

Genomic Signatures of Somatic Mutation and Selection in a Bud Sport Population of *Bougainvillea × buttiana* ‘Miss Manila’

Hongyan Meng

Fujian Institute of Subtropical Botany

Qun Zhou

13599535812@163.com

Fujian Institute of Subtropical Botany

Duchao Chen

Xiamen Chaozhiran Horticulture Co., Ltd

Bayan Huang

Xiamen Weiguanshu Biotech Co., Ltd

Mingqiong Zheng

Fujian Institute of Subtropical Botany

Wanqi Zhang

Fujian Institute of Subtropical Botany

Research Article

Keywords: *Bougainvillea*, bud sport, genotyping-by-sequencing (GBS), population genomics, genetic differentiation, selection signal, somatic mutation.

Posted Date: January 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8303143/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background In *bougainvillea* breeding, bud sports (somatic mutations) offer a rapid method for developing new varieties by altering specific traits while preserving the elite genetic background of the original cultivar. However, a comprehensive understanding of their genomic architecture, genetic differentiation from conventional varieties, and the molecular mechanisms underlying their formation remains limited. This study aims to elucidate the population genomic characteristics of bud sports derived from the commercial variety *Bougainvillea ×buttiana* 'Miss Manila'.

Results Using genotyping-by-sequencing (GBS) on 39 accessions (27 bud sports and 12 conventional varieties), we identified 64,810 high-quality SNPs. Population genomic analyses revealed significantly reduced genetic diversity (nucleotide diversity π and expected heterozygosity H_e) in the bud sports compared to conventional varieties, consistent with their clonal origin. Principal component, phylogenetic, and ADMIXTURE analyses (optimal $K=4$) demonstrated clear genetic differentiation and distinct population structures between the two groups. The bud sport population possessed far fewer private alleles and exhibited a less negative Tajima's D value. Genome scans identified regions with high genetic differentiation (F_{ST}) and signals of selective sweeps (XP-CLR) in bud sports. Functional analysis highlighted differentially mutated genes between the groups, involving pathways such as ubiquitin-mediated proteolysis and RNA transport, with genes like Bou_119143 (UDP-rhamnose rhamnosyltransferase 1) showing high mutation frequency specifically in bud sports.

Conclusions Our results demonstrate that bud sports of 'Miss Manila' are genetically distinct, clonal lineages with a narrow genetic base shaped by somatic mutation and selection. This study provides the first population-genomic evidence for the genetic essence of *bougainvillea* bud sports, confirming them as efficient sources for germplasm innovation. The identified genomic regions and candidate genes lay a foundation for marker-assisted selection and future molecular breeding in *bougainvillea*.

1. Introduction

The global ornamental horticulture industry thrives on continuous innovation and the introduction of new plant varieties, with genetic diversity serving as the cornerstone for breeding cultivars endowed with enhanced aesthetic value, environmental resilience, and market appeal (1). *Bougainvillea*, a quintessential tropical and subtropical genus, has cemented its status as a landscape and pot plant staple worldwide, celebrated for its vividly colored bracts, prolonged flowering period, and remarkable drought tolerance. The continuous enhancement of its commercial value largely depends on the ongoing selection of new varieties with novel bract colors, variegated foliage, or compact plant architecture, which in turn drives the continuous optimization and innovation of breeding techniques(2).

In *bougainvillea* breeding practice, sexual hybridization and somatic bud sports are two major sources of variation(2). Traditional breeding predominantly relies on sexual hybridization, a technique that integrates favorable genetic material from different parents through meiotic recombination. Indeed, the three fundamental species of *bougainvillea*—*B. glabra*, *B. spectabilis*, and *B. peruviana*—along with their

hybrid derivatives *B. ×buttiana*, *B. ×spectoperuviana*, and *B. ×spectoglabra*, form the genetic framework of the hundreds of horticultural varieties available today(3, 4). Hybridization can aggregate superior genes from different parents in terms of bract color, plant architecture, and stress resistance, achieving unprecedented trait combinations(2, 5). However, hybrid breeding also has significant limitations, such as a lengthy juvenile phase, highly heterozygous genetic backgrounds, difficulties in genetically dissecting complex traits, and the susceptibility of breaking up well-adapted gene combinations in segregating progeny(4, 6).

In contrast, selection based on somatic mutations (bud sports) offers an efficient alternative pathway. Bud sports can directly generate specific trait variations while completely preserving the elite comprehensive traits and genetic background of the original cultivar, enabling a "one-step" variety improvement(5, 7, 8). This effectively overcomes the issues of trait segregation and long breeding cycles associated with sexual reproduction. For example, in *bougainvillea*, important cultivars like 'Mrs. Butt' have given rise to several new varieties, including 'Scarlet Queen', through a series of bud sports, significantly enriching *bougainvillea*'s varietal diversity(9, 10).

In recent years, with the rapid development of genomic technologies, significant breakthroughs have been made in *bougainvillea* research. The completion of a chromosome-level reference genome for *B. glabra* has provided a crucial foundation for analyzing its biological characteristics at the molecular level(11). Furthermore, large-scale genome sequencing of the hybrid cultivar *B. × buttiana* 'Mrs Butt' revealed whole-genome duplication events during its evolution, providing valuable clues for understanding variety origin and the betalain pigment metabolism pathway(12). Moreover, population genomics studies, using reduced-representation genome sequencing on 84 *bougainvillea* accessions, not only classified them into six genetic subgroups (B1–B6) but also detected extensive gene flow between these subgroups. Through selective sweep analysis, these studies further found that metabolic pathways related to defense response and cell division regulation (such as terpenoid and triterpenoid biosynthesis, glycerophospholipid metabolism, etc.) were under significant selection in bud sport varieties, offering new perspectives for revealing the molecular mechanisms of bud sport occurrence and the domestication direction of *bougainvillea* (13).

Despite significant progress in existing research, most current work still focuses on macro-level genetic diversity analysis between species or varieties. There remains a lack of systematic and in-depth research on the internal genomic structure of bud sport populations derived from a single variety, the fine-scale genetic differentiation between them and conventional cultivars, the selection signals driving the formation of bud sport traits, and the related key genes.

To accurately address the aforementioned questions, this study established a population of 39 *Bougainvillea* accessions for systematic analysis. The population comprised two parts: a core group of 27 bud sport mutants derived from the commercial variety *B. × buttiana* 'Miss Manila' and a control group of 12 representative cultivars obtained through conventional breeding. Genotyping-by-sequencing (GBS) technology was employed to generate 64,810 high-quality SNP markers, based on which a

systematic analysis was conducted on the genomic architecture, selection signals, and genetic differentiation characteristics of the bud sport population. This study reveals, for the first time from a population genomics perspective, the genetic essence of *bougainvillea* bud sports, providing not only a theoretical basis for bud sport breeding but also laying the foundation for precision breeding driven by marker-assisted selection and key gene cloning.

2. Materials and Methods

2.1 Plant Materials

A total of 39 *Bougainvillea* accessions were analyzed in this study (Table S1). The collection was comprised exclusively of two forms of this variety. The core population of 27 somatic mutants (bud sports) derived from the cultivar *Bougainvillea* × *buttiana* 'Miss Manila', numbered from S1 to S27. These accessions, which display distinct flower color variations and leaf morphological traits, were collected from diverse cultivation areas across the Philippines, Malaysia, and China (Taiwan, Fujian, Guangdong, and Yunnan provinces/regions). The control set included 12 representative cultivars obtained through conventional breeding, covering several species and hybrids such as *B. spectabilis*, *B. glabra*, and *B. × buttiana*. These cultivars are stable, non-sport genotypes and were used for comparative analysis with the mutant accessions.

All plant materials examined in this study were sourced from the long-term living collections of the National Germplasm Repository for Bougainvillea in China, situated within the Xiamen Botanical Garden (Xiamen, Fujian, China). These collections consist exclusively of cultivated varieties; no wild populations were included. Consequently, no physical voucher specimens have been deposited in a public herbarium. All accessions are maintained as permanent living plants in the repository, with complete documentation including collection origin, cultivation history and identification records archived and available for verification upon request. The taxonomic identification of all materials was verified by Prof. Kefu Huang, a senior taxonomic expert affiliated with the National Germplasm Repository for Bougainvillea, following the taxonomic criteria outlined in *Flora of China* (14).

Young leaves were sampled from all accessions in December 2023, immediately flash-frozen in liquid nitrogen, and stored at -80°C for subsequent DNA analysis. .

2.2 DNA Extraction and Genotyping-by-Sequencing (GBS)

Genomic DNA was isolated from the young leaf tissue using the DNasecure Plant Kit (TIANGEN, China) according to the manufacturer's instructions. DNA quality and concentration were assessed via 1% agarose gel electrophoresis, NanoDrop spectrophotometry (Thermo Fisher Scientific, USA), and Qubit 4 Fluorometry (Thermo Fisher Scientific, USA). Only high - quality DNA samples (concentration > 50 ng/μL, OD_{260/280} ≈ 1.8–2.0, OD_{260/230} ≈ 2.0–2.5) were used for further analysis.

Library preparation was performed using the Super-GBS protocol (15). Briefly, approximately 100 ng of genomic DNA from each sample was digested with PstI - HF and MspI restriction enzymes (New England Biolabs, USA). Following digestion, sample - specific barcoded adapters (Illumina) were ligated to the fragments. Fragments of 300–700 bp were size-selected using magnetic beads (Beckman Coulter, USA) and then amplified by PCR. The final qualified libraries were pooled and sequenced on an Illumina NovaSeq 6000 platform (Illumina, USA) to generate 150 bp paired - end reads.

2.3 SNP Calling and Filtering

Raw sequencing reads were demultiplexed and subjected to quality control using Fastp v0.20.0 (16) to remove adapters and low-quality sequences, resulting in clean reads. Given the availability of the reference genome for *B. × buttiana* 'Mrs. Butt', a reference-guided SNP discovery pipeline was implemented in Stacks v2.4(17). This process involved building a catalog of consensus loci from all samples and then genotyping each sample against this catalog. The initial raw SNPs were stringently filtered using Vcftools v0.1.16 (18) with the following criteria: loci had to be diploid, possess a total read depth (DP) ≥ 4 , a minor allele frequency (MAF) ≥ 0.01 , and a genotyping missing rate $< 20\%$. The resulting high-quality SNP set was used for all downstream population genetic analyses.

2.4 Population Genetic Analysis

Genetic relationships among the 39 accessions were investigated using multiple complementary approaches. A neighbor - joining (NJ) phylogenetic tree was constructed based on genetic distances with 1000 bootstrap replicates using Treebest v1.9.2 (19) and visualized with FigTree v1.4.4(20). Principal Component Analysis (PCA) was conducted with GCTA v1.26.0(21) to visualize genetic clustering, and the results were plotted in R v4.3.1. Population structure was inferred using ADMIXTURE v1.3.0(22) for a hypothetical number of ancestral populations (K) ranging from 1 to 10. The optimal K value was determined by the lowest cross - validation error.

Genetic diversity parameters, including expected heterozygosity (H_e), observed heterozygosity (H_o), polymorphism information content (PIC), and nucleotide diversity (π), were calculated using the R package *genepop* (23). Genetic differentiation between populations was quantified by the pairwise fixation index (F_{st}) using the R package *StAMPP* (24). The number of private alleles between the two populations was calculated using the *pegas* package in R(25), and the differences between populations were compared via a t-test with a significance level of $\alpha = 0.05$.

2.5 Selection Signal and Neutrality Test Analysis

Selective sweep regions between the two populations were detected using XP-CLR v1.0 software (26), with the conventional population as the reference population and the bud mutation population as the target population. The parameter settings were as follows: window size = 50 kb, step size = 10 kb, and maximum linkage disequilibrium (LD) distance = 50 kb. The XP-CLR score for each window was calculated, and windows with the top 5% of scores were defined as potential selective sweep regions.

Tajima's D neutrality test was used to analyze the evolutionary dynamics of the populations. Tajima's D values were calculated separately for the two populations (window size: 50 kb, step size: 10 kb) using the *pegas* package in R (25), and 1000 simulation tests were performed to evaluate the significance of deviations from neutrality. Differences in Tajima's D values between the two populations were compared via an independent-samples t-test with a significance level of $\alpha = 0.0001$.

2.6 Functional Enrichment and Differential Mutation Gene Analysis

Using "mutation frequency of bud mutation population – mutation frequency of conventional population" as the indicator, differential analysis was performed with the *DESeq2* package in R (27). The screening criteria were set as follows: false discovery rate (FDR) < 0.01 and absolute value of mutation frequency difference > 0.5.

KEGG pathway enrichment analysis was conducted on the differentially mutated genes using the *clusterProfiler* package in R(28), with the KEGG annotation information of the *Bougainvillea* reference genome as the background. The enrichment significance threshold was set at FDR < 0.05, and bar plots of the top 10 significantly enriched pathways were generated.

Gene Set Enrichment Analysis (GSEA) was performed using the *gseKEGG* function of the *clusterProfiler* package. First, "difference in mutation frequency between the bud mutation population and the conventional population" was used as the gene ranking indicator to construct a gene ranking list. With the KEGG pathway gene sets of the *Bougainvillea* reference genome as the background (consistent with the KEGG enrichment analysis), the number of permutations was set to 1000 and the significance threshold to FDR < 0.05.

3. Results

3.1 Sequencing Data Statistics and SNP Discovery

Reduced-representation genome sequencing (RRGS) was performed on 39 germplasm samples of *B. × buttiana* 'Miss Manila', comprising 27 bud sport varieties and 12 conventional non-sport varieties, using the Illumina NovaSeq PE150 platform. A total of 275,325,940 raw reads corresponding to 40.48 Gb of raw base pairs (ranging from 0.63 to 2.63 Gb per sample). Following stringent quality control, which included the removal of adapter sequences, low-quality reads, and reads with excessive ambiguous bases, 261,217,220 clean reads (33.96 Gb of clean bases) were retained. The average proportion of clean reads was 83.55% (range: 80.83%–87.04%), yielding an average of 0.67 Gb of high-quality data per sample, sufficient for subsequent genetic analyses (Table S2). Quality assessment indicated high data reliability, with an average GC content of 41.55% (range: 39.05%–48.21%), Q20 of 96.96% (range: 95.27%–97.84%), and Q30 of 91.98% (range: 88.22%–93.96%). Subsequent SNP calling and filtering identified 64,810 high-confidence SNPs.

3.2 Genetic Diversity of SNP Markers and Population Comparison

Genetic diversity was assessed across all 39 germplasms using the 64,810 high-quality SNPs. The Polymorphism Information Content (PIC) and nucleotide diversity (π) both exhibited right-skewed distributions, with mean values of 0.139 and 0.166, respectively, indicating a preponderance of loci with low to moderate polymorphism (Fig. 1A, C). The comparison between expected heterozygosity (H_e , mean = 0.164) and observed heterozygosity (H_o , mean = 0.070) revealed that H_o was consistently lower than H_e across most loci (Fig. 1B), a pattern potentially indicative of underlying population substructure. The Minor Allele Frequency (MAF) spectrum showed a balanced composition, with 47.9% of SNPs classified as low-frequency variants ($MAF < 0.05$) and 52.1% as common variants ($MAF \geq 0.05$) (Fig. 1D).

Comparative analysis between the bud sport population ($n=27$) and the conventional population ($n=12$) revealed significantly reduced genetic diversity in the former. The bud sport group showed markedly lower values ($P < 0.001$) for nucleotide diversity (π), expected heterozygosity (H_e), and PIC compared to the conventional group (Fig. 1E). This consistent reduction across multiple metrics indicates a narrower genetic base in the bud sports, consistent with their origin from somatic mutations of a single progenitor cultivar.

3.3 Population Genetic Structure

The population genetic structure of the *B.xbuttiana* 'Miss Manila' varieties was investigated using multiple complementary approaches. Principal Component Analysis (PCA) revealed clear genetic separation between the sport mutants and conventional varieties (Fig. 2A). The first two principal components explained 28.96% (PC1) and 14.81% (PC2) of the total genetic variation, respectively. While the conventional varieties showed a dispersed distribution, the sport mutants formed a tight, distinct cluster, indicating substantial genetic differentiation between the two groups.

To determine the optimal number of genetic clusters, we performed population structure analysis using ADMIXTURE with K values ranging from 1 to 10. The cross-validation error reached a minimum at $K=4$ (Fig. 2B), indicating that four ancestral populations best explained the genetic structure. At this optimal K value, the analysis identified four distinct genetic clusters (POP1-POP4; Fig. 2C). Conventional varieties were primarily assigned to POP1 and POP3, while sport mutants predominantly constituted POP2 and POP4. Notably, two conventional accessions (SJM145-1 and SJM013-1) showed substantial ancestral components from POP2 and POP4, respectively, suggesting potential phylogenetic affinities with the sport groups.

The genetic partitioning observed in the structure analysis was further supported by a neighbor-joining phylogenetic tree (Fig. 2D), which exhibited a topology highly consistent with the ADMIXTURE results. The sport mutants and conventional varieties formed distinct clades, with the sport mutants primarily clustering into two branches corresponding to POP2 and POP4 from the structure analysis.

Collectively, these analyses demonstrate a well-defined genetic structure and significant divergence between the sport mutant and conventional variety groups, supporting their distinct genetic origins and evolutionary histories.

3.4 Characterization of Bud Sport-Specific Mutations

The distribution of high-quality SNPs was analyzed across the top 30 longest scaffolds, providing broad genomic coverage despite an uneven distribution with enrichment in specific regions (Fig. 3A).

Analysis of private alleles revealed a substantial disparity between the populations (Fig. 3B). The conventional population contained 23,797 private alleles, vastly outnumbering the 335 found in the bud sport population ($p < 0.0001$). This difference reflects their distinct histories: the bud sports, derived clonally from a single progenitor, possess a limited and recent gene pool, restricting the accumulation of private alleles. In contrast, the conventional varieties represent a more diverse genetic background.

Consistent with this, the average heterozygosity of the bud sport population was significantly lower than that of the conventional population ($p < 0.01$; Fig. 3C), a direct consequence of the founder effect and clonal propagation, which limit the introduction of new genetic variation.

3.5 Genetic Differentiation and Signatures of Selection

To explore genetic divergence and potential selective pressures, we analyzed genetic differentiation (F_{ST}), selective sweeps (XP-CLR), and neutrality (Tajima's D) between the bud sport and conventional populations (Fig. 3D–F).

Pairwise F_{ST} analysis revealed multiple genomic regions with significantly elevated differentiation (Fig. 3D), highlighting loci that have substantially diverged between the two populations.

Genome-wide scanning with XP-CLR identified several regions with high scores in the bud sport population, indicating potential selective sweeps(29) (Fig. 3E). These regions are candidates for positive selection, potentially associated with novel horticultural traits arising from somatic mutation.

Tajima's D analysis showed negative mean values for both populations, consistent with population expansion or purifying selection(30). However, the bud sport population exhibited a significantly less negative Tajima's D than the conventional population ($P < 0.0001$; Fig. 3F), suggesting that the conventional population experienced a stronger or more recent expansion, or maintains more low-frequency variants. The less negative value in bud sports aligns with their origin from a limited number of somatic mutants and subsequent clonal propagation.

3.6 KEGG Enrichment and Differential Mutated Gene Analysis

To elucidate the biological processes underlying the differentiation between the bud sport and conventional populations, functional annotation of the detected genetic variations was performed, followed by KEGG pathway enrichment analysis of the top 10 most enriched pathways in each

population. The results revealed 7 shared enriched pathways between the two groups, accounting for 70% of the top 10 pathways in both the bud sport and conventional populations. These core pathways include Ubiquitin-mediated proteolysis (ko04120), RNA degradation (ko03018), Fatty acid biosynthesis (ko00061), as well as Ribosome biogenesis in eukaryotes (ko03008), Inositol phosphate metabolism (ko00562), RNA transport (ko03013), and Endocytosis (ko04144) (Fig. 4A, B). However, no significant difference in enrichment significance was found between the populations for these pathways (FDR > 0.05; Fig. 4C).

We identified individual genes with significant differences in mutation frequency between the groups (FDR < 0.01; Fig. 4D, Table S3). Several genes showed markedly higher mutation frequencies in the bud sport population, including *Bou_106388* (DNA-binding protein), *Bou_24586* (60S ribosomal protein), *Bou_144753* (zinc finger protein), *Bou_93970* (SelT-like protein), and *Bou_119143* (UDP-rhamnose rhamnosyltransferase 1), with frequency differences exceeding 0.64.

Conversely, several genes had significantly higher mutation frequencies in the conventional population. Notably, *Bou_96858*, *Bou_129238*, *Bou_105951* (E3 ubiquitin-protein ligase RWD3), and *Bou_9911* were absent in bud sports (frequency = 0) but present at high frequency (0.83) in conventional varieties. Other genes with higher frequencies in conventional populations included *Bou_65905* (transmembrane protein 64-like), *Bou_29416* (carboxylesterase), *Bou_27687* (cytochrome c oxidase subunit 5b), *Bou_133689* (actin-related protein 2/3 complex subunit 3), and *Bou_112008* (COP9 signalosome complex subunit 7).

The functional diversity of these differentially mutated genes, spanning protein degradation, transcriptional regulation, metabolism, and cellular structure, suggests multiple potential mechanisms contributing to the phenotypic divergence between bud sport and conventional populations.

4. Discussion

The genomic era has provided unprecedented tools for deciphering the origin and evolution of ornamental plants. Our super-GBS analysis of a *Bouainvillea* population comprising both a core bud sport group derived from *B. × buttiana* 'Miss Manila' and a control group of conventionally bred varieties offers a comprehensive perspective on the genetic underpinnings of bud sports, affirming their critical and irreplaceable role in the breeding of this species.

4.1 Unique Genetic Lineage Shaped by Asexual Reproduction and Population Bottleneck

Multi-dimensional population genetic analyses, including principal component analysis (PCA), ADMIXTURE-based population structure inference, and neighbor-joining (NJ) phylogenetic tree construction, consistently revealed that the *B. × buttiana* 'Miss Manila' bud mutation population constitutes a genetically distinct and cohesive lineage, exhibiting significant genetic differentiation from conventional non-bud mutation cultivars (Fig. 2). This pattern is highly consistent with the known horticultural breeding history of bud sports: all these bud sports originated from somatic mutations in the single parental cultivar *B. × buttiana* 'Miss Manila' and were subsequently propagated asexually via

cuttings or grafting. Strict asexual reproduction inherently restricts gene flow and the accumulation of novel genetic variations(31), resulting in a significantly narrower genetic base of the bud mutation population compared to conventional cultivars (Fig. 1E).

Quantitatively, the bud mutation population exhibited significantly lower nucleotide diversity (π), expected heterozygosity (H_e), and polymorphism information content (PIC) ($P < 0.001$). These metrics are all typical genomic signatures of a population that experienced a bottleneck followed by asexual expansion (32). The substantial difference in private alleles further corroborates this conclusion: the conventional population contained 23,797 private alleles, whereas the bud mutation population had only 335 ($p < 0.0001$; Fig. 3B). This sharp reduction in private alleles aligns with recent studies documenting widespread private allele depletion in asexually reproducing lineages(33, 34). Despite the sterility of most *Bougainvillea* cultivars, conventional populations likely originated from multiple genetic backgrounds and occasionally undergo outcrossing, enabling them to maintain a relatively diverse gene pool and facilitate the accumulation of unique alleles(3, 11). In contrast, as asexual lineages, bud sports inherit only a limited set of alleles from their parental cultivar, with novel variations arising solely from rare somatic mutations. This explains the scarcity of private alleles and reduced heterozygosity in the bud mutation population (Fig. 3C).

Notably, ADMIXTURE analysis revealed that the ancestry components of two conventional germplasms (SJM145-1 and SJM013-1) exhibit overlap with those of the bud mutation clusters (POP2 and POP4) (Fig. 2C). This suggests a potential phylogenetic relationship between them, implying they may have originated from a common ancestor. The result indicates that the evolutionary history of *bougainvillea* germplasm resources remains complex, even within seemingly distinct clusters.

4.2 Evolutionary Dynamics: Selection Signals and Population History

Although asexual reproduction constrains the variation potential of bud mutation genomes, our analyses reveal that the evolution of these genomes is not a random process; rather, it is jointly shaped by population history and selection pressures. Pairwise F_{ST} analysis identified several genomic regions with significantly elevated genetic differentiation between bud mutation and conventional varieties (Fig. 3D), suggesting that divergence at these loci may be driven by genetic drift or selection(35). Complementing this, XP-CLR scanning uncovered distinct signatures of selective sweeps within the bud mutation population (Fig. 3E), indicating that these genomic regions have likely undergone positive selection(26). This selection is presumably linked to artificial selection for favorable horticultural traits during domestication and cultivation, such as bract coloration, plant architecture, and flowering phenology. Collectively, these findings underscore the notion that bud mutations are not entirely random in their evolutionary trajectory: directional artificial selection for novel and favorable phenotypes acts to screen and fix beneficial somatic mutations within asexual lineages, ultimately leading to the formation of the selective sweeps observed in our genomic analyses.

Tajima's D analysis further revealed distinct demographic histories between the two populations. Both groups showed negative mean Tajima's D values (Fig. 3F), consistent with population expansion or

purifying selection(36, 37). However, the bud mutation population exhibited a significantly less negative Tajima's D ($P < 0.0001$; Fig. 3F). This pattern aligns with the evolutionary model for bud mutation populations: such mutations arise from a small set of somatic mutants that subsequently undergo clonal propagation. In this scenario, clonal expansion constrains the introduction of new low-frequency genetic variants, thereby preventing the excessive accumulation of rare alleles that would otherwise shift Tajima's D to more negative values . In contrast, the more negative Tajima's D values observed in the conventional population imply either historical population expansion or a higher abundance of rare variants(38). These patterns align with the population's genetically diverse background and the potential for occasional outcrossing events .

Together, these results highlight the divergent evolutionary trajectories of the two groups: bud mutations represent a **young, genetically constrained lineage** governed by asexual reproduction and artificial selection, whereas conventional varieties constitute an **ancient, dynamic population** with a more complex demographic history.

4.3 Functional Implications of Differentially Mutated Genes

This study identified genes with significantly different mutation frequencies between bud sport and conventional varieties, offering critical insights into the molecular basis of their phenotypic differentiation. These genes span diverse functional domains, including pigment biosynthesis, transcriptional regulation, protein degradation, and cellular metabolism, indicating that the development of bud sport-specific traits involves the synergistic action of multiple pathways (Fig. 4D).

A particularly notable finding is the significantly elevated mutation frequency of the *Bou_119143* gene (encoding UDP-rhamnose rhamnosyltransferase 1) in the bud sport population. As a member of the Caryophyllales, *Bougainvillea* produces betalains (rather than anthocyanins) as the primary pigments responsible for bract coloration(12). While this enzyme is traditionally known to catalyze the glycosylation of flavonoids(39-42), emerging evidence indicates that UDP-rhamnose-dependent glycosyltransferases are highly likely to modify betalain precursors or betalain (12, 43). This glycosylation represents a key post-translational modification that regulates the stability, water-solubility, and intracellular localization of betalains, thereby directly modulating pigment intensity and color variation. Given that bract color is a primary target of selection in *Bougainvillea* bud sports, mutations in *Bou_119143* likely alter the glycosylation pattern of betalains and flavonoids, contributing to the diverse bract color phenotypes (e.g., pink, red, and purple) observed in the *B. × buttiana* 'Miss Manila' bud sport lineage. Similarly, the increased mutation frequencies of genes involved in transcriptional regulation (e.g., *Bou_106388*, encoding a DNA-binding protein, and *Bou_144753*, encoding a zinc finger protein) suggest that alterations in gene regulatory networks may drive broader phenotypic variation in important horticultural traits such as plant architecture, branching pattern, and flowering time (44-46).

Conversely, several genes were completely devoid of mutations in the bud sport population despite reaching mutation frequencies up to 83% in conventional varieties. These include *Bou_105951* (encoding

the E3 ubiquitin-protein ligase RFWD3) and *Bou_112008* (encoding COP9 signalosome subunit 7). This pattern implies the existence of genetic constraints specific to the asexual *B. × buttiana* 'Miss Manila' lineage. E3 ubiquitin-protein ligases are central regulators of protein turnover, cell cycle progression, and stress responses (47). The COP9 signalosome regulates the activity of cullin-RING ligase (CRL) families of E3 ubiquitin ligase complexes, and play critical roles in regulating gene expression, cell proliferation, and cell cycle (48). Consequently, mutations in these essential pathways may be incompatible with the survival or adaptability of this clonal background. Their elimination in bud sports underscores the action of purifying selection in pruning deleterious genetic variation from the asexual genome.

4.4 Pathway-Level Analysis and Phenotypic Implications

Comparative KEGG pathway enrichment analysis of the top 10 ranked pathways for each population revealed a landscape of shared core functions alongside group-specific signatures. Seven pathways were common to both groups, including “Ubiquitin mediated proteolysis,” “RNA degradation,” “Ribosome biogenesis in eukaryotes,” “Fatty acid biosynthesis,” “Inositol phosphate metabolism,” “RNA transport,” and “Endocytosis.” This substantial overlap underscores a conserved foundation of essential cellular processes related to protein homeostasis, RNA metabolism, and basic cellular functions.

Despite this common core, each population exhibited distinct pathway enrichments. The bud sport population uniquely featured “Other glycan degradation,” “mRNA surveillance pathway,” and “Circadian rhythm-plant.” These specificities may be more directly linked to observed phenotypic divergence. For instance, enhanced “glycan degradation” could influence cell wall dynamics or pigment modification(49, 50), while alterations in “circadian rhythm” pathways are known to affect key horticultural traits like flowering time and photomorphogenesis(51). In contrast, the conventional population uniquely enriched pathways such as “Lysine biosynthesis,” “Synthesis and degradation of ketone bodies,” and “Phosphatidylinositol signaling system,” reflecting a broader engagement with primary metabolism and complex signal transduction(52-54), a pattern consistent with its more diverse genetic background and sexual reproductive history.

Crucially, while these overlapping core pathways were significantly enriched in both groups, their enrichment levels did not differ statistically between the populations (Fig. 4C). This key result indicates that the major phenotypic differentiation between bud sports and conventional varieties is not driven by large-scale changes in the utilization of fundamental biological pathways. Instead, it is likely orchestrated by targeted mutations within specific genes of these conserved pathways, such as *Bou_119143* in modification of betalains and flavonoids, complemented by the influence of the population-specific pathways noted above. Therefore, we conclude that bud sport variation arises predominantly from precise genetic modifications in key functional nodes against a backdrop of conserved core cellular machinery, rather than from a wholesale reorganization of metabolic or regulatory networks.

4.5 Limitations and Future Research Directions

While this study provides insights, some limitations exist. Although a chromosome-level reference genome of *B. glabra* has been published(11), it lacks a complete gene annotation file, precluding its direct use for function-related analyses. Consequently, this study ultimately adopted the scaffold-level reference genome of *B. × buttiana* "Mrs. Butt" with available gene annotations for analysis(12). However, the scaffold-level assembly characteristics (Fig. 3A) still limited the fine mapping of selective sweep regions and the precise chromosomal anchoring of mutated genes, hindering further improvement in the resolution of relevant genetic analyses. Additionally, candidate genes require validation through functional assays such as expression analysis or genetic transformation. Future work integrating precise phenotypic data with genomic approaches like GWAS would help definitively link genetic variants to key horticultural traits.

5. Conclusion

This study analyzed 39 *Bougainvillea* accessions using genotyping-by-sequencing (GBS). It revealed that the bud sports of 'Miss Manila' are clonal lineages with a narrow genetic basis shaped jointly by somatic mutations, artificial selection, and purification selection. These bud sports exhibit significant genetic differentiation from conventional cultivars and a distinct population structure. Meanwhile, we identified selected genomic regions and high-frequency mutated genes, confirming that bud sport breeding is an efficient approach for *Bougainvillea* germplasm innovation. This research deepens our understanding of the mechanism underlying bud sport formation and establishes the core role of bud sports in *Bougainvillea* genetic.

Declarations

Clinical trial number

Not applicable.

Ethics, Consent to Participate, and Consent to Publish declarations

Not applicable

Competing Interest

The authors declare no competing financial interest.

Funding

This work was supported by the joint research project of the natural science foundation of Xiamen , China (3502Z202573339)

CRediT authorship contribution statement

Hongyan Meng: Conceptualization, Investigation, Data curation, Methodology, Writing-review and editing. **Qun Zhou:** Supervision, Data curation, Project administration. **Duchao Chen:** Supervision, Resource. **Bayan Huang:** Investigation, Writing-original draft. **Mingqiong Zheng:** Investigation. **Wanqi Zhang:** Investigation.

Data availability

The datasets generated and analysed during the current study are available in the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA) repository, under accession number CRA057799 (<https://ngdc.cncb.ac.cn/gsa/browse/CRA057799>)

Acknowledgements

The authors would like to express their gratitude to the anonymous reviewers for their valuable comments and suggestions on this work.

References

1. Mekapogu M, Song H-Y, Lim S-H, Jung J-A. Genetic Engineering and Genome Editing Advances to Enhance Floral Attributes in Ornamental Plants: An Update. *Plants*. 2023;12(23):3983.
2. Datta SK, Jayanthi R, Janakiram T. Bougainvillea. In: Datta SK, Gupta YC, editors. *Floriculture and Ornamental Plants*. Singapore: Springer Singapore; 2020. p. 1-34.
3. Datta SK. Breeding of Bougainvillea: past, present, and future. *The Nucleus*. 2022;65(2):239-54.
4. Lin X, Lee SY, Ni J, Zhang X, Hu X, Zou P, et al. Comparative Analyses of Chloroplast Genome Provide Effective Molecular Markers for Species and Cultivar Identification in Bougainvillea. *International journal of molecular sciences*. 2023;24(20):15138.
5. Zhang W, Zhou Q, Lin J, Ma X, Dong F, Yan H, et al. Transcriptome analyses shed light on floral organ morphogenesis and bract color formation in Bougainvillea. *BMC Plant Biology*. 2022;22(1):97.
6. Li H, Junhai N, Sun L, Li Y, Leng Q, Chen J, et al. Ploidy Levels and DNA Contents of Bougainvillea Accessions Determined by Flow Cytometry Analysis. *HortScience*. 2022;57:1502-6.
7. Ban S, Jung JH. Somatic Mutations in Fruit Trees: Causes, Detection Methods, and Molecular Mechanisms. *Plants* [Internet]. 2023; 12(6):[1316 p.].
8. Ferreira MD, Rocha AD, Nascimento FD, Oliveira WD, Soares JM, Rebouças TA, et al. The Role of Somaclonal Variation in Plant Genetic Improvement: A Systematic Review. *Agronomy* [Internet]. 2023; 13(3):[730 p.].
9. Ohno S, Makishima R, Doi M. Post-transcriptional gene silencing of CYP76AD controls betalain biosynthesis in bracts of bougainvillea. *Journal of Experimental Botany*. 2021;72(20):6949-62.

10. Salam P, Bhargav V, Gupta D, Nimbolkar P. Evolution in *Bougainvillea* (*Bougainvillea Commers.*) - A review. *Journal of Applied and Natural Science*. 2017;9:1489-94.
11. Lan L, Li H, Xu S, Xu Y, Leng Q, Zhang L, et al. Chromosome-level and haplotype-resolved genome assembly of *Bougainvillea glabra*. *Scientific Data*. 2025;12(1):107.
12. Lan L, Zhao H, Xu S, Kan S, Zhang X, Liu W, et al. A high-quality *Bougainvillea* genome provides new insights into evolutionary history and pigment biosynthetic pathways in the Caryophyllales. *Horticulture Research*. 2023;10(8):uhad124.
13. Ma X, Wang H, Hu Y, Hu Q, Qian R, Zheng J, et al. ddRAD-seq data reveals the significant population structure and potential causes of population divergence in *Bougainvillea*. *Scientific Reports*. 2025;15(1):34200.
14. Committee FoCE. *Flora of China*: Beijing: Science Press; St. Louis: Missouri Botanical Garden Press; 2007-2019. Available from: <http://www.floraofchina.org>; www.iplant.cn/foc.
15. Wang Q, Liu Y, Yan L, Chen L, Li B. Genome-Wide SNP Discovery and Population Genetic Analysis of *Mesocentrotus nudus* in China Seas. 2021;Volume 12 - 2021.
16. Chen S. fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. 2025;4(5):e70078.
17. Rochette NC, Rivera-Colón AG, Catchen JM. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular ecology*. 2019;28(21):4737-54.
18. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8.
19. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*. 2009;19(2):327-35.
20. FigTree Development T. FigTree v1.4.4 Official Website 2025 [updated 2025-05-02. Available from: <http://tree.bio.ed.ac.uk/software/figtree/>.
21. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*. 2011;88(1):76-82.
22. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*. 2009;19(9):1655-64.
23. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular ecology resources*. 2008;8(1):103-6.
24. Pembleton LW, Cogan NO, Forster JW. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular ecology resources*. 2013;13(5):946-52.
25. Rotaru LI, Surleac M. PeGAS: a versatile bioinformatics pipeline for antimicrobial resistance, virulence and pangenome analysis. *Bioinformatics advances*. 2025;5(1):vbaf165.

26. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Genome research*. 2010;20(3):393-402.
27. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. *Journal of visualized experiments : JoVE*. 2021(175).
28. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Cambridge (Mass))*. 2021;2(3):100141.
29. Panigrahi M, Rajawat D, Nayak SS, Ghildiyal K, Sharma A, Jain K, et al. Landmarks in the history of selective sweeps. *Animal Genetics*. 2023;54(6):667-88.
30. Meleshko O, Martin Michael D, Flatberg Kjell I, Stenøien Hans K, Korneliussen Thorfinn S, Szövényi P, et al. Linked Selection and Gene Density Shape Genome-Wide Patterns of Diversification in Peatmosses. *Evolutionary Applications*. 2024;17(8):e13767.
31. Reusch TBH, Baums IB, Werner B. Evolution via somatic genetic variation in modular species. *Trends in Ecology & Evolution*. 2021;36(12):1083-92.
32. Jaron KS, Parker DJ, Anselmetti Y, Tran Van P, Bast J, Dumas Z, et al. Convergent consequences of parthenogenesis on stick insect genomes. *Science Advances*. 2022;8(8):eabg3842.
33. VanWallendael A, Alvarez M, Franks SJ. Patterns of population genomic diversity in the invasive Japanese knotweed species complex. *American journal of botany*. 2021;108(5):857-68.
34. Brandt A, Tran Van P, Bluhm C, Anselmetti Y, Dumas Z, Figuet E, et al. Haplotype divergence supports long-term asexuality in the oribatid mite *Oppeia nova*. *Proceedings of the National Academy of Sciences*. 2021;118(38):e2101485118.
35. Hasan A, Whitlock MC. FST and genetic diversity in an island model with background selection. *PLOS Genetics*. 2024;20(12):e1011225.
36. Vasilarou M, Alachiotis N, Garefalaki J, Beloukas A, Pavlidis P. Population Genomics Insights into the First Wave of COVID-19. *Life (Basel, Switzerland)*. 2021;11(2).
37. Zhong L, Zhu Y, Olsen KM. Hard versus soft selective sweeps during domestication and improvement in soybean. *Molecular ecology*. 2022;31(11):3137-53.
38. Dong X, Xiang Y, Li Y, Zhang Y. Lineage-specific evolution and resistance-virulence divergence in *Klebsiella pneumoniae* ST268: a global population genomic analysis. *Antimicrobial agents and chemotherapy*. 2025;69(10):e0070325.
39. Hu S, Wang B, Pei L, Wang J, Gan Y, Jiang L, et al. Advances and Challenges in Biomanufacturing of Glycosylation of Natural Products. 2024;10(7):349.
40. Ren C, Qian J, Wang Y, Xie L, Guo Y, Chen K, et al. Flavonoid UDP-glycosyltransferase in plants: functional identification, substrate recognition mechanism, and biotechnology application. *Phytochemistry Reviews*. 2025;24(5):4451-74.
41. Jiang N, Dillon FM, Silva A, Gomez-Cano L, Grotewold E. Rhamnose in plants - from biosynthesis to diverse functions. *Plant Science*. 2021;302:110687.

42. Hsu YH, Tagami T, Matsunaga K, Okuyama M, Suzuki T, Noda N, et al. Functional characterization of UDP-rhamnose-dependent rhamnosyltransferase involved in anthocyanin modification, a key enzyme determining blue coloration in *Lobelia erinus*. *The Plant journal : for cell and molecular biology*. 2017;89(2):325-37.
43. Polturak G, Heinig U, Grossman N, Battat M, Leshkowitz D, Malitsky S, et al. Transcriptome and Metabolic Profiling Provides Insights into Betalain Biosynthesis and Evolution in *Mirabilis jalapa*. *Molecular Plant*. 2018;11(1):189-204.
44. Hussain A, Liu J, Mohan B, Burhan A, Nasim Z, Bano R, et al. A genome-wide comparative evolutionary analysis of zinc finger-BED transcription factor genes in land plants. *Scientific Reports*. 2022;12(1):12328.
45. Shalmani A, Muhammad I, Sharif R, Zhao C, Ullah U, Zhang D, et al. Zinc Finger-Homeodomain Genes: Evolution, Functional Differentiation, and Expression Profiling Under Flowering-Related Treatments and Abiotic Stresses in Plants. *Evolutionary Bioinformatics*. 2019;15:1176934319867930.
46. Bollier N, Gonzalez N, Chevalier C, Hernould M. Zinc Finger-Homeodomain and Mini Zinc Finger proteins are key players in plant growth and responses to environmental stresses. *Journal of Experimental Botany*. 2022;73(14):4662-73.
47. Yang Q, Zhao J, Chen D, Wang Y. E3 ubiquitin ligases: styles, structures and functions. *Molecular Biomedicine*. 2021;2(1):23.
48. Qin N, Xu D, Li J, Deng XW. COP9 signalosome: Discovery, conservation, activity, and function. *Journal of Integrative Plant Biology*. 2020;62(1):90-103.
49. Su G, Lin Y, Wang C, Lu J, Liu Z, He Z, et al. Expansin SIExp1 and endoglucanase SICel2 synergistically promote fruit softening and cell wall disassembly in tomato. *The Plant Cell*. 2024;36(3):709-26.
50. Kato S, Hayashi M, Kitagawa M, Kajiura H, Maeda M, Kimura Y, et al. Degradation pathway of plant complex-type N-glycans: identification and characterization of a key α 1,3-fucosidase from glycoside hydrolase family 29. *Biochemical Journal*. 2018;475(1):305-17.
51. Mujahid M, Ambreen A, Zarlashat Y, Sarfraz Z, Iqbal MS, Waheed A, et al. Integration of Light and Circadian Signaling in Plant Gene Regulatory Networks: Implications for Photomorphogenesis and Stress Adaptation. *Biology (Basel)*. 2025;14(10).
52. Soares da Costa TP, Hall CJ, Panjekar S, Wyllie JA, Christoff RM, Bayat S, et al. Towards novel herbicide modes of action by inhibiting lysine biosynthesis in plants. *eLife*. 2021;10:e69444.
53. Liao P, Lung S-C, Chan WL, Bach TJ, Lo C, Chye M-L. Overexpression of HMG-CoA synthase promotes *Arabidopsis* root growth and adversely affects glucosinolate biosynthesis. *Journal of Experimental Botany*. 2020;71(1):272-89.
54. Munnik T, Testerink C. Plant phospholipid signaling: "in a nutshell". *Journal of Lipid Research*. 2009;50:S260-S5.

Figures

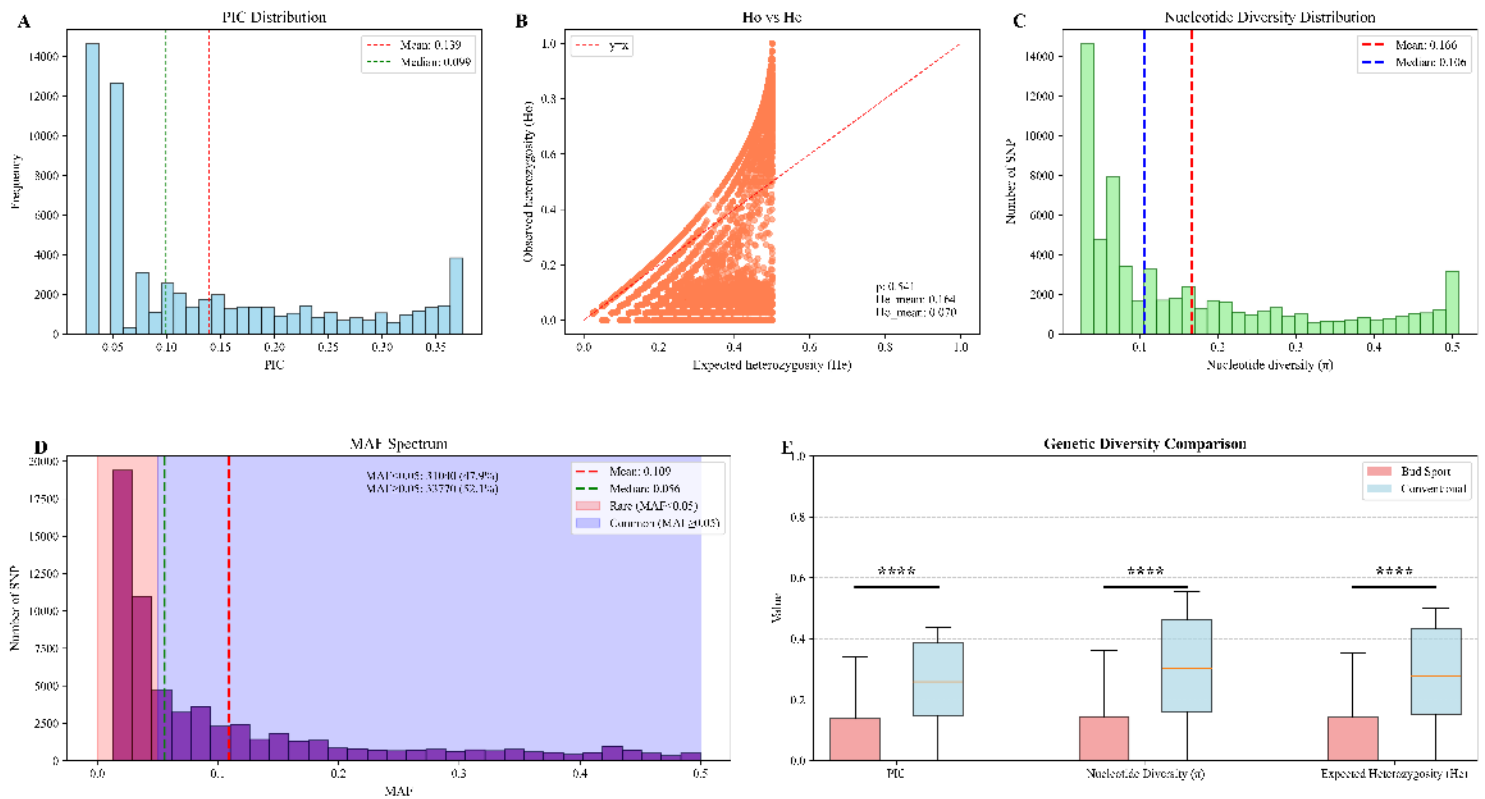


Figure 1

Genetic Diversity Analysis of 39 *B. x buttiana* 'Miss Manila' Variants.

(A) Distribution of the Polymorphic Information Content (PIC) values across the analyzed genome-wide markers. The red and green vertical dashed lines represent the mean and median PIC values, respectively. (B) Comparison of observed heterozygosity (H_o) and expected heterozygosity (H_e) among the variants. (C) Genome-wide distribution of nucleotide diversity (π). (D) Minor Allele Frequency (MAF) spectrum of all identified SNPs. (E) Comparative summary of key genetic diversity parameters, including nucleotide diversity (π), expected heterozygosity (H_e), and Polymorphic Information Content (PIC). Statistical significance (****) indicates $P < 0.0001$.

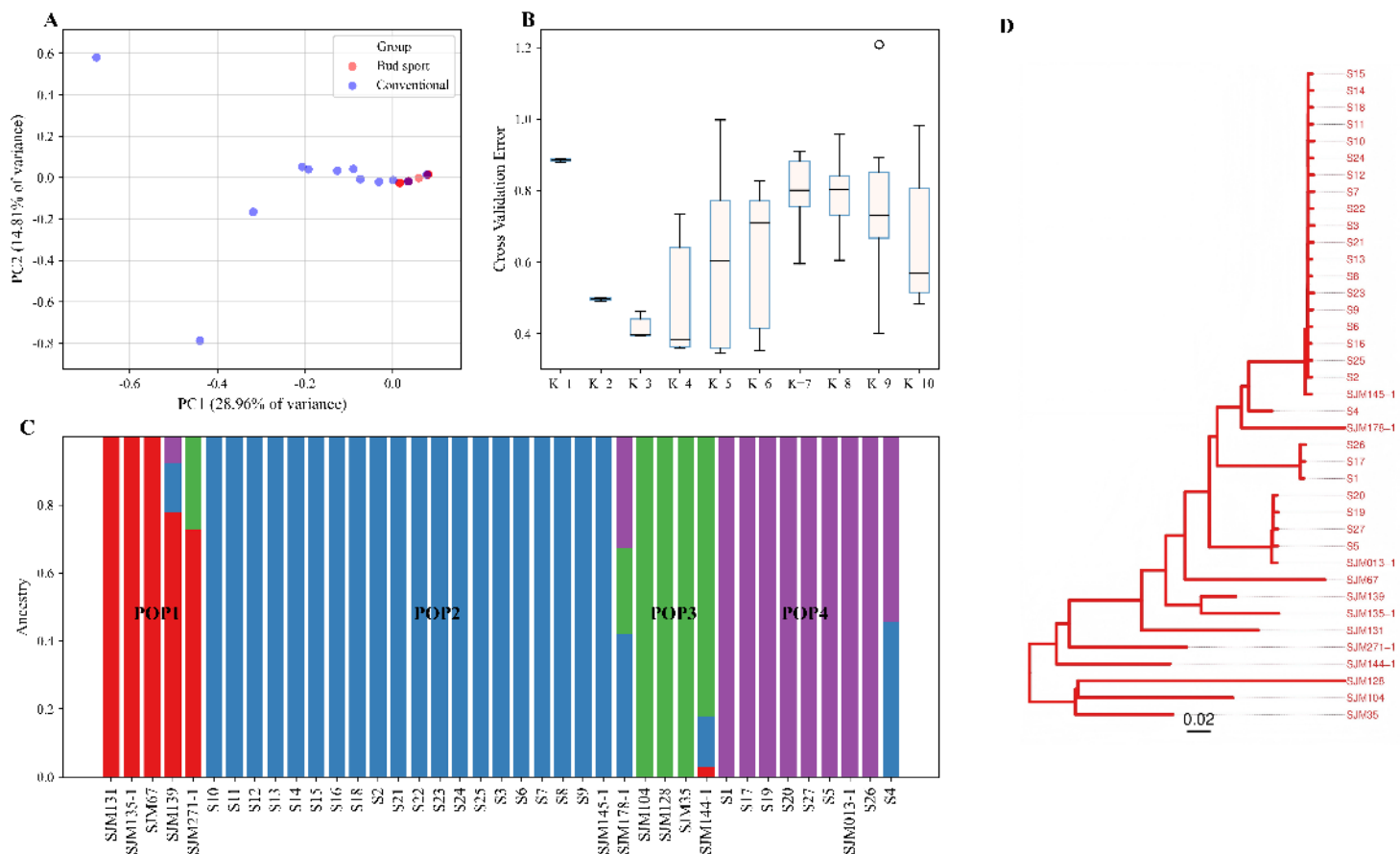


Figure 2

Population Genetic Structure of *B. x butiana* 'Miss Manila' Variants.

(A) Principal components analysis (PCA) plot based on genome-wide SNPs. The first two principal components (PC1 and PC2) are shown, illustrating the genetic relationship and stratification among individuals. (B) Determination of the optimal genetic cluster number (K). Cross-validation (CV) error across a range of K values (from 1 to 10) from the ADMIXTURE analysis. The K value with the minimum CV error (K=4) was selected as optimal. (C) Population structure bar plot inferred by ADMIXTURE at the optimal K=4. Each vertical bar represents an individual, and the colored segments represent the estimated proportion of ancestry from each of the K genetic clusters. (D) Phylogenetic tree constructed from genome-wide SNPs, depicting the evolutionary relationships among the *B. x butiana* 'Miss Manila' accessions.

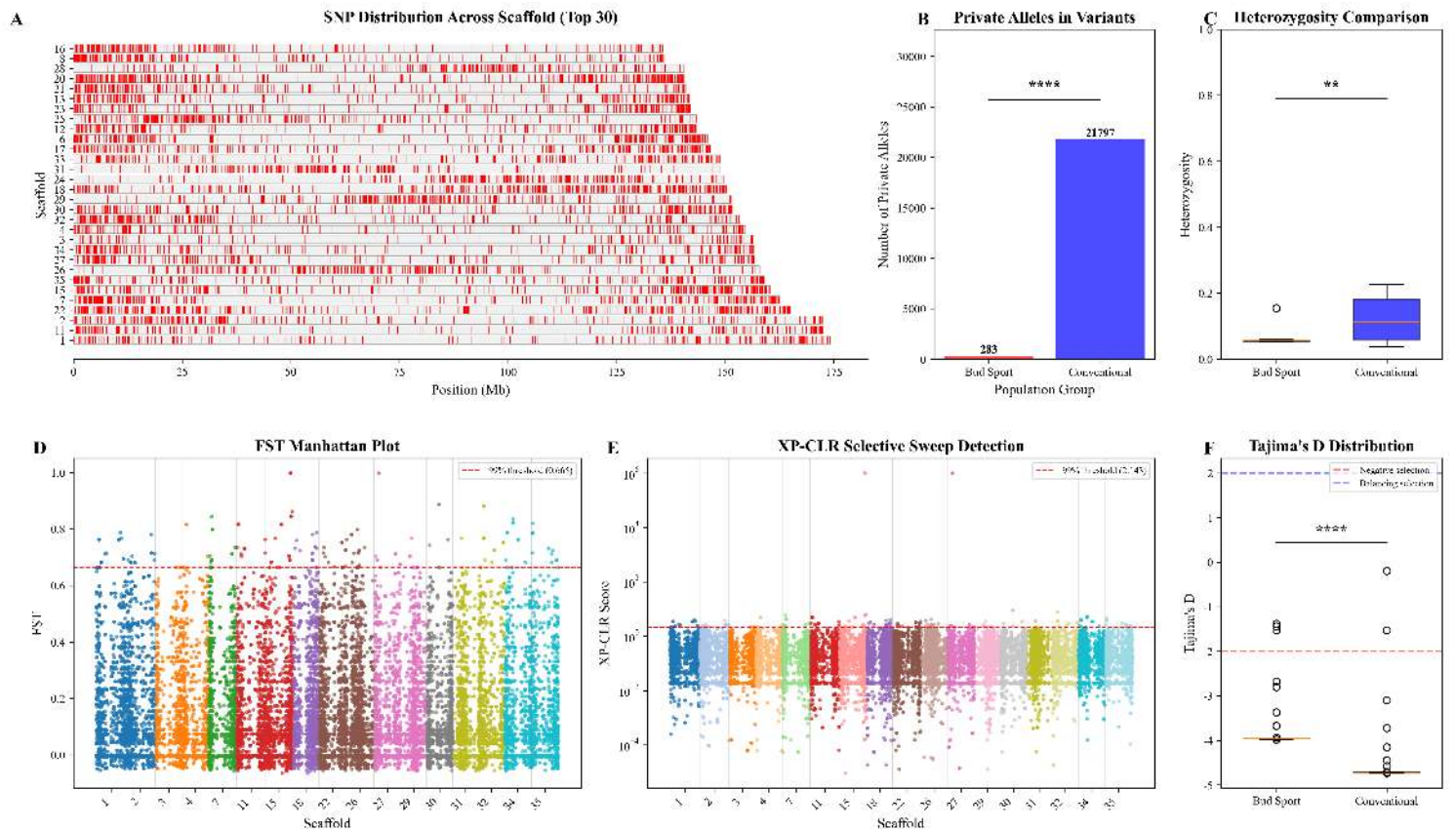


Figure 3

Genomic Landscape of Bud Sport Mutations and Signatures of Selection.

(A) Genome-wide distribution of single nucleotide polymorphisms (SNPs) across all scaffolds, illustrating the density and genomic context of genetic variations. (B) Comparison of private alleles between the bud sport and conventional populations. The bar chart shows a significantly lower proportion of private alleles in the bud sport group (****, $P < 0.0001$). (C) Comparison of heterozygosity levels between the bud sport and conventional groups. Bud sports exhibit a significant change in heterozygosity (**, $P < 0.01$), demonstrating the impact of mutations on genetic diversity. (D) Manhattan plot of population differentiation (F_{ST}) between bud sport and conventional populations. The dashed horizontal line indicates the genome-wide significance threshold. (E) Results of the Cross-Population Composite Likelihood Ratio (XP-CLR) test, identifying genomic regions that have undergone selective sweeps in the bud sport lineage. (F) Comparison of Tajima's D values in bud sport (red) and conventional (blue) populations. The comparison reveals a significantly different allele frequency spectrum between the two groups (****, $P < 0.0001$), reflecting distinct demographic or selective histories.

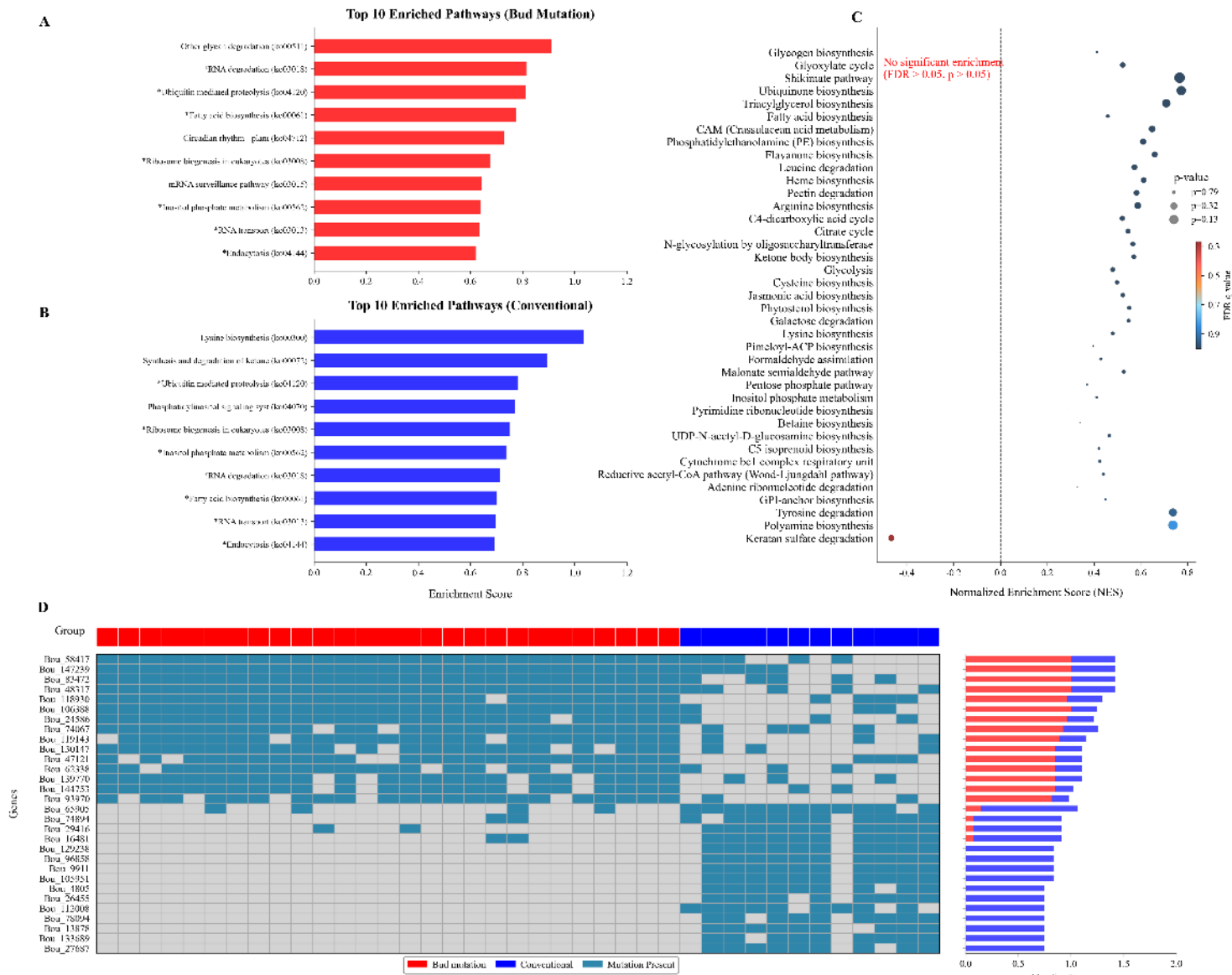


Figure 4

Comparative KEGG Enrichment and Mutational Landscape Between Bud Sport and Conventional Populations.

(A, B) KEGG pathway enrichment analysis of the top 10 significantly enriched pathways in the (A) bud sport and (B) conventional populations, respectively; pathways labeled with an asterisk (*) are common to both populations. (C) Gene Set Enrichment Analysis (GSEA) comparing pathway enrichment patterns between the two populations. The x-axis represents the Normalized Enrichment Score (NES), and the y-axis shows the signaling pathways. The size of each point corresponds to the statistical significance ($-\log_{10}(p\text{-value})$), and the color indicates the false discovery rate (FDR). (D) Distribution of genes with significant allele frequency differences between bud sport and conventional populations. The central heatmap (cyan) denotes the mutation status (presence/absence) of each gene in the respective groups. The adjacent bar plot illustrates the mutation frequency of each gene, with red and blue bars representing the bud sport and conventional groups, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.xlsx](#)