin and invasion of the British Isles by a self-
781 incompatible species, Oxford ragwort (*Senecio squalidus* L., Asteraceae). Biol.
782 Invasions *11*, 1145.

783 29. Barone, G., Domina, G., Bartolucci, F., Galasso, G., and Peruzzi, L. (2022). A
784 nomenclatural and taxonomic revision of the *Senecio squalidus* Group (Asteraceae).
785 Plants (Basel) *11*.

786 30. Hind, N., and King, C. (2022). *Senecio squalidus*: Compositae. Curtis's Botanical
787 Magazine *39*, 113-134.

788 31. Abbott, R.J., and Lowe, A.J. (2004). Origins, establishment and evolution of new
789 polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. Biol.
790 J. Linn. Soc. *82*, 467-474.

791 32. Kim, M., Cui, M.-L., Cubas, P., Gillies, A., Lee, K., Chapman, M.A., Abbott, R., and
792 Coen, E. (2008). Regulatory genes control a key morphological and ecological trait
793 transferred between species. Science *322*, 1116-1119.

794 33. Hegarty, M., Coate, J., Sherman-Broyles, S., Abbott, R., Hiscock, S., and Doyle, J.
795 (2013). Lessons from natural and artificial polyploids in higher plants. Cytogenet.
796 Genome Re.s *140*, 204-225.

797 34. Crisp, P., and Jones, B.M.G. (1978). Hybridization of *Senecio squalidus* and *S.*
798 *viscosus* and introgression of genes from the diploid into tetraploid Senecio species.
799 Ann. Bot. *42*, 937-944.

800 35. Abbott, R.J., and Forbes, D.G. (1993). Outcrossing rate and self-incompatibility in the
801 colonizing species *Senecio squalidus*. Heredity *71*, 155-159.

802 36. Hiscock, S.J. (2000). Self-incompatibility in *Senecio squalidus* L. (Asteraceae). Ann.
803 Bot. *85*, 181-190.

804 37. Brennan, A.C., Harris, S.A., and Hiscock, S.J. (2013). The population genetics of
805 sporophytic self-incompatibility in three hybridizing *Senecio* (Asteraceae) species
806 with contrasting population histories. Evolution *67*, 1347-1367.

807 38. Brennan, A.C., Harris, S.A., Tabah, D.A., and Hiscock, S.J. (2002). The population
808 genetics of sporophytic self-incompatibility in *Senecio squalidus* L. (Asteraceae) I: S
809 allele diversity in a natural population. Heredity *89*, 430-438.

810 39. Baker, H.G. (1955). Self-Compatibility and establishment after 'Long-Distance'
811 dispersal. Evolution *9*, 347-349.

812 40. Pannell, J.R., Auld, J.R., Brandvain, Y., Burd, M., Busch, J.W., Cheptou, P.O.,
813 Conner, J.K., Goldberg, E.E., Grant, A.G., Grossenbacher, D.L., et al. (2015). The
814 scope of Baker's law. New Phytol. *208*, 656-667.

41. Pannell, J.R., and Barrett, S.C.H. (1998). Baker's Law revisited: reproductive assurance in a metapopulation. Evolution *52*, 657-668.

42. Brennan, A.C., Harris, S.A., and Hiscock, S.J. (2003). Population genetics of sporophytic self-incompatibility in *Senecio squalidus* L. (Asteraceae) II: a spatial autocorrelation approach to determining mating behaviour in the presence of low S allele diversity. Heredity *91*, 502-509.

43. Brennan, A.C., Harris, S.A., and Hiscock, S.J. (2003). The population genetics of sporophytic self-incompatibility in *Senecio squalidus* L. (Asteraceae): avoidance of mating constraints imposed by low S-allele number. Philos. Trans. R. Soc. Lond. B Biol. Sci. *358*, 1047-1050.

44. Brennan, A.C., Tabah, D.A., Harris, S.A., and Hiscock, S.J. (2011). Sporophytic self-incompatibility in *Senecio squalidus* (Asteraceae): S allele dominance interactions and modifiers of cross-compatibility and selfing rates. Heredity *106*, 113-123.

45. Hiscock, S.J., McInnis, S.M., Tabah, D.A., Henderson, C.A., and Brennan, A.C. (2003). Sporophytic self-incompatibility in *Senecio squalidus* L. (Asteraceae) – the search for S. J. Exp. Bot. *54*, 169-174.

46. Walter, G.M., Abbott, R.J., Brennan, A.C., Bridle, J.R., Chapman, M., Clark, J., Filatov, D., Nevado, B., Ortiz-Barrientos, D., and Hiscock, S.J. (2020). *Senecio* as a model system for integrating studies of genotype, phenotype and fitness. New Phytol. *226*, 326-344.

47. Chapman, M.A., Hiscock, S.J., and Filatov, D.A. (2013). Genomic divergence during speciation driven by adaptation to altitude. Mol. Biol. Evol. *30*, 2553-2567.

48. Osborne, O.G., Batstone, T.E., Hiscock, S.J., and Filatov, D.A. (2013). Rapid speciation with gene flow following the formation of Mt. Etna. Genome Biol. Evol. *5*, 1704-1715.

49. Brennan, A.C., Bridle, J.R., Wang, A.L., Hiscock, S.J., and Abbott, R.J. (2009). Adaptation and selection in the *Senecio* (Asteraceae) hybrid zone on Mount Etna, Sicily. New Phytol. *183*, 702-717.

50. Osborne, O.G., Chapman, M.A., Nevado, B., and Filatov, D.A. (2016). Maintenance of species boundaries despite ongoing gene flow in ragworts. Genome Biol. Evol. *8*, 1038-1047.

51. Wong, E.L.Y., Nevado, B., Osborne, O.G., Papadopulos, A.S.T., Bridle, J.R., Hiscock, S.J., and Filatov, D.A. (2020). Strong divergent selection at multiple loci in two closely related species of ragworts adapted to high and low elevations on Mount Etna. Mol. Ecol. *29*, 394-412.

52. Brennan, A.C., Hiscock, S.J., and Abbott, R.J. (2014). Interspecific crossing and genetic mapping reveal intrinsic genomic incompatibility between two *Senecio* species that form a hybrid zone on Mount Etna, Sicily. Heredity *113*, 195–204.

53. Chapman, M.A., Hiscock, S.J., and Filatov, D.A. (2016). The genomic bases of morphological divergence and reproductive isolation driven by ecological speciation in *Senecio* (Asteraceae). J. Evolution. Biol. *29*, 98-113.

54. Hegarty, M.J., Barker, G.L., Brennan, A.C., Edwards, K.J., Abbott, R.J., and Hiscock, S.J. (2008). Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in *Senecio*. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 3055-3069.

55. Hegarty, M.J., Barker, G.L., Brennan, A.C., Edwards, K.J., Abbott, R.J., and Hiscock, S.J. (2009). Extreme changes to gene expression associated with homoploid hybrid speciation. Mol. Ecol. *18*, 877-889.

56. Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2006). Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. Curr. Biol. *16*, 1652-1659.

57. Sell, P., Murrell, G., and Walters, S.M. (2006). Flora of Great Britain and Ireland: Volume 4, Campanulaceae - Asteraceae (Cambridge University Press).

58. Ross, R.I.C. (2010). Local adaptation and adaptive divergence in a hybrid species complex in *Senecio*. PhD thesis, University of Oxford.

59. Stroh, P., Walker, K., Humphrey, T., Pescott, O., and Burkmar, R. (2023). Plant Atlas 2020. Mapping Changes in the Distribution of the British and Irish Flora (Princeton University Press).

60. Preston, C.D., Pearman, D., Dines, T.D., and Isles, B.S.o.t.B. (2002). New Atlas of the British & Irish Flora: An Atlas of the Vascular Plants of Britain, Ireland, the Isle of Man and the Channel Islands (Oxford University Press).

61. Sibthorp, J. (1794). Flora Oxoniensis, exhibens plantas in agro Oxoniensis sponte crescentes, secundum Systema Sexuale Distributas. Oxoni Typis Academicus, Oxford.

62. Bennett, M.D., and Smith, J.B. (1976). Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. Lond. B Biol. Sci. *274*, 227-274.

881 63. Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N.M., Salman-Minkov, A.,
882 Mayzel, J., Chay, O., and Mayrose, I. (2015). The Chromosome Counts Database
883 (CCDB) - a community resource of plant chromosome numbers. New Phytol. *206*,
884 19-26.
885 64. BlobToolKit browser, *Senecio squalidus* (available at
886 https://blobtoolkit.genomehubs.org/view/Senecio%20squalidus/dataset/CAJVGC01/re
887 port#Settings).
888 65. Genome-note server, *Senecio squalidus* (available at https://genome-note-
889 higlass.tol.sanger.ac.uk/l/?d=LzVoH3QgQq6VhzSIjSxELg).
890 66. Kandziora, M., Kadereit, J.W., and Gehrke, B. (2017). Dual colonization of the
891 Palaearctic from different regions in the Afrotropics by *Senecio*. J. Biogeogr. *44*, 147-
892 157.
893 67. Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, L.E., and Funk,
894 V.A. (2019). A fully resolved backbone phylogeny reveals numerous dispersals and
895 explosive diversifications throughout the history of Asteraceae. Proc. Natl. Acad. Sci.
896 USA *116*, 14083-14088.
897 68. Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-
898 Brière, C., Owens, G.L., Carrère, S., Mayjonade, B., et al. (2017). The sunflower
899 genome provides insights into oil metabolism, flowering and Asterid evolution.
900 Nature *546*, 148-152.
901 69. Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J.,
902 and Rieseberg, L.H. (2008). Multiple paleopolyploidizations during the evolution of
903 the Compositae reveal parallel patterns of duplicate gene retention after millions of
904 years. Mol. Biol. Evol. *25*, 2445-2455.

70. Barker, M.S., Li, Z., Kidder, T.I., Reardon, C.R., Lai, Z., Oliveira, L.O., Scascitelli, M., and Rieseberg, L.H. (2016). Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am. J. Bot. *103*, 1203-1211.

71. Grant, V. (1958). The regulation of recombination in plants. Cold Spring Harb. Symp. Quant. Biol. *23*, 337-363.

72. Grant, V. (1981). Plant speciation, 2nd Edition (New York: Columbia University Press).

73. Kamau, E., and Charlesworth, D. (2005). Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant *Arabidopsis lyrata*. Curr. Biol. *15*, 1773-1778.

74. Schumer, M., Cui, R., Rosenthal, G.G., and Andolfatto, P. (2015). Reproductive isolation of hybrid populations driven by genetic incompatibilities. PLoS Genet. *11*, e1005041.

75. Rieseberg, L.H., Van Fossen, C., and Desrochers, A.M. (1995). Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature *375*, 313-316.

76. Ungerer, M.C., Baird, S.J., Pan, J., and Rieseberg, L.H. (1998). Rapid hybrid speciation in wild sunflowers. Proc. Natl. Acad. Sci. USA *95*, 11757-11762.

77. Buerkle, C.A., and Rieseberg, L.H. (2008). The rate of genome stabilization in homoploid hybrid species. Evolution *62*, 266-275.

78. Sun, Y.S., Lu, Z.Q., Zhu, X.F., and Ma, H. (2020). Genomic basis of homoploid hybrid speciation within chestnut trees. Nat. Commun. *11*.

79. Schumer, M., Xu, C., Powell, D.L., Durvasula, A., Skov, L., Holland, C., Blazier, J.C., Sankararaman, S., Andolfatto, P., Rosenthal, G.G., et al. (2018). Natural

selection interacts with recombination to shape the evolution of hybrid genomes. Science *360*, 656-660.

80. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737-746.

81. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050-1054.

82. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896-2898.

83. Rao, Suhas S.P., Huntley, Miriam H., Durand, Neva C., Stamenova, Elena K., Bochkov, Ivan D., Robinson, James T., Sanborn, Adrian L., Machol, I., Omer, Arina D., Lander, Eric S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

84. Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M., and Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput. Biol. *15*, e1007273.

85. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, 1207.3907.

86. Challis, R., Richards, E., Rajan, J., Cochrane, G., and Blaxter, M. (2020). BlobToolKit - interactive quality assessment of genome assemblies. G3 (Bethesda) *10*, 1361-1374.

953 87. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly
954 of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18.
955 88. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and
956 Greiner, S. (2017). GeSeq - versatile and accurate annotation of organelle genomes.
957 Nucleic Acids Res. *45*, W6-W11.
958 89. Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW
959 (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of
960 organellar genomes. Nucleic Acids Res. *47*, W59-W64.
961 90. Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: a multiplatform
962 graphical user interface for sequence alignment and phylogenetic tree building. Mol.
963 Biol. Evol. *27*, 221-224.
964 91. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker Open-4.0 (available at
965 http://www.repeatmasker.org).
966 92. Smit, A.F.A., and Hubley, R. RepeatModeler Open-1.0 (available at
967 http://www.repeatmasker.org).
968 93. Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez
969 Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline
970 designed for emerging model organism genomes. Genome Res. *18*, 188-196.
971 94. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.
972 95. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome
973 assembly and annotation completeness. Methods Mol. Biol. *1962*, 227-245.
974 96. Beier, S., Thiel, T., Munch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a
975 web server for microsatellite prediction. Bioinformatics *33*, 2583-2585.
976 97. UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic
977 Acids Res. *49*, D480-D489.

98. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

99. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

100. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. protoc. *8*, 1494-1512.

101. Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science *320*, 486-488.

102. Tang, H., Krishnakumar, V., and Li, J. (2015). jcvi: JCVI utility libraries (v0.5.7). Zenodo *https://doi.org/10.5281/zenodo.31631*.

103. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

104. Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. *9*, 286-298.

105. Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.

106. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268-274.

107. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518-522.

108. Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics *19*, 153.

109. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586-1591.

110. dos Reis, M., and Yang, Z. (2011). Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. *28*, 2161-2172.

111. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol.. *67*, 901-904.

112. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

113. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

114. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987-2993.

115. Ferretti, L., Raineri, E., and Ramos-Onsins, S. (2012). Neutrality tests for sequences with missing data. Genetics *191*, 1397-1401.

116. Rezvoy, C., Charif, D., Gueguen, L., and Marais, G.A. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics *23*, 2188-2189.

117. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.

118. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084-1097.

119. Shin, J.H., Blay, S., McNeney, B., and Graham, J. (2006). LDheatmap: an R function for graphical display of pairwise Linkage Disequilibria between Single Nucleotide Polymorphisms. J. Stat. Soft. *16*.

S. squalidus (UK)

1690s — The three taxa are present in Badminton [20]

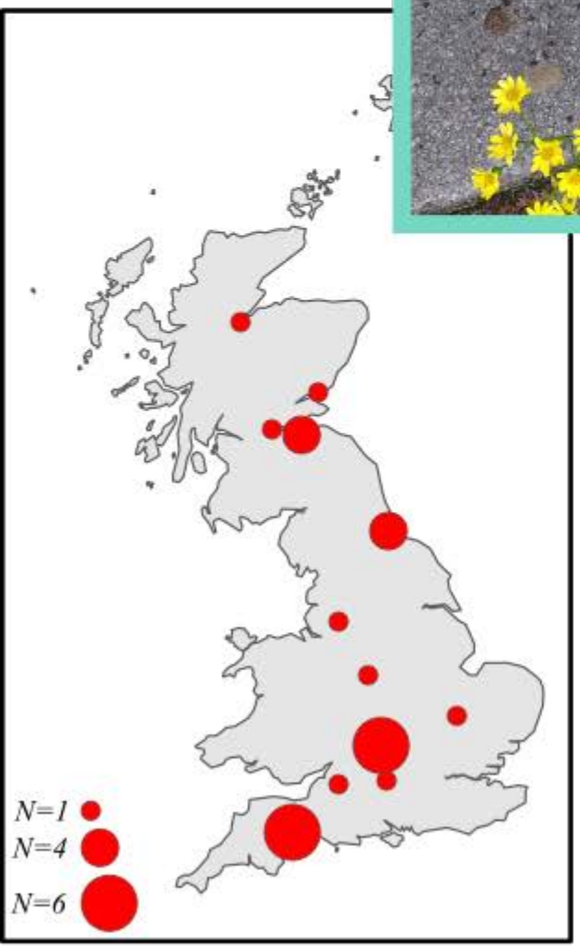
1700..1719 — Hybrid material introduced into Oxford BG from Badminton. *S. chrysanthemifolius* also present in Oxford BG [20]

1736 — *S. squalidus* type material sent to Linnaeus. Likely only this species still in cultivation at Oxford BG [20]

1794 — *S. squalidus* plants growing outside Oxford BG walls [61; but see fn. 3 in ref 20]

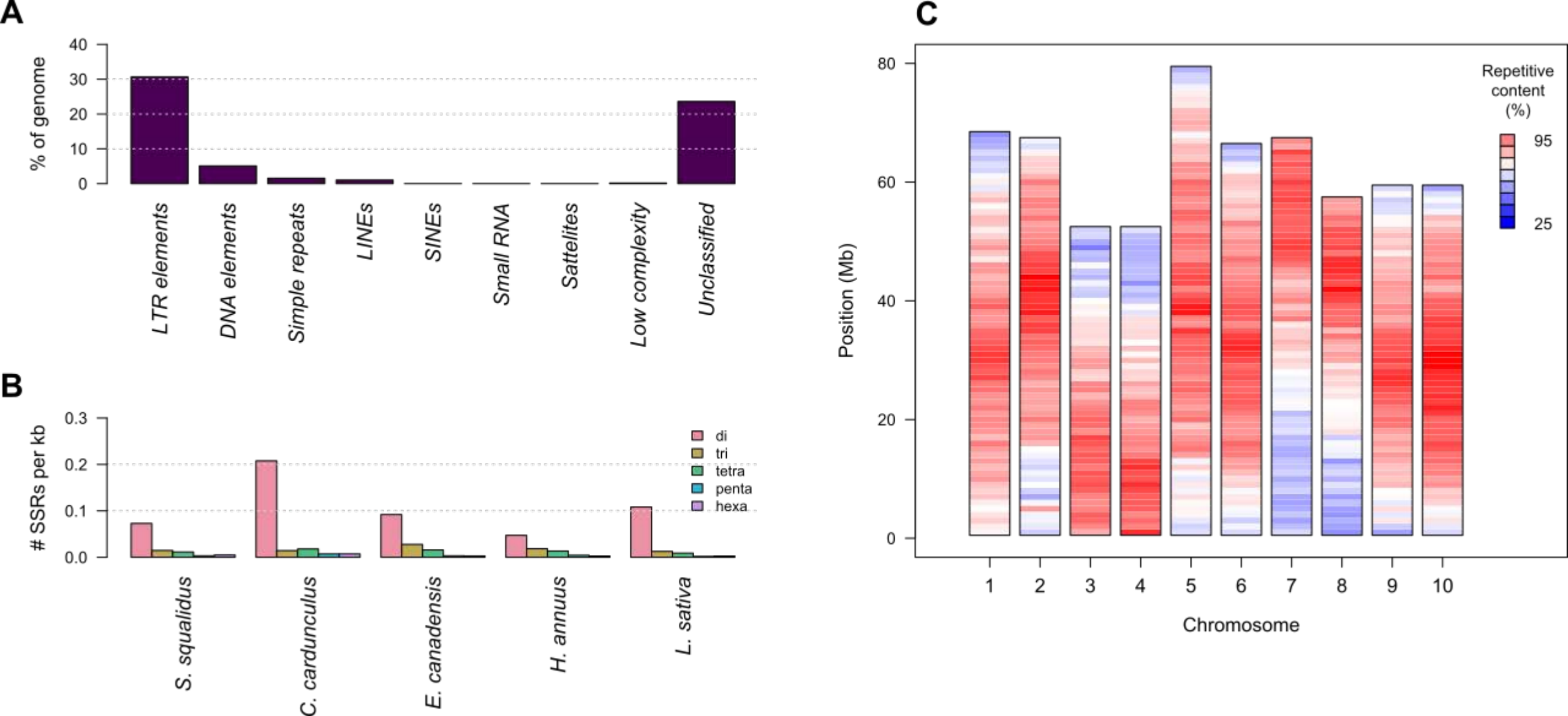
1820..1850 — From 1820~1850 onwards, *S. squalidus* starts to spread outside Oxford [26, 27]

1999 — By 1999, *S. squalidus* is present in most of Great Britain [59, 60]

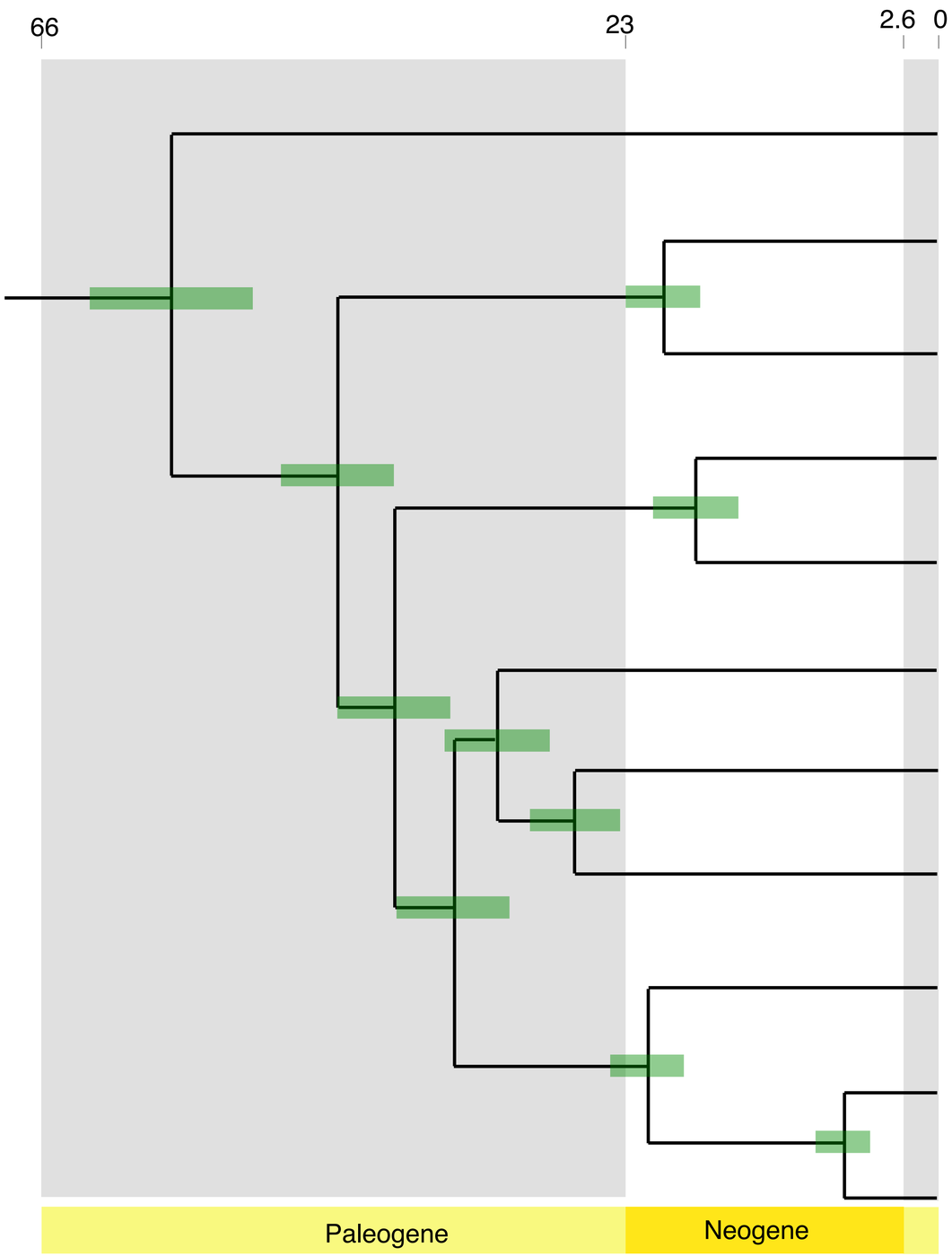


S. aethnensis (Mt Etna, > 2000m)

S. chrysanthemifolius (Mt Etna, < 1000m)

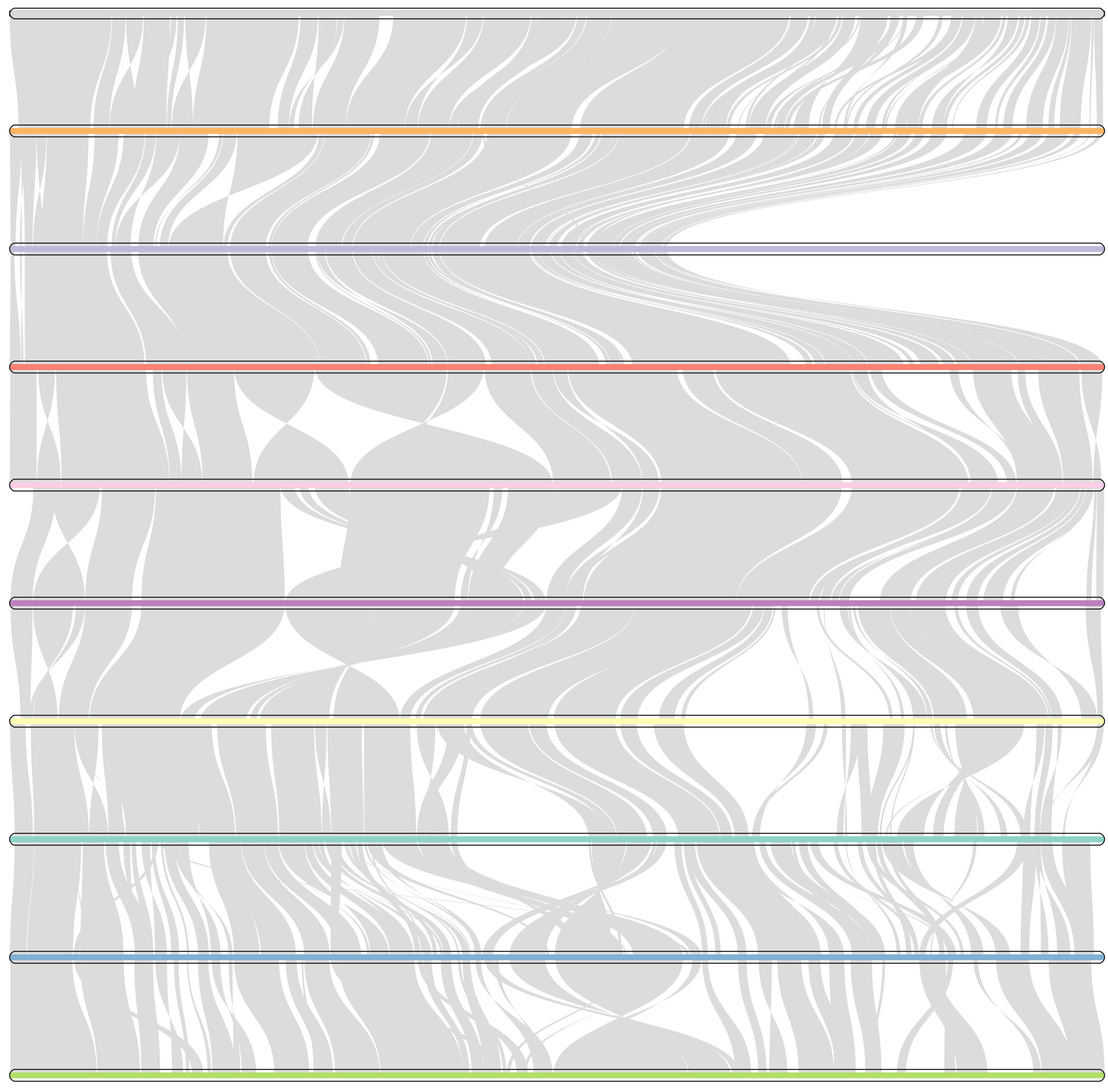

A
% of genome
40
30
20
10
0
LTR elements
DNA elements
Simple repeats
LINEs
SINEs
Small RNA
Sattelites
Low complexity
Unclassified
B
SSRs per kb
0.3
0.2
0.1
0.0
di
tri
tetra
penta
hexa
S. squalidus
C. cardunculus
E. canadensis
H. annuus
L. sativa
C
Position (Mb)
80
60
40
20
0
Repetitive content (%)
95
25
1
2
3
4
5
6
7
8
9
10
Chromosome

A

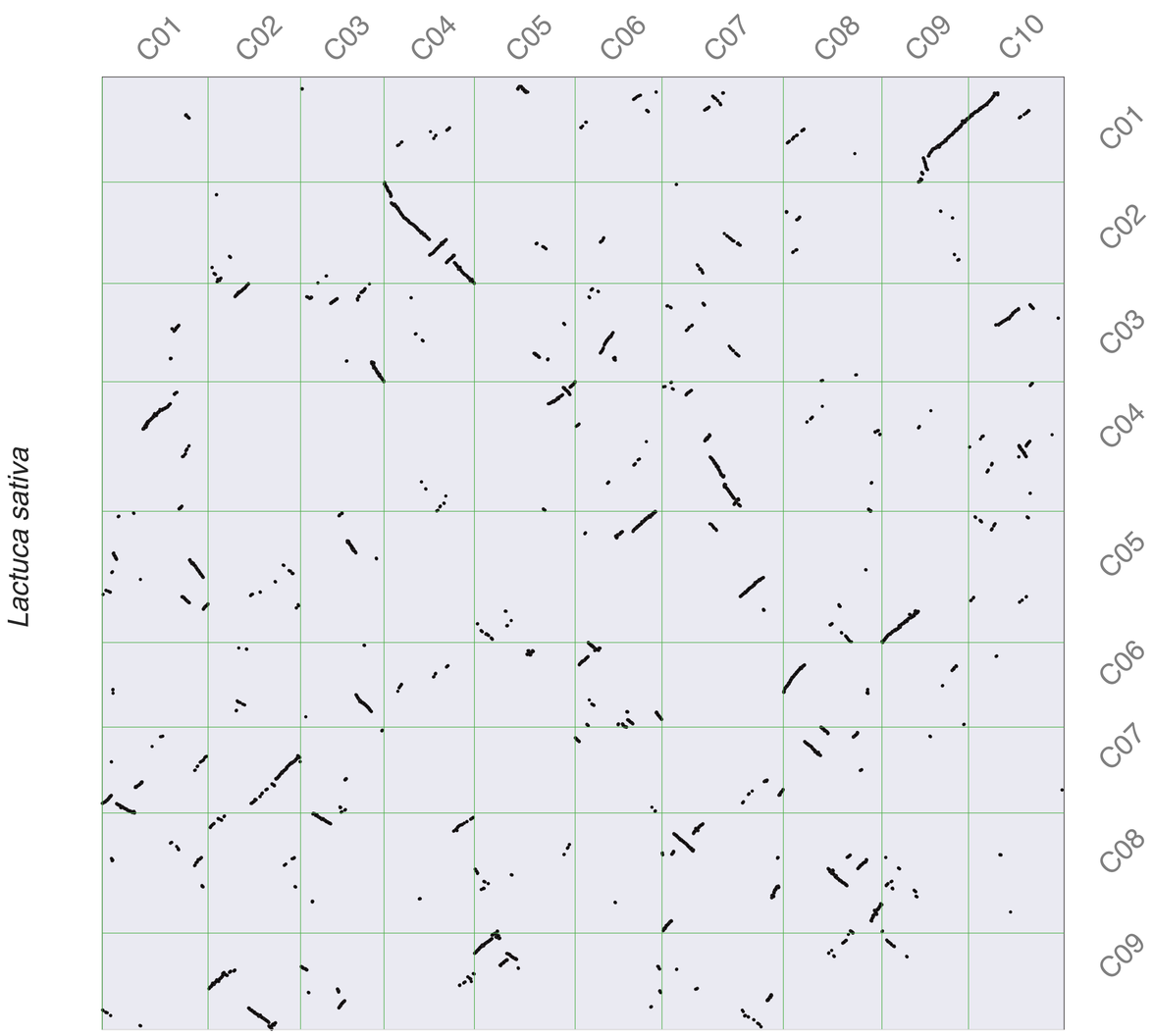

66
23
2.6
0
Paleogene
Neogene

B

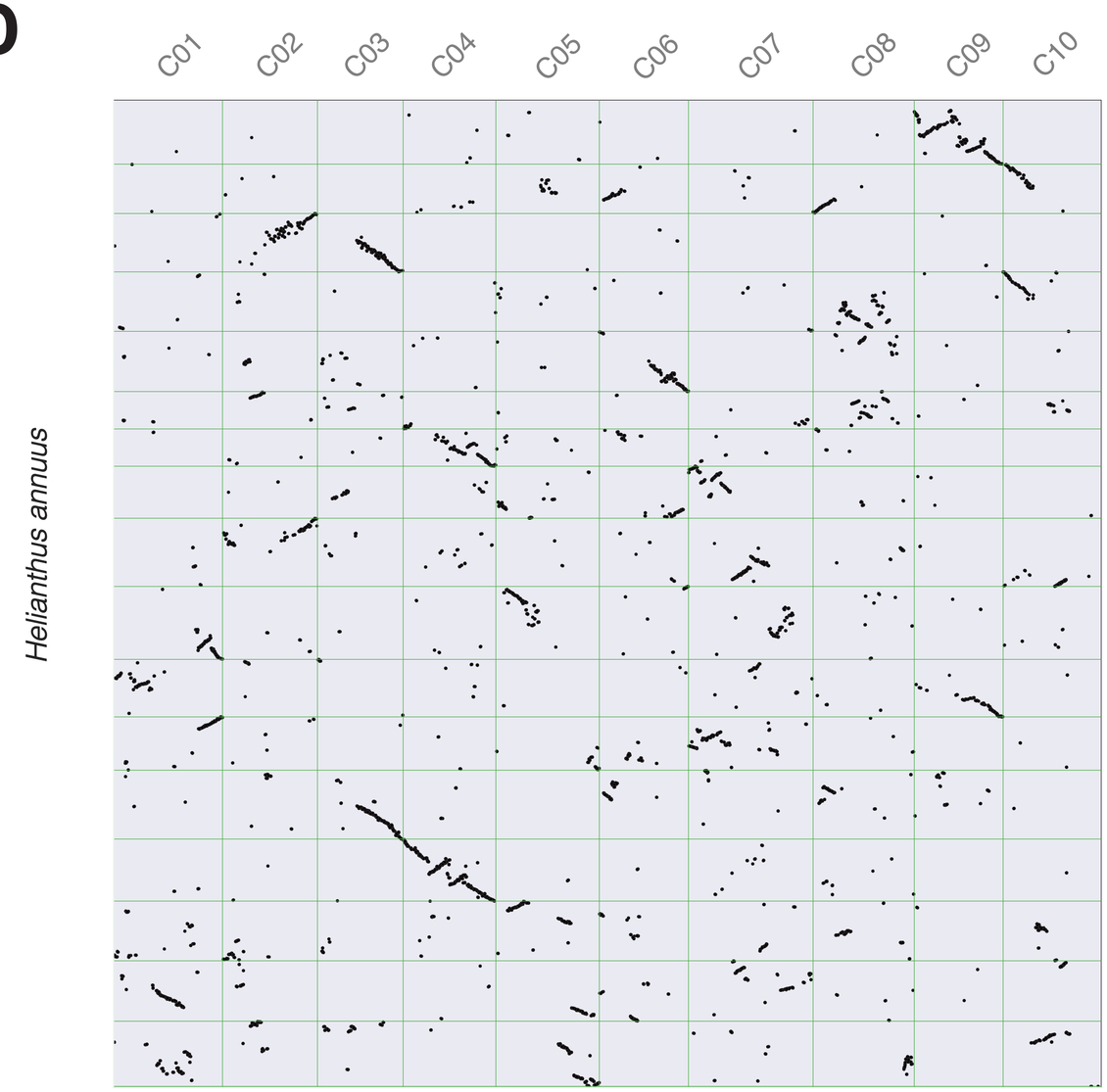

Barnadesia spinosa
Cynara cardunculus
Carthamus tinctorius
Taraxacum kok-saghyz
Lactuca sativa
Senecio squalidus
Chrysanthemum seticuspe
Erigeron canadensis
Mikania micrantha
Helianthus annuus
Scalesia atractyloides



C


C01
C02
C03
C04
C05
C06
C07
C08
C09
C10
Lactuca sativa
Senecio squalidus

D


C01
C02
C03
C04
C05
C06
C07
C08
C09
C10
C11
C12
C13
C14
C15
C16
C17
Helianthus annuus
Senecio squalidus