LETTER

OPEN

doi:10.1038/nature22380

The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution

Hélène Badouin[1]*, Jérôme Gouzy[1]*, Christopher J. Grassa[1,2]*, Florent Murat[3], S. Evan Staton[2], Ludovic Cottret[1], Christine Lelandais-Brière[4,5], Gregory L. Owens[2], Sébastien Carrère[1], Baptiste Mayjonade[1], Ludovic Legrand[1], Navdeep Gill[2], Nolan C. Kane[2,6], John E. Bowers[7], Sariel Hubner[2,8,9], Arnaud Bellec[10], Aurélie Bérard[11], Hélène Bergès[10], Nicolas Blanchet[1], Marie-Claude Boniface[1], Dominique Brunel[11], Olivier Catrice[1], Nadia Chaidir[2,12], Clotilde Claudel[13], Cécile Donnadieu[14], Thomas Faraut[15], Ghislain Fievet[1], Nicolas Helmstetter[10], Matthew King[2,16], Steven J. Knapp[17], Zhao Lai[18,19], Marie-Christine Le Paslier[11], Yannick Lippi[1], Lolita Lorenzon[1], Jennifer R. Mandel[20], Gwenola Marage[1], Gwenaëlle Marchand[1], Elodie Marquand[11], Emmanuelle Bret-Mestries[21], Evan Morien[2], Savithri Nambeesan[22], Thuy Nguyen[2,23], Prune Pegot-Espagnet[1], Nicolas Pouilly[1], Frances Raftis[2], Erika Sallet[1], Thomas Schiex[24], Justine Thomas[1], Céline Vandecasteele[14], Didier Varès[1], Felicity Vear[3], Sonia Vautrin[10], Martin Crespi[4,5], Brigitte Mangin[1], John M. Burke[7], Jérôme Salse[3], Stéphane Muños[1]§, Patrick Vincourt[1]§, Loren H. Rieseberg[2,18]§ & Nicolas B. Langlade[1]§

The domesticated sunflower, *Helianthus annuus* L., is a global oil crop that has promise for climate change adaptation, because it can maintain stable yields across a wide variety of environmental conditions, including drought[1]. Even greater resilience is achievable through the mining of resistance alleles from compatible wild sunflower relatives[2,3], including numerous extremophile species[4]. Here we report a high-quality reference for the sunflower genome (3.6 gigabases), together with extensive transcriptomic data from vegetative and floral organs. The genome mostly consists of highly similar, related sequences[5] and required single-molecule real-time sequencing technologies for successful assembly. Genome analyses enabled the reconstruction of the evolutionary history of the Asterids, further establishing the existence of a whole-genome triplication at the base of the Asterids II clade[6] and a sunflower-specific whole-genome duplication around 29 million years ago[7]. An integrative approach combining quantitative genetics, expression and diversity data permitted development of comprehensive gene networks for two major breeding traits, flowering time and oil metabolism, and revealed new candidate genes in these networks. We found that the genomic architecture of flowering time has been shaped by the most recent whole-genome duplication, which suggests that ancient paralogues can remain in the same regulatory networks for dozens of millions of years. This genome represents a cornerstone for future research programs aiming to exploit genetic diversity to improve biotic and abiotic stress resistance and oil production, while also considering agricultural constraints and human nutritional needs[8,9].

As the only major crop domesticated in North America, with its sun-like inflorescence that inspired artists, the sunflower is both a social icon and a major research focus for scientists. In evolutionary biology, the *Helianthus* genus is a long-time model for hybrid speciation and adaptive introgression[10]. In plant science, the sunflower is a model for understanding solar tracking[11] and inflorescence development[12]. Despite this large interest, assembling its genome has been extremely difficult as it mainly consists of long and highly similar repeats. This complexity has challenged leading-edge assembly protocols for close to a decade[13].

To finally overcome this challenge, we generated a 102× sequencing coverage of the genome of the inbred line XRQ using 407 single-molecule real-time (SMRT) cells on the PacBio RS II platform. Production of 32 million very long reads allowed us to generate a genome assembly that captures 3 gigabases (Gb) (80% of the estimated genome size) in 13,957 sequence contigs. Four high-density genetic maps were combined with a sequence-based physical map to build the sequences of the 17 pseudo-chromosomes that anchor 97% of the gene content (Fig. 1 and Supplementary Note 1.1–1.6). This compares favourably to an assembly of another sunflower genotype (HA412-HO; Supplementary Note 1.7), based on second-generation sequencing data, in which 2 Gb of sequence are placed in 816,854 contigs and 31,392 scaffolds. The sunflower genome encodes 52,232 inferred protein-coding genes and 5,803 spliced long non-coding RNAs (lncRNAs, Supplementary Note 2.1). To build the first small-RNA-mediated regulatory network for the sunflower, we identified 123 microRNA (miRNA) genes that we classified into 43 families (Supplementary Data 1), including 16 novel families. Sixty-three lncRNAs and 1,020 mRNAs are predicted to be miRNA targets, including 71 loci that probably produce secondary phased short-interfering RNAs (siRNAs, Supplementary Note 2.2).

More than three quarters of the sunflower genome consisted of long terminal repeat retrotransposons (LTR-RTs), of which 59% belong to the *Gypsy* evolutionary lineage. Sunflower LTR-RT lineages are predominantly young and exhibit minimal sequence divergence owing to significant expansion in the past one million years[5]. This pattern contrasts with that of DNA transposons, where the greatest density of

[1]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France. [2]Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada. [3]INRA/UBP UMR 1095 GDEC (Genetics, Diversity and Ecophysiology of Cereals), Clermont Ferrand 63100, France. [4]Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, University of Paris-Saclay, 91405 Orsay, France. [5]Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, University of Paris-Diderot, Sorbonne Paris-Cité, 91405 Orsay, France. [6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309-0334, USA. [7]Department of Plant Biology, Miller Plant Sciences, University of Georgia, Athens, Georgia 30602, USA. [8]Department of Biotechnology, Tel-Hai Academic College, Upper Galilee 12210, Israel. [9]MIGAL - Galilee Research Institute, PO box 831, Kiryat Shmona 11016, Israel. [10]INRA, Centre National de Ressources Génomiques Végétales, F-31326 Castanet-Tolosan, France. [11]INRA, US 1279 EPGV/CEA/CNG, Evry, France. [12]Dow AgroSciences LLC, Indianapolis, Indiana 46268, USA. [13]Biogemma, 31700 Mondonville, France. [14]INRA, GeT-PlaGe, Genotoul, Castanet-Tolosan, France. [15]INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France. [16]DuPont Pioneer, Johnston, Iowa 50131, USA. [17]Department of Plant Sciences, University of California, Davis, California 95616, USA. [18]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [19]Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana 47405, USA. [20]Department of Biological Sciences, University of Memphis, Memphis, Tennessee 38152, USA. [21]TERRES INOVIA, UMR Arche INRA/ENSAT F-31320 Castanet-Tolosan, France. [22]Department of Horticulture, University of Georgia, Athens, Georgia 30602, USA. [23]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [24]MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

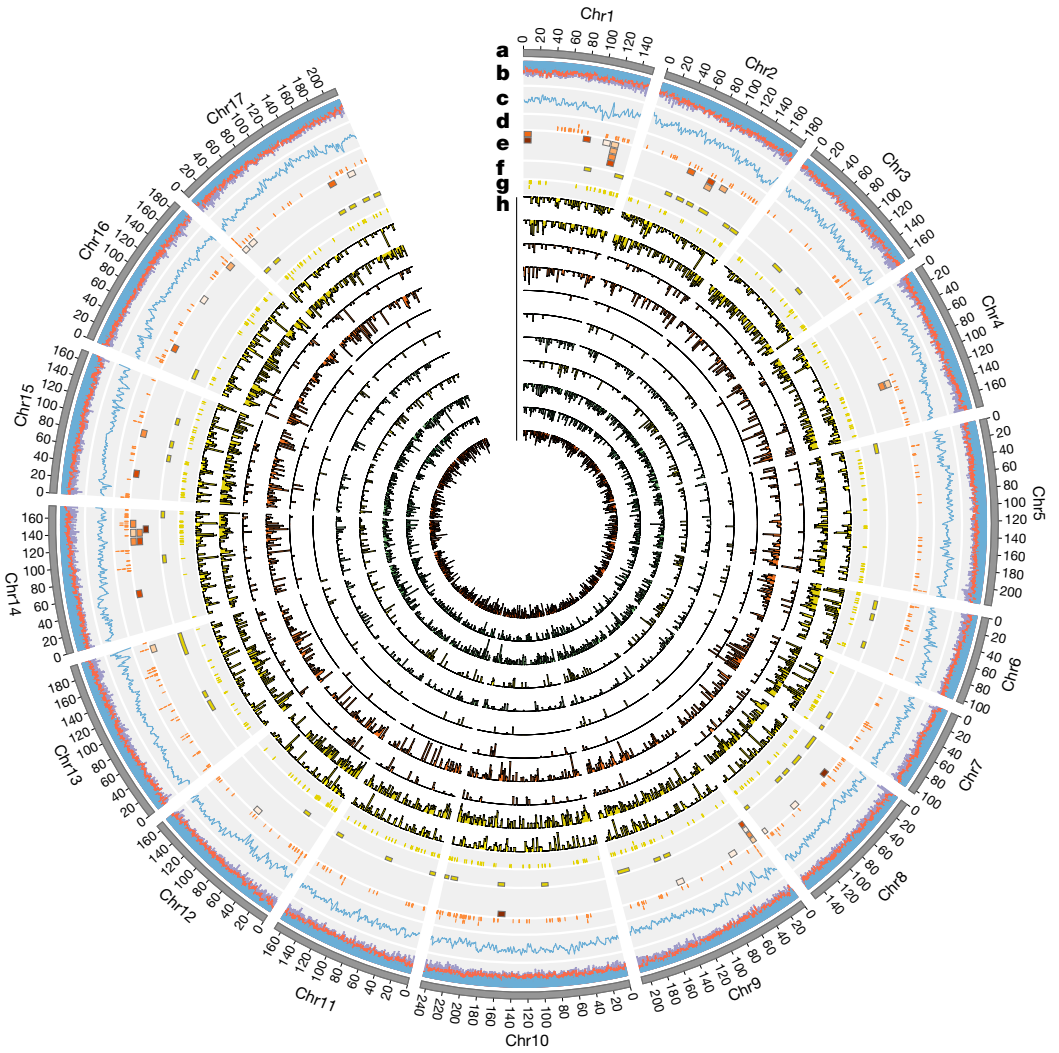


Figure 1 | The sunflower genome assembly allows integration of diversity, genetics and expression data. a, Circular representation of the pseudomolecules. **b**, Density of two families of long retrotransposon terminal repeats (5.25–9.5 kb in blue and 9.5–12.25 kb in red) and genes (purple). **c**, SNP density in 80 lines of the domesticated sunflower. **d**, Locations of genes mapping to oil metabolic pathways. **e**, QTLs for seven oil-related traits, whereby the colour (from light to dark) indicates the trait: palmitate, linoleate, oil content, oleate, phytosterol, stearate and tocopherol. **f**, Regions associated with flowering time in domesticated sunflowers. **g**, Location of homologues of *A. thaliana* flowering genes. **h**, Expression of organ-specific genes (from outside to inside tracks: pollen, stamen, pistil, disc floret ovary, ray floret ovary, disc floret corolla, bract, ray floret ligule, leaf, stem and root).

insertions is 2–4 million years old (Extended Data Fig. 1). The LTR-RTs in the sunflower exhibit non-random patterns of chromosomal distribution and are predominantly intact (Extended Data Fig. 2, Supplementary Figs 2.3.1, 2.3.2 and Supplementary Note 2.3). We found that LTR sequences display an elevated transition-to-transversion ratio, similar to that of maize[14], probably reflecting the outcomes of epigenetic silencing. We discovered that more than 6,000 transposons have acquired gene fragments, and Helitron transposons contained significantly more gene fragments than other transposon types ($P = 2 \times 10^{-16}$). In addition, 8% of Helitrons contained more than one gene fragment, with the most commonly acquired sequences being related to metabolism and defence (Supplementary Table 2.3.4). These findings highlight the creative potential of transposons and provide tools for understanding gene function in this model system.

To assess the palaeohistory of the Asterid family, we performed a comparative genomic investigation of the sunflower with lettuce[15] and artichoke[16] as representatives of Asterids II, coffee as a representative of Asterids I (ref. 17) and grape[18] as an outgroup. The grape genome is considered to be the closest modern representative of the ancestral eudicot karyotype (AEK) consisting of 7 (pre-γ ancestor) or 21 (post-γ ancestor) protochromosomes, with γ indicating the ancestral whole-genome triplication of the Eudicots (WGT-γ)[19]. We identified orthologous genes between the sunflower and grape–coffee–lettuce–artichoke as well as paralogous genes within the sunflower (Supplementary Data 2 and Supplementary Note 3.1), coffee and artichoke genomes. In addition to WGT-γ (common with grape, artichoke, lettuce, coffee and sunflower) we established that sunflower, lettuce and artichoke experienced a whole-genome triplication (WGT-1)[15,16], which has recently been proposed as independent genome duplications that are close in time[6]. A minimum of 3 chromosomal fissions and 57 chromosomal fusions were necessary for the lettuce to reach its current structure of 9 chromosomes, and 14 fissions and 60 fusions for the artichoke to reach 17 modern chromosomes. The sunflower experienced a much more complex evolutionary history with a lineage-specific whole-genome duplication (WGD-2, around 29 million years ago), in addition to the shared ancestral WGT-γ (dating back to around 122–164 million years ago) and WGT-1 (around 38–50 million years ago), plus 17 chromosomal fissions and 126 chromosomal fusions that finally shaped the present-day karyotype of 17 chromosomes (Fig. 2a). The K_s distribution (Fig. 2b) of paralogues clearly illustrates the different rounds (WGD-2, WGT-1 and WGT-γ[7]) of polyploidization events experienced by the sunflower so that for any ancestral

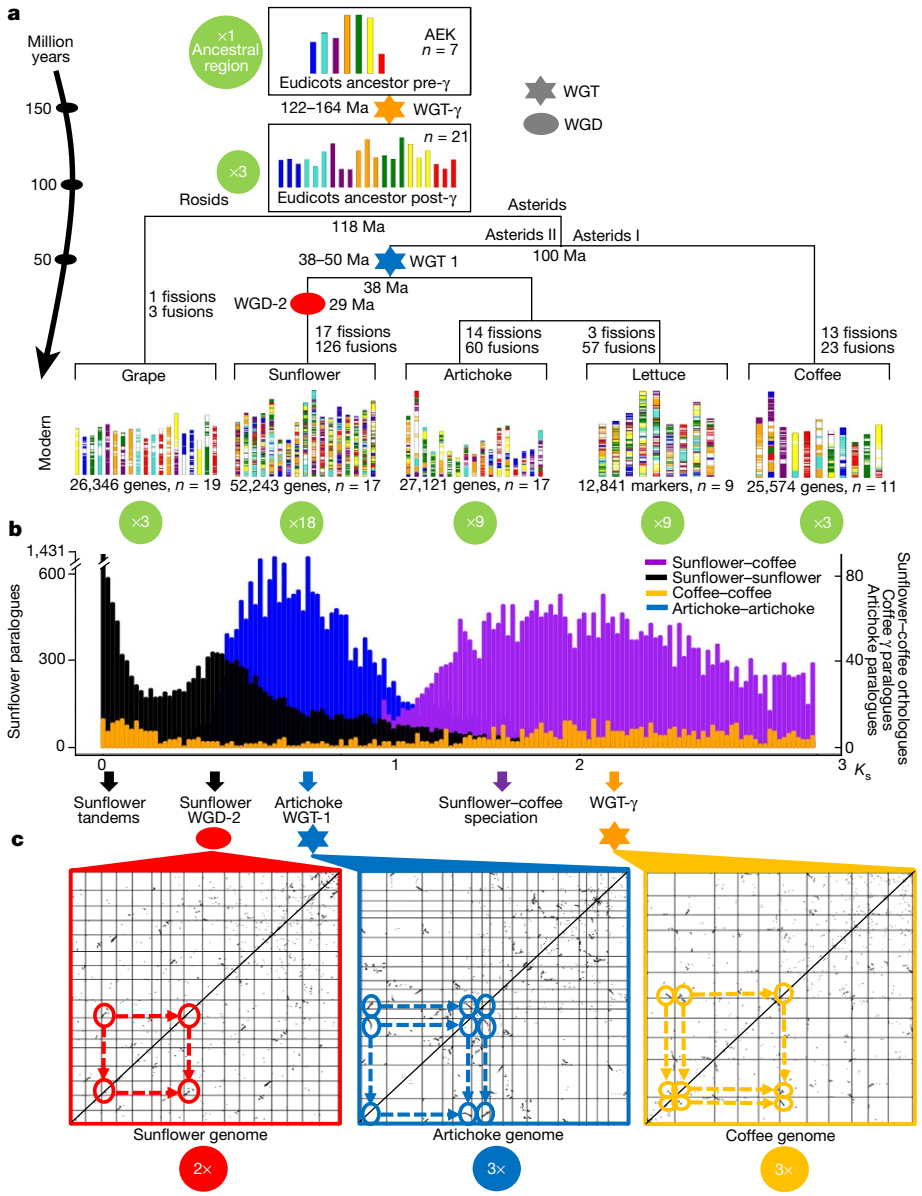


Figure 2 | Sunflower evolutionary history. a, Evolutionary scenario of the Asterids (sunflower, artichoke, lettuce and coffee) from the AEKs of 21 (post-WGT-γ) and 7 (pre-WGT-γ) protochromosomes. The modern genomes are illustrated at the bottom with the different colours reflecting the origin from the seven ancestral chromosomes from the $n = 7$ AEK (top). Polyploidization events are shown with coloured dots (duplications) and stars (triplications), along with the shuffling events (fusions and fissions). The time scale is shown on the left (million years). **b**, K_s distributions. Left *y* axis, sunflower paralogues (black); right *y* axis, coffee paralogues (orange), artichoke paralogues (blue) and sunflower–coffee orthologues (purple). Polyploidization (WGT-1, WGD-2 and WGT-γ) and speciation (sunflower–coffee) events are referenced on the *x* axis. **c**, Dot plots of paralogues in sunflower, artichoke and coffee genomes illustrating, respectively, WGD-2 (1–2 chromosomal relationships in red circles), WGT-1 (1–3 relationships in blue circles) and WGT-γ (1–3 relationships in brown circles) events.

region from the $n = 7$ AEK, a maximum number of 18 inherited regions are currently expected to be found in the modern sunflower genome. The dot plots (Fig. 2c) illustrate the paralogues inherited from WGD-2 in the sunflower genome (2–2 diagonal relationships), the paralogues deriving from WGT-1 in the artichoke genome (3–3 diagonal relationships) and finally the WGT-γ paralogues in the coffee genome (3–3 diagonal relationships). Thus, for any ancestral regions from the AEK (post-γ $n = 21$) the complete repertoire of 6–3–1–3 orthologous regions in the sunflower–artichoke–coffee–lettuce, respectively, is provided (Extended Data Fig. 3 and Supplementary Data 3).

