

ORIGINAL ARTICLE OPEN ACCESS

Comparative Population Genomics of Relictual Caribbean Island *Gossypium hirsutum*

Weixuan Ning¹  | Guanjing Hu^{2,3}  | Daojun Yuan⁴  | Mark A. Arick II⁵  | Chuan-Yu Hsu⁵ | Zenaida V. Magbanua⁵  | Olga Pechanova⁵ | Daniel G. Peterson⁵  | Yating Dong⁶  | Joshua A. Udall⁷  | Corrinne E. Grover¹  | Jonathan F. Wendel¹ 

¹Ecology, Evolution, and Organismal Biology Department, Iowa State University, Ames, Iowa, USA | ²Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China | ³State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, China | ⁴College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China | ⁵Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, Mississippi, USA | ⁶School of Life Sciences, Zhejiang Chinese Medical University, Hangzhou, China | ⁷Crop Germplasm Research Unit, USDA/Agricultural Research Service, College Station, Texas, USA

Correspondence: Jonathan F. Wendel (jfw@iastate.edu)

Received: 4 June 2025 | **Revised:** 30 October 2025 | **Accepted:** 5 January 2026

Keywords: crop domestication | *Gossypium hirsutum* | habitat fragmentation | interspecific introgression | wild cotton

ABSTRACT

Gossypium hirsutum is the world's most important source of cotton fibre, yet the diversity and population structure of its wild forms remain largely unexplored. The complex domestication history of *G. hirsutum* combined with reciprocal introgression with a second domesticated species, *G. barbadense*, has generated a wealth of morphological forms and feral derivatives of both species and their interspecies recombinants, which collectively are scattered across a large geographic range in arid regions of the Caribbean basin. Here we assessed genetic diversity within and among populations from two Caribbean islands, Puerto Rico ($n=43$, five sites) and Guadeloupe ($n=25$, one site), which contain putative wild or introgressed forms. Using whole-genome resequencing data and a phylogenomic framework derived from a broader genomic survey, we parsed individuals into feral derivatives and truly wild forms. Feral cottons display uneven levels of genetic and morphological resemblance to domesticated cottons, with diverse patterns of genetic variation and heterozygosity. These patterns are inferred to reflect a complex history of interspecific and intraspecific gene flow that is spatially highly variable in its effects. Wild cottons in both Caribbean islands appear to be relatively inbred, especially the Guadeloupe samples. Our results highlight the dynamics of population demographics in relictual wild cottons that experienced profound genetic bottlenecks associated with repeated habitat destruction superimposed on a natural ecogeographical distribution comprising widely scattered populations. These results have implications for conservation and utilisation of wild diversity in *G. hirsutum*.

1 | Introduction

Most modern crop plants have undergone thousands of years of domestication, during which human-mediated directional selection has led to variously severe genetic

bottlenecks associated with the favoring of desired morphological traits (Alam and Purugganan 2024; Andersson and Purugganan 2022; Purugganan and Fuller 2009). In contrast, hybridization and introgression, both as natural processes and during crop breeding, serve to increase genetic diversity

Corrinne E Grover and Jonathan F Wendel have contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Molecular Ecology* published by John Wiley & Sons Ltd.

through combining genetic pools from related crop species (e.g., Kathe et al. 2019). These two countervailing processes complicate our ability to unravel the history and genomic structure of modern cultivated species (Gepts 2014). Insights into the domestication history and molecular changes in crop plants thus require analyses of naturally occurring genomic variation across the spectrum, from truly wild species through the various stages spanning original crop domestication to the development of modern cultivars (Alseikh et al. 2021; Olsen and Wendel 2013).

As humans colonised and spread across the Americas, early domesticators became attracted to a number of useful wild plant species (Harlan 1971), including two cotton species (Viot and Wendel 2023), *Gossypium hirsutum* (genome designation AD₁) and *G. barbadense* (AD₂). These species have seeds with special single-celled epidermal seed trichomes, colloquially known as cotton “fibres”, which are short and light brown in their natural wild forms. Human utilisation and domestication of these two wild species have a long history, over 7000 years in *G. barbadense* and over 4000 years in *G. hirsutum* (Brubaker et al. 1999; Brubaker and Wendel 1994; Splitstoser et al. 2016; Wendel et al. 1992; Westengen et al. 2005). This long period of protracted domestication and subsequent improvement occurred in parallel in these two species, which in the wild exist in complete allopatry, with *G. barbadense* originating in South America and *G. hirsutum* in Central America (Viot and Wendel 2023). In both cases, strong human-mediated selection transformed the wild plant architecture from perennial, multi-branched outcrossing shrubs bearing sparse, relatively short fibres into modern, high-yielding annualised selfing crop plants bearing the strong, fine, long fibres that form the basis of the contemporary global cotton trade (Wendel et al. 1992).

Modern cultivars of *G. hirsutum* presently account for the majority of cotton produced globally. Despite this agricultural significance, little is known about the genetic diversity present within the wild forms. This knowledge gap limits our ability to fully elucidate the consequences and mechanisms of domestication and constrains the effective use of genomic resources for crop improvement (Hu et al. 2021). As with many other important crop plants, truly wild forms of *G. hirsutum* are relatively rare (d’Eeckenbrugge and Lacape 2014; Fryxell 1979), and these can be easily confused with both (1) early domesticated (semi-domesticated) forms that variously overlap in their ‘wild’ cotton characteristics (Brubaker et al. 1999; Brubaker and Wendel 1994; Fryxell 1979; Yuan et al. 2021), and (2) feral derivatives that escaped from cultivation over the millennia and became reestablished in various parts of the presumed ancestral species range (Alavez et al. 2021; Brubaker et al. 1999; Brubaker and Wendel 1994; Fryxell 1979; Wendel et al. 1992). These semi-domesticated and feral forms of *G. hirsutum* are quite common throughout the historical and modern cotton growing regions of the subtropics, including much of Central America, the Caribbean, and northern South America. Because of the complexities introduced by protracted domestication across the millennia, geographic diffusion throughout much of the tropical and subtropical Americas, and repeated escape from domestication as reestablished feral populations, the nature of what constitutes truly wild cotton has been fraught and widely discussed

(Brubaker et al. 1999; Brubaker and Wendel 1994; d’Eeckenbrugge and Lacape 2014; Fryxell 1979; Hutchinson 1951; Wendel et al. 1992). This complexity notwithstanding, some efforts have been made to organise this diversity into geomorphological clusters, for example, the seven “races” of wild or feral cotton described by Hutchinson (1951).

Compounding the confusion as to what comprises truly wild populations, the parallel domestication of *G. barbadense* and *G. hirsutum* was accompanied by similar processes of geographic diffusion and escape into nature as reestablished feral derivatives (Percy and Wendel 1990; Wang et al. 2019; Wendel et al. 1992). Although these two species were initially domesticated in different continents, that is, in the Yucatán Peninsula for *G. hirsutum* (Brubaker and Wendel 1994; d’Eeckenbrugge and Lacape 2014; Wendel et al. 1992) and northwest South America for *G. barbadense* (Percy and Wendel 1990; Westengen et al. 2005), they came into contact as a result of thousands of years of human-mediated germplasm diffusion and trade, becoming sympatric in a broad region of the Caribbean and other adjacent locations in northern South America and parts of Central America (Brubaker and Wendel 1994; d’Eeckenbrugge and Lacape 2014; Wendel et al. 1992). This sympatry provided the opportunity for ancient intermingling and introgression (Brubaker et al. 1993; Stephens and Phillips 1972), which now has been documented using extensive whole genome resequencing surveys (Yuan et al. 2021). However, without knowing the genetic composition of truly wild representatives of both species, the inference of introgression is challenging, and thus the extent of introgression between the two species remains unclear.

To shed the light on the nature and extent of the remaining wild cotton genetic diversity, we recently have been collecting *G. hirsutum* from many parts of its indigenous range, with the goal of phylogenomically contextualising this diversity within a global survey of mostly germplasm bank samples, which we previously completed for the species as a whole (Yuan et al. 2021). Our focus here is on the relictual natural populations of *G. hirsutum* on two Caribbean islands, Puerto Rico and Guadeloupe (Figure 1A). The previous survey (Yuan et al. 2021) suggested that wild cotton may exist near the southwestern coastal tip of Puerto Rico. In addition, Ano et al. (1982) described two types of *G. hirsutum* coexisting in the easternmost part (Pointe des Châteaux) of Guadeloupe (Figure 1), one with several morphological characteristics that are more typical of *G. barbadense*, and the other, more extensive population with features that characterise wild *G. hirsutum* (Fryxell 1979), specifically the presence of small capsules and seeds, the latter covered with brown to dark green fuzz and khaki-coloured fibre.

Here we study the newly collected cotton samples from both of these islands using whole genome resequencing of field-collected individuals ($n=25$ and $n=43$ for Guadeloupe and Puerto Rico, respectively). The sequencing data are analysed in the context of our previous global survey of germplasm bank accessions (Yuan et al. 2021) and compared to a recently identified wild cotton population from Mound Key, Florida (Ning et al. 2024). We were particularly interested in assessing: (1) whether the *G. hirsutum* populations in Guadeloupe or Puerto Rico are truly wild, or if instead they represent derivatives from some earlier state of cotton

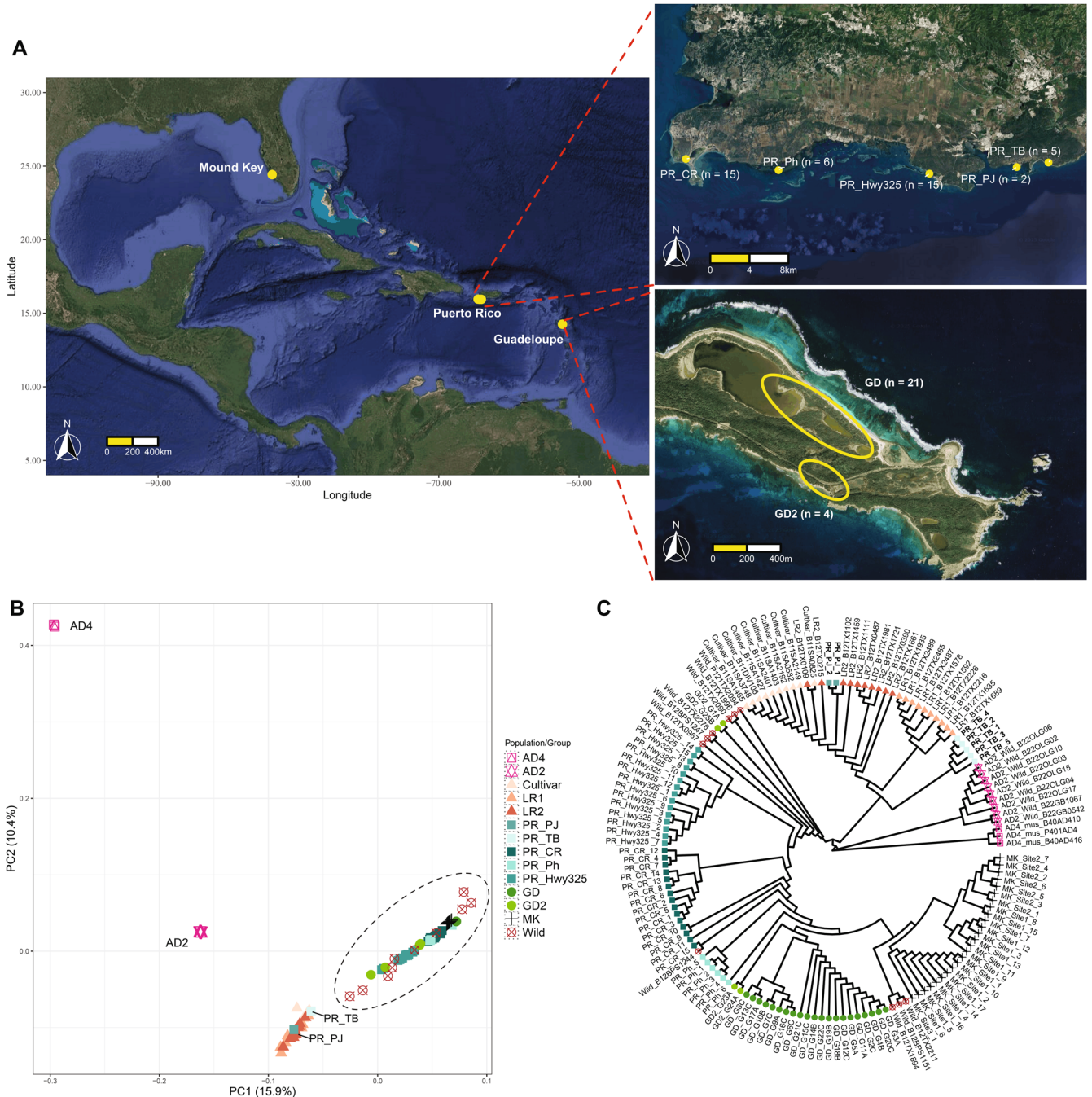


FIGURE 1 | (A) The distribution of field-collected *G. hirsutum* samples from Mound Key, Florida, and the Caribbean islands of Puerto Rico and Guadeloupe. Puerto Rican samples (right top) were collected from five sites along the SW coastline, from east to west: CaboRojo (PR_CR), Pitahaya (PR_Ph), Highway325 (PR_Hwy325), PuntaJacinto (PR_PJ), and TamarindoBeach (PR_TB), dispersed only a few km between each site (see scale bar). The 25 Guadeloupe samples (right bottom) were genetically partitioned into two groups, GD ($n=21$) and GD2 ($n=4$), and were collected from one site at Pointe des Châteaux, Guadeloupe. (B) and (C) PCA and neighbour-joining trees showing genetic relationships among Caribbean cottons and *G. mustelinum* (AD4), *G. barbadense* (AD2), Mound Key (MK), and genetic groups from Yuan et al. (2021): Cultivar, Landrace1 (LR1), Landrace2 (LR2), and Wild groups. Maps plotted using R package ggmap (Kahle and Wickham 2013) and resources from Map Data 2025 Google map (<https://www.google.com/maps>).

cultivation in the Caribbean; (2) how much genetic diversity exists within and between these populations and how different the populations are from each other; and (3) how much heterozygosity there is genome-wide in an average individual, and what this might reveal about population demographic and inbreeding histories.

2 | Materials and Methods

2.1 | Whole Genome Resequencing

In March 2018, seeds from 25 *G. hirsutum* individuals were collected by G. A. from the Pointe des Châteaux site of Guadeloupe

(Figure 1A) and sent to Iowa State University where they were planted in the greenhouse (Table S1). Once mature, specimens from each individual were deposited at the Ada Hayden herbarium (ISC; Voucher IDs: from 455774 to 455797). In January 2023, leaves from 43 *G. hirsutum* individuals were collected by JFW from five sites in southeastern Puerto Rico: CaboRojo (PR_CR), Pitahaya (PR_Ph), Highway325 (PR_Hwy325), PuntaJacinto (PR_PJ), and TamarindoBeach (PR_TB) (Figure 1A), which were immediately dried and stored in silica gel for later DNA extraction.

For the Guadeloupe samples, genomic DNA was extracted from fresh young leaves via Qiagen DNeasy Plant Mini Kit following the manufacturer's protocol. Extracted DNA concentration and quality were measured using the Qubit 1× dsDNA HS Assay Kit on a Qubit 4 Fluorometer and NanoDrop 2000 Spectrophotometer. DNA libraries were constructed and sequenced by BGI Genomics (Hong Kong) using the BGISEQ-500 system to produce 150 bp paired-end reads.

For the Puerto Rican samples, the silica dried leaf tissues (30–50 mg) were ground into fine powder, and genomic DNA was extracted using the Qiagen Plant DNeasy Mini Kit. Purity of the extracted DNA was assessed using a NanoDrop One Spectrophotometer, and DNA integrity was confirmed via agarose gel electrophoresis. Library preparation used 300–500 ng of genomic DNA per sample in the NEBNext Ultra II FS DNA Library Prep Kit with 3-cycle PCR barcoding and enrichment procedure. Genomic libraries were pooled in equimolar amounts and sequenced on a single lane of the Illumina NovaSeq S4 system with 150 bp paired-end reads by Novogene Corporation.

2.2 | Experimental Setup and Genome Variant Calling

To determine the wild or domesticated status of cottons from the Caribbean islands, we analysed 68 individuals (25 Guadeloupe and 43 Puerto Rico) within the context of a global *G. hirsutum* germplasm framework (Yuan et al. 2021). We selected 10 representative individuals each from the inferred natural clusters representing Wild, Landrace1 (LR1), Landrace2 (LR2), and Cultivar groups. In addition, we included all 25 individuals from a recently identified wild *G. hirsutum* population at Mound Key, Florida (Ning et al. 2024). To select the outgroups for genomic introgression tests, we downloaded sequencing data for wild *G. barbadense* (genome designation AD₂; $n = 18$) and *G. darwinii* (AD₅; $n = 5$) from Yuan et al. (2021), and a more distantly related species *G. mustelinum* (AD₄; $n = 3$) (Tables S1 and S2).

Raw reads were trimmed using Trimmomatic V.0.39 (Bolger et al. 2014) to remove sequencing adapters and low-quality bases ('ILLUMINACLIP:Adapters.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:75'). Trimmed paired-end reads were mapped to the *de novo* assembled reference genome of wild *G. hirsutum* accession TX2094 v2.0 (unpub. data) via BWA v0.7.17 with flag '-R' to add read group information (Li and Durbin 2009). For samples split over several lanes, multiple bam files were generated, each containing relevant read group information (<https://github.com/Wendellab/CaribbeanAD1>). The

Sentieon DNaseq variant-calling pipeline (Kendig et al. 2019) was applied to genotype variants from each bam file. Specifically, duplicated reads were removed ('--rmdup') and small indels were realigned ('--algo realigner') to increase precision. From the final realigned bam file, we performed base quality score recalibration ('--algo QualCal') and calculated coverage depth ('--algo CoverageMetrics'), before calling genomic variants ('--algo Haplotyper --emit_mode gvcf') for all 159 samples.

Using these gVCF files, we joint-genotyped samples independently using Sentieon ('--algo GVCFTyper') in seven groups: (1) Mound Key, (2) Puerto Rico, (3) Guadeloupe, (4) the group of 40 reference *G. hirsutum* germplasm samples described above, and (5,6,7) the three outgroup species. For each VCF, we filtered for biallelic SNPs with average read coverage between 10 and 100 reads per site ('--remove-indels --max-missing-count 0 --max-alleles 2 --min-meanDP 10 --max-meanDP 100 --mac 2') and invariant sites with the same read coverage ('--remove-indels --max-maf 0 --min-meanDP 10 --max-meanDP 100') via VCFtools v0.1.16; filtered variant and invariant sites were then remerged ('concat') via bcftools v1.19 (Li 2011) to create the final VCF for each group.

2.3 | Identification of Outgroups and Feral Cottons

To select *G. barbadense* accessions most likely to be truly wild among the 18 sequenced samples previously identified as putatively wild (Yuan et al. 2021), and acknowledging the possible genomic introgression from *G. hirsutum* into both wild *G. barbadense* and *G. darwinii* (Wendel and Percy 1990; Yuan et al. 2021), we first combined the VCFs from Mound Key, *G. darwinii*, *G. barbadense* and *G. mustelinum* ('merge') and refiltered the biallelic sites ('-m2 -M2 -i 'F_MISSING=0' -q 0.001:minor') via bcftools (8,090,831 sites). SNP sites with high levels of pairwise linkage disequilibrium (LD) were removed (i.e., LD-pruned SNPs) via PLINK v1.9 ('--indep-pairwise 50 10 0.1') (Purcell et al. 2007). The LD-pruned VCF (520,520 sites) was used to explore population genetic structure via Principal Component Analysis (PCA) ('--pca 20') and neighbour-joining tree analysis ('--distance square 1-ibs') using PLINK and the ape v5.8.1 package (Paradis et al. 2004) in R v4.4.1. Using the same set of filtered SNPs, we analysed population genetic structure using LEA v3.18.0 (Frichot and François 2015), with K (the number of ancestral populations) ranging from 1 to 10, with each K replicated 10 times. Because *G. darwinii* is sister to *G. barbadense* (Grover et al. 2015), we required the selected wild *G. barbadense* to be both monophyletic in the neighbour-joining tree and sister to *G. darwinii* (details see Figure S1 and Table S2), leaving nine remaining *G. barbadense* samples.

VCFs containing all 145 *G. hirsutum* samples (Tables S1 and S2) from Puerto Rico, Guadeloupe, and Mound Key, as well as the four inferred groups from Yuan et al. (2021), were merged with those nine wild *G. barbadense* and three *G. mustelinum* accessions. This combined dataset was used to evaluate the *G. hirsutum* samples for potentially cryptic feral representatives among the Caribbean cottons. Using the same steps as above, we filtered the 35,453,021 biallelic SNPs from the joined VCF to retain only 4,200,388 high-confidence, LD-pruned sites for subsequent PCA and neighbour-joining tree analyses. Results

were plotted in R using ggplot2 v3.5.1 (Wilkinson 2011) and gg-tree v3.14.0 (Yu et al. 2017).

After identifying two Puerto Rican cotton populations (PuntaJacinto and TamarindoBeach) as putatively non-wild feral derivatives due to their placement among domesticated *G. hirsutum* groups, we removed those two populations (7 individuals) from the analysis of wild population diversity along with the 40 germplasm collection samples from Yuan et al. (2021), which are of less certain provenance. This subset of 86 samples, representing only putatively wild cotton populations, was extracted via bcftools ('view -S -m2 -M2 -i 'F_MISSING=0' -q 0.001:minor') from the combined VCF of 145 samples, to retain only the Mound Key cottons, the Guadeloupe cottons, and the remaining non-feral Puerto Rican cottons. Relationships among these putative populations were evaluated by PCA using the surviving 1,039,022 filtered SNPs, with the first three PCs visualised via a modified script 3D-PCA-plot.py (<https://github.com/Siavash-cloud/3D-PCA-plot>). In addition, we calculated genetic relatedness among these 86 samples using the PLINK identity-by-descent test ('--genome') and plotted the PI_HAT values higher than 0.3 with ggplot2.

2.4 | Genomic Introgression Tests

To evaluate potential gene flow among *G. hirsutum* populations and reciprocal genomic introgression with *G. barbadense*, a dataset of 128 individuals represented by 33,774,815 sites containing 3,292,507 independent biallelic SNPs ('--indep-pairwise 50 10 0.1') was extracted from the full dataset of the combined 145 samples and 35 million SNPs described above. The selected samples included 25 Mound Key, 25 Guadeloupe, 36 wild Puerto Rico (excluding PuntaJacinto and TamarindoBeach), 30 of the 40 *G. hirsutum* samples (omitting the 10 wild samples) based on Yuan et al. (2021), 9 wild *G. barbadense*, and 3 *G. mustelinum* samples.

Using the ~3.3 million filtered biallelic SNPs, we first estimated the number of ancestral populations ($K = 1-18$) via LEA, with each K replicated 10 times. The best K was selected based on the lowest cross-entropy value (Frichot and François 2015). From these results and using geographic considerations, we defined 10 groups/populations among the sampled *G. hirsutum* individuals.

Gene flow was estimated using the maximum likelihood-based approach implemented in TreeMix v1.13 (Pickrell and Pritchard 2012) by first calculating the allele frequency at each of the 3.3 million SNP sites in PLINK (--freq) for each population/group/species. For TreeMix, the four *G. mustelinum* individuals were assigned as outgroup ('-root') and a block size of 1000 SNPs per window ('-k') with bootstrap replications ('-bootstrap') was used (<https://github.com/Wendellab/CaribbeanAD1>). In addition, each 'm' (i.e., number of estimated gene flow) had five replicate runs, and the best migration model was selected using the R package OptM (Fitak 2021). The final network was visualised using the R script "plotting_funcs.R" in TreeMix.

Putative genomic introgression was evaluated via the ABBA-BABA (or D-statistic-based) method using Dsuite (Malinsky

et al. 2021). Specifically, the topology of the maximum likelihood TreeMix non-network was rooted using the outgroup *G. mustelinum* ('-k 1000', '-root') and with the final filtered 3.3 million SNPs as input, we calculated the tree branches that showed strong introgression signals ('Fbranch') between 10 *G. hirsutum* populations/groups or *G. barbadense*.

2.5 | Phylogenetic Analysis of Plastomes

The conserved and uni-parentally inherited whole chloroplast DNA (i.e., the plastome) has been widely used in cotton comparative genomics (Chen et al. 2017; Wendel and Albert 1992; Wu et al. 2018; Yan et al. 2024). We *de novo* assembled plastomes using GetOrganelle v1.7.7.1 (Jin et al. 2020) ('get_organelle_from_reads.py'), with default settings, for all 159 samples in this study. These assemblies, along with 339 previously published cotton plastome sequences (Yan et al. 2024), were first aligned via MAFFT v7.508 (Katoh and Standley 2013) to check assembly quality. One sample (AD5_dar_BYU50004) had a large number of mismatch sites, likely due to contamination, and hence was removed from further analysis.

For plastome annotation, one Guadeloupe sample (GD_G7C) was selected and annotated with an online tool GeSeq (Tillich et al. 2017), using a *G. tomentosum* reference sequence (NC_016690) as a basis (Xu et al. 2012). The resulting annotation file was used to transfer annotations to the remaining 157 aligned samples via PAG (Qu et al. 2019). Using a customised script (<https://github.com/Wendellab/CaribbeanAD1>), we then extracted 110 chloroplast gene sequences present across all 158 individuals, including 81 protein-coding genes and 29 tRNA-coding genes. Each gene was aligned via MAFFT, and all gap sites ('-nogaps') were removed via trimAl v1.5 (Capella-Gutiérrez et al. 2009). We then concatenated ('concat') all trimmed gene sequences into one sequence via SeqKit v2.9.0 (Shen et al. 2016).

We constructed cpDNA-based phylogenies for these 158 samples using (1) an alignment of the whole plastome, with the internal repeat B (IRB) region removed using the annotation file of GD_G7C and SeqKit, and gap sites trimmed by trimAl; and (2) a concatenated sequence of the 110 genic loci. Maximum Parsimony trees were constructed using PAUP* v4 (Swofford 1998) and employing a Heuristic search with random addition and tree-bisection-reconnection (TBR) branch-swapping. Given the slow evolutionary rate of cpDNA, we also tabulated the number of parsimony informative sites supporting the backbone of the phylogenies in PAUP* using 'describeTrees'. In addition, we calculated a haplotype network via PopArt v1.7 (Leigh and Bryant 2015) using the TCS method (Templeton et al. 1992) for both datasets. Complementing the above analyses, the same datasets were used for maximum likelihood phylogenetic reconstructions using IQTree2 (Minh et al. 2019) with bootstrapping ('-B 1000').

2.6 | Estimation of Genomic Diversity

To quantify genomic diversity within and among populations, we first excluded five outlier samples that did not group with other samples collected from the same site (four in GD2 and

one PR_Ph_6 from Pitahaya site; see Results) on the PCA. VCFs containing both variant and invariant sites for the remaining 121 samples from the Caribbean, Mound Key, and the four germplasm groups were merged via bcftools. To better characterise genomic diversity for wild cottons with small, inbred populations, such as those from Guadeloupe cotton, we additionally removed fixed heterozygosity sites across all individuals in each collection site via bcftools ('--exclude "F_PASS(GT='het')=1"'), to reduce the pseudo-heterozygosity genotypes that arise from mapping errors due to repetitive regions or reference genome biases. These samples were divided into nine groups based on genetic relationships inferred by LEA (see Results). For each group, genome-wide nucleotide diversity (π) was calculated using pixy v2 (Bailey et al. 2025), with a sliding-window size of 10 kbp. To ensure that comparisons between different populations/groups were not affected by sample size differences, we randomly downsampled each population/group to five individuals (matching the smallest group, Pitahaya (Ph), $n = 5$) and replicated this downsampling 20 times. Final π values were calculated by counting the average values of each sliding window across all 20 replicates. Using the same downsampling and window setting, pairwise differences of sequence diversity (d_{xy}) and population genetic differentiation (F_{st}) among the nine populations/groups were also calculated in pixy.

2.7 | Assessment of Homozygosity, Heterozygosity, and Linkage Disequilibrium

The length and frequency of long uninterrupted homozygous sites (i.e., identical haplotypes) can be informative with respect to population structure, history, and breeding system (e.g., Ceballos et al. 2018; Kumar et al. 2020). To explore this in cotton, long runs of homozygosity (ROH) were identified for each of 121 samples using 26,112,044 filtered biallelic SNPs (described above) via the 'slidingRUNS.run' function in the R package detectRUNS v0.9.6 (Biscarini et al. 2018). This method applied sliding window-based detection, similar to PLINK ROH detection (Meyermans et al. 2020), by specifying the sliding window size as 15 (windowSize=15), the threshold of overlapping windows of the same state to 0.05, the minimum number of SNPs to 10, the maximum number of heterozygous SNPs and the maximum number of missing SNPs in a sliding window both to 1 (maxOppWindow=1, maxMissWindow=1), the maximum distance between consecutive SNPs to 1 Mbp, the minimal length of ROH to 0.25 Mbp, and the lowest SNP density per kbps to 0.001. Using a customised R script (<https://github.com/Wendellab/CaribbeanAD1>), we tabulated ROH into three length categories: 0.25 to 1 Mbp, 1 to 2 Mbp, and larger than 2 Mbp, and calculated the proportion of total ROH lengths compared to the whole genome using the reference TX2094 V2 for each individual, that is, the ROH inbreeding coefficients (F_{ROH}).

The proportion of heterozygous sites among the 26 million SNPs was estimated using VCFtools ('--het') for the 121 samples. Correlations between linkage disequilibrium (r^2) and all pairs of SNP physical distance (in bp) were estimated for the nine groups using the same set of SNPs and PopLDdecay (Zhang et al. 2019). To account for the sample size differences in LD

decay comparison, we randomly downsampled each population to five individuals (PR_Pitahaya $n = 5$) and repeated this 20 times, then calculated the average r^2 across all replicates and for each distance window.

2.8 | Population Demographic Inference

Whole-genome resequencing data offers the possibility of exploring past demographic processes or selection using contemporary allele frequencies (Marchi et al. 2021). Toward that end, population demographics of the four cotton populations that had larger sample sizes and were inferred to be wild (Mound Key, Guadeloupe, and wild Puerto Rican cottons from Highway325 and CaboRojo; see Results) were first analysed by calculating Tajima's D (Tajima 1989) with a sliding-window size of 10 kbp via pixy v2 using a previous filtered, combined-VCF file that contains variant and invariant sites as input.

To more precisely characterise site allele frequencies, we used VCFtools to subset ('--keep') and filter ('--remove-indels --max-missing-count 0 --min-alleles 2 --max-alleles 2 --min-meanDP 10 --max-meanDP 100 --maf 0.001') the raw output VCF files from Sentieon for a total of 81 samples collected from each island, using bcftools to retain biallelic sites and exclude sites with fixed heterozygosity ('--exclude "F_PASS(GT='het')=1"'). The observed site frequency spectra (SFS) were tabulated via VCFtools ('--counts') and awk command line. In addition, the expected SFS under neutrality was estimated using θ_w (Hudson 2015).

Changes in effective population size (N_e) over evolutionary time can be indicative of bottlenecks or population expansion (Mather et al. 2020). To reconstruct N_e histories in wild cotton populations, we applied SMC₊₊ (Terhorst et al. 2017), which leverages both sequentially Markovian coalescence and the site frequency spectrum across all sampled individuals to improve resolution of recent demographic changes. We converted the VCF containing ~26 million filtered SNPs into SMC₊₊ input format ('vcf2smc') with a bed file to mask the reference genome repetitive transposable element regions ('--mask') for five populations: PR_CaboRojo ($n = 15$), PR_Hwy325 ($n = 15$), PR_Pitahaya ($n = 5$), Mound Key ($n = 25$), and Guadeloupe ($n = 21$). N_e changes were estimated from the output and plotted using a previously noted Malvaceae mutation rate 4.56×10^{-9} (Grover et al. 2017) and a generation time ('-g') of 1 year.

3 | Results

3.1 | Inferring Truly Wild *G. barbadense* for Outgroup Purposes

Given that reciprocal introgression is known to have occurred between *G. barbadense* and *G. hirsutum* (Brubaker et al. 1993; Stephens and Phillips 1972; 2021), selecting *G. barbadense* accessions that are known to be truly wild is essential for testing potential genomic introgression from *G. barbadense* into wild *G. hirsutum*. Using 0.5 million independent, whole-genome, biallelic SNPs extracted from 25 Mound Key (wild)

G. hirsutum, 5 *G. darwinii*, 3 *G. mustelinum*, and 18 *G. barbadense* individuals previously identified as putatively wild (Yuan et al. 2021), we identified nine *G. barbadense* samples that represented a monophyletic clade. Although all wild *G. barbadense* and *G. darwinii* formed one cluster in the first two principal components based on the whole-genome data (39.9% variance explained by PC1 and 25.5% by PC2) (Figure S1A), the population genetic structure analysis suggested that nine *G. barbadense* individuals (dashed box, Figure S1C) were more uniform, particularly for $K=4$. These samples were also monophyletic in the neighbour-joining tree (Figure S1B), containing one type of chloroplast DNA (see plastome analysis, below), and were collectively sister to the truly wild, non-domesticated endemic species from the Galapagos Islands, *G. darwinii* (Grover et al. 2015; Wendel and Percy 1990). These samples, accordingly, were used to represent purely wild *G. barbadense*.

3.2 | Genetic Structure and Relationships Among Major *G. hirsutum* Genetic Groups

Whole-genome resequencing yielded high coverage for both Guadeloupe ($\geq 46\times$; average $48\times$, $n=25$) and Puerto Rican samples ($\geq 20\times$; average $25\times$, $n=43$) (Table S1). Combining the genetic data from the total of 145 samples from the Caribbean ($n=68$), Mound Key ($n=25$), previously identified germplasm groups ($n=40$; Yuan et al. 2021, see methods), purely wild *G. barbadense* ($n=9$), and the phylogenetic outgroup *G. mustelinum* to the other allopolyploid *Gossypium* species ($n=3$), we extracted 4.2 million high-quality, genome-wide, independent biallelic SNPs to assess their genetic relationships.

Principal component analysis (PCA) broadly showed four groups in the first two PCs (Figure 1B), accounting for 15.9% and 10.4% of the genetic variance, respectively. In addition to the two groups comprising *G. mustelinum* and *G. barbadense*, *G. hirsutum* accessions were divided into two groups, one containing the domesticated cottons comprising Landrace1 (LR1), Landrace2 (LR2), and Cultivars, and the other containing all Wild and Mound Key (MK) cottons (denoted by a dashed circle in Figure 1B). Thus, there is a primary genetic division between purely wild *G. hirsutum* and other germplasm that has been through a long history of domestication and possible introgression with *G. barbadense*. Notably, eight samples collected from two Puerto Rico sites: PuntaJacinto (PJ) and TamarindoBeach (TB) grouped with the domesticated cottons, whereas all remaining Puerto Rico (PR) and Guadeloupe (GD) samples nested with the other wild cottons from Mound Key and elsewhere. In addition, compared to Mound Key, all Caribbean cottons collected from both islands exhibited larger variation within the Wild cotton cluster.

A neighbour-joining tree constructed from the same set of 4.2 million SNPs supported a similar depiction and resolution of the four groups: *G. mustelinum*, *G. barbadense*, wild cottons, and domesticated cottons (Figure 1C). Within domesticated cottons, TamarindoBeach (PR_TB) formed a clade that was sister to all Landrace1 samples, and PuntaJacinto (PR_PJ) was nested with the clade that included interdigitating samples representing groups of Landrace2 and Cultivars. As for wild

cottons, only Mound Key, Puerto Rico site Hwy325 and the main Guadeloupe group (GD, $n=21$) were monophyletic. By contrast, Puerto Rican cottons collected from sites CaboRojo (PR_CR) and Pitahaya (PR_Ph) were mixed and paraphyletic to each other, and four outlier samples (GD2) from Guadeloupe exhibited two divergent lineages, with one (GD2_G24A with GD2_G23A) sister to the main Guadeloupe group (GD, $n=21$), and the other one (GD2_G25B with GD2_G1A) nested among all other wild cottons (Figure 1A,C). These results are interpreted as genetic evidence that the TamarindoBeach and PuntaJacinto populations represent feral derivatives of domesticated cottons that escaped to become reestablished as a component of the native vegetation (Alavez et al. 2021; Fryxell 1979; Wendel et al. 1992), and suggest some admixture among the cottons from different locations.

3.3 | Detecting Outliers Among Wild Caribbean Cottons

To further explore genetic relationships of the outliers among wild Caribbean cottons, we reconstructed the PCA using approximately one million filtered SNPs from 86 samples of wild cotton from Puerto Rico, Guadeloupe, and Mound Key. The first principal component (Figure 2A) explained 23.8% of the variation and separated Mound Key from the remaining Caribbean cottons. In PC2 (12.01%) and PC3 (8.5%), the Caribbean cottons clustered into three loose groups: (1) the Guadeloupe cotton with a main group (GD, $n=21$) and four more diverged outliers (GD2, $n=4$), (2) the Puerto Rican population Highway325 (PR_Hwy325), and (3) the remaining group of Puerto Rican cottons (CaboRojo and Pitahaya). The four Guadeloupe outlier samples (GD2) were grouped in two sets of two individuals each (GD2_G25B with GD2_G1A, and GD2_G24A with GD2_G23A). In addition, the sample PR_Ph_6 collected from the Puerto Rican Pitahaya site (Figure 1A) was closer in multivariate space to the Guadeloupe main cotton group compared to the other five Pitahaya samples.

Neighbour-joining tree (Figure 2B) recapitulates the PCA patterns, whereby most wild cotton populations were monophyletic, with the aforementioned exceptions of Pitahaya (PR_Ph_6) and the four Guadeloupe outlier samples (GD2). Specifically, two samples (GD2_G24A and GD2_G23A) formed a separate lineage sister to the Guadeloupe main group (GD, $n=21$) and notably adjacent to a lineage composed of solely sample PR_Ph_6, whereas the other two outliers (GD2_G1A and GD2_G25B) formed a lineage that was in between the Mound Key and PR_Hwy325 clades.

Genetic kinship analysis was explored using the metric PI_HAT for the 86 cotton samples used in this analysis. The relatedness within the main populations of Guadeloupe (GD) and Pitahaya ($n=5$) showed mean PI_HAT values of 0.8, respectively, indicating a high level of kinship. This is in contrast to Mound Key cottons whose two geographically close (about 200 m) subpopulations (Ning et al. 2024) displayed an average PI_HAT of 0.57 (Figure S2). Notably, the two sets of four Guadeloupe outliers exhibited lower levels of relatedness, both with an average PI_HAT of around 0.46. Puerto Rican cottons from the CaboRojo and Highway325 sites had the lowest relatedness among individuals within sites, being approximately 0.36 and 0.42, respectively.

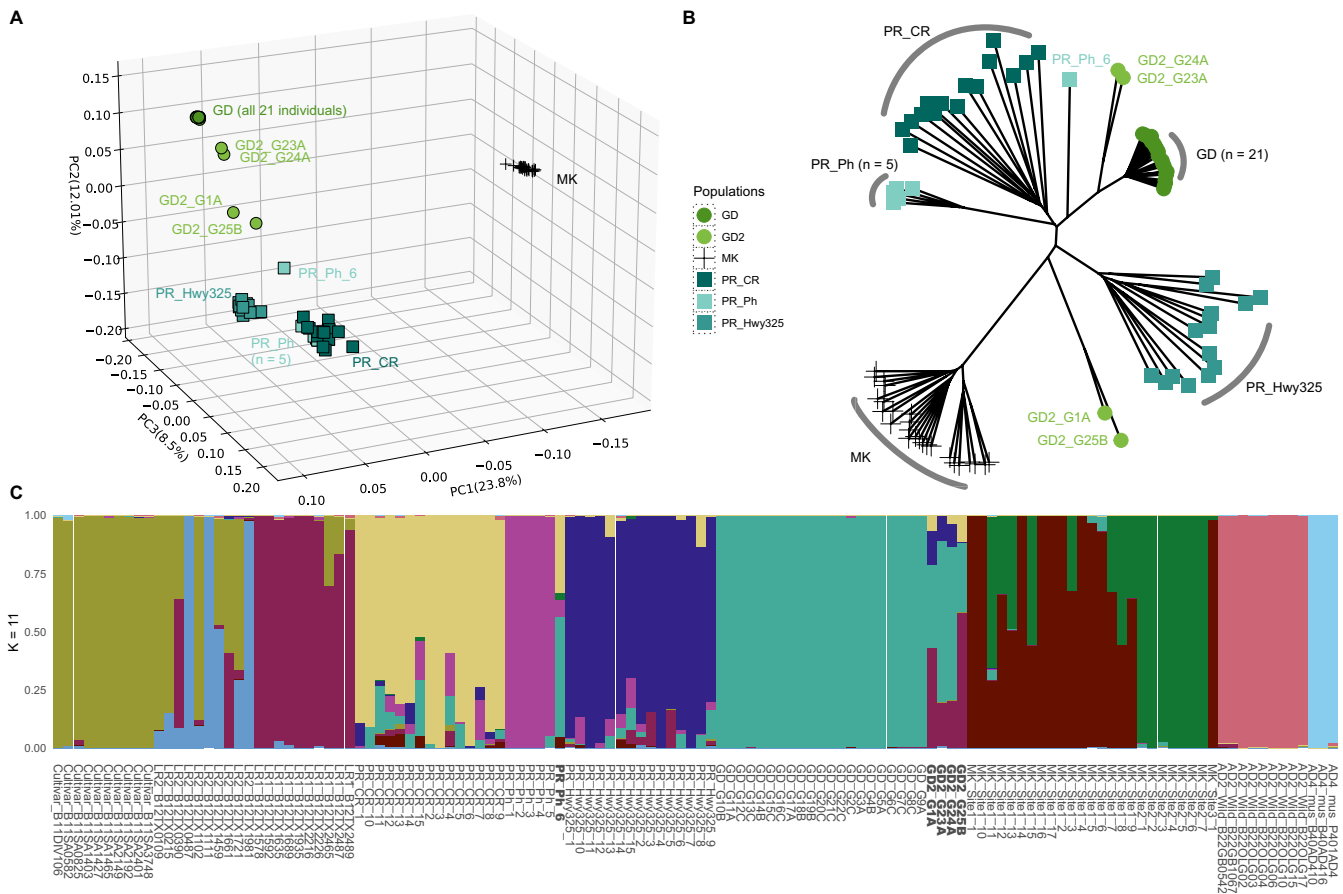


FIGURE 2 | Genetic relationships inferred from (A) PCA and (B) neighbour-joining tree of 61 Caribbean and 25 Mound Key *G. hirsutum* samples. Each accession/population is represented by a different symbol. (C) LEA genetic structure plot when the number of ancestral populations (K) = 11. The x-axis is labelled with individual code (Table S1) and the y-axis represents the proportion of estimated ancestral population and is filled by different colours.

3.4 | Frequent Gene Flow Among Caribbean Cottons

To assess the effects of potential inter- and intraspecific gene flow in shaping genetic relationships and diversity of the Caribbean cottons, we included the additional three domesticated *G. hirsutum* gene pools (Landrace1, Landrace2, and Cultivars) and the outgroups (*G. barbadense* and *G. mustelinum*) in a new analysis. In total, 3.3 million independent biallelic SNPs were extracted for a total of 128 samples.

Population structure analysis by LEA identified 11 as the optimal number of ancestral populations (K value) (Figure S3), with the results (Figure 2C) reiterating the inferences from the PCA and NJ phylogenetic analyses (Figure 2A,B) described above. Each outgroup species formed its own cluster, as expected, and the domesticated cottons were divided into three clusters, as previously noted (Yuan et al. 2021), with more mixed ancestral signals detected in Landrace2. Among wild cottons, most samples clustered based on their collection sites, except for PR_Ph_6 and four Guadeloupe outliers (GD2). Notably, PR_Ph_6 exhibited over 50% ancestral signals from the Guadeloupe main population (GD), and the four Guadeloupe outliers (GD2) had variable proportions of their ancestry shared with Landrace1, rising to over 50% in GD2_G25B and GD2_G1A. Unlike Puerto Rico Pitahaya

(*n* = 5) and the Guadeloupe (*n* = 21) main population (GD), which had mostly uniform ancestry, Mound Key cottons exhibited two distinct ancestral populations, as previously reported (Ning et al. 2024). In contrast to these observations of site-specificity, the Puerto Rican cottons from both CaboRojo (CR) and Highway325 (Hwy325) exhibited a combination of site-specific major ancestral genetic background with mixed signals from multiple other sources.

Using the same set of SNPs as above, we estimated genomic introgression among all 128 samples using TreeMix (Figure 3A). We first grouped the samples based on their genetic structure (Figure 2C), from which we separated the outliers from GD2 and PR_Ph_6 (described above) as two additional groups and merged the two Mound Key subpopulations, for a total of 12 populations/groups. The best TreeMix model estimated the number of migration events (*m*) = 7 (Figure S4). Compared to the topology of the maximum likelihood tree when *m* = 0 (the phylogenetic tree on the left side of the heatmap in Figure 3B), the topology of the TreeMix result, when *m* = 7, placed four Guadeloupe outliers (GD2) as sister to Landrace1 among all domesticated cotton groups (Figure 3A), as expected from the above results and indicating mixed ancestry. The best TreeMix model also supported three gene flow branches from Landrace1 into Puerto Rican populations Highway325 (Hwy325), Pitahaya (Ph), and CaboRojo (CR), which is also supported by the ancestry

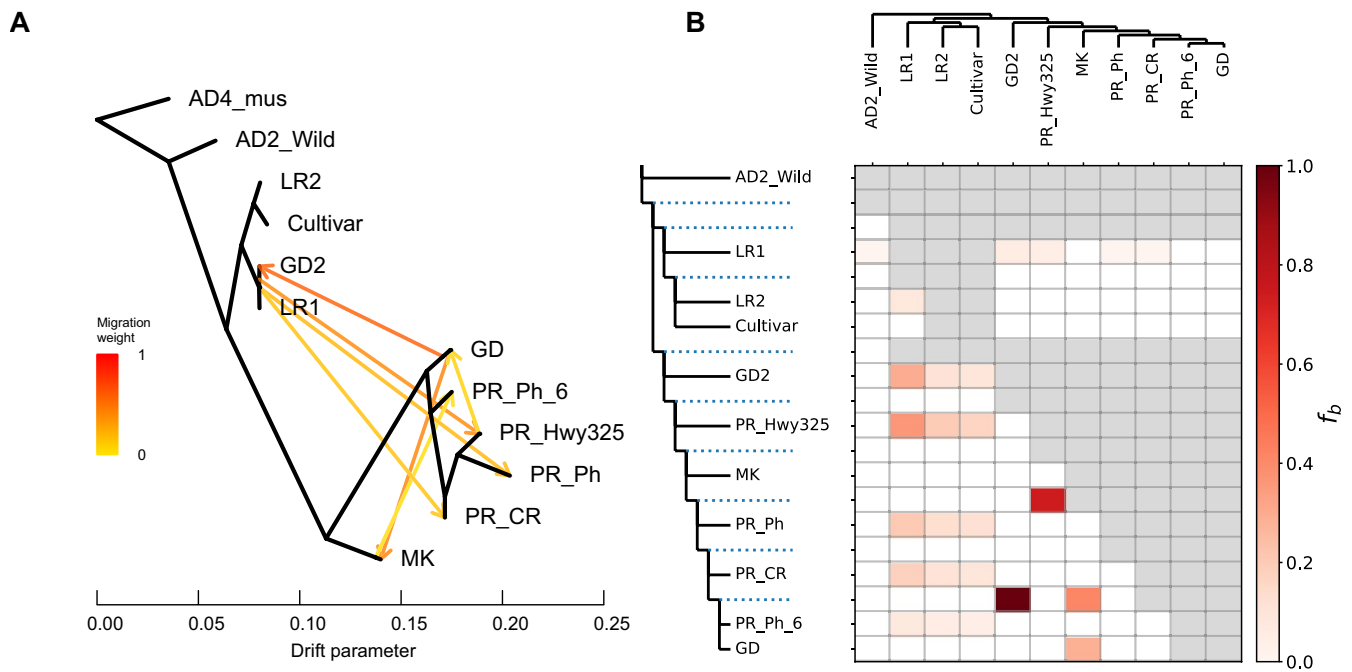


FIGURE 3 | Genomic introgression tests between *Gossypium mustelinum* (AD4_mus), *G. barbadense* (AD2_wild), three domesticated cottons from germplasm collections: Landrace1 (LR1), Landrace2 (LR2), and Cultivar, and wild Caribbean and Mound Key (MK) cottons. (A) TreeMix with seven migration edges shown by seven branches with arrows indicating gene flow direction, and the colour of the migration edges indicates migration weight. (B) Dsuite heatmap shows the level of introgression between the maximum likelihood tree nodes on the top and left. Darker colour indicates values of branch introgression (f_b).

structure for Highway325 and CaboRojo but is not apparent in Pitahaya (Figure 2C). Two additional gene flow branches with high migration weight were suggested from the Guadeloupe main population (GD) into both the GD2 and Mound Key cottons; the former of which is expected due to proximity, but the latter is neither expected from geographic proximity nor from likely ancestry. This unexpected inference of gene flow from Guadeloupe into Mound Key likely reflects other population evolutionary processes, such as shared recent ancestry, genetic drift in small localised populations, and regional dispersal dynamics associated with cyclonic storms. Finally, the model suggests both gene flow from the Mound Key cottons into the Puerto Rican Pitahaya site outlier (PR_Ph_6), and from Puerto Rican Highway325 population into the Guadeloupe main population. Considering their ancestry reconstructions, however, the TreeMix-suggested gene flow from Mound Key into PR_Ph_6 may derive instead from admixture with other Puerto Rican cottons (e.g., CaboRojo and Highway325) that exhibit some shared ancestry with unsampled, and perhaps extinct Mound Key-like populations, whereas introgression of Highway325 into GD is not supported by the structure analysis, perhaps indicating that it is an artefact of analysis in these small, isolated populations.

We further analysed the potential for genomic introgression using the ABBA-BABA test based Dsuite (Figure 3B). The results showed that all Puerto Rico samples and the four Guadeloupe outliers (GD2) experienced different levels of gene flow from domesticated cottons, especially from Landrace1 accessions. Similar to the neighbour-joining tree and LEA structure results (Figure 2B,C), when assuming no gene flow, the phylogeny from TreeMix grouped the outlier PR_Ph_6 with the Guadeloupe main population (GD). The strongest introgression

signal ($f_b = 1$) was detected between the group containing PR_Ph_6 and Guadeloupe main population (GD) and the four geographically adjacent Guadeloupe outliers (GD2). Moreover, the ancestral node of all Caribbean wild cottons (excluding GD2) to Puerto Rican cotton Highway325 also showed high introgression probabilities ($f_b = 0.7$). Similar to the TreeMix result, there was also a gene flow signal detected between Mound Key and the Guadeloupe main population (GD $n = 21$), but given the tiny size of these islands and the great distance between them, we interpret this observation to reflect a combination of shared ancestry and population structuring (see Discussion).

3.5 | Reticulation Revealed Through Plastome Phylogenomics

Chloroplast DNA is maternally inherited in *Gossypium* and has been shown to be useful in phylogenetics and analyses of introgression (Brubaker et al. 1993; Chen et al. 2017; Wendel and Albert 1992; Wu et al. 2018; Yan et al. 2024). Among the 158 samples analysed here, the average plastome length was 160,406 bp, similar to that previously reported (Lee et al. 2006). For phylogenetic analysis, we removed one of the large inverted repeats and all indels, resulting in a final aligned length of 134,121 bp. Maximum Parsimony (MP) (Figure 4A; Figure S5A) and maximum likelihood (ML) (Figure S6A,B) analysis of this alignment contained 287 parsimony informative sites, only 91 of which were within the 110 cpDNA coding loci. This extraordinarily low level of nucleotide diversity with the chloroplast genome is not unexpected given its well-known low rate of sequence evolution, but there are few comparable datasets for this level of intraspecific sampling.

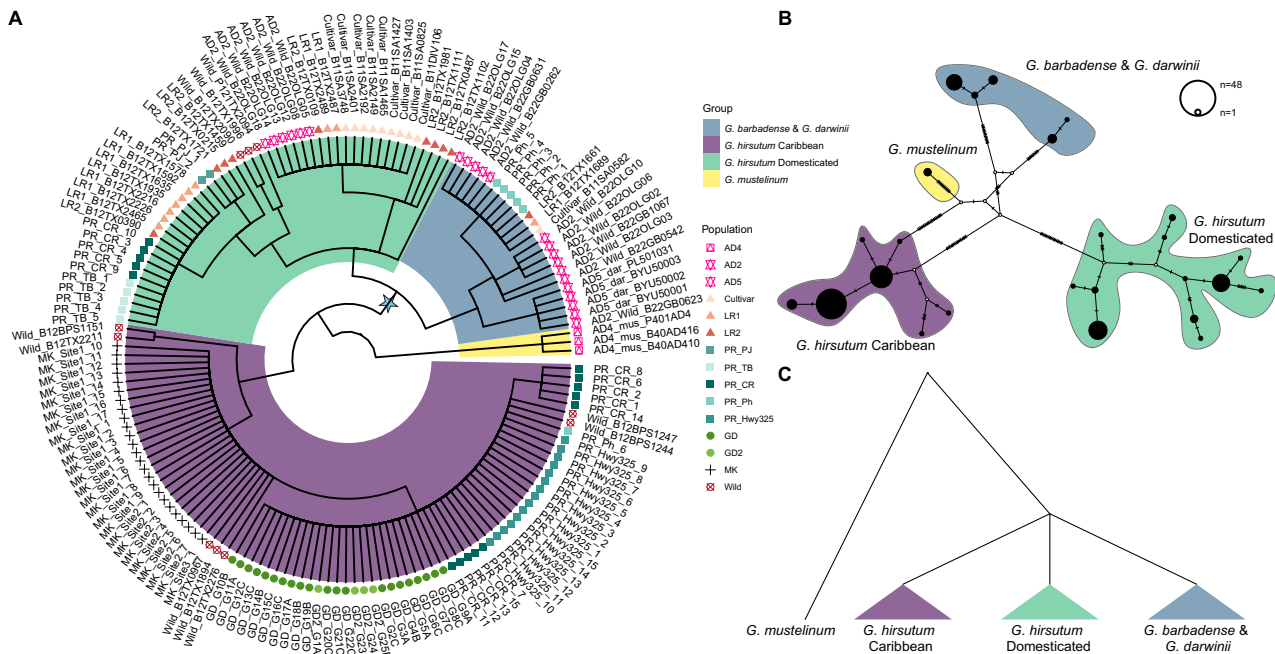


FIGURE 4 | (A) Maximum parsimony phylogenetic tree for 158 *Gossypium* samples using 110 extracted gene sequences. Each tree tip is labelled by a shape and an individual ID that represents their genetic group, which was assigned using genome-wide SNPs (see Figure 1A and Figure S1A). Clades are coloured based on assigned groups. (B) Haplotype network using a concatenation of sequences from the 110 cpDNA genes. The lines on the branch represent the nuclear changes between groups. Colours represent their cpDNA main types. (C) A cartoon tree to show genetic relationships between the three main plastome groups inferred.

Phylogenetic reconstruction using either the full genome or genes only, and MP or ML methods, both supported a similar topology and overall species relationship (Figure 4A; Figures S5A and S6A,B). Rooting with the outgroup *G. mustelinum* (AD4) (Wendel and Albert 1992; Hu et al. 2021), the 158 samples were resolved into three major clades (Figure 4A), namely: (1) a clade that mostly includes wild *G. barbadense* and *G. darwinii*; (2) a ‘Domesticated’ clade containing the majority of cultivated *G. hirsutum*; and (3) a ‘Caribbean’ clade, containing most of the putatively wild *G. hirsutum* collected from the Caribbean islands. Notably, however, both the ‘*G. barbadense* + *G. darwinii*’ and the ‘Domesticated *G. hirsutum*’ clades exhibited some unexpected membership, and the two *G. hirsutum* clades were not monophyletic.

We further explored relationships among these plastomes using haplotype network reconstruction of both the 110 cpDNA loci and the plastomes (Figure 4B; Figure S5B). In both cases these networks recapitulate the three primary groups found by MP phylogenetic analysis; however, the haplotype network failed to resolve relationships among these three main groups (Figure 4B). Importantly, although the phylogeny supported the ‘Domesticated *G. hirsutum*’ clade as sister to the two sister clades of ‘*G. barbadense* + *G. darwinii*’ clade rather than the ‘Caribbean’ clade, only two parsimony informative sites (including one homoplasious site) supported the node labelled with a star in Figure 4A, which separates the wild *G. hirsutum* samples from the ‘*G. barbadense* and *G. darwinii*’ and ‘Domesticated’ clades. Accordingly, for the purpose of this study, we treated those three groups as a polytomy (Figure 4C). This result serves as a striking reflection of the relative recency of species divergence and domestication when viewed on the scale of cpDNA sequence evolution.

Although the composition of each clade was largely consistent, notable exceptions were identified that may suggest historical introgression leading to chloroplast capture (Wendel and Albert 1992). While the ‘*G. barbadense* + *G. darwinii*’ clade contained all nine *G. barbadense* samples that we judged to be wild (see above) and all *G. darwinii* samples, it also contained *G. hirsutum* samples from the Puerto Rican population Pitahaya (Ph; $n=5$) and three domesticated *G. hirsutum* individuals previously sampled from the National Cotton germplasm collection (Yuan et al. 2021). Likewise, the ‘Domesticated’ group was largely composed of *G. hirsutum*, albeit with the unexpected inclusion of some putatively wild *G. hirsutum* and *G. barbadense* samples. As expected, the Cultivars, Landrace1, and Landrace2 *G. hirsutum* samples from Yuan et al. (2021) Puerto Rico feral cottons from the TamarindoBeach (TB) and PuntaJacinto (PJ) sites, were all included in the ‘Domesticated’ *G. hirsutum* clade. Interestingly, however, three Wild samples from Yucatán, Mexico sequenced by Yuan et al. (2021) and five individuals from the wild cotton CaboRojo site (CR; Puerto Rico) were also included within the ‘Domesticated’ clade, as were six wild *G. barbadense* samples with mixed ancestral signals from more than one population (Figure S1C), potentially indicating a history of hybridization or introgression. All remaining wild samples, including those from Mound Key and other Caribbean islands (including seven Wild samples from germplasm collections) grouped together to form the ‘Caribbean’ clade.

3.6 | High Genetic Diversity Preserved Among Caribbean Cottons

Genome-wide nucleotide diversity (π) was estimated for 121 individuals, including all four *G. hirsutum* groups from Yuan

et al. (2021), Mound Key cottons, and all Caribbean wild cottons (Figure 5A). The four outlier samples from Guadeloupe (GD2) and the single outlier sample from Puerto Rico Pitahaya site (PR_Ph_6) were excluded due to their mixed genomic signals and paraphyletic placement in the NJ tree (Figure 2B,C). Nucleotide diversity within populations (Figure 5A) for the four *G. hirsutum* groups from Yuan et al. (2021) was, on average, highest for the Wild group (2.57×10^{-3}), followed by similar levels in Landrace2 (1.91×10^{-3}) and Landrace1 (1.92×10^{-3}), with the lowest diversity being found in the Cultivars (9.84×10^{-4}), as expected given the strong genetic bottlenecks accompanying cotton domestication (Brubaker and Wendel 1994; Wendel et al. 1992). Notably, however, one of the chromosomes (D08) in the cultivars exhibited elevated diversity (3.25×10^{-3}), possibly resulting from the known structural variation in the Cultivars (e.g., Hu et al. 2025).

Among the additional wild cotton populations analysed here, two Puerto Rican populations Highway325 (1.98×10^{-3}) and CaboRojo (1.90×10^{-3}) had similar diversity as in the two landrace groups, whereas Mound Key cottons (1.04×10^{-3}) exhibited

nucleotide diversities in between those of the landraces and Cultivars. Moreover, these cottons also exhibited large variation between chromosomes, which reiterated the higher divergence observed between individuals within each population (Figure 2C). In contrast, the main Guadeloupe population (GD) and cottons from the Puerto Rico Pitahaya site (PR_Ph) showed similarly exceptionally low diversity and lower levels of variation between chromosomes, 5.13×10^{-4} and 7.11×10^{-4} , respectively, reflecting their highly inbred nature (Figure 2A; Figure S2).

Pairwise nucleotide differences (d_{xy}) were calculated between all pairs of populations (Figure 5B). Compared to all three domesticated cotton groups (Landrace1, Landrace2, and Cultivars), the d_{xy} (in dash boxes) in the Wild group ($> 3.71 \times 10^{-3}$), Caribbean wild cottons ($> 3.36 \times 10^{-3}$), and Mound Key cottons ($> 4.28 \times 10^{-3}$) all showed sequence divergence at the upper end of the range. These data mirror the PCA results (Figure 1B), in which the domesticated cottons and wild cottons exhibit greater sequence divergence. Among wild cotton populations, Puerto Rican site Highway325 cottons exhibited higher sequence divergence ($> 2.28 \times 10^{-3}$) from all other cotton

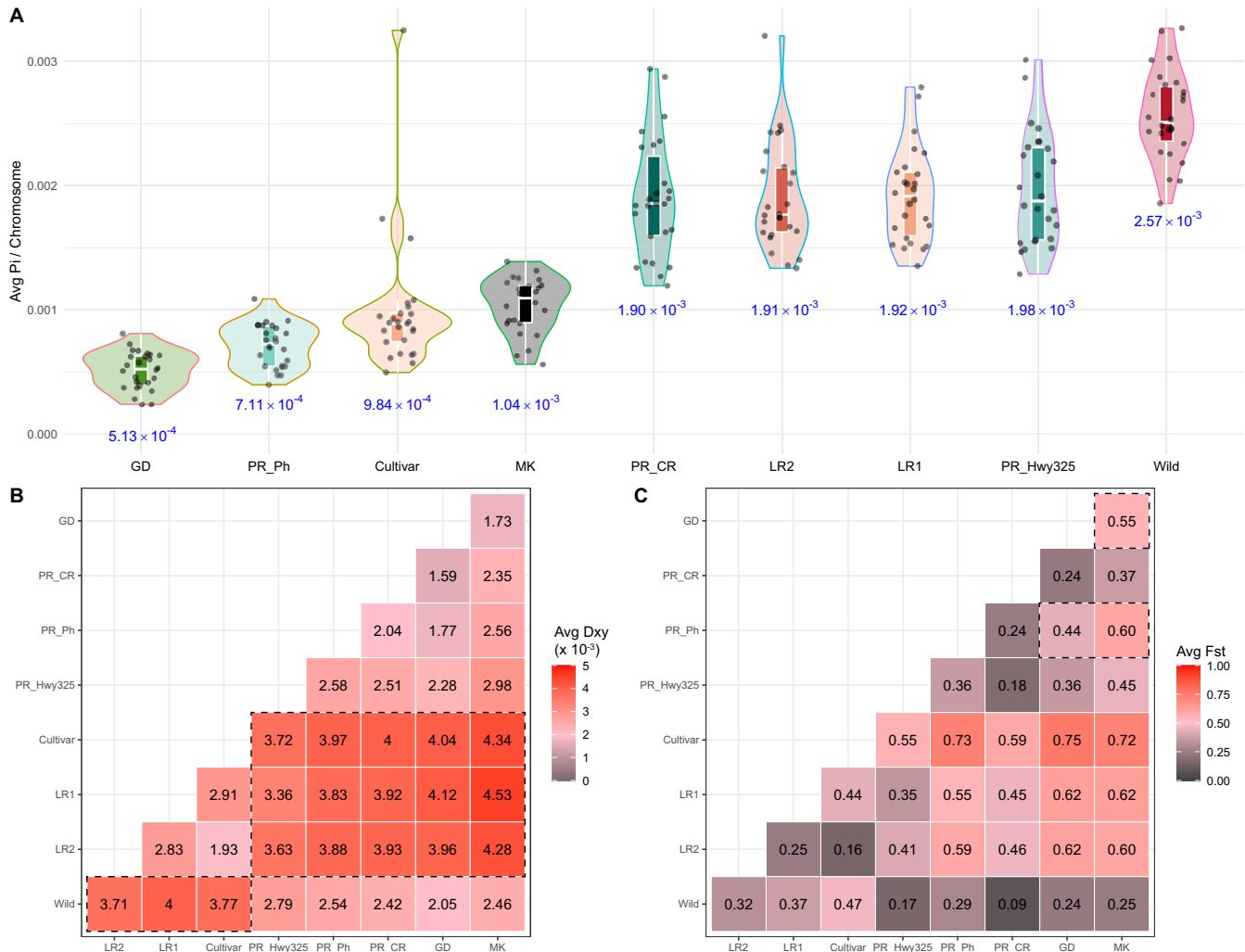


FIGURE 5 | Genetic diversity and divergence measurements of four *G. hirsutum* groups and wild cottons from populations in the Caribbean and Mound Key. (A) Average nucleotide diversity (π) per chromosome. Populations/groups are distinguished by shape and colour. Mean π value across all chromosomes for each group is labelled below. Heatmap plot of averaged pairwise comparisons of (B) d_{xy} and (C) F_{st} between nine populations/groups. Deeper colour represents higher differences between each pair, and actual values are labelled in each comparison.

populations, and especially when compared to Mound Key cottons ($d_{xy} = 2.98 \times 10^{-3}$). By contrast, relatively lower divergence ($d_{xy} < 1.73 \times 10^{-3}$) is found between the Guadeloupe (GD) and Mound Key populations, and for samples from the two Puerto Rican populations at CaboRojo and Pitahaya.

Population differentiation (F_{st}) between wild and domesticated cottons showed two distinct patterns (Figure 5C). When compared to the three domesticated cotton groups (Landrace1, Landrace2, and Cultivars), the wild cotton populations from Mound Key, Puerto Rico Pitahaya site and Guadeloupe exhibited higher F_{st} (0.55–0.75), whereas the Highway325 and CaboRojo site Puerto Rican cottons showed a lower F_{st} range (0.35–0.59). These results reiterate those from analyses of genetic structure (Figure 2C), introgression (Figure 3A,B), and plastome phylogenomics (Figure 4A), in which Highway325 and CaboRojo contained more mixed ancestral signals than other wild cottons. Among the putatively wild cotton populations from Mound Key, Puerto Rico, and Guadeloupe, all three pairs (in dash boxes) had F_{st} values between 0.44 and 0.60.

3.7 | Levels of Homozygosity and Heterozygosity

Long runs of homozygosity (ROH), contiguous homozygous regions inherited from both parents, are informative about

inbreeding and demographic history. Although ROH-based measures of inbreeding have been widely applied in humans and domesticated animals (reviewed in Shafer and Kardos 2025), their use in plants remains limited (Bemmels et al. 2025; e.g., Kumar et al. 2020). Using 26 million biallelic SNPs from 121 samples, we found the number of ROH per genome ranged from 76 to 978, with genome-wide ROH coverage varying from 1.5% to 15.3% (Figure 6A; Figure S7B). Most ROH were detected in short tracts below 1 Mb (Figure 6A), and, as might be expected, samples with higher numbers of ROH also had larger genome-wide ROH coverage (Figure S7B). As a metric of inbreeding coefficient based on ROH (F_{ROH}), the F_{ROH} values were the highest in the Cultivars (0.103), Landrace2 (0.092), and Guadeloupe main population (GD; 0.085), followed by the Wild group (0.055) and Landrace1 (0.051), with the populations from Puerto Rico exhibiting much lower homozygosity levels (0.039 Pitahaya ($n = 5$); 0.034 CaboRojo, and 0.029 Highway325). The lowest F_{ROH} (0.025) was found in the Mound Key population (Figure 6C; Figure S7A), which seems counterintuitive given its small population size and geographic isolation from other cottons.

Complementing the ROH analysis, we also estimated linkage disequilibrium (LD) to infer differences in historical demography that influence SNP co-occurrence (e.g., inbreeding; Figure 6B). Previous analysis of LD decay (Ning et al. 2024) revealed that LD remains the highest for Cultivars where pairwise SNP distances

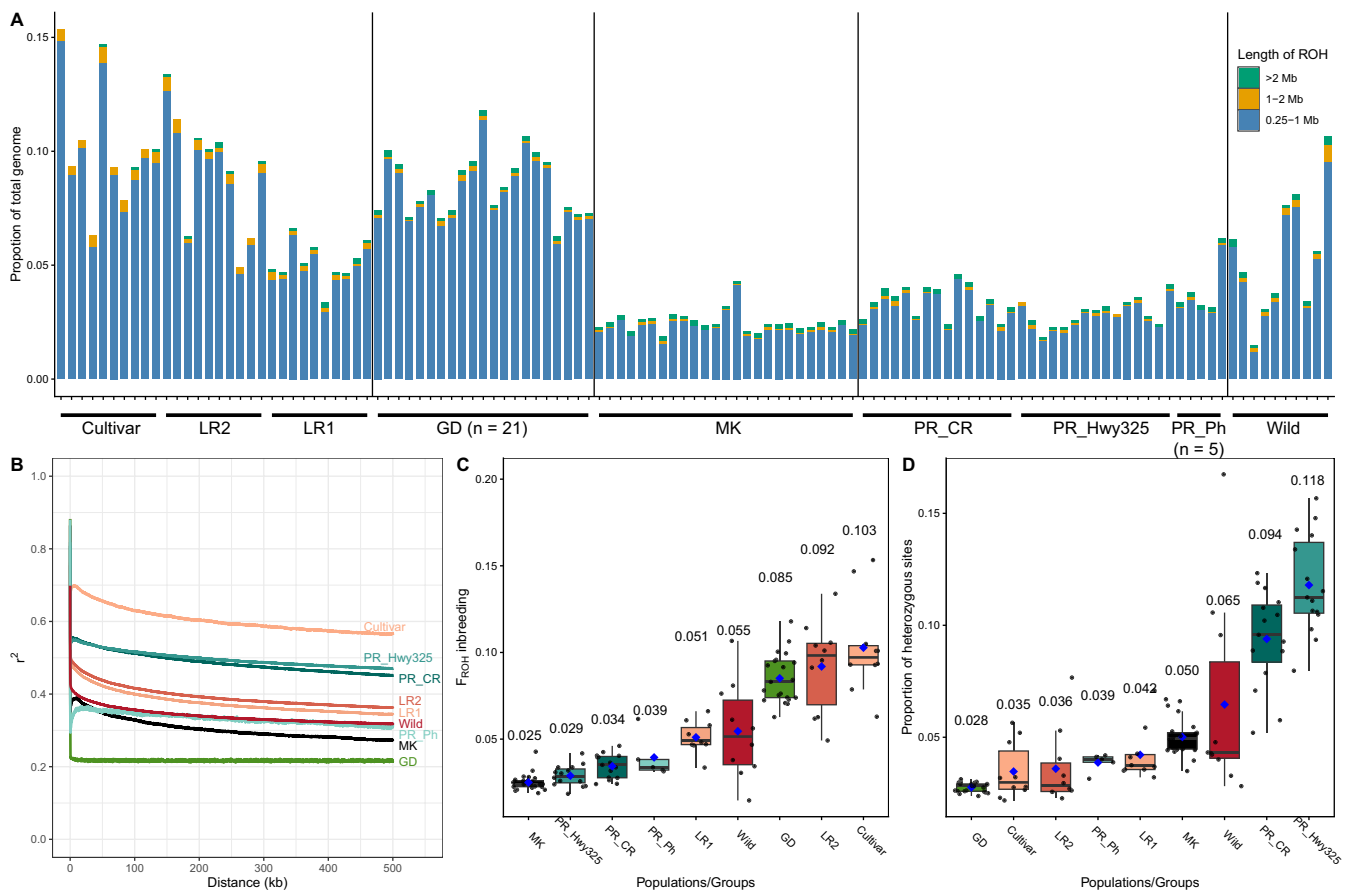


FIGURE 6 | (A) Total genome proportion of runs of homozygosity (ROH) in each individual. Bar colour shows the ROH in different length categories. (B) Linkage disequilibrium (LD) decay shown by the correlations of r^2 and the distance (kb) between SNP pairs. Each group/population is labelled by a different colour. (C) F_{ROH} inbreeding coefficient and (D) Proportions of heterozygosity sites for nine populations/groups. The average F_{ROH} is shown by a blue dot and labelled above each boxplot.

reach 500 kbp (i.e., the slowest decay), followed by faster and overlapping LD decay slopes in Landrace1 and Landrace2, with the Wild group showing the fastest decay. Among the newly sampled wild cotton populations, LD in Highway325 and CaboRojo decayed the slowest, falling to levels intermediate between the landraces and the Cultivars. Conversely, Pitahaya from Puerto Rico and the Wild group exhibited a similar decay pattern, both lower than that of the Mound Key cottons. Interestingly, the Guadeloupe samples showed a flat r^2 trend which had no correlation with physical distances of SNPs.

As a final measure of genome-wide diversity, we calculated the proportion of heterozygous sites in each genome. Among wild cotton populations, the highest value was observed for the two Puerto Rican populations with a history of admixture, that is, Highway325 (11.8%) and CaboRojo (9.4%), where heterozygosity was even higher than the Wild cotton group (6.5%) (Figure 6D). In contrast, the Pitahaya population from Puerto Rico (PR_Ph; 3.9%) showed less than half the genome-wide heterozygosity relative to Highway325 or CaboRojo. The Mound Key cottons (5.0%) exhibited larger proportions of heterozygous sites than the Pitahaya population. As expected, given the strong selection and inbreeding accompanying true-breeding line development in cotton cultivars, the domesticated cottons had the lowest

overall heterozygosity (Figure 6B,C). In particular, the Cultivar (3.5%) and Landrace2 (3.6%) groups had a lower number of heterozygous sites compared to Landrace1 (4.2%). Notably, the Guadeloupe main population (GD), despite being a wild cotton population, showed the lowest proportion of heterozygous sites (2.8%). This may reflect a small effective population size as well as a recent population bottleneck and a protracted history of localised inbreeding.

3.8 | Understanding Divergence of Wild Cottons Under Coalescent Model

To investigate the evolutionary history of wild *G. hirsutum* populations, we first used Tajima's D as a proxy to assess deviations of rare allele distributions from neutral allele frequency distributions (Nielsen 2001). Given the known sensitivity of this statistic to sample size (Subramanian 2016), the Puerto Rican Pitahaya cotton population ($n=5$) was removed from the analysis. The average Tajima's D (Figure 7A) values were negative for all four tested populations, whereas comparing to Puerto Rico CaboRojo (-0.369) and Highway325 (-0.153), cotton populations from Mound Key (-1.18) and Guadeloupe (-2.25) showed strong negative Tajima's D values, indicating large genome region with an

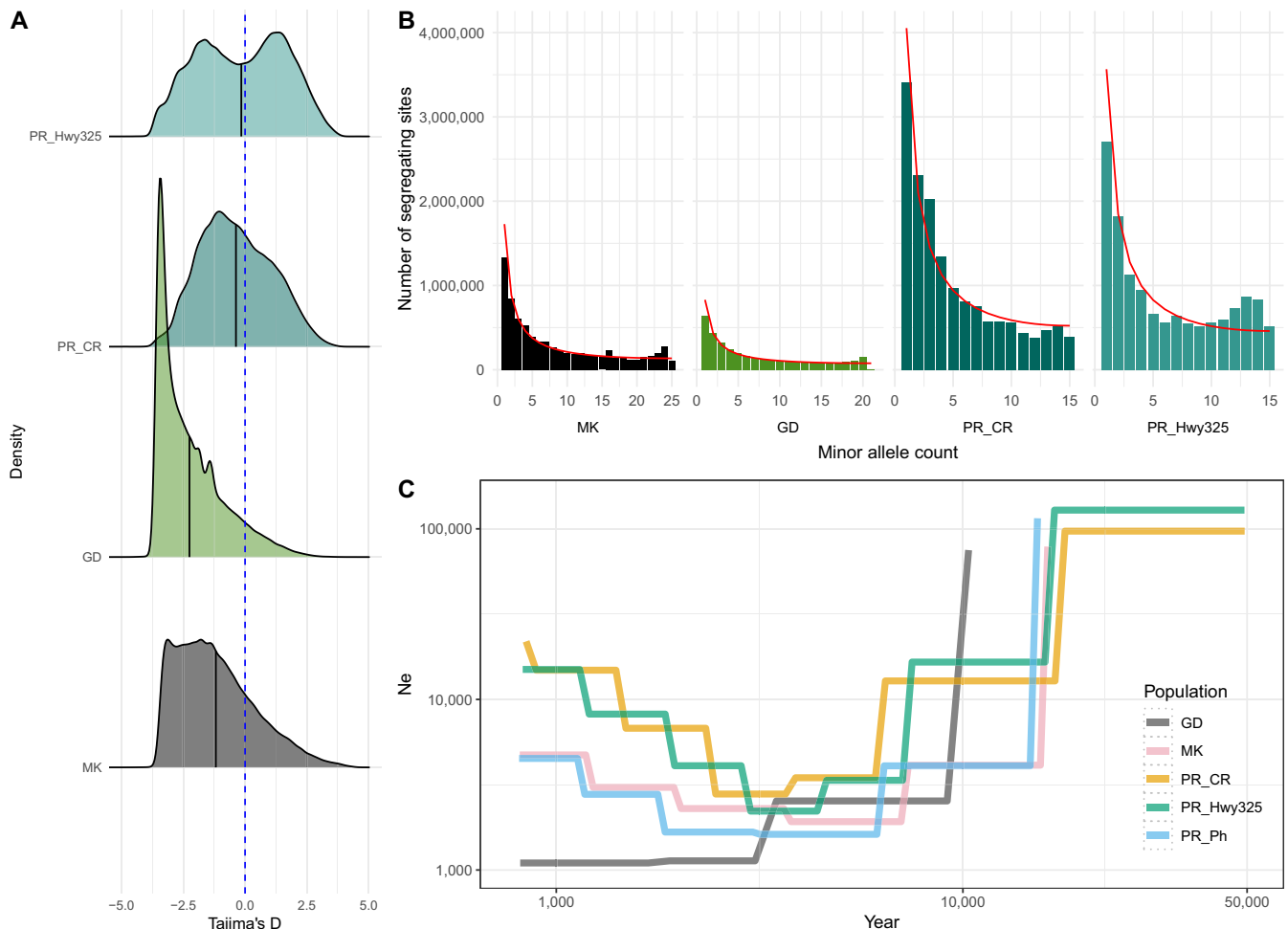


FIGURE 7 | (A) Density plot of Tajima's D . (B) The folded minor allele frequency plot against the proportions of the total variant sites for four populations, with the expected variant site distribution along different minor allele frequencies plotted as a red line. (C) SMC_{++} analysis of inferred effective population size change (y-axis) over evolutionary time (x-axis).

excess of rare alleles compared to the neutral model, usually reflecting recent population expansion or purifying selection.

As it is well understood that selection, admixture population history, subpopulation structure, and nonrandom sampling could affect Tajima's D estimation (Moeller et al. 2007; Nielsen 2001; Ramos-Onsins and Rozas 2002), we further explored the observed and expected site frequency spectrum (SFS) to reveal allele frequencies and distributions in four populations, where deviations from the expected allele distribution can also provide insight into selection or changes in population size (Beichman et al. 2018; Pavlidis and Alachiotis 2017). Since the ancestral allele is unknown, we calculated the observed folded SFS for each population based on the minor allele frequency (MAF). Notably, Guadeloupe cottons showed only around half to one fourth of variable sites (3.4M) when compared to Mound Key cottons (7.6M), Puerto Rican cottons Highway325 (13.7M) and CaboRojo (15.5M) (Figure 7B), further underscoring its general uniformity.

When compared to the expected SFS (calculated using Watterson's theta based neutral modelling (Hudson 2015)), all four wild cotton populations exhibited a deficiency of rare alleles (the first left peak of the SFS plots in Figure 7B), which combined with their negative Tajima's D values, may indicate a history of recent population expansion following a historical bottleneck. Notably, cotton populations from Mound Key and the two Puerto Rican sites Highway325 and CaboRojo all exhibited an excess of intermediate to high frequency alleles, suggesting recent admixture followed by inbreeding in these populations. Guadeloupe (GD) exhibited overall low genetic variation, a strong deficiency of high frequency alleles (last peak), and a strong negative Tajima's D value, perhaps suggesting recent population expansion from a severe bottleneck.

Finally, we also applied SMC_{++} to infer long-term effective population size (N_e) changes, acknowledging that the strong population structure from inbreeding, low genetic variation from recent strong bottlenecks, admixture between populations, and short divergence time may violate model assumptions and likely reduce our resolution or confound our results (Bansal and Nichols 2025; Wang et al. 2016). SMC_{++} analysis showed a strong declining trend for all five cotton populations starting around 20,000 years ago, reducing populations from an estimated 100,000 individuals to less than 10,000 in most cases. While the reduction in N_e observed in the two Puerto Rican cotton populations CaboRojo (CR) and Highway325 (Hwy325) appeared less severe and faster to recover, the history of admixture in these two cotton populations (Figure 6D; Figure 3A,B) can obscure coalescent events and thus inflate population size estimates (Figure 7C). By contrast, the Mound Key (MK), Puerto Rico Pitahaya (Ph), and Guadeloupe (GD) populations exhibited generally low N_e after the initial reduction, although MK and Ph may have somewhat recovered from the bottleneck starting around 2000 years ago. Population size estimates for the Guadeloupe population remained low, consistent with the small, relatively uniform population found on the geographically isolated Guadeloupe island, which has also experienced a human-induced population crash associated with habitat destruction and range restriction on this island. These estimated N_e changes, however, could also reflect subpopulation structure

between wild cottons, or the general violations of the neutral model (reviewed by Bansal and Nichols 2025) and should be interpreted with caution.

4 | Discussion

Plant domestication represents one of the most transformative shifts in human history, enabling the development of agriculture and the rise of complex societies (Meyer and Purugganan 2013). Through selective propagation, wild species were genetically altered to produce more uniform, desirable phenotypes that comprise the elite cultivars grown today. An unintended and unfortunate consequence of this selection process, however, has been the winnowing of genetic diversity, potentially limiting the ability of modern crops to adaptively respond to pathogens, pests, and environmental stressors. Accordingly, the genetic resources represented by wild crop relatives are a valuable resource for introducing traits and genetic variability into modern inbred crops (Brozynska et al. 2016; Hajjar and Hodgkin 2007; Warschefsky et al. 2014; Zhang et al. 2017). For many crop species, however, ancestral wild gene pools have been depleted by habitat destruction or other human-mediated impacts, or face extinction (Ford-Lloyd et al. 2011; Maxted et al. 2010).

Given the vital role that *G. hirsutum* plays in global commerce, it is remarkable that so little is understood about the species as a wild plant. Prior to human disturbance and domestication, in natural settings wild cotton occurred as a woody perennial shrub to small tree in the arid to semiarid scrub forests that characterize some of the leeward coastal habitats in a broad swath of the Caribbean, ranging from perhaps Trinidad and Tobago in the south to as far north as Tampa Bay, Florida, with the best developed population systems in the northern part of the Yucatán Peninsula (d'Eeckenbrugge and Lacape 2014; Fryxell 1979; Hu et al. 2021; Stephens and Phillips 1972; Wendel et al. 1992). Although this aggregate geographical range is extensive, actual populations typically are widely scattered and may be highly inbred. Complicating our understanding of the status and diversity of wild populations has been the effects of human history, which includes both widespread cultivation of cotton in agricultural settings and habitat destruction of the native vegetation in which wild cotton occurs, leading to further habitat fragmentation. Finally, cultivated cottons are interfertile with wild forms, thus leading to gene flow into natural settings and the establishment of feral derivatives in native vegetation. Due to these many complexities, understanding the domestication genetics and diversity of truly wild *G. hirsutum* has been challenging. A corollary is that there is little understanding of the actual proportion and source of wild diversity that has been captured or represented in modern elite gene pools.

Here we use a previously generated phylogenomic framework (Yuan et al. 2021) to evaluate the genome-wide diversity and population structure in newly collected wild and putatively wild populations of *G. hirsutum* from the Caribbean. Analysis of field-collected cotton from multiple sites in Puerto Rico and Guadeloupe reveals novel pockets of cotton genomic diversity not yet described, which expands our understanding of wild diversity in the species. Using population genomics, we quantify diversity, divergence, and possible gene flow dynamics

among different island populations, as well as introgression from a second domesticated polyploid cotton species found in the Caribbean, *G. barbadense*. Our results confirm that populations of Caribbean *G. hirsutum* consist of three broad groups: purely wild *G. hirsutum*, feral derivatives of previous stages in the domestication continuum, and plants derived from possible hybridization and introgression from *G. barbadense*. Below we discuss the origins and diversity of Caribbean wild cottons.

4.1 | Genomic Diversity Among Domesticated Cottons

A recent phylogenomic context for understanding *G. hirsutum* diversity was generated by Yuan et al. (2021) using mostly germplasm from the Germplasm Resources Information Network (GRIN; Byrne et al. 2018)). This analysis grouped accessions into four broadly encompassing genetic groups, one comprising obsolete as well as modern cultivars, and three others representing collections from the wild and indigenous (pre-Columbian) domesticated range. These three groups, each heterogeneous, were termed Landrace1, Landrace2, and Wild, with Landraces 1 and 2 also aligning loosely with geography, the former broadly encompassing the Caribbean and the latter consisting of early domesticated forms from Central America. The utility of this framework was recently demonstrated by using it to verify that cotton populations from Mound Key, Florida are truly wild as opposed to feral (Ning et al. 2024) and to demonstrate that this population represented a novel or previously unknown pocket of genetic diversity.

Here we extend this analysis to demonstrate the utility of this framework to distinguish other “wild-looking” feral derivatives from truly wild cottons. These feral plant populations are highly heterogeneous and are thought to have been derived from primitively domesticated forms that repeatedly escaped from human into natural settings over the past couple of millennia, as cotton cultivation spread from its ancestral home to encompass much of the drier American subtropics and tropics, up to and including more recent escapes and introgressants (from agricultural fields) in the post-colonial era from larger-scale cotton field agriculture. Using representatives from the framework provided by Yuan et al. (2021), we found that all samples from two sites in Puerto Rican (PuntaJacinto and TamarindoBeach) are genomically similar to the domesticated and landrace cottons, suggesting that they represent feral cottons or early-domesticated forms (Figure 1C) rather than truly wild relicts. This conclusion was further supported by plastome variation, which revealed that cottons from these two sites contain the same cpDNA type as the domesticated cottons (the ‘Domesticated’ group in Figure 4A), which was distinct from the wild cottons. Notably, the morphology of these plants was ambiguous, in that they form multi-branched shrubs, consistent with wild cotton plant architecture, but with seed trichomes (“fibres”) that indicate the influence of domesticated cotton, in that they were white and longer than in truly wild forms. This quasi-wild phenotype is, in fact, quite common, and it also has been noted that feral cotton may regain wild cotton architecture after escaping cultivation and becoming reestablished in natural or disturbed sites (Brubaker and Wendel 1994; d’Eeckenbrugge and Lacape 2014; Fryxell 1979; Hutchinson 1951).

Similar to other crops (reviewed in Zhang et al. 2017), domestication led to a series of genetic diversity reductions from wild cottons to early domesticated cottons (i.e., landraces), and eventually to the modern cultivars (Figure 5A). Despite this decline, domesticated cotton gene pools contain a substantial proportion of novel genetic variants that were absent in wild cottons (Ning et al. 2024). Principal component analysis and neighbour-joining trees support a close genetic relationship between the two groups, placing the domesticated cottons as sister to wild cottons (Figure 1B,C). These findings highlight the reciprocal introgression history between *G. hirsutum* and *G. barbadense*, with all domesticated *G. hirsutum* containing different levels of genomic introgression signals from *G. barbadense* (Yuan et al. 2021). Although nuclear gene flow was not directly detected in this study, plastome variation provided additional genetic introgression evidence (Figure 4A), in which the ‘Domesticated’ group contained six *G. barbadense* samples that had more mixing signals in nuclear genetic structure analysis (Figure 5C), and the ‘*G. barbadense* & *G. darwinii*’ group had three domesticated *G. hirsutum* individuals from Landraces and Cultivars. This reciprocal introgression likely contributed to increased genomic variation in cultivated *G. hirsutum* and resulted in higher divergence levels compared to the wild relatives.

4.2 | Origins of Semi-Wild Cottons

When the feral derivatives are excluded, all remaining wild-collected cotton samples included in this study grouped with the Wild germplasm group from Yuan et al. (2021) (Figure 1B,C), using both PCA and phylogenetics. This result notwithstanding, population structure analysis revealed that some of these have signals of mixed ancestry. That is, unlike the general uniformity exhibited by the Mound Key cottons and most of the cottons from Guadeloupe and Puerto Rico site Pitahaya (Ph), cotton collected from the Puerto Rican sites Highway325 (Hwy325) and CaboRojo (CR) exhibited patterns of genetic variation that implicate an origin or influence from more than one population, as did the small population from Guadeloupe (GD2). Specifically, these “semi-wild” cottons contained different levels of ancestral signals from a proximal wild cotton population along with mixed signals from various other sources, including the domesticated pool (Figure 2C). Genomic introgression tests highlighted Landrace1 (and related groups) as a possible source of introgression (Figure 3A,B), congruent with the localization of that group in the Caribbean (Yuan et al. 2021). Notably, the morphology of these semi-wild cottons was consistent with the wild cotton phenotype, for example, multi-branched architecture and seeds with short-brownish fibres, providing morphological evidence of historical genomic introgression.

Further analysis of these samples revealed varying degrees of “contamination” from domesticated sources. In Guadeloupe, for example, the presence of domesticated alleles among the four semi-wild samples (i.e., GD2) was so pervasive that TreeMix grouped GD2 with Landrace1, suggesting that the likely direction of introgression was from the wild (main) GD population into a Landrace1-like background (Figure 3A). Notably, when collecting these, Georges Ano noted (Wendel, pers. comm.) that these four cottons appeared as “hybride

Marie Galante—Yucatanense”, the former representing a semi-domesticated race of *G. hirsutum* known from (and named for) the nearby island of Marie-Galante (Ano et al. 1983) and the latter representing the truly wild form known as *G. hirsutum* race Yucatanense (Hutchinson 1951; Fryxell 1979; Wendel et al. 1992). The presence of both “primitive” (i.e., race Yucatanense) *G. hirsutum* and *G. hirsutum* race Marie-Galante on Guadeloupe (Ano et al. 1982), along with a history of cotton cultivation during the colonial period (Hoy 1962), offers an explanation for the introduction of domesticated alleles into a predominantly wild background. Given our observations and the history of both wild and early domesticated forms on Guadeloupe, it is reasonable to suggest that these semi-wild cottons originate from gene flow between the wild cottons that inhabit Guadeloupe and one (or more) feral forms. Moreover, the four semi-wild outliers in Guadeloupe (GD2) form two divergent lineages (GD2_G25B with GD2_G1A, and GD2_G24A with GD2_G23A) in the neighbour-joining tree (Figure 2B), perhaps reflecting historical differences in genomic recombination and segregation (Figure 2C). Perhaps more striking is the observation that these four samples (collected within meters of one another) were growing at a site that is closely adjacent to the Guadeloupe main population (GD), yet no signature of admixture with domesticated cottons was evident in the GD population.

A similar interpretation applies to the semi-wild cottons found on Puerto Rico, which also are inferred here as having mixed ancestry between natural, wild populations of *G. hirsutum* and feral remnants of previously cultivated material. Among the Puerto Rican sites, both CaboRojo and Highway325 were composed solely of cottons with varying proportions of mixed ancestry (Figure 2A,B). The Highway325 population exhibited strong introgression signals from Landrace1 (Figure 3B), which supports hybrid ancestry for the Highway325 cottons. By contrast, the CaboRojo population also exhibited some (lesser) evidence of introgression from Landrace1; however, it also exhibited evidence of plastome introgression that was absent in the Highway325 cottons. Specifically, three different plastome types were found among the CaboRojo cottons (Figure 4A), including five individuals with the ‘Domesticated’ type, and the remaining 10 individuals containing the ‘Caribbean’ type (Figure 4). Therefore, while the semi-wild cottons from the Puerto Rican sites CaboRojo and Highway325 both likely result from introgression with feral cottons, the timing and contemporary outcomes of this contact appear variable between sites.

4.3 | Genomic Diversity of Semi-Wild Cottons

Gene flow from the domesticated crops to their wild progenitors has been documented in other crop plants (reviewed in Gepts 2014), such as common bean (Papa and Gepts 2003), maize (Hufford et al. 2013), barley (Russell et al. 2011), rice (Chen et al. 2004; Jin et al. 2018), and wheat (Nyine et al. 2020). These introgression events alter the genomic landscape and genomic diversity of wild crop relatives by mixing previously separated gene pools. As an example of this, although the Puerto Rican CaboRojo and Highway325 cottons had elevated heterozygosity and nucleotide variations relative to other wild

cotton populations (consistent with gene flow; Figure 5A; Figure 6B), these two semi-wild cotton populations also exhibited similar long runs of homozygosity as inbred wild cotton populations and elevated LD, the latter exceeding that of both landraces (Figure 6D). These seemingly contradictory observations reveal another dimension to the genomic history of the Puerto Rican cottons, that is, the presence of feral *G. barbadense* on the island. Cotton cultivation in Puerto Rico predates Christopher Columbus; however, it was not until the late 18th century when large-scale production commenced (Rodríguez et al. 1956). By the early 20th century, *G. barbadense*, a South American native, was in broadscale cultivation while the relictual populations of wild *G. hirsutum* faced targeted eradication (Stephens 1976); however, introgression between the two species was also employed in developing new cotton varieties (Percy 2009; Stephens 1975; Viot and Wendel 2023). Previous studies have noted that the history of genomic introgression from *G. barbadense* into Landrace1 (Yuan et al. 2021) remains evident, while others note the propensity for hybrid breakdown in crosses between *G. hirsutum* and *G. barbadense* (Hu et al. 2019; S. Stephens 1949). With these considerations, it is tempting to suggest that the high nucleotide variation and slower LD decay observed in the cottons from CaboRojo and Highway325 reflects additional historical admixture events (e.g., Nyine et al. 2020); that is, these cottons may represent derivatives of relictual populations of wild *G. hirsutum* that encountered feral cottons which contained genomic segments of *G. barbadense*. This complex ancestry is grounded in the history of the island, and accounts for the observations of increased genetic diversity and long LD. The slower LD decay trend in these two populations may be attributed to reduced efficiency of recombination in regions of interspecific introgression; however, this speculation needs experimental testing.

4.4 | Divergence Among Wild Cottons

Understanding genetic diversity among crop wild relatives has been an important goal in agriculture (Bohra et al. 2022). However, only limited plant crop wild relatives have been studied for levels and patterns of genomic diversity using whole genome data across multiple wild populations or landraces (e.g., wheat [Cheng et al. 2024] and maize [Hufford et al. 2012]). Complementing previous efforts to understand wild cotton diversity using different tools and sampling regimes (d’Eeckenbrugge and Lacape 2014; Ning et al. 2024; Wendel et al. 1992; Yuan et al. 2021), the present study reveals hidden genetic diversity within the Caribbean islands Guadeloupe (GD, $n = 21$) and Puerto Rico (Pitahaya site). These populations, like the Mound Key cottons (Ning et al. 2024), are found exclusively in arid to seasonally arid coastal regions, where they form small, inbred populations spanning just a few hundred meters along the coastline. Although the N_e estimates in these small inbred populations may be influenced by artefacts arising from sub-population structure (reviewed by Bansal and Nichols 2025) or reference genome biases (Akopyan et al. 2025), all three truly wild cotton populations exhibited similar patterns of low and declining population sizes, the latter being concomitant with the sea-level rise in the Caribbean since the Glacial Maximum that flooded many existing habitats (Khan et al. 2017). As

expected by this reduction in N_e and the generally small population sizes, all three wild populations exhibited low genetic nucleotide diversity (Figure 5A), although each was characterised by its own unique demographic history.

Florida Mound Key cottons, which consist of two subpopulations (Figure 2C) as previously reported (Ning et al. 2024), showed high population divergence from both the Caribbean wild cottons (Pitahaya) and Guadeloupe (GD), with F_{st} 0.60 and 0.55 (Figure 5D), respectively. Mound Key cotton also showed a diverged plastome type sister to all Caribbean cottons (Figure 4A). In addition, Mound Key cottons had a negative Tajima's D relative to equilibrium expectation (Figure 7A), which may indicate the population expansion of Mound Key cotton from initial seed dispersal to the island and population establishment. Perhaps Mound Key cottons belong more generally to a regionally genetically more coherent set of wild cotton populations from southernmost Florida, with the Mound Key cottons retaining different portions of the broader ancestral polymorphism, and therefore appearing as separate populations.

Populations at the Puerto Rican Pitahaya site and in Guadeloupe (GD) both contained only one inbred population with lower F_{st} (0.44) when compared to Mound Key cotton (Figure 5D). These two populations showed similar levels of nucleotide diversity and heterozygosity (Figures 5A and 6D), with the Guadeloupe main population particularly characterised by its genome homogeneity. The lack of variation between individuals in Guadeloupe may also have led to insufficient genetic diversity to differentiate LD at short or long distances and, therefore, create a flat LD decay trend (Figure 6C). In addition, the lack of rare alleles in Guadeloupe also resulted in a strong negative Tajima's D , suggesting the presence of new mutations after a severe bottleneck, the latter supported by the relative uniformity in the Guadeloupe population. Overall, this population appears highly inbred, likely composed of closely related individuals that went through a severe recent bottleneck or represent a recent founding event. Although some genomic regions maintained pre-bottleneck dated polymorphisms from recent ancestry (e.g., Brandvain et al. 2014), a large proportion of genomic regions have markedly reduced genetic variation.

In contrast, although all five individuals of the Pitahaya population had plastome type that grouped with wild *G. barbadense* in the 'AD2_AD5' group (Figure 4A), their nuclear data showed no clear evidence of introgression history from *G. barbadense*. Given the uniform ancestry in the Pitahaya main population (Figure 2C), this might indicate ancient introgression from *G. barbadense* in the Pitahaya populations, or recent introgression from domesticated *G. hirsutum* containing a *G. barbadense* plastome combined with strong genetic drift (Bemmels et al. 2025). An additional Puerto Rican outlier (PR_Ph_6) similarly exhibited mixed ancestry; instead of grouping with all other five samples from the same Puerto Rico Pitahaya site, it showed over 50% mix of ancestral signals from both Guadeloupe main population (GD) and CaboRojo (Figure 2C). Different from all other Pitahaya individuals that showed a plastome type nested in the '*G. barbadense* + *G. darwinii*' group (Figure 4A), the outlier Pitahaya_6 plastome was identical to all Guadeloupe and

Highway325 types, indicating possible gene flow from different islands. Given the limited sample size, this may also reflect additional gene flow resulting from human activities or natural dispersal events.

4.5 | Conservation Insights

As a biodiversity hotspot (Maunder et al. 2008), the Caribbean region, with more than 700 islands, provides a vast yet highly fragmented habitat, fostering the dispersal and diversification of wild cottons. The diverged microhabitats in each island and large effects from drift as well as geographic isolation provide new opportunities for novel adaptive alleles to evolve in each small relictual population. As a result, many discrete populations with unique genetic content may remain unrecognised among the islands.

Echoing the previous ecological niche modelling results (d'Eeckenbrugge and Lacape 2014), the distribution of semi-wild or wild cottons in Puerto Rico and Guadeloupe inhabits a narrow range (Figure 1A) in drier coastal regions. Puerto Rican cottons were collected at multiple sites, and it appears that a few of these are arguably truly wild. In addition, the wild cottons from Guadeloupe described here were collected from one small region across perhaps several hundred meters, but these were almost identical siblings. These rare and naturally uncommon wild cottons indicate their vulnerability to climate change-induced perturbations from rising sea levels and storms, or habitat eradication from human activities. In addition, with documented precolonial human activities that started 6000–7000 years ago (Fitzpatrick and Keegan 2007), and the fact that cotton farming was at one time a dominant component of agriculture in both islands (Viot and Wendel 2023), wild cottons experienced and may continue to face genomic contamination from feral cotton derivatives, such as the semi-wild cottons mentioned in this study.

Although sampling within other Caribbean islands, such as Cuba and Jamaica, would provide greater insight into *G. hirsutum* diversity, we expect that the remaining populations on these islands would exhibit similar patterns of diversity reduction and introgression in regions of sympatry with domesticated cotton. One can easily envision that prior to human-mediated disturbance, wild cottons were once more broadly distributed in the drier coastal regions throughout the Caribbean basin, as shown by ecological niche modelling (d'Eeckenbrugge and Lacape 2014). Populations most likely became established by long-distance dispersal and were eliminated by seasonal storms and hurricanes, which together generated an aggregately large and shifting range for the species, but one in which populations were often highly isolated and small in size. One can also imagine that these pressures intensified as humans spread throughout these islands, transforming the landscape in the process. Thus, the present relative rarity further constrains natural patterns of diversity, with reproductive isolation between demes increasing generalised inbreeding and genetic depauperatization. This dynamic, quantified to a certain extent in the present study, serves to justify the need for greater collection, evaluation, and germplasm preservation of this valuable cultivated cotton species.

Author Contributions

J.F.W. conceived the project, secured the funding, and collected the samples. W.N. co-designed the experimental analysis, wrote the manuscript, and created all figures, with assistance from co-last authors C.E.G. and J.F.W. W.N., C.E.G., G.H., and D.Y. contributed to data analysis. M.A.A. and J.A.U. contributed to reference genome assembly and annotation. Y.D., C.H., Z.V.M., and O.P. extracted DNA and generated genomic libraries. C.-Y.H., M.A.A., and D.G.P. contributed to genome sequencing.

Acknowledgements

We thank our collaborator and long-time cotton aficionado Georges Ano for collecting the Guadeloupe cotton samples. We thank Iowa State University Research IT for technical support; Kathleen Foster-Wendel for assistance with field collections; and Matthew Hufford and Xinlin Yuan for discussion. We also thank Kenneth McCabe and Anna Tuchin for assistance with plant care. We especially thank Emma Miller for growing the Guadeloupe plants, an experience she assures us was as unpleasant as it was botanically productive. This study was supported by the National Science Foundation Plant Genome Program (141589) and by Cotton Incorporated (22-605) to J.F.W. and USDA ARS Non-Assistance Cooperative Agreements 58-6066-0-066 and 58-6066-0-064 to D.G.P. The research reported in this paper is partially supported by the HPC@ISU equipment at Iowa State University, some of which has been purchased through funding provided by NSF under MRI grants number 1726447 and MRI2018594. N/A

Funding

This work was supported by National Science Foundation Plant Genome Program, 141589. Cotton Incorporated, 22-605. USDA ARS Non-Assistance Cooperative Agreements, 58-6066-0-066, 58-6066-0-064.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Bioinformatic pipelines are uploaded to: <https://github.com/WendellLab/CaribbeanAD1>; All sequence data generated in this study are available: Guadeloupe cotton (PRJNA603025) and Puerto Rico cotton (PRJNA1226603).

References

- Akopyan, M., M. Genchev, E. E. Armstrong, and J. A. Mooney. 2025. "Reference Genome Choice Compromises Population Genetic Analyses." *Cell* 188: 6939.
- Alam, O., and M. D. Purugganan. 2024. "Domestication and the Evolution of Crops: Variable Syndromes, Complex Genetic Architectures, and Ecological Entanglements." *Plant Cell* 36, no. 5: 1227–1241.
- Alavez, V., Á. P. Cuervo-Robayo, E. Martínez-Meyer, and A. Wegier. 2021. "Eco-Geography of Feral Cotton: A Missing Piece in the Puzzle of Gene Flow Dynamics Among Members of *Gossypium hirsutum* Primary Gene Pool." *Frontiers in Ecology and Evolution* 9: 653271.
- Alseikh, S., F. Scossa, W. Wen, et al. 2021. "Domestication of Crop Metabolomes: Desired and Unintended Consequences." *Trends in Plant Science* 26, no. 6: 650–661.
- Andersson, L., and M. Purugganan. 2022. "Molecular Genetic Variation of Animals and Plants Under Domestication." *Proceedings of the National Academy of Sciences of the United States of America* 119, no. 30: e2122150119.

- Ano, G., J. Fersing, and J.-M. Lacape. 1983. "Les Cotonniers de l'Ile de Marie-Galante." *Coton et Fibres Tropicales* 38, no. 2: 201–208.
- Ano, G., J. Schwendiman, J. Fersing, and J. M. Lacape. 1982. "Les Cotonniers Primitifs *G. hirsutum* Race Yucataense de la Pointe des Châteaux en Guadeloupe et l'origine Possible des Cotonniers Tétraplôides du Nouveau Monde." *Coton et Fibres Tropicales* 37, no. 4: 327–332.
- Bailey, N., L. Stevison, and K. Samuk. 2025. "Correcting for Bias in Estimates of θ_w and Tajima's D From Missing Data in Next-Generation Sequencing." *Molecular Ecology Resources* 25, no. 6: e14104.
- Bansal, J. K., and R. A. Nichols. 2025. "Can Genomic Analysis Actually Estimate Past Population Size?" *Trends in Genetics: TIG* 41, no. 7: 559–567. <https://doi.org/10.1016/j.tig.2025.03.003>.
- Beichman, A. C., E. Huerta-Sanchez, and K. E. Lohmueller. 2018. "Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms." *Annual Review of Ecology, Evolution, and Systematics* 49, no. 1: 433–456.
- Bemmels, J. B., S. Starko, B. L. Weigel, et al. 2025. "Population Genomics Reveals Strong Impacts of Genetic Drift Without Purging and Guides Conservation of Bull and Giant Kelp." *Current Biology: CB* 35, no. 3: 688–698.
- Biscarini, F., P. Cozzi, G. Gaspa, and G. Marras. 2018. "detectRUNS: Detect Runs of Homozygosity and Runs of Heterozygosity in Diploid Genomes." <https://cran.r-project.org/web/packages/detectRUNS/index.html>.
- Bohra, A., B. Kilian, S. Sivasankar, et al. 2022. "Reap the Crop Wild Relatives for Breeding Future Crops." *Trends in Biotechnology* 40, no. 4: 412–431.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30, no. 15: 2114–2120.
- Brandvain, Y., A. M. Kenney, L. Flagel, G. Coop, and A. L. Sweigart. 2014. "Speciation and Introgression Between *Mimulus nasutus* and *Mimulus guttatus*." *PLoS Genetics* 10, no. 6: e1004410.
- Brozynska, M., A. Furtado, and R. J. Henry. 2016. "Genomics of Crop Wild Relatives: Expanding the Gene Pool for Crop Improvement." *Plant Biotechnology Journal* 14, no. 4: 1070–1085.
- Brubaker, C. L., F. M. Borland, and J. F. Wendel. 1999. "The Origin and Domestication of Cotton." In *Cotton: Origin, History, Technology and Production*, edited by C. W. Smith and J. T. Cothren, 3–31. John Wiley & Sons.
- Brubaker, C. L., J. A. Koontz, and J. F. Wendel. 1993. "Bidirectional Cytoplasmic and Nuclear Introgression in the New World Cottons *Gossypium barbadense* and *G. hirsutum* (Malvaceae)." *American Journal of Botany* 80, no. 10: 1203–1208.
- Brubaker, C. L., and J. F. Wendel. 1994. "Reevaluating the Origin of Domesticated Cotton (*Gossypium hirsutum*; Malvaceae) Using Nuclear Restriction Fragment Length Polymorphisms (RFLPs)." *American Journal of Botany* 81, no. 10: 1309–1326.
- Byrne, P. F., G. M. Volk, C. Gardner, M. A. Gore, P. W. Simon, and S. Smith. 2018. "Sustaining the Future of Plant Breeding: The Critical Role of the USDA-ARS National Plant Germplasm System." *Crop Science* 58, no. 2: 451–468.
- Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25, no. 15: 1972–1973.
- Ceballos, F. C., P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson. 2018. "Runs of Homozygosity: Windows Into Population History and Trait Architecture." *Nature Reviews. Genetics* 19, no. 4: 220–234.
- Chen, L. J., D. S. Lee, Z. P. Song, H. S. Suh, and B.-R. Lu. 2004. "Gene Flow From Cultivated Rice (*Oryza sativa*) to Its Weedy and Wild Relatives." *Annals of Botany* 93, no. 1: 67–73.

- Chen, Z., C. E. Grover, P. Li, et al. 2017. "Molecular Evolution of the Plastid Genome During Diversification of the Cotton Genus." *Molecular Phylogenetics and Evolution* 112: 268–276.
- Cheng, S., C. Feng, L. U. Wingen, et al. 2024. "Harnessing Landrace Diversity Empowers Wheat Breeding." *Nature* 632, no. 8026: 823–831.
- d'Eeckenbrugge, G. C., and J.-M. Lacape. 2014. "Distribution and Differentiation of Wild, Feral, and Cultivated Populations of Perennial Upland Cotton (*Gossypium hirsutum* L.) in Mesoamerica and the Caribbean." *PLoS One* 9, no. 9: e107458.
- Fitak, R. R. 2021. "OptM: Estimating the Optimal Number of Migration Edges on Population Trees Using Treemix." *Biology Methods & Protocols* 6, no. 1: bpab017.
- Fitzpatrick, S. M., and W. F. Keegan. 2007. "Human Impacts and Adaptations in the Caribbean Islands: An Historical Ecology Approach." *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 98, no. 1: 29–45.
- Ford-Lloyd, B. V., M. Schmidt, S. J. Armstrong, et al. 2011. "Crop Wild Relatives—Undervalued, Underutilized and Under Threat?" *Bioscience* 61, no. 7: 559–565.
- Frichot, E., and O. François. 2015. "LEA: An R Package for Landscape and Ecological Association Studies." *Methods in Ecology and Evolution/British Ecological Society* 6, no. 8: 925.
- Fryxell, P. A. 1979. *The Natural History of the Cotton Tribe (Malvaceae, Tribe Gossypieae)*. TAMU Press.
- Gepts, P. 2014. "The Contribution of Genetic and Genomic Approaches to Plant Domestication Studies." *Current Opinion in Plant Biology* 18: 51–59.
- Grover, C. E., M. A. Arick, J. L. Conover, A. Thrash, G. Hu, and W. S. Sanders. 2017. "Comparative Genomics of an Unusual Biogeographic Disjunction in the Cotton Tribe (Gossypieae) Yields Insights Into Genome Downsizing." *Genome Biology and Evolution* 9, no. 12: 3328–3344.
- Grover, C. E., J. P. Gallagher, J. J. Jareczek, et al. 2015. "Re-Evaluating the Phylogeny of Allopolyploid *Gossypium* L." *Molecular Phylogenetics and Evolution* 92: 45–52.
- Hajjar, R., and T. Hodgkin. 2007. "The Use of Wild Relatives in Crop Improvement: A Survey of Developments Over the Last 20 Years." *Euphytica/Netherlands Journal of Plant Breeding* 156, no. 1–2: 1–13.
- Harlan, J. R. 1971. "Agricultural Origins: Centers and Noncenters: Agriculture May Originate in Discrete Centers or Evolve Over Vast Areas Without Definable Centers." *Science* 174: 468–474.
- Hoy, D. R. 1962. "Changing Agricultural Land Use on Guadeloupe, French West Indies." *Annals of the Association of American Geographers* 52, no. 4: 441–454.
- Hu, G., C. E. Grover, J. Jareczek, et al. 2021. "Evolution and Diversity of the Cotton Genome." In *Cotton Precision Breeding*, edited by M.-U. Rahman, Y. Zafar, and T. Zhang, 25–78. Springer International Publishing.
- Hu, G., Z. Wang, Z. Tian, et al. 2025. "A Telomere-To-Telomere Genome Assembly of Cotton Provides Insights Into Centromere Evolution and Short-Season Adaptation." *Nature Genetics* 57, no. 4: 1031–1043.
- Hu, Y., J. Chen, L. Fang, et al. 2019. "*Gossypium barbadense* and *Gossypium hirsutum* Genomes Provide Insights Into the Origin and Evolution of Allotetraploid Cotton." *Nature Genetics* 51, no. 4: 739–748.
- Hudson, R. R. 2015. "A New Proof of the Expected Frequency Spectrum Under the Standard Neutral Model." *PLoS One* 10, no. 7: e0118087.
- Hufford, M. B., P. Lubinsky, T. Pyhäjärvi, M. T. Devengenzo, N. C. Ellstrand, and J. Ross-Ibarra. 2013. "The Genomic Signature of Crop-Wild Introgression in Maize." *PLoS Genetics* 9, no. 5: e1003477.
- Hufford, M. B., X. Xu, J. van Heerwaarden, et al. 2012. "Comparative Population Genomics of Maize Domestication and Improvement." *Nature Genetics* 44, no. 7: 808–811.
- Hutchinson, J. B. 1951. "Intra-Specific Differentiation in *Gossypium hirsutum*." *Heredity* 5, no. 2: 161–193.
- Jin, J.-J., W.-B. Yu, J.-B. Yang, et al. 2020. "GetOrganelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly of Organelle Genomes." *Genome Biology* 21, no. 1: 241.
- Jin, X., Y. Chen, P. Liu, et al. 2018. "Introgression From Cultivated Rice Alters Genetic Structures of Wild Relative Populations: Implications for in Situ Conservation." *AoB Plants* 10, no. 1: plx055.
- Kahle, D., and H. Wickham. 2013. "Ggmap: Spatial Visualization With ggplot2." *R Journal* 5, no. 1: 144.
- Katche, E., D. Quezada-Martinez, E. I. Katche, P. Vasquez-Teuber, and A. S. Mason. 2019. "Interspecific Hybridization for *Brassica* Crop Improvement." *Crop Breeding, Genetics and Genomics* 1, no. 1. https://cbgg.hapres.com/htmls/CBGG_1055_Detail.html.
- Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30, no. 4: 772–780.
- Kendig, K. I., S. Baheti, M. A. Bockol, et al. 2019. "Sentieon DNaseq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy." *Frontiers in Genetics* 10: 736.
- Khan, N. S., E. Ashe, B. P. Horton, A. Dutton, R. E. Kopp, and G. Brocard. 2017. "Drivers of Holocene Sea-Level Change in the Caribbean." *Quaternary Science Reviews* 155: 13–36.
- Kumar, S., C. H. Deng, M. Hunt, et al. 2020. "Homozygosity Mapping Reveals Population History and Trait Architecture in Self-Incompatible Pear (*Pyrus* spp.)." *Frontiers in Plant Science* 11: 590846.
- Lee, S.-B., C. Kaitanis, R. K. Jansen, et al. 2006. "The Complete Chloroplast Genome Sequence of *Gossypium hirsutum*: Organization and Phylogenetic Relationships to Other Angiosperms." *BMC Genomics* 7: 61.
- Leigh, J. W., and D. Bryant. 2015. "Popart: Full-Feature Software for Haplotype Network Construction." *Methods in Ecology and Evolution* 6, no. 9: 1110–1116.
- Li, H. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation From Sequencing Data." *Bioinformatics* 27, no. 21: 2987–2993.
- Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment With Burrows–Wheeler Transform." *Bioinformatics* 25, no. 14: 1754–1760.
- Malinsky, M., M. Matschiner, and H. Svardal. 2021. "Dsuite—Fast D-Statistics and Related Admixture Evidence From VCF Files." *Molecular Ecology Resources* 21, no. 2: 584–595.
- Marchi, N., F. Schlichta, and L. Excoffier. 2021. "Demographic Inference." *Current Biology: CB* 31, no. 6: R276–R279.
- Mather, N., S. M. Traves, and S. Y. W. Ho. 2020. "A Practical Introduction to Sequentially Markovian Coalescent Methods for Estimating Demographic History From Genomic Data." *Ecology and Evolution* 10, no. 1: 579–589.
- Maunder, M., A. Leiva, E. Santiago-Valentín, et al. 2008. "Plant Conservation in the Caribbean Island Biodiversity Hotspot." *Botanical Review; Interpreting Botanical Progress* 74, no. 1: 197–207.
- Maxted, N., S. Kell, Á. Toledo, et al. 2010. "A Global Approach to Crop Wild Relative Conservation: Securing the Gene Pool for Food and Agriculture." *Kew Bulletin* 65, no. 4: 561–576.
- Meyer, R. S., and M. D. Purugganan. 2013. "Evolution of Crop Species: Genetics of Domestication and Diversification." *Nature Reviews. Genetics* 14, no. 12: 840–852.
- Meyermans, R., W. Gorssen, N. Buys, and S. Janssens. 2020. "How to Study Runs of Homozygosity Using PLINK? A Guide for Analyzing

- Medium Density SNP Data in Livestock and Pet Species.” *BMC Genomics* 21, no. 1: 94.
- Minh, B. Q., J. Trifinopoulos, D. Schrempf, H. A. Schmidt, and R. Lanfear. 2019. “IQ-TREE Version 2.0: Tutorials and Manual Phylogenomic Software by Maximum Likelihood”.
- Moeller, D. A., M. I. Tenaillon, and P. Tiffin. 2007. “Population Structure and Its Effects on Patterns of Nucleotide Polymorphism in *Teosinte* (*Zea mays ssp. parviglumis*).” *Genetics* 176, no. 3: 1799–1809.
- Nielsen, R. 2001. “Statistical Tests of Selective Neutrality in the Age of Genomics.” *Heredity* 86, no. Pt 6: 641–647.
- Ning, W., K. M. Rogers, C.-Y. Hsu, et al. 2024. “Origin and Diversity of the Wild Cottons (*Gossypium hirsutum*) of Mound Key.” *Florida. Scientific Reports* 14, no. 1: 1–12.
- Nyine, M., E. Adhikari, M. Clinesmith, K. W. Jordan, A. K. Fritz, and E. Akhunov. 2020. “Genomic Patterns of Introgression in Interspecific Populations Created by Crossing Wheat With Its Wild Relative.” *G3 (Bethesda, Md.)* 10, no. 10: 3651–3661.
- Olsen, K. M., and J. F. Wendel. 2013. “Crop Plants as Models for Understanding Plant Adaptation and Diversification.” *Frontiers in Plant Science* 4: 290.
- Papa, R., and P. Gepts. 2003. “Asymmetry of Gene Flow and Differential Geographical Structure of Molecular Diversity in Wild and Domesticated Common Bean (*Phaseolus vulgaris* L.) From Mesoamerica.” *Theoretical and Applied Genetics* 106, no. 2: 239–250.
- Paradis, E., J. Claude, and K. Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” *Bioinformatics* 20, no. 2: 289–290.
- Pavlidis, P., and N. Alachiotis. 2017. “A Survey of Methods and Tools to Detect Recent and Strong Positive Selection.” *Journal of Biological Research (Thessalonike, Greece)* 24, no. 1: 7.
- Percy, R. G. 2009. “The Worldwide Gene Pool of *Gossypium barbadense* L. and Its Improvement.” In *Genetics and Genomics of Cotton*, 53–68. Springer US.
- Percy, R. G., and J. F. Wendel. 1990. “Allozyme Evidence for the Origin and Diversification of *Gossypium barbadense* L.” *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 79, no. 4: 529–542.
- Pickrell, J. K., and J. K. Pritchard. 2012. “Inference of Population Splits and Mixtures From Genome-Wide Allele Frequency Data.” *PLoS Genetics* 8, no. 11: e1002967.
- Purcell, S., B. Neale, K. Todd-Brown, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81, no. 3: 559–575.
- Purugganan, M. D., and D. Q. Fuller. 2009. “The Nature of Selection During Plant Domestication.” *Nature* 457, no. 7231: 843–848.
- Qu, X.-J., M. J. Moore, D.-Z. Li, and T.-S. Yi. 2019. “PGA: A Software Package for Rapid, Accurate, and Flexible Batch Annotation of Plastomes.” *Plant Methods* 15, no. 1: 50.
- Ramos-Onsins, S. E., and J. Rozas. 2002. “Statistical Properties of New Neutrality Tests Against Population Growth.” *Molecular Biology and Evolution* 19, no. 12: 2092–2100.
- Rodríguez, J. P., R. G. Caro, and R. V. Romero. 1956. “Cotton Variety Trials in Puerto Rico.” *Journal of Agriculture of the University of Puerto Rico* 40: 49.
- Russell, J., I. K. Dawson, A. J. Flavell, et al. 2011. “Analysis of > 1000 Single Nucleotide Polymorphisms in Geographically Matched Samples of Landrace and Wild Barley Indicates Secondary Contact and Chromosome-Level Differences in Diversity Around Domestication Genes.” *New Phytologist* 191, no. 2: 564–578.
- Shafer, A. B. A., and M. Kardos. 2025. “Runs of Homozygosity and Inferences in Wild Populations.” *Molecular Ecology* 34, no. 3: e17641.
- Shen, W., S. Le, Y. Li, and F. Hu. 2016. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” *PLoS One* 11, no. 10: e0163962.
- Splitstoser, J. C., T. D. Dillehay, J. Wouters, and A. Claro. 2016. “Early Pre-Hispanic Use of Indigo Blue in Peru.” *Science Advances* 2, no. 9: e1501623.
- Stephens, S. 1949. “The Cytogenetics of Speciation in *Gossypium*. I. Selective Elimination of the Donor Parent Genotype in Interspecific Backcrosses.” *Genetics* 34, no. 5: 627.
- Stephens, S. G. 1975. “A Reexamination of the Cotton Remains From Huaca Prieta, North Coastal Peru.” *American Antiquity* 40, no. 4: 406–419.
- Stephens, S. G. 1976. “The Origin of Sea Island Cotton.” *Agricultural History* 50, no. 2: 391–399.
- Stephens, S. G., and L. L. Phillips. 1972. “The History and Geographical Distribution of a Polymorphic System in New World Cottons.” *Biotropica* 4, no. 2: 49.
- Subramanian, S. 2016. “The Effects of Sample Size on Population Genomic Analyses—Implications for the Tests of Neutrality.” *BMC Genomics* 17, no. 1: 123.
- Swofford, D. L. 1998. “Phylogenetic Analysis Using Parsimony.” and relevely.net. http://andrelevely.net/bioinfo/paup/PAUP_Cmd_ref_v2.pdf.
- Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123, no. 3: 585–595.
- Templeton, A. R., K. A. Crandall, and C. F. Sing. 1992. “A Cladistic Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction Endonuclease Mapping and DNA Sequence Data. III. Cladogram Estimation.” *Genetics* 132, no. 2: 619–633.
- Terhorst, J., J. A. Kamm, and Y. S. Song. 2017. “Robust and Scalable Inference of Population History From Hundreds of Unphased Whole Genomes.” *Nature Genetics* 49, no. 2: 303–309.
- Tillich, M., P. Lehwark, T. Pellizzer, et al. 2017. “GeSeq - Versatile and Accurate Annotation of Organelle Genomes.” *Nucleic Acids Research* 45, no. W1: W6–W11.
- Viot, C. R., and J. F. Wendel. 2023. “Evolution of the Cotton Genus, *Gossypium*, and Its Domestication in the Americas.” *Critical Reviews in Plant Sciences* 42, no. 1: 1–33.
- Wang, J., E. Santiago, and A. Caballero. 2016. “Prediction and Estimation of Effective Population Size.” *Heredity* 117: 193–206.
- Wang, M., L. Tu, D. Yuan, et al. 2019. “Reference Genome Sequences of Two Cultivated Allotetraploid Cottons, *Gossypium hirsutum* and *Gossypium barbadense*.” *Nature Genetics* 51, no. 2: 224–229.
- Warschefsky, E., R. V. Penmetza, D. R. Cook, and E. J. B. von Wettberg. 2014. “Back to the Wilds: Tapping Evolutionary Adaptations for Resilient Crops Through Systematic Hybridization With Crop Wild Relatives.” *American Journal of Botany* 101, no. 10: 1791–1800.
- Wendel, J., and R. Percy. 1990. “Allozyme Diversity and Introgression in the Galapagos Islands. Endemic *Gossypium darwinii* and Its Relationship to Continental *G. barbadense*.” *Biochemical Systematics and Ecology* 18: 517–528.
- Wendel, J. F., and V. A. Albert. 1992. “Phylogenetics of the Cotton Genus (*Gossypium*): Character-State Weighted Parsimony Analysis of Chloroplast-DNA Restriction Site Data and Its Systematic and Biogeographic Implications.” *Systematic Botany* 17, no. 1: 115–143.
- Wendel, J. F., C. L. Brubaker, and A. E. Percival. 1992. “Genetic Diversity in *Gossypium hirsutum* and the Origin of Upland Cotton.” *American Journal of Botany* 79, no. 11: 1291–1310.
- Westengen, O. T., Z. Huamán, and M. Heun. 2005. “Genetic Diversity and Geographic Pattern in Early South American Cotton Domestication.”

Wilkinson, L. 2011. “ggplot2: Elegant Graphics for Data Analysis by WICKHAM.” *H. Biometrics* 67, no. 2: 678–679.

Wu, Y., F. Liu, D.-G. Yang, et al. 2018. “Comparative Chloroplast Genomics of *Gossypium* Species: Insights Into Repeat Sequence Variations and Phylogeny.” *Frontiers in Plant Science* 9: 376.

Xu, Q., G. Xiong, P. Li, et al. 2012. “Analysis of Complete Nucleotide Sequences of 12 *Gossypium* Chloroplast Genomes: Origin and Evolution of Allotetraploids.” *PLoS One* 7, no. 8: e37128.

Yan, X.-L., S.-L. Kan, M.-X. Wang, et al. 2024. “Genetic Diversity and Evolution of the Plastome in Allotetraploid Cotton (*Gossypium* spp.).” *Journal of Systematics and Evolution* 62, no. 6: 1118–1136.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. 2017. “Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data.” *Methods in Ecology and Evolution* / *British Ecological Society* 8, no. 1: 28–36.

Yuan, D., C. E. Grover, G. Hu, et al. 2021. “Parallel and Intertwining Threads of Domestication in Allopolyploid Cotton.” *Advanced Science* 8, no. 10: 2003634.

Zhang, C., S.-S. Dong, J.-Y. Xu, W.-M. He, and T.-L. Yang. 2019. “PopLDdecay: A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files.” *Bioinformatics* 35, no. 10: 1786–1788.

Zhang, H., N. Mittal, L. J. Leamy, O. Barazani, and B.-H. Song. 2017. “Back Into the Wild—Apply Untapped Genetic Diversity of Wild Relatives for Crop Improvement.” *Evolutionary Applications* 10, no. 1: 5–24.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Population genetic structure analysis used for outgroup selection and interpretation of diversity within *G. barbadense*, using (A) PCA, (B) Cross-entropy values vs. the best ancestral populations (K), and (C) neighbour-joining analysis, LEA structure results (with K = 4) and cpDNA type. In (C), tree tips are labelled with sample ID. The nine *G. barbadense* individuals selected are highlighted with a black dash box in the phylogeny and genetic structure plot. **Figure S2:** Heatmap plot of genetic relatedness ($PI_HAT > 0.3$) between all 86 samples from Mound Key (MK), Guadeloupe (GD; $n = 21$), and Puerto Rico (PR_CR, PR_Ph with $n = 5$, PR_Hwy325). Richer colour intensity indicates higher degree of relatedness. **Figure S3:** Cross-entropy values for the structure plot. Each dot represents an average cross-entropy value for each number of ancestral populations (K) across 10 replicates; standard errors are represented by a vertical bar across each dot. **Figure S4:** TreeMix model selection with migration edges ranging from 0 to 8, with each model was replicated 10 times. The best model ($m = 7$) was selected based on Δm . **Figure S5:** (A) Maximum parsimony phylogenetic tree for all 158 samples using the whole plastome and excluding one of the two copies of the large inverted repeat. Each tree tip is labelled by a shape and an individual ID that represents their genetic group, which was assigned using genome-wide SNPs (see Figure 1A and Figure S1A). Clades are coloured based on assigned groups. (B) Haplotype network using the same plastome alignment, and the colours represent their cpDNA main types. **Figure S6:** Maximum likelihood trees for (A) 110 loci dataset and (B) whole plastome. Nodes with bootstrap value > 95 black dots, and nodes between 85 and 95 black dots. **Figure S7:** (A) Averaged proportions of ROH across all individuals in each population/group. (B) Correlations between the number of ROH and the proportion of the genome occupied by ROH. **Table S1:** mec70239-sup-0002-Supplementarytables.xlsx. *Gossypium hirsutum* samples included in this study and their average sequencing depth. **Table S2:** Outgroup samples from *Gossypium barbadense*, *G. mustelinum*, and *G. darwinii* samples included in this.