nature plants

ARTICLES

https://doi.org/10.1038/s41477-018-0329-0

Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance

Sariel Hübner[1,2,3*], Natalia Bercovich[1], Marco Todesco[1], Jennifer R. Mandel[4], Jens Odenheimer[5], Emanuel Ziegler[5], Joon S. Lee[1], Gregory J. Baute[1], Gregory L. Owens[1,6], Christopher J. Grassa[1,7], Daniel P. Ebert[1,8], Katherine L. Ostevik[1,9], Brook T. Moyers[1,10], Sarah Yakimowski[1], Rishi R. Masalia[11], Lexuan Gao[1], Irina Ćalić[11], John E. Bowers[11], Nolan C. Kane[1,12], Dirk Z. H. Swanevelder[13], Timo Kubach[5], Stephane Muños[14], Nicolas B. Langlade[14], John M. Burke[11] and Loren H. Rieseberg[1]

Domesticated plants and animals often display dramatic responses to selection, but the origins of the genetic diversity underlying these responses remain poorly understood. Despite domestication and improvement bottlenecks, the cultivated sunflower remains highly variable genetically, possibly due to hybridization with wild relatives. To characterize genetic diversity in the sunflower and to quantify contributions from wild relatives, we sequenced 287 cultivated lines, 17 Native American landraces and 189 wild accessions representing 11 compatible wild species. Cultivar sequences failing to map to the sunflower reference were assembled de novo for each genotype to determine the gene repertoire, or 'pan-genome', of the cultivated sunflower. Assembled genes were then compared to the wild species to estimate origins. Results indicate that the cultivated sunflower pan-genome comprises 61,205 genes, of which 27% vary across genotypes. Approximately 10% of the cultivated sunflower pan-genome is derived through introgression from wild sunflower species, and 1.5% of genes originated solely through introgression. Gene ontology functional analyses further indicate that genes associated with biotic resistance are over-represented among introgressed regions, an observation consistent with breeding records. Analyses of allelic variation associated with downy mildew resistance provide an example in which such introgressions have contributed to resistance to a globally challenging disease.

Many crop plants exhibit a strong reduction in genetic variation relative to their wild progenitors due to population bottlenecks and strong artificial selection during domestication[1]. Genetic variation has been eroded further over the past century with the transition from traditional varieties adapted to their local environment to uniform elite lines. The reduced diversity of cultivated lines represents a substantial constraint to breeding[2–4], increasing interest in tapping the enormous reservoirs of genetic variation found in crop wild relatives. However, the use of crop wild relatives in breeding can also reintroduce undesired traits that were eliminated during domestication. Development of genomic information for a crop and its wild relatives can illuminate the genetic basis and ancestry of both beneficial and undesirable traits at high resolution[5,6]. Pan-genomes, which capture a broader representation of the genomic variation contained in a gene pool[7,8], represent an especially useful resource for research and breeding. However, developing a pan-genome for crops is challenging owing to the size and complexity of their genomes, so such analyses are often restricted to a few representative genotypes[9–11] or a reduced fraction of the genome[12,13]. Here, we describe the development of a comprehensive pan-genome for sunflower, a globally important oil crop.

Sunflower (*Helianthus annuus* L.) exhibits typical domestication syndrome, including dramatic morphological and ecological differences from its wild ancestor[14,15]. The transition from a wild progenitor (also *H. annuus*) to the cultivated form about 4,000 years ago in North America[16,17], and later to elite lines during the nineteenth and twentieth centuries[14,18,19], progressively narrowed the available genetic variation for breeding[20,21]. Fortunately, sunflower is cross-compatible with many of its wild relatives, which has permitted the introgression of beneficial traits from the wild[20,22–25]. For example, the transition from an open-pollinated crop to hybrid production in the early 1970s[18,26] involved the introgression of (1) a mitochondrial variant from *Helianthus petiolaris* that causes cytoplasmic male sterility to create female (HA) lines[27,28]; (2) a nuclear restorer of fertility (*Rf*) locus from *H. petiolaris* to produce male (RHA) lines; and (3) recessive branching from wild *H. annuus* to extend pollen production of male lines for fertilization[29]. These features distinguish

[1]Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada. [2]Department of Biotechnology, Tel-Hai Academic College, Upper Galilee, Israel. [3]MIGAL—Galilee Research Institute, Kiryat Shmona, Israel. [4]Department of Biological Sciences, University of Memphis, Memphis, TN, USA. [5]SAP SE, Dietmar-Hopp-Allee 16, Walldorf, Germany. [6]Department of Integrative Biology, University of California, Berkeley, CA, USA. [7]Harvard University Herbaria , Cambridge, MA, USA. [8]The Beef Industry Centre, University of New England, Armidale, New South Wales, Australia. [9]Department of Biology , Duke University, Durham, NC, USA. [10]Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO, USA. [11]Department of Plant Biology, Miller Plant Sciences, University of Georgia, Athens, Georgia, USA. [12]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA. [13]Agricultural Research Council, Biotechnology Platform, Private Bag X05, Onderstepoort, South Africa. [14]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France.
*e-mail: sarielh@migal.org.il

between female and male lines in cultivated sunflower and facilitate commercial hybrid production.

Recently, genome assemblies were published for two sunflower cultivars[30], which has allowed us to conduct a comprehensive analysis of genomic diversity in cultivated sunflower and the contributions from wild sunflower relatives. We specifically examine the extent, partitioning and ancestry of genetic diversity captured in cultivated sunflower by exploring a single-nucleotide polymorphism (SNP) data set and pan-genome derived from whole-genome sequencing of 287 lines that make up the cultivated sunflower association mapping (SAM) population[31]. The SAM population is mainly comprised of modern oil and confectionary (non-oil) cultivars, which were released after the transition to hybrid breeding in the early 1970s (Supplementary Fig. 1 and Supplementary Table 10). Next, we use sequence data from an additional 17 Native American landraces and 189 genotypes representing 11 annual and perennial congeners—including species representing the primary (no crossing barriers), secondary (some barriers) and tertiary (special techniques required for breeding) gene pools—to test whether introgression from wild relatives has significantly expanded the gene repertoire of cultivated sunflower. Finally, we ask whether these introgressions have contributed to enhanced disease resistance, as predicted by breeding records, and identify candidate genes contributing to downy mildew resistance.

Results

Allelic diversity and evidence of selection in cultivated sunflower. We sequenced 493 sunflower accessions, including modern cultivars, landraces and wild relatives yielding a total of 11.98 trillion base pairs (bp) of raw sequence (Supplementary Fig. 1). Sequence coverage varied between accessions, with cultivated accessions (modern cultivars and landraces) sequenced to a depth of 5–25× and wild accessions to 1–5× depth. Variant calling was targeted to genic regions based on the HA412-HO.v.1.1 assembly[30], which was available when this study was initiated. HA412-HO is a high-oleic oilseed female line developed by the US Department of Agriculture[32] and is available from the North Central Regional Plant Introduction Station in Ames, IA, USA (PI 642777). Although an improved assembly was recently reported for another genotype (XRQ, a female oil line developed by the Institut National de la Recherche Agronomique (INRA), Paris, France), the two assemblies offer equivalent coverage of the gene space as estimated by BUSCO[33]. A total of 10,522,743 raw variants were detected across all accessions, of which 5,830,734 SNPs were kept after filtering.

Among *H. annuus* accessions, gene diversity declined along the domestication gradient ($He_{wild} = 0.12 \pm 0.16$; $He_{landraces} = 0.08 \pm 0.15$; $He_{cultivars} = 0.06 \pm 0.13$) with significant differences between groups (wild–landraces: $t_{welch} = 17.80$, $P = 1.02 \times 10^{-15}$; wild–cultivars: $t_{welch} = 27.37$, $P < 2 \times 10^{-16}$; landraces–cultivars: $t_{welch} = 14.66$, $P = 2.85 \times 10^{-15}$). The cultivated sunflower data set was further filtered using population-level filters and exclusion of lines with relatively high levels of heterozygosity, leaving a total of 675,291 high-quality SNPs across 239 accessions of the SAM population.

Genome-wide diversity was highly structured in the SAM population and clearly distinguishes market types (oil versus non-oil) and heterotic groups (male versus female) (Supplementary Fig. 2), as previously reported based on a smaller SNP set[34]. Within modern cultivars, diversity was significantly lower in oil varieties than in non-oil varieties ($t_{welch} = 2.80$, $P = 0.01$), and slightly higher (although not significantly) in male than female lines ($t_{welch} = 0.21$, $P = 0.83$).

Reduced nucleotide diversity is expected in regions affected by artificial selection. To identify these regions in the SAM population, the genome was screened at a resolution of 1-Mb sliding windows (Fig. 1 and Supplementary Fig. 3). Overlaps between the lowest 1% quantiles of reduced diversity (π per Mb) and a negative Tajima's D score were observed for eight windows on five chromosomes (Ha1, Ha5, Ha8, Ha13 and Ha14; Supplementary Table 1). Coalescent simulations under a neutral model indicate that overlap between low Tajima's D and low diversity is expected by chance for 0.07% of windows, corresponding to 2.5 windows in the observed data set. Thus, some of the regions putatively targeted by artificial selection may represent false positives. To further support these results, a composite analysis, in which all genome scan statistics were combined into a single measure, was conducted using the Mahalanobis distance with rank-based P values (MD-rank-P) of each raw score (see Supplementary Methods). Overall, results were consistent with the overlapping statistics approach and additional candidate regions on chromosomes Ha4, Ha9, Ha11 and Ha15 were identified by the composite analysis.

To further characterize the putative signal of selection, we tested for differentiation and selection between subgroups in the cultivated gene pool that were bred to enhance functional diversity and hybrid vigour in the case of female and male lines or for different purposes in the case of oil and non-oil types, and compared the results with the genomic scans for selection within groups. A strong signal of differentiation between males and females was observed on chromosomes Ha8, Ha10 and Ha13 (Fig. 1 and Supplementary Fig. 3). These analyses were complemented with subgroup-specific analyses of selection using a composite-likelihood ratio test, which further supported the signal of selection in male (RHA) but not female (HA) lines on chromosomes Ha8, Ha10 and Ha13 (Supplementary Figs. 3 and 4). In addition, a composite analysis (MD-rank-P) for each group was conducted separately using all corresponding statistics including the measure of population differentiation F_{ST} (Supplementary Tables 2 and 3). Both analytical approaches (overlapping statistics and the composite measure of selection) identified a region on chromosome Ha10 where the branching locus was previously mapped[34]. In this region, low diversity was observed in female lines, as well as evidence for a stronger selection signal in male lines as expected (Ha10: 25–26 Mb, highest MD-rank-P score = 1,060). In addition, a strong signal of selection was observed in male lines on chromosome Ha13 (Ha13: 197–198 Mb, second highest MD-rank-P score = 1,043) where the fertility restoration locus was previously identified[35].

We then tested for differentiation and selection between oil and non-oil lines based on overlaps between the lowest 1% quantiles of reduced diversity (π per Mb), a negative Tajima's D score and high F_{ST} at 1-Mb windows (Fig. 1 and Supplementary Fig. 3). These analyses revealed 36 regions with a mean $F_{ST} > 0.26$ compared with the genome-wide average of $F_{ST} = 0.07$. This approach was further complemented with subgroup-specific analysis of selection using a composite-likelihood ratio test, which further supported the signal of selection in oil lines, but not in non-oil lines, on chromosomes Ha2, Ha9, Ha14, Ha15 and Ha17 (Supplementary Fig. 4). In addition, a composite measure approach (MD-rank-P) was conducted across all statistics and identified, among the top 1% quantile, 30 candidate regions in oil lines and 28 regions in non-oil lines as targets of selection (Supplementary Tables 4 and 5). The candidate regions overlap with 11 oil-related quantitative trait loci (QTLs) previously identified on chromosomes Ha1, Ha9, Ha13, Ha14, Ha15 and Ha17 (ref. [30]). Candidate genes, including those encoding plant lipid transfer proteins (Ha2: 205700866 and Ha11: 14452832), were identified in regions with the highest F_{ST} ($F_{ST} > 0.34$) and MD-rank-P values (Supplementary Tables 4 and 5 and Supplementary Fig. 3).

The recombination rate varied widely within chromosomes in all four subgroups (Supplementary Figs. 5 and 6). Overall, a higher median recombination rate (ρ per kb) was observed on chromosomes Ha11 and Ha17 across all four subgroups, whereas chromosome Ha10, which includes the introgression associated with branching in male lines, was characterized by a significantly

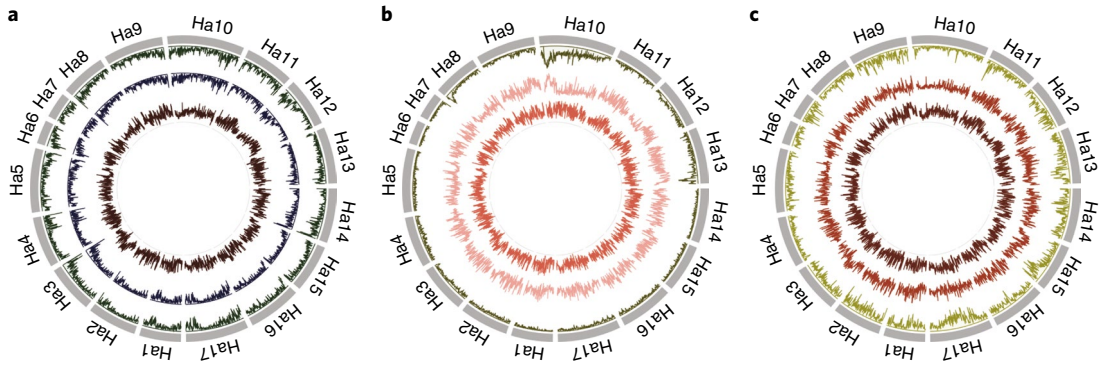

Fig. 1 | Genomic variation in the SAM population based on 675,291 SNPs detected across all lines. Positive values in each track are directed inwards. **a**, Genome scan statistics were calculated in 1-Mb windows across all accessions. Tracks in the circular view (outwards) correspond to Tajima's *D* (brown), π (blue) and SNP density (green). The outer grey blocks in each figure represent chromosomes (same order). **b**, Genome scans for male (RHA) versus female (HA) lines. Tracks correspond to (outwards) Tajima's *D* for female lines (red), Tajima's *D* for male lines (light red) and F_{ST} between male and female lines (green). **c**, Genome scans for oil versus non-oil lines. Tracks correspond to (outwards) Tajima's *D* for oil lines (brown), Tajima's *D* for non-oil lines (red) and F_{ST} between oil and non-oil lines (green).

($F_{d.f.=3}=922$, $P<0.0001$) higher recombination rate among female lines ($HA_{median}=0.36$) than among other subgroups ($RHA_{median}=0.18$; $oil_{median}=0.20$; $non\text{-}oil_{median}=0.11$) in the collection.

Development of the cultivated sunflower pan-genome. Genes in the cultivated gene pool that are not represented in the HA412-HO.v1.1 genome assembly were identified by de novo assembly of reads that did not map to the reference sequence. Each of the 287 cultivated accessions (cultivated lines and landraces) was assembled independently. Altogether, 420,624 sequences with a minimum length of 200 bp were assembled across all accessions with an N50 of 986 bp. After a clustering step to remove redundancy, 114,164 sequences were obtained with an N50 of 1,323 bp. This set of sequences was aligned to the HA412-HO reference assembly to remove sequences that were already present but had previously failed to align. A total of 17,061 de novo-assembled genes were absent from the reference genome and passed all filtering steps (see Supplementary Methods). Annotating this set of genes using the annotations available for the HA412-HO reference genome and the plant protein database indicated that 5% were complete genes (that is, cover >90% of the predicted protein length). When combined with previously annotated genes in the reference genome, 61,205 genes were obtained, which comprise the cultivated sunflower pan-genome. However, this is probably an underestimate of the true pan-genome owing to the possible combining of close paralogues. Indeed, our approach estimates 47,848 genes in the XRQ reference rather than the 52,232 protein-coding genes reported by Badouin et al.[30].

Next, sequences from all cultivated accessions were mapped to the ensemble of genes represented in the pan-genome. After filtering hits with low alignment scores (bit-score < 200, alignment length < 150 bp), low-confidence or redundant annotations and genes encoding unknown proteins, a total of 45,302 genes were kept, of which 2,700 are de novo-assembled genes (Fig. 2).

We used these 45,302 well-characterized genes to explore the presence or absence of variation across the cultivated gene pool (see Supplementary Methods). The majority of genes (32,917 (72.7%)) represent so-called core genes that occur in >95% of the accessions in the cultivated gene pool. The remaining genes seem to be dispensable, with 2,464 genes (5.4%) found in <5% of the accessions. Introgressed genes are expected to enrich the rare gene fraction among dispensable genes, resulting in a bimodal frequency distribution (Fig. 2). Gene expression was evident for 70% of dispensable genes based on available transcription data[36], suggesting that most dispensable genes are functional.

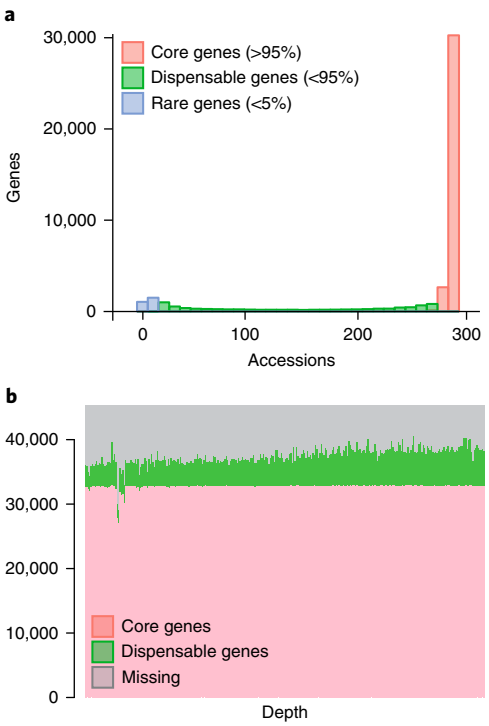

Fig. 2 | The cultivated sunflower pan-genome. a, Histogram representing the abundance of genes in the pan-genome divided into three bins: core genes found across >95% of the accessions, dispensable genes and rare genes found in <5% of the accessions. **b**, Bar plot representing the number of core genes, dispensable genes and missing genes per cultivated accession detected in relation to the estimated sequencing depth.

To test how well the pan-genome represents the gene presence–absence diversity across the cultivated sunflower gene pool, a saturation analysis was conducted. Results indicate saturation of gene accumulation after about 60% of accessions are included in the analysis, although rare genes continue to be added at a slower rate as more accessions are added (Supplementary Fig. 9).

Identifying wild introgressions and their potential functional role. To identify potential introgressed genes identified in each

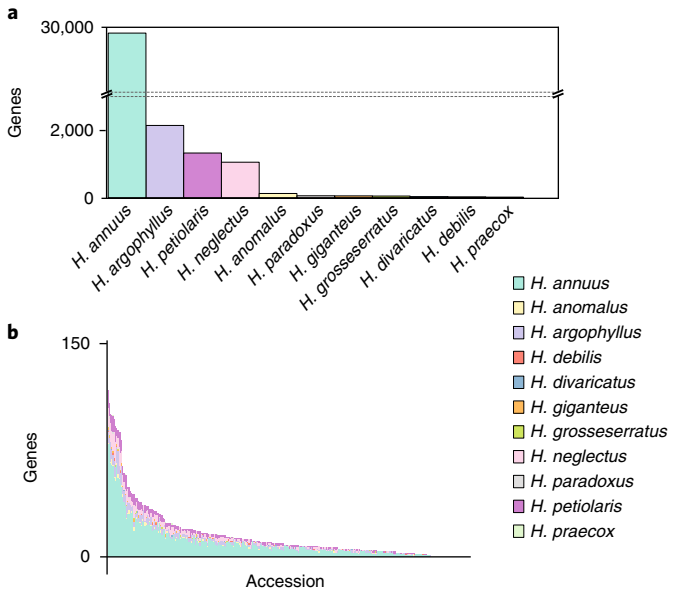

Fig. 3 | Average number of genes in the cultivated sunflower pan-genome assigned to the wild source. a, Bar plot for the overall number of genes contributed by each wild species. **b**, Bar plot for the number of genes identified as introgressions in modern cultivated accessions. Colours correspond to the wild source of introgression. Colours for both plots are indicated in the legend.

cultivated accession and assign them to a wild donor, a comprehensive comparison of gene sequences between each of the 17 Native American landraces, 189 wild accessions and 270 accessions of the cultivated SAM population was conducted for the 45,302 genes using BLAST. After filtering hits, results were summarized in a matrix, which allowed us to determine the likely ancestry of each gene in the cultivated sunflower pan-genome. Although the number of individuals sequenced varied across the putative donor species (Supplementary Figs. 1 and 2), introgression assignments accord closely with expectations based on known genetic relationships and breeding records rather than with sample size. As anticipated, the majority of hits (66.3%) corresponded to the primary gene pool, wild *H. annuus*, followed by *Helianthus argophyllus* (sister to *H. annuus*), *H. petiolaris* and *Helianthus neglectus* in the secondary gene pool, with 4.7%, 2.9% and 2.3% of the hits, respectively (Fig. 3a). Hits from each of the remaining eight species were found to be <1%. In all, 10.6% of the pan-genome seems to derive from the secondary gene pool. Wide variation was observed between cultivated lines for introgression intensity due to the breeding history. For example, the accession SAM207 (PI_619204, RHA 419) was previously reported[37] as being highly introgressed due to crossing with *H. argophyllus* and was shown to harbour a large number of *H. argophyllus* introgressions by gene assignment analysis (Supplementary Table 11). However, these results should be viewed with caution as the BLAST approach does not distinguish between the retention of ancestral polymorphism and introgression[38]. Thus, some false positives are likely.

To determine whether the gene repertoire of cultivated sunflower has expanded owing to introgression, we asked whether any of the ~45,000 genes validated across the cultivated gene pool were absent in Native American landraces from which modern cultivars arose, but present in one or more wild species. A total of 636 such genes were found across the cultivated gene pool (Fig. 3b) and further investigated for their genomic location and possible functional role(s) in sunflower improvement (Supplementary Figs. 10–12). Overall, introgressed genes from wild species were dispersed across all chromosomes. Among the genes identified as introgressions, 15 were de novo-assembled genes not found in the reference genome assembly HA412-HO.v1.1. Despite their absence from the reference genome, linkage disequilibrium between these genes and neighbouring loci could potentially indicate their genomic physical position. Thus, we tested for linkage disequilibrium between SNPs in the SAM population and the presence or absence of the de novo-assembled genes (Supplementary Fig. 11). Significant associations ($P < 1.5 \times 10^{-7}$) were identified for 14 of the 15 genes. Likely genomic locations (based on the strongest association) are provided in Supplementary Table 8.

To validate these results, we analysed admixture in the subset of 239 highly inbred SAM lines versus three wild species (29 *H. annuus* accessions, 29 *H. petiolaris* accessions and 10 *H. argophyllus* accessions) for which we had sufficient sequence data to implement PCAdmix[39], a principal component analysis-based approach for inferring patterns of mixed ancestry along chromosomes. The PCAdmix analysis indicated that approximately 9% of the cultivated gene pool derives from introgression with the secondary gene pool, which is similar to that estimated from the more comprehensive BLAST comparison (above), and introgressed regions detected by the two approaches largely overlap (81%). Conversely, some introgressed regions that the BLAST analysis assigned to other donor species are mistakenly inferred as derived from *H. argophyllus*, *H. petiolaris* or *H. annuus* by PCAdmix owing to restriction of the analysis to these three species, illustrating the value of including all possible donors in admixture analyses.

Next, gene ontology (GO) terms were determined for the 636 genes to infer their potential functional role in sunflower improvement. Among the 140 biological process categories found, 25 categories were significantly enriched (Supplementary Fig. 12). These include categories related to biotic stress response, such as response to biotic stimulus (Fisher's exact test = 10.953, $P = 0.003$), defence response (Fisher's exact test = 4.764, $P = 0.028$) and chitinase activity (Fisher's exact test = 10.737, $P = 0.017$). These results are consistent with reports from sunflower breeding programmes that crosses involving wild species have been most commonly employed to introduce disease resistance genes into modern sunflower cultivars[25]. As expected, if introgressions were driven by artificial selection, the average recombination rate (ρ) was reduced by 64% for introgressed genes relative to all other genes (note that recombination rates could not be tested for 134 of 636 introgressed genes because of insufficient SNP density).

Identification of resistance genes to downy mildew. To further test whether introgression from wild relatives has contributed to improved disease resistance in modern cultivars, we conducted a genome-wide association study (GWAS) for downy mildew resistance using the SAM population (Supplementary Fig. 13). Seedlings from all lines in the SAM population were inoculated with downy mildew spores and the level of susceptibility was recorded (see Supplementary Methods). To identify genomic regions associated with downy mildew resistance, we searched for associations in the SNP data set using a GWAS framework (Fig. 4). The mixed linear model used in the analysis included correction for population structure (Supplementary Fig. 7) using the first four eigenvectors as covariates, and the kinship matrix between accessions as a random effect (Supplementary Fig. 14). Inflation of *P* values was well controlled ($\lambda = 1.003$), and an adjustment of the significance threshold to account for multiple comparisons was set at 2.9×10^{-7} by simpleM and 1.5×10^{-7} by permutation tests.

Seven significant associations with downy mildew resistance were found (Fig. 4 and Supplementary Table 9), four of which passed both significance thresholds ($<1.5 \times 10^{-7}$) and three that passed only the simpleM threshold ($<2.9 \times 10^{-7}$). Note that there are two independent associations on Ha13 and three on Ha17 (Supplementary Table 9). Candidate genes underlying significant associations with

downy mildew resistance are listed in Supplementary Table 9 and include, for example, a Tify 9-like protein gene (*TIFY9*) and a syntaxin gene (*SYP132*). Other SNP associations did not pass the significance thresholds corrected for multiple comparisons, although a strong signal coupled with elevated linkage disequilibrium was observed, suggesting that the multiple test correction may be overly conservative. These associations and underlying genes are reported in the Supplementary Information, as they may be of interest to other researchers and sunflower breeders (Supplementary Table 9).

Discussion

Genetic diversity is required for populations to respond to natural and/or artificial selection. Until recently, responses to selection were thought to result exclusively from changes in frequencies of sequence variants. However, it is now recognized that closely related individuals often vary in other genomic features that may also underlie selective responses[9–11]. Sunflower is armed with a substantial amount of genetic variation despite domestication and improvement bottlenecks in the past 4,000 years[14,16–19]. This may be a consequence of its broad geographical range, large effective population size, outcrossing mating system and, as discussed below, the widespread use of hybridization and introgression during improvement[23,26]. Here, we provide a new perspective on genetic diversity in cultivated sunflower by analysing sequence polymorphisms (SNPs) and gene content variation (that is, the pan-genome) across the cultivated gene pool.

Analyses of the SNP data recovered the four main categories of sunflower cultivars (male oil, male non-oil, female oil and female non-oil), as expected. In addition, using various genome scan approaches, we were able to identify regions of the genome that were affected by artificial selection across all lines, as well regions that have been subjected to divergent selection between groups. For example, as expected, male and female lines are most differentiated at the branching locus on chromosome Ha10 and the male fertility restoration locus on chromosome Ha13, which are the two loci required to implement hybrid breeding programmes[26].

Analyses of the SNP data also permitted inferences about the phylogenetic origins of most (80%) of the sequence diversity in the modern cultivated lines that make up the SAM population. Approximately two-thirds of gene sequences in the SAM population seem to derive from the primary gene pool (wild *H. annuus*) and 10.6% from the secondary gene pool (other wild species that form partially fertile hybrids with cultivated sunflower). Previous analyses of transcriptome sequences estimated that, on average, introgressions cover 10% of modern cultivar genomes[26], which is similar to the estimates provided here. Dissection of expression records previously reported by Badouin et al.[30] for the XRQ genotype (Supplementary Information) indicates that introgressed genes have higher expression levels in ovaries than non-introgressed genes, but the opposite pattern was observed in leaves and seeds, and no differences were seen for eight other tissues (Supplementary Fig. 17b). Thus, introgression from wild species has contributed importantly to the sequence and functional diversity found in sunflower.

We also found considerable variation in gene content across cultivated sunflower genomes, with more than one-fourth of the pan-genome comprising 'dispensable' genes. These results are comparable with the *Brassica oleracea* and *Brachypodium distachyon* pan-genomes in which 20% and 27% of genes were identified as dispensable, respectively[5,40]. Although the expression profiles of dispensable genes were not tested explicitly here, over two-thirds of dispensable genes had expression records reported by Badouin et al.[30]. Possibly, the proportion of dispensable genes is overestimated due to low sequencing depth for some samples. However, we suspect that including non-coding regions in the analysis would further increase the dispensable fraction of the genome, because non-coding sequences are less likely to be maintained by selection.

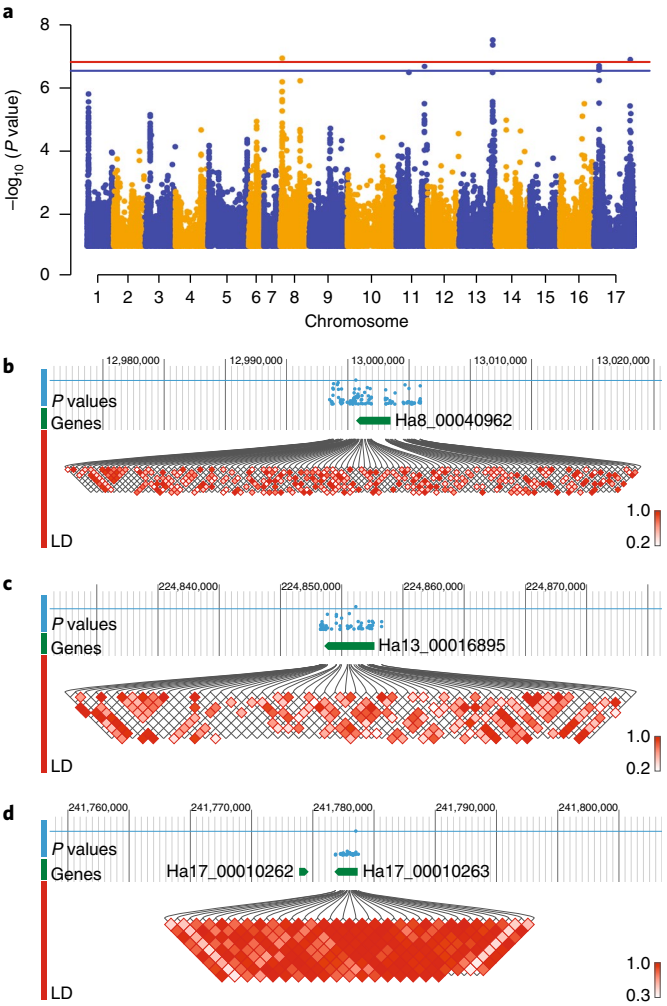

Fig. 4 | Genome-wide association mapping of downy mildew resistance in the SAM population. a, Manhattan plot for genome-wide association of downy mildew resistance using the 675,291 SNP data set. The corrected significance threshold at 5% using the simpleM algorithm is indicated by the horizontal blue line and the permutation-based threshold is indicated by the horizontal red line. **b**–**d**, Zoomed-in view on genomic regions (25 kb) where the three most significant hits, based on the association test conducted with EMMAX, were identified on chromosomes Ha8 (**b**), Ha13 (**c**) and Ha17 (**d**). Each plot includes genomic coordinates, the Manhattan plot in which the blue horizontal line corresponds to the simpleM significance threshold, the location of genes in the region (green) and the level of linkage disequilibrium (LD) between SNPs (colour scale on the right).

In addition, ~5% of the genes in the pan-genome were classified as unique; that is, found in <5% of the accessions. Further analysis of the pan-genome revealed that 636 genes (~1.5%) arose exclusively via introgressions from wild relatives during improvement. Thus, it is clear that introgression has increased both the SNP diversity and the gene repertoire of cultivated sunflower.

The sunflower breeding literature indicates that sunflower wild relatives are frequently tapped by breeders to increase tolerance to various environmental stresses, especially common sunflower diseases, such as downy mildew, white mould and rust, and so on[25]. Gene annotations and the associated ontologies are consistent with the focus of breeders on disease resistance, in that genes associated with biotic stress are over-represented among the

introgressed gene fraction. These findings are similar to earlier reports in maize, in which introgressions from wild relatives were associated with changes in pigmentation and trichome density that are thought to improve tolerance to highland conditions[41]. Likewise, in *Brachypodium* defence-related genes were enriched among the dispensable fraction of the pan-genome[40].

Our high-resolution genomic data for the SAM population are also expected to yield new and more precise information regarding the genetic basis of traits of interest, including resistance to both biotic and abiotic stress. Owing to the conservative approach taken to correct for multiple comparisons, only seven QTLs passed one or both significance thresholds, although additional suggestive QTLs were also observed. In the present paper, for example, GWAS detected associations for downy mildew resistance on chromosomes Ha1, Ha6, Ha8 and Ha13 that had been previously identified in biparental populations[42], as well as new associations on chromosomes Ha3, Ha11, Ha16 and Ha17. Several associations overlap with introgressions from wild species, including the Ha11 and Ha16 QTLs, which seem to be derived from wild *H. annuus* and *H. neglectus*, respectively. Within genomic regions associated with downy mildew resistance, several putative pathogen defence genes were identified, including a disease resistance gene in the NBS-LRR family on chromosome Ha3, a pathogen-responsive signalling gene (encoding lectin receptor-like kinase[43]) on chromosome Ha6, genes encoding a B-box zinc finger[44], Tify 9-like protein[36], AHBP-1B[45] and a mitogen-activated protein kinase[46] on chromosome Ha8, and a gene encoding serine/threonine-protein kinase[47] on chromosome Ha13. In addition, the direct defence genes encoding the syntaxin SYP132 (ref. [48]) and GDSL-motif lipase[49] were detected on chromosomes Ha11 and Ha16, respectively. Although the association between these genes and downy mildew resistance was determined based on SNPs called within them, we cannot rule out the possibility that closely linked regions are causative because of linkage disequilibrium in the SAM population.

Our results provide further support for the contribution of introgressions from wild species to disease resistance in sunflower and add to a growing literature on the value of primary and secondary wild germplasm in crop improvement[41,50–52]. We further show that such introgressions not only enhance allelic diversity but they also increase the total number of genes in a crop pan-genome. Thus, the gene content of an introgression is not necessarily predictable from a reference sequence.

Despite its clear utility to understanding crop genetic variation, assembling a pan-genome is still challenging analytically. The conservative approach and stringent homology cut-offs used in the present study probably reduced the comprehensiveness of the assembled sunflower pan-genome. Likewise, the low sequencing depth available for some genotypes no doubt limited both the number and the completeness of genes identified. To more fully represent the genes in the pan-genome, as well as to capture presence–absence variation in non-coding regions and repetitive sequences, it will be necessary to increase sequencing depth and refine the homology search strategy[53]. Advances in long-read sequencing have the potential to alleviate some of the challenges associated with pan-genome assembly, especially the sensitivity and specificity of structural variation detection[54]. Given the ongoing advances in sequencing technology, we suspect that a pan-genome approach will become standard in comparative genomics studies, facilitating a more complete view of the genetic variation represented in a gene pool, especially novel genomic components associated with traits of interest.

Methods

Resequencing the SAM population and wild relatives. Sunflower germplasm collections comprise ~40,000 cultivated and wild accessions globally and served as the basis for selecting accessions in this study. Altogether, 493 accessions were chosen, including 287 accessions from the cultivated SAM population, which captures ~90% of the allelic diversity in the cultivated sunflower gene pool[31] (Supplementary Fig. 1 and Supplementary Table 10). All lines can be obtained from the USDA's National Plant Germplasm System and/or the INRA. Most of the accessions in the SAM population are modern cultivars that were released after the transition to hybrid breeding in the early 1970s and include 51 female confectionary lines, 64 female oil lines, 25 male confectionary lines and 66 male oil lines. However, the population also includes a small number of open-pollinated varieties and a few lines that are not easily classified (Supplementary Fig. 1 and Supplementary Table 10). In addition to the SAM population, 17 Native American landraces and 189 wild accessions were chosen for sequencing, including the following 11 annual and perennial congeners: *H. annuus* from the primary gene pool; *H. petiolaris*, *H. neglectus*, *H. argophyllus*, *H. anomalus*, *H. debilis*, *H. paradoxus* and *H. praecox* from the secondary gene pool; and *H. divaricatus*, *H. grosseserratus* and *H. giganteus* representing the tertiary gene pool (Supplementary Table 10 and Supplementary Figs. 1 and 2). Genomic DNA was extracted from each accession as described in Mandel et al.[34], libraries were prepared following Illumina's TruSeq protocol and sequencing was conducted on the Illumina HiSeq platform with 100-bp paired-end reads. Sequence coverage varied between accessions, in which cultivated accessions (modern cultivars and landraces) were sequenced to a coverage of 5–25× and wild accessions were sequenced to 1–5×.

Calling SNPs across all samples. Raw sequence data generated for each accession in the SAM population were processed and cleaned using Trimmomatic v.036 (ref. [55]) and aligned to the HA412-HO.v.1.1 reference assembly. Alignment of clean reads was conducted with an aligner developed by SAP SE[56], which is optimized to use available main memory for faster indexing, minimize the cache miss ratio to improve performance and optimize parallel code execution. Alignment files were processed to remove potential PCR duplicates using picard v.2.5 (ref. [57]). To reduce the computational intensity of the variant calling process, we targeted genic regions and a 2.5-kb flanking sequence. Before variant calling, low-quality alignments and reads mapping to highly repetitive regions were removed.

Variants including SNPs and insertions or deletions (indels) were called across all 493 accessions in one batch using a haplotype-sensitive algorithm implemented in the open source software FreeBayes[58]. The cultivated gene pool was further filtered to include only highly inbred accessions from the SAM population and to remove variants with a quality of<30, >30% missing data, minor allele frequency of <5%, minimum genotype depth of 1 read and maximum of 30, maximum observed heterozygosity of 10%, and strand-biased or direction-biased alleles. The sum of depth of coverage across all accessions was set between 1,000 and 2,300 to further remove outliers in the depth distribution. In addition, SNPs found close to a gap (<5 bp) and indels were removed from the data set using vcffilter[59] and vcftools[60].

Following this set of filters, a panel of high-quality and trustable SNPs across 239 highly inbred accessions from the SAM population was obtained for the GWAS analysis and genome scans. Genomic scans were calculated in 1-Mb sliding windows and included nucleotide diversity (π), Tajima's D, SNP density, selective sweeps and recombination rate (see Supplementary Methods). These genomic scans were applied to the SAM population data set and subgroups corresponding to oil versus non-oil varieties, male (RHA) and female (HA) lines, representing the four major breeding types in sunflower. In addition, genomic scans for differentiation (Weir–Cockerham F_{ST}) were conducted between subgroups (that is, male versus female and oil versus non-oil).

Assembly and annotation of the cultivated sunflower pan-genome. The HA412-HO.v1.1 (ref. [30]) reference sequence and annotations were used to guide the assembly of the cultivated sunflower pan-genome using a conservative approach[12]. Following the alignment of reads from each accession in the cultivated gene pool to the reference, unmapped and poorly mapped reads (defined as an edit distance of ≥8 for a read pair) were extracted. These reads were assembled de novo for 270 accessions that were well classified and characterized from the SAM population and an additional 17 landraces independently using the Ray assembler[61] with a range of k-mers between 13 and 51 to enable the assembly of contigs from low-coverage data. All sequences were blasted against the UniVec database[62] to remove potential contaminants. Assembled contigs shorter than 200 bp were removed from further analysis. Remaining contigs were aligned to the reference genome to identify sequences that were reassembled and are already present in the reference. Contigs with >75% similarity along >75% of the alignment length were considered as represented in the reference genome and were excluded from further analysis. These steps were executed for each accession in the cultivated gene pool separately. Next, all contigs from all accessions were pooled into one data set that represents all dispensable sequences not found in the reference genome. To cluster overlapping sequences and avoid redundancy, the pooled data set was processed using the software CD-HIT[63] with a similarity threshold of 95%, keeping the longest contig at each cluster. Remaining sequences were searched against the NCBI nr database to look for potential assembled contaminants. A total of 526 sequences were identified as potential contaminants and excluded from further analysis. To annotate the non-redundant sequences from the pooled data set, annotations of the HA412-HO reference genome were integrated with the plant protein database[64] to create a more complete annotation database. Annotations for the non-redundant de novo-assembled sequences were determined using blastx

with a minimum bit-score threshold of 200 to ensure a high quality of annotations. Only the best hit was kept while giving priority to the HA412-HO hits associated with each query. The final outcome of this process was a set of annotated genes that corresponded to the cultivated sunflower pan-genome, including core and dispensable genes.

To determine the full genetic repertoire for each of the cultivated accessions, raw reads from each accession were aligned to the assembled pan-genome. Genes that were not annotated with high confidence were removed from further analysis and overlapping annotations were merged based on protein identity, keeping the longest sequence for each gene. To obtain a uniform and comparable similarity score for each gene in each accession, alignments were assembled to contigs using the mpileup command in SAMtools[65] followed by the seq command in seqtk[66] to convert the fastq file to fasta. Recovered sequences were aligned to the pan-genome using blastn and filtered at a minimum bit-score of 200, minimum alignment length of 150 bp and minimum identity of 75%. For each sequence, we verified that the alignment and pan-genome identities matched, and the corresponding bit-score of the best hit was kept as a measure of similarity between a gene in each accession to the pan-genome sequences.

Localizing the genomic position of the unmapped de novo-assembled genes. To localize the genomic position of de novo-assembled genes that are absent from the reference genome, we took advantage of linkage disequilibrium between these genes and SNPs of known genomic location in the SAM population. First, the identity matrix generated for the pan-genome was transformed into a presence–absence matrix, in which a similarity score was considered as presence (1) and a missing score as absence (0). Unmapped, de novo-assembled genes were extracted from the presence–absence matrix and were associated with SNPs in a genome-wide association framework. To associate SNPs with the presence or absence of genes, we used the software EMMAX[67] with the genes presence–absence score as 'phenotypes', SNP calls as genotypes and kinship among the cultivated lines calculated as identity by state as a random effect in the mixed linear model. To correct the false-positive rate due to multiple comparisons, the simpleM algorithm[68] was used at a significance level of 5%.

Identifying the wild source of introgressed genes in the pan-genome. Two different approaches were used to identify the potential wild source of genes comprising the cultivated sunflower pan-genome. In the first approach, genes of each of the 287 cultivated accessions (270 cultivated lines and 17 landraces) were compared to whole-genome sequences generated for the 189 wild accessions representing 11 congeneric species. Comparisons were conducted following the same procedure described above for the pan-genome in the cultivated gene pool. Briefly, raw reads from wild accessions were cleaned, trimmed and aligned to the cultivated sunflower pan-genome. A consensus sequence was called for each wild accession alignment separately. A comprehensive comparison of sequences from all wild accessions to all cultivated accessions was performed with blastn and results were filtered using a minimum bit-score of 200, a minimum alignment length of 150 bp and a minimum of 75% identity. Of the remaining hits, only the best hit was kept as a potential source of introgression. The pairwise identity matrix was further filtered to remove wild introgressions supported by only one cultivated accession. To distinguish gene sequences brought into the cultivated gene pool by domestication versus those that were introgressed during improvement, landrace accessions were used as a reference. All gene sequences found in landraces were considered as derived from domestication and were removed from downstream analyses.

As a validation approach, we used the SNP data set called across all accessions to identify introgressions from *H. argophyllus* and *H. petiolaris*, with wild *H. annuus* included as the most probable source of variation. Analyses were conducted using the software PCAdmix[39]. PCAdmix is an algorithm designed to infer local ancestry by classifying and projecting segments of SNPs in the admixed genomes onto the basis of ancestral individuals and smoothing the signal using the Viterbi algorithm within an HMM framework. Analysis was restricted to three wild species and the SAM population as the admixed individuals because increasing the number of wild species populations reduced the overall number of overlapping SNPs, causing a major reduction in the number of segments that could be considered in the analysis. Prior to the PCAdmix analysis, variants were filtered to remove low-quality SNPs ($Q < 30$), strand-biased or direction-biased alleles, SNPs close to a gap (<5 bp) and indels. In addition, all SNPs were phased in Beagle v4.1 (ref. [69]) for each species separately. No pruning was set in the analysis, and recombination rates were inferred within the PCAdmix analysis. Assignments of segments to wild donors were converted to bed format and plotted using the Sushi package[70] in R.

GOs. To elucidate the possible functional role of the introgressed genes from wild congeners, we used the GO annotations available for the HA412-HO genome for genes found in the reference genome, and the GOA database[71] for de novo-assembled genes. GO biological processes were clustered and visualized using the web-server REVIGO[72]. REVIGO's clustering algorithm finds a single representative GO term for clusters of semantically similar GO terms, thus resulting in a reduced, non-redundant GO term set (that is, superclusters). The size of each supercluster reflects the term group abundance. Fisher's exact test was implemented in R to test for enrichment of GO terms among introgressed genes compared to the pan-genome as a whole.

Downy mildew resistance experiment. *Plasmopara halstedii* race 734 (accession 1916 from Gnadenthal, Manitoba) was provided by R. Khalid (AgCanada, Morden, Manitoba, Canada). The isolate was propagated in susceptible cultivar HA89 for further infections. Seeds from the SAM population were sterilized, germinated and infected using the whole-immersion method[73–75]. Seeds were scarified and dehulled to prevent contamination or concomitant infections and also to synchronize their germination speed. Four-day-old healthy seedlings, with a developed root of about 1–2-cm long, were infected with fresh *P. halstedii* suspension from race 734 at a concentration of 20,000 zoosporangia per ml and were eventually transferred to soil, watered daily and grown in a growth chamber under controlled conditions (19 °C, 16 h of light) for 10 days. Finally, plants were incubated in 100% relative humidity for 24 h in darkness and scored for susceptibility.

This experiment was conducted in a randomized block design with three replicates. Each block included a treated plant and a control for each accession. As a measure of each plant's response to inoculation with downy mildew, the following parameters were scored: (1) the percentage of the cotyledon area covered by spores; (2) spore density in a single cotyledon measured as the total number of sporangia normalized to the cotyledon area; and (3) the percentage of leaf area covered by spores. Finally, resistant plants were considered as those showing no sign of infection.

Genome-wide association mapping of downy mildew resistance. GWAS was performed to identify SNPs associated with downy mildew resistance in the SAM population. For the resistance phenotype, we used the quadratic normalization of the per cent spore density, which provided a robust quantitative estimate of resistance/susceptibility. GWAS was performed using the software EMMAX[67] with genotype as a fixed effect, and the kinship matrix between accessions calculated as identity by state was included as a random effect in a mixed linear model. To correct for population stratification, principal component analysis was computed from the genotypic data using the SNPRelate package[76] in R, and the first four eigenvectors were used as covariates in the association mapping model. False-positive inflation due to multiple comparisons was corrected with the simpleM algorithm and 500,000 permutations, both at a genome-wide significance level of 5%.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All raw sequence data are stored in the Sequence Read Archive (SRA) under Bioproject PRJNA353001 for cultivars and PRJNA397453 for wild and landrace. Accession numbers for each sample are listed in Supplementary Table 10. The SNP data set in vcf format, pan-genome sequences in fasta format and genome scan statistics in bed files format can be downloaded from https://www.sunflowergenome.org/.

Received: 13 July 2017; Accepted: 15 November 2018;
Published online: 31 December 2018

References

1. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
2. Harlan, J. R. *Crops and Man* 2nd edn (American Society of Agronomy, Madison, 1992).
3. Ladizinsky, G. *Plant Evolution Under Domestication* (Springer, Dordrecht, 1998).
4. Galluzzi, G., Van Duijvendijk, C., Collette, L. Azzu, N. & Hodgkin, T. *Biodiversity for Food and Agriculture. Contributing to Food Security and Sustainability in a Changing World* (FAO, 2011).
5. Golicz, A. A., Batley, J. & Edwards, D. Towards plant pangenomics. *Plant Biotechnol. J.* **14**, 1099–1105 (2016).
6. Barabaschi, D. et al. Next generation breeding. *Plant Sci.* **242**, 3–13 (2016).
7. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
8. Medini, D. et al. Microbiology in the post-genomic era. *Nat. Rev. Microbiol.* **6**, 419–430 (2008).
9. Gan, X. et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
10. Li, Y. H. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
11. Lin, K. et al. Beyond genomic variation—comparison and functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics* **15**, 250 (2014).
12. Hirsch, C. N. et al. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**, 121–135 (2014).

13. Lu, F. et al. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* **6**, 6914 (2015).
14. Burke, J. M., Tang, S., Knapp, S. J. & Rieseberg, L. H. Genetic analysis of sunflower domestication. *Genetics* **161**, 1257–1267 (2002).
15. Burke, J. M., Knapp, S. J. & Rieseberg, L. H. Genetic consequences of selection during the evolution of cultivated sunflower. *Genetics* **171**, 1933–1940 (2005).
16. Harter, A. V. et al. Origin of extant domesticated sunflowers in eastern North America. *Nature* **430**, 201–205 (2004).
17. Smith, B. D. Eastern North America as an independent center of plant domestication. *Proc. Natl Acad. Sci. USA* **103**, 12223–12228 (2006).
18. Korell, M., Mosges, G. & Friedt, W. Construction of a sunflower pedigree map. *Helia* **15**, 7–16 (1992).
19. Putt, E. D. in *Sunflower Technology and Production* Vol. 35 (ed. Schneiter A. A.) 1–19 (American Society of Agronomy, Madison, 1997).
20. Rauf, S. Breeding sunflower (*Helianthus annuus* L.) for drought tolerance. *CBCS* **3**, 29–44 (2008).
21. Mayrose, M., Kane, N. C., Mayrose, I., Dlugosch, K. M. & Rieseberg, L. H. Increased growth in sunflower correlates with reduced defences and altered gene expression in response to biotic and abiotic stress. *Mol. Ecol.* **20**, 4683–4694 (2011).
22. Seiler, G. J. Utilization of wild sunflower species for the improvement of cultivated sunflower. *Field Crops Res.* **30**, 195–230 (1992).
23. Seiler, G. J., Qi, L. L. & Marek, L. F. Utilization of sunflower crop wild relatives for cultivated sunflower improvement. *Crop Sci.* **57**, 1083–1101 (2017).
24. Ma, G. J., Markell, S. G., Song, Q. J. & Qi, L. L. Genotyping-by-sequencing targeting of a novel downy mildew resistance gene Pl_{20} from wild *Helianthus argophyllus* for sunflower (*Helianthus annuus* L.). *Theor. Appl. Genet.* **30**, 1519–1529 (2017).
25. Dempewolf, H. et al. Past and future use of wild relatives in crop breeding. *Crop Sci.* **57**, 1070–1082 (2017).
26. 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).
27. Leclercq, P. Cytoplasmic male sterility in sunflower. *Ann. Amelior. Plant* **19**, 99–106 (1969).
28. Kinman, M. L. New developments in the USDA and state experiment station sunflower breeding programs. In *Proc. 4th Int. Sunflower Conference* 181–183 (International Sunflower Association, 1970).
29. Miller, J. F. & Fick, G. N. in *Sunflower Technology and Production* Vol. 35 (ed. Schneiter A. A.) 441–496 (American Society of Agronomy, Madison, 1997).
30. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
31. Mandel, J. R., Dechaine, J. M., Marek, L. F. & Burke, J. M. Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theor. Appl. Genet.* **123**, 693–704 (2011).
32. Miller, J. F., Gulya, T. J. & Vick, B. A. Registration of three maintainer (HA 456, HA 457, and HA 412 HO) high-oleic oilseed sunflower germplasms. *Crop Sci.* **46**, 2728 (2006).
33. 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).
34. Mandel, J. R. et al. Association mapping and the genomic consequences of selection in sunflower. *PLoS Genet.* **9**, e1003378 (2013).
35. Baute, G. J. *Genomics of Sunflower Improvement: From Wild Relatives to a Global Oil Seed*. Dissertation, Univ. British Columbia (2015).
36. Chung, H. S. & Howe, G. A. A critical role for the TIFY motif in repression of jasmonate signaling by a stabilized splice variant of the JASMONATE ZIM-domain protein JAZ10 in *Arabidopsis*. *Plant Cell* **21**, 131–145 (2009).
37. Miller, J. F., Gulya, T. J. & Seiler, G. J. Registration of five fertility restorer sunflower germplasms. *Crop Sci.* **42**, 989 (2002).
38. Kane, N. C. et al. Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* **63**, 2061–2075 (2009).
39. Brisbin, A. et al. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* **84**, 343–364 (2013).
40. Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
41. Hufford, M. B. et al. The genomic signature of crop-wild introgression in maize. *PLoS Genet.* **9**, e1003477 (2013).
42. Jocic, S. et al. Towards sustainable downy mildew resistance in sunflower. *Helia* **35**, 61–72 (2012).
43. Vaid, N., Macovei, A. & Tuteja, N. Knights in action: lectin receptor-like kinases in plant development and stress responses. *Mol. Plant* **6**, 1405–1418 (2013).
44. Gupta, S. K., Rai, A. K., Kanwar, S. S. & Sharma, T. R. Comparative analysis of zinc finger proteins involved in plant disease resistance. *PLoS ONE* **7**, e42578 (2012).
45. Zhang, Y., Fan, W., Kinkema, M., Li, X. & Dong, X. Interaction of NPR1 with basic leucine zipper protein transcription factors that bind sequences required for salicylic acid induction of the *PR-1* gene. *Proc. Natl Acad. Sci. USA* **96**, 6523–6528 (1999).
46. Asai, T. et al. MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature* **415**, 977–983 (2002).
47. Romeis, T. Protein kinases in the plant defence response. *Curr. Opin. Plant Biol.* **4**, 407–414 (2001).
48. Kalde, M., Nuhse, T. S., Findlay, K. & Peck, S. C. The syntaxin SYP132 contributes to plant resistance against bacteria and secretion of pathogenesis-related protein 1. *Proc. Natl Acad. Sci. USA* **104**, 11850–11855 (2007).
49. Oh, I. S. et al. Secretome analysis reveals an *Arabidopsis* lipase involved in defense against *Alternaria brassicicola*. *Plant Cell* **17**, 2832–2847 (2005).
50. Hajjar, R. & Hodgkin, T. The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* **156**, 1–13 (2007).
51. Zamir, D. Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* **2**, 983–989 (2001).
52. Qi, L. L., Foley, M. E., Cai, X. W. & Gulya, T. J. Genetics and mapping of a novel downy mildew resistance gene, Pl_{18}, introgressed from wild *Helianthus argophyllus* into cultivated sunflower (*Helianthus annuus* L.). *Theor. Appl. Genet.* **129**, 741–752 (2016).
53. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).
54. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
55. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
56. Schapranow, M. P., Häger, F., Fähnrich, C., Ziegler, E. & Plattner, H. In-memory computing enabling real-time genome data analysis. *Int. J. Adv. Life Sci.* **6**, 11–29 (2014).
57. Picard v.2.5.0 (Broad Institute, 2016); http://broadinstitute.github.io/picard
58. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
59. *vcflib* (MIT, 2015); https://github.com/vcflib/vcflib
60. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
61. Boisvert, S., Laviolette, F. & Corbeil, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* **17**, 1519–1533 (2010).
62. *UniVec Database* (NCBI, 2016); ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec
63. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
64. Dong, Q., Shannon, D. S. & Brendel , V. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.* **32**, D354–D359 (2004).
65. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Seqtk v.1.0 (MIT, 2013); https://github.com/lh3/seqtk
67. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
68. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**, 361–369 (2008).
69. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
70. Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**, 2808–2810 (2014).
71. *Gene Ontology Annotation (GOA) Database* (EMBL-EBI, 2016); www.ebi.ac.uk/GOA
72. Supek, F. et al. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
73. Gulya, T. J. Everything you should know about downy mildew testing but were afraid to ask. In *Proc. 18th Sunflower Research Workshop* 39–48 (National Sunflower Association, 1996).
74. Cohen, Y. & Sackston, W. E. Factors affecting infection of sunflowers by *Plasmopara halstedii*. *Can. J. Bot.* **51**, 15–22 (1973).
75. Trojanová, Z., Sedlářová, M., Gulya, T. J. & Lebeda, A. Methodology of virulence screening and race characterization of *Plasmopara halstedii*, and resistance evaluation in sunflower—a review. *Plant Pathol.* **66**, 171–185 (2017).
76. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

Acknowledgements

We thank the Genome Quebec Innovation Centre and UBC's Biodiversity Research Centre for conducting the sequencing, M. Heffernan for the development of the statistical pipeline for GWAS, K. Rashid for providing downy mildew isolates,

W. Cheung for project coordination and assistance with experimental work, and SAP SE for computing resources. Funding was provided by Genome Canada and Genome BC (LSARP2014-223SUN), the NSF Plant Genome Program (DBI-0820451 and DBI-1444522) and the International Consortium for Sunflower Genomic Resources.

Author contributions

J.R.M., M.T., G.J.B., C.J.G., D.P.E., K.L.O., S.Y, B.T.M. and N.C.K. contributed to DNA sample collection and data production. N.B. and M.T. mapped downy mildew resistance. S.H., J.O. and E.Z. conducted the alignments and variant calling. S.H. performed the genome scans, pan-genome analyses, introgression analyses and GWAS. J.S.L. conducted the expression analyses. G.L.O. archived the data. J.E.B., I.C., L.G. and R.R.M. optimized the SNP data set for GWAS. L.H.R., J.M.B., N.B.L., S.M., T.K. and D.Z.H.S. designed the experiments and coordinated the project. S.H., N.B., J.M.B. and L.H.R. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0329-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Saril Hubner, Loren Rieseberg

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software or code were used to collect data. Databases from which data was retrieved are NCBI, PlantDB, UniVec; GOA. Links are given in the references.
Data analysis	Trimmomatic v.036; picard v.2.5; samtools v1.1; freebayes v1.1.0; SplitsTree v.4.14.5; bcftools v1.5; vcffilter; vcftools v.0.1.17; popgenome v.2.6.1; sweeD v.3.0; LDHat v.2.2; Ray v.2.3.1; CD-Hit v.4.6; ms (Oct2017), BLAST v.2.2.31; seqtk v.1.0; EMMAX beta-07Mar2010; simpleM v.2; PCAdmix v.1; Beagle v4.1; SAP-aligner; Sushi v.1.14; Revigo; SNPRelate v.1.10.2; ggpubr v.0.1.6; Blast2GO v4.1.9. Specific commands are provided in the supplementaries.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequence data is stored in the Sequence Read Archive (SRA) under Bioproject PRJNA353001, for cultivars, and PRJNA397453 for wild and landrace. Accession numbers for each sample are listed in Supplementary Table S9.
The SNP dataset in vcf format, pan-genome sequences in fasta format and introgression table can be downloaded from https://www.sunflowergenome.org/.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The study included the sunflower association mapping population (SAM) representing ca. 90% of the allelic diversity in cultivated sunflower. The collection was described previously in Mandel et al. 2011, 2013 as a sufficient sample for GWA analyses. The entire collection was used and therefore no sample size calculation was conducted. Other germplasm collections used in the study include Native American landraces and wild accessions representing 11 annual and perennial congeners. All accessions for which whole genome sequences is available were included in the analyses. Specific caveats associated with reduced sample size are provided along the the text.
Data exclusions	Exclusion of samples was pre-established and carried out before the GWAS analysis. Sunflower inbred cultivars with outliers of heterozygosity were excluded.
Replication	All attempts at replication were successful. All raw data, genome scans results, genotypes data are now available. In addition, we provide comprehensive tables to describe each accession and results. The explicit command that was used to produce the genotypes, de-novo assemblies and sequence comparisons are provided in the supplements.
Randomization	Cultivated lines were grouped into types (oil/non-oil; male/female) based on the information available for each line (Mandel et al. 2013)
Blinding	All analyses were conducted using serial code (SAM1:288) which provide no information about the corresponding accession. Most of the analyses are not focused on a specific accession. In places where specific accessions are described, all conclusions were drawn based on the obtained results and not the identify of the accession.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging