













ORIGINAL ARTICLE

Deep resequencing unveils novel SNPs, InDels, and large structural variants for the clonal fingerprinting of sweet orange [*Citrus sinensis* (L.) Osbeck]

Davide Scaglione¹  | Angelo Ciacciulli²  | Stefano Gattolin^{3,4}  | Marco Caruso²  |
 Fabio Marroni^{5,6}  | Giuseppina Las Casas²  | Irena Jurman⁶  |
 Grazia Licciardello^{2,7}  | Antonino Felice Catara⁷  | Laura Rossini^{4,8}  |
 Concetta Licciardello²  | Michele Morgante^{5,6} 

¹IGA Technology Services s.r.l., Udine, Italy

²CREA Research Centre Olive, Fruit and Citrus Crops, Acireale, Italy

³CNR-National Research Council of Italy, Institute of Agricultural Biology and Biotechnology, Milan, Italy

⁴PTP Science Park, Lodi, Italy

⁵Dipartimento di Scienze Agro-alimentari, Ambientali e Animali, Università degli Studi di Udine, Udine, Italy

⁶IGA-Istituto di Genomica Applicata, Udine, Italy

⁷Parco Scientifico e Tecnologico della Sicilia, Catania, Italy

⁸Department of Agricultural and Environmental Sciences (DISAA), University of Milan, Milan, Italy

Correspondence

Concetta Licciardello, CREA Research Centre Olive, Fruit and Citrus Crops, Corso Savoia 190, 95024 Acireale, Italy.
 Email: concetta.licciardello@crea.gov.it

Michele Morgante, Dipartimento di Scienze Agro-alimentari, Ambientali e Animali, Università degli Studi di Udine, Via delle Scienze 206, 33100, Udine, Italy.
 Email: michele.morgante@uniud.it

Assigned to Associate Editor Mukesh Jain.

Funding information

Qualitrace—Messa a punto e validazione di tool genetici e chimici per la tracciabilità integrata e la valorizzazione della qualità dell'Arancia Rossa di Sicilia IGP, Grant/Award Number: MIPAAF

Abstract

The large phenotypic variability characterizing the sweet orange [*Citrus sinensis* (L.) Osbeck] germplasm arose from spontaneous somatic mutations and led to the diversification of major groups (common, acidless, Navel, and pigmented). Substantial divergence also occurred within each varietal group. The genetic basis of such variability (i.e., ripening time, fruit shape, color, acidity, and sugar content) is largely uncharacterized, and therefore not exploitable for molecular breeding. Moreover, the clonal nature of all sweet orange accessions hinders the traceability of propagation material and fruit juice using low-density molecular markers. To build a catalog of somatic mutations in Italian varieties, 20 accessions were sequenced at high coverage. This allowed the identification of single nucleotide polymorphisms (SNPs), structural variants (SVs), and large hemizygous deletions, specific to clones or varietal groups. A panel of 239 SNPs was successfully used for genotyping 221 sweet

Abbreviations: CNV, copy number variant; HRM, high-resolution melting; LTR, long terminal repeat; