Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates

Adam Nunn\(^1,2\), Isaac Rodríguez-Arévalo\(^3,4\), Zenith Tandukar\(^5\), Katherine Frels\(^5,6\), Adrián Contreras-Garrido\(^7\), Pablo Carbonell-Bejerano\(^7\), Panpan Zhang\(^8,9\), Daniela Ramos Cruz\(^3,4\), Katharina Jandrasits\(^3,4\), Christa Lanz\(^7\), Anthony Brusa\(^5\), Marie Mirozou\(^8,9\), Kevin Dorn\(^1,10,11\), David W Galbraith\(^12\), Breeze A. Jarvis\(^13\), John C. Sedbrook\(^13\), Donald L. Wyse\(^5\), Christian Otto\(^1\), David Langenberger\(^1\), Peter F. Stadler\(^2,14\), Detlef Weigel\(^7\), M. David Marks\(^10\), James A. Anderson\(^5\), Claude Becker\(^3,4\)* and Ratan Chopra\(^5,10,16\)

\(^{1}\)ecSeq Bioinformatics GmbH, Leipzig, Germany
\(^{2}\)Department of Computer Science, Leipzig University, Leipzig, Germany
\(^{3}\)Genetics, Faculty of Biology, Ludwig Maximilians University, Martinsried, Germany
\(^{4}\)Gregor Mendel Institute of Molecular Plant Biology GmbH, Austrian Academy of Sciences (OAW), Vienna BioCenter (VBC), Vienna, Austria
\(^{5}\)Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, MN, USA
\(^{6}\)Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE, USA
\(^{7}\)Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany
\(^{8}\)Institut de Recherche pour le Développement, UMR232 DIADE, Montpellier, France
\(^{9}\)Laboratory of Plant Genome and Development, University of Perpignan, Perpignan, France
\(^{10}\)Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN, USA
\(^{11}\)USDA-ARS, Soil Management and Sugarbeet Research, Fort Collins, CO, USA
\(^{12}\)BIOS Institute, Arizona Cancer Center, Department of Biomedical Engineering, University of Arizona, School of Plant Sciences, Tucson, AZ, USA
\(^{13}\)School of Biological Sciences, Illinois State University, Normal, IL, USA
\(^{14}\)Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

**Summary**

*Thlaspi arvense* (field pennycress) is being domesticated as a winter annual oilseed crop capable of improving ecosystems and intensifying agricultural productivity without increasing land use. It is a selfing diploid with a short life cycle and is amenable to genetic manipulations, making it an accessible field-based model species for genetics and epigenetics. The availability of a high-quality reference genome is vital for understanding pennycress physiology and for clarifying its evolutionary history within the Brassicaceae. Here, we present a chromosome-level genome assembly of var. MN106-Ref with improved gene annotation and use it to investigate gene structure differences between two accessions (MN108 and Spring32-10) that are highly amenable to genetic transformation. We describe non-coding RNAs, pseudogenes and transposable elements, and highlight tissue-specific expression and methylation patterns. Resequencing of forty wild accessions provided insights into genome-wide genetic variation, and QTL regions were identified for a seedling colour phenotype. Altogether, these data will serve as a tool for pennycress improvement in general and for translational research across the Brassicaceae.

**Introduction**

Native to Eurasia, field pennycress (*Thlaspi arvense L.*) is a member of the Brassicaceae family and is closely related to the oilseed crop species rapeseed (*Brassica rapa* and *Brassica napus* L.), camelina (*Camelina sativa* L.) and the wild plant *Arabidopsis thaliana* (Beilstein et al., 2010; Warwick et al., 2002). It is an emerging oil feedstock species with the potential to improve sustainability of cold climate cropping systems through use as a cash cover crop (Boateng et al., 2010; Chopra et al., 2018; Sedbrook et al., 2014). Pennycress is extremely winter hardy (Warwick et al., 2002) and can be planted in traditional fallow periods following summer annuals such as wheat, maize or soya bean (Cubins et al., 2019; Johnson et al., 2015; Ott et al., 2019; Phippen and Phippen, 2012). By providing a protective living cover from the harvest of the previous summer annual crop through early spring, pennycress prevents soil erosion and nutrient loss, which in turn protects surface and below-ground water sources, suppresses early-season weed growth, and provides a food source for pollinators (Del Gatto et al., 2015; Johnson et al., 2015; Weyers et al., 2019, 2021). The short life cycle allows for harvest in May or June in temperate regions, with reported seed yields ranging from 750 to 2400 kg/ha (Cubins et al., 2019; Moore et al., 2020). Following harvest, an additional crop of summer annuals can be grown in a double-crop system that provides increased total seed yields and beneficial ecosystem
services (Johnson et al., 2015; Phippen and Phippen, 2012; Thomas et al., 2017). The pennycress seed contains an average of 30%–35% oil, and the fatty acid profile is conducive to producing biofuels (Fan et al., 2013; Moser, 2012; Moser et al., 2009). Seed oil also has the potential to be converted into an edible oil and protein source (Chopra et al., 2020b; Claver et al., 2017; McGinn et al., 2019).

*Thlaspi arvense* is a homoygous diploid species (2n = 2x = 14) (Mulligan, 1957) and is predominantly self-pollinating (Mulligan and Kevan, 1973), suggesting that breeding efforts could proceed with relative ease and speed. It is amenable to genetic transformation using the floral dip method (McGinn et al., 2019), and its diploid nature with many one-to-one gene correspondence with *A. thaliana* (Chopra et al., 2018) could provide an avenue for gene discovery followed by field-based phenotypic validation. Indeed, several agronomic and biochemical traits have already been identified in pennycress using this translational approach, including traits crucial for de novo domestication of *T. arvense* such as transparent testa phenotypes (Chopra et al., 2018), early flowering (Chopra et al., 2020b), reduced shatter (Chopra et al., 2020b) and seed oil composition traits (Chopra et al., 2020b; Esfahanian et al., 2021; Jarvis et al., 2021; McGinn et al., 2019). Field pennycress could thus serve as a de novo-domesticated oilseed crop for the cooler climates of the world and at the same time as a new dicotyledonous model for functional genetics studies. Its amenability for translational research constitutes a clear advantage vis-a-vis *A. thaliana*.

However, to establish *T. arvense* as a genetic model and a crop, it is important to develop genomic resources that will help explore the spectrum of genetic diversity, the extent and patterns of gene expression, genetic structure and untapped genetic potential for crop improvement.

Here, we describe a set of new resources developed for research and breeding communities, including a high-quality, chromosome-level genome assembly of *T. arvense* var. MN106-Ref, representing ~97.5% of the estimated genome size of 539 Mbp. We provide robust annotations of both protein-coding and non-coding genes, including putative transfer RNA (tRNA), ribosomal RNA (rRNA) and small nucleolar RNA (snoRNA) predictions, alongside small RNA-producing loci, transposable element (TE) families and predicted pseudogenes. From transcriptome data based on a panel of eleven different tissues and life stages, we built a gene expression atlas. In combination with whole-genome DNA methylation profiles of both roots and shoots, this provides a basis for exploring gene regulatory and/or epigenetic mechanisms within pennycress. A comprehensive analysis of forty resequenced pennycress accessions highlights the nucleotide diversity in these collections, alongside gene variants and population structure. Finally, by means of modified bulked-segregant analysis (BSA), we identified quantitative trait loci (QTL) associated with seedling colour phenotype, exemplifying the usefulness of this resource. The genome and resequencing information presented in this study will increase the value of pennycress as a model and as tool for translational research and accelerate pennycress breeding through the discovery of genes affecting important agronomic traits.

**Results**

An improved reference genome sequence

The genome of *T. arvense* var. MN106-Ref was assembled *de novo* from 476X (256 Gb) depth PacBio Sequel II continuous long reads (CLRs) (38 kb N50). The initial assembly attempts exceeded the genome size by ~53% with respect to the range of 459–540 Mb total size estimated from flow cytometry and k-mer analyses (Table S1). Reducing the duplicated fraction, polishing and scaffolding/resc scaffolding using several approaches resulted in a final assembly of ~526 Mb, corresponding to ~97.5% of the upper limit of the flow cytometry-based estimate and representing an improvement of ~20% relative to the original assembly size. Scaffolding/resc scaffolding of the genome assembly was achieved using Bionano optical, Hi-C contact, genetic linkage and comparative synteny maps. The final genome contains 964 scaffolds, with ~83.6% of the total estimated size represented by seven large scaffolds, in agreement with the haploid chromosome number, demonstrating a vast improvement in overall contiguity and bringing the assembly to chromosome level. The coding space is 98.7% complete on the basis of conserved core eukaryotic single-copy genes (BUSCO), with 92.1% being single copy and 6.6% duplicated. Full descriptive statistics of the final version in comparison with *T. arvense* v1 are given in Table 1; intermediary versions are summarized in Table S2.

The seven largest scaffolds are all characterized by high gene density towards both telomeres and a high density of repeats and TEs in the pericentromeric and centromeric regions (Figure 1, Figure S1). While the protein-coding gene fraction of the genome is similar in size to other closely related Brassicaceae (Wang et al., 2011), the large repetitive fraction suggests an increased genome size driven by TE expansion (Beric et al., 2021). In addition, the spatial distribution of sRNA loci followed the gene density but was concentrated predominantly at the boundary between genes and TEs.

In addition to the duplicate-containing contigs, alignments of the raw CLR reads to the new genome revealed the presence of what appeared to be a small number of collapsed repeats in scaffolds 1, 3, 5 and 7, which were typically larger than 25 kbp and indicative of misassembly in these loci (Figure S2). Further investigation revealed an overlap with tandem repeat clusters of 185 and 285 rRNA annotations at those loci on scaffolds 3 and 5, and a large supersatellite of 5S rRNA on scaffold 1. In addition,

**Table 1** Full descriptive statistics comparing the previously published *T. arvense* v1 assembly with the present version *T. arvense* v2

| Assembly category | *T. arvense* v1 | *T. arvense* v2 |
|-------------------|----------------|----------------|
| No. of contigs    | 44 109         | 4714           |
| Largest contig    | ~              | 41.6 Mbp       |
| contig N50        | 0.02 Mbp       | 13.3 Mbp       |
| No. of scaffolds  | 6768           | 964            |
| No. of scaffolds (≥50 000 bp) | 1807 | 607 |
| Largest scaffold  | 2.4 Mbp        | 70.0 Mbp       |
| Total length      | 343 Mbp        | 526 Mbp        |
| Total length (≥50 000 bp) | 276 Mbp | 514 Mbp |
| GC (%)            | 37.99          | 38.39          |
| N50               | 0.14 Mbp       | 64.9 Mbp       |
| NG50              | 0.05 Mbp       | 64.9 Mbp       |
| N75               | 0.06 Mbp       | 61.0 Mbp       |
| NG75              | ~              | 55.2 Mbp       |
| L50               | 561            | 4              |
| LG50              | 1678           | 4              |
| L75               | 1469           | 6              |
| LG75              | ~              | 7              |
| No. of Ns per 100 kbp | 5165.00 | 0.51 |

© 2022 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 20, 944–963
there were corresponding genes associated with organellar DNA at those loci on scaffolds 3 and 7, indicating either erroneous incorporation of plastome sequence during assembly or genuine nuclear integrations of plastid DNA (NUPTs) (Michalovova et al., 2013).

