horticulturae

MDPI

Article

Comparative Pangenomics of the MATE Family Reveals Gene Content Variation and Identifies a Citrate-Transporter-Proximal Candidate Group in *Vaccinium*

Bin Li [1,2], Qingping Yi [1,2] and Ruiyi Fan [1,2,*]

[1] College of Food and Biology, Jingchu University of Technology, Jingmen 448000, China; bingli@jcut.edu.cn (B.L.); yiqingping@jcut.edu.cn (Q.Y.)
[2] Hubei Engineering Research Center for Specialty Flowers Biological Breeding, Jingmen 448000, China
* Correspondence: fanruiyi@outlook.com or fanruiyi@jcut.edu.cn

Highlights

What are the main findings?

- Across 34 blueberry and cranberry genomes, MATE gene content varied, while a group related to known citrate transporters was represented by 66 genome-supported loci.
- At least one locus from this group was detected in each assembly. Genome searches also showed that incomplete annotations explained some apparent gene absences.

What are the implications of the main findings?

- The candidate loci were largely dispersed across genomes, often retained similar neighbouring genes, and showed a conserved gene order across six *Vaccinium* species.
- Comparisons of gene sequences, upstream regions, and root expressions identified complementary candidates for functional studies.

Abstract

Plant multidrug and toxic compound extrusion (MATE) transporters participate in citrate transport, aluminium detoxification, and mineral homeostasis, yet their diversity and genomic organisation in *Vaccinium* remain poorly understood. In this study, we compared the MATE family across 34 blueberry and cranberry genomes and examined a candidate group related to experimentally characterised citrate transporters. The high-confidence family set contained 3479 proteins, and new MATE orthogroups continued to be detected as blueberry genomes were added. Genome searches showed that incomplete or non-standard annotations explained some of the apparent absences: four of the 12 tested orthogroups changed their prevalence category after gene-model reconstruction. In contrast, the candidate group was represented by 66 genome-supported loci, with at least one detected in each crop assembly. These loci were largely dispersed, often retained a similar neighbouring-gene order, and formed eight direct collinear links across six chromosome-scale *Vaccinium* genomes. A codon model estimated a global d_N/d_S of 0.876, which did not differ significantly from one. Comparisons of 2 kb upstream regions revealed sequence differences among the assemblies, while two public root transcriptomes showed increased expression assigned to *CUFF.34706* at pH 6.5 in 'Emerald' and 'Reka'. Interaction tests did not establish genotype-specific pH responses. Together, these results distinguish variable MATE gene content from a consistently detected candidate group and identify sequences for further studies of transport and regulation.

Check for updates

Academic Editor: Gunārs Lācis

Received: 3 September 2026
Revised: 11 October 2026
Accepted: 13 October 2026
Published: 16 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: *Vaccinium*; MATE transporter; pangenome; comparative genomics; gene content variation; MCScanX; chromosome-scale collinearity; root transcriptome; pH response

1. Introduction

Plant MATE transporters move diverse compounds across membranes, including organic acids, flavonoids, and xenobiotics [1,2]. Some members export citrate from root apices and reduce aluminium toxicity, whereas others transport citrate within the plant to support iron movement and mineral homeostasis [3–10]. These functions are particularly relevant to *Vaccinium* species, which are adapted to acidic soils and show impaired growth and nutrient acquisition as root-zone pH rises [11]. Understanding the diversity of MATE genes may therefore help identify candidates involved in balancing nutrients and responding to changing soil pH.

The diversity of MATE functions also makes candidate identification difficult. Family membership alone does not establish citrate transport, and related proteins may differ in their substrate or physiological role [1–10]. Moreover, a single reference genome cannot show whether a gene is widely retained or varies among cultivars. The available blueberry and cranberry pangenomes allow this question to be addressed across multiple genomes [12]. Studies in other crops have likewise revealed extensive variation in gene presence and annotated gene numbers [13–15]. Combining these resources with experimentally characterised MATE proteins provides a basis for examining both family-wide diversity and specific candidate groups.

Differences in transport-related traits can arise through changes in the gene number, genomic position, or expression. In maize, tandem triplication of ZmMATE1 is associated with increased expression and aluminium tolerance [16]. In barley, an upstream insertion redirects HvAACT1 expression to the root apex [17]. In rice, ART1 regulates OsFRDL4 and other aluminium-response genes [18–20]. These findings suggest that gene content and upstream sequence variation should be examined together when selecting MATE candidates in *Vaccinium*.

In this study, we investigated MATE diversity across 34 blueberry and cranberry genomes, with particular attention to a group related to known citrate-transport and metal-homeostasis proteins. We first examined family-wide gene content and assessed whether apparent absences could result from incomplete annotation. We then assessed the candidate group through phylogeny, genome searches, and a conserved gene order, extending the comparison to six chromosome-scale *Vaccinium* genomes. Finally, coding sequences, annotated upstream regions, and two public root transcriptomes were compared to identify candidates for functional studies. This approach addresses how a broadly variable transporter family can contain a group consistently detected across the sampled genomes.

2. Materials and Methods

2.1. Genome Resources and Annotation Preparation

The primary dataset comprised 34 public genomes from the Genome Database for *Vaccinium* (GDV): 23 blueberry and 11 cranberry assemblies. Thirty-two assemblies came from Yocca et al. [12], and the chromosome-level *V. corymbosum* 'Draper' v1.0 and *V. macrocarpon* 'Stevens' v1.0 were added as crop references [21,22]. Genome FASTA, GFF3 annotation, and protein FASTA files were downloaded for each assembly. Source URLs, file sizes, and SHA-256 checksums are recorded in Supplementary File S1; the 102 source files totalled 55,795,271,598 bytes. Assembly quality and the available information on phasing are summarised in Supplementary File S2. Blueberry and cranberry were labelled 4*x* and

2*x*, respectively, to describe the sampled materials, not to normalise gene counts. Cultivar names are used in the main text and figures, while B01–B23 and C01–C11 remain in scripts and supplementary tables as assembly identifiers (Supplementary File S1).

Protein, transcript, and gene identifiers were standardised while preserving their source genome and isoform relationships. The combined dataset contained 2,768,241 uniquely identified protein records. MATE proteins were linked to their transcript and gene coordinates in the GFF3 files, and isoforms from the same locus were not counted as separate loci. Gene, transcript, and protein counts agreed for 33 genomes. One missing source protein in 'Weymouth' could not be reconstructed as a complete open reading frame and was left unresolved.

2.2. MATE Identification and Functional Reference Proteins

The hidden Markov model for Pfam domain PF01554.24 (named MatE) was searched against all standardised proteins with PyHMMER 0.12.1 [23,24]. The recommended 24.9-bit gathering threshold was applied to both protein and domain scores. The broad catalogue included proteins with at least one MatE domain above this threshold; matches passing the protein-level threshold only were excluded. The high-confidence, or strict, set required complete genomic coordinates, a length of 350–650 amino acids, and exactly two qualifying PF01554 domains. A broader sensitivity set used the same coordinate and length criteria but required only one qualifying domain. All 22 experimentally characterised reference proteins passed the length criterion; 21 contained two qualifying domains, whereas AtEDS5 contained one. Reference proteins were retained irrespective of the strict filtering rule to represent known functional diversity (Supplementary File S3).

Three reference sets were used at different stages. The initial 12 proteins were used for the citrate-focused PHMMER search and the archived DeepTMHMM analyses. An expanded set of 16 proteins supported the gene-model review, and the final set of 22 proteins from nine experimentally characterised functional groups was used to classify blueberry's MATE phylogeny. The final set comprised SbMATE, HvAACT1, ZmMATE1, OsFRDL4, BoMATE, TaMATE1B, AtMATE/DTX42, AtFRD3, VuMATE1, VuMATE2, MtMATE66, MtMATE69, AtTT12, MtMATE1, MtMATE2, LaMATE, AtALF5, AtDTX1, AtEDS5, AtDTX50, NtJAT1, and NtMATE1. Their experimentally established functions guided candidate selection.

2.3. Phylogenetic Analysis and Candidate-Group Selection

Proteins most similar to the functional references and the lengths of 400–700 amino acids were included in the main phylogenetic analysis. This review range was slightly wider than the strict 350–650-amino-acid family filter, allowing moderately extended but alignable proteins to be examined. Severe fragments and unusually long proteins, including possible gene fusions, were reviewed separately. Sequences were aligned with FAMSA/pyFAMSA 0.7.0 [25], and columns with >80% gaps were removed. Maximum-likelihood trees were inferred with IQ-TREE 3.1.3 [26] and ModelFinder [27], using 1000 ultrafast-bootstrap replicates [28], 1000 SH-like approximate likelihood-ratio replicates [29], and the seed 20260820. Strong branch support required SH-aLRT $\geq$80 and ultrafast bootstrap (UFBoot) $\geq$95.

The gene-model-review tree used 388 sequences and 654 retained alignment columns under PMB+R5. For the 34-genome comparison, the longest proteins from each of the 133 high-confidence blueberry orthogroups were combined with the 22 references, aligned with FAMSA, and analysed under VT+F+R6 with the same support tests. All 204 strict proteins from the six orthogroups closest to citrate-related references were then analysed to test whether one representative adequately described each orthogroup. For each protein, its nearest reference group was compared with that of the longest representa-

tive, and the clustering of candidate proteins was examined. Phylogenetic placement, branch support, PF01554 domain structure, and predicted membrane topology defined the citrate-transporter-proximal MATE candidate group, referred to below as the focal group. Neighbouring-gene conservation was used separately to corroborate individual loci (Supplementary File S4).

Protein motifs were identified in the same 155 sequences used for the representative phylogeny with MEME Suite 5.5.9 [30]. The zero-or-one occurrence per sequence model was used to search for 15 motifs within 6–50 amino acids using the seed 20260829. Motif positions and site *p* values were extracted from the XML output. The 133 representative transcripts were matched to their source GFF3 files, retaining their gene, transcript, exon, CDS, and UTR coordinates. Exon–intron structures were displayed in transcriptional orientation and scaled within each transcript. Exact genomic coordinates and transcript spans are provided in Supplementary File S5.

2.4. Orthogroup Identification and Pangenome Analysis

An orthogroup is a set of genes inferred to descend from one ancestral gene. The strict and broader MATE sets were analysed separately for blueberry and cranberry with OrthoFinder 2.5.5 [31], its bundled DIAMOND 2.0.13 for all-versus-all protein comparisons [32], and bundled MCL 14-137 for Markov clustering. The broader set tested sensitivity to the family-filtering criteria. Within each crop, core orthogroups occurred in all genomes, soft-core orthogroups in at least 90% but not all genomes, shell orthogroups in at least three genomes but below the soft-core threshold, and cloud orthogroups in one or two genomes. Orthogroup identifiers were assigned independently to blueberry and cranberry.

Pangenome accumulation was calculated separately for the two crops under both family definitions. Genome order was randomised 1000 times with the seed 20260828. At each sample size, mean pan, core, and newly detected orthogroup counts were recorded with empirical 95% intervals. Heaps models, $P(N) = kN\gamma$, were fitted in log space for 2–23 blueberry and 2–11 cranberry genomes. To assess sensitivity to genome order, the accumulation and new-orthogroup decay exponents were also fitted to each of the 1000 randomised blueberry orders. The 2.5th and 97.5th percentiles defined the 95% permutation intervals. These estimates describe discovery within the sampled assemblies rather than unlimited species-wide diversity or uncertainty in a population sample.

To assess the effect of incomplete annotation, six shell and six cloud orthogroups were selected across the prevalence range in the 23-blueberry strict matrix. Up to three high-quality proteins from different assemblies were selected per orthogroup, giving 26 queries. Each was aligned to all 23 blueberry genomes with miniprot 0.18-r281 [33], using the same splice-aware settings as had been used for focal-locus reconstruction (--gff --aln --trans -j 1 -G 50000 -p 0.5 -N 20 --outn=20 --outs=0.70 --outc=0.80).

The baseline criteria required at least 90% query coverage and amino-acid identity; a predicted length of 350–650 amino acids; no frameshifts, internal stops, or unknown residues; and at least two PF01554 domains above the 24.9-bit threshold. The intended orthogroup also had to give the best DIAMOND match among the 3479 strict proteins, with a bit-score margin of at least 2% over the best alternative orthogroup. Models overlapping a strict MATE gene assigned to another orthogroup were rejected. Accepted models for apparent absences either lacked a public annotation or overlapped a PF01554 model excluded by the strict filter. Multiple models from one orthogroup in an assembly contributed a single presence call. Sensitivity was tested at 95% coverage/95% identity, 98% coverage/98% identity, and 100% coverage/99% identity, with all other criteria remaining unchanged (Supplementary File S6).

2.5. Membrane Topology and Gene-Model Reconstruction

Membrane topology was initially analysed before the final 34-genome dataset was assembled. PHMMER in PyHMMER 0.12.1 was run against the initial 12 references with protein and domain E-value thresholds of 1×10^{-5}. Of 6232 PF01554-positive proteins from the original 32 genomes, 535 had their best full-sequence score against a citrate-related reference rather than a non-citrate control. Twenty-three contained unknown X residues and were excluded, leaving 512 *Vaccinium* proteins with standard amino acids only. These proteins and the 12 references were analysed with DeepTMHMM 1.0.57 in three archived batches [34]. The three focal proteins subsequently added from 'Draper' and 'Stevens' were evaluated in a fourth run. The two 'Draper' proteins had 11 predicted membrane-spanning regions and the 'Stevens' protein had 12; none had a predicted signal peptide. The ten additional proteins in the final 22-reference set were not included in these topology analyses.

Loci with strong genomic similarity but fragmented or implausible public protein models were re-examined with miniprot 0.18-r281 [33]. Reconstructed models had to meet the coverage and sequence-quality criteria, contain two PF01554 domains, have a plausible DeepTMHMM topology, and cluster with the focal group. Five blueberry and ten cranberry loci passed these filters (Supplementary File S7). The new models were used for analysis without changing the released assembly or annotation files. An N-rich interval in 'Chandler' and two unknown residues in 'Harding' prevented acceptance of complete models for these loci.

2.6. Locus Counts and Neighbouring-Gene Conservation

After defining the focal group, individual loci were examined using genome sequence evidence, reconstruction of fragmented models, and neighbouring-gene conservation. Public annotations and PF01554 proteins guided the searches, while well-assembled intervals and consistent local gene orders distinguished supported loci from unresolved regions. Non-overlapping supported loci were counted for each assembly (Supplementary File S8). These are reported as assembly-supported locus counts and not read-depth-confirmed copy numbers, as matched reads and ploidy-aware normalisation were unavailable for the full dataset.

For each supported locus, up to ten annotated genes on either side were ordered relative to the focal gene in its transcriptional orientation. Their proteins were compared by all-versus-all PHMMER in PyHMMER 0.12.1, requiring protein and domain E-values $\leq 1 \times 10^{-5}$, $\geq$50% coverage of both proteins, and $\geq$25% aligned identity. Reciprocal best qualifying matches between genomes were grouped into conserved neighbouring-gene families. Strong support required at least three shared families on the same side of the focal gene and a conserved order of at least three families. Partial support required at least two same-side shared families in conserved order without meeting both strong-support criteria. Within-genome pairs were also classified by their physical positions. Pairs on different scaffolds or chromosomes with conserved local order were described as dispersed retention; they were not assigned to a specific duplication event. Sensitivity was tested for the 51 pairs from the original 32-genome set by varying the partial cutoff from one to four shared families and the strong cutoff from two to five, all while requiring a conserved order. The baseline criteria were also tested without this requirement (Supplementary File S9). The additional 'Draper' pair was evaluated separately with chromosome-scale MCScanX evidence (Supplementary File S4).

2.7. Cross-Species Comparison and Chromosome-Scale Collinearity

Four diploid species extended the comparison beyond cultivated blueberry and cranberry: *V. darrowii* (GCA_020921065.1) [35], *V. duclouxii* (SGLD20220023) [36], *V. vitis-idaea* 'Red Candy' (Figshare dataset 23926251) [37], and *V. bracteatum* (GWHBPBP00000000) [38]. Their genomes, annotations and proteomes were obtained from NCBI GenBank, PGDD/GDV, Figshare, and Genome Warehouse (Supplementary File S10). Together with 'Draper' and 'Stevens', they formed the six-genome chromosome-scale comparison. The scaffold-rich 'Stanley' and 'Early Black' genomes were retained for family screening but not for chromosome-scale analysis. The four additional species were not included in the blueberry or cranberry accumulation curves.

The same PF01554, coordinate, length, and domain-count criteria were applied to all species. Candidate proteins from the chromosome-level crop references and the existing cross-species representatives were combined with the 22 functional references in a 54-tip tree under VT+F+R6, with 1000 UFBoot and 1000 SH-aLRT replicates. Three proteins from 'Draper' and 'Stevens' met the focal-group criteria: two qualifying PF01554 domains, a plausible 10–14-pass topology, and supported placement near citrate-transport references. A second 'Stevens' protein close to MtMATE69 was retained as a related comparator.

For chromosome-scale collinearity, the longest released protein per gene was retained, giving 291,626 species-prefixed models from the six genomes. Genome-wide similarity was calculated with DIAMOND 2.1.9 [32] in sensitive mode, with E-value $\leq 1 \times 10^{-5}$ and a maximum of 25 targets per query. MCScanX [39] was run jointly for within- and between-species comparisons using a match score of 50, a gap penalty of -1, a minimum block size of five anchors, a maximum of 25 gaps, and an overlap window of five. The source commit was 0956cb8f900c152e2be8ba3196829e50ab454b94 (Supplementary File S11).

Focal loci were matched to the genome-wide MCScanX output, and blocks containing these genes were summarised by their species pair. Local displays included up to six genes on each side and the anchor pairs reported by MCScanX. An additional OrthoFinder analysis examined local gene order around loci outside the recovered genome-wide blocks.

2.8. Coding-Sequence Evolution

Codon analysis used the 21-sequence comparison set linked to the W8520 transcript model *CUFF.34706*. Coding sequences were extracted from six public GFF3 models and 15 accepted miniprot models in the correct strand and exon order. Each translated exactly to its corresponding protein and contained no ambiguous bases or internal stops. An unverified terminal valine, shared by ten cranberry protein exports, was omitted from the codon alignment rather than inferred from genomic sequence.

The 499-amino-acid alignment was converted to a 1497-nucleotide codon alignment. IQ-TREE 3.1.3 used GY+F3X4+G4, seed 20260821, and 1000 replicates for each support test. Pairwise d_N and d_S were estimated by the Nei–Gojobori method with Jukes–Cantor correction [40]. All 210 comparisons are provided in Supplementary File S12, although ratios were summarised only when at least 400 codons were comparable and $0.01 \leq d_S \leq 1.0$. Seventy-three comparisons had $d_S < 0.01$ and were excluded, leaving 137 estimates. These were treated as descriptive comparisons between homologous loci because one-to-one orthology was unresolved. PAML codeml 4.10.10 [41] was also fitted to the same alignment and IQ-TREE topology using the one-ratio (M0), fixed-$\omega = 1$, nearly neutral (M1a), positive-selection (M2a), beta (M7) and beta-and-ω (M8) models. The global ω estimate and its test against one came from M0 versus the fixed-$\omega = 1$ model.

2.9. Upstream-Sequence Comparisons

For each supported locus with known orientation, 2000 bp immediately upstream of the annotated gene start were extracted in the direction of transcription. We use the term annotated 2 kb upstream regions because transcription start sites were not experimentally mapped. Regions were screened for ambiguous bases and compared only after passing the sequence-quality and coding-model filters. Blueberry comparisons used 'Duke', 'Legacy', and 'Stanley', whereas the expression datasets used 'Emerald', *V. arboreum*, 'Reka', and 'Spartan'. Because the materials differed, upstream sequences and expressions were analysed separately rather than tested for their associations.

Eligible upstream regions were aligned separately for blueberry and cranberry with MAFFT 7.526 using L-INS-i [42]. 'Legacy' and 'Black Veil' served as the corresponding references. Direct comparison required at least 90% of known bases to align with the reference. Single-nucleotide differences and contiguous insertion/deletion events were counted after excluding ambiguous N positions. The 'Stevens' region was assessed by the same criteria. Exact sequence classes were defined from the full 2 kb sequences within each crop. Both strands were scanned for the ART1-like consensus GGNVS (N = A/C/G/T, V = A/C/G and S = C/G) and its GGYMS subset (Y = C/T and M = A/C). Each eligible region was shuffled 1000 times, preserving mononucleotide composition, with the seed 20261001. Empirical enrichment p values were calculated as (1 + shuffled counts at least as large as the observed count)/1001, followed by Bonferroni correction across the 17 regions separately for each motif. These matches annotate sequence features rather than validated regulatory sites (Supplementary File S13).

2.10. Short-Read Root Transcriptome Analysis

The short-read dataset, retained as PH01 in data files, contained 16 paired-end Illumina libraries from PRJNA353989/SRP093536. *V. arboreum* and *V. corymbosum* 'Emerald' were treated at pH 4.5 or 6.5, with four biological replicates per species–pH group [11]. Plants received eight weeks of hydroponic treatment without controlled aluminium addition. FastQC 0.12.1 was used for read quality assessment, and Cutadapt 5.2 for adapter and quality trimming [43]. Trimming retained 349,429,634 read pairs. This study reanalysed public reads and generated no new RNA-seq data.

Reads were aligned with Subjunc 2.1.1 [44] to the diploid W8520 May 2013 genome and its annotation of 63,840 genes and 70,581 transcript models. Across the 16 libraries, 219,895,076 fragments were mapped (62.93% overall; median 68%, range 30–75%). SRR5036124 ('Emerald', pH 6.5, replicate 2) had a 30% mapping rate and was flagged as a low-quality outlier (Supplementary File S14). featureCounts 2.1.1 [45] tests found heterogeneous read orientations, so primary counts used uniquely mapped paired fragments assigned to exons without a strand requirement. No other annotated gene, on either strand, overlapped *CUFF.34706* at scaffold01060:65910–77037 in the source GFF3; this interval check was separate from the MATE transcript 31-mer analysis. Each species was analysed separately with edgeR 4.10.1 in R 4.6.1 [46,47], using filterByExpr, trimmed mean of M-values (TMM) normalisation, robust dispersion estimation, and robust quasi-likelihood models. The contrast was pH 6.5 minus pH 4.5. A four-group model tested sensitivity and the species–pH interaction. Absolute expression was not compared between species, and unique reference alignments were not assumed to resolve individual copies in the tetraploid cultivar.

2.11. Long-Read Root Transcriptome Analysis

The long-read dataset, retained as PH02 in data files, comprised 12 Oxford Nanopore full-length cDNA runs from PRJNA1037260, with three biological replicates per cultivar–

pH group [48]. The associated publication identifies the deposited R and S materials as the high-pH-tolerant cultivar 'Reka' and the sensitive cultivar 'Spartan'. Both were grown at pH 4.5 and 6.5 without an imposed aluminium treatment. All FASTQ files matched the official European Nucleotide Archive sizes and MD5 checksums [49]. Dataset codes were retained for traceability, while the sequencing type and cultivar names were used in the narrative.

Reads were aligned to the W8520 transcriptome with Magic-BLAST 1.7.2 [50]. Assignment required a unique top-scoring gene, ≥85% identity, ≥20% query coverage, and ≥200 aligned nucleotides. Four-group edgeR models tested within-cultivar pH responses and the cultivar–pH interaction. Assignment specificity was examined by comparing *CUFF.34706* with 117 other W8520 MATE-linked genes using diagnostic 31-mers, alignment scores, identity, and coverage. All 605 reads aligning to *CUFF.34706* were competitively realigned with BLASTN 2.17.0+ [51] to the 21-sequence comparison set, *CUFF.34706*, and representatives of the other 117 genes. This tested whether the signal mapped to the focal group or another MATE branch. Read retention was also assessed at a query coverage of ≥80% and an identity of ≥90% or ≥95% while still requiring a unique top-scoring W8520 gene (Supplementary File S15). The resulting sparse counts were used to evaluate assignment retention; differential expression was not refitted at these stricter thresholds.

2.12. Integration of Genomic and Expression Evidence

The analyses followed the workflow in Figure S1 (Supplementary File S16). PF01554 domains established family membership while phylogeny, branch support and membrane topology identified the focal group. Genome searches, gene-model reconstruction, and local gene orders then assessed individual loci. Root expression and sequence comparisons were used to select candidates for functional studies and not to establish their transport substrates.

2.13. Statistical Thresholds and Reproducibility

Unless stated otherwise, multiple testing used the Benjamini–Hochberg correction [52], with significance defined as a false discovery rate (FDR) ≤ 0.05. Branch support, sequence-quality filters, and locus decisions were recorded in machine-readable tables. The archive also contains source URLs, accessions, file sizes, checksums, software versions, seeds, and session information. Repository access is given in the Data Availability Statement.

Analyses were repeated from archived commands, and manuscript values were checked against their corresponding machine-readable outputs.

ChatGPT (OpenAI, GPT-5.6) was used to assist with writing and debugging the bioinformatics analysis code and with language refinement of the manuscript. The authors have reviewed, verified, and revised the AI-assisted code and text and have checked the analytical outputs against the underlying data; they take full responsibility for the accuracy and integrity of the work.

3. Results

3.1. Standardised Screening Identifies the Vaccinium MATE Family

Uniform PF01554 screening of the 2,768,241 proteins identified 6549 proteins with at least one qualifying MatE domain. Each could be linked to its source transcript and gene coordinates. Broad-family counts ranged from 228 to 278 per blueberry genome and from 51 to 62 per cranberry genome, including 260 in 'Draper' and 57 in 'Stevens'. These annotation-based counts cannot be interpreted as a ploidy-only difference: assembly completeness, haplotype representation, and annotation pipelines may also contribute

(Supplementary File S2). A further 158 proteins passed only the protein-level threshold and were excluded, as no individual domain met the criterion.

Applying the coordinate, length, and domain filters produced 3479 high-confidence proteins: 3058 from blueberry and 421 from cranberry. The broader sensitivity set contained 4032 proteins. These standardised sets provided a common basis for comparing gene contents and phylogenetic relationships across genomes.

3.2. MATE Gene Content Varies, Although Some Apparent Absences Reflect Annotation

The strict MATE set formed 133 blueberry and 34 cranberry orthogroups. Blueberry contained 28 core, 12 soft-core, 41 shell, and 52 cloud groups, whereas cranberry contained 14 core, 11 soft-core, and eight shell groups, as well as one cloud group (Figure 1). New groups were still being detected as blueberry genomes were added; across 1000 random genome orders, the 23rd genome added a mean of 1.771 orthogroups. The Heaps curve was $P(N) = 54.11N^{0.285}$ (log-space $R^2 = 0.9994$). The 95% permutation intervals were 0.237–0.339 for γ and 0.378–0.958 for the new-group decay exponent α, whose mean-curve estimate was 0.708. Thus, discovery depended partly on the genome order, but this did not reach saturation in the sampled blueberry genomes (Figure S2). Cranberry was closer to saturation, with $\gamma = 0.019$ and $R^2 = 0.9869$. The broader family set showed the same contrast ($\gamma = 0.338$ for blueberry and 0.022 for cranberry). In the original 22-blueberry set, apparent orthogroup absences correlated positively with N_{50} ($\rho = 0.61$, $p = 0.003$) and negatively with a strict MATE gene number ($\rho = -0.85$). Assembly configuration was confounded with these associations, preventing attribution to continuity alone.

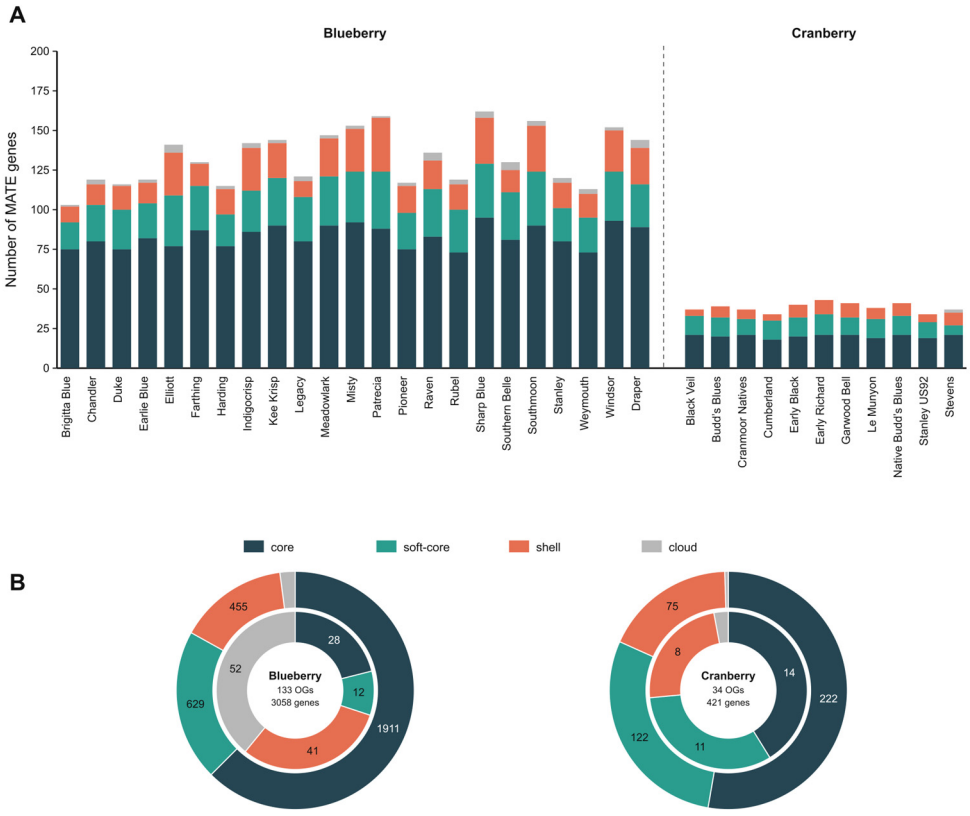


Figure 1. MATE pangenome composition across 34 *Vaccinium* assemblies. (**A**) Annotated strict-set genes assigned to core, soft-core, shell, and cloud orthogroups in each assembly. Groups were inferred separately for blueberry and cranberry; the dashed line separates the two crops. (**B**) Nested summaries showing orthogroup counts in the inner rings and annotated gene counts in the outer rings. Presence–absence variation describes orthogroup detection in assemblies, and annotated gene counts are not read-depth-confirmed copy numbers.

Genome searches showed that incomplete annotations explained some apparent absences. Of 76 orthogroup–assembly combinations scored as present, 68 were recovered (89.5%), providing an internal sensitivity check. Among 200 combinations scored as absent, 36 (18.0%) contained an accepted model at ≥90% coverage and identity; 23 (11.5%) remained supported at 100% coverage and ≥99% identity. Of the 64 models recovered at the baseline threshold, 63 overlapped public PF01554 annotations excluded by the strict filter. One OG0000070 model in 'Draper' occurred in an unannotated interval. Four of the 12 tested orthogroups changed category: OG0000036 and OG0000040 moved from shell to core, and OG0000085 and OG0000129 from cloud to shell. Eight retained their original categories across the tested thresholds (Figure S3; Supplementary File S6). The annotation-based matrix therefore overstated absence for some groups, but this did not explain all observed variations.

Phylogenetic analysis identified a smaller group within this variable family. The tree included 133 blueberry orthogroup representatives and 22 functional references; 98 out of 152 internal branches met SH-aLRT ≥80 and UFBoot ≥95, with median support values of 97.05 and 100. Six orthogroups lay closest to citrate-transport or metal-homeostasis references and included two core, one soft-core, one shell, and two cloud groups. All 204 proteins in these six groups were nearest to these reference classes, and 197 agreed with the reference class assigned to their longest representative. Four orthogroups were monophyletic whereas OG0000005 and OG0000017 were not, and all seven assignment differences occurred in OG0000017. Within these six groups, OG0000017 linked the narrower focal group to *CUFF.34706*. It contained 67 annotated blueberry genes in 22 out of 23 assemblies, but phylogenetic, domain, and locus reviews reduced these to 55 supported loci: 53 in the original 22 genomes and two in 'Draper' (Figure 2A; Figure S4; Supplementary File S4). The 21-sequence comparison set and 66-locus crop-wide catalogue describe different subsets and analysis stages of this focal group. In the orthogroup-representative tree, the OG0000017 representative was grouped with BoMATE, AtMATE/DTX42, and MtMATE66 (Figure 2A). This placement supports candidate prioritisation but does not establish substrate specificity.

Protein motifs and gene structures further characterised the six reference-related orthogroups. The 15 motifs were 15–50 amino acids long, had E-values no greater than 3.1×10^{-523}, and occurred in 19–142 sequences. Each of the six representatives carried three to five motifs. M12 and M13, both 50 amino acids long, occurred in all six representatives and in the same 13 functional references—12 citrate-related proteins and LaMATE—but not in the other 127 blueberry representatives. All 133 representative transcripts were mapped to their source annotations. Their CDS exon counts ranged from one to 16 (median seven), compared with 12–16 (median 15) in the six representatives (Figure S5; Supplementary File S5).

Most full-length proteins in the wider search set had predicted multiple membrane-spanning regions. The original batches produced valid predictions for 512 *Vaccinium* proteins and 12 references. The three focal proteins added from 'Draper' and 'Stevens' had 11, 11, and 12 predicted membrane-spanning regions without signal peptides, consistent with a full-length MATE architecture.

