Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree

Tuomas Hämälä[a,1,2], Eric K. Wafula[b,1], Mark J. Guiltinan[c,d], Paula E. Ralph[b], Claude W. dePamphilis[b,d], and Peter Tiffin[a,2]

[a]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108; [b]Department of Biology, The Pennsylvania State University, University Park, PA 16802; [c]Department of Plant Sciences, The Pennsylvania State University, University Park, PA 16802; and [d]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802

Edited by Loren H. Rieseberg, University of British Columbia, Vancouver, Canada, and approved June 28, 2021 (received for review February 12, 2021)

Genomic structural variants (SVs) can play important roles in adaptation and speciation. Yet the overall fitness effects of SVs are poorly understood, partly because accurate population-level identification of SVs requires multiple high-quality genome assemblies. Here, we use 31 chromosome-scale, haplotype-resolved genome assemblies of *Theobroma cacao*—an outcrossing, long-lived tree species that is the source of chocolate—to investigate the fitness consequences of SVs in natural populations. Among the 31 accessions, we find over 160,000 SVs, which together cover eight times more of the genome than single-nucleotide polymorphisms and short indels (125 versus 15 Mb). Our results indicate that a vast majority of these SVs are deleterious: they segregate at low frequencies and are depleted from functional regions of the genome. We show that SVs influence gene expression, which likely impairs gene function and contributes to the detrimental effects of SVs. We also provide empirical support for a theoretical prediction that SVs, particularly inversions, increase genetic load through the accumulation of deleterious nucleotide variants as a result of suppressed recombination. Despite the overall detrimental effects, we identify individual SVs bearing signatures of local adaptation, several of which are associated with genes differentially expressed between populations. Genes involved in pathogen resistance are strongly enriched among these candidates, highlighting the contribution of SVs to this important local adaptation trait. Beyond revealing empirical evidence for the evolutionary importance of SVs, these 31 de novo assemblies provide a valuable resource for genetic and breeding studies in *T. cacao*.

structural variants | de novo assembly | genetic load | local adaptation | cacao

For more than a century, genomic structural variants (SVs) have been recognized as an important source of functional variation (1–3). Such variants influence the presence, quantity, position, and/or direction of nucleotide sequence, commonly affecting a larger proportion of the genome than single-nucleotide polymorphisms (SNPs) (4–8). SVs also can have large phenotypic effects (9–12) and contribute to adaptation and speciation (13, 14). However, genome-scale analyses have revealed that SVs usually segregate at low frequencies and are depleted from functional regions of the genome, indicating strong purifying selection (4, 7, 15, 16). Despite this general observation, processes responsible for the fitness effects are poorly understood. The fitness consequences of SVs could be due to direct effects on gene function, through disruption of coding regions or regulatory elements (17–19), or the effects may be indirect, arising from suppression of recombination (20, 21).

The suppression of recombination by SVs might play an important role in local adaptation, as theory predicts that SVs can shelter locally beneficial alleles from gene flow by preventing or reducing the formation of viable crossovers within chromosomal heterozygotes (22, 23). Consistent with this expectation, SVs—particularly inversions—have been associated with locally beneficial phenotypes in multiple species, including *Mimulus guttatus* (24), *Zea mays* (25), *Boechera stricta* (26), and *Helianthus annuus* (27). However, the suppression of recombination also has a downside, as it reduces the effective population size (N_e) of the arrangements. The lower N_e can, in turn, weaken the efficacy of purifying selection and thus increase the accumulation of deleterious mutations within the SVs (28, 29).

To examine the fitness consequences of SVs, we constructed chromosome-scale, haplotype-resolved de novo assemblies for 31 wild-collected accessions of *Theobroma cacao* (hereafter cacao). Most previous work on the evolutionary role of SVs has been conducted on short-lived, selfing, and/or domesticated species. By contrast, cacao is a predominantly outcrossing and long-lived perennial with diverse accessions originating from undomesticated populations (30, 31). The combination of wild-collected accessions and SV detection based on diploid assemblies—the gold standard of SV detection (32)—makes our dataset uniquely suited for studying the evolutionary impact of SVs in natural populations. We ask the following questions: What are the overall fitness effects of SVs? Do SVs influence gene expression, suggestive of impaired

Significance

Genomic structural variants (SVs) are frequent contributors to adaptation and speciation, but our understanding of their overall fitness consequences is limited, with data and analyses primarily available for humans and short-lived domesticated species. Here, we use 31 high-quality genome assemblies to study the evolutionary impact of SVs among natural populations of *Theobroma cacao*. We find that most SVs are deleterious and thus constrain adaptation. These detrimental effects likely arise as a direct result of impaired gene function and as an indirect result of suppressed recombination. Yet we detect several SVs that may contribute to local adaptation mainly through traits involved in pathogen resistance. Overall, we provide important insight into processes underlying the fitness effects of SVs in natural populations.

Author contributions: T.H., M.J.G., C.W.d., and P.T. designed research; T.H., E.K.W., and P.E.R. performed research; T.H. contributed new reagents/analytic tools; T.H. and E.K.W. analyzed data; and T.H., E.K.W., and P.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]T.H. and E.K.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: tuomas.hamala@gmail.com or ptiffin@umn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2102914118/-/DCSupplemental.

Published August 18, 2021.

gene function? Do SVs accumulate deleterious nucleotide variants as a result of suppressed recombination? What proportion of SVs are likely contributors to local adaptation?

Results

To identify SVs segregating among 31 wild-collected cacao accessions (Fig. 1*A*), we constructed de novo genome assemblies using 10x Genomics linked-read technology (*SI Appendix*, Fig. S1). Each of our 62 haplotype-specific genome assemblies had a total size between 341 and 387 Mb, with scaffold N50 between 35 and 41 Mb and contig N50 between 94 and 188 Kb (Datasets S1 and S2). The assemblies also captured >96% of the universally conserved single-copy benchmark (BUSCO) genes (Dataset S3) and showed high collinearity with two cacao reference genomes, Matina 1-6 version 1.1 (33) and Criollo B97-61/B2 version 2.0 (34) (*SI Appendix*, Figs. S2 and S3). Overall, our chromosome-scale, haplotype-resolved genome assemblies yielded better qualitative metrics than the two previously published reference assemblies (Fig. 1*B* and *SI Appendix*, Table S1).

By aligning each of the 62 assemblies against the Criollo reference genome, we identified five types of SVs: insertion (INS), deletion (DEL), tandem duplication (DUP), inversion (INV),

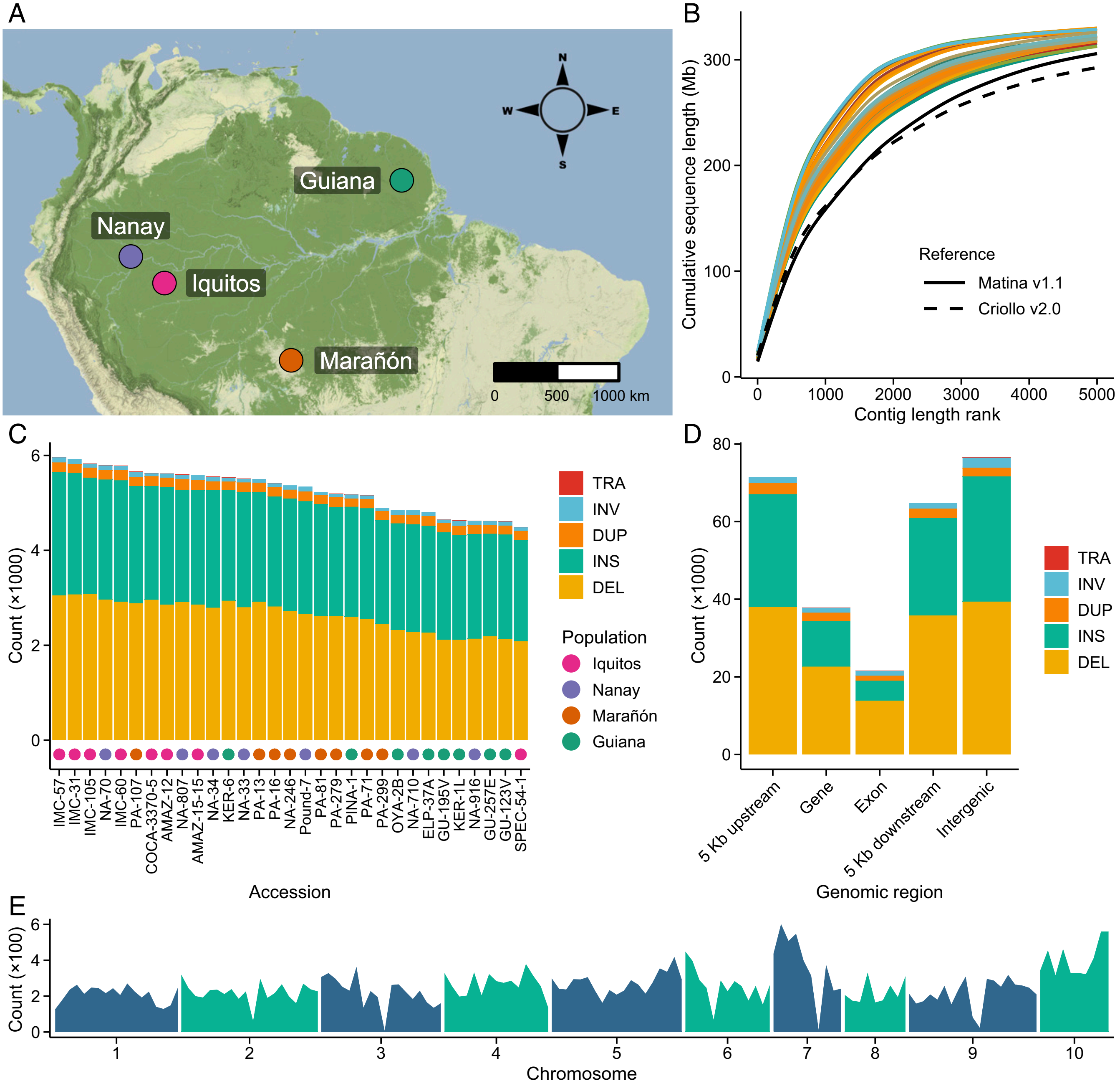

Fig. 1. SVs identified from 62 high-quality genome assemblies. (*A*) Approximate locations of the study populations. (*B*) Sequence contiguity (based on the 5,000 longest contigs) between the two published reference assemblies (black lines) and our 31 assemblies (one haplotype per accession; colored lines). For each of our assemblies, the cumulative sequence length increases faster than in the reference genomes, indicating higher contiguity. (*C*) The number of SVs identified for each accession. (*D*) The number of SVs overlapping different genomic features, counted individually for each accession. (*E*) The number of uniquely located SVs at different regions of the genome, counted in windows of 2 Mb.

and translocation (TRA). We further used the diploid assemblies to distinguish between heterozygotes (SV found in one haplotype) and homozygotes (SV found in both haplotypes). We identified 36,303 uniquely located SVs longer than 50 bp (cumulative total of 163,423 SVs), with 4,610 to 5,963 variants per accession (Fig. 1*C* and *SI Appendix*, Table S2). The proportion of heterozygous SVs per accession was consistent with SNP data (Pearson's $r = 0.97$), although homozygosity was generally higher for SVs than for SNPs (*SI Appendix*, Table S3). In total, these SVs cover a much larger portion of the cacao genome than SNPs and short (<50 bp) insertions/deletions (INDELs) (125 versus 15 Mb). As expected, shorter variants were more common than long ones: 15,829 were <1 Kb, 14,104 were 1 to 10 Kb, 6,186 were 10 to 100 Kb, 152 were 100 Kb to 1 Mb, and 32 were >1 Mb (*SI Appendix*, Fig. S4 and Table S4). Although most SVs were in intergenic regions (Fig. 1*D* and *SI Appendix*, Table S5), 64% of genes annotated in the Criollo genome had SVs overlapping coding regions or putative regulatory elements (≤5 Kb up- or downstream of genes). SVs were present in nearly all parts of the genome, but some regions harbored a considerably higher than average number of SVs (Fig. 1*E*), indicative of SV hotspots. Overall, we detected abundant structural variation in the 31 accessions, suggesting that SVs can have a considerable impact on evolution in cacao.

SVs Are Overall Deleterious. To determine the overall fitness effects of SVs, we examined their frequency patterns. We did this for each SV type, but as there are very few translocations (Fig. 1 *C* and *D*), we excluded them from these and subsequent analyses. We further combined INS and DEL into a joined INDEL type because the mutations underlying INDELs cannot be defined based on a reference genome (e.g., INS in the query could be a DEL in the reference), which might influence the fitness inferences. Minor allele frequency spectra (AFS) showed that SVs segregate at considerably lower frequencies than synonymous or nonsynonymous SNPs (Fig. 2*A*), suggesting that SVs have, on average, more detrimental effects on fitness than SNPs (35). We note, however, that the skewed AFS also may arise from a shift in the mutation-selection balance associated with a limited N_e, a high rate of point mutations, and a considerably lower SV mutation rate (*SI Appendix*, Fig. S5). Nevertheless, simulations conditioned on demography and mutation rate indicated that >85% of new SVs are strongly deleterious, compared to 56% of nonsynonymous SNPs (*SI Appendix*, Fig. S6). The stronger fitness effects of SVs were also supported by a large frequency difference between SVs overlapping functional and (presumably) nonfunctional elements, which was ~6 times greater than the difference observed between nonsynonymous and synonymous SNPs (Fig. 2*B*).

The strong depletion of SVs overlapping functional elements is suggestive of impaired gene function, and we next examined whether relaxed selective constraint allows SVs to be retained at these genes. To do so, we estimated the strength of purifying selection by calculating the ratio of nonsynonymous to synonymous nucleotide diversity (π_N/π_S) within cacao and the ratio of nonsynonymous to synonymous nucleotide divergence between cacao and a close relative, *Herrania umbratica* (d_N/d_S). As purifying selection tends to reduce the frequencies of functional variants, increased π_N/π_S and d_N/d_S is an indication of relaxed selective constraint. We found that both genes located within the

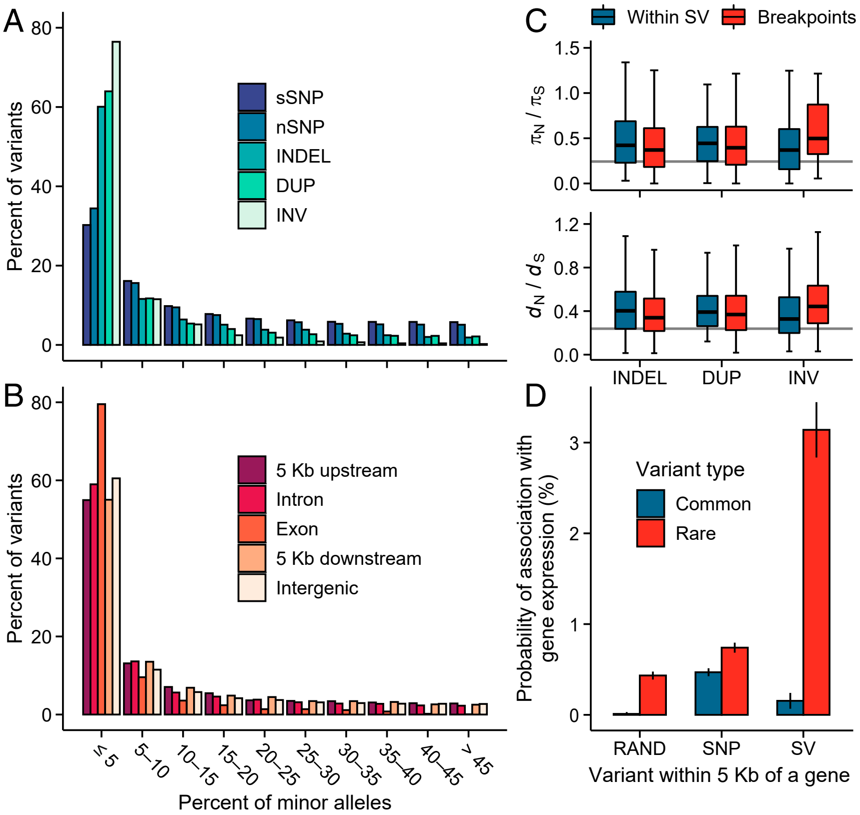

Fig. 2. Fitness effects of SVs. (*A*) AFS for each SV type, compared to synonymous (sSNP) and nonsynonymous (nSNP) nucleotide variants. (*B*) AFS for SVs overlapping different genomic features. (*C*) Measures of selective constraint at genes affected by SVs. Shown are the ratio of nonsynonymous to synonymous nucleotide diversity (π_N/π_S) and the ratio of nonsynonymous to synonymous nucleotide divergence (d_N/d_S) for genes within the SVs and for genes overlapping the SV breakpoints. Gray horizontal lines show medians of control genes. (*D*) The probability that a variant within 5 Kb of a gene is associated with its expression. The SV results are compared to randomized data (RAND) and SNPs. Common variant: MAF > 0.05, rare variant: MAF ≤ 0.05. Error bars show 95% CIs.

SVs and genes overlapping the SV breakpoints had higher-than-expected estimates of π_N/π_S and d_N/d_S (Fig. 2*C*; $P \leq 0.0001$, Wilcoxon rank-sum test). The strength of the effect differed among the SV types: genes overlapping INV breakpoints had higher π_N/π_S and d_N/d_S than genes within INVs, whereas the opposite was true for INDELs and DUPs (Fig. 2*C*). These results indicate that, on average, genes overlapping SVs are under weaker selective constraint, presumably because the majority have nonessential roles in physiology and development.

SVs Influence Gene Expression. To further understand the potential fitness effects of SVs, we studied their impact on gene expression. We first examined how common SVs (minor allele frequency [MAF] > 0.05) influence the expression of nearby genes (≤5 Kb away). Using a naïve quantification approach, we detected 214 genes (3% of tested genes) that were differentially expressed between accessions carrying the major and minor arrangements (*SI Appendix*, Fig. S7). However, after controlling for potential bias resulting from reduced mapping accuracy around SVs and population structure among accessions, considerably fewer SVs harbored evidence of influencing the expression of nearby genes: out of 8,004 tested pairs, only 12 (0.1%) passed the nominal false discovery rate (*Q* value) threshold of 0.1. For these 12 genes, the SV genotypes explained 39 to 84% of the expression variance among accessions (*SI Appendix*, Fig. S8). Although only a small number of genes showed strong signals of their expression being affected by SVs, by randomly assigning genotypes to genes, we found that common SVs were more often associated with gene expression than expected by chance (Fig. 2*D*; $P = 1 \times 10^{-9}$, likelihood ratio test [LRT]). However, repeating the analysis using SNPs revealed that common SVs were ∼3 times less likely to influence gene expression than common SNPs (Fig. 2*D*; $P = 4 \times 10^{-6}$, LRT).

We next assessed the potential of rare SVs (MAF ≤ 0.05) to influence gene expression. Compared to common variants, rare SVs are more likely to have arisen recently and have a detrimental impact on gene expression. By contrasting the expression of the minor arrangements against the distribution of expression values from the major arrangements, we detected 405 SVs (3% of tested SVs) with a putative effect on gene expression (*SI Appendix*, Fig. S8). In contrast to the pattern observed among common variants, rare SVs were ∼4 times more likely to influence gene expression than rare SNPs (Fig. 2*D*; $P < 2 \times 10^{-16}$, LRT). Although we only detected few INVs associated with gene expression (*SI Appendix*, Fig. S8), data on allele-specific expression (ASE) revealed that genes with ASE were overrepresented at INV heterozygotes (*SI Appendix*, Fig. S9), suggesting that INVs influence the expression of genes within them. Taken together, these results suggest that common SVs, which are likely neutral or beneficial, have only a subtle influence on gene expression. Rare SVs, by contrast, show a stronger tendency to impact gene expression, which likely impairs gene function and contributes to the deleterious effects of SVs.

Genetic Load Is Increased at Inversions. Besides directly impairing gene function, the deleterious effects of SVs may arise indirectly through the accumulation of detrimental nucleotide variants. This prediction particularly concerns INVs, as they often result in complete loss of recombinant haplotypes (14, 20, 21). Consistent with the effects of suppressed recombination, we found elevated nucleotide differentiation between the INV arrangements (Fig. 3*A*; $P < 2 \times 10^{-16}$, Wilcoxon rank-sum test) as well as increased genetic linkage within the arrangements (Fig. 3*B*; $P < 2 \times 10^{-16}$, Wilcoxon rank-sum test).

The suppressed recombination is expected to reduce the N_e of the arrangements, potentially resulting in greater accumulation of deleterious nucleotide variants (28). Indeed, INVs harbored a higher proportion of derived alleles at functional sites than collinear regions (Fig. 3*C*; $P < 2 \times 10^{-16}$, LRT), suggestive of increased genetic load (35). Similar increase of derived alleles was found at putatively neutral sites (Fig. 3*C*; $P < 2 \times 10^{-16}$, LRT), likely reflecting weaker background selection at INVs compared to collinear regions (see *SI Appendix*, Fig. S10 for simulation results). We note that, if INVs mainly capture nonessential genes, they may contain more derived alleles even in the absence of SVs. However, the minor INV arrangements also had higher derived allele frequencies than the major arrangements (Fig. 3*C*; $P < 2 \times 10^{-16}$, LRT), indicating that N_e at these regions is reduced due to suppressed recombination (28). The increased genetic load was further supported by mutational effects predicted with SIFT4G (36), which identified an increase of deleterious SNPs at INVs compared to collinear regions (Fig. 3*D*; $P < 2 \times 10^{-16}$, LRT) as well as an increase at minor arrangements compared to major arrangements (Fig. 3*D*; $P = 1 \times 10^{-5}$, LRT). Together, our results support theoretical predictions that suppressed recombination leads to an increase of genetic load at INVs. This increased load is greater in the minor than the major arrangements, likely due to both their lower frequency and lower levels of recombination resulting from the minor arrangements being more often in a heterozygous state (*SI Appendix*, Fig. S10).

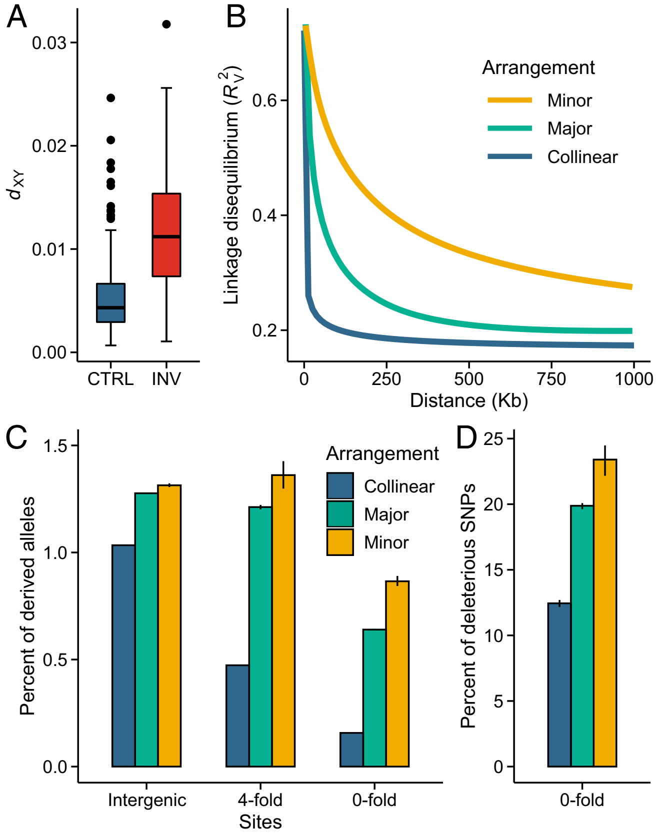

Fig. 3. Genetic load at inversions. (*A*) Absolute nucleotide differentiation (d_{XY}) between the major and minor homozygotes (INV), compared to random collinear regions of equal size (CTRL). (*B*) Average decay of linkage disequilibrium as a function of physical distance in collinear regions and in the major and minor INV homozygotes (estimated using the same sample size). (*C*) Percentage of derived nucleotide alleles in collinear regions and in the major and minor INV homozygotes. Results are divided into intergenic regions (>5 Kb from each gene), synonymous sites (4-fold), and nonsynonymous sites (0-fold). Error bars show 95% CIs. (*D*) The percentage of nonsynonymous SNPs predicted to be deleterious. Error bars show 95% CIs.

SVs Contribute to Local Adaptation in Cacao. Although SVs, as a class, appear to be deleterious and thus constrain adaptation, some may contribute to local adaptation. The 31 accessions in this study were sampled from four environmentally different locations (*SI Appendix*, Figs. S11 and S12), providing an opportunity to identify SVs that contribute to adaptive divergence. We note that SVs show similar, although less pronounced, patterns of population structure than SNPs (Fig. 4*A* and *SI Appendix*, Fig. S13). The weaker signal of population structure seen for SVs (F_{ST}: INDEL = 0.22, DUP = 0.20, INV = 0.13) compared to SNPs (0.24) does not appear to result from gene flow but rather SVs having, on average, a lower MAF than SNPs. By conditioning the variants on MAF, we found that population structure is generally more pronounced at SVs than at SNPs (*SI Appendix*, Fig. S14).

To identify SVs that potentially contribute to local adaptation, we compared F_{ST} estimates of SVs to neutral variants simulated under a demographic model conditioned on N_e, divergence time, and migration rate estimates for these populations (Fig. 4*B*). In each pairwise population comparison, between 136 and 291 SVs had F_{ST} estimates exceeding the 99th percentile of their simulated distributions (total of 846 unique SVs). We consider these SVs as candidates contributing to local adaptation. A large majority of the SVs were INDELs, but five INVs also were identified as outliers. These INVs potentially contribute to adaptive divergence between closely related population pairs Iquitos and Nanay (three INVs) and Marañon and Guiana (two INVs), suggesting that suppressed recombination between the INV arrangements (Fig. 4*C*) may shelter locally beneficial alleles from gene flow. These INVs capture four genes involved in stress responses, cell differentiation, and protein biosynthesis (Dataset S4).

To understand the potential phenotypic effects of the outlier SVs, we examined gene ontology (GO) annotations of the 769 genes within 5 Kb of the 864 candidate SVs. These genes were enriched for 22 GO terms ($Q < 0.1$, hypergeometric test), including terms related to defense against bacteria and oomycetes (*SI Appendix*, Fig. S15). In 140 of these genes, the SV genotypes explained a higher-than-expected proportion of expression variance among accessions (Fig. 4*D* and Dataset S5). Using data on all tested SVs (MAF > 0.05) within 5 Kb of genes, we found that the selection outliers were ∼6 times more likely to affect gene expression than other SVs (*SI Appendix*, Fig. S16; $P = 2 \times 10^{-16}$, LRT). Five GO terms were enriched among the 140 differentially expressed outlier genes, of which "defense response to bacterium" ($Q = 1 \times 10^{-5}$, hypergeometric test) was the most highly enriched. To determine if the same genes would be found through SNP-based analyses, we repeated the outlier analysis using SNPs (weighted F_{ST} estimated for each gene). Although genes identified by the SV-based analysis had higher-than-average SNP-based F_{ST} estimates (*SI Appendix*, Fig. S17), only 47 of the 769 genes were identified as selection outliers using both approaches. We also found no overlap between GO terms enriched among the SV and SNP outliers (*SI Appendix*, Fig. S18).

To further study the role of SVs in local adaptation, we utilized publicly available whole-genome data from 126 cacao accessions (Fig. 4*E*), sampled across the species' native range (30). Of the assembly-based SVs, 3,311 were segregating (MAF > 0.05) in this larger dataset, and our ancestry-based genome scan approach identified 45 as candidates contributing to local adaptation (Fig. 4*F*). Genes annotated as involved in "defense response to bacterium" were strongly enriched ($Q = 8 \times 10^{-6}$, hypergeometric test) among the genes ≤5 Kb from these SVs (Dataset S6). Together, our results suggest that SVs contribute to local adaptation in cacao and that the beneficial fitness effects are likely conferred though traits involved in pathogen resistance.

Discussion

Genomic structural variants (SVs) have been repeatedly implicated in adaptive divergence between populations (24–27), suggesting that SVs have ubiquitous roles in ecological and evolutionary processes (13, 14, 20). Although most empirical studies have focused on these positive effects, the majority of SVs are likely deleterious—either because of their direct impact on gene function or because of suppressed recombination. To better understand the mechanisms behind the fitness effect of SVs in natural populations, we generated high-quality genome assemblies for 31 wild-collected cacao accessions, allowing us to identify a wide range of SVs. Among the 62 genomes, we detected over 160 K SVs, which together cover ∼8 times more of the cacao genome than SNPs and short INDELs.

Consistent with theoretical predictions, the vast majority of these SVs bore signatures of purifying selection, indicating constraint on adaptation. By associating SVs with gene expression, we found that many rare SVs influence the expression of adjacent genes, which likely impairs gene function and contributes to the deleterious impact of SVs. Theory predicts that SVs, particularly inversions, also may constrain adaptation by accumulating deleterious nucleotide variants as a result of suppressed recombination (22, 28). We found empirical support for this prediction by showing that inversions harbored an increased proportion of derived, deleterious alleles at functional sites. In addition, inversion breakpoints were overrepresented in genes under weak purifying selection. There is no reason to assume that inversions are more likely to arise in genes under relaxed selective constraint, but rather this result is consistent with inversions being purged from essential genes. Although the loss of recombinant haplotypes at inversions may occasionally be beneficial (22, 23), our results suggest that an overwhelming majority of inversions are detrimental to fitness.

Despite these overall deleterious effects, a small proportion of SVs carried footprints of local adaptation. The local adaptation candidates had an increased tendency to influence gene expression and were enriched for genes involved in pathogen resistance. The same genes were not identified through SNP-based analyses, demonstrating that SVs need to be considered to gain a full understanding of how selection shapes genomic variation. In crop species, SVs underlie many agronomically important traits (8–10, 15, 19, 37, 38), and natural loss-of-function alleles created by short INDELs have been implicated in drought adaptation in *Arabidopsis* (39). Similarly, impaired gene function by SVs may have locally beneficial effects, if there exists a cost of expressing a phenotype. A cost of pathogen resistance is well documented in plants (40), making it a prime target for SV-mediated local adaptation. Indeed, resistance genes have been linked to SVs in multiple plant species (41), and, in *Arabidopsis*, high SV diversity in resistance genes is likely maintained by balancing selection (16), which may also arise from local adaptation. For example, of the top five local adaptation candidates identified in our analysis (Fig. 4*D*), four have annotated functions involved in pathogen resistance: Tc01v2_g014660, flavonoid glucosyltransferase (42); Tc02v2_g026140, receptor-like protein kinase (43); and Tc02v2_g026130 and Tc07v2_g011300, NBS-LRR (44).

In sum, our analyses provide support for theoretical predictions that most SVs are selected against, likely due to their direct effects on coding sequence or expression as well as due to accumulation of deleterious nucleotide variants. Nevertheless, a subset of SVs bore signatures of local adaptation. By extending the pangenomic approach (41) to wild-collected genotypes, we have demonstrated how SVs can both constrain and facilitate adaptation in natural populations. The genome-scale analyses in a long-lived, undomesticated species broadens the perspective of the fitness effects of SVs that has emerged from work in domesticated species (6, 8, 19, 45, 46) as well as from studies that have focused on individual selectively beneficial SVs (24–27).

Materials and Methods

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

We used 10x Genomics linked-read technology to construct chromosome-scale, haplotype-resolved genome assemblies for 31 wild-collected cacao accessions. With ∼65-fold raw read coverage of each genome, we assembled

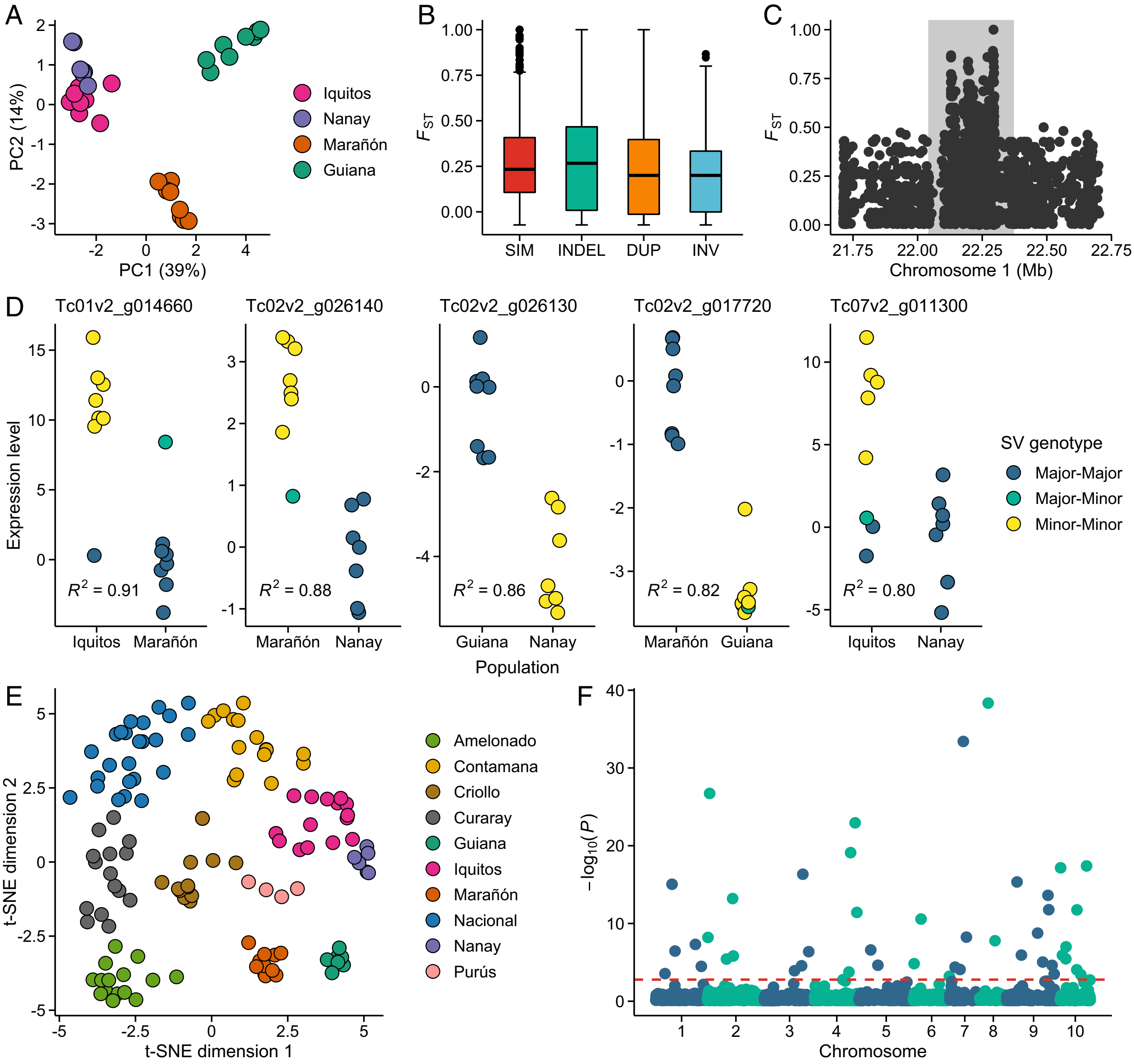

Fig. 4. SVs and local adaptation. (*A*) Variation along the first two eigenvectors of a principal components analysis conducted with SVs. (*B*) The distribution of F_{ST} estimates for the three SV types compared to simulated neutral samples (SIM). (*C*) Example of a nonrecombining haplotype block, caused by a 330-Kb INV (shaded area). Shown are SNP-based F_{ST} estimates between the Iquitos and Nanay populations. (*D*) Expression level at genes affected by selection outlier SVs. Shown are five genes with the largest proportion of expression variance explained by the SVs (R^2). (*E*) Relationship between 126 resequenced cacao accessions (30) used in genotyping our assembly-based SVs. Accessions are labeled according to their largest contributing ancestral population. Shown are t-distributed stochastic neighbor embedding (t-SNE) projections performed on genome-wide SNP data (*SI Appendix*, Fig. S19 shows t-SNE on SV data.) (*F*) *P* values from an ancestry-based genome scan, conducted using 3,011 SVs genotyped for the 126 accessions. Red dashed line: $Q < 0.1$.

the linked reads into phased contigs and scaffolds with Supernova (47) and then used RaGOO (48) to anchor them to the cacao cultivar Matina 1-6 (version 1.1) reference chromosomes (33). We aligned each of the 62 haplotype-specific assemblies against the Criollo B97-61/B2 (version 2.0) reference genome (34) with MUMmer4 (49) and identified SVs from the alignments using MUM&Co (50). We assessed the strength of purifying selection acting on SVs by estimating allele frequency spectra for SVs overlapping different genomic regions and by inferring the distribution of fitness effects using fit∂a∂i (51).

To quantify gene expression, we collected leaf samples from trees maintained in a common garden in Costa Rica. We compiled data on SVs within the vicinity of genes (≤5 Kb) and tested whether common (MAF > 0.05) and rare (MAF ≤ 0.05) SVs were associated with the expression of these genes. Common SVs were analyzed using linear regression models, with DNA-sequencing (DNA-seq) read counts and 10 genotype-based principal components as cofactors to control for reduced mapping accuracy around SVs and population structure among samples, respectively. We tested whether gene expression is affected by rare SVs using absolute median *Z* normalization and the normal cumulative distribution function. For both analyses, we used false discovery rate–based *Q* values (52) to account for multiple testing.

We utilized short-read sequencing data from standard Illumina sequencing to identify SNPs that localized within inversions. We estimated genetic differentiation between the major and minor inversion arrangements using d_{XY} (53), quantified linkage disequilibrium within the arrangements using a method by Mangin et al. (54), and determined the extent of genetic load at the arrangements by estimating the proportion of derived and deleterious alleles at functional sites (35). Ancestral versus derived alleles were inferred using genomic sequences from a closely related species, *H. umbratica*, as an out-group. We predicted mutational effects with SIFT4G (36) to distinguish between deleterious and tolerated SNPs in the inverted regions.

We studied how SVs contribute to local adaptation by searching for highly differentiated variants between the populations. Population differentiation was quantified using the F_{ST} estimator by Hudson et al. (55), and the observed estimates were compared against neutral samples simulated with msprime (56) to detect outliers. We assessed how the local adaptation candidates influence gene function by testing for an association with gene expression using linear models that controlled for technical variation using DNA-seq read counts. We also used Paragraph (57) to genotype our assembly-based SVs using publicly available resequencing data from 126 accessions (30) and searched for differentiation outliers among this larger dataset. To do so, we used multiple linear regression and the Mahalanobis distance to find SVs that best explain the distribution of ancestry proportions among the accessions. Ancestry proportions were estimated with ADMIXTURE (58). We then examined whether SVs involved in the same biological processes, as determined using GO, have been under selection across the species range.

Data Availability. Raw sequence data are available at the National Center for Biotechnology Information Sequence Read Archive under accession number PRJNA558793. Genome assemblies are available at Dryad at the following link: https://doi.org/10.5061/dryad.rfj6q579s. Custom scripts for conducting the analyses are available at GitHub at the following link: https://github.com/thamala/cacaoSV.

ACKNOWLEDGMENTS. We thank J. H. Marden, S. N. Maximova, D. Zhang, B. M. Tyler, and N. Winters for comments and discussions during the project, A. S. Fister and M. E. Leandro-Muñoz for sample collection, L. Landherr Shaeffer and M. G. Perryman for technical support, N. M. Springer and two anonymous reviewers for their comments on improving the manuscript, and Centro Agronómico Tropical de Investigación y Enseñanza (CATIE) for providing access to the cacao germplasm. Computational resources were provided by the Institute for Computational and Data Sciences at The Pennsylvania State University and the Minnesota Supercomputing Institute at the University of Minnesota. This work was supported by the NSF Grant IOS-1546863 and the US Department of Agriculture National Institute of Food and Agriculture, Federal Appropriations under Project PEN04569 and accession number 1003147.

1. A. H. Sturtevant, The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.* **14**, 43–59 (1913).
2. B. McClintock, *Cytological Observations of Deficiencies Involving Known Genes, Translocations and an Inversion in Zea mays* (University of Missouri, College of Agriculture, Agricultural Experiment Station, 1931).
3. C. B. Bridges, The Bar “gene” a duplication. *Science* **83**, 210–211 (1936).
4. 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).
5. W. B. Jiao, K. Schneeberger, Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).
6. Y. Liu *et al*., Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
7. R. L. Collins *et al*.; Genome Aggregation Database Production Team; Genome Aggregation Database Consortium, A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
8. P. Qin *et al*., Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e16 (2021).
9. T. Sutton *et al*., Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* **318**, 1446–1449 (2007).
10. A. Studer, Q. Zhao, J. Ross-Ibarra, J. Doebley, Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**, 1160–1163 (2011).
11. H. Nishikawa *et al*., A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat. Genet.* **47**, 405–409 (2015).
12. C. Küpper *et al*., A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2016).
13. C. Mérot, R. A. Oomen, A. Tigano, M. Wellenreuther, A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* **35**, 561–572 (2020).
14. K. Huang, L. H. Rieseberg, Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front. Plant Sci.* **11**, 296 (2020).
15. Y. Zhou *et al*., The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
16. M. Göktay, A. Fulgione, A. M. Hancock, A new catalog of structural variants in 1,301 *A. thaliana* lines from Africa, Eurasia, and North America reveals a signature of balancing selection at defense response genes. *Mol. Biol. Evol.* **38**, 1498–1511 (2021).
17. C. D. Hirsch, N. M. Springer, Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* **1860**, 157–165 (2017).
18. C. Chiang *et al*.; GTEx Consortium, The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
19. M. Alonge *et al*., Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
20. L. H. Rieseberg, Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
21. B. A. Rowan *et al*., An ultra high-density *Arabidopsis thaliana* crossover map that refines the influences of structural variation and epigenetic features. *Genetics* **213**, 771–787 (2019).
22. M. Kirkpatrick, N. Barton, Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
23. S. Yeaman, Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E1743–E1751 (2013).
24. D. B. Lowry, J. H. Willis, A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
25. Z. Fang *et al*., Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics* **191**, 883–894 (2012).
26. C. R. Lee *et al*., Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nat. Ecol. Evol.* **1**, 1–13 (2017).
27. M. Todesco *et al*., Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
28. E. L. Berdan, A. Blanckaert, R. K. Butlin, C. Bank, Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS Genet.* **17**, e1009411 (2021).
29. P. Jay *et al*., Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nat. Genet.* **53**, 288–293 (2021).
30. O. E. Cornejo *et al*., Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Commun. Biol.* **1**, 167 (2018).
31. T. Hämälä *et al*., Gene expression modularity reveals footprints of polygenic adaptation in *Theobroma cacao*. *Mol. Biol. Evol.* **37**, 110–123 (2020).
32. M. Mahmoud *et al*., Structural variant calling: The long and the short of it. *Genome Biol.* **20**, 246 (2019).
33. J. C. Motamayor *et al*., The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* **14**, r53 (2013).
34. X. Argout *et al*., The cacao Criollo genome v2.0: An improved version of the genome for genetic and functional genomic studies. *BMC Genomics* **18**, 730 (2017).
35. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
36. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
37. L. G. Maron *et al*., Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5241–5246 (2013).
38. J. Y. Choi *et al*., Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* **21**, 21 (2020).
39. J. G. Monroe *et al*., Drought adaptation in *Arabidopsis thaliana* by extensive genetic loss-of-function. *eLife* **7**, 2–21 (2018).
40. J. K. M. Brown, J. C. Rant, Fitness costs and trade-offs of disease resistance and their consequences for breeding arable crops. *Plant Pathol.* **62** (S1), 83–95 (2013).
41. 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).
42. M. Langlois-Meurinne, C. M. M. Gachon, P. Saindrenan, Pathogen-responsive expression of glycosyltransferase genes *UGT73B3* and *UGT73B5* is necessary for resistance to *Pseudomonas syringae* pv tomato in *Arabidopsis*. *Plant Physiol.* **139**, 1890–1901 (2005).
43. E. R. Morris, J. C. Walker, Receptor-like protein kinases: The keys to response. *Curr. Opin. Plant Biol.* **6**, 339–342 (2003).
44. L. McHale, X. Tan, P. Koehl, R. W. Michelmore, Plant NBS-LRR proteins: Adaptable guards. *Genome Biol.* **7**, 212 (2006).
45. M. Jayakodi *et al*., The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).
46. S. Walkowiak *et al*., Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
47. N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, D. B. Jaffe, Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
48. M. Alonge *et al*., RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
49. G. Marçais *et al*., MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
50. S. O'Donnell, G. Fischer, MUM&Co: Accurate detection of all SV types through whole-genome alignment. *Bioinformatics* **36**, 3242–3243 (2020).
51. B. Y. Kim, C. D. Huber, K. E. Lohmueller, Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics* **206**, 345–361 (2017).
52. J. D. Storey, A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).
53. M. Nei, *Molecular Evolutionary Genetics* (Columbia University Press, 1987).
54. B. Mangin *et al*., Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* **108**, 285–291 (2012).
55. R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
56. J. Kelleher, A. M. Etheridge, G. McVean, Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
57. S. Chen *et al*., Paragraph: A graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
58. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).