Comparative genomics

Exploiting information from the genome of Eutrema salsugineum (Yang et al., 2013), a closely related species (Franzke et al., 2011) with a much smaller genome (241 Mbp) but the same karyotype \(n = 7\), aided during rescaffolding (see methods; Figure S3) and confirmed synteny of the seven largest scaffolds in the two species (Figure S4). There is a large-scale synteny between the two genomes, with the exception of some regions on scaffolds 2, 3, 6 and 7. This could be due to the low gene density observed in the \(T. \) arvense genome towards the centre of each chromosome and/or the high presence of dispersed repeats in those regions.

Chromosome evolution in the Brassicaceae has been studied through chromosome painting techniques, and 24 chromosome blocks (A-X) have been defined from an ancestral karyotype of \(n = 8\) (Murat et al., 2015; Schranz et al., 2006). We identified the 24 blocks in \(T. \) arvense based on gene homology and synteny between \(T. \) arvense and \(A. \) thaliana (Figure 2). While in general the distribution of the chromosomal blocks resembles that in the close relatives \(E. \) salsugineum and \(S. \) parvula, some blocks are rearranged in a small section at the end of the scaffold representing chromosome 1 and at the beginning of chromosome 6. The first case involves the transposition of a small part of block C in between A and B, while chromosome 6 has a possible inversion between the blocks O and W when compared to \(E. \) salsugineum and \(S. \) parvula. Overall, despite having an increase in genome size compared with \(E. \) salsugineum and \(S. \) parvula, \(T. \) arvense conserves all the ancestral Brassicaceae karyotype blocks. The synteny analysis also revealed intra-chromosomal rearrangements, but no obvious inter-chromosomal rearrangements.

Genome annotation

Transcriptome assembly

We sequenced total cDNA with strand-specific RNA-seq from eleven tissues, including rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques, old green siliques, green seeds, mature seeds, seed pods, roots of 1-week-old seedlings and shoots of 1-week-old seedlings (Table S3). Reads from each tissue sample were aligned to the genome with unique mapping rates between 76% and 91%, with the exception of old green siliques (19%), green seed (59%) and mature seed (12%). The majority of unmapped reads in each case were due to insufficient high-quality read lengths. We constructed independent tissue-specific transcriptome assemblies and combined them into a multi-sample de novo assembly, yielding 30 650 consensus transcripts. These were further refined by prioritizing isoforms supported by Iso-seq data, resulting in 22 124 high-quality consensus transcripts to inform gene models.

Protein-coding genes

In addition to the expression data, gene models were informed by protein homology using a combined database of Viridiplantae from UniProtKB/Swiss-Prot (Boutet et al., 2007) and selected Brassicaceae from RefSeq (Pruitt et al., 2012). Following initial
training and annotation by *ab initio* gene predictors, protein-coding loci were further annotated with InterPro to provide PFAM domains, which were combined with a BLAST search to the UniProtKB/Swiss-Prot Viridiplantae database to infer gene ontology (GO) terms. In accordance with MAKER-P recommendations (Campbell et al., 2014), the final set of 27,128 protein-coding loci was obtained by filtering out those with an annotation edit distance (AED) score of 1 unless they also contained a PFAM domain. Approximately 95% of loci had an AED score <0.5 (Figure S5), demonstrating a high level of support with the available evidence, and 21,171 (~78%) were annotated with a PFAM domain. Analysis of gene orthologs and paralogs among related Brassicaceae confirmed the close relationship with *E. salsugineum*, with the protein-coding fraction occupying a genome space comparable to related species (Figure 3a). A total of 4433 gene duplication events were recorded with OrthoFinder, comparable to *E. salsugineum* (5108), but fewer than in *B. rapa* (11,513), for example.

The full descriptive statistics are given in Table 2, in comparison with the original T_arvense_v1 annotation (Dorn et al., 2015) lifted over to the new genome with Liftoff v1.5.2 (Shumate and Salzberg, 2020), where applicable. Gene feature distributions are comparable between T_arvense_v1 and the present assembly of MN106-Ref (hereafter referred to as T_arvense_v2; Figure S6). A total of 4433 gene duplication events were recorded with OrthoFinder, comparable to *E. salsugineum* (5108), but fewer than in *B. rapa* (11,513), for example.

The full descriptive statistics are given in Table 2, in comparison with the original T_arvense_v1 annotation (Dorn et al., 2015) lifted over to the new genome with Liftoff v1.5.2 (Shumate and Salzberg, 2020), where applicable. Gene feature distributions are comparable between T_arvense_v1 and the present assembly of MN106-Ref (hereafter referred to as T_arvense_v2; Figure S6). A total of 4433 gene duplication events were recorded with OrthoFinder, comparable to *E. salsugineum* (5108), but fewer than in *B. rapa* (11,513), for example.

**Non-coding loci**

In addition to the protein-coding gene annotations, we annotated non-coding RNA (ncRNA) genes, pseudogenes, and TEs. Descriptive annotation statistics are summarized in Table 2. While many of these annotation features in *T. arvense* were similar to those found in other plant species, we observed several unique patterns, which we will describe in detail below. ncRNA annotations were inferred from sequence motifs (tRNA, rRNA, snoRNA) or from sequencing data (siRNA, miRNA). We predicted clusters of both 5S rRNA and tandem repeat units of 18S and 28S rRNA with RNAmmer (Lagesen et al., 2007), often in relative proximity to loci identified with Tandem Repeats Finder v4.09.1 (Benson, 1999) and putatively associated with centromeric repeat motifs (not shown). Of the largest seven scaffolds, only scaffolds 4 and 7 carried no such annotations. Notably, several large clusters of 5S rRNA genes were interspersed throughout the pericentromeric region of scaffold 1, whereas the remaining four scaffolds contained 18S and 28S rRNA gene annotations. Finally, we identified 243 homologs from 114 snoRNA families.

**sRNA annotation**

We identified 19,386 siRNA loci. More than 98% of these loci corresponded to heterochromatic 23- to 24-nt siRNA loci, with only 196 producing 20- to 22-nt siRNAs. The siRNA loci were expressed unevenly across tissues, as inferred from prediction with data from different tissues. Only 2938 loci were shared across all four tissues studied (rosette leaves, roots, inflorescences and pollen). Inflorescences were the major contributor with 6728 private loci. Despite these differences between tissues, we
observed similar overall patterns in terms of locus length, expression (Figure 3c) and complexity (Figure S7).

Altogether, sRNA loci accounted for \( \sim 8 \) Mbp or \( \sim 1.5\% \) of the assembled genome. Of the seven largest scaffolds, where the majority of genes are located, the total coverage of siRNA loci ranged between 1.5% and 2% and the loci appeared to be preferentially concentrated at the boundary between TEs and the protein-coding gene fraction of the genome. To further explore this, we partitioned the seven largest scaffolds into gene-enriched and gene-depleted regions, based on a median of 14 genes per Mbp and a mean of 54.2 genes per Mbp. We defined gene-enriched loci as those above and gene-depleted loci as those below the mean. At the chromosomal level, sRNA loci correlated with gene-enriched and gene-depleted regions, based on a median of 14 genes per Mbp and a mean of 54.2 genes per Mbp. We defined gene-enriched loci as those above and gene-depleted loci as those below the mean. At the chromosomal level, sRNA loci correlated with gene-enriched regions and were scarce in regions with high TE content. This trend is in contrast to that observed in A. thaliana (Hardcastle et al., 2018) but resembles what has been observed, for example, in maize (He et al., 2013) and tomato (Tomato Genome Consortium, 2012).

Phased secondary sRNAs (phasiRNAs) are a class of secondary sRNAs that, due to the way they are processed, produce a distinct periodical pattern of accumulation (Axtell, 2013b). In the T. arvense genome, we observed 139 loci with such phased patterns. In contrast to the general notion that phasiRNAs are typically 21 nt long (Lunardon et al., 2020), we found 24-nt sRNAs to be dominant in 133 of these loci.

**MicroRNAs**

MicroRNA (miRNA)-encoding genes were predicted using a combination of ShortStack and manual curation (see Methods). We identified 72 miRNA-producing loci, with 53 that were already known from other species, and 19 appeared to be species-specific. Most of the identified families were produced from only one or two loci, with miR156 and miR166 being produced by the most loci, with eight and five family members, respectively. A total of 21 out of 25 families in T. arvense are found in other rosids, and three (miR161, miR157 and miR165) only in other Brassicaceae. One family, miR817, is also present in rice. There is a strong preference for 5’-U at the start of both unique and conserved miRNAs (Figure S8), in line with previous reports (Voinnet, 2009). The expression level of both conserved and novel miRNA families was compared between tissues, showing that the ten most highly expressed across all tissues are conserved families, whereas novel miRNA demonstrates a marginal tendency to be more lowly expressed or with potential for differential expression (Figure S9).

**sRNA loci**

When we overlaid the sRNA loci with our annotated genomic features, most sRNAs localized to the intergenic space, but a substantial fraction, especially 20- to 22-nt sRNAs, were produced...
from intronic sequences (Figure S10a). Helitrons make up only 1.5% of the genome space, yet more than 5% of sRNA biogenesis loci overlap with this type of TE. Most sRNA loci (93.0%) fell within 1.5 kb of annotated genes or TEs (Figure S10b,c). As expected, 23- to 24-nt sRNAs were more frequently associated with TEs, whereas 20- to 22-nt sRNAs were more often produced by coding genes (Axtell, 2013a).

Pseudogenes

In accordance with the MAKER-P protocol, pseudogenes (Ψ) were predicted in intergenic DNA with the ShuLab pseudogene pipeline (Zou et al., 2009). A total of 44 490 set II pseudogenes were annotated, exceeding those in A. thaliana (~3700) or rice (~7900) by one order of magnitude. We identified 35 818 pseudogenes overlapping with TEs, and 8672 pseudogenes that were either concentrated in intergenic space or more towards the protein-coding gene complement of the genome, and thus perhaps less likely to have arisen from retrotransposition. Approximately 59.2% of these contained neither a non-sense nor a frameshift mutation, indicating either (i) that the regulatory sequences of the pseudogenes were silenced first, (ii) a pseudo-exon that may be linked to another non-functional exon, or (iii) a possible undiscovered gene.

Transposable elements

In total, we identified 423 251 TEs belonging to 10 superfamilies and covering ~61% of the genome (Figure 3d). Retrotransposons (75% of all TEs are Gypsy elements; 10% Copia; 4% LINE) by far outnumbered DNA transposons (3% Helitrons; 1% hAT; 2% CACTA; 1% PI/R-Harbinger; 2% MuLE). A detailed breakdown of repeats is shown in Table S4. As the most abundant retrotransposon superfamily, Gypsy elements accounted for 46% of the total genome space, which is consistent with a high abundance observed in the pericentromeric heterochromatin of E. salsugineum, where centromere expansion is thought to have been caused by Gypsy proliferation (Zhang et al., 2020). In addition, we identified 359 protein-coding genes located fully within TE bodies that could represent Pack-TYPE elements and contribute to gene shuffling (Catoni et al., 2018). Among these elements, 153 were intersecting with mutator-like elements, suggesting they correspond to Pack-MULE loci. TEs were located primarily in low gene density regions, while the fraction of TE-contained genes was randomly distributed.

Expression atlas

With cDNA sequences from 11 different tissues or developmental stages, we could annotate tissue-specific expression patterns. The complete expression atlas is provided in Data S1. We evaluated the relative extent of tissue-specific gene expression using the Tau (τ) algorithm (Yanai et al., 2005), from the normalized trimmed mean of M-value (TMM) counts in all tissues (Robinson and Oshlack, 2010). To preclude potential biases caused by substantial differences in library size, we excluded low-coverage samples from mature seeds and old green siliques. In total, 4045 genes had high or even complete tissue specificity (τ = 0.8–1.0), while 5938 genes had intermediate specificity (0.2–0.8) and 6107 had no or low specificity (0–0.2); the remaining genes were ignored due to missing data. The relative breakdown of each specificity fraction by tissue type is shown in Figure 4a, with ‘roots’, ‘green seeds’ and ‘inflorescences’ representing the tissues with the greatest proportion of high or complete specificity genes. The relative log2(TMM) expression values of the top 30 most highly expressed genes in each tissue, given a high or complete specificity score, are plotted in Figure 4b with respect to the overall mean expression per gene across all included tissues. These include, for example, genes with homology to EXTSN2 (EXT2; A. thaliana) in ‘roots’, CRUCIFERIN (BnC1; B. napus) in ‘green seeds’, and PECTINesterase inhibitor 1 (PME1; A. thaliana) in ‘inflorescences’ and ‘open flowers’ (Data S2).

DNA methylation

Cytosine methylation (also commonly referred to as DNA methylation) is a prevalent epigenetic mark in plant genomes and is often associated with heterochromatin and transcriptional inactivation of TEs and promoters, but also with higher and more stable expression when present in gene bodies (Zhang et al., 2018). In plants, DNA methylation occurs in three cytosine contexts, CG, CHG and CHH (where H is any base but G), with the combined presence of CG, CHG and CHH methylation usually indicative of heterochromatin formation and TE silencing, while gene body methylation consists only of CG methylation (Bewick and Schmitz, 2017). In the light of the high TE density in T. arvense, we analysed genome-wide DNA methylation by whole-genome bisulphite sequencing (WGBS) in shoots and roots of 2-week-old seedlings. Genome-wide, 70% of cytosines were methylated in the CG context, 47% in the CHG context and 33% in the CHH context. In line with findings in other Brassicaceae, methylation at CG sites was consistently higher than at CHG and CHH (Figure 1a; Figure S11). When we compared the WGBS data against the genome annotation, high levels of DNA methylation (mostly 5mCG) localized with regions of dispersed repeats and TEs in the centre of the chromosomes. Conversely, methylation was depleted in gene-rich regions (Figure 1a,b). In line with this, DNA methylation was consistently high along TEs, particularly in the CG context (Figure 4c). In contrast to E. salsugineum (Bewick

Table 2  Summary of feature annotations in comparison with the original version T_arvense_v1

| Type                          | T_arvense_v1 | T_arvense_v2 | diff. |
|-------------------------------|--------------|--------------|-------|
| (A) Protein-coding genes      |              |              |       |
| Total number of loci          | 27 390       | 27 128       | -262  |
| Total number of unique loci   | 4780         | 5034         | +254  |
| Total number of transcript isofoms | –         | 30 650       | +30 650 |
| Number of matching loci with changes in CDS | –       | –            | +14 102 |
| Number of matching loci with changes in UTR(s) | –         | –            | +22 559 |
| Loci containing one or more PFAM domain | –     | 21 171       | +21 171 |
| Loci annotated with one or more GO term | –    | 13 074       | +13 074 |
| (B) Non-coding genes          |              |              |       |
| tRNA                          | –            | 1148         | +1148 |
| tRNA clusters (<25 kbp)       | –            | 63           | +63   |
| snoRNA                        | –            | 243          | +243  |
| Small interfering RNA (sRNA)  | –            | 19 373       | +19 373 |
| MicroRNA (miRNA)              | –            | 72           | +72   |
| (C) Other gene types          |              |              |       |
| Pseudogenes (set II %)        | –            | 44 490       | +44 490 |
| Transposable element genes    | –            | 423 251      | +423 251 |
et al., 2016; Niederhuth et al., 2016), DNA methylation dropped only slightly in regions flanking TEs, which might be related to the overall dense TE content in *T. arvense*.

In contrast to TE and promoter methylation, gene body methylation (gbM) is generally associated with medium-to-high gene expression levels (Zhang et al., 2006; Zilberman et al., 2006). gbM occurs in ~30% of protein-coding genes in *A. thaliana*, with DNA methylation increasing towards the 3′-end of the gene (Zhang et al., 2006). The *T. arvense* relative *E. salsugineum* lacks gbM (Bewick et al., 2016; Niederhuth et al., 2016). gbM was also largely absent in *T. arvense* (Figure 4d), suggesting that gbM was lost at the base of this clade.

**Applications towards crop improvement**

**Genetic variation in a pennycress collection**

Knowledge of genetic diversity within wild populations is an essential process for improvement and domestication of new crop species. We analysed a geographically broad sample of forty accessions (Figure S12) using whole-genome resequencing to characterize population structure and variation in germplasm available for breeding. We identified a total of 13 224 528 variants with QD value of ≥2000. Of these, 12 277 823 (92.8%) were SNPs, 426 115 (3.2%) were insertions, and 520 590 (3.9%) were deletions relative to the reference genome. Across all variants, 661 156 (2.9%) were in exons, with 340 132 synonymous, 314 075 nonsynonymous and 6949 non-sense changes. STRUCTURE analysis of both indel and SNP data sets resulted in optimal models of *k* = 3 populations (Figure S13). Both data sets assigned the three lines of Armenian descent, which were highly distinct and had the largest genetic distance to the other accessions, to a single discrete population with limited to no gene flow to the other populations. These results are consistent with previous reports in pennycress (Frels et al., 2019) and were further supported by whole-genome dendrograms (Figure 5a).

We also calculated linkage disequilibrium (LD) among 2 518 379 genome-wide markers and chromosome-specific markers using TASSEL v5.2.75 (Bradbury et al., 2007) with a sliding window of...
40 markers. The r-squared values were plotted against the physical distance with a LOESS curve fitted to the data to show LD decay (Figure S14). Genome-wide, LD decayed to an r-squared value ($r^2$) of 0.2 over 6.2 kbp (Hill and Weir, 1988), which is comparable to LD decay reported in related Brassica species at $r^2 = 0.3$, including B. rapa (2.1 kbp) (Wu et al., 2019) and B. napus (12.1 kbp) (Lu et al., 2019).

**Gene structure variation in pennycress accessions**

The natural variation present in germplasm is an important source of alleles to facilitate breeding efforts and presents an opportunity to understand the evolution of gene families and adaptation within a species. To understand these in a more targeted approach, we sequenced on the PacBio Sequel platform the transcriptomes of two accessions, MN108 and Spring32-10, that are amenable to transformation and gene editing (McGinn et al., 2019), using RNA from leaves, roots, seeds, flowers and siliques. We constructed de novo reference transcriptomes using the Iso-seq3 pipeline, resulting in 25 296 and 26 571 accession-specific isoforms for MN108 and Spring32-10, respectively. These transcriptomes were then polished using the raw reads and processed through the SQANTI3 pipeline (Tardaguila et al., 2018) to characterize the genes and isoforms identified in each of the accessions. We identified 212 of 220 unique genes and 3780 of 3857 unique isoforms for MN108 and Spring32-10 respectively compared with the new reference. Transcripts mapping to the known reference denoted by ‘Full Splice Match’ (FSM) and ‘Incomplete Splice Match’ (ISM) accounted for 28.7% and 30.6% of all transcript models in MN108 and Spring32-10, respectively (Figure 5b, c). Transcripts of the antisense, intergenic and genic intron categories collectively accounted for a total of 12.0% (MN108) and 11.2% (Spring32-10). About ~15% of all identified transcripts were novel isoforms when compared to the reference transcriptome for T. arvense_v2.

**Mapping a pale seedling phenotype**

From a segregating population with a high oleic pennycress (fae-1/rod1-1) background (Chopra et al., 2020b), we identified pale seedling lines (Figure 5d). This phenotype segregated in a Mendelian fashion. To determine the genetic control for this phenotype, we separately pooled genomic DNA from 20 wild-type and 20 pale plants. We processed sequence data obtained from each of these pools through the MutMap pipeline (Sugihara et al., 2019) from each of these pools through the MutMap pipeline (Sugihara et al., 2019) to identify polymorphisms that could cause a pale seedling phenotype in Arabidopsis (Niittylä et al., 2012) identified polymorphisms that might have deleterious effects on function of genes in this region (Table S5). The most obvious candidate is MEX1, which causes a pale seedling phenotype in A. thaliana (Niittylä et al., 2004).

**Discussion**

In this study, we report a high-quality reference genome assembly and annotation for T. arvense (var. MN106-Ref), a newly domesticated oilseed crop for the cooler climates of the world. The improved genome assembly, containing seven chromosome-level scaffolds, revealed two main features: a landscape characterized by a large repetitive fraction populated with TEs and pseudogenic loci in pericentromeric regions, and a gene complement similar in size to other Brassicaceae and densely concentrated towards the telomeres (Figure 1). Previous annotations were enriched with additional gene models for protein-coding loci, and now include non-coding genes for tRNAs, rRNAs, snoRNAs, siRNAs and miRNAs, alongside predicted pseudogenes and TEs (Table 2). These newly improved assembly features will allow for efficient combining of traits and help accelerate future breeding as it would provide knowledge about the gene localization and the linkage of genes of interest. For example, the improved genome assembly has revealed that multiple domestication syndrome genes (ALKENYL HYDROXYALKYL PRODUCING 2-like, TRANSPARENT TESTA 8, EARLY FLOWERING 6) (Figure S1) are located on a single chromosome.

Improved genomic resources can facilitate general understanding of plant biology and evolutionary biology while aiding plant breeding and crop improvement (Scheben et al., 2016). For example, pennycress and Arabidopsis share many key features that made Arabidopsis the most widely studied model plant system (Meinke et al., 1998). The use of Arabidopsis for translational research and for identifying potential gene targets in T. arvense is possible and has been extensively validated (Chopra et al., 2018, 2020a, 2020b; Jarvis et al., 2021; McGinn et al., 2019). Previous studies have suggested that over a thousand unique genes in T. arvense are represented by multiple genes in Arabidopsis and vice versa. Our comparative genomics by way of synteny with E. salsugineum (Yang et al., 2013) revealed a high level of agreement, particularly between the protein-coding fraction of the genome, represented as conserved blocks in the largest seven scaffolds relative to the ancestral karyotype in Brassicaceae (Murat et al., 2015; Figure 2). The detailed description of gene synteny between T. arvense and other Brassicaceae provides insights into the evolutionary relevance of T. arvense within lineage II of Brassicaceae. In addition, the difference in genome size between T. arvense and other species, despite the reduced level of gene duplication and the 1:1 gene relationship, can be explained by the large repetitive fractions present throughout both the centromeric and pericentromeric regions. In the absence of whole-genome duplication events, these repetitive fractions indicate that the increased genome size may be a consequence of active TE expansion. This is therefore suggestive of a mechanism by which deleterious retrotransposon insertions must be mitigated in T. arvense. This could be explained by the high proportion of Gypsy retrotransposons in this species, usually located in heterochromatic regions, or by integration site selection (Sultana et al., 2017), or otherwise by silencing by small RNA activity and/or DNA methylation (Bucher et al., 2012; Sigman and Slotkin, 2016). Given the relatively high error rate of PacBio CLR reads (~10% before correction) with respect to circular consensus sequencing (CCS), the repetitive fraction would also help to explain the initial overestimation of the assembly size as a result of duplicated contigs. We also detected several loci with highly overrepresented read coverage indicative of repeat collapsing during the assembly process, often intersecting with 5S, 18S and 28S rDNA annotations. Such regions are difficult even for current long read technologies due to the large size of the tandem repeat units.

With the availability of improved genomic resources, increasing interest has turned towards understanding tissue-specific gene regulation to reduce pleiotropic effects upon direct targeting of genes during crop improvement. In this study, we have generated a resource using mRNA-seq, sRNA-seq and WGBS to gain insights into genes and their associated regulatory landscape. These data sets help elucidate the extent of tissue specificity and provide useful
Adam Nunn et al.
information for gene modification targets. For example, fatty acid desaturase 2 gene (FAD2; Ta12495 – T.arvense_v1) is involved in the oil biosynthesis pathway and is expressed in many different tissues analysed in this study (Data S1). FAD2 gene knockout should result in higher levels of oleic acid in the seed oil and provide an opportunity for pennycress oil to be used in food applications. It has been observed, however, that knockout mutants in pennycress display delayed growth and reduced seed yields in spring types (Jarvis et al., 2021), and reduced winter survival in the winter types (Chopra et al., 2019), as a purported consequence of its broad expression profile. Similarly, genes such as AOP2-LIKE (Tarvense_05380 – T.arvense_v2) have been targeted to reduce glucosinolates in pennycress seed meal for food and animal feed applications (Chopra et al., 2020b). However, AOP2-LIKE, too, is expressed in many tissues during development, which might explain why knockout plants with reduced glucosinolate content are reportedly more susceptible to insect herbivores such as flea beetles feeding on rosette leaves and root tissues (Marks et al., 2021). Our tissue-specific expression data suggest that, to overcome this challenge, one could alternatively target genes such as Glucosinolate Transporter 1 (GTR1; Tarvense_14683), which is expressed specifically in reproductive tissues (Data S1). This might achieve the desired reductions of seed glucosinolates while avoiding developmental defects. Such approaches have been effectively used in Arabidopsis and many Brassica species (Ander sen and Halkier, 2014; Nour-Eldin et al., 2012).

Finally, the forty resequenced accessions described here provide a rich source of variants that reflect the genetic diversity and population structure of the species in the collection (Figure 5a). Further evaluations of transcriptome sequences showed ample variation in the transcripts from two separate lines – MN108 and Spring32-10 – that are highly amenable to transformation and highlighted the potential for developing pan-genomes in the future. These genomic resources will facilitate genetic mapping studies in pennycress in both natural populations and mutant panels. We have identified genomic regions associated with a pale leaf mutant in pennycress seedlings using a modified BSA-Seq approach in this study (Figure 5d).

Over the last few years, significant efforts have been made towards the discovery of crucial traits and translational research in pennycress, centring on MN106-Ref and the gene space information generated by Dorn et al. (2013) and Dorn et al. (2015). In this study, we continued to generate genomic tools for this accession, with improved contiguity and high-quality annotations to make T. arvense var. MN106-Ref more accessible as a field-based model species for genetics and epigenetics studies and to provide tools for this new and extremely hardy winter annual cash cover crop. However, the assembly of additional accessions can only help to further enrich the resources available for the study of pennycress. In parallel to this study, a Chinese accession of T. arvense (YUN_Tarv_1.0) was assembled using Oxford Nanopore, Illumina HiSeq and Hi-C sequencing (Geng et al., 2021). This timely availability of an additional frame of reference opens the door to a pan-genomic approach in evolutionary research and allows for the better characterization of structural variants moving forward. Furthermore, the use of different sequencing technologies and assembly software provides an additional avenue to correct misassemblies and base calling errors in either case. The overall longer contigs assembled with PacBio CLR, for example, and the consideration of various genetic map data in addition to Hi-C provides a greater resolution of scaffolds particularly throughout the centromere and pericentromeric regions (Figure S15). The reduced error rate of PacBio CCS (used for polishing) is also reflected in the overall k-mer content, which is measured with a two-order magnitude higher consensus quality over scaffolds representing chromosomes and ~99% overall completeness for T_arvense_v2 (Tables S5-S8), indicative of high-quality, error-free sequences more appropriate for variant calling, for instance. Geng et al. (2021) also reported WGS analysis on forty Chinese accessions and reported an LD decay of 150 kbp at an r-squared value (r²) of 0.6, which is considerably higher than the values determined on the forty accessions in this study, as well as those reported for related Brassica species (Lu et al., 2019; Wu et al., 2019). We believe the combination of resources will allow us to investigate the differences that might exist between accessions originating from different geographic locations around the world and help provide further insight into structural variations and evolutionary dynamics.

In conclusion, the T.arvense_v2 assembly offers new insights into the genome structure of this species and of lineage II of Brassicaceae more generally, and it provides new information and resources relevant for comparative genomic studies. The tools presented here provide a solid foundation for future studies in an alternative model species and an emerging crop.

Methods

Seeds for the reference genome development

Seeds from a small natural population of T. arvense L. were collected near Coates, MN by Dr. Wyse, and the accession number MN106 was assigned to this population. We propagated a single plant for ten generations from this population, and we refer to this line as MN106-Ref.

Sample collection, library preparation and DNA sequencing for assembly

PacBio CLR library

Plants were cultivated, sampled and prepared at the Max Planck Institute for Developmental Biology (Tübingen, Germany). Plant seeds were stratified in the dark at 4 °C for 4–6 day prior to planting on soil. Samples were collected from young rosette leaves of T. arvense var. MN106-Ref seedlings, cultivated for 2 weeks under growth chamber conditions of 16–23 °C, 65% relative humidity and a light/dark photoperiod of 16 h:8 h under 110–140 µmol/m²/s light. High molecular weight (HMW) DNA was obtained following nucleus isolation and DNA extraction with the Circulomics Nanobind Plant Nuclei Big DNA Kit according to the protocol described in Workman et al. (2018) and (Workman et al., 2019). A total of 11 extractions from 1.5–2 g frozen leaves each were processed in that way, yielding a pooled sample with a total of 12 µg of DNA by Qubit® 2.0 fluorometer.

Figure 5  (a) Dendrogram representing the forty wild accessions in our study showing three distinct subpopulations, inferred from STRUCTURE analysis (Figure S13). (b,c) Variation of transcript isoforms for MN108 (b) and Spring32-10 (c) accessions based on SQANTI3 analysis. (d) A pale phenotype segregating in an improved pennycress line (fae-1/rod1-1) was analysed with a modified bulked-segregant analysis, and the QTL region associated with this phenotype was mapped using the MutMap approach.

© 2022 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 20, 944–963

Chromosome-level T. arvense genome assembly 953
array (Bionano Genomics, San Diego, CA) and automatically and labelled DNA samples were loaded onto the NanoChannel stained as previously described (Lam et al., 2012). The stained HMW DNA molecules. The nicked DNA molecules were then prepared following the proximo Hi-C plant protocol (Phase Genomics (San Diego, CA)). The Hi-C library was evaluated with BUSCO (Simão et al., 2015). The Hi-C data were aligned with bwa-mem through a 26G needle using a 1-mL syringe, resulting in an 85-kb peak size sample as estimated by FEMTO Pulse Analyzer (Agilent Technologies, Santa Clara, CA). A large insert gDNA library for PacBio Sequel II CLR sequencing was prepared using the SMRTbell® Express Template Preparation Kit 2.0. The library was size-selected for >30 kb using BluePippin with a 0.75% agarose cassette (Sage Science) and loaded into one Sequel II SMRT cell at a 32 pM concentration. This yielded a genome-wide sequencing depth of approximately 476X over ~6.9 million polymerase reads with a subread N50 of ~38 bp.

**PacBio CCS library**

MN106-Ref plants were grown in growth chambers at the University of Minnesota. Individual plants were grown to form large rosettes for isolating DNA. Approximately 25 g of tissue was harvested and submitted to Intac Genomics (Saint Louis, MO) for high molecular weight DNA extraction. This yielded a pooled sample with a total of 269 ng of DNA by Qubit® (Thermo Fisher Scientific, Waltham, MA) estimation, and high DNA purity with a mean absorbance ratio of 1.87 at 260/280 nm and 2.37 at 260/230 nm, as measured by Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA). To further clean up the high molecular weight DNA, we used Salt-Chloroform Wash protocol recommended by PacBio. This yielded a total of 12.1 ng/μL of high-quality DNA for library preparation. A large insert gDNA library was prepared, and 15 kb High Pass Size Selection on Pippin HT was performed at the University of Minnesota Genomics Center (Minneapolis, MN). These libraries were sequenced on 4 SMRT cells using PacBio Sequel II (Pacific Biosciences, Menlo Park).

**Bionano library**

High molecular weight DNA was isolated from young leaves and nicking endonuclease – BspQI was chosen to label high-quality HMW DNA molecules. The nicked DNA molecules were stained as previously described (Lam et al., 2012). The stained and labelled DNA samples were loaded onto the NanoChannel array (Bionano Genomics, San Diego, CA) and automatically imaged by the Irys system (Bionano Genomics, San Diego, CA).

**Hi-C library**

The MN106-Ref plant tissue used for PacBio CCS was submitted to Phase Genomics (San Diego, CA). The Hi-C library was prepared following the proximo Hi-C plant protocol (Phase Genomics, San Diego, CA), and the libraries were sequenced to 116X depth on an Illumina HiSeq 2500 (Illumina, San Diego, CA) instrument with 125-bp paired-end reads.

**Illumina PCR-free library**

Libraries for PCR-free short read sequencing were prepared from MN106-Ref genomic DNA using the TruSeq DNA PCR-Free Low Throughput Library Prep Kit (Illumina, San Diego, CA) in combination with TruSeq DNA Single Indexes Set A (Illumina, San Diego, CA) according to the manufacturer’s protocol. We prepared two libraries, with average insert sizes of 350 bp and 550 bp, respectively. Samples were sequenced to 125X depth (~66 Gb) on an Illumina HiSeq 2500 (Illumina, San Diego, CA) with default options, aside from cluster runtime configuration and the settings corOutCoverage=50, minReadLength=5000, minOverlapLength=4000, correctedErrorRate=0.04 and genomeSize=539 m, which were selected based on the characteristics of the library. Canu performs consensus-based read correction and trimming, resulting in a curated set of reads that were taken forward for assembly (Figure S16).

The resulting assembly overestimated the genome size by approximately 53% (Table S2), which we surmised was likely due to uncorrected sequencing errors in the remaining fraction of reads, in which Canu was able to assemble into independent, duplicated contigs. Analysis of single-copy orthologs from the Eucodiotyledons odb10 database with BUSCO v3.0.2 (Simão et al., 2017) revealed a high completeness of 98.4% and a duplication level of 23.6% (Table S6). Subsequent alignment of the reads to the assembly using minimap2 v2.17 (Li, 2018) and purge_dups v1.0.1 (Guan et al., 2020) presented bimodal peaks in the read depth distribution, indicative of a large duplicated fraction within the assembly (Figure S17). As efforts to collapse this duplicated fraction using assembly parameters were unsuccessful, and purge_dups is intended to correct duplication arising from heterozygosity (which does not apply in T. arvense), the fraction was reduced by manual curation instead. Contigs starting from the left-hand side of the read depth distribution were consecutively removed until reaching an approximation of the estimated genome size, with any contigs containing non-duplicated predicted BUSCO genes kept preferentially in favour of discarding the next contig with lower read depth in the series.

The deduplicated assembly from Canu was polished with the PacBio Sequel II HiFi CCS reads using two iterations of RACON v1.4.3 (Vaser et al., 2017), prior to repeat reassembly. Bionano maps were used to build de novo scaffolds using the polished assembly; hybrid scaffolds were generated using the de novo Bionano maps and the assembly (https://bionanogenomics.com/support-page/data-analysis-documentation/). To further resolve repetitive regions and improve assembly contiguity, the bionano-scaffolded assembly was integrated into the HERA pipeline (Du and Liang, 2019). The Hi-C data were aligned with bwa-mem v0.7.17 (Li and Durbin, 2009), PCR duplicates were marked with picard tools v1.83 (http://broadinstitute.github.io/picard), and the quality was assessed with the hic_qc.py tool of Phase Genomics (https://github.com/phasegenomics/hic_qc). The assembly was then scaffolded with the Hi-C alignments using SALSA v2.2 (Ghurye et al., 2017) and subsequently polishing with the PCR-free Illumina data using two iterations of PILON v1.23 (Walker et al., 2014). The final assembly was the result of a meta-assembly with quick-merge v0.3 (Chakraborty et al., 2016), which combined the current assembly with an earlier draft version assembled using Canu 1.8 (Koren et al., 2017) directly from the PacBio CCS reads and polished only with the Illumina PCR-free short-reads, following an almost identical workflow, in order to help address the possibility of misassembly arising from technical sources and improve overall contiguity. This resulting assembly was evaluated with BUSCO (Simão et al., 2015) and QUAST

© 2022 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 20, 944-963
v5.0.2 (Gurevich et al., 2013). Intermediate assembly statistics are given in comparison with (i) immediately after Canu, and (ii) the final version after rescaffolding (Table S2).

**Genome size estimation using flow cytometry and k-mer-based approach**

The nuclei of field pennycress line MN106-Ref, Arabidopsis thaliana, maize and tomato were stained with propidium iodide, and fluorescent signals were captured using a Becton-Dickinson FACSCanto flow cytometer (https://www.bd biosciences.com/). DNA content for all four species that corresponded to G0/1 nuclei is listed in Table S1. The genome size of Arabidopsis is 135 Mb, and therefore, the genome size of pennycress was calculated to be 501 ± 33 Mb. Using the Illumina HiSeq2500 platform, we obtained ~100× PCR-free reads, which were used for subsequent K-mer analysis using Jellyfish (Marçais and Kingsford, 2011). The 101-mer frequency distribution curve exhibited a peak at 22 k-mer, and analysis showed that the total number of K-mers was 11 403 836 319. Using the formula of genome size estimation, we found the pennycress genome ranged from 459 to 540 Mb.

**Development of genetic maps for rescaffolding**

To improve the contiguity and correct misassemblies, we developed two genetic linkage maps using F2 populations. The first linkage map was derived from a cross between a wild Minnesota accession ‘MN106-Ref’ and a genetically distant Armenian accession ‘Ames32867’. The resulting F1 plants were allowed to self-fertilize, and seeds from a single plant were collected and propagated to the F2 generation. Approximately 500 mg fresh tissue was collected from 94 individuals in the F2 population. The tissue was desiccated using silica beads and pulverized using a TissueLyser. DNA was isolated with the BioSprint DNA Plant Kit (Qiagen, Valencia, CA). The F2 population along with the two parental genotypes was genotyped with genotyping by sequencing at the University of Minnesota Genomics Center (Minneapolis, MN). Each sample was digested with the BtgI_BglII restriction enzyme combination, barcoded and sequenced on the Illumina NovaSeq S1 (single-end 101 bp) yielding 1 237 890 mean reads per sample. The raw reads were demultiplexed based on the barcode information and aligned to the most recent iteration of the assembled genome. Further investigation by way of synteny comparison with a closely related species, *Eutrema salsugineum* (Yang et al., 2013), revealed that several of these likely occurred during scaffolding as orientation errors. Some of these errors could also be supported in comparison with the recent assembly of a Chinese accession (YUN_Tarv1.0) of *T. arvense*. Consequently, we manually introduced breakpoints at selected loci in the assembled genome where they were supported by at least two sources of data from whole-genome alignments to YUN_Tarv1.0, synteny maps to *E. salsugineum* (derived from reciprocal best blast), genetic linkage maps (wild-derived and EMS mutation based) and Hi-C contact maps. These were cross-examined with minimap2 alignments of PacBio CLR reads to the genome, an overview of corresponding gene distributions produced by Liftoff v1.5.2 (Shumate and Salzberg, 2020) and the resulting synteny analysis to *E. salsugineum*. The resulting contigs were then rescaffolded with ALLMAPS v1.1.5 (Tang et al., 2015) to produce the final assembly, integrating both the synteny map and genetic map data and manually discounting contigs that were supported by only single markers. The final assembly statistics in comparison with previous intermediary stages are given in Table S2.

**Comparative genomics**

**Genome sequences**

*Arabidopsis thaliana* (Araport 11), *Schrenkiiella parvula* (v2.2) and *Arabidopsis lyrata* (v2.1) genome sequences and gene annotation were downloaded from Phytome (Goodstein et al., 2012). The *Eutrema salsugineum* gene annotation was obtained from Phytome and lifted over the assembly GenBank GCA_000325905.2.

**Genome alignments and synteny analysis**

The genome alignments between the different versions of the *T. arvense* assembly to *E. salsugineum* were done using MUMmer v4.0.0 (Marçais et al., 2018) with a minimal length of 200 nt and followed by filtering for 1:1 matches and removing alignments smaller than 1000 bp. To identify the interspecies gene orthologs and syntenic relationships between *T. arvense* and other species,
we used MCScan in the JCVI utility library (https://github.com/tanghaibao/jcvi; Tang et al., 2008). The ortholog relationships were obtained using the proteinic translation of the CDS and using the argument \( --cScore=0.99 \). To define the syntenic blocks and the corresponding genomic coordinates, we used the parameters \( --minSpan=15 \) and \( --minSize=5 \). The genomic coordinates from the syntenic blocks were parsed to draw the syntenic relationships using Circos v0.69-8 (Krzywinski et al., 2009).

To determine the different ancestral Brassicaceae chromosomal blocks (ABKs), we took the ortholog relationship between each gene in \( T. arvense \) and \( A. thaliana \) from the synteny analysis, and compared it with a gene list derived from Murat et al. (2015) where each ortholog gene of \( A. thaliana \) had an assigned ABK block (Murat et al., 2015).

**Genome annotation**

**Tissue preparation for RNA sequencing**

Thlaspi arvense var. MN106-Ref seeds were surface-sterilized with chlorine gas for 1 h and stratified for 3 days at 4 °C. For seedling-stage RNA extractions, seeds were plated on 1/2 MS medium supplemented with 1% plant agar and stratified for 3 days at 4 °C. For all other tissue collections, plants were sown on soil and grown in a climate-controlled growth chamber in long-day conditions (16/8-h light/dark at 21/16 °C, light intensity 140 \( \mu \text{E/m}^2\text{s} \), with 60% relative humidity); plants were watered twice per week. Two weeks after germination, plants growing on soil were vernalized at 4 °C in the dark for 4 weeks, then moved back to the growth chamber. Samples were collected from 11 different tissues in three biological replicates (two in case of mature seeds); for each replicate, we pooled tissue from two individuals. Tissues included the following: one-week-old shoots (from plate culture), one-week-old roots (from plate culture), rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques (about 0.5 × 0.5 cm), older green siliques (about 1 × 1 cm), seed pods, green seeds and mature seeds.

**RNA extraction and sequencing**

Total mRNA was extracted using the RNeasy Plant Kit (Qiagen, Valencia, CA) and treated with DNase I using the DNA-free Kit DNase Treatment and Removal Reagents (Ambion by Life Technologies, Carlsbad, CA), following the manufacturer’s protocols. cDNA libraries were constructed using the NEBNext Ultra II Directional RNA Library Prep Kit (New England BioLabs, Ipswich, MA, USA Inc.) for Illumina following the manufacturer’s protocol. Libraries were sequenced on a HiSeq 2500 instrument (Illumina, San Diego, CA) as 125-bp paired-end reads.

**Transcriptome assembly**

Following quality control and adapter clipping with cutadapt (Martin, 2011), biological replicates for each of eleven tissue types from Illumina mRNA-seq libraries were aligned independently using STAR v2.5.3a (Dobin et al., 2013), then merged according to tissue type, prior to assembly by a reference-based approach. Each assembly was performed using Ryuto v1.3m (Gatter and Stadler, 2019), and consensus reconstruction was then performed using TACO v0.7.3 (Niknafs et al., 2017) to merge tissue-specific transcriptome assemblies. PacBio Iso-seq libraries from MN106-Ref were refined, clustered and polished following the Iso-seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq), prior to alignment with STARlong and isoform collapsing using the cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake) suite. The Iso-seq data were later leveraged together with the Illumina mRNA-seq data to prioritize convergent isoforms using custom in-house scripting.

**Genome annotation**

The final assembly was annotated using the MAKER-P v2.31.10 (Campbell et al., 2014, 2015) pipeline on the servers provided by the Epiderivate project, at ecSeq Bioinformatics GmbH (Leipzig, Germany). Protein sequences were obtained from the Viridiplantae fraction of UniProtKB/Swiss-Prot and combined with RefSeq sequences derived from selected Brassicaceae: Arabidopsis thaliana, Brassica napus, Brassica rapa, Camelina sativa and Raphanus sativus. TEs were obtained from RepetDB (Amselem et al., 2019) for selected plant species: Arabidopsis lyrata, Arabidopsis thaliana, Arabis alpina, Brassica rapa, Capsella rubella and Schrenkia parvula (Extrema parvulum). Repeat library construction was carried out using RepeatModeler v1.0.11 (Smit and Hubley, 2008) following basic recommendations from MAKER-P (Campbell et al., 2014). Putative gene fragments were filtered out following BLAST search to the combined Swiss-Prot + RefSeq protein database after exclusion of hits from RepetDB. The de novo library was combined with a manually curated library of plant sequences derived from repbase (Bao et al., 2015). Genome masking is performed with RepeatMasker v4.0.9 (Smit, 2004) as part of the MAKER-P pipeline. Protein-coding genes, non-coding RNAs and pseudogenes were annotated with the MAKER-P pipeline following two iterative rounds under default settings, using (i) transcript isoforms from Illumina mRNA-seq and PacBio Iso-seq data, (ii) protein homology evidence from the custom Swiss-Prot + RefSeq protein database and (iii) the repeat library and TE sequences for masking. The initial results were used to train gene models for ab initio predictors SNAP v2006-07-28 (Korf, 2004) and Augustus v3.3.3 (Stanke et al., 2006), which were fed back into the pipeline for the subsequent rounds. The final set of annotations was filtered based on Annotation Edit Distance (AED) < 1 except in cases with corresponding PFAM domains, as derived from InterProScan v5.45-80.0 (Jones et al., 2014). The tRNA annotation was performed with tRNAscan-SE v1.31 (Lowe and Eddy, 1997) and the rRNA annotation with RNAmer v1.2 (Lagesen et al., 2007). The snoRNA homologs were derived using Infernal v1.1.4 (Navrocki and Eddy, 2013) from plant snoRNA families described in Patra Bhattacharya et al. (2016). A small phylogeny based on gene orthologs and duplication events in comparison with selected Brassicaceae (A. lyrata, A. thaliana, B. rapa, S. parvula and E. salsugineum) was performed with OrthoFinder v2.5.2, and the resulting species tree is rooted using STRIDE (Emms and Kelly, 2017) and inferred from all genes using STAG (Emms and Kelly, 2018).

**Transposible element annotation**

Two de novo annotation tools, EDTA v1.7.0 (Ou et al., 2019) and RepeatModeler v2.0 (Flynn et al., 2020), were used to annotate TEs independently. For EDTA, the following parameters were used in addition to defaults: --species others, --step all, --sensitive 1, --anno 1, and --evaluate 1. For RepeatModeler2, the additional parameters were -engine ncbi and -LTRStruct. The outputs of both tools were evaluated by manual curation. First, we used tblastn to align each TE consensus with the transposase database obtained from repbase, and the retrotransposon domains (GAG,
Pol, Env, etc.) were viewed one by one with dotter (Sonnhammer and Durbin, 1995). Sequences with multiple paralogs were mapped back to the genome and manually extended to determine the full-length boundary of each TE. A total of 107 full-length, representative Copia and Gypsy families were successfully evaluated. The TE consensus from RepeatModeler2 was selected as the most accurate model based on full-length paralogs. RepeatMasker was then used to construct the GFF3-like file from the FASTA file from RepeatModeler2, with the optional settings: 
```
-e ncbi -q -no_is -norma -nolow -div 40 -cutoff 225
```
The perl script rmOutToGFF3.pl was used to generate the final GFF3 file.

sRNA plant material

Seeds were sterilized by overnight incubation at −80 °C, followed by 4 h of bleach treatment at room temperature (seeds in open 2 mL tube in a desiccator containing a beaker with 40 mL chlorine-based bleach (−5% DanKlorix, Colgate-Palmolive, New York, NY) and 1 mL HCl (32%; Carl Roth, Karlsruhe, Germany)). For rosette, inflorescence and pollen, seeds were stratified in the dark at 4 °C for six days prior to planting on soil, then cultivated under growth chamber conditions of 16–23 °C, 65% relative humidity and a light/dark photoperiod of 16 h:8 h under 110–140 μmol/m²/s light. Rosette leaves were harvested after two weeks of growth. For inflorescence and pollen, 6-week-old plants were vernalized for 4 weeks at 4 °C:8 h under 110–140 μmol/m²/s light. Rosette leaves were harvested after two weeks of growth. For inflorescence and pollen, 6-week-old plants were vernalized for 4 weeks at 4 °C for a light/dark photoperiod of 12 h:12 h under 110–140 μmol/m²/s light. Two weeks after bolting, inflorescence and pollen were collected. Pollen grains were collected by vortexing open flowers in 18% sucrose for 5 min following by centrifugation at 3000g for 3 min in a swinging bucket rotor. For root samples, seeds were stratified for 6 days at 4 °C in the dark on ½ MS media. Plants were grown in 3–4 mL ½ MS medium plates in long day (16 h) at 16 °C. Root samples were collected 12–14 days after stratification.

sRNA extraction and library preparation

Total RNA was extracted by freezing collected samples with liquid nitrogen and grinding with a mortar and pestle with TRIzol reagent (Life Technologies, Carlsbad, CA). Then, total RNA (1 μg) was treated with DNase I (Thermo Fisher Scientific, Waltham, MA) and used for library preparation. Small RNA libraries were prepared as indicated by the TruSeq Small RNA Library Prep Kit (Illumina, San Diego, CA), using 1 μg of total RNA as input, as described by the TruSeq RNA sample prep V2 guide (Illumina, San Diego, CA). Size selection was performed using the BluePippin System (SAGE Science, Massachusetts). Single-end sequencing was performed on a HiSeq 3000 instrument (Illumina, San Diego, CA).

sRNA locus annotation

Raw FASTQ files were processed to remove the 3’-adapter and quality-controlled with trim_galore v0.6.6 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using `trim_galore -q 30 --small_rna`. Read quality was checked with FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reference annotation of sRNA loci was created following the steps indicated by Lunardon et al. (2020). In short, each library was aligned to the reference genome independently using ShortStack v3.8.5 (Axtell, 2013b), with default parameters, to identify clusters of sRNAs de novo with a minimum expression threshold of 2 reads per million (RPM). sRNA clusters from all libraries of the same tissue were intersected using BEDTools v2.26.0 multiIntersectBed (Quinlan and Hall, 2010) with default parameters, and only those loci present in at least three libraries were retained. For each tissue, sRNA clusters 25 nt apart were padded together with the bedtools merge -d option. sRNA loci whose expression was <0.5 RPM in all libraries of each tissue were also removed. Finally, sRNA loci for all different tissues were merged in a single file retaining tissue of origin information with bedtools merge -o distinct options. miRNAs predicted by the ShortStack tool were manually curated (Appendix S1) following the criteria of Axtell (2013b): maximum hairpin length of 300 nt; ≥75% of reads mapping to the hairpin must belong to the miRNA/miRNA* duplex; for the miRNA/miRNA* duplex, no internal loops allowed, two-nucleotide 3’ overhangs, maximum five mismatched bases and only three of which are nucleotides in asymmetric bulges; and mature miRNA sequence should be between 20 and 24 nt.

Expression atlas

Gene expression was measured from the same tissue-specific STAR alignments taken prior to merging biological replicates for transcript assembly, excluding coverage outliers ‘mature seed’ and ‘green old silique’. A total of 27 samples from 9 tissues were therefore considered for gene expression analysis. Raw counts were generated using subread featureCounts v2.0.1 (Liao et al., 2014) and subsequently normalized using the trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010) derived from edgeR v3.34 (Robinson et al., 2010). Averaged expression counts by group were taken for tissue specificity evaluation using the Tau (τ) algorithm (Yanai et al., 2005), as implemented in the R package tspec v0.99.0 (https://drdr.io/github/roonyongsalbi/tspec/), which provides a measure of τ in the range of 0 - 1, where 0 is non/low specificity, and 1 indicates high/absolute specificity.

DNA methylation

We extracted genomic DNA from roots and shoots of 2-week-old seedlings grown on ½ MS medium with 0.8% agar and 0.1% DMSO. Seedlings were grown vertically in 16-h/8-h light/dark cycle; at the time of sampling, roots were separated from shoot tissue with a razor blade and the plant tissue was flash-frozen in liquid nitrogen. Genomic DNA was extracted from ground tissue using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Libraries for WGBS were prepared using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs). Adapter-ligated DNA was treated with sodium bisulphite using the EpiTect Plus Bisulfite Kit (Qiagen, Hilden, Germany) and amplified using the Kapa HiFi Uracil + ReadyMix (Roche, Basel, Switzerland) in 10 PCR cycles. WGBS libraries were sequenced on an Illumina HiSeq2500 instrument with 125-bp paired-end reads.

The WGBS libraries were processed using the nf-core/methylseq v1.5 pipeline (10.5281/zenodo.2555454) combining bwa-meth v0.2.2 (Pedersen et al., 2014) as an aligner and MethylDackel v0.5.0 (https://github.com/dpryan79/MethylDackel) for the methylation calling. The default parameters were used for the entire workflow with the exception of the methylation calling where the following arguments were used: --D 1000 --maxVariantFrac 0.4 --minOppositeDepth 5 --CHG --CHH --nOT 3,3,3,3 --nOB 3,3,3,3 --d 3. Only cytosines with a minimum coverage of 3x were kept for the subsequent analysis.
Further comparisons between the methylated cytosines and the genome annotation were performed using BEDtools v2.27.1 (Quinlan and Hall, 2010).

Population genomics

DNA from forty pennycress accessions was extracted from approximately 500 mg of leaf tissue pooled from five plants using a plant genomic DNA kit (Epoch Life Science). DNA was then subjected to whole-genome sequencing on an Illumina Novaseq sequencer (2 × 125 bp). Raw reads were then aligned to the new reference genome (T_arvense_v2) using bwa-mem (Li and Durbin, 2009). The aligned files were processed with Samtools and Picard tools, and variants were called using GATK HaplotypeCaller v3.3.0 (Ren et al., 2018). Variants were annotated using Snpeff 5.0e (Cingolani et al., 2012). Data sets for both Indel and SNP panels were trimmed based on LD prior to population genomic analysis using PLINK v1.9 (Purcell et al., 2007) with the parameter --indep-pairwise 1000 5 0.5. Population structure for both SNP and indel data was then characterized using the admixture model and independent allele frequencies in STRUCTURE v2.3.4 (Pritchard et al., 2000). Dendrograms of both SNP and Indel data were generated under the UPGMA method using the R package poppr (Kamvar et al., 2014).

The forty accessions were planted in a three replication, randomized complete block design in a greenhouse maintained at 21/20 °C and 16 hour days. Ten seeds per replicate were planted in 13.3-cm² pots in Sungrow propagation potting mix. Seedlings were thinned to one plant per pot after emergence. Winter annual accessions require vernalization to induce flowering, so all winter accessions were placed in a growth chamber maintained at 4 °C with 16-h light for a period of 21 days about 4 weeks after emergence. Spring annual accessions were planted approximately five weeks after winter accessions. Data for days to flowering were collected on 34 accessions that germinated as the number of days that elapsed from the date of emergence to the appearance of the first flower. The vernalization requirement for winter accesses explains the large differences in mean number of days to flowering between spring and winter accessions. Additional phenotypes and data associated with these sequenced accessions are available in Data S6.

Structural variants using Iso-seq data

Single-molecule real-time (SMRT) isoform sequencing (Iso-seq) based on PacBio (PacBio Biosciences, Menlo Park, CA) generated reads was used to investigate unambiguous full-length isoforms for two pennycress wild accessions, MN108 and Spring32-10. Total RNA extraction was performed on the green seed, hypocotyl, seedling root and flower tissues from pennycress plants grown in a climate-controlled growth chamber maintained 21/20 °C during 16-h:8-h day–night setting. Approximately 250 ng of total RNA was obtained and subjected to the Iso-seq Express Library Workflow (PacBio Biosciences, Menlo Park, CA). cDNA is synthesized from full-length mRNA with the NEBNext Single Cell/ Low Input RNA Prep Kit (New England Biolabs, Ipswich, MA) followed by PCR amplification. The amplified cDNA is converted into SMRTbell templates using the PacBio SMRTbell Express Template Prep Kit 2.0 for sequencing on the Sequel System. Sequencing was performed at the University of Minnesota Genomics Center Facility (Minneapolis, MN).

The polished high-quality FASTA file obtained from Iso-seq3 was aligned to pennycress version 2 (T_arvense_v2) with minimap2 (Li, 2018). The resulting SAM file was sorted and collapsed using the cDNA_Cupcake package to obtain an input GFF file such that each transcript has exactly one alignment and at most one ORF prediction. Sgaant v3.3.0-qc, py, part of the SQANTI3 package (Tardagula et al., 2018), was deployed on the resulting GFF file along with the reference genome in the FASTA format and a GTF annotation file. This returned a reference corrected transcriptome, transcript-level and junction-level files with structural and quality descriptors, and a QC graphical report. Among the splice junction sites, SQANTI3 defines canonical junctions such as AT–AC, GC–AG and GT–AG, whereas all others are classified as non-canonical splice junctions.

Linkage disequilibrium analysis

Linkage disequilibrium (LD) among genome-wide markers and chromosome-specific markers was calculated with TASSEL v5.2.75 (Bradbury et al., 2007) with a sliding window size of 40 markers with 100 734 460 total comparisons. The r-squared values obtained via the linkage disequilibrium function in TASSEL were plotted against the physical distance with a LOESS curve fitted to the data to show LD decay (Figure S14).

Bulked-segregation sequencing and MutMap analysis

Bulked-segregant analysis (BSA) (Michelmore et al., 1991) coupled with whole-genome sequencing (BSA-Seq) was performed to locate genomic region harbouring the gene responsible for the pale mutant phenotype in pennycress (Figure 5d). Two pools were created with one pool containing leaf tissue from 20 individual pale mutants and the other pool consisting of wild-type individuals that did not exhibit the pale phenotype. DNA was extracted from fresh pennycress leaves using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA). Both pools were sequenced on an Illumina HiSeq 2000 instrument using 2 × 125 base-paired reads at the University of Minnesota Genomics Center (Minneapolis, MN). The reads were analysed using the MutMap pipeline (Sugihara et al., 2020), and the QTL region was surveyed for candidate genes.

Comparison with YUN_Tarv_1.0

Synteny between T_arvense_v2 and YUN_Tarv_1.0 was assessed with minimap2 alignments and the resulting dotplot generated with the R package dotPlotly (https://github.com/tvooroten/dotPlotly). The k-mer analysis of quality and completeness was carried out for each assembly with Merqury v1.3 (Rhie et al., 2020; Table S7, S8 and S9), using both the PCR-free Illumina HiSeq reads generated in this study and those obtained from Geng et al. (2021) under the accession SRR14757813 in the NCBI Sequence Read Archive.

Acknowledgements

We thank Win Phippen, Thomas Gatter, Prabin Bajgain, Korbinian Schneeberger, Raül Wijffjes, MPI DB Genome Center, Vienna Biocenter Core Facilities (VBCF), University of Minnesota Genomics Center (UMGC), Minnesota Supercomputing Institute (MSI), and all EpiDiverse network members and beneficiaries. We acknowledge the hard work of many who contributed to this study including Brett Heim, Krishan Rai, Nicole Folstad, Matthew A. Ott, Shweta Jain and many others.
Funding

This material is based upon work that is supported by the Minnesota Department of Agriculture (J.A., K.F., R.C.) and by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award numbers 2018-67009-27374 (J.A., R.C., K.F.), and 2019-67009-29004 (M.D.M., J.S.) and the Agriculture and Food Research Initiative Competitive Grant No. 2019-69012-29851 (M.D.M., R.C., J.S.). This research was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Science Program grant no. DE-SC00211286 (M.D.M., R.C.). This work was further funded by the Austrian Academy of Sciences (C.B., I.R.A., K.J., D.R.C.); the Max Planck Society (D.W., A.C.G., P.C.B., C.L.); the European Union’s Horizon 2020 research and innovation programme by the European Research Council (ERC), Grant Agreement No. 716823 ‘FEAR-SAP’ (I.R.A., C.B.), by the Marie Skłodowska-Curie ETN ‘EpiDiverse’, Grant Agreement No. 764965 (D.R.C., C.B.) and by Marie Sklodowska-Curie, Grant Agreement MSCA-IF No 797460 (P.C.B.); and the German Federal Ministry of Education and Research BMBF, Grant No. 031A538A, de.NBI-RBC (A.N., P.F.S.).

Conflicts of interest

The authors declare potential competing interests as intellectual property applications have been submitted on some of the genes discussed in this study.

Author contributions

RC, AN, KF and CB conceived the study. RC and AN led the genome assembly and evaluation, assisted by IRA and PCB. IRA performed the comparative genomics analysis of synteny during genome rescaffolding and in the final evaluation. AN led the genome annotation and performed analysis for protein-coding genes, non-coding genes (tRNA, rRNA, snRNA) and pseudogenes. ACG performed small RNA library sequencing, annotation and analysis, supervised by DW. PZ and ACG performed the transposable element annotation, supervised by DW and MM. AN performed the gene expression analysis and evaluation of tissue specificity. CB and KJ provided PCR-free libraries. RC performed k-mer analysis for genome estimation. KC and RC provided the CCS libraries, which were prepared by the UMGC. CB and IRA provided the DNA methylation libraries and analysis. RC, ZT, MDM and KF developed linkage mapping populations, designed primers, performed genotyping and built genetic maps. KC, RC and KD generated resources for Hi-C, Bionano and resequencing of accessions. KF and ZT phenotyped resequenced accessions. RC performed SNP analysis of resequenced datasets. ZT performed the linkage disequilibrium decay analysis. AB performed population genomics. RC and BJ prepared samples for Iso-Seq libraries. RC and ZT performed gene structure variation analysis. RC and MDM performed bulk-segregant analysis. The PacBio CLR library was prepared and sequenced by PCB and AN under the guidance of CL. DR prepared and sequenced mRNA-seq libraries. RC, AN, CB, ACG, IRA and ZT wrote the manuscript. All authors reviewed and approved the manuscript.

Data availability statement

The assembly and all NGS-based raw data are deposited in the ENA Sequence Read Archive Repository (www.ebi.ac.uk/ena/) under study accession number PRJEB46635. The summary of data provided by each institute and corresponding application is described in Table S10.

References

Amselem, J., Cornut, G., Choixne, N., Alaux, M., Alfama-Depauw, F., Jamilou, V. and Maumus, F. (2019) RepeatDB: a unified resource for transposable element references. Mob. DNA, 10, 6.
Andersen, T.G. and Halkier, B.A. (2014) Upon bolting the GTR1 and GTR2 transporters mediate transport of glucosinolates to the inflorescence rather than roots. Plant Signal. Behav., 9, e27740.
Axtell, M.J. (2013a) Classification and comparison of small RNAs from plants. Annu. Rev. Plant Biol. 64, 137–159.
Axtell, M.J. (2013b) ShortStack: comprehensive annotation and quantification of small RNA genes. RNA, 19, 740–751.
Baq, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA, 6, 11.
Beilstein, M.A., Nagalvingum, N.S., Clemens, M.D., Manchester, S.R. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA, 107, 18724–18728.
Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.
Benic, A., Mabry, M.E., Harkess, A.E., Brose, J., Schranz, M.E., Conant, G.C., Edger, P.P. et al. (2021) Comparative phylogenetics of repetitive elements in a diverse group of flowering plants (Brassicales). G3 Genes|genomes|genetics, 11(7), https://doi.org/10.1093/g3journal/jkab140.
Bewick, A.J., Li, L., Niederhuth, C.E., Willing, E.-M., Hofmeister, B.T., Shi, X., Wang, L. et al. (2016) On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. USA, 113, 9111–9116.
Bewick, A.J. and Schmitz, R.J. (2017) Gene body DNA methylation in plants. Curr. Opin. Plant Biol. 36, 103–110.
Boateng, A.A., Mullen, C.A. and Goldberg, N.M. (2010) Producing stable pyrolisis liquids from the oil-seed Fresciskaces of mustard family plants: pennycress (Thlaspi arvense L.) and Camelina (Camelina sativa). Energy Fuels, 24, 6624–6632.
Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot. Methods Mol. Biol. 406, 89–112.
Bradbury, P.J., Zhang, Z., Kroon, D.E., Castellanos, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSLE: software for association mapping of complex traits in diverse samples. Bioinformatics, 23, 2633–2635.
Bucher, E., Reinders, J. and Mirouzé, M. (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. Curr. Opin. Plant Biol. 15, 503–510.
Campbell, M.S., Holt, C., Moore, B. and Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinform. 48, 4.11.1–39.
Campbell, M.S., McVean, G., Carson, H., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Ikai, L. et al. (2014) MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164(2), 513–524. https://doi.org/10.1104/pp.113.230144.
Catoni, M., Jonesman, T., Cerruti, E. and Paszkowski, J. (2018) Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. Nucleic Acids Res. 47, 1311–1320.
Chakraborty, M., Baldwin-Brown, J.G., Long, A.D. and Emerson, J.J. (2016) Contiguous and accurate de novo assembly of metazoan genomes with modest read coverage. Nucleic Acids Res. 44, e147.
Chopra, R., Folstad, N., Lyons, J. and Umusalov, T. (2019) The adaptable use of Brassica NIRS calibration equations to identify pennycress variants to facilitate the rapid domestication of a new winter oilseed crop. Ind. Crops Prod. 128, 55–61.
Chopra, R., Folstad, N. and Marks, M.D. (2020a) Combined genotype and fatty-acid analysis of single small field pennycress (Thlaspi arvense) seeds increases the throughput for functional genomics and mutant line selection. Ind. Crops Prod. 156, 112823.
Chopra, R., Johnson, E.B., Daniels, E., McGinn, M., Dorn, K.M., Esfahaniyan, M., Folstad, N. et al. (2018) Translational genomes using Arabidopsis as a model enables the characterization of pennycress genes through forward and

© 2022 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 20, 944–963
Chromosome-level *T. arvense* genome assembly

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Marks, M.D., Chopra, R. and Sedbrook, J.C. (2021) Technologies enabling rapid crop improvements for sustainable agriculture: example pennycress (*Thlaspi arvense* L.). *Emerg. Top Life Sci.* **5**, 325–335.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

McGinn, M., Phippen, W.B., Chopra, R., Bansal, S., Jarvis, B.A., Phippen, M.E., Dorn, K.M. et al. (2019) Molecular tools enabling pennycress (*Thlaspi arvense*) as a model plant and oilseed cover crop. *Plant Biotechnol. J.* **17**(4), 776–788. https://doi.org/10.1111/pbi.13014

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koonneef, M. (1998) *Arabidopsis thaliana*: a model plant for genome analysis. *Science*, **282**, 662–679.

Michalovova, M., Vyskot, B. and Kejnovsky, E. (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity*, **111**, 314–320.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA*, **88**, 9828–9832.

Moore, S.A., Wells, M.S., Gesch, R.W., Becker, R.L., Rosen, C.J. and Wilson, M.L. (2020) Pennycress as a cash cover-crop: improving the sustainability of sweet corn production systems. *Agronomy*, **10**, 614.

Moser, B.R. (2012) Biodiesel from alternative oilseed feedstocks: camelina and field pennycress. *Biofuels*, **3**, 193–209.

Moser, B.R., Knotte, G., Vaughn, S.F. and Isbell, T.A. (2009) Production and evaluation of biodiesel from field pennycress (*Thlaspi arvense L*). *Oil Energy Fuels*, **23**, 4149–4155.

Mulligan, G.A. (1957) Chromosome numbers of Canadian weeds. I. *Can. J. Bot.* **35**, 779–789.

Mulligan, G.A. and Kean, P.G. (1973) Color, brightness, and other floral characteristics affecting insects to the blossoms of some Canadian weeds. *Can. J. Bot.*, **51**, 1939–1952.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R. et al. (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**(1), https://doi.org/10.1186/s13059-015-0814-y

Navr rock, E.P. and Edsby, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Niederhuth, C.E., Bewick, A.J., Li, J., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A. et al. (2016) Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.*, **17**, 194.

Nittäyä, T., Messeri, G., Trevisan, M., Chen, J., Smith, A.M. and Zeeman, S.C. (2004) A previously unknown maltose transporter essential for starch degradation in leaves. *Science*, **303**, 87–89.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M. and Iyer, M.K. (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods*, **14**, 68–70.

Nour-Eldin, H.H., Andersen, T.G., Burow, M., Madsen, S.R., Jorgensen, M.E., Olsen, C.E. and Dreyer, I. (2012) NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature*, **488**, 531–534.

Ott, M.A., Eberle, C.A., Thom, M.D., Archer, D.W., Forcella, F., Gesch, R.W. and Wyse, D.L. (2019) Economics and agronomics of relay-cropping pennycress and Camelina with soybean in Minnesota. *Agron. J.*, **111**, 1281–1292.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B. et al. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, **20**(1), https://doi.org/10.1186/s13059-019-1905-y

Patra Bhattacharya, D., Canzler, S., Kehr, S., Hertel, J., Grosse, I. and Stadler, P.F. (2016) Phylogenetic distribution of plant snRNA families. *BMC Genom.*, **17**, 969.

Pedersen, B.S., Eyring, K., De, S., Yang, I.V. and Schwartz, D.A. (2014) Fast and accurate alignment of long bisulfite-seq reads. *arXiv* [q-bio.GN]
Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 485, 635-641.

Vasara, R., Savić, I., Nagarajan, N. and Sikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746.

Voinnet, O. (2009) Origin, biogenesis, and activity of plant microRNAs. Cell, 136, 669–687.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelbel, A., Sathikumar, S., Cuomo, C.A. et al. (2014) Pilot: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One, 9 (11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y. et al. (2011) The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43(10), 1035–1039. https://doi.org/10.1038/ng.919

Warwick, S.I., Francis, A. and Susko, D.J. (2002) The biology of Canadian weeds. 9. Thlaspi arvense L. (updated). Can. J. Plant Sci. 82, 803–823.

Weyers, S.L., Gesch, R.W., Forcella, F., Eberle, C.A., Thom, M.D., Matthees, H.L., Ott, M. et al. (2021) Surface runoff and nutrient dynamics in cover crop-soybean systems in the Upper Midwest. J. Environ. Qual. 50(1), 158–171. https://doi.org/10.1002/jeq2.20135

Weyers, S., Thom, M., Forcella, F., Eberle, C., Matthees, H., Gesch, R., Ott, M. et al. (2019) Reduced potential for nitrogen loss in cover crop-soybean relay systems in a cold climate. J. Environ. Qual. 48(3), 660–669. https://doi.org/10.2134/jeq2018.09.0350

Workman, R., Fedak, R., Kilburn, D., Hao, S., Liu, K. and Timp, W. (2019) High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing v1 (protocols.io/dvbug2vzn). protocols.io.

Workman, R.E., Myrka, A.M., Wong, G.W., Tseng, E., Welch, K.C. and Timp, W. (2018) Single-molecule, full-length sequencing profiles provide insight into the extreme metabolism of the ruby-throated hummingbird Archilochus colubris. Gigascience, 7(3), gyx009. https://doi.org/10.1093/gigascience/gyx009

Wu, D., Liang, Z., Yan, T., Xu, Y., Xuan, L., Tang, J., Zhou, G. et al. (2019) Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. Molecular Plant, 12(1), 30–43. https://doi.org/10.1016/j.molp.2018.11.007

Yanai, I., Benjamini, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A. et al. (2005) Genome-wide microarray transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics, 21(5), 650–659. https://doi.org/10.1093/bioinformatics/bti042

Yang, R., Jarvis, D.E., Chen, H., Beilstein, M.A., Grimwood, J., Jenkins, J., Shi, S. et al. (2013) The reference genome of the halophytic plant Eremea salugineum. Front. Plant Sci. 4, https://doi.org/10.3389/fpls.2013.00046

Zhang, H., Lang, Z. and Zhu, J.-K. (2018) Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. 19, 489–506.

Zhang, S.-J., Liu, L., Yang, R. and Wang, X. (2020) Genome size evolution mediated by gypsy retrotransposons in Brassicaceae. Genom. Proteom. Bioinform. 18, 321–332.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.L., Chen, H., Henderson, I.R. et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell, 126(6), 1189–1201. https://doi.org/10.1016/j.cell.2006.08.003

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2006) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. 39, 61–69.

Zou, C., Lehti-Shiu, M.D., Thibaudeau-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.-H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol. 151, 3–15.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.
Table S3  Alignment statistics of mRNA-seq reads prior to merging by tissue type.
Table S4  Detailed per-class statistics of the transposable element fraction of the T. arvense genome.
Table S5  Description of genes identified in the QTL region (Scaffold_6: 63.85–63.95 Mbp) of the BSA analysis of pale seedling phenotype in pennycress.
Table S6  BUSCO statistics on (a) initial assembly, immediately after CANU, and (b) final assembly. Both are derived from orthologs to the Eudicotyledons odb10 database.
Table S7  Merqury k-mer (k = 21) analysis of Illumina HiSeq reads sequenced from the accession in YUN_Tarv_1.0, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.
Table S8  Merqury k-mer (k = 21) analysis of Illumina HiSeq reads (PCR-free) sequenced from the accession MN106-Ref, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.
Table S9  Merqury k-mer (k = 21) analysis of each total assembly showing relative completeness of k-mers present in each read set from Illumina HiSeq.
Table S10  Summary of data provided by each institute and corresponding application.
Appendix S1  Manual curation of predicted miRNAs.
Data S1  Normalized read counts for the genes expressed in each of the tissues analysed (See excel file). Tau values are incorporated in each of the genes to highlight the specificity.
Data S2  Top 30 most-expressed genes in each tissue, relative to the mean across all tissues, from the subset of genes with a high/absolute tau specificity score.
Data S3  Location of SNPs and the primers used in the genotyping of EMS-based population for development of linkage map.
Data S4  Genetic map developed using an F2 population derived from MN106 and Ames32867.
Data S5  Genetic map developed using an F2 population derived from MN106 and 2019-M2-111.
Data S6  Phenotypes, total reads, and coverage associated with the accessions used for GWAS.