Gene-model reconstruction recovered focal proteins that had been fragmented or missed in public annotations. Alongside 47 unchanged public proteins, five blueberry and ten cranberry models were accepted. Each cranberry reconstruction was 496 amino acids in length, fully covered its selected query, lacked frameshifts, internal stops and unknown residues, and contained two PF01554 domains. All had 11 predicted membrane-spanning regions and no signal peptide. These analytical models were kept separate from the released annotations.

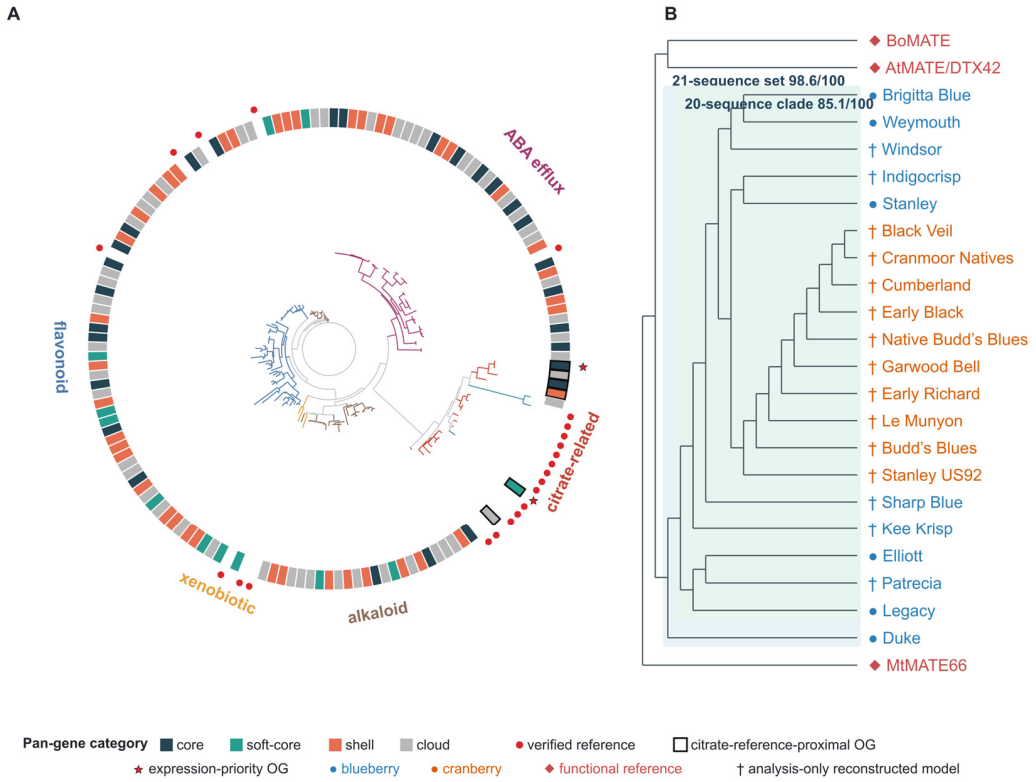


Figure 2. Phylogeny of blueberry MATE proteins and the focal candidate group. (**A**) Maximum-likelihood tree of 133 high-confidence orthogroup representatives and 22 experimentally characterised reference proteins. Branch colours show the nearest reference group, and the outer ring shows pangenome categories. Red circles indicate reference proteins, black outlines identify the six orthogroups nearest citrate-related references, and red stars identify groups containing the best genomic matches to the two root-expression transcripts. (**B**) Cladogram pruned from the gene-model-review tree to the 21 comparison sequences and three citrate-related references. Shading marks the 20-sequence clade (SH-aLRT/UFBoot = 85.1/100) and the broader set including 'Duke' (98.6/100); daggers identify reconstructed models. The 155-tip and 388-tip trees used VT+F+R6 and PMB+R5, respectively, with 1000 replicates per support test. Figure S5 shows motifs and representative exon–intron structures.

In the gene-model-review tree, the ten cranberry reconstructions clustered with five unchanged blueberry proteins and five repaired blueberry models in a 20-sequence clade (SH-aLRT/UFBoot = 85.1/100). Including 'Duke' gave the 21-sequence comparison set, supported at 98.6/100 (Figure 2B). 'Chandler' and 'Harding' remained incomplete because of an N-rich interval and two unknown amino acids, respectively. Accepted public or reconstructed models were available for 30 of the original 32 genomes; the other two were assessed from their genome sequences and local gene orders.

3.3. Chromosome-Level References Connect Crop and Cross-Species Comparisons

Adding the chromosome-level 'Draper' and 'Stevens' references expanded the crop dataset to 23 blueberry and 11 cranberry genomes and connected it to the cross-species analysis. 'Draper' contributed 144 high-confidence MATE proteins and 'Stevens' contributed 37. Blueberry OG0000017 remained in the soft-core group, with 67 annotated genes in 22 of the 23 genomes, whereas the corresponding cranberry group was detected in all 11 genomes.

Phylogeny and membrane topology distinguished three focal proteins from a related comparator in these references. Two 'Draper' proteins on VaccDscaff5 and VaccDscaff10 and one 'Stevens' protein on chromosome 9 were nearest to MtMATE66. All three contained two PF01554 domains and 11–12 predicted membrane-spanning regions. A second 'Stevens'

protein on chromosome 7 was nearest to MtMATE69 and was retained as a related metal-homeostasis comparator.

MCScanX placed 216,391 of the 291,626 genes (74.20%) in collinear blocks across the six genomes. It identified 21,524 blocks—16,481 between species and 5043 within species—and 778,130 anchor pairs. Eight of the 11 focal proteins occurred in at least one block. Together, they occurred in 41 blocks and formed eight direct links between focal genes (Figure 3; Supplementary Files S4 and S11). Conserved gene order thus connected candidate loci in cultivated blueberry and cranberry with related regions in the other *Vaccinium* species.

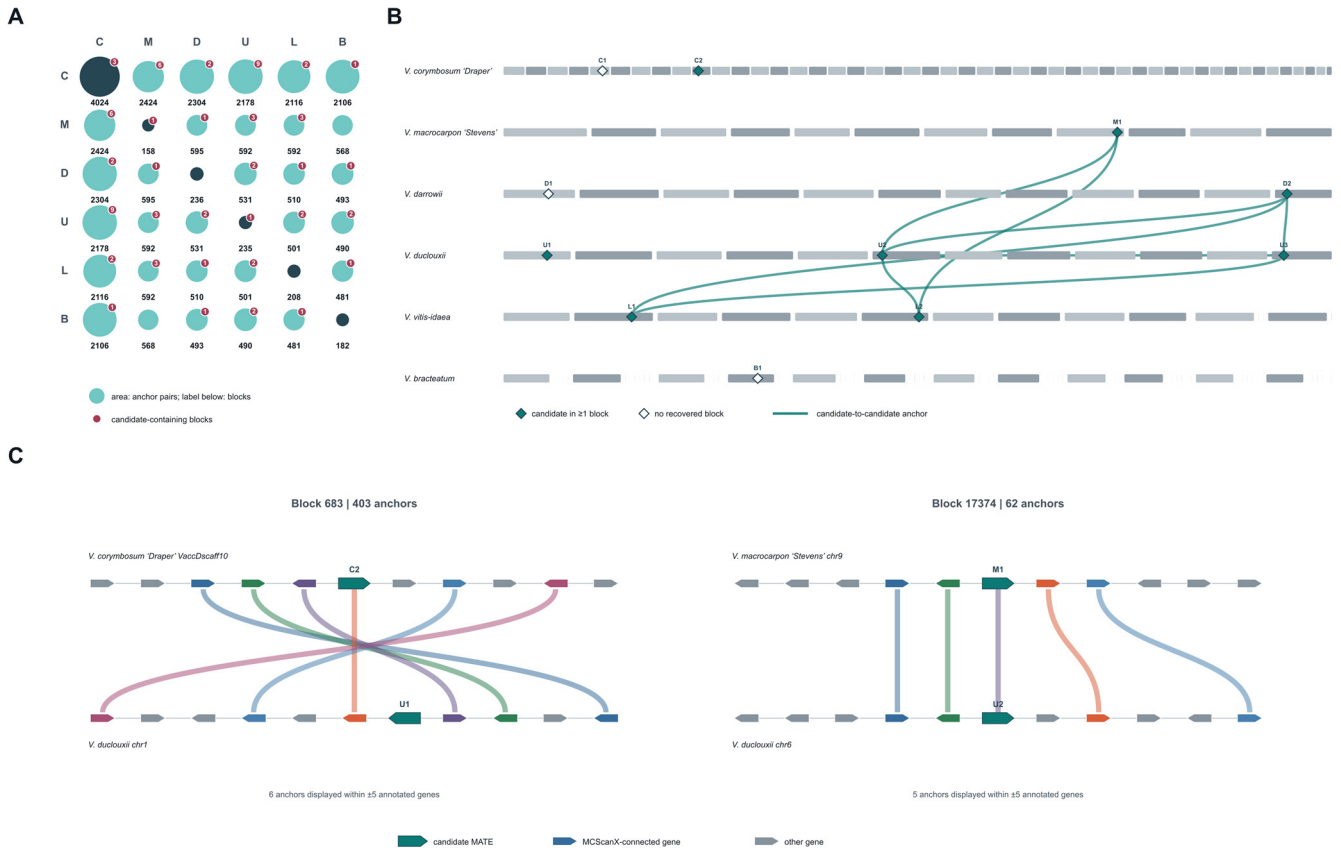


Figure 3. Chromosome-scale collinearity across two crop references and four additional *Vaccinium* species. (**A**) Pairwise MCScanX summaries. Dark circles show within-species comparisons and turquoise circles show between-species comparisons. Circle area shows anchor-pair number, values below circles provide block counts, and red satellite values provide blocks containing focal genes. (**B**) Chromosome positions of 11 focal proteins. Filled diamonds identify proteins in MCScanX blocks, open diamonds identify those outside recovered blocks, and teal curves show direct links between focal genes. (**C**) The strongest focal-gene-associated block linking each crop reference to *V. duclouxii*. Block 17374 links the 'Stevens' chr9 focal gene to the *V. duclouxii* chr6 focal gene. In block 683, the 'Draper' VaccDscaff10 focal gene is linked to a neighbouring MATE-like model alongside the *V. duclouxii* chr1 focal locus. Teal arrows identify focal genes; other matching colours identify connected anchors, and grey arrows indicate other neighbouring genes. Horizontal spacing represents gene order. Across the six genomes, eight of the 11 focal proteins occurred in blocks and formed eight direct links. C, *V. corymbosum* 'Draper'; M, *V. macrocarpon* 'Stevens'; D, *V. darrowii*; U, *V. duclouxii*; L, *V. vitis-idaea*; B, *V. bracteatum*.

3.4. The Focal Group Is Detected in Every Crop Genome

At least one focal-group locus was detected in each of the 34 crop genomes, including those with incomplete public models. Annotation reviews, genome searches, reconstruction, and neighbouring-gene analyses identified 63 loci in the original dataset; 47 public models, 15 reconstructed models, and the assembly-gap-supported 'Chandler' locus gave 62 usable

proteins. Two intact 'Draper' loci and one intact 'Stevens' locus increased the total to 66 assembly-supported loci (Table 1; Figure 4). Sixty-five were classified as intact or probable, while 'Chandler' remained limited by an assembly gap.



Figure 4. Focal-group retention and local gene order. (**A**) Assembly-supported locus counts and intact/probable model counts in 23 blueberry and 11 cranberry genomes. Cultivar names label the assemblies; 'Chandler' has an assembly-gap-limited locus. (**B**) Gene order around representative four-locus ('Duke'), three-locus ('Brigitta Blue'), and two-locus ('Legacy') assemblies. Focal genes are shown in teal; other matching colours and bands identify shared neighbouring-gene families, while grey genes lack a displayed family match. Colours distinguish families and do not indicate functional classes. Horizontal spacing represents gene order. (**C**) Fifty-two within-genome pairs; 51 strongly and one partly supported pair occurred on different scaffolds or chromosomes, with no tandem or proximal pairs. The two 'Draper' loci occur on separate pseudomolecules in a 374-anchor MCScanX block. Figure S7 presents the original 63-locus comparison, and Supplementary File S4 documents the three loci from the added crop references.

The 23 blueberry assemblies contributed 55 focal loci: seven assemblies contained one, four contained two, eight contained three, and four contained four. Each of the 11 cranberry assemblies contained one. These distributions were summarised separately since tetraploid blueberry and diploid cranberry differ in both ploidy and assembly properties. Within the original 22-blueberry set, the locus count was not significantly correlated with N_{50}, the annotation gene number, or a strict MATE family size (all $|\rho| \leq 0.11$, $p \geq 0.62$; Supplementary File S2). The source resource used reference-guided assemblies informed by phased

blueberry references [12], but independent cultivar-specific phasing was not documented for these 22 assemblies. A phased-versus-unphased comparison was therefore not possible. Northern and southern highbush assemblies also differed substantially in continuity (mean N_{50} 37.0 versus 1.4 Mb), so a comparison of their occupancy was evaluated under both family definitions as a sensitivity analysis (Figure S6). BUSCO [53] completeness with eudicots_odb10 ranged from 88.8% to 96.0% in blueberry and from 79.5% to 88.4% in cranberry, showing that the latter assemblies were generally less complete.

Table 1. Genome-supported focal loci and their distribution across assemblies.

Metric	Result
Genome assemblies	34 (23 blueberry; 11 cranberry)
Assembly-supported loci in the focal lineage	66
Intact or probable loci	65
Assemblies without a supported locus	0
Genomes with at least 1/2/3/4 loci	34/16/12/4
Genomes with exactly 1/2/3/4 loci	18/4/8/4
Locus pairs within the same genome	52
Cross-scaffold pairs	52
Pairs with strongly/partially conserved gene order	51/1
Tandem or proximal pairs	0

Note: Exact-count categories summarise all 34 crop assemblies; the 18 single-locus assemblies comprise seven blueberry and all 11 cranberry assemblies.

The focal loci were dispersed and generally retained similar neighbouring-gene order. The 16 genomes with multiple loci produced 52 pairs on different scaffolds or chromosomes: 51 from the original 32 genomes and one from 'Draper'. Fifty-one pairs had strong support, and one 'Patrecia' pair had partial support; none were tandem or proximal. In the original set, 50 pairs had strong baseline support, and one had partial support. At least 48 of the 51 retained strong support under every tested threshold (Supplementary File S9). The two 'Draper' loci were assessed separately and lay on different pseudomolecules within a 374-anchor MCScanX block. The direct anchor joined the VaccDscaff10 focal model to an adjacent MATE-like model beside the VaccDscaff5 focal locus. The homoeologous-group assignment of these released scaffold identifiers was not independently verified; separate pseudomolecules therefore do not exclude retention on homologous chromosome copies. Figure S7 shows the complete original 63-locus comparison, and Supplementary File S4 provides the extension.

3.5. Codon Analysis Does Not Detect Significant Selection Across the Focal Group

The 137 stability-filtered pairwise comparisons gave a median Nei–Gojobori d_N/d_S of 0.4407, and the 123 within-clade comparisons among them gave 0.4123 (Figure 5A; Supplementary File S12). Near-identical cranberry sequences lowered this summary, and another 73 comparisons were excluded because $d_S < 0.01$. The tree-based codon model gave a global ω of 0.876, which did not differ significantly from one (M0 versus fixed $\omega = 1$: $2\Delta\ln L = 0.34$, $p = 0.56$). Site models M2a/M8 nominally identified a contiguous block of positively selected sites, but ω_2 reached the model ceiling, and the pattern reflected fixed differences between blueberry and cranberry rather than recurrent substitutions. Taken together, these analyses did not provide robust evidence for purifying selection or recurrent positive selection in the comparison set.

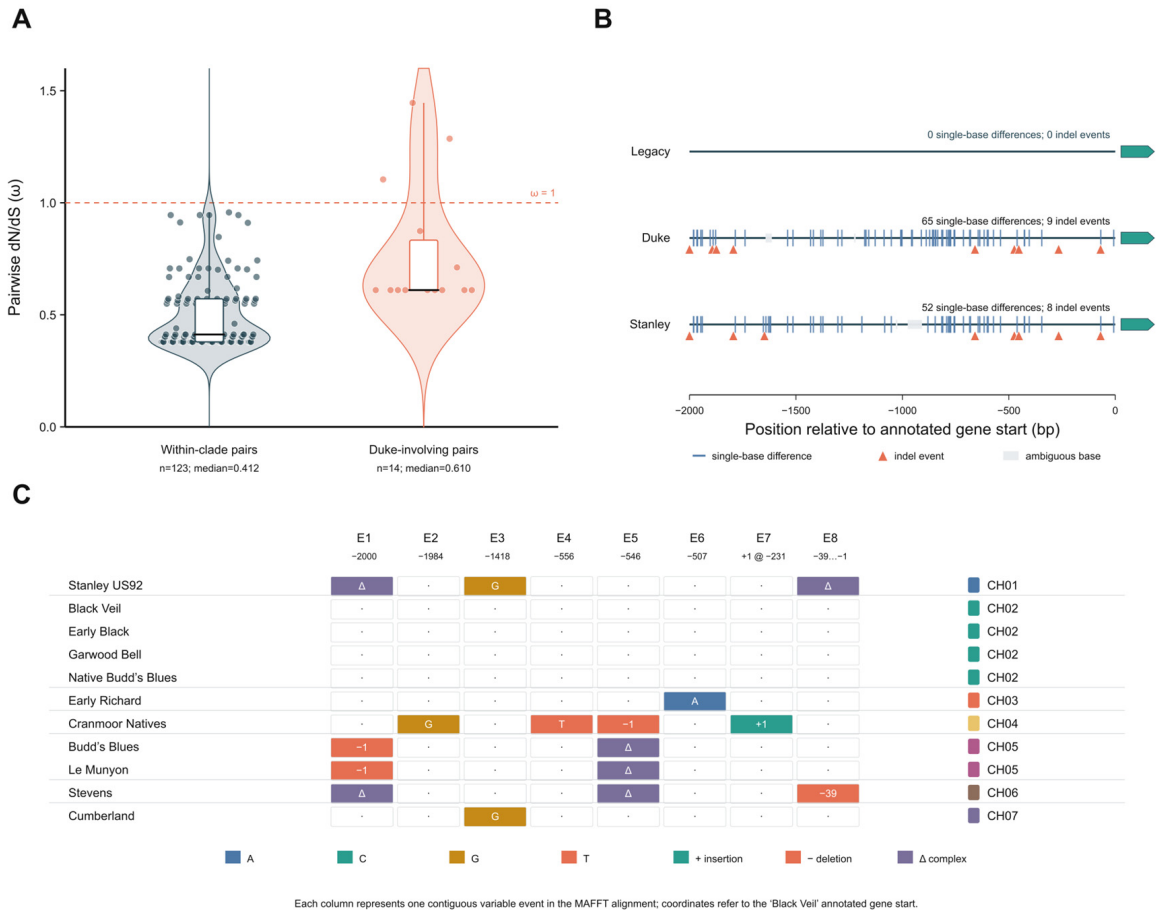


Figure 5. Coding-sequence comparisons and upstream variations in the focal group. (**A**) Pairwise Nei–Gojobori d_N/d_S for 123 within-clade comparisons and 14 involving 'Duke'. The dashed line marks $\omega = 1$. All within-clade estimates and 11 'Duke'-involving estimates were below one; these are descriptive comparisons among non-independent homologues. The codon model gave global $\omega = 0.876$, not significantly different from one ($p = 0.56$). (**B**) Differences in the three comparable blueberry 2 kb upstream regions, mapped relative to the annotated 'Legacy' gene start. Blue ticks indicate single-nucleotide differences, orange triangles indicate contiguous indels and pale segments indicate ambiguous bases. (**C**) Eight variable events in the 11 cranberry upstream regions, ordered by seven exact sequence classes. Colours show substitution or indel states relative to 'Black Veil'; dots show the reference state. Supplementary File S13 contains the complete matrix of 935 assembly-by-variable-column entries, and Figure S8 shows comparability for all six blueberry regions.

3.6. Upstream Regions Differ Between Assemblies

Six blueberry loci passed the coding-model filters, but only 'Duke', 'Legacy', and 'Stanley' met the ≥0.90 known-base comparability threshold for their upstream regions (Figure S8). Relative to 'Legacy', 'Duke' differed at 65 known bases and had nine insertion/deletion events, whereas 'Stanley' differed at 52 bases and had eight events (Figure 5B). These provide sequence contrasts for later regulatory studies. They were not tested against root expression because the expression datasets used different cultivars.

All 11 cranberry upstream regions passed the comparability threshold against 'Black Veil' (fractions 0.9805–1.0000) and formed seven exact sequence classes. 'Stevens' had no single-nucleotide differences across 1961 comparable bases but differed through three insertion/deletion events spanning 78 alignment columns. Variation across the 11 assemblies was summarised as eight contiguous events (Figure 5C; Supplementary File S13). Neither ART1-like short motif showed consistent enrichment across regions after composition-preserving background comparison. GGNVS was locally enriched in 'Sharp Blue' after

correction across 17 regions (adjusted $p = 0.034$) but not in the others. The sequence classes and motif matches describe assembly-level variation; they are not population haplotypes or validated regulatory differences.

3.7. CUFF.34706-Linked Root Expression Increases at Higher pH in Two Datasets

The short-read dataset showed a broader root pH response in 'Emerald' than in *V. arboreum*. At FDR ≤ 0.05 and $|\log2 \text{ fold change}| \geq 1$, 4411 genes responded in 'Emerald', compared with 380 in *V. arboreum*, consistent with the original study [11]. Of 349,429,634 trimmed read pairs, 219,895,076 mapped to W8520. After filtering, 32,304 and 23,172 genes were tested in the two materials.

Expression assigned to *CUFF.34706* increased significantly in 'Emerald' at pH 6.5. Mean TMM-normalised counts per million (CPM) rose from 1.942 to 6.673 (log2 fold change = 1.725, FDR = 0.0182), whereas the change in *V. arboreum* was not significant (log2 fold change = 0.692, FDR = 0.504). The species–pH interaction was also not significant (FDR = 0.339). *CUFF.34706* is the diploid W8520 transcript model used to assign expression to the previously defined focal group, rather than to a resolved copy in 'Emerald'. Its low mean abundance (TMM CPM 1.9–6.7) should be considered when interpreting the fold change.

The long-read dataset independently resolved cultivar and pH effects. All 12 files passed the size and MD5 checks. Replicates formed four tight clusters in multidimensional scaling (MDS), with cultivar and pH separated mainly along different dimensions. Magic-BLAST detected 31,895 W8520 genes, and 12,876 passed edgeR filtering. The common dispersion was 0.005174, corresponding to a biological coefficient of variation of 0.07193. The focal-gene expression contrasts in both datasets are summarised in Table 2.

Table 2. *CUFF.34706*-linked root expression responses in the short- and long-read datasets.

Data Type and Material	Replicates per pH	Mean TMM CPM, pH 4.5	Mean TMM CPM, pH 6.5	log2 Fold Change	FDR
Illumina short-read, *V. arboreum*	4	2.577	4.210	0.692	0.504
Illumina short-read, *V. corymbosum* 'Emerald'	4	1.942	6.673	1.725	0.0182
Nanopore long-read, 'Reka'	3	18.89	39.11	1.119	0.00123
Nanopore long-read, 'Spartan'	3	33.43	41.02	0.291	0.329

CUFF.34706-linked expression also increased at pH 6.5 in 'Reka', from 18.89 to 39.11 mean TMM CPM (log2 fold change = 1.119, FDR = 0.00123). 'Spartan' showed a smaller, non-significant change (log2 fold change = 0.291, FDR = 0.329), and the cultivar–pH interaction was not significant (FDR = 0.102). All 605 supporting reads had *CUFF.34706* as their unique best-scoring W8520 gene. Competitive mapping assigned all reads to the focal group rather than the other 117 MATE-linked genes; *CUFF.34706* remained the best overall match for 600 reads.

Thus, both datasets showed increased *CUFF.34706*-linked expression at pH 6.5 within 'Emerald' and 'Reka' (Figure 6). The non-significant interaction tests, however, did not establish that the pH response differed between species or cultivars.

3.8. Complementary Candidates Are Identified for Functional Studies

Combining phylogeny, gene models, local gene orders, upstream sequences, and root expressions identified four complementary representatives. 'Patrecia' represents the supported 20-sequence clade and is a priority for clone resequencing. 'Legacy' provides an intact public model and the blueberry upstream reference. 'Stanley' provides a high-quality upstream contrast with 'Legacy', while 'Duke' provides a phylogenetic contrast outside the 20-sequence clade.

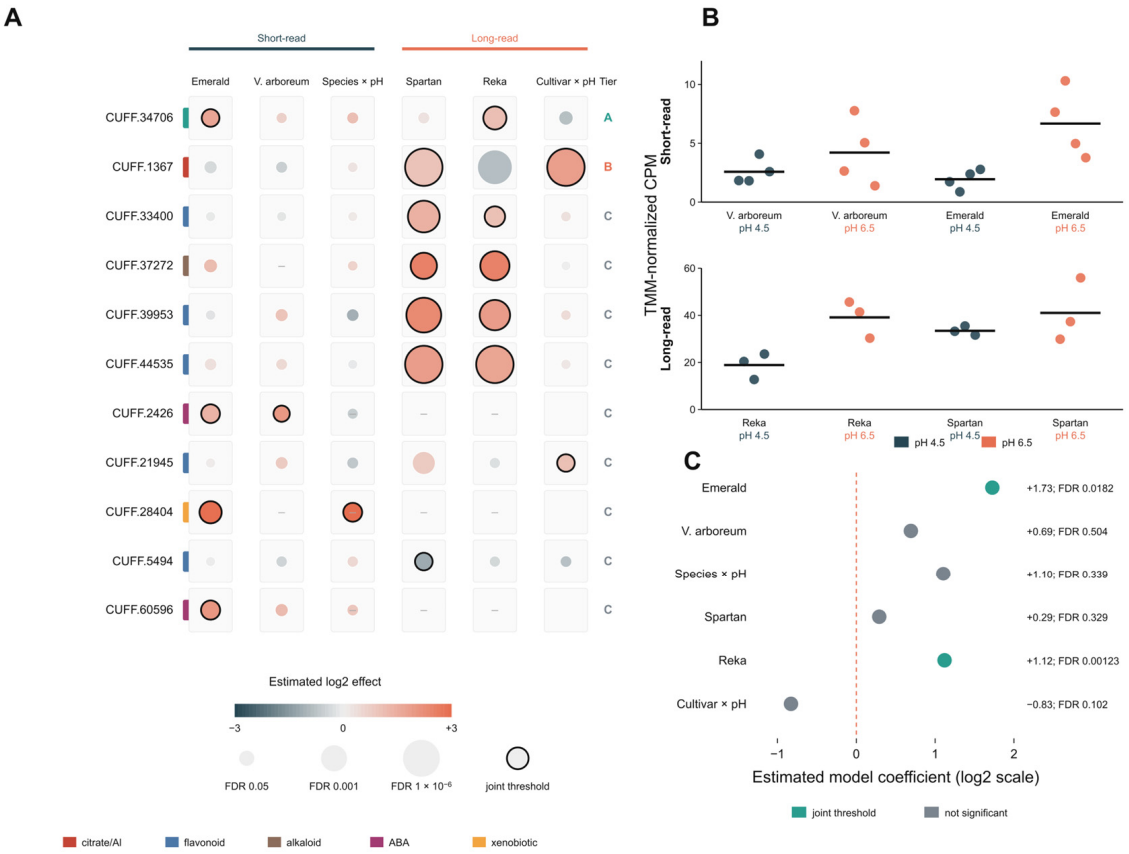


Figure 6. Root expression responses to pH. (**A**) Expression matrix for 11 MATE-associated W8520 genes in priority tiers A–C. The short-read block shows 'Emerald', *V. arboreum*, and the species–pH interaction; the long-read block shows 'Spartan', 'Reka' and the cultivar–pH interaction. Circle colour represents the estimated log2 coefficient: the pH 6.5-versus-4.5 contrast for individual materials, or the difference between their pH responses for interactions. Circle size represents −log10 FDR; black outlines indicate FDR $\leq$ 0.05 and an absolute coefficient $\geq$1. The adjacent strip shows the similarity-based reference group, and *CUFF.34706* is highlighted in teal. (**B**) TMM-normalised expression assigned to *CUFF.34706* for each sample; horizontal lines show group means. (**C**) *CUFF.34706* coefficients and FDR values. Expression increased significantly in 'Emerald' and 'Reka', while the species–pH (FDR = 0.339) and cultivar–pH (FDR = 0.102) interactions were not significant. All 605 supporting long reads mapped competitively to the focal group.

These representatives address different questions rather than carrying the same functional prediction. 'Legacy' and 'Patrecia' can be used to examine proteins from the supported clade, 'Duke' to test the more distant comparator, and 'Stanley' to compare upstream sequences. The *CUFF.34706*-linked expression signal adds a root pH context for the group, thereby guiding later transport, localisation, and regulatory studies.

4. Discussion

4.1. A Consistently Detected Candidate Group Within a Variable MATE Family

The main finding is a contrast between variable family-wide MATE gene content and consistent detection of the focal group. New blueberry orthogroups continued to appear as genomes were added, yet at least one focal locus was found in every blueberry and cranberry assembly. Pangenome studies have shown the value of examining multiple genomes rather than relying on one reference [12–15]. Here, this approach identifies a candidate group that remains detectable despite variation in the wider family. Interpretation of the between-crop count difference must also consider ploidy, haplotype representation, and annotation quality.

Genome searches also show why annotation-based absence needs careful interpretation. Models were recovered from 18.0% of the tested orthogroup–assembly combinations originally scored as absent, changing the category of four of the 12 orthogroups. Nearly all the recovered models that overlapped existing PF01554 annotations failed the strict filter, implicating fragmented or non-standard models rather than wholly unannotated sequence. Nevertheless, eight orthogroups retained their categories. Annotation differences therefore explain part, but not all, of the observed variation; the matrix should not be read directly as a history of gene gain and loss.

The cross-species comparison extended the genomic evidence beyond cultivated blueberry and cranberry. Related focal proteins occurred in *V. corymbosum*, *V. macrocarpon*, *V. darrowii*, *V. duclouxii*, *V. vitis-idaea* and *V. bracteatum*. Eight out of 11 proteins occurred in collinear blocks and formed eight direct links, connecting the phylogenetic pattern with conserved chromosome regions. Analysis of all 204 proteins in the six reference-related orthogroups also broadly supported the representative-based classification. Together, these results strengthen candidate selection across *Vaccinium* without assigning transport function from sequence similarity alone.

Family-wide discovery and focal-group retention describe different aspects of the data. The first concerns variation in annotated gene content; the second combines sequence placement with evidence from the genome around each locus. Additional genomes may reveal further family diversity, while matched genomic reads could refine locus counts. Their contrast does not by itself demonstrate different selection regimes.

4.2. Dispersed Loci Retain Similar Genomic Surroundings

The focal group was not organised into local tandem clusters. All 52 within-genome pairs occurred on different scaffolds or chromosomes, with strong neighbouring-gene support for 51 and partial support for one. The two ‘Draper’ loci also lay on separate pseudomolecules within a 374-anchor MCScanX block. Local gene order and chromosome-scale collinearity therefore support retention in related genomic regions, but separation across pseudomolecules alone does not distinguish chromosome-copy retention from segmental duplication.

This differs from the tandem triplication of ZmMATE1 associated with increased gene dosage and aluminium tolerance in maize [16]. In *Vaccinium*, the data instead show dispersed loci with conserved neighbouring genes. Such regions may reflect polyploid or segmental duplication, but assigning individual loci to these events will require gene-tree/species-tree reconciliation, resolved haplotypes, and matched read depths.

4.3. Upstream Sequence Differences Provide Candidates for Regulatory Studies

Coding-sequence analysis did not detect robust evidence of selection in the 21-sequence set. The low pairwise d_N/d_S summary was influenced by near-identical cranberry sequences, whereas the codon-model estimate of 0.876 did not differ significantly from one ($2\Delta\ln L = 0.34$, $p = 0.56$). The site-model signal coincided with fixed blueberry–cranberry differences and an $\omega 2$ estimate at the model ceiling, limiting its interpretation as recurrent positive selection. These results neither establish a strong coding constraint nor demonstrate that regulatory change is the main source of diversification.

The upstream regions nevertheless provide direct sequence contrasts worth examining. Compared with ‘Legacy’, ‘Duke’ and ‘Stanley’ differed by 52–65 known bases and eight to nine indel events. The 11 cranberry regions formed seven sequence classes, and ‘Stevens’ differed from ‘Black Veil’ mainly through three indels. Upstream variation can alter MATE expression, as shown for HvAACT1 in barley [17], but the present comparisons identify sequence differences rather than their regulatory effects. Transcription start sites

were not mapped, and the sequenced assemblies differed from the expression materials. Genotype-matched upstream and expression studies are therefore the next step for testing these contrasts.

4.4. Root Expression Links the Candidate Group to Higher pH Conditions

The two root datasets provide a complementary biological context: *CUFF.34706*-linked expression increased at pH 6.5 in 'Emerald' and 'Reka'. Although interaction tests did not establish species- or cultivar-specific responses, the repeated increase across two sequencing platforms makes the group a useful target for follow-up work. Diagnostic 31-mers and competitive mapping support assignment to this group, but the diploid reference does not resolve its individual copies in tetraploid cultivars. Because near-neutral pH impairs nutrient acquisition in acid-adapted blueberry [11], the expression pattern is compatible with a role in citrate distribution or mineral homeostasis.

Citrate export is one possible explanation, although known MATE proteins also mediate internal citrate distribution and iron translocation [3,4]. The pH treatments did not include aluminium, so their expression responses should not be interpreted as aluminium tolerance. 'Legacy', 'Patrecia', 'Stanley', and 'Duke' offer complementary sequences for distinguishing transport properties from coding and upstream differences.

The main remaining questions are substrate identity, cellular location and the link between upstream variation and expression. Clone resequencing, localisation assays, and direct transport measurements can establish the properties of the selected proteins. Genotype-matched reporter and expression studies under controlled pH or aluminium treatments can then determine which sequence differences have biological effects. The present work supplies a defined candidate set and comparative framework for these studies rather than a validated breeding intervention.

5. Conclusions

The *Vaccinium* MATE family combines variable gene content with a candidate group consistently detected across 34 blueberry and cranberry assemblies. Genome searches showed that incomplete annotation explains some apparent absences, while local gene order and chromosome-scale collinearity supported the focal loci. Codon analysis did not detect a significant departure from neutral expectation, but upstream comparisons revealed assembly-level sequence differences. Two independent root datasets further associated *CUFF.34706*-linked expression with a higher pH in 'Emerald' and 'Reka'. These complementary findings identify 'Legacy', 'Patrecia', 'Stanley', and 'Duke' for further investigation of transport and regulation. Direct functional assays and genotype-matched expression studies will determine the physiological roles of these candidates.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/horticulturae12101284/s1, Supplementary File S1: Genome resources and checksums; Supplementary File S2: Assembly quality, phasing information, and focal-locus counts; Supplementary File S3: PF01554 screening and functional references; Supplementary File S4: Integrated family, pangenome, and phylogenetic analyses; Supplementary File S5: MEME motifs and representative gene structures; Supplementary File S6: Matched genome searches for 12 shell/cloud orthogroups; Supplementary File S7: Public and reconstructed candidate models; Supplementary File S8: Locus support and neighbouring-gene conservation; Supplementary File S9: Assembly and local gene-order sensitivity analyses and pangenome permutation intervals; Supplementary File S10: Comparisons across additional *Vaccinium* species; Supplementary File S11: Chromosome-scale MCScanX results; Supplementary File S12: Coding-sequence validation and d_N/d_S analyses; Supplementary File S13: Upstream sequences and ART1-like motif analysis; Supplementary File S14: Short-read root transcriptome quality control and differential expression; Supplementary File S15:

Long-read expression and mapping analyses; Supplementary File S16: Figures S1–S8; Figure S1. Analytical workflow. The pipeline proceeds from the input data (34 crop genomes and four additional *Vaccinium* references) through family screening, orthogroup inference, phylogeny, focal-locus audit and the four evidence lines, to the experimental representatives prioritized for functional testing. Figure S2. MATE pan-orthogroup accumulation in blueberry and cranberry across 1000 genome-order permutations. Mean pan-orthogroup numbers and empirical 95% intervals are shown as the number of sampled blueberry or cranberry genomes increased; dashed curves show the fitted Heaps models. The 23rd blueberry genome added a mean of 1.771 orthogroups, and the blueberry fit, $P(N) = 54.11N^{0.285}$ (log-space $R^2 = 0.9994$), indicated continued discovery within the sampled genomes. The cranberry exponent was 0.019 ($R^2 = 0.9869$) and is presented as descriptive evidence that discovery was much closer to saturation in the sampled cranberry genomes. These curves describe the available assemblies and are not species-wide extrapolations. Across the same 1000 randomized orders of the 23-genome blueberry matrix, the 95% permutation interval was 0.237–0.339 for γ; the mean-curve estimate of the new-orthogroup decay exponent α was 0.708 (95% permutation interval 0.378–0.958). These intervals measure sensitivity to genome order and do not constitute a formal open-versus-closed pangenome test. Figure S3. Matched genome-based assessment of annotation-derived shell and cloud MATE occupancy in 23 blueberry assemblies. (A) Annotated prevalence and genome-supported prevalence lower bounds for six shell and six cloud orthogroups. Base-gate lower bounds include accepted models with at least 90% query coverage and 90% amino-acid identity; strict-endpoint lower bounds require 100% query coverage and at least 99% identity. All accepted models also passed protein-length, sequence-integrity, two-domain PF01554, reciprocal-orthogroup and public-model-context gates. (B) Recovery of annotated-present positive controls and the fraction of annotated-absent orthogroup-by-assembly combinations with an accepted genome-supported model across four sequence-similarity thresholds. Values are assembly-level presence calls; multiple models within the same orthogroup and assembly are counted once. Reconstructed models are analysis-only and do not constitute revised public annotations or direct evidence of functional copy number. Detailed model- and orthogroup-level evidence is provided in Supplementary File S6. Figure S4. Enlarged reference-anchored phylogeny of the blueberry MATE family. The fixed-model maximum-likelihood tree contains one representative from each of 133 high-confidence blueberry orthogroups and 22 accession-verified functional MATE references. Branch colours denote homogeneous nearest-reference groups, whereas grey branches join mixed groups. The outer ring indicates core, soft-core, shell and cloud classification. Red circles mark verified references, black outlines mark the six citrate-transporter-proximal orthogroups and red stars mark orthogroups containing the best genomic matches to the two root-expression transcripts. The analysis used VT+F+R6 with 1000 ultrafast-bootstrap and 1000 SH-aLRT replicates. Reference proximity prioritizes candidates but does not assign transported substrate. Figure S5. Tree-aligned de novo protein motifs and representative exon–intron structures of the blueberry MATE family. Rows follow the fixed maximum-likelihood tree containing one representative from each of 133 strict blueberry orthogroups and 22 accession-verified functional MATE references. The two narrow annotation columns show the topology-based nearest-reference group and blueberry pan-gene class; neither assigns function. Protein motifs were recovered from the exact 155 unaligned tree sequences with MEME Suite 5.5.9 under a zero-or-one-occurrence-per-sequence model, using 15 motifs 6–50 amino acids long and seed 20260829. Motif blocks use amino-acid coordinates. Pale rows mark the six citrate-transporter-proximal blueberry orthogroups. Exon–intron structures derive from the source GFF3 files for the 133 blueberry representatives and are normalized within transcripts to preserve exon order; exact coordinates are supplied in Supplementary File S5. Functional-reference rows have no corresponding blueberry gene structure. Figure S6. Family-definition sensitivity of the assembly-confounded NHB–SHB comparison. (A) Strict-set PAV-frequency comparison. (B) Sensitivity-set PAV comparison. (C) Sensitivity-set annotation-derived count comparison. (D) Log_2 fold-change concordance after mapping strict orthogroups to sensitivity orthogroups by shared gene membership. Because NHB and SHB labels are perfectly aligned with different assembly configurations in the available genomes, these analyses assess the reproducibility of annotation-derived patterns and are not interpreted as

evidence of a biological breeding-group difference. Figure S7. Complete pairwise neighbouring-gene evidence audit for the original 63 reconciled candidate-group loci. Of 1953 unique locus pairs, 1874 had strong support, 17 had partial support and 62 had no support. All no-support comparisons involved the assembly-gap-limited B02 locus. The three loci added with the chromosome-level 'Draper' and 'Stevens' reference extension are documented separately in Supplementary File S4; representative within-genome gene order is shown in main Figure 4. Figure S8. Comparability audit for six blueberry 2-kb upstream regions relative to B10/'Legacy'. The dashed line marks the predefined 0.90 known-base comparison threshold. B03, B10 and B20 passed; B08, B13 and B17 had insufficient homologous overlap and were excluded from the sequence-difference tracks in main Figure 5. The complete derived-data and code archive is available at Figshare: https://figshare.com/articles/dataset/Vaccinium_MATE_pangenome_data_and_code/34291467 (accessed on 12 October 2026).

Author Contributions: Conceptualisation, R.F.; methodology, R.F. and B.L.; software, B.L.; formal analysis, B.L.; investigation, B.L.; resources, R.F.; writing—original draft preparation, B.L. and R.F.; writing—review and editing, B.L., Q.Y. and R.F.; visualisation, B.L.; supervision, R.F.; project administration, R.F.; funding acquisition, R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Jingmen City Innovation and Development Joint Fund of the Hubei Provincial Natural Science Foundation, grant number JCZRLH202601519; the Key Project of the Science and Technology Research Program of the Department of Education of Hubei Province, grant number D20244303; and the PhD Startup Foundation of Jingchu University of Technology, grant number YYZ202512.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The 34 crop genome assemblies, annotations, and proteomes are available through the Genome Database for *Vaccinium*; the 'Draper' and 'Stevens' source URLs and SHA-256 hashes are listed in Supplementary File S1. The additional-species resources are available under NCBI GenBank GCA_020921065.1, PGDD/GDV SGLD20220023, Figshare dataset 23926251, and Genome Warehouse GWHBPBP00000000. Public root transcriptome data are available under PRJNA353989 and PRJNA1037260. Supporting tables and analyses are supplied in the Supplementary Materials. A complete reproducibility archive containing derived data, alignments, trees, analysis scripts, and associated manifests is available at Figshare: https://figshare.com/articles/dataset/Vaccinium_MATE_pangenome_data_and_code/34291467 (accessed on 12 October 2026).

Acknowledgments: The authors thank the Plant Biological Breeding Team within the College of Food and Biology at Jingchu University of Technology for its support. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.6) to assist with bioinformatics analysis, code development, and language editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

CDS, coding sequence; CPM, counts per million; FDR, false discovery rate; GDV, genome database for *Vaccinium*; MATE, multidrug and toxic compound extrusion; MCScanX, Multiple Collinearity Scan toolkit X; MDS, multidimensional scaling; PAV, presence–absence variation; PH01, short-read root pH dataset PRJNA353989; PH02, long-read root pH dataset PRJNA1037260; SH-aLRT, SH-like approximate likelihood-ratio test; TMM, trimmed mean of M-values; UFBoot, ultra-fast bootstrap.

References

1. Omote, H.; Hiasa, M.; Matsumoto, T.; Otsuka, M.; Moriyama, Y. The MATE proteins as fundamental transporters of metabolic and xenobiotic organic cations. *Trends Pharmacol. Sci.* **2006**, *27*, 587–593. [CrossRef] [PubMed]
2. Takanashi, K.; Shitan, N.; Yazaki, K. The multidrug and toxic compound extrusion (MATE) family in plants. *Plant Biotechnol.* **2014**, *31*, 417–430. [CrossRef]
3. Durrett, T.P.; Gassmann, W.; Rogers, E.E. The FRD3-mediated efflux of citrate into the root vasculature is necessary for efficient iron translocation. *Plant Physiol.* **2007**, *144*, 197–205. [CrossRef] [PubMed]
4. Yokosho, K.; Yamaji, N.; Ueno, D.; Mitani, N.; Ma, J.F. OsFRDL1 is a citrate transporter required for efficient translocation of iron in rice. *Plant Physiol.* **2009**, *149*, 297–305. [CrossRef] [PubMed]
5. Magalhaes, J.V.; Liu, J.; Guimarães, C.T.; Lana, U.G.P.; Alves, V.M.C.; Wang, Y.H.; Schaffert, R.E.; Hoekenga, O.A.; Piñeros, M.A.; Shaff, J.E.; et al. A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nat. Genet.* **2007**, *39*, 1156–1161. [CrossRef] [PubMed]
6. Furukawa, J.; Yamaji, N.; Wang, H.; Mitani, N.; Murata, Y.; Sato, K.; Katsuhara, M.; Takeda, K.; Ma, J.F. An aluminum-activated citrate transporter in barley. *Plant Cell Physiol.* **2007**, *48*, 1081–1091. [CrossRef] [PubMed]
7. Liu, J.; Magalhaes, J.V.; Shaff, J.; Kochian, L.V. Aluminum-activated citrate and malate transporters from the MATE and ALMT families function independently to confer *Arabidopsis* aluminum tolerance. *Plant J.* **2009**, *57*, 389–399. [CrossRef] [PubMed]
8. Maron, L.G.; Piñeros, M.A.; Guimarães, C.T.; Magalhaes, J.V.; Pleiman, J.K.; Mao, C.; Shaff, J.; Belicuas, S.N.J.; Kochian, L.V. Two functionally distinct members of the MATE (multi-drug and toxic compound extrusion) family of transporters potentially underlie two major aluminum tolerance QTLs in maize. *Plant J.* **2010**, *61*, 728–740. [CrossRef] [PubMed]
9. Yokosho, K.; Yamaji, N.; Ma, J.F. An Al-inducible MATE gene is involved in external detoxification of Al in rice. *Plant J.* **2011**, *68*, 1061–1069. [CrossRef] [PubMed]
10. Wu, X.; Li, R.; Shi, J.; Wang, J.; Sun, Q.; Zhang, H.; Xing, Y.; Qi, Y.; Zhang, N.; Guo, Y.D. *Brassica oleracea* MATE encodes a citrate transporter and enhances aluminum tolerance in *Arabidopsis thaliana*. *Plant Cell Physiol.* **2014**, *55*, 1426–1436. [CrossRef] [PubMed]
11. Payá-Milans, M.; Nunez, G.H.; Olmstead, J.W.; Rinehart, T.A.; Staton, M. Regulation of gene expression in roots of the pH-sensitive *Vaccinium corymbosum* and the pH-tolerant *Vaccinium arboreum* in response to near neutral pH stress using RNA-Seq. *BMC Genom.* **2017**, *18*, 580. [CrossRef] [PubMed]
12. Yocca, A.E.; Platts, A.; Alger, E.; Teresi, S.; Mengist, M.F.; Benevenuto, J.; Ferrão, L.F.V.; Jacobs, M.; Babinski, M.; Magallanes-Lundback, M.; et al. Blueberry and cranberry pangenomes as a resource for future genetic studies and breeding efforts. *Hortic. Res.* **2023**, *10*, uhad202. [CrossRef] [PubMed]
13. Bayer, P.E.; Golicz, A.A.; Scheben, A.; Batley, J.; Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **2020**, *6*, 914–920. Correction in *Nat. Plants* **2020**, *6*, 1389. [CrossRef] [PubMed]
14. Golicz, A.A.; Bayer, P.E.; Barker, G.C.; Edger, P.P.; Kim, H.; Martinez, P.A.; Chan, C.K.K.; Severn-Ellis, A.; McCombie, W.R.; Parkin, I.A.P.; et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **2016**, *7*, 13390. [CrossRef] [PubMed]
15. Montenegro, J.D.; Golicz, A.A.; Bayer, P.E.; Hurgobin, B.; Lee, H.T.; Chan, C.K.K.; Visendi, P.; Lai, K.; Doležel, J.; Batley, J.; et al. The pangenome of hexaploid bread wheat. *Plant J.* **2017**, *90*, 1007–1013. [CrossRef] [PubMed]
16. Maron, L.G.; Guimarães, C.T.; Kirst, M.; Albert, P.S.; Birchler, J.A.; Bradbury, P.J.; Buckler, E.S.; Coluccio, A.E.; Danilova, T.V.; Kudrna, D.; et al. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5241–5246. [CrossRef] [PubMed]
17. Fujii, M.; Yokosho, K.; Yamaji, N.; Saisho, D.; Yamane, M.; Takahashi, H.; Sato, K.; Nakazono, M.; Ma, J.F. Acquisition of aluminium tolerance by modification of a single gene in barley. *Nat. Commun.* **2012**, *3*, 713. [CrossRef] [PubMed]
18. Yamaji, N.; Huang, C.F.; Nagao, S.; Yano, M.; Sato, Y.; Nagamura, Y.; Ma, J.F. A zinc finger transcription factor ART1 regulates multiple genes implicated in aluminum tolerance in rice. *Plant Cell* **2009**, *21*, 3339–3349. [CrossRef] [PubMed]
19. Tsutsui, T.; Yamaji, N.; Ma, J.F. Identification of a cis-acting element of ART1, a C2H2-type zinc-finger transcription factor for aluminum tolerance in rice. *Plant Physiol.* **2011**, *156*, 925–931. [CrossRef] [PubMed]
20. Arbelaez, J.D.; Maron, L.G.; Jobe, T.O.; Piñeros, M.A.; Famoso, A.N.; Rebelo, A.R.; Singh, N.; Ma, Q.; Fei, Z.; Kochian, L.V.; et al. ALUMINUM RESISTANCE TRANSCRIPTION FACTOR 1 (ART1) contributes to natural variation in aluminum resistance in diverse genetic backgrounds of rice (*O. sativa*). *Plant Direct* **2017**, *1*, e00014. [CrossRef] [PubMed]
21. Colle, M.; Leisner, C.P.; Wai, C.M.; Ou, S.; Bird, K.A.; Wang, J.; Wisecaver, J.H.; Yocca, A.E.; Alger, E.I.; Tang, H.; et al. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience* **2019**, *8*, giz012. [CrossRef] [PubMed]
22. Diaz-Garcia, L.; Garcia-Ortega, L.F.; González-Rodríguez, M.; Delaye, L.; Iorizzo, M.; Zalapa, J. Chromosome-Level Genome Assembly of the American Cranberry (*Vaccinium macrocarpon* Ait.) and Its Wild Relative *Vaccinium microcarpum*. *Front. Plant Sci.* **2021**, *12*, 633310. [CrossRef] [PubMed]

23. Potter, S.C.; Luciani, A.; Eddy, S.R.; Park, Y.; Lopez, R.; Finn, R.D. HMMER web server: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W200–W204. [CrossRef] [PubMed]
24. Larralde, M.; Zeller, G. PyHMMER: A Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **2023**, *39*, btad214. [CrossRef] [PubMed]
25. Deorowicz, S.; Debudaj-Grabysz, A.; Gudyś, A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.* **2016**, *6*, 33964. [CrossRef] [PubMed]
26. Wong, T.K.F.; Ly-Trong, N.; Ren, H.; Demotte, P.; Baños, H.; Roger, A.J.; Susko, E.; Bielow, C.; De Maio, N.; Goldman, N.; et al. IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. *Mol. Biol. Evol.* **2026**, *43*, msag117. [CrossRef] [PubMed]
27. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef] [PubMed]
28. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522. [CrossRef] [PubMed]
29. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321. [CrossRef] [PubMed]
30. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef] [PubMed]
31. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 238. [CrossRef] [PubMed]
32. Buchfink, B.; Reuter, K.; Drost, H.G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **2021**, *18*, 366–368. [CrossRef] [PubMed]
33. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **2023**, *39*, btad014. [CrossRef] [PubMed]
34. Hallgren, J.; Tsirigos, K.D.; Pedersen, M.D.; Almagro Armenteros, J.J.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv* **2022**. [CrossRef]
35. Cui, F.; Ye, X.; Li, X.; Yang, Y.; Hu, Z.; Overmyer, K.; Brosché, M.; Yu, H.; Salojärvi, J. Chromosome-level genome assembly of the diploid blueberry *Vaccinium darrowii* provides insights into its subtropical adaptation and cuticle synthesis. *Plant Commun.* **2022**, *3*, 100307. [CrossRef] [PubMed]
36. Zeng, T.; He, Z.; He, J.; Lv, W.; Huang, S.; Li, J.; Zhu, L.; Wan, S.; Zhou, W.; Yang, Z.; et al. The telomere-to-telomere gap-free reference genome of wild blueberry *Vaccinium duclouxii* provides its high soluble sugar and anthocyanin accumulation. *Hortic. Res.* **2023**, *10*, uhad209. [CrossRef] [PubMed]
37. Hirabayashi, K.; Debnath, S.C.; Owens, G.L. Unveiling the evolutionary history of lingonberry (*Vaccinium vitis-idaea* L.) through genome sequencing and assembly of European and North American subspecies. *G3 Genes Genomes Genet.* **2024**, *14*, jkad294. [CrossRef] [PubMed]
38. Yang, L.; Li, M.; Shen, M.; Bu, S.; Zhu, B.; He, F.; Zhang, X.; Gao, X.; Xiao, J. Chromosome-level genome assembly and annotation of the native Chinese wild blueberry *Vaccinium bracteatum*. *Fruit Res.* **2022**, *2*, 8. [CrossRef]
39. Wang, Y.; Tang, H.; DeBarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.H.; Jin, H.; Marler, B.; Guo, H.; et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **2012**, *40*, e49. [CrossRef] [PubMed]
40. Nei, M.; Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **1986**, *3*, 418–426. [CrossRef] [PubMed]
41. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef] [PubMed]
42. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
43. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **2011**, *17*, 10–12. [CrossRef]
44. Liao, Y.; Smyth, G.K.; Shi, W. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **2013**, *41*, e108. [CrossRef] [PubMed]
45. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef] [PubMed]
46. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
47. Lun, A.T.L.; Chen, Y.; Smyth, G.K. It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods Mol. Biol.* **2016**, *1418*, 391–416. [CrossRef] [PubMed]
48. Liu, C.; Wei, X.; Yang, Y.; Liu, Y.; Yuan, X.; Tao, C. The full-length transcriptome analysis of blueberry under different pH conditions. *BMC Plant Biol.* **2025**, *25*, 732. [CrossRef] [PubMed]
49. Harrison, P.W.; Ahamed, A.; Aslam, R.; Alako, B.T.F.; Burgin, J.; Buso, N.; Courtot, M.; Fan, J.; Gupta, D.; Haseeb, M.; et al. The European Nucleotide Archive in 2020. *Nucleic Acids Res.* **2021**, *49*, D82–D85. [CrossRef] [PubMed]

50. Boratyn, G.M.; Thierry-Mieg, J.; Thierry-Mieg, D.; Busby, B.; Madden, T.L. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinform.* **2019**, *20*, 405. [CrossRef] [PubMed]
51. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef] [PubMed]
52. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
53. Manni, M.; Berkeley, M.R.; Seppey, M.; Simão, F.A.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **2021**, *38*, 4647–4654. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.