The evolution of the cultivated sunflower progressed in two steps, domestication by native North Americans, followed by breeding involving selection on traits related to modern agricultural production. We applied an integrative approach to identify candidate genes for two major breeding traits: flowering time and seed oil content and quality. Sunflower gene networks were reconstructed with a supervised orthology-based transfer of knowledge from model species for both traits. Network genes that co-localized with genomic regions associated with variation in the traits of interest were further investigated by exploiting new information on paralogy relationships, expression and diversity data. We generated and integrated 58 transcriptomes for the roots, stem, leaves and eight floral organs (Fig. 1h, Extended Data Fig. 4 and Supplementary Data 5, 6), and for the leaves and/or roots following application of nine hormones and three abiotic stress treatments (Supplementary Note 4.1–4.3). In addition, we re-sequenced 80 domesticated lines (10–20× coverage) (Supplementary Note 5.1, 5.2). The integrative web interface Heliagene provides visualization, querying tools for data mining and network exploration for the community (https://www.heliagene.org).

Reconstructing the flowering-time genetic network in sunflower is of particular interest, because it is a key trait in crop production and the best-adapted flowering time has been selected in each cropping area during the breeding phase. Taking advantage of a recently developed database of flowering-time gene networks in *Arabidopsis thaliana*[20], we identified 485 orthologues and in-paralogues (that is, paralogues post-dating speciation) for 270 flowering-time genes in the sunflower genome (Extended Data Fig. 5, Supplementary Data 7 and Supplementary Note 6.2). There were several sunflower in-paralogues for 180 *Arabidopsis* genes, illustrating the complexity of regulatory networks in sunflower.

Previous investigations of flowering-time architecture in the sunflower[21], using more limited genomic data, focused on the transition from the wild sunflower to early domesticates. Whether flowering-time variation among modern lines involves the same genomic regions and gene families has broad implications for understanding pre- and post-domestication selection. Furthermore, the identification of ohnologous regions (that is, regions originating from whole-genome duplication) in the sunflower genome offers an excellent opportunity to determine the extent of functional diploidization for a quantitative trait in a complex genome. We used genome-wide association studies (GWAS) to dissect the genetic basis of flowering-time variation in a set of 480 F_1 hybrids obtained from 72 inbred lines, identifying 35 genomic regions associated with flowering time (Extended Data Fig. 5a and Supplementary Note 6.1). Comparison with flowering-time quantitative trait loci (QTLs) associated with domestication[21] suggests that similar genomic regions are responsible for variation among modern cultivars (Supplementary Note 6.2), possibly because selection during domestication has not been intense enough to eliminate variation at those loci, or because introgressions during sunflower breeding have reintroduced wild alleles[22]. The genomic architecture of flowering time has been shaped by the most recent whole-genome duplication (WGD-2), with more pairs of duplicated blocks associated with flowering time than is expected by chance (Extended Data Fig. 5b, Extended Data Table 1 and Supplementary Note 7). Therefore, even ancient ohnologues remain involved in the same regulatory networks and complete functional diploidization after whole-genome duplication may take long to achieve. Our integrative approach also highlights new candidate genes such as a newly discovered *AGL24* in-paralogue, which directly colocalizes with single-nucleotide polymorphisms (SNPs) associated with flowering time and new *FT* paralogues (Extended Data Fig. 5c and Supplementary Note 6.2). This analysis therefore provides insights into the architecture of flowering time in domesticated sunflowers and provides a major resource for breeding programs.

Seed oil content and quality have been under selection during sunflower improvement[23] and continue to be a primary target of breeding programs. To determine the genetic bases of these traits, we reconstructed a genome-scale metabolic network for the sunflower (Extended Data Fig. 6a and Supplementary Note 8.1) and extracted metabolic pathways involved in oil synthesis, yielding a total of 429 genes mapped onto 125 reactions, corresponding to 12 pathways (Extended Data Fig. 6b). A review of the literature on sunflower-oil synthesis showed that our network captured all 40 genes that have already been described (Supplementary Data 8), demonstrating the sensitivity of the approach.

To find evidence of selection during sunflower breeding, we mapped resequencing data of 80 genotypes and measured differentiation (F_{st}) between oil and non-oil (for example, confectionary) types of domesticated lines (Supplementary Note 8.2). Genes of the oil metabolic network were enriched in the top differentiated genes, suggesting that we had successfully identified relevant candidates for oil improvement. We found 46 oil genes in 32 genomic regions corresponding to previously identified QTLs for seven oil-related traits (Supplementary Note 8.2). Nine of these genes were highly differentiated between high- and low-oil lines (Extended Data Fig. 6c), including *FAD2-1*, which has been shown to be under selection during post-domestication[24]. Another, *HPPD*, had already been found to co-localize with a QTL for the vitamin E precursor tocopherol[25]. Our data suggest that this gene may have been targeted by selection. The remaining seven genes mainly mapped onto the diacylglycerol and linoleate biosynthesis pathways (Extended Data Fig. 6d, e). In particular, a member of the PAP2 superfamily, which is involved in biosynthesis of fatty acid precursors[26] and controls total lipid content in micro-algae[27], was predominantly expressed in seeds and co-localized with a QTL for total oil content. It therefore constitutes a strong candidate to improve this character (Extended Data Fig. 6f).

The availability of this reference genome and companion resources will not only strengthen interest in the sunflower as a model for ecological and evolutionary studies, but will also accelerate breeding programs. In addition to the genome-wide association study of flowering time presented here, precisely mapping loci that contribute to other ecologically and agriculturally important traits in wild and domesticated individuals will enable precision breeding through marker-assisted and genomic selection[28,29]. Functional validation of GWAS candidates will provide insights into the molecular mechanisms underlying variation in these traits[30]. The sunflower now has the potential to become a model crop for climate change adaptation, which can be achieved by exploiting genome-enabled systems biology and multi-disciplinary analyses of interactions between abiotic stressors, pathogen attacks and agronomic practices.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 November 2016; accepted 16 April 2017.
Published online 22 May 2017.

1. Kane, N. C. & Rieseberg, L. H. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* **175,** 1823–1834 (2007).
2. Zamir, D. Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* **2,** 983–989 (2001).
3. Fernández-Martínez, J., Melero-Vara, J., Munõz-Ruz, J., Ruso, J. & Domínguez, J. Selection of wild and cultivated sunflower for resistance to a new broomrape race that overcomes resistance of the Or_5 gene. *Crop Sci.* **40,** 550–555 (2000).
4. Seiler, G. J. Wild annual *Helianthus anomalus* and *H. deserticola* for improving oil content and quality in sunflower. *Ind. Crops Prod.* **25,** 95–100 (2007).
5. Staton, S. E. *et al.* The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. *Plant J.* **72,** 142–153 (2012).
6. Barker, M. S. *et al.* Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *Am. J. Bot.* **103,** 1203–1211 (2016).
7. Barker, M. S. *et al.* Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25,** 2445–2455 (2008).
8. Challinor, A. J., Ewert, F., Arnold, S., Simelton, E. & Fraser, E. Crops and climate change: progress, trends, and challenges in simulating impacts and informing adaptation. *J. Exp. Bot.* **60,** 2775–2789 (2009).
9. Lobell, D. B. *et al.* Prioritizing climate change adaptation needs for food security in 2030. *Science* **319,** 607–610 (2008).
10. Rieseberg, L. H., Van Fossen, C. & Desrochers, A. M. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature* **375,** 313–316 (1995).
11. Vandenbrink, J. P., Brown, E. A., Harmer, S. L. & Blackman, B. K. Turning heads: the biology of solar tracking in sunflower. *Plant Sci.* **224,** 20–26 (2014).
12. Tähtiharju, S. *et al.* Evolution and diversification of the *CYC/TB1* gene family in Asteraceae—a comparative study in *Gerbera* (Mutisieae) and sunflower (Heliantheae). *Mol. Biol. Evol.* **29,** 1155–1166 (2012).
13. Kane, N. C. *et al.* Progress towards a reference genome for sunflower. *Botany* **89,** 429–437 (2011).
14. Vitte, C. & Bennetzen, J. L. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl Acad. Sci. USA* **103,** 17638–17643 (2006).
15. Truco, M. J. *et al.* An ultra-high-density, transcript-based, genetic map of lettuce. *G3 (Bethesda)* **3,** 617–631 (2013).
16. Scaglione, D. *et al.* The genome sequence of the outbreeding globe artichoke constructed *de novo* incorporating a phase-aware low-pass sequencing strategy of F_1 progeny. *Sci. Rep.* **6,** 19427 (2016).
17. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345,** 1181–1184 (2014).
18. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).

19. Salse, J. Ancestors of modern plant crops. *Curr. Opin. Plant Biol.* **30,** 134–142 (2016).
20. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* **44** (D1), D1167–D1171 (2016).
21. Blackman, B. K. *et al.* Contributions of flowering time genes to sunflower domestication and improvement. *Genetics* **187,** 271–287 (2011).
22. Baute, G. J., Kane, N. C., Grassa, C. J., Lai, Z. & Rieseberg, L. H. Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytol.* **206,** 830–838 (2015).
23. Chapman, M. A. & Burke, J. M. Evidence of selection on fatty acid biosynthetic genes during the evolution of cultivated sunflower. *Theor. Appl. Genet.* **125,** 897–907 (2012).
24. Merah, O. *et al.* Genetic analysis of phytosterol content in sunflower seeds. *Theor. Appl. Genet.* **125,** 1589–1601 (2012).
25. Haddadi, P. *et al.* Genetic dissection of tocopherol and phytosterol in recombinant inbred lines of sunflower through quantitative trait locus analysis and the candidate gene approach. *Mol. Breed.* **29,** 717–729 (2012).
26. Carman, G. M. & Han, G.-S. Roles of phosphatidate phosphatase enzymes in lipid metabolism. *Trends Biochem. Sci.* **31,** 694–699 (2006).
27. Deng, X. D., Cai, J. J. & Fei, X. W. Involvement of phosphatidate phosphatase in the biosynthesis of triacylglycerols in *Chlamydomonas reinhardtii*. *J. Zhejiang Univ. Sci. B* **14,** 1121–1131 (2013).
28. Bolger, M. E. *et al.* Plant genome sequencing — applications for crop improvement. *Curr. Opin. Biotechnol.* **26,** 31–37 (2014).
29. Kang, Y. J. *et al.* Translational genomics for plant breeding with the genome sequence explosion. *Plant Biotechnol. J.* **14,** 1057–1069 (2016).
30. Curtin, S. J. *et al.* Validating genome-wide association candidates controlling quantitative variation in nodulation. *Plant Physiol.* **173,** 921–931 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Kuhn for sharing his expertise in PacBio sequencing and H. Witsenboer for his help with the production of the Fingerprint-based physical map; the Genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing help and computing resources, the common services of the LIPM for their support, and Genome Quebec Innovation Centre and Canada's Michael Smith Genome Science Centre for 454 and Illumina sequencing; M. Scascitelli, M. Stewart, D. Ebert, J. Roeder, H. Shaffer, E. Gudger, B. Hsieh, S. Jackson, S. Rounsley, C. Feuillet, B. Barbazuk and M. Barker for their help and advice during the Genome Canada/Genome BC project; and D. Swanevelder for contributing to the sequencing of the sunflower association mapping populations; members of the International Consortium for Sunflower Genomics resources (2012–2015): Advanta, BASF, Biogemma, Dow, KWS, Pioneer and Syngenta companies and their sunflower project leaders; F. Bonnafous for the development of the statistical pipeline for GWAS and P. Castellanet, C. Henry, M. Laporte, J. Piquemal, M. Coque and T. André for the coordination of flowering time phenotyping on the sunflower hybrid panel (GWAS). This project was funded by the French National Research Agency (SUNYFUEL/ANR-07-GPLA-0022 and SUNRISE/ANR-11-BTBR-0005 projects), by the Midi-Pyrénées Region, the European Fund for Regional Development, the French Fund for Competitiveness Clusters (FUI), the Genoscope SystemSun project, Genome Canada and Genome BC's Applied Genomics Research in Bioproducts or Crops (ABC) Competition, the NSF Plant Genome Program (DBI-0820451) and the International Consortium for Sunflower Genomics Resources.

Author Contributions F.M., S.E.S., L.C., C.L.-B. and G.L.O. contributed equally to this work. M.C., B.Man., J.M.B. and J.S. contributed equally to this work. A.Bel., H.Be. and N.H. prepared BAC libraries. B.May. developed the DNA-extraction protocol for PacBio sequencing. B.May., C.V. and C.D. performed PacBio sequencing. A.Bér., D.B., D.V., E.Ma., E.B.-M., G.Marc., G.Mara., J.R.M., J.T., L.Lo., M.-C.B., M.-C.L.P., N.B., N.B.L., N.P., S.N., S.V., Y.L. and Z.L. contributed to DNA/RNA sample collection and data production. O.C. performed flow cytometry experiments. N.G., T.N. and N.C.K. built the physical map and integrated the physical and genetic maps. C.J.G., S.M., J.E.B. and J.M.B. developed genetic maps. J.G. assembled the XRQ genome. C.J.G. built the XRQ pseudomolecules. C.C. performed quality control of XRQ pseudomolecules. C.J.G., J.E.B., N.C.K., S.H. and M.K. assembled the HA412-HO genome. J.G., E.S. and T.S. annotated protein-coding genes and miRNA (XRQ). S.E.S. annotated the HA412-HO genome. S.C., J.G., F.R., M.K., T.F., C.J.G., J.E.B., N.C.K., N.G., T.N., N.C., E.Mo. developed bioinformatics resources. L.Le., E.Ma. and G.F. performed bioinformatics analyses. G.L.O. conducted ancestry analyses. P.V. designed the GWAS hybrid panel. B.Man., N.B.L. and P.V. designed the GWAS experiment. F.V. developed the XRQ inbred line. B.Man., P.P.-E. conducted the GWAS analysis. L.C. conducted metabolism analyses. F.M. and J.S. conducted palaeo-evolution analyses. S.E.S. conducted repeat analyses. C.L.-B. and M.C. conducted small-RNA analyses. H.Ba., S.M. and N.B.L. performed integrated analyses on flowering time and oil metabolism. H.Ba. and N.B.L. performed transcriptomic analysis. H.Ba. performed analysis of sunflower ohnologues. S.J.K. contributed to the genome consortium coordination. N.B.L., L.H.R., P.V., S.M., J.M.B. and J.G. designed experiments and coordinated the project. L.H.R. coordinated the sunflower genome consortium. H.Ba., N.B.L. and L.H.R. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to N.B.L. (nicolas.langlade@inra.fr).

Reviewer Information *Nature* thanks A. Paterson, J. Schmutz and Y. Van der Peer for their contribution to the peer review of this work.

METHODS

A full description of the Methods can be found in the Supplementary Information. No statistical methods were used to predetermine sample size. The genome-wide association experiments were fully randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Genome sequencing and assembly of the XRQ genotype. *Sequencing.* The DNA of the INRA inbred genotype XRQ (Supplementary Note 1.1) was extracted following a previously published protocol[31], and sequenced using 407 SMRT cells with P6/C4 chemistry. Subreads were obtained using the SMRT Analysis RS.Subreads.1 pipeline (Supplementary Note 1.2). In total 32.8 million subreads were generated with an N_{50} of 13.7 kb and a mean length of 10.3 kb. The targeted genome coverage of 102× was obtained with 367 Gb of raw sequence (340 Gb of subread data).

Assembly. The PBcR wgs8.3rc1 assembly pipeline[32] was used to perform the correction of reads, WGS 8.3 to assemble the corrected reads and quiver[33] to polish the consensus sequence after the construction of the pseudomolecules (see below). However, to overcome challenges associated with the sunflower genome assembly, substantial parameter tuning, code modification and software development were required and these are described in Supplementary Note 1.3–1.7.

Physical map construction, genetic map construction and assembly of pseudomolecules. To develop a robust physical map for the sunflower that could be used to help to place sequence contigs on chromosomes and determine the physical length of gaps between them, bacterial artificial chromosome (BAC) libraries were constructed for genotype HA412-HO by the French Plant Genome Resource Center (http://cnrgv.toulouse.inra.fr/en/library/sunflower). We used 382,464 clones from the three BAC libraries to develop a 12.5× physical map, which was integrated with high-density genetic maps (see below). The resulting physical map covers approximately 3.3 Gb (around 92.5% of the 3.6 Gb genome) and is publicly available at https://www.sunflowergenome.org/.

We developed several high-density genetic maps that we used for correctly placing and ordering BAC and sequence contigs on chromosomes, as well as for the association and QTL analyses. While individual maps had gaps with no mappable markers owing to identity by descent, this problem was minimized by the use of multiple mapping populations (Supplementary Note 1.5). The pseudomolecules were assembled as described in Supplementary Note 1.6, leading to a final assembly of 17 pseudomolecules and 1,509 unanchored contigs. A web browser of this genome assembly is available at https://www.heliagene.org/HanXRQ-SUNRISE/.

Sequencing, assembly and annotation of the genome of another genotype, HA412-HO, is presented in Supplementary Note 1.7.

Annotation of protein-coding genes and lncRNAs. Gene models were predicted using EuGene 4.2 (ref. 34) embedded in a new and fully automated pipeline that integrates probabilistic sequence model training, genome masking, transcript- and protein-alignment computation and alternative splice site detection. The plant early release of BUSCO (release July 2015)[35] was run on the set of predicted transcripts, and it detected 92% of complete gene models (590 complete single copy and 291 duplicated, respectively) plus 10 additional fragmented gene models.

Protein-coding genes were annotated using a three-step process, taking into account reciprocal best hits in the SwissProt and TAIR10 (ref. 36) databases (12,360 sunflower proteins), protein-domain content using Interpro (26,646 sunflower proteins), and similarity with plant proteomes (Ensembl release 30) or coverage of the transcript with RNA-sequencing data (1,200 predicted proteins with similarities in other plant proteomes without expression support, 1,832 with similarities in other plant proteomes with expression support and 8,542 gene models supported by expression data, but without significant hits with other plant proteomes). The remaining 1,663 predicted proteins remained completely uncharacterized. Details of the gene prediction and annotation process are provided in Supplementary Note 2.1.

Annotation of small RNA. To identify *H. annuus* miRNA genes, we constructed a small-RNA library using mixed RNAs from the various organs in control conditions (as for RNA sequencing) and sequenced them using Illumina GAIIx (oriented single-end 50 nucleotides (nt)). A total of 139 million reads were obtained that classically displayed a size distribution with two peaks of 21 and 24 nt small RNAs (Supplementary Note 2.2). Genome-wide prediction of miRNAs was performed combining Shortstack version 3.4 (ref. 37) and an adapted version of the pipeline described in ref. 38, post-processed with the stringent criteria proposed by MiRBase[39]. Targets of miRNA were predicted using miRanda version 3.0 (http://www.microrna.org).

Annotation of repeats. LTR-RTs were annotated with an in-house pipeline that uses LTRharvest[40] and LTRdigest[41]. DNA transposons were annotated with a custom pipeline that includes the 'gt tirvish' command, which is part of the GenomeTools suite[42]. The age of LTR-RTs was determined by obtaining a likelihood divergence estimate between the LTRs with baseml from PAML[43] and using this divergence value (hereafter d) to calculate the LTR-RT age with the equation $T = d/2r$, where $r = 1 \times 10^{-8}$ (ref. 44). The total transposable element content was estimated to be 74.7 ± 0.08% (mean ± s.d.) on the basis of analyses with Transposome from random sequence reads (Supplementary Table 2.3.3). The detailed annotation pipeline of repeated elements is described in Supplementary Note 2.3.

Sunflower palaeogenomics. A comparative analysis was performed with sunflower, artichoke[16], coffee[17] and lettuce[15] and with grape[18] as the outgroup. Identification of orthology and paralogy relationships, measurements of sequence divergence and estimation of divergence time through the level of synonymous substitutions were performed as detailed in Supplementary Note 3.1 on the basis of the methods described in ref. 45 and the Timetree web service to estimate speciation dates (http://www.timetree.org/). Speciation events were dated to 38 million years ago (Ma) for sunflower–artichoke, 100 Ma for sunflower–coffee and 118 Ma for sunflower–grape. Palaeoploidization events were dated to 122–164 Ma for WGT-γ, 38–50 Ma for WGT-1 and 29 Ma for WGD-2.

Ancestry of the sunflower genome. To identify introgressed regions in the XRQ and HA412-HO genome assemblies, we used previously published transcriptome sequences[22] from 60 genotypes representing native North-American landraces (that is, early domesticates), and several wild species that are probable donors to modern cultivated lines based on pedigree information, *H. argophyllus*, *H. petiolaris* and *H. tuberosus* (Supplementary Table 3.2.1). Raw reads were aligned to the genome assemblies and filtered as described in Supplementary Note 3.2. To identify introgressed regions in the genomes of XRQ and HA412-HO we used the 'site-by-site' linkage admixture model in STRUCTURE[46] (Supplementary Note 3.2). Genome-wide and window estimates of introgression are provided in Supplementary Table 3.2.2 and Supplementary Figs 3.2.1, 3.2.2.

Transcriptome sequencing and analysis. We generated 58 paired-end RNA-sequencing libraries to measure expression in 11 sunflower organs, the responses to hormonal and osmotic and salt treatments in roots and leaves, as well as response to variable water status (Supplementary Note 4.1). Library sequencing was done with Illumina HiSeq, reads were mapped with the glint software (https://forge-dga.jouy.inra.fr/projects/glint) and only the best scoring pair(s) of reads was(were) kept. Expression measurements and normalization were performed as described in Supplementary Note 4.2. Organ-specificity was measured by computing a specificity index, Tau[47], on the normalized expression score. We identified sets of organ-specific genes and regulators (transcription factors and lncRNAs) (Extended Data Fig. 4 and Supplementary Note 4.2). Analysis of differential expression in response to hormones and stress treatments were performed with the glm model of EdgeR[48] as detailed in Supplementary Note 4.2. Gene Ontology enrichment tests were carried out with Blast2GO Pro (one-sided Fisher's exact tests, false discovery rate of <0.05).

Resequencing of domesticated lines. We resequenced 80 lines of the sunflower mapping population (SAM) that represent the diversity of the cultivated sunflower. Statistics on resequenced lines are provided in Supplementary Table 5.1.1. Seventy-two parent lines of the 480 hybrids used in a genome-wide association analysis of flowering time were also resequenced. The paired-end libraries were resequenced with Illumina HiSeq, read mapping was performed with the glint software (https://forge-dga.jouy.inra.fr/projects/glint) and SNP calling with VarScan[49] (Supplementary Notes 5.1, 5.2).

Identification of flowering time orthologues and in-paralogues. Flowering time genes in *A. thaliana* were retrieved from a recently developed database, FLOR-ID[20], which includes 295 protein-coding genes and 11 miRNA genes and describes their interactions. We built gene clusters for a set of seven species, namely *H. annuus*, *A. thaliana*, *Cynara cardunculus*, *Oryza sativa*, *Hordeum vulgare*, *Brassica rapa* and *Populus trichocarpa*, chosen to be consistent with a previous study that identified orthologues for more than 30 flowering-time genes in the sunflower[21], adding the proteome of the recently sequenced member of Asterids II *C. cardunculus*[16]. To identify orthologues and in-paralogues (that is, paralogues post-dating speciation) of *A. thaliana* genes, we built and visually examined trees for the clusters defined above (Supplementary Note 6.2) and manually screened BLAST reports on the sunflower genome browser. We identified 485 orthologues and in-paralogues (Supplementary Data 7). A genome-wide association study of flowering time was performed on a set of 480 hybrids obtained from 72 inbred genotypes (Supplementary note 6.1), and colocalization of flowering-time orthologues with flowering time QTLs was assessed with bedtools[50].

Analysis of paralogues dating from the most recent whole-genome duplication (WGD-2). Correlation of expression between WGD-2 paralogues was assessed quantitatively by measuring the Pearson correlation coefficient and qualitatively by counting the number of pairs of paralogues that belong to the same co-expression modules based on a weighted gene co-expression network constructed with WGCNA (Supplementary Note 7). Significance was tested with 1,000 permutations

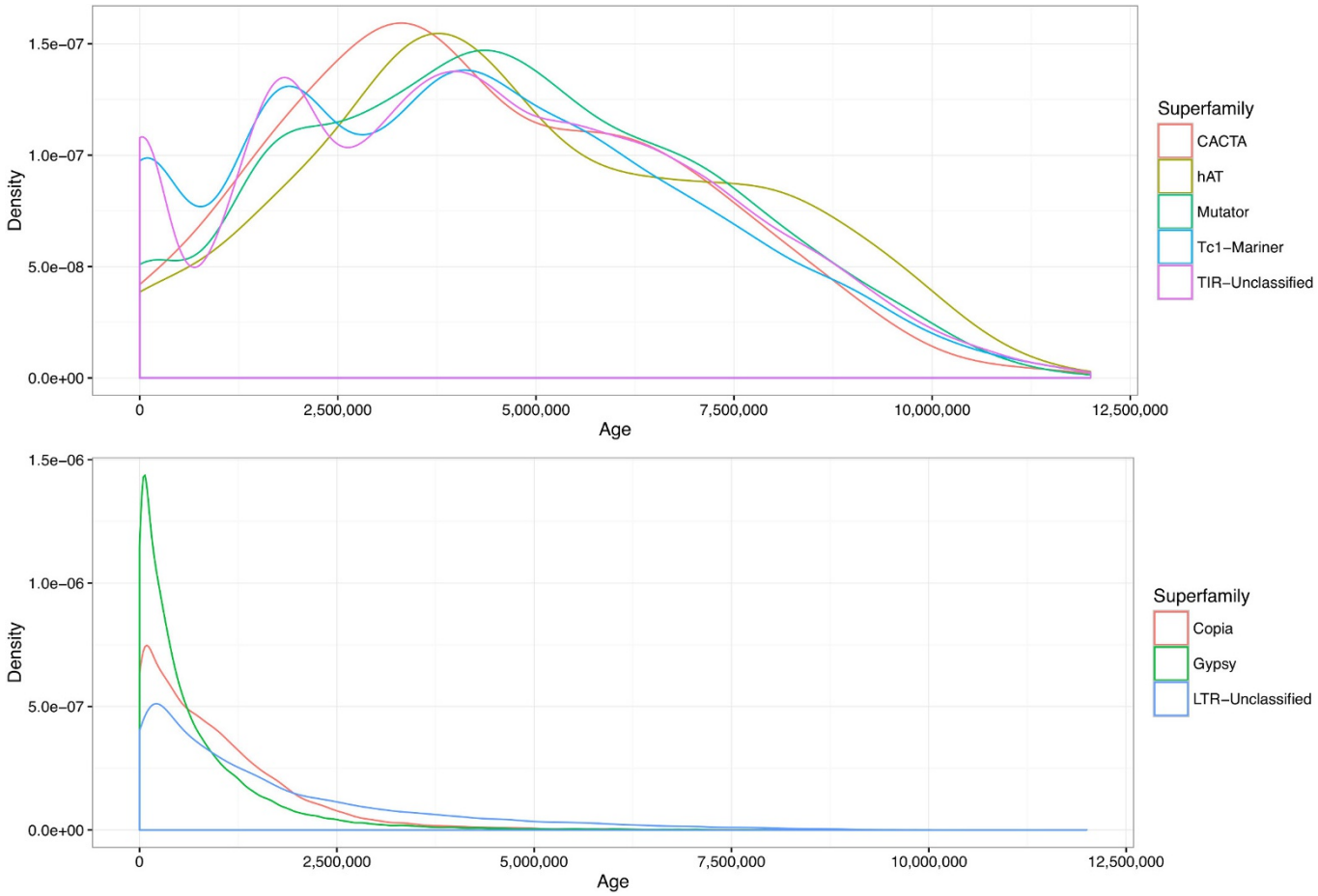
of the genes in the expression matrix. The level of functional diploidy of the genome for flowering time was measured as the number of pairs of WGD-2 paralogous genes or paralogous genomic regions for which both members of the pair (that is, both paralogous genes or both paralogous genomic regions) intersected with genomic intervals corresponding to flowering-time QTLs. Paralogous blocks were identified by a chaining approach detailed in Supplementary Note 7. Observed counts were compared to a null distribution obtained from 1,000 permutations of flowering-time QTLs for several sets of parameters (Extended Data Table 1, Supplementary Note 7).

Reconstruction of oil metabolic pathways. The metabolic annotation of protein sequences was performed with the E2P2 software (version 3.0, https://dpb.carnegiescience.edu/labs/rhee-lab/software). We used the pathway-tools software[51] to infer biochemical reactions and metabolic pathways from the protein annotations. The super pathway of sunflower oil metabolism was created on the basis of the main components of the known sunflower oil metabolism by merging 16 pathways, and it includes 125 reactions, 160 metabolites and 429 genes (Supplementary Note 8.1). Web resources for exploring the sunflower metabolism network are available at https://www.heliagene.org/HanXRQ-SUNRISE/data/analyses/metabolism.

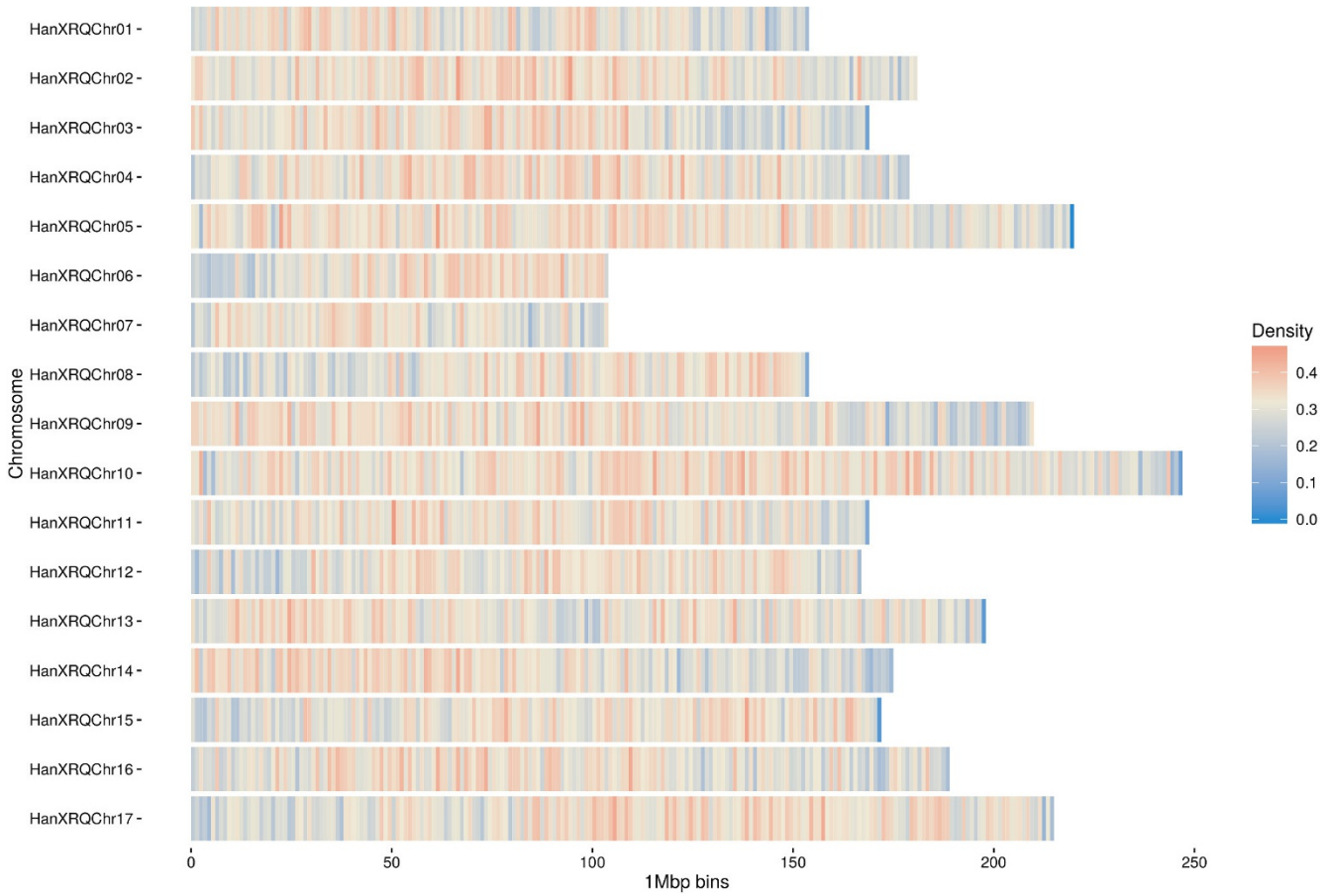
Integrative candidate genes analysis for oil metabolism. We measured the F_{st} (ref. 52) between lines cultivated for oil production and other lines (mainly confectionary for human consumption) with egglib version 2 (ref. 53). Genes of the oil super pathway that possessed an F_{st} score above the 95th percentile were further examined. Forty-nine previously published QTLs[54–56] were mapped to the XRQ genome assembly and 5 Mb were added at the flanks of the mapped markers to define the QTL coordinates and assess colocalization with candidate genes (Supplementary Note 8.2).

Data availability. This whole genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession MNCJ00000000. Transcriptome and resequencing sequence reads have been deposited in the SRA database as studies SRP092899, SRP092742, SRP093222 and SRP095974.

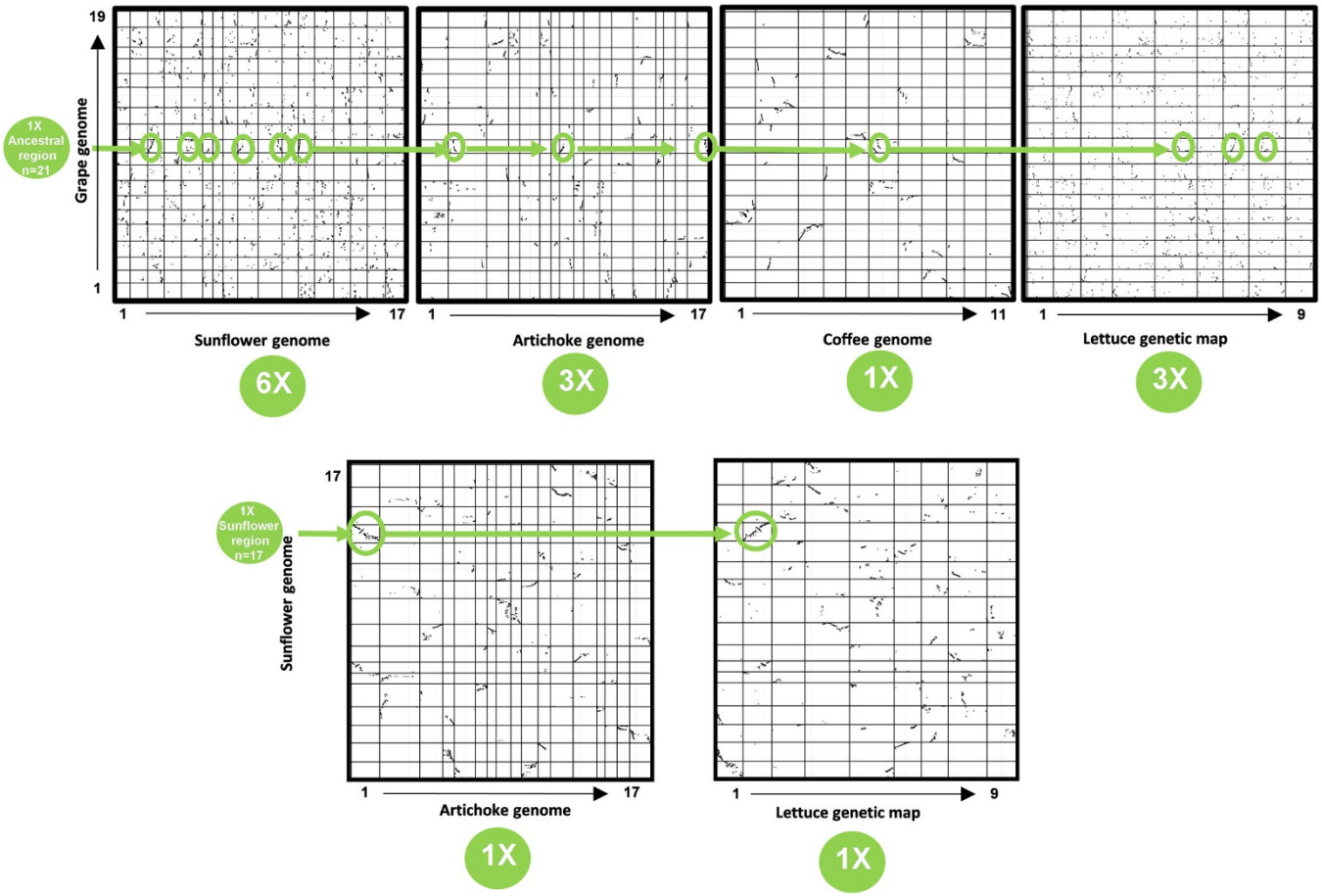
31. Mayjonade, B. *et al.* Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **61,** 203–205 (2016).
32. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33,** 623–630 (2015).
33. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).
34. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* **3,** 87–97 (2008).
35. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
36. Lamesch, P. *et al.* The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40,** D1202–D1210 (2012).
37. Axtell, M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19,** 740–751 (2013).
38. Formey, D. *et al.* The small RNA diversity from *Medicago truncatula* roots under biotic interactions evidences the environmental plasticity of the miRNAome. *Genome Biol.* **15,** 457 (2014).
39. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–D73 (2014).
40. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 18 (2008).
41. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Res.* **37,** 7002–7013 (2009).
42. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10,** 645–656 (2013).
43. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
44. Strasburg, J. L. & Rieseberg, L. H. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*—large effective population sizes and rates of long-term gene flow. *Evolution* **62,** 1936–1950 (2008).
45. Salse, J., Abrouk, M., Murat, F., Quraishi, U. M. & Feuillet, C. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.* **10,** 619–630 (2009).
46. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
47. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–659 (2005).
48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
49. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285 (2009).
50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
51. Karp, P. D., Paley, S. & Romero, P. The pathway tools software. *Bioinformatics* **18** (Suppl 1), S225–S232 (2002).
52. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132,** 583–589 (1992).
53. De Mita, S. & Siol, M. EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* **13,** 27 (2012).
54. Ebrahimi, A. *et al.* QTL mapping of seed-quality traits in sunflower recombinant inbred lines under different water regimes. *Genome* **51,** 599–615 (2008).
55. Pérez-Vich, B. *et al.* Molecular basis of the high-palmitic acid trait in sunflower seed oil. *Mol. Breed.* **36,** 43 (2016).
56. Premnath, A., Narayana, M., Ramakrishnan, C., Kuppusamy, S. & Chockalingam, V. Mapping quantitative trait loci controlling oil content, oleic acid and linoleic acid content in sunflower (*Helianthus annuus* L.). *Mol. Breed.* **36,** 106 (2016).



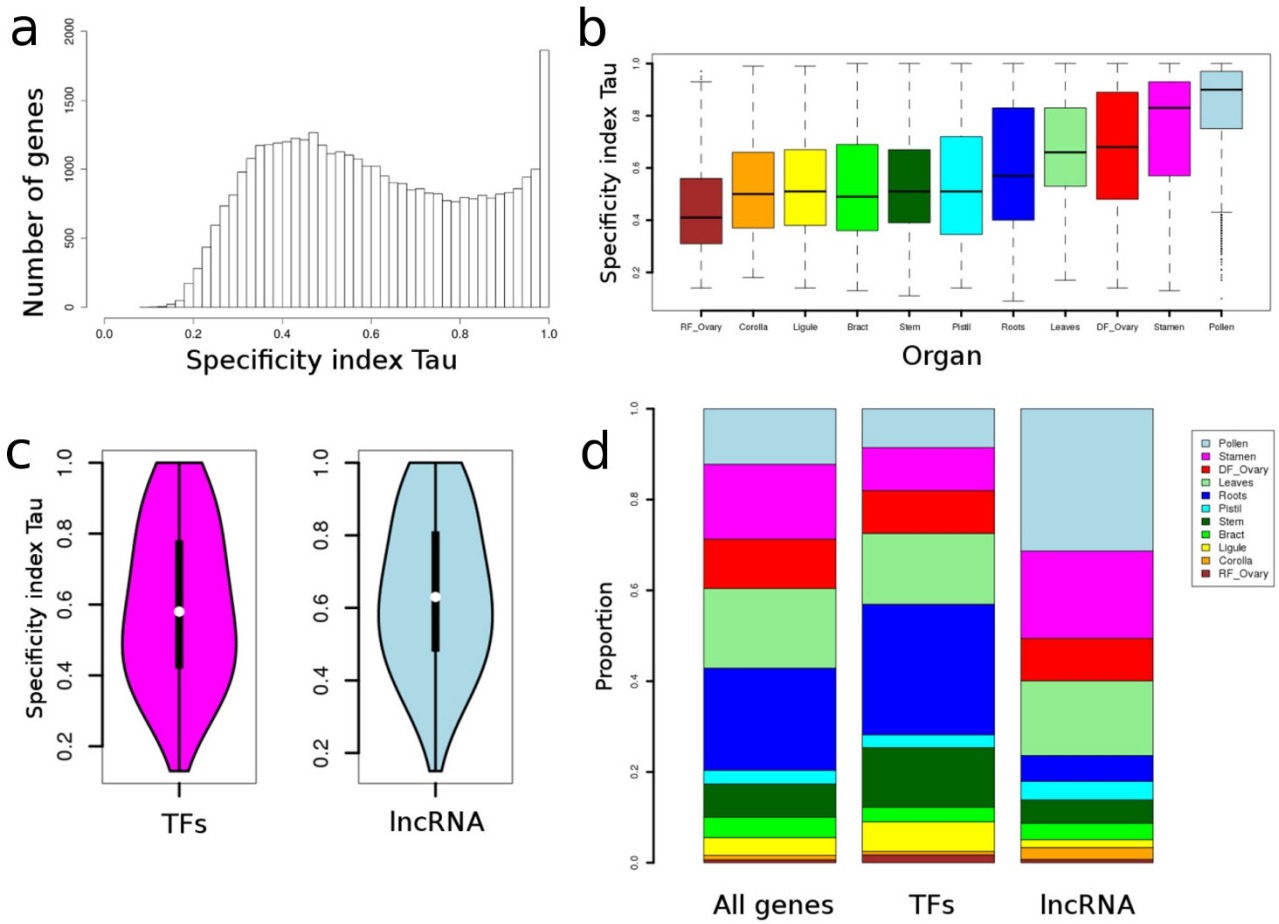
Extended Data Figure 1 | Age distribution of transposons in the sunflower. The *x* axis represents the age of insertions in millions of years, the *y* axis is the density of insertions at a given time point. Top, the age distribution of each superfamily of subclass I of the Class II transposons (the terminal inverted repeat transposons). Bottom, the age distribution of LTR-RT superfamilies.



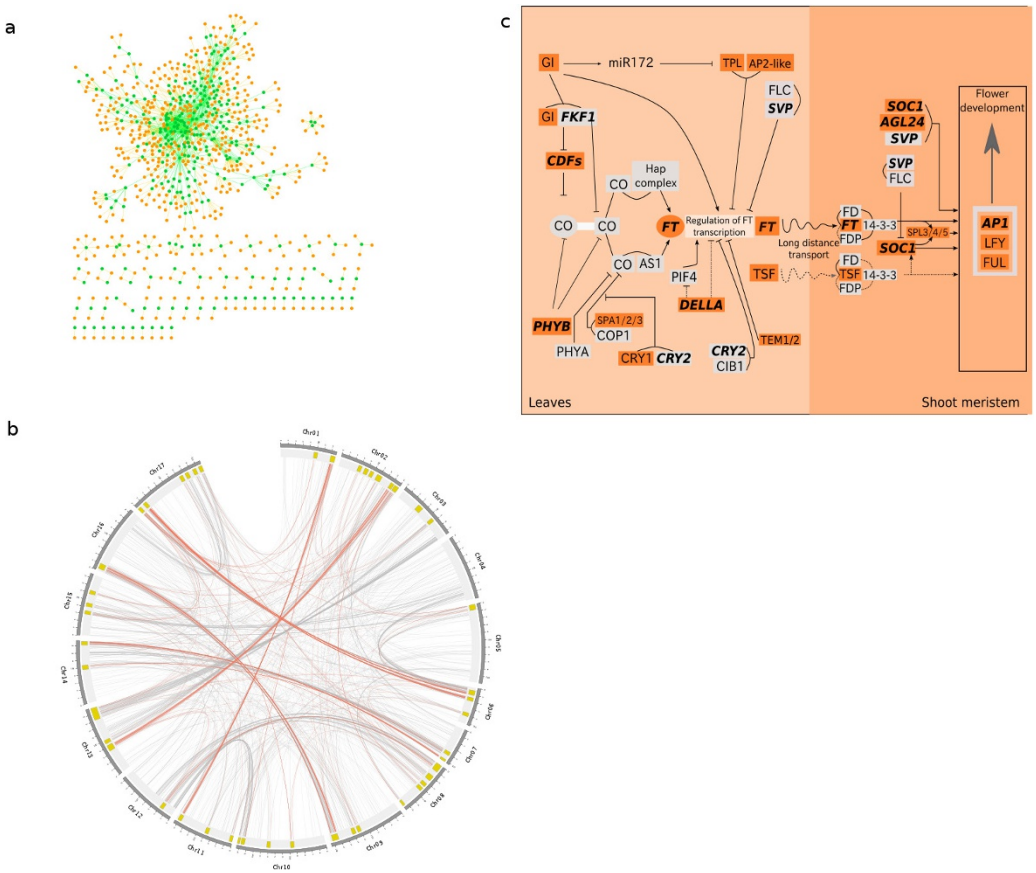
Extended Data Figure 2 | The density of LTR-RTs in 1 Mb bins per chromosome. The scale represents a fraction, where 1.0 is 100% of a given bin.



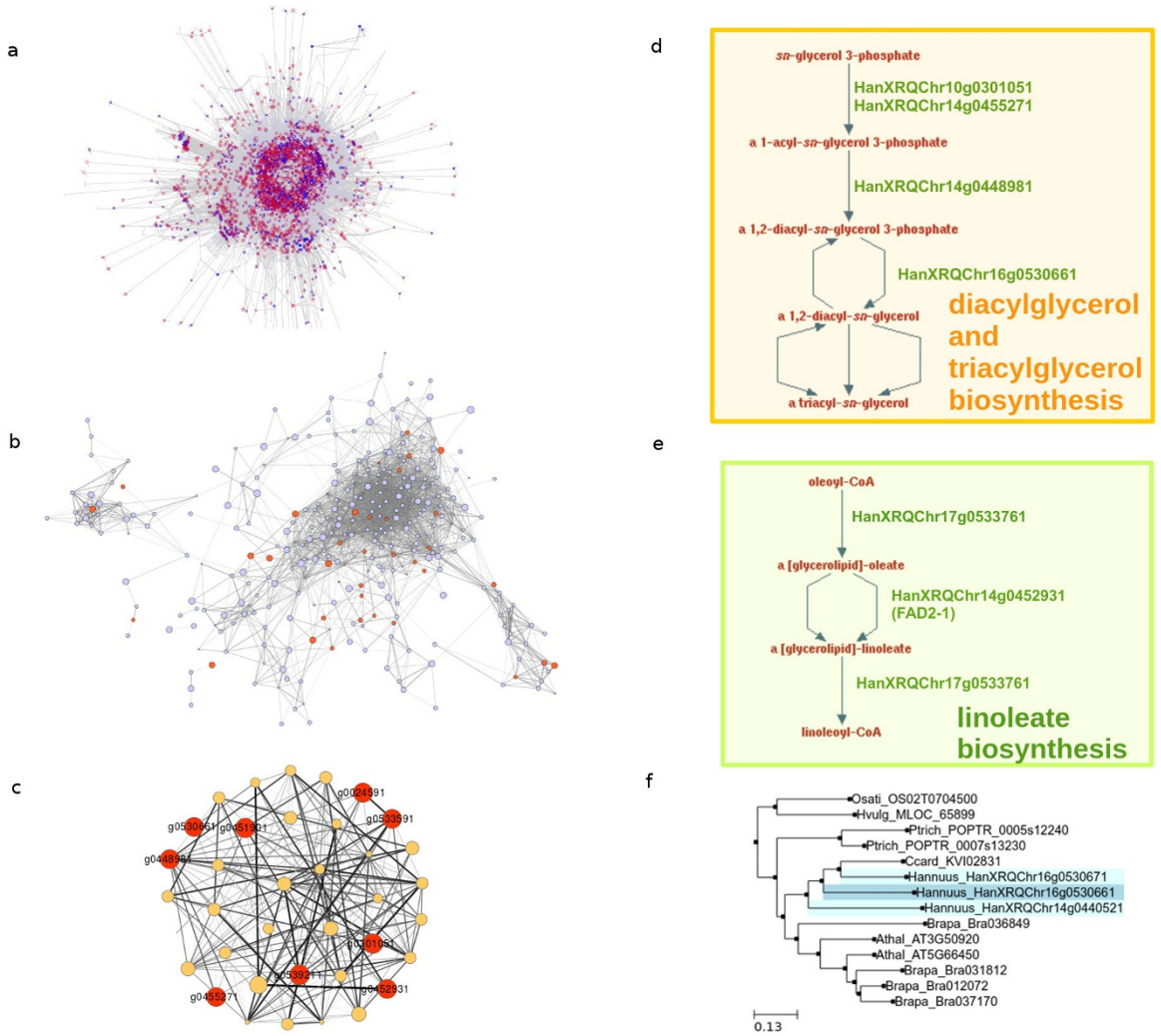
Extended Data Figure 3 | Comparison of grape–sunflower–artichoke–coffee–lettuce genomes. Top, dot plots of orthologues between the grape genome (*y* axis, as a representative of the $n = 21$ post-γ ancestor) and, from left to right, the sunflower (1–6 chromosomal relationships inherited from WGT-1 and WGD-2), artichoke (1–3 chromosomal relationships deriving from WGT-1), coffee (1–1 chromosomal relationships illustrating the absence of a coffee-specific WGD, despite WGT-1) genomes and the lettuce genetic map (1–3 chromosomal relationships deriving from WGT-1). Bottom, dot plots of orthologues between the sunflower genome (*y* axis, $n = 17$ chromosomes) and artichoke (*x* axis, $n = 17$ chromosomes) and lettuce (*x* axis, $n = 9$ chromosomes) genomes with 1–1 chromosomal relationships.



Extended Data Figure 4 | Organ-specific expression in the sunflower transcriptome. a, Histogram of the specificity index Tau in expressed genes. **b**, Box plot distribution of the specificity index Tau in 11 different organs. The different organs are represented with the following colours: Ray floret ovary, dark brown; disc floret corolla, orange; ray floret ligule, yellow; bract, bright green; stem, dark green; pistil, bright blue; roots, dark blue; leaves, light green; disc floret ovary (seeds), red; stamens, magenta; pollen, light blue. **c**, Violin plot of the specificity index Tau for transcription factors (TFs, magenta) and long non-coding RNA (lncRNA, light blue). **d**, Cumulative bar plot showing the organ distribution of specific genes (left), transcription factors (middle) and lncRNA (right). Colours are the same as in **b**.



Extended Data Figure 5 | Integrative analysis of flowering time. **a**, Flowering time network in the sunflower. Flowering time genes of *A. thaliana* and their interactions are drawn in green. Sunflower genes and orthology relationships with *A. thaliana* genes are shown in orange. **b**, Genomic architecture of flowering time in the domesticated sunflower. Outer ring, location of genomic regions associated with flowering time. Inner ring, links between ohnologues of a sunflower-specific whole-genome duplication (WGD-2), limited to genes located in regions associated with flowering time. Links between ohnologues of WGD-2 that are both located in regions associated with flowering time are drawn in red, other links are drawn in grey. **c**, Pathway of the integration of flowering signals in meristem (simplified pathway adapted from ref. 20). The bright orange backgrounds indicate genes for which at least one sunflower orthologue was located in a region associated with flowering time. Bold italic genes indicates genes for which we identified additional in-paralogues compared to a previous study using more limited genomic data[21]. Simple arrows represent positive regulation and other arrows negative regulation. Curved lines between genes represent protein–protein complexes.



Extended Data Figure 6 | Integrative analysis of oil metabolism. **a**, Whole-metabolic network (3,821 reactions and 475 pathways). Genes are coloured by expression levels in developing seeds. **b**, Co-expression network of oil metabolic pathway. Genes that co-localize with QTLs are coloured in orange. **c**, Sub-network with genes from **b** co-localizing with QTLs. Node size is proportional to F_{st} between lines cultivated for oil production and other domesticated lines. Genes with an F_{st} in the top 5% are coloured in dark orange. **d**, Mapping of candidate genes (orange genes from **c**) on the pathways of diacylglycerol and triacylglycerol biosynthesis. **e**, Mapping of candidate genes on the pathway of linoleate biosynthesis. **f**, Tree of a gene cluster including a candidate gene of the PAP2 superfamily, involved in the synthesis of fatty acid precursors (**d**). Athal, *Arabidopsis thaliana*; Brapa, *Brassica rapa*; Ccard, *Cynara cardunculus*; Hvulg, *Hordeum vulgare*; Osati, *Oryza sativa*; Ptrich, *Populus trichocarpa*.

Extended Data Table 1 | Link between the genomic architecture of flowering time and the most recent whole-genome duplication experienced by the sunflower

a

parameters	number of pairs	summary statistics of a distribution based on 1000 permutations					
		mean	median	p5	p95	p99	p99.5
fIBP=5Mbp; fo=0.5, minBlkSize=0	55	43.629	43.0	34.0	55.0	58.0	59.005
fIBP=5Mbp; fo=0.5, minBlkSize=10000	26	18.847	19.0	12.0	26.0	29.0	30.0
fIBP=5Mbp; fo=0.5, minBlkSize=100000	23	13.157	13.0	8.0	19.0	21.0	22.0
fIBP=5Mbp; fo=0.5, minBlkSize=1000000	12	4.51	4.0	2.0	8.0	10.0	11.0
fIBP=5Mbp; fo=10-9, minBlkSize=10000	30	23.388	23.0	16.0	31.0	35.0	36.005
fIBP=5Mbp; fo=10-9, minBlkSize=100000	27	17.201	17.0	11.0	24.0	27.0	28.0
fIBP=5Mbp; fo=10-9, minBlkSize=1000000	15	8.037	8.0	4.0	13.0	15.0	15.0
fIBP=1Mbp; fo=10-9, minBlkSize=10000	10	6.369	6.0	3.0	10.0	12.0	12.005
fIBP=10Mbp; fo=10-9, minBlkSize=10000	62	46.906	47.0	36.0	58.05	63.0	65.005
fIBP=10Mbp; fo=0.5, minBlkSize=100000	43	27.407	27.0	20.0	36.0	39.0	40.0

b

parameters	number of pairs	summary statistics of a distribution based on 1000 permutations					
		mean	median	p5	p95	p99	p99.5
fIBP=1Mbp	47	26.371	26.0	16.95	37.0	42.02	46.005
fIBP=5Mbp	344	210.46	210.0	164.0	262.0	287.0	299.005
fIBP=10Mbp	780	474.064	472.0	380.95	573.1	635.0	650.025

a, The number of pairs of genomic regions originating from a sunflower-specific whole genome duplication where both blocks are associated with flowering time. **b**, Number of pairs of paralogues originating from a sunflower-specific whole genome duplication where both genes are associated with flowering time. The observed number of pairs is indicated, as well as summary statistics of a distribution based on 1,000 permutations of the genomic regions associated with flowering time. fIBP, number of Mb added around the SNPs associated with flowering time to set the limits of the genomic regions associated with flowering time; fo, minimum fraction of blocks overlapping with flowering-time associated regions; minBlkSize, minimum block size.