The genomics of linkage drag in inbred lines of sunflower

Kaichi Huang, Mojtaba Jahani, Jérôme Gouzy, Alexandra Legendre, Sébastien Carrere, José Miguel Lázaro-Guevara, Eric Gerardo González Segovia, Marco Todesco, Baptiste Mayjonade, Nathalie Rodde, et al.

PNAS RESEARCH ARTICLE | AGRICULTURAL SCIENCES

The genomics of linkage drag in inbred lines of sunflower

Kaichi Huang (黄恺驰)[a,b,1,2], Mojtaba Jahani[a,b,1], Jérôme Gouzy[c], Alexandra Legendre[c], Sébastien Carrere[c], José Miguel Lázaro-Guevara[a,b], Eric Gerardo González Segovia[a,b], Marco Todesco[a,b], Baptiste Mayjonade[c], Nathalie Rodde[d], Stéphane Cauet[d], Isabelle Dufau[d], S. Evan Staton[a,b,e], Nicolas Pouilly[c], Marie-Claude Boniface[c], Camille Tapy[c], Brigitte Mangin[c], Alexandra Duhnen[c], Véronique Gautier[f], Charles Poncet[f], Cécile Donnadieu[g], Tali Mandel[h], Sariel Hübner[h], John M. Burke[i], Sonia Vautrin[d], Arnaud Bellec[d], Gregory L. Owens[j], Nicolas Langlade[c,3], Stéphane Muños[c,3], and Loren H. Rieseberg[a,b,2,3]

Edited by Susan McCouch, Cornell University, Ithaca, NY; received June 9, 2022; accepted September 18, 2022

Crop wild relatives represent valuable sources of alleles for crop improvement, including adaptation to climate change and emerging diseases. However, introgressions from wild relatives might have deleterious effects on desirable traits, including yield, due to linkage drag. Here, we analyzed the genomic and phenotypic impacts of wild introgressions in inbred lines of cultivated sunflower to estimate the impacts of linkage drag. First, we generated reference sequences for seven cultivated and one wild sunflower genotype, as well as improved assemblies for two additional cultivars. Next, relying on previously generated sequences from wild donor species, we identified introgressions in the cultivated reference sequences, as well as the sequence and structural variants they contain. We then used a ridge-regression best linear unbiased prediction (BLUP) model to test the effects of the introgressions on phenotypic traits in the cultivated sunflower association mapping population. We found that introgression has introduced substantial sequence and structural variation into the cultivated sunflower gene pool, including >3,000 new genes. While introgressions reduced genetic load at protein-coding sequences, they mostly had negative impacts on yield and quality traits. Introgressions found at high frequency in the cultivated gene pool had larger effects than low-frequency introgressions, suggesting that the former likely were targeted by artificial selection. Also, introgressions from more distantly related species were more likely to be maladaptive than those from the wild progenitor of cultivated sunflower. Thus, breeding efforts should focus, as far as possible, on closely related and fully compatible wild relatives.

introgression | linkage drag | plant breeding | structural variation | sunflower

Significance

Wild relatives of crops often contain traits such as pest resistance that can be tapped to improve cultivated varieties. This is typically accomplished by crossing wild plants with cultivars, followed by backcrossing (introgression) to remove unwanted genetic material from the crop. However, it can be challenging to remove all unwanted genes, a phenomenon called linkage drag. In this paper, we generate and analyze reference sequences and trait data for sunflower to examine the consequences of linkage drag. We find that crop wild introgression has increased the genetic diversity of the crop gene pool. However, introgression has negatively affected yield and quality-related traits. We conclude by discussing potential strategies to minimize linkage drag.

Domestication—the process that transformed wild plants into highly productive crops—is arguably the most important innovation in human history (1). Not only did it spark explosive population growth and the establishment of modern civilization (2), but it also laid the foundation for the theory of evolution (3), thereby unifying the life sciences (4). While domestication and subsequent improvement have proven spectacularly successful in modifying plant architecture and enhancing yield (5), such changes often come with a cost, including losses of genetic diversity (6, 7), increases in genetic load (8), and reductions in resistance to biotic and abiotic stress (9, 10). This is of increasing concern in the 21st century, as environmentally resilient cultivars are needed to cope with a more hostile climate, while minimizing use of costly external inputs, such as fertilizer, pesticides, and water.

Fortunately, diversity lost during domestication and improvement may be regained by tapping the gene pools of crop wild relatives (CWRs). The potential utility of such wild germplasm has long been recognized by plant biologists and breeders (11–13), leading to global efforts to collect and conserve CWRs, in addition to the crops themselves (14). Likewise, breeding programs often include systematic efforts to introduce wild genetic material into domesticated breeding lines (15, 16). While many such efforts have focused on enhancing disease resistance (17), CWRs also have been used to increase nutritional quality, boost yield, and enhance resistance to abiotic stressors, such as drought, salt, and flooding (16, 18–20). Economic analyses have confirmed the value of such an approach. For example, a 2013 analysis of 32 crops estimated current benefits from CWR traits to be ~$68 billion annually, with potential future benefits of ~$196 billion annually (21).

Despite the clear value of CWR traits for crop improvement, there are downsides. The introduction of wild genetic material into cultivated lines typically occurs via repeated backcrossing or introgression (12). This process is not only time-consuming, but it also can be hampered by reproductive barriers that interfere with crosses or that reduce the fitness of hybrid offspring (22, 23). In addition, the resulting introgressions may have undesirable impacts on nontarget crop traits (24). While this can be due to

Author contributions: J.M.B., G.L.O., N.L., S.M., and L.H.R. designed research; K.H., M.J., J.G., A.L., M.T., B. Mayjonade, N.R., I.D., N.P., M.-C.B., C.T., V.G., C.P., C.D., T.M., S.H., S.V., and A.B. performed research; J.G. contributed new reagents/analytic tools; K.H., M.J., J.G., S. Carrere, J.M.L.-G., E.G.G.S., S. Cauet, S.E.S., B. Mangin, and A.D. analyzed data; and K.H., M.J., J.G., N.L., S.M., and L.H.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]K.H. and M.J. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: kaichi.huang@botany.ubc.ca or lriesebe@mail.ubc.ca.

[3]N.L., S.M., and L.H.R. contributed equally to this work.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2205783119/-/DCSupplemental.

Published March 27, 2023.

negative pleiotropic effects of the target alleles, adverse effects appear to be more frequently caused by linked alleles that are deleterious in the crop genetic background (24, 25), a phenomenon called linkage drag. Plant breeders typically monitor the size and location of introgressions with molecular markers and focus their efforts on fully compatible wild relatives [i.e., members of the primary gene pool (26)] to reduce the impact of the linkage drag (12, 27, 28). However, in large plant genomes, regions of low recombination are widespread, making it difficult to reduce the sizes of some introgressions (29–31). Also, key traits may be found outside of the primary gene pool, making it necessary to tap less compatible wild relatives (e.g., ref. 32). The latter are classified as the secondary gene pool if they can intercross with the crop and produce at least some partially fertile hybrids (26). More distantly related species that require technological interventions to produce hybrid offspring are referred to as the tertiary gene pool (26). In sunflower, for example, the primary gene pool comprises cultivated and wild accessions of *Helianthus annuus*, which are fully fertile in crosses (33). The secondary gene pool consists of members of the annual sunflower clade to which *H. annuus* belongs, and crosses with cultivated sunflower result in partially sterile hybrids due to chromosome-pairing abnormalities at meiosis (34). All other members of the genus are classified as belonging to the tertiary gene pool (35), and embryo culture is typically employed to generate hybrids with cultivated sunflower (36).

The causes of linkage drag are assumed to be like those that contribute to species differences in natural populations. These include the genetic changes responsible for phenotypic divergence, as well as various kinds of hybrid incompatibilities (23, 24). Introgressions with strongly negative effects are likely purged by selection during early breeding steps, so those successfully incorporated into the cultivated gene pool should be less harmful. However, the cumulative effects of such introgressions on the genomes and phenotypes of cultivated plants are not fully understood. Introgression has been shown to reduce genetic load in maize (37) and sorghum (38) and to introduce gene presence/absence polymorphisms in sunflower (39), thereby increasing the size of its pan-genome (40). However, a more complete analysis of the genomic impacts (e.g., changes in gene content, structural variation, and so forth) of such introgressions can be aided by generation and analyses of multiple high-quality reference genomes.

Here, we analyze the phenotypic and genomic effects of linkage drag in inbred sunflower lines. CWRs have been widely employed in sunflower breeding (17, 41), and recent genomic studies have estimated that circa (ca.) 10% of the cultivated gene pool is derived from wild introgressions (40, 42). While most such introgressions are from wild *H. annuus*, the fully compatible progenitor of the cultivated sunflower, there are significant contributions from other species as well (*SI Appendix*, Table S1), making it feasible to compare the effects of introgression from the primary and secondary gene pools.

To estimate the impacts of linkage drag, we first sequenced and assembled reference genomes for seven cultivated and one wild sunflower genotype and improved the assemblies for two previously sequenced cultivars (43). Then, using resequencing data previously generated for a diverse panel of wild donor species (40, 44), we identified introgressions in the cultivar genomes and examined their impacts on sequence and structural variation in the cultivated sunflower gene pool. Lastly, we determined the locations of introgressions in the cultivated sunflower association mapping (SAM) population (45) and conducted a genome-wide association study (GWAS) to detect significant associations between markers for each introgression (i.e., introgression variants) and variation in 16 phenotypic traits, including quality traits, such as oil percentage in seeds; developmental traits, such as flowering time; and yield-related traits, such as head weight. Further analyses using a ridge-regression best linear unbiased prediction (BLUP) model allowed us to quantify the cumulative effects of introgressions on phenotypic variation in the inbred sunflower lines.

As expected, we found that introgressions increased sequence and structural polymorphism in the cultivated gene pool and reduced genetic load at protein-coding sequences. On the other hand, introgressions typically reduced quality and yield traits. We also found that higher-frequency introgressions have larger effects than low-frequency introgressions, possibly indicating that the former have been targeted by artificial selection. Lastly, introgressions from the secondary gene pool had much larger negative effects than those from the primary gene pool.

Results

To identify SVs and introgressions in cultivated sunflowers, we constructed de novo genome assemblies using Pacific Biosystems (PacBio) sequencing for seven inbred cultivated lines and one wild *H. annuus* genotype (Table 1, *SI Appendix*, Table S1, and Dataset S1). Five of these assemblies were further scaffolded by using Bionano optical mapping. We also improved the quality of previously sequenced assemblies (43) for the HA412-HO inbred line using Illumina, 10×, and Hi-C sequencing (*SI Appendix*, Table S1) and for the XRQ inbred line using the PacBio/Bionano combination described above. The nine cultivated lines represent a large part of cultivated sunflower genetic diversity present in the world's gene banks (*SI Appendix*, Fig. S1) (46).

All genomes were assembled into 17 pseudomolecules, corresponding to the 17 chromosomes in sunflower. Each of our chromosome-level genome assemblies had a total size between 3,002 and 3,226 Mb, with N50 of 172 to 187 Mb (Table 1 and Dataset S2). The total number of genes per genome, after stringent filtering, ranged from 44,640 for XRQv2 to 63,048 genes for HA300 (*SI Appendix*, Table S5). The assemblies captured 85.9 to 97.9% of the universally conserved single-copy benchmark (BUSCO) genes (Table 1 and *SI Appendix*, Table S4). BUSCO percentages were positively correlated with sequence depth, rather than gene number, with the lowest BUSCO scores observed for LR1 and OQP8, which were sequenced to ca. 13× depth, whereas the highest BUSCO scores were seen for HA412-HOv2 and XRQv2, which were sequenced to 251× depth and 172× depth, respectively (Table 1 and Dataset S1). The genomes showed high collinearity without large interchromosomal translocations (*SI Appendix*, Figs. S2–S6). Overall, our chromosome-scale genome assemblies yielded better qualitative metrics than the two published reference assemblies (43).

In general, 74 to 83% of the genomes are composed of transposable elements (TEs), with 70 to 73% of these being long terminal repeat–retrotransposons (LTR-RTs) (*SI Appendix*, Table S6). In agreement with previous studies of the cultivated sunflower genome (47), there is a major bias in TE composition toward *Gypsy* (50–60% of total TEs) and *Copia* (13–18% of total TEs) elements, while Class II TEs (DNA transposons) were much lower in abundance relative to LTR-RTs, comprising <13% of each genome (*SI Appendix*, Table S6). The genomic distributions of LTR-RTs in the newly generated assemblies are similar to those reported for the first reference genomes for cultivated sunflower (*SI Appendix*, Figs. S7–S15) (43).

Table 1. Description of new or improved reference genomes for sunflower (*H. annuus*)

Genotype/ version	Type	Sequencing technology	Sequence depth*	Scaffolding technology	N50, Kb	Assembly size, Kb	Complete BUSCO genes, %
HA412-HO v2	Cultivar, maintainer	Illumina paired-end, mate pair & 10X linked reads	251×	Hi-C sequencing	187,414	3,226,370	97.9
XRQ v2	Cultivar, maintainer	PacBio CLR, Illumina paired-end	172×	Bionano optical mapping	176,491	3,010,048	97.4
PSC8 v1	Cultivar, restorer	PacBio CLR, Illumina paired-end	66×	Bionano optical mapping	179,999	3,057,327	94.5
RHA438 v1	Cultivar, restorer	PacBio CCS	55×	Bionano optical mapping	177,554	3,095,288	96.7
IR v1	Cultivar, maintainer	PacBio CCS	60×	Bionano optical mapping	179,325	3,047,956	97.1
HA89 v1	Cultivar, maintainer	PacBio CCS	34×	Bionano optical mapping	175,389	3,002,007	97.3
LR1 v0.9	Cultivar, maintainer	PacBio CCS	13×	Reference-guided	174,126	3,154,038	85.9
OQP8 v0.9	Cultivar, restorer	PacBio CCS	13×	Reference-guided	177,187	3,119,769	88.1
HA300 v0.9	Cultivar, maintainer	PacBio CCS	10×	Reference-guided	171,505	3,025,264	90.3
PI659440 v1	Wild	PacBio CCS	41×	Bionano optical mapping	181,076	3,162,322	96.5

*Polished sequence data.

By mapping published whole-genome sequences (*SI Appendix*, Table S7) (40, 44) from native North American landraces and five wild possible donor species to each genome assembly, we determined the ancestry of each cultivated line and estimated the locations and likely parentage of introgressions. Only a small portion (2 to 8%) of each genome was admixed (Fig. 1, *SI Appendix*, Fig. S16, and Dataset S3), which is similar to previous estimates for the XRQ and HA412-HO genomes (43). All cultivated genomes possessed more introgressions from the primary gene pool (primary introgressions) than those from the secondary gene pool (secondary introgressions).

Sunflower is a hybrid crop, and CWRs were used to develop cytoplasmic male sterile "female" lines and branching, fertility-restoring "male" lines for hybrid production. The male-restorer lines (PSC8, OQP8, and RHA438) generally had more introgressions than the female-maintainer lines (HA412, XRQ, IR, HA89, LR1, and HA300) (*SI Appendix*, Fig. S16). Consistent with breeding records and previous findings (40, 42, 48, 49), the restorer lines had substantial introgression from wild *H. annuus* on chromosome 10 (chr10), which underlies apical branching, as well as an introgression near the distal end of chr13, where the restorer of fertility locus (*Rf1*) of the common PET1 male sterile cytoplasm is located (Fig. 1). However, while the restorer allele in PSC8 and OQP8 was derived from *Helianthus petiolaris*, as expected (50), an introgression from wild *H. annuus* was found in RHA438 at the region, suggesting possible different origins of fertility restoration in cultivated sunflower. The majority (~68%) of the primary introgressions were unique to one genotype, and only a small proportion (<0.1%) were shared across all nine genomes. Almost all secondary introgressions were unique to one genotype.

We identified single-nucleotide polymorphisms (SNPs) and small (<50 bp) insertions/deletions (InDels), as well as different types of structural variants (SVs), including large (>50 bp) InDels, copy-number variants (CNVs), inversions, and translocations through the alignment of the high-contiguity cultivar genome assemblies (HA412-HOv2, XRQv2, PSC8, RHA438, IR, and HA89). In total, we identified 12,036,913 SNPs and 3,005,855 small InDels across 17 chromosomes using the HA412-HOv2 genome as the reference (Fig. 1). We also detected 70,612 to 84,709 large InDels, 32,668 to 47,706 CNVs, 4,776 to 7,738 translocations, and 261 to 301 inversions (>1 kb) between each genome and the HA412-HOv2 reference (Fig. 1 and Dataset S4). About 68.6% of large InDels and 80.5% of CNVs appear to be associated with the movement of TEs (*SI Appendix*, Fig. S17), consistent with a dominant role for transposons in driving the structural divergence of genomes. After merging, 532 polymorphic inversions with a total size of 200 Mb were identified across the cultivars, including a 21-Mb region (156 to 177 Mb) on chr5 that corresponded to the largest section of a cluster of inversions previously identified in wild *H. annuus* (Fig. 1*J*) (44).

Introgression Introduced Substantial Sequence and Structural Variation into the Cultivated Sunflower Gene Pool. We compared densities of SNPs and small InDels between regions with introgression in one to five genomes (polymorphic introgressed regions) and those without introgression in any of the six highly contiguous cultivar genomes (nonintrogressed). We calculated densities of SNPs and small InDels in nonoverlapping windows of 500 kb using the HA412-HOv2 genome as the reference and compared between polymorphic introgressed regions and nonintrogressed regions. Overall, regions polymorphic for primary or secondary introgressions had more SNPs and small InDels than nonintrogressed regions (Fig. 2 *A* and *B*). Secondary introgressions had more SNPs and small InDels than primary introgressions, although the differences were not significant. Analyses of 287 individuals comprising the cultivated SAM population (see below) revealed that introgressed regions also possessed significantly higher numbers of SNPs compared to nonintrogressed regions, and secondary introgressions

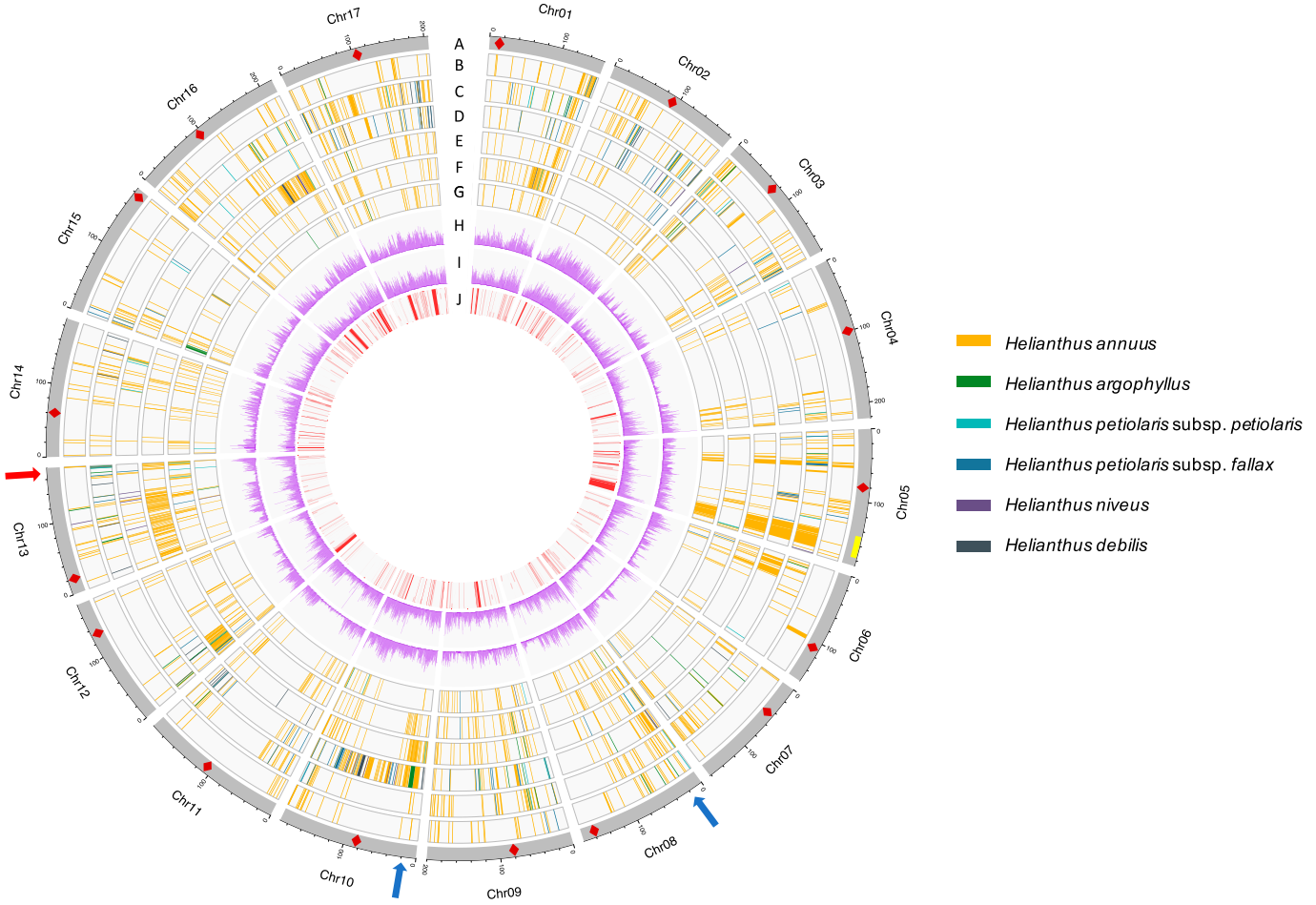

Fig. 1. Introgressions and genetic variants of the high-contiguity cultivated sunflower genome assemblies. (*A*) Chromosomes of the HA412-HOv2 reference. Diamonds mark approximate positions of centromeres. (*B*–*G*) Introgressions in HA412-HO, XRQ, PSC8, RHA438, IR, and HA89, respectively, projected to the Ha412-HOv2 reference. Colored bars represent introgressions from different wild donors: primary gene pool: *H. annuus*; secondary gene pool: *H. argophyllus*, *H. petiolaris* subsp. *petiolaris*, *H. petiolaris* subsp. *fallax*, *H. niveus*, and *H. debilis*. (*H* and *I*) Density of SNPs (*H*) and small InDels (*I*) (number/500 kb; 0 to 10,000 for SNPs and 0 to 2,000 for small InDels). (*J*) Inversions identified in genome assemblies. Blue and red arrows indicate locations of major branching loci and the restorer of fertility locus (*Rf1*) identified in previous studies, respectively. Yellow arc indicates the location of a cluster of inversions identified in ref. 44. Regions of introgression less than 1 Mb were thickened to 1 Mb for visualization.

displayed significantly more SNPs than primary introgressions (*SI Appendix*, Fig. S18).

Wild introgressions also introduced large (>50 bp) insertions and deletions (large InDels) into the cultivated sunflower gene pool. In each pair of genome comparisons with the HA412-HOv2 reference, both primary and secondary introgressions had significantly higher numbers of large InDels compared to regions without introgression (Fig. 2*C*). Conversely, introgressions had significantly fewer CNVs than nonintrogressed regions (Fig. 2*D*). We suspect that this is due to the reduced strength of purifying selection on TE copy number in the cultivated gene pool.

Across the six high-contiguity genomes, chromosomal inversions had an overlap of 58 Mb with primary introgressions and 5.7 Mb with secondary introgressions, which is significantly higher than a random distribution in both cases (primary introgressions: $P < 0.001$; secondary introgressions: $P = 0.0269$). In each pair of genome comparisons with the HA412-HOv2 reference, the number of inversions introduced from the primary gene pool varied from 0.24 to 0.43 per Mb, which is significantly ($P < 0.01$) higher than that in nonintrogressed regions (0.07 to 0.08/Mb). More inversions were introduced from the secondary than from the primary gene pool in each genome, except in HA89, where no inversions were found in secondary introgressions (*SI Appendix*, Fig. S19).

Introgression Reduced Genetic Load. We estimated the effect of introgression on genetic load by calculating the ratio of the number of alternative stop codons (P_{nonsense}) and the number of nonsynonymous mutations (P_{nonsyn}) in 500-kb sliding windows (51). The statistic was negatively correlated with recombination rate (*SI Appendix*, Fig. S20), in accord with previous understanding of the role of recombination in eliminating deleterious mutations (31). $P_{\text{nonsense}}/P_{\text{nonsyn}}$ of polymorphic primary introgressions was lower in null recombination-rate regions than that of nonintrogressed regions and comparable to nonintrogressed regions in regions of reduced and high recombination rate (*SI Appendix*, Fig. S20). Secondary introgressions displayed a trend toward reduced load (i.e., lower $P_{\text{nonsense}}/P_{\text{nonsyn}}$ ratios) compared to nonintrogressed regions, but the sample size was too small to draw conclusions. Analyses of 287 individuals in the cultivated SAM population (see below) provided clearer results. While $P_{\text{nonsense}}/P_{\text{nonsyn}}$ was also negatively correlated with recombination rate in this dataset (Fig. 3*A*), primary introgressions displayed significantly lower $P_{\text{nonsense}}/P_{\text{nonsyn}}$ than nonintrogressed regions in all recombination-rate categories,

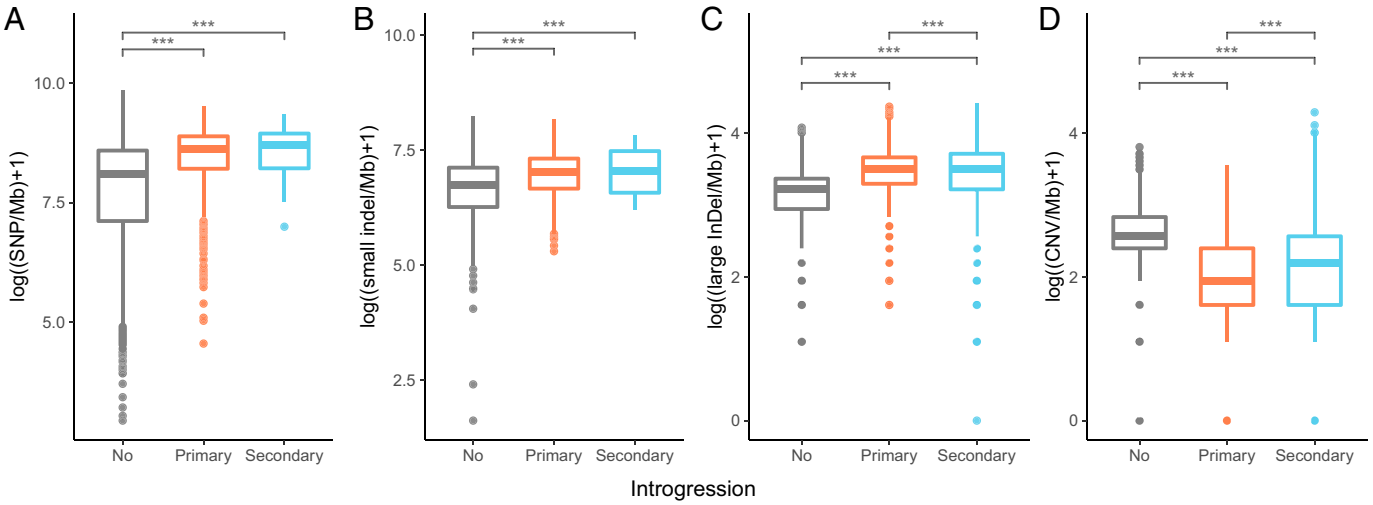

Fig. 2. Densities of SNPs (*A*), small InDels (<50 bp) (*B*), large InDels (>50 bp) (*C*), and CNVs (*D*) in regions without introgression, regions with introgressions from the primary gene pool (primary introgressions), and regions from the secondary gene pool (secondary introgression). The densities of SNPs and small InDels were calculated in nonoverlapping windows of 500 kb by using the HA412-HOv2 genome as the reference. Densities of large InDels and CNVs were calculated in 10,000 samplings of 500-kb windows in each type of region between each genome and the HA412-HOv2 reference. Asterisks denote significance in independent *t* tests. ***$P < 0.001$.

and secondary introgressions had significantly lower $P_{nonsense}/P_{nonsyn}$ than nonintrogressed regions in regions of null and reduced recombination rate (Fig. 3*B*).

Introgressions Introduced Gene Presence/Absence Polymorphisms. A total of 77,334 genes were obtained across the 10 genome assemblies, among which 75,791 were present in the 9 genomes of cultivars. Altogether, 31,099 genes in the pan-genome displayed presence–absence variation (PAV) between genomes. After filtering based on synteny, we retained 75,369 genes with coordinate information for homologs, 29,948 of which showed PAV.

We found that introgressions introduced significantly more gene PAVs than nonintrogressed regions, but gene PAVs from primary and secondary introgressions did not differ significantly, except in several pairs (Fig. 4). The total number of genes introduced by primary introgressions ranged from 889 for HA300 to 4,323 for RHA438, respectively, whereas between 26 (HA89) and 1,800 (OQP8) genes were introduced by secondary introgressions (*SI Appendix*, Fig. S21). On average, 12% of the PAVs result from primary introgressions and 5% from secondary introgressions. Across the nine cultivar genomes, a total of 3,187 genes were introduced by introgression from CWRs. Unsurprisingly, the number of new genes introduced by introgression is closely correlated with total amount of introgression detected in a genome, so we see more new genes resulting from introgression in the restorer lines (PSC8, RHA438, and OQP8) than in maintainer lines (*SI Appendix*, Fig. S21).

In addition to new genes, introgressions often lack genes that are present in syntenic nonintrogressed regions (Fig. 4*B*). Primary introgressions resulted in 383 (HA300) to 1,577 (RHA438) missing genes, whereas between 22 (HA89) and 2,095 (OQP8) gene absences were caused by secondary introgressions (*SI Appendix*, Fig. S21). About 17 to 32% of the gene

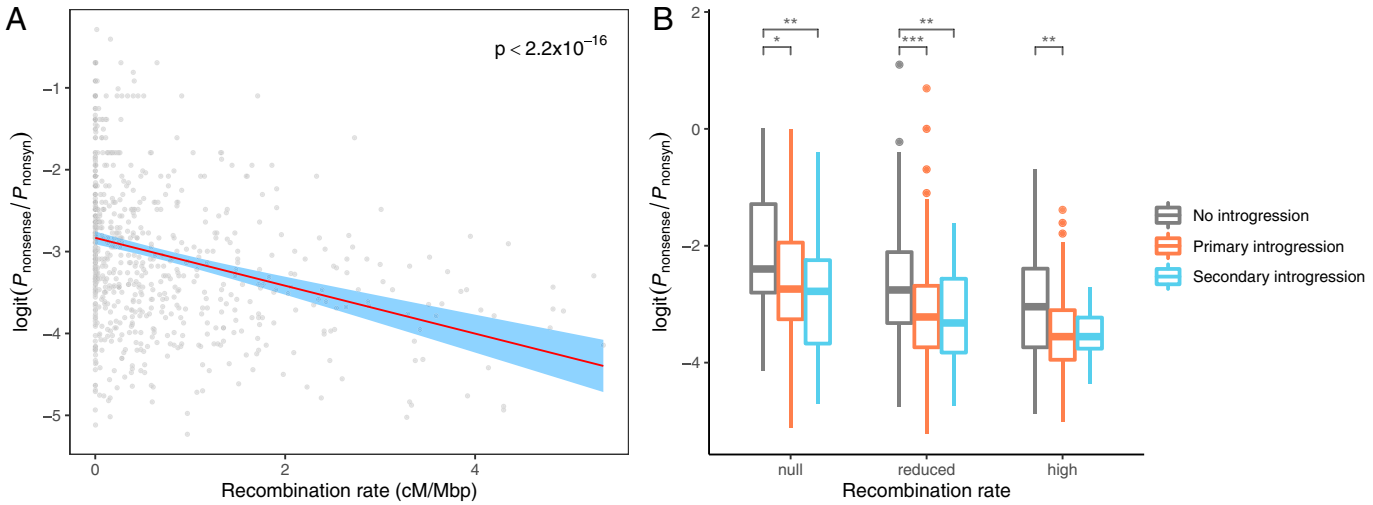

Fig. 3. Ratio of alternative stop codons and nonsynonymous mutations ($P_{nonsense}/P_{nonsyn}$) in the cultivated SAM population. (*A*) Recombination rate and $P_{nonsense}/P_{nonsyn}$. The red line denotes the best-fit linear regression line with 95% CIs shaded in blue. (*B*) $P_{nonsense}/P_{nonsyn}$ in regions without introgression, regions with introgressions from the primary gene pool (primary introgressions), and regions from the secondary gene pool (secondary introgressions) in the cultivated SAM population. $P_{nonsense}/P_{nonsyn}$ was calculated in nonoverlapping windows of 500 kb. Windows of each recombination rate category (high: >2 cM/Mb, reduced: 0.01 to 2 cM/Mb, null: <0.01 cM/Mb) were compared separately. Asterisks denote significance in independent *t* test. *$0.05 > P > 0.01$; **$0.01 > P > 0.001$; ***$P < 0.001$.

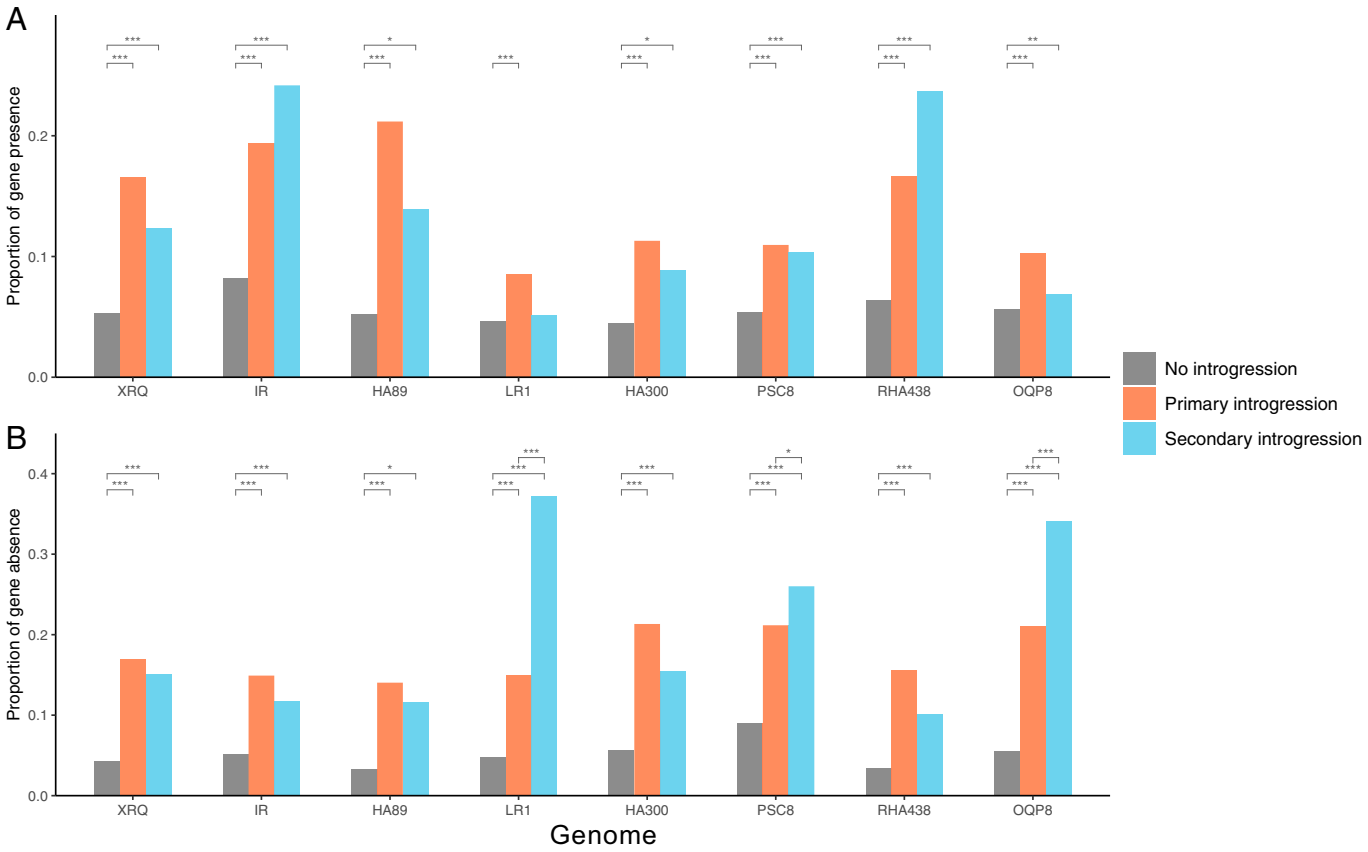

Fig. 4. Proportions of gene presence (*A*) and gene absence (*B*) in introgressed and nonintrogressed regions in each cultivar genome compared to the HA412-HOv2 reference. Asterisks denote significance in independent *t* test. *0.05 > *P* > 0.01; **0.01 > *P* > 0.001; ***$P < 0.001$.

absences in primary introgressions had a homolog present in the wild *H. annuus* (PI659440) genome, indicating that many of such missing genes represent gene PAVs in the wild donor species.

Because sunflower is a hybrid crop, and complementation of PAVs appears to contribute importantly to heterosis (52), we quantified the extent of introgressed PAV complementation that would be expected in hybrids formed from each combination of inbred cultivars. Across each pair of the sequenced lines, we predicted 536 to 2,921 cases of complementation of introgressed gene PAVs. Predicted PAV complementation in hybrids between maintainer and restorer lines was significantly higher than those between maintainers ($P < 0.001$), but no difference was found relative to pairs of restorers ($P = 0.258$; *SI Appendix*, Fig. S22).

Introgressions in the Cultivated SAM Population. We generated a dataset of 1,348,829 SNPs using published sequence data for 287 individuals in the SAM population (40, 45), as well as the aforementioned whole-genome sequences from landraces and five possible wild donor species (*SI Appendix*, Table S7), and determined the locations and source of introgressions in each of the 287 cultivated genotypes. All samples contained putative introgressions, and all chromosomes appeared to have experienced introgression in at least one SAM sample (*SI Appendix*, Fig. S23). The amount of introgression in each sample varied from 0.4 to 11%, with a number of samples having large blocks of introgression (Dataset S5). On average, each sample had ca. 3% of the genome covered with introgressions from the primary gene pool and 0.1% derived from the secondary gene pool, which is similar to the estimates from the genome assemblies, but lower than previously estimated for the SAM population using a different method (40). Maintainer lines possessed comparable amounts of introgression to open-pollinated lines (2.9% vs. 2.6%, $P = 0.296$), while restorer lines had more introgression than maintainer lines, on average (3.8% vs. 2.9%, $P < 0.001$). Maintainer and restorer lines showed distinct patterns of introgression on the first half of chr8, a substantial portion of chr10, part of chr12, and the end of chr13, broadly consistent with previously identified regions of high divergence between these groups (39, 40, 42). Small regions of introgression from the secondary gene pool were identified at the end of chr13 in most of the restorer lines, but not in maintainers. These regions roughly correspond to the introgression from *H. petiolaris* in the PSC8 genome, corroborating previous findings of the *Rf1* restorer allele at this position (42, 48).

Using these datasets, we evaluated the presence or absence of introgressions in 1-kb nonoverlapping windows across the genome. We took this approach to account for the fact that most introgressions are fragmented by recombination as they are incorporated in the cultivated sunflower gene pool and to permit GWAS and various population genomic analyses. A total of 505,038 and 5,243 introgression variants were detected at a ≥3% minor-allele frequency cutoff for primary (wild *H. annuus*) and secondary germplasm donors, respectively (*SI Appendix*, Fig. S23).

We then performed GWAS of the introgression variants for 16 traits that were previously phenotyped (52–54) in common gardens at three locations (Watkinsville, GA; Ames, IA and Vancouver, BC) using a model that corrects the population structure and familial relatedness. Our results revealed that introgressions have a significant effect on the phenotypic variation in

the SAM population (*SI Appendix*, Fig. S24). After merging genome-wide association (GWA) outliers in the range of 10 Mb, introgression intervals were found to underlie 27 quantitative trait loci (QTLs) for 12 phenotypic traits (*SI Appendix*, Table S8 and Fig. S23). Of these, 23 (85.18%) were introgressed from primary germplasm (wild *H. annuus*), while 4 (14.81%) were introgressed from secondary germplasm. The introgressed QTLs reduced head diameter and head weight, but increased plant biomass, number of branches, anthocyanins in disk florets, number of days to flowering, dry leaf weight, oil percentage, seed size, dry stem weight, and anthocyanins in stigmas. For stem diameter, introgressed QTLs with negative and positive effects were found. The 27 QTLs were not fully independent. A primary introgression near the beginning of chr10 that introduced branching into restorer lines also affects oil content, seed size, head diameter, and head weight.

However, GWAS does not consider the effects of introgression variants that fall below a stringent significance threshold. Therefore, we employed a ridge-regression BLUP model to estimate phenotypic effects across all introgression variants.

Then, to assess whether introgressions overall have a significant impact on the 16 phenotypic traits, we compared the average value of introgression marker effects to a null distribution. Our results indicated that introgressions overall have negative effects on quantitative traits associated with yield, including head diameter, head weight, leaf area, leaf weight, seed size, seed weight, stem diameter, and stem weight (Fig. 5). This pattern was seen for introgressions from both the primary (wild *H. annuus*) and secondary gene pools. In contrast, biomass, branching, and specific leaf area (SLA) showed an increase in the trait value for introgressions from both gene pools. Branching was introgressed into restorer lines to prolong the flowering period for hybrid production, and increased SLA is thought to be associated with higher growth rates (55), so both changes can be viewed as potentially desirable. We also observed gene-pool-specific effects for stigma and disk-floret anthocyanins and oil percentage; primary introgressions increase anthocyanin content and reduce oil percentage, whereas introgressions from secondary germplasm do not cause significant change (Fig. 5). Lastly, a comparison of effect sizes of introgression variants from the primary vs. secondary gene pool indicates that the latter have much larger effects on average (*SI Appendix*, Fig. S25).

Next, we asked whether the frequency of introgression variants was correlated with their effect size. Higher-frequency introgressions are more likely to have been targets of artificial selection, so we were especially interested in the potential for linkage drag associated with such introgressions. We found a significant correlation ($P < 0.05$) between the frequency and the effect size of introgression variants from both the primary and secondary gene pools across all traits and common gardens (Fig. 6 and *SI Appendix*, Fig. S26). In general, higher-frequency introgressions have larger phenotypic effects than lower-frequency introgressions. Changes in beta coefficients were mostly consistent between donor gene pools: Biomass, branching, SLA, oil percentage, and stigmas anthocyanins had positive beta values for introgressions from both the primary and secondary gene pools, whereas negative beta values were observed for the other traits.

Complementation of Introgressions in Simulated Hybrids from the SAM Population. As shown above, introgressions introduce gene PAVs, as well as other potentially maladaptive variants. Thus, complementation of introgressions in hybrids offers a potential means for reducing linkage drag and may contribute to heterosis (52). Therefore, we quantified the extent of predicted introgression complementation in hybrids by simulating all pairwise cross combinations in the SAM population. As

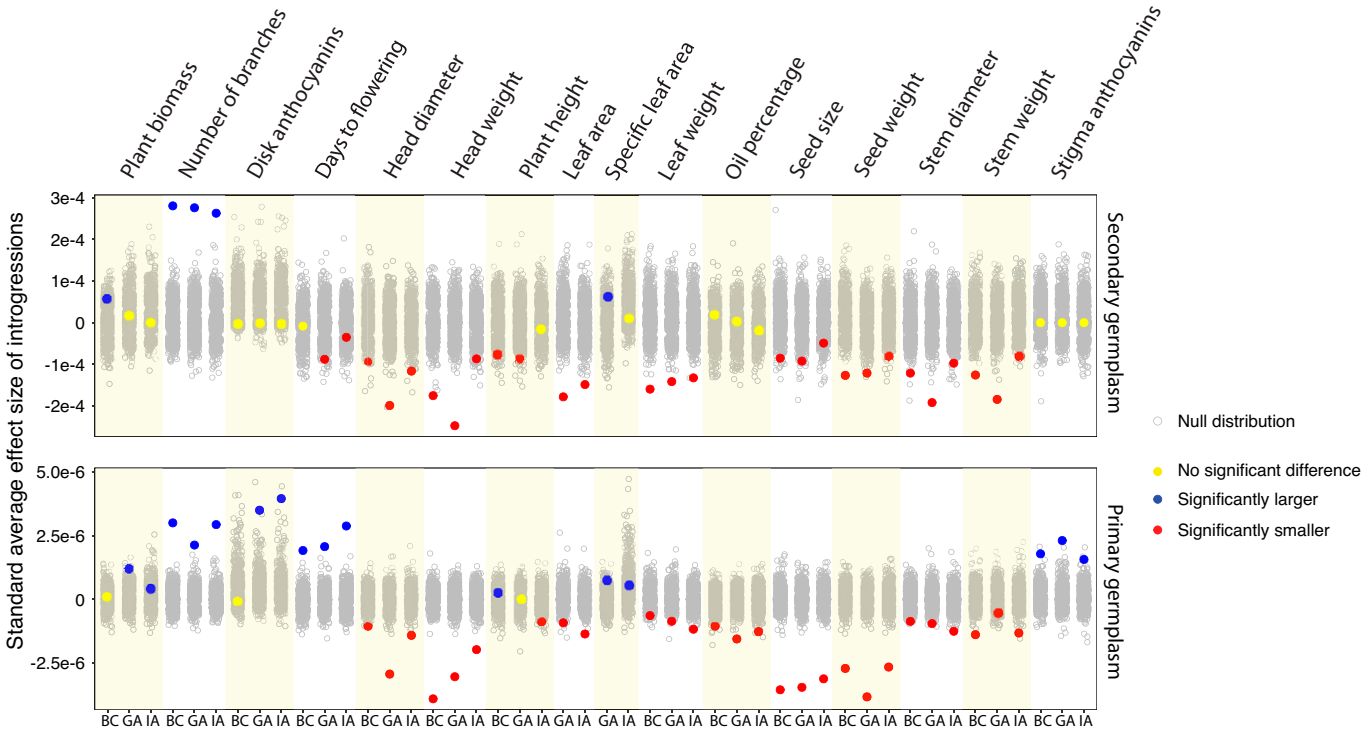

Fig. 5. Standardized average effect sizes of introgression variants from the ridge-regression BLUP model (calculated from *Z*-score-normalized trait values) in the SAM population. Gray dots show the null distribution of effect sizes. Red, blue, and yellow represent the decreasing, increasing, or neutral effects of introgressions on phenotypic traits at *P* value < 0.05. BC, British Columbia; GA, Georgia; IA, Iowa.

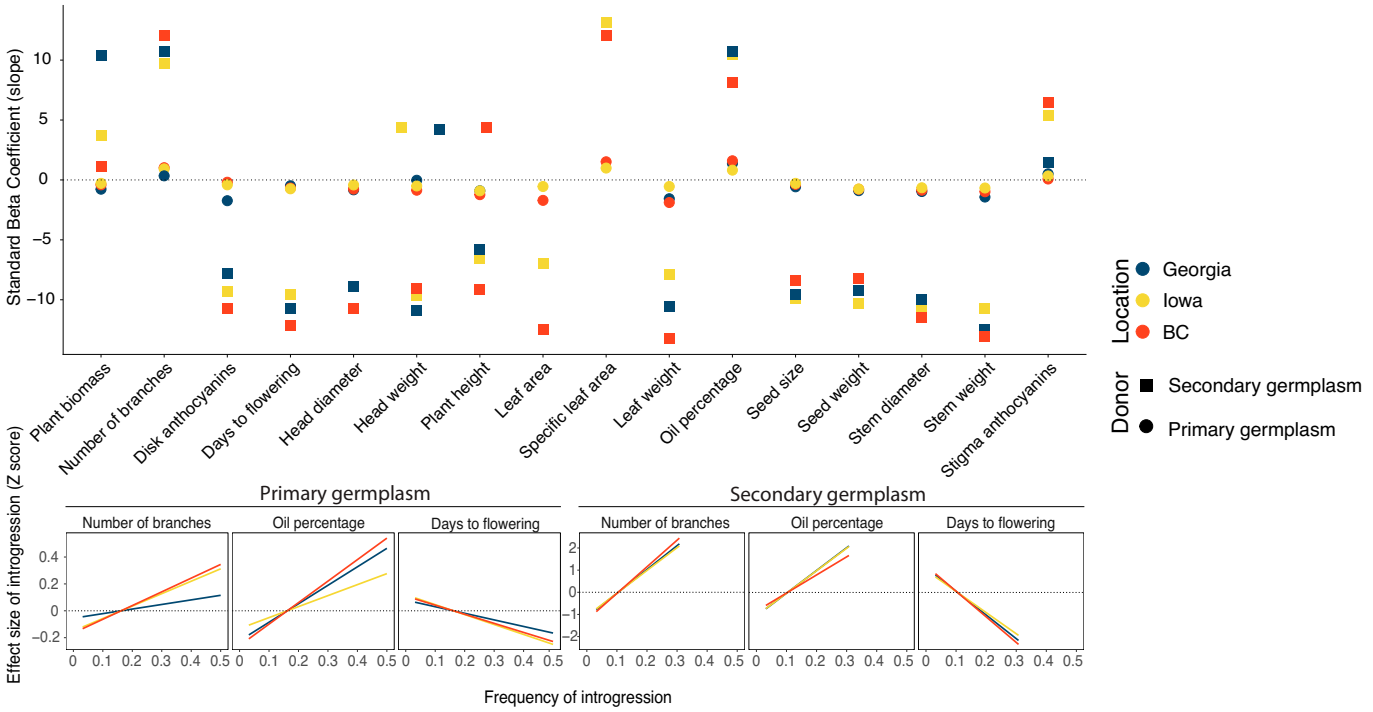

Fig. 6. Results from a linear regression model where X = introgression frequency in SAM population and Y = introgression effect on phenotype trait (in the ridge-regression BLUP model). (*A*) The standard beta coefficient of all traits in three common garden experiments. (*B*) A fitted linear regression line for branching, oil percentage, and days to flowering for introgressions from primary and secondary germplasm.

expected, the highest predicted complementation was for hybrids between maintainer (HA) and restorer (RHA) lines, whereas the lowest complementation was predicted for HA–HA, HA–Other, and Other–Other hybrids (*SI Appendix*, Fig. S27). The predicted complementation in RHA–RHA hybrids was surprisingly high and not different from HA-RHA lines for secondary introgressions. Information on simulated crosses is summarized in two interactive plots for the primary (https://sunflowergenome.org/interactive/SAM_cross_primary_germplasm.html) and secondary (https://sunflowergenome.org/interactive/SAM_cross_2nd_germplasm.html) germplasm, which can be used, for example, to select cross-combinations that are predicted to maximize complementation.

Discussion

Genomic Resources for Sunflower. For the past two decades, the plant biology community has made substantial investments into the generation of genomic tools and resources for crops and their wild relatives, especially high-quality reference sequences (56). These investments are now bearing dividends, ranging from exciting new discoveries about plant domestication (57), to the genetic dissection of key ecological and agronomic traits (58, 59), to increases in the speed and precision of plant breeding (60). Despite these successes, the goalposts have moved. Plant genomes have been shown to vary remarkably in their content and structure, even within species (61, 62), and these differences often underlie variation in phenotypic traits (63). Thus, tens or even hundreds of reference-quality genomes are needed to fully understand the genomic basis of phenotypic variation (63, 64). Here, we report progress toward this goal by providing eight new chromosome-level genomes for sunflower along with significant improvements of two previously published sunflower genomes (43). These 10 publicly available genomes, which encompass much of the genetic space in the cultivated sunflower gene pool (*SI Appendix*, Fig. S1), represent a valuable resource for sunflower research and breeding.

While the genomes were sequenced and assembled by using different sequencing technologies and depths, we were able to obtain chromosome-level assemblies for all genotypes, even with sequencing depth as low as 10×, when using PacBio HiFi reads and reference-guided assembly (for HA300; Table 1). We did see a trade-off between lower sequence coverage and BUSCO scores, suggesting that the quality of gene annotation suffers at lower sequencing depths. However, excellent BUSCO scores were obtained with sequence depth in the 30x range with HiFi reads, which may represent an optimal balance between sequencing cost and genome quality.

The cultivated genomes range from 3.02 to 3.23 Gb in size, with the wild genome at 3.16 Gb falling in the middle. Thus, domestication in sunflower does not appear to been accompanied by a change in genome size. On the other hand, the 10 genomes are ca. 15% smaller than previous genome size estimates for *H. annuus* (which included HA89, one of the genomes sequenced here) based on Feulgen staining (65) and flow cytometry (66). Given that the two different scaffolding approaches (Bionano optical mapping and Hi-C sequencing) employed in the present study resulted in similar genome size estimates, we suspect that previous work overestimated the size of the sunflower genome.

Synteny comparisons of the six high-contiguity genomes did not reveal large (>10 Mb) chromosomal rearrangements between the genomes, except for one 21-Mb inversion, consistent with previous findings (43, 44). However, we did find millions of small InDels, thousands of deletions and insertions, and hundreds of inversions. We also detected numerous differences in gene content, with ∼40% of the 77,334 genes in the sunflower pan-genome varying between genomes. This is higher than the 27% previously reported based on resequencing data from the SAM population (40), possibly because the present study is based on comparisons of fully assembled reference genomes. Estimates of the proportion of genes displaying

presence–absence polymorphisms in other plant species range from 15 to 66% (62, 67), so the level of polymorphism in sunflower is not unusual. Like other plant species, gene presence–absence polymorphisms have been shown to play an important functional role in sunflower. For example, Todesco et al. (44) showed that a PAV for *HaFT1* was responsible for a 77-d shift in flowering time between two ecotypes of the silverleaf sunflower. More recently, Lee et al. (52) found that the complementation of PAVs in sunflower hybrids was the primary cause of heterosis.

Genomic Consequences of Introgression. Analyses of the 10 genomes provide insights regarding the sources of variation among them. Consistent with previous reports, about three-quarters of the sunflower genome is made up of LTR transposons and other TEs, and many of the differences between genomes result from variability in the accumulation, movement, and elimination of TEs (43). Also, sunflower is the product of a whole-genome duplication event ∼29 million years ago (43, 68, 69), and the differential retention of duplicated sequences likely contributes to genomic diversity as well.

Introgression from wild relatives represents another potential source of variation (39, 40). By examining the location and parentage of introgressions in the cultivated genomes, we were able to show that introgressed regions have greater diversity than non-introgressed regions, as measured in terms of SNPs, small InDels, deletions, insertions, inversions, and gene PAVs. The impact of the introgressions was most pronounced for the latter, with introgressions accounting for about 17% of PAVs. This is qualitatively similar to wheat, where differences in the gene content of introgressions from divergent donors appears to cause reduced performance (70). Introgressions also reduced genetic load at protein-coding genes and variation in CNVs, possibly because of relaxed purifying selection in the cultivated gene pool. CNVs in sunflower are mostly caused by variation in TE copy number, which may explain why introgression affects them differently than gene PAVs.

A previous study of the SAM population showed that the absence allele at PAVs often has deleterious impacts on yield-associated traits (52), and we speculate that they may be the primary genetic cause of linkage drag. The genetic architecture of linkage drag has implications for mitigation strategies. If the maladaptive allele is commonly the absence variant of a PAV, then it could be complemented in hybrids containing the domesticated allele, whereas an allele that was maladaptive for other reasons (e.g., additive effect polygenes) is unlikely to be rescued in hybrids. Unfortunately, we were unable to directly test this hypothesis in the present study because the SAM population mainly comprises inbred lines. However, simulation of cross combinations revealed that crosses between HA and RHA lines typically show the highest complementation of introgressions and introgressed PAVs. Thus, if the PAVs are the main cause of linkage drag, then current breeding practices would minimize linkage drag and maximize heterosis. However, complementation of primary and secondary introgressions did not exceed 52% and 71%, respectively, for any cross, possibly indicating that significant gains in productivity could be made by breeding to further enhance complementation of introgressions. Surprisingly, we found that complementation in RHA–RHA crosses could be high as well, perhaps due to multiple origins of fertility restoration in cultivated sunflowers. More generally, our results offer guidance regarding potentially favorable and potentially unfavorable cross combinations.

Phenotypic Consequences of Introgression. Introgressions from the primary gene pool (i.e., wild *H. annuus*) had a significant impact on all 16 traits phenotyped in the SAM population, whereas those from the secondary gene pool affected 13 of the 16 traits (Fig. 5). This is unsurprising since introgressions from wild *H. annuus* are much more frequent in the SAM population than those from the secondary gene pool. On the other hand, the effect sizes of secondary introgressions are much larger on average than those from wild *H. annuus* (*SI Appendix*, Fig. S25).

Examination of the direction of effects of the introgressions indicates that most reduce desirable agronomic trait values, especially for quantitative traits that correlate closely with yield, including head diameter, head weight, seed size, and seed weight, though there are exceptions. For example, introgressions typically increase SLA, which is often used as a surrogate for plant growth in high-throughput phenotyping studies (55). In addition, introgressions show an increase in biomass, but this appears to be a by-product of increased branching, which has been introduced into restorer lines to prolong flowering and, thus, pollen shed. Lastly, while introgressions may negatively affect traits on average, there can be individual introgressions with effects in the opposite direction. For example, an introgression on chr10 from wild *H. annuus* that is associated with increased branching also results in increased oil content and seed size (*SI Appendix*, Table S8). Overall, however, introgressions from wild *H. annuus* negatively affected the latter two traits.

An unexpected result was that higher-frequency introgressions had larger effects on traits (both positive and negative). We speculate that such high-frequency introgressions have been directly targeted by artificial selection. In some instances, the trait we phenotyped was likely the target of selection (e.g., branching and oil content), whereas maladaptive quantitative trait values are most likely the product of linkage drag for qualitative traits such as disease resistance that were not phenotyped in the present study.

Conclusions

In summary, by utilizing a combination of high-quality reference genomes and genotypic and phenotypic analyses of the SAM population, we provide an assessment of the impact of linkage drag on the cultivated sunflower genome and on the performance of inbred sunflower lines. We find that, despite the numerous benefits deriving from tapping CWRs, such as the introduction of desirable traits and genetic and phenotypic variation (17, 20), there can be downsides, including reductions in yield-related traits. We speculate that this is largely due to the introduction of variation in gene content; cultivars containing introgressions not only have new genes, but they also are missing genes that would otherwise be present, which can have deleterious consequences (52).

So, what strategies can be employed to mitigate the effects of linkage drag? Marker-assisted selection is widely employed to reduce the sizes of introgressed regions (27, 70), although this can be challenging in genomic regions of low recombination, such as near the branching locus on chr10. Genome editing and other biotechnology approaches have the potential to introduce favorable alleles without linkage drag (71), although we recognize that the application of such approaches is currently limited by regulatory and socio-political factors (72). If the genetic factors underlying linkage drag are mostly recessive, such as would be the case for missing genes, then hybrid production offers an effective strategy for ameliorating linkage drag, although (at least in sunflower) greater attention should be paid toward enhancing the complementation of introgressions in hybrids. Lastly, our results indicate that introgressions

from distantly related species are much more problematic than those from the fully compatible wild progenitor of cultivated sunflower. Thus, linkage drag could be ameliorated by focusing breeding efforts on closely related and fully compatible wild relatives. While certain desirable traits might not be expressed in close relatives, many of the underlying alleles may exist in the primary gene pool, albeit at a lower frequency. If so, there is a growing potential for the use of bioinformatics approaches to identify compatible GenBank germplasm containing the allele(s) of interest (73). Furthermore, natural introgression from the secondary gene pool into the primary gene pool may provide a source of alleles that have already been purged of deleterious incompatibilities and show reduced linkage drag.

Materials and Methods

For full materials and methods, see *SI Appendix, SI Text*.

Diversity Analyses. To show the relationships of the nine sequenced inbred lines to cultivated sunflower genetic diversity, we positioned them in genetic space using principal components analysis (*SI Appendix*, Fig. S1) based on unpublished genotypic data comprising 16,048 SNP markers genotyped on 2,850 cultivated lines.

Nucleic Acid Extractions, Library Preparations, and Sequencing. For DNA sequencing, high-molecular-weight DNA was extracted from young leaves by using several different protocols, including a modified cetyltrimethylammonium bromide protocol for HA412-HO, magnetic bead extraction for the remaining cultivated genotypes, and the QIAGEN Genomic-tip 100/g procedure for PI659440.

For the HA412-HOv2 genome [which is an updated version of the HA412-HO genome (43)], paired-end and mate-pair libraries were generated and sequenced by using Illumina sequencing technology to a total depth of 214× (Dataset S1). In addition, 10× Genomics Chromium libraries were prepared and sequenced by using Illumina to 37× depth (Dataset S1).

For XRQv2 [which is an updated version of the XRQ genome (43)] and the newly sequenced genotypes, library preparation and sequencing employed PacBio technology (Dataset S1).

Scaffolding. To enable chromosome-level scaffolding of the HA412-HOv2 genome, Hi-C libraries were generated by Dovetail Genomics and sequenced to 49× depth by the McGill University and Génome Québec Innovation Centre. For the XRQv2, PSC8, IR, RHA438, PI659440, and HA89 genomes, scaffolding was aided by Bionano optical mapping.

Genome Assembly. De novo assembly was conducted by using different protocols depending on the genotype, the accuracy of raw sequence data, and the bioinformatics tools available at the time when each genotype was sequenced (Dataset S2). In brief, the HA412-HOv2 genome was assembled with DeNovoMAGIC version 3 (v3) (NRGene Technologies) and scaffolded by using Hi-C sequencing data (Dovetail Genomics) and the HiRise assembler (74).

Contigs for XRQv2, PSC8, IR, and RHA428 were generated by using a meta-assembly approach (75), whereas assembly of the other genomes used canu v2 (76). A first scaffolding step was performed for six genomes (XRQv2, PSC8, IR, RHA438, PI659440, and HA89) by using BNG optical maps, and AllMaps (77) was used to anchor the sequences on the 17 chromosomes for all nine PacBio genomes.

Genome Annotation. Gene models were predicted using the EuGene pipeline (78). Previous RNA sequencing data generated with Illumina sequencing technology (RNA-seq) (43) or PacBio sequencing technology (Iso-Seq) (National Center for Biotechnology Information accession no. PRJNA517222) were used for functional annotation of the HA412-HOv2, XRQv2, and PSC8 genomes. We generated Iso-Seq data for the IR, RHA438, PI659440, and HA89 lines, which were employed for the annotation of each genome. Iso-Seq data for HA89 were used to annotate the LR1, OQP9, and HA300 genomes. Details of the annotation processes, along with assessment results generated with BUSCO v5.1.2 (-m prot -l embryophyta_odb10) software (79), are provided in Dataset S3.

To ensure that we were not overestimating gene-content variation among the 10 sunflower genomes, we developed a pipeline to filter out gene fragments resulting from TE activity and other genomic processes (https://github.com/megahitokiri/Sunflower_annotation_Snakemake) (80). At each step, parameters were fine-tuned by comparison with a set of functionally well-characterized genes to ensure the filtering was not overly aggressive. First, we employed the Extensive de novo TE Annotator to find areas with high content of repeated elements (81). Gene models whose exonic or 3′ untranslated regions overlapped more than 75% with TEs or other repetitive sequences were filtered out. The remaining gene models were further filtered to remove those with pseudogene marks, lacking introns, or that predicted proteins of less than 50 amino acids in length (*SI Appendix*, Table S5).

Identification of Sequence and SVs. Because reference-guided scaffolding of the low-depth genomes (LR1, OQP8, and HA300) can cause spurious results, we only included the six high-contiguity cultivar genomes (HA412-HOv2, XRQv2, PSC8, RHA438, IR, and HA89) to identify sequence variants and SVs. Each of the other five genomes was aligned to the HA412-HOv2 reference by using the nucmer4 program in MUMmer v4 (82). We identified SNPs and small InDels within unambiguous alignment blocks using the show-snps program in MUMmer. We then used SyRI v1.4 (83) to parse the results of MUMmer to identify candidate inversions and intrachromosomal and interchromosomal translocations. Large InDels and CNVs were identified by using SVMU (84).

Identification of Gene Presence and Absence Variation. A pan-genome was constructed for *H. annuus* by using the Roary pan-genome assembler (85), modified to handle eukaryotic gene models, using a minimum threshold for detection of 90%, no splitting of paralogs, and PRANK core genes alignment. Representative sequences generated by Roary were mapped to each reference genome using GMAP (86) with the parameters "-t 12 -O -n 1 -f 2 –min-trimmed-coverage = 0.90 –min-identity = 0.90" to integrate possible missing genes into the annotations.

Identification of Introgressions To identify introgressed regions in the genome assemblies of cultivated sunflower, we employed published resequencing data (*SI Appendix*, Table S7) (40, 44) from native North American landraces and five wild sunflower species (*H. annuus*, *Helianthus argophyllus*, *H. petiolaris*, *Helianthus niveus*, and *Helianthus debilis*) that are probable donors to modern cultivated lines based on breeding records and previous studies (40, 41, 43, 49). For each assembly, raw reads of 48 landrace and wild samples were aligned to the genome, and a VCF file was generated by using a GATK pipeline (*SI Appendix*, *SI Text*). Introgressed regions in the genomes were identified using the "site-by-site" linkage admixture model in STRUCTURE (87, 88).

Genetic Variation Analysis. The densities of SNPs and small InDels were calculated by using vcftools (89) in nonoverlapping 500-kb windows. Windows overlapping with >50% with primary or secondary introgressed regions in at least one, but not all, genomes were defined as polymorphic introgressed windows. Densities of SNPs and small InDels were then compared between polymorphic introgressed regions and nonintrogressed regions. We further annotated functional SNPs by using snpEff v5.0c (90) and calculated the ratio of the number of alternative stop codons ($P_{nonsense}$) and the number of nonsynonymous mutations (P_{nonsyn}) in the 500-kb windows and compared polymorphic introgressed windows and nonintrogressed windows within the same recombination rate category (high: > 2 cM/Mb, reduced: 0.01 to 2 cM/Mb, and null: <0.01 cM/Mb). For the SAM population, we defined polymorphic introgressed windows as those with minor allele frequency (MAF) > 0.01. SNP density and $P_{nonsense}/P_{nonsyn}$ were then calculated in nonoverlapping windows of 500 kb and compared in the same way as for the genome assemblies.

For large InDels and CNVs, in each pair of genomes, we randomly sampled fragments of 500 kb 10,000 times within polymorphic primary introgressed regions, polymorphic secondary introgressed regions, and nonintrogressed regions, respectively. Densities of large InDels and CNVs were calculated and compared between these regions.

We permutated the locations of the inversions identified across the genome assemblies 10,000 times and calculated how often the overlapping size with primary introgressions and secondary introgressions exceeded the observed value, respectively. In each pair of genomes, an inversion was defined as introgression-introduced if one orientation of the inversion overlapped with primary or

secondary introgressions, while the other orientation did not. The incidences of inversions were calculated for polymorphic primary introgressed regions, polymorphic secondary introgressed regions, and regions without introgression.

Effects of Introgression on Gene PAV. To determine how introgression affected gene content, we filtered the table of gene presence–absence polymorphism based on synteny between the genomes as determined by MUMmer4 (82). Using the synteny-filtered table of gene presence–absence polymorphisms, as well as the introgressions identified in each genome, we assigned a single introgression value for each gene in a genome if >50% of the gene overlapped with regions of primary or secondary introgressions. Each missing copy in a genome was assigned an introgression value if the corresponding MUMmer alignment overlapped >50% with regions of primary or secondary introgressions. We compared each of the cultivar genomes to the HA412-HOv2 reference and examined the presence/absence of genes in introgressed and nonintrogressed regions.

Effects of Introgressions on Phenotypic Variation in the SAM Population. We made use of 287 cultivated accessions in the SAM population, which includes close to 90% of cultivated sunflower genetic diversity (45) and was previously sequenced to 5 to 25x depth (40). All 287 accessions, as well as the aforementioned 48 landrace and wild samples (*SI Appendix*, Table S7), were mapped to the HA412-HOv2 reference genome, and an SNP dataset was generated by using a pipeline similar to that described above (*SI Appendix*, *SI Text*). We then used the SNP dataset to identify introgressions from the primary and secondary germplasm in all accessions using the software package PCAdmix (91). Introgression variants were identified by assessing the presence or absence of introgressions in 1-kb nonoverlapping windows across the genome of each sample and filtered for MAF $\geq$ 3%.

We employed data for 16 traits that were generated as part of a common garden study (52–54) and identified associations between introgression variants and the phenotypic traits using EMMAX (92). To further explore the signature of linkage drag on phenotypic data, a ridge-regression BLUP model was used to estimate the effect of each introgression variant on a given trait with the mixed.solve function in R package rrBLUP (93). The mixed.solve function calculates maximum-likelihood/restricted maximum likelihood solutions for mixed models of the form:

$$y = 1\beta + Zg + \varepsilon,$$

where **y** is a vector of the phenotypic trait; **Z** is an incidence matrix containing the allelic states of the introgression variants (**Z** = {−1, 1, 0}); −1 and 1 represent homozygous nonintrogressed and introgressed genotypes at a locus, respectively, and 0 represents the heterozygous state; **β** is a vector of fixed effects; **g** is the vector of introgression variants effects; and **ε** is a vector of residuals.

We extracted **g** as a vector of observed introgression variant effects calculated in the ridge-regression BLUP model. To determine if the overall average effect of **g** is statistically different from the null hypothesis of no effect, average effect values were compared to a null distribution.

To construct the null distribution, we kept the **Z** matrix constant and shuffled the position of observed phenotypes in the **y** vector of the ridge-regression BLUP model. The scheme was repeated 10,000 times, and the average value of estimated **g** in each round was used to build the null distribution of the introgression variant effect. We assessed significance by asking how often the observed average effect exceeds the bounds of the null distribution by calculating the two-tail *P* value.

A linear model (Y ∼ X) was fit to evaluate the effects of frequency on the phenotypic impact of introgression, where Y is the vector of introgression variant effects in the ridge-regression BLUP model and X is a vector of introgression variant frequencies in the SAM population. The beta coefficient of X (slope) can therefore represent the contribution of frequency to the direction and effect size of introgression variants.

Simulation of Cross-Combinations in SAM Population to Predict Complementation of Introgressions. To quantify the extent to which introgressions would be complemented in hybrids, we simulated crosses between all possible combinations of the 286 lines making up the SAM population. Because there was residual heterozygosity in some lines, each cross combination was simulated 100 times to consider different combinations of introgression variants from the two parents. Average heterozygosity for the introgressions for each cross combination was then calculated to provide a measure of predicted complementation for both the primary and secondary gene-pool introgressions. An interactive heatmap-plot format was employed to summarize the simulation results.

We then asked which type of cross was likely to result in the highest level of introgresssion complementation. For this, the SAM population was classified into three main groups: HA, RHA, and other lines. One-way ANOVAs ($P < 0.001$) and Tukey tests (99.9% confidence level) were used to determine if there was a significant difference in the average complementation value of crosses between the different groups.

Data, Materials, and Software Availability. Genome assemblies and annotations are available at https://sunflowergenome.org/assembly-data/ (94), https://www.heliagene.org/HanXRQr2.0-SUNRISE/ (95), and the National Center for Biotechnology Information (NCBI) (SRP373013, SRP373021, SRP373269, SRP373521, SRP373523, PRJNA345532, MNCJ02000000, JAMPTQ000000000, JANGYF000000000, JANGYE000000000, JANRFI000000000, JANJOV000000000, JAMPTL000000000, JANRFJ000000000, JAMPTO000000000) (*SI Appendix*, Table S9) (96, 97). Raw sequences were deposited in NCBI (*SI Appendix*, Table S9) (97). Custom scripts for the analyses are available at GitHub (80, 98, 99).

ACKNOWLEDGMENTS. We thank NRGene and DoveTail Genomics for assembly and scaffolding, respectively, of the HA412-HOv2 genome; and Greg Baute for comments and discussions during the project. This work was supported by the International Consortium of Sunflower Genomics; China Scholarship Council Scholarship 201506380099 (to K.H.); Secretaría de Educación, Ciencia, Tecnología e Innovación de la Ciudad de México (to E.G.G.S.); NSF Plant Genome Program Grant IOS-1444522 (to J.M.B. and L.H.R.); and a National Science and Engineering Research Council of Canada Discovery Grant (to L.H.R.). We are grateful to Compute Canada and the GenoToul bioinformatics platform of Toulouse-Occitanie for providing computing and storage resources.

Author affiliations: [a]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [b]Biodiversity Research Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [c]Laboratoire des Interactions Plantes-Microbes-Environnement, Centre national de la recherche scientifique (CNRS), Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Université de Toulouse, Castanet-Tolosan, F-31326 France; [d]Centre National de Ressources Génomiques Végétales (CNRGV), Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Castanet-Tolosan, F-31326 France; [e]Research and Development Department, NRGene Canada Inc., Saskatoon, SK S7N 3R3, Canada; [f]Gentyane Genomic Platform, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Clermont Ferrand, 63000 France; [g]Plateforme Génome et Transcriptome (GeT-PlaGe), Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Castanet-Tolosan, F-31326 France; [h]MIGAL Galilee Research Institute, Tel-Hai Academic College, Upper Galilee, 11016 Israel; [i]Department of Plant Biology, University of Georgia, Athens, GA 30602; and [j]Department of Biology, University of Victoria, Victoria, BC V8W 2Y2, Canada

1. J. Diamond, Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).
2. J. Diamond, C. Renfrew, Guns, germs, and steel: The fates of human societies. *Nature* **386**, 339–339 (1997).
3. C. Darwin, *The Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (John Murray, London, ed. 6, 1859).
4. T. Dobzhansky, Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **35**, 125–129 (1973).
5. L. T. Evans, *Crop Evolution, Adaptation and Yield* (Cambridge University Press, New York, 1993).
6. S. Tang, S. J. Knapp, Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflower. *Theor. Appl. Genet.* **106**, 990–1003 (2003).
7. C. K. Khoury *et al.*, Crop genetic erosion: Understanding and responding to loss of crop diversity. *New Phytol.* **233**, 84–118 (2022).
8. B. T. Moyers, P. L. Morrell, J. K. McKay, Genetic costs of domestication and improvement. *J. Hered.* **109**, 103–116 (2018).
9. V. Smedegaard-Petersen, K. Tolstrup, The limiting effect of disease resistance on yield. *Annu. Rev. Phytopathol.* **23**, 475–490 (1985).
10. M. Mayrose, N. C. Kane, I. Mayrose, K. M. Dlugosch, L. H. Rieseberg, 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).
11. J. R. Harlan, Our vanishing genetic resources: Modern varieties replace ancient populations that have provided genetic variability for plant breeding programs. *Science* **188**, 618–621 (1975).
12. S. D. Tanksley, S. R. McCouch, Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* **277**, 1063–1066 (1997).
13. S. McCouch *et al.*, Agriculture: Feeding the future. *Nature* **499**, 23–24 (2013).
14. D. L. Plucknett, N. J. Smith, J. T. Williams, A. N. Murthi, *Gene Banks and the World's Food* (Princeton University Press, Princeton, NJ, 1987).

15. D. Zamir, Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* **2**, 983–989 (2001).
16. S. Hübner, M. B. Kantar, Tapping diversity from the wild: From sampling to implementation. *Front Plant Sci* **12**, 626565 (2021).
17. H. Dempewolf *et al.*, Past and future use of wild relatives in crop breeding. *Crop Sci.* **57**, 1070–1082 (2017).
18. A. Gur, D. Zamir, Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol.* **2**, e245 (2004).
19. R. Hajjar, T. Hodgkin, The use of wild relatives in crop improvement: A survey of developments over the last 20 years. *Euphytica* **156**, 1–13 (2007).
20. E. Warschefsky, R. V. Penmetsa, D. R. Cook, E. J. von Wettberg, Back to the wilds: Tapping evolutionary adaptations for resilient crops through systematic hybridization with crop wild relatives. *Am. J. Bot.* **101**, 1791–1800 (2014).
21. PwC, *Crop Wild Relatives: A Valuable Resource for Crop Development* (PricewaterhouseCoopers, London, 2013).
22. L. C. Moyle, E. B. Graham, Genetics of hybrid incompatibility between *Lycopersicon esculentum* and *L. hirsutum*. *Genetics* **169**, 355–373 (2005).
23. D. Tao, K. L. McNally, Y. Koide, K. Matsubara, Reproductive barriers and gene introgression in rice species. *Front Plant Sci* **12**, 699761 (2021).
24. J. Chitwood-Brown, G. E. Vallad, T. G. Lee, S. F. Hutton, Characterization and elimination of linkage-drag associated with Fusarium wilt race 3 resistance genes. *Theor. Appl. Genet.* **134**, 2129–2140 (2021).
25. K. P. Voss-Fels *et al.*, Linkage drag constrains the roots of modern wheat. *Plant Cell Environ.* **40**, 717–725 (2017).
26. J. R. Harlan, J. M. de Wet, Toward a rational classification of cultivated plants. *Taxon* **20**, 509–517 (1971).
27. N. D. Young, S. D. Tanksley, RFLP analysis of the size of chromosomal segments retained around the Tm-2 locus of tomato during backcross breeding. *Theor. Appl. Genet.* **77**, 353–359 (1989).
28. A. Frary, T. M. Fulton, D. Zamir, S. D. Tanksley, Advanced backcross QTL analysis of a *Lycopersicon esculentum* x *L. pennellii* cross and identification of possible orthologs in the Solanaceae. *Theor. Appl. Genet.* **108**, 485–496 (2004).
29. E. Rodgers-Melnick *et al.*, Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3823–3828 (2015).
30. T. Brazier, S. Glémin, Diversity and determinants of recombination landscapes in flowering plants. bioRxiv [Preprint] (2022). https://www.biorxiv.org/content/10.1101/2022.03.10.483889v1 (Accessed 9 June 2022).
31. K. Huang *et al.*, Mutation load in sunflower inversions is negatively correlated with inversion heterozygosity. *Mol. Biol. Evol.* **39**, msac101 (2022).
32. P. Duriez *et al.*, A receptor-like kinase enhances sunflower resistance to *Orobanche cumana*. *Nat. Plants* **5**, 1211–1215 (2019).
33. C. B. Heiser, D. M. Smith, S. B. Clevenger, W. C. Martin, The North American sunflowers (*Helianthus*). *Mem. Torrey Bot. Club* **22**, 1–218 (1969).
34. J. M. Chandler, C.-C. Jan, B. H. Beard, Chromosomal differentiation among the annual *Helianthus* species. *Syst. Bot.* **11**, 354–371 (1986).
35. M. B. Kantar *et al.*, Ecogeography and utility to plant breeding of the crop wild relatives of sunflower (*Helianthus annuus* L.). *Front Plant Sci* **6**, 841 (2015).
36. J. M. Chandler, B. H. Beard, Embryo culture of *Helianthus* hybrids 1. *Crop Sci.* **23**, 1004–1007 (1983).
37. L. Wang *et al.*, The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).
38. O. Smith *et al.*, A domestication history of dynamic adaptation and genomic deterioration in *Sorghum*. *Nat. Plants* **5**, 369–379 (2019).
39. G. L. Owens, G. J. Baute, S. Hubner, L. H. Rieseberg, Genomic sequence and copy number evolution during hybrid crop development in sunflowers. *Evol. Appl.* **12**, 54–65 (2018).
40. S. Hübner *et al.*, Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants* **5**, 54–62 (2019).
41. G. J. Seiler, L. L. Qi, L. F. Marek, Utilization of sunflower crop wild relatives for cultivated sunflower improvement. *Crop Sci.* **57**, 1083–1101 (2017).
42. G. J. Baute, N. C. Kane, C. J. Grassa, Z. Lai, L. H. Rieseberg, 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).
43. H. Badouin *et al.*, The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
44. M. Todesco *et al.*, Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
45. J. R. Mandel, J. M. Dechaine, L. F. Marek, J. M. Burke, 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).
46. S. Terzić *et al.*, Gene banks for wild and cultivated sunflower genetic resources. *OCL Oilseeds Fats Crops Lipids* **27**, 1–14 (2020).
47. S. E. Staton *et al.*, The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. *Plant J.* **72**, 142–153 (2012).
48. L. Gentzbittel *et al.*, A composite map of expressed sequences and phenotypic traits of the sunflower (*Helianthus annuus* L.) genome. *Theor. Appl. Genet.* **99**, 218–234 (1999).
49. F. Vear, Changes in sunflower breeding over the last fifty years. *OCL Oilseeds Fats Crops Lipids* **23**, 1–8 (2016).
50. P. Leclercq, Une sterilite male cytoplasmique chez le tournesol. *Ann. Amelior. Plant.* **19**, 99–106 (1969).
51. S. Renaut, L. H. Rieseberg, The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Mol. Biol. Evol.* **32**, 2273–2283 (2015).
52. J. S. Lee *et al.*, Expression complementation of gene presence/absence polymorphisms in hybrids contributes importantly to heterosis in sunflower. *J. Adv. Res.*, 10.1016/j.jare.2022.04.008 (2022).
53. J. R. Mandel *et al.*, Association mapping and the genomic consequences of selection in sunflower. *PLoS Genet.* **9**, e1003378 (2013).
54. S. U. Nambeesan *et al.*, Association mapping in sunflower (*Helianthus annuus* L.) reveals independent control of apical vs. basal branching. *BMC Plant Biol.* **15**, 84 (2015).
55. S. M. Weraduwage *et al.*, The relationship between leaf area growth and biomass accumulation in *Arabidopsis thaliana*. *Front Plant Sci* **6**, 167 (2015).
56. M. Thudi *et al.*, Genomic resources in plant breeding for sustainable agriculture. *J. Plant Physiol.* **257**, 153351 (2021).
57. M. D. Purugganan, Evolutionary insights into the nature of plant domestication. *Curr. Biol.* **29**, R705–R714 (2019).
58. T. Kuroha *et al.*, Ethylene-gibberellin signaling underlies adaptation of rice to periodic flooding. *Science* **361**, 181–186 (2018).
59. A. A. Temme, K. L. Kerr, R. R. Masalia, J. M. Burke, L. A. Donovan, Key traits and genes associate with salinity tolerance independent from vigor in cultivated sunflower. *Plant Physiol.* **184**, 865–880 (2020).
60. J.-L. Jannink, A. J. Lorenz, H. Iwata, Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genomics* **9**, 166–177 (2010).
61. L. Lei *et al.*, Plant pan-genomics comes of age. *Annu. Rev. Plant Biol.* **72**, 411–435 (2021).
62. P. E. Bayer, A. A. Golicz, A. Scheben, J. Batley, D. Edwards, Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
63. J. L. Gage *et al.*, Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. *Plant Genome* **12**, 180069 (2019).
64. R. Della Coletta, Y. Qiu, S. Ou, M. B. Hufford, C. N. Hirsch, How the pan-genome is changing crop genomics and improvement. *Genome Biol.* **22**, 3 (2021).
65. L. E. Sims, H. J. Price, Nuclear DNA content variation in *Helianthus* (Asteraceae). *Am. J. Bot.* **72**, 1213–1219 (1985).
66. E. J. Baack, K. D. Whitney, L. H. Rieseberg, Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. *New Phytol.* **167**, 623–630 (2005).
67. M. B. Hufford *et al.*, De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
68. M. S. Barker *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).
69. M. S. Barker *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).
70. M. Hao *et al.*, The resurgence of introgression breeding, as exemplified in wheat improvement. *Front Plant Sci* **11**, 252 (2020).
71. K. Kawall, New possibilities on the horizon: Genome editing makes the whole genome accessible for changes. *Front Plant Sci* **10**, 525 (2019).
72. S. Friedrichs *et al.*, An overview of regulatory approaches to genome editing in agriculture. *Biotech. Res. Innovation* **3**, 208–220 (2019).
73. D. Guerra *et al.*, Extensive allele mining discovers novel genetic diversity in the loci controlling frost tolerance in barley. *Theor. Appl. Genet.* **135**, 553–569 (2022).
74. N. H. Putnam *et al.*, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
75. O. Raymond *et al.*, The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
76. S. Koren *et al.*, Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
77. H. Tang *et al.*, ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
78. E. Sallet, J. Gouzy, T. Schiex, "EuGene: An automated integrative gene finder for eukaryotes and prokaryotes" in *Gene Prediction of Methods in Molecular Biology*, M. Kollmar, Ed. (Methods in Molecular Biology, Springer, New York, 2019), vol. **1962**, pp. 97–120.
79. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
80. J. M. Lázaro-Guevara, Code for "The genomics of linkage drag in inbred lines of sunflower." GitHub. https://github.com/megahitokiri/Sunflower_annotation_Snakemake. Deposited 3 January 2022.
81. S. Ou *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
82. G. Marçais *et al.*, MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
83. M. Goel, H. Sun, W.-B. Jiao, K. Schneeberger, SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
84. M. Chakraborty, J. J. Emerson, S. J. Macdonald, A. D. Long, Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
85. A. J. Page *et al.*, Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
86. T. D. Wu, C. K. Watanabe, GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
87. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
88. D. Falush, M. Stephens, J. K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
89. P. Danecek *et al.*; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
90. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
91. A. Brisbin *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 (2012).
92. H. M. Kang *et al.*, Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
93. J. B. Endelman, Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
94. K. Huang *et al.*, Data from "The genomics of linkage drag in inbred lines of sunflower." Sunflower Genome Database. https://sunflowergenome.org/assembly-data. Deposited 3 July 2020.
95. K. Huang *et al.*, Data from "The genomics of linkage drag in inbred lines of sunflower." INRA Sunflower Bioinformatics Resources. https://www.heliagene.org/HanXRQr2.0-SUNRISE. Deposited 13 July 2020.
96. K. Huang *et al.*, Data from "The genomics of linkage drag in inbred lines of sunflower." National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/MNCJ02000000. Deposited 13 July 2020.
97. K. Huang *et al.*, Data from "The genomics of linkage drag in inbred lines of sunflower." National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA345532. Deposited 9 September 2022.
98. K. Huang, Code for "The genomics of linkage drag in inbred lines of sunflower." GitHub. https://github.com/hkchi/Linkage-drag. Deposited 30 August 2022.
99. M. Jahani, Code for "The genomics of linkage drag in inbred lines of sunflower." GitHub. https://github.com/m-jahani/LINKAGE_DRAG. Deposited 19 April 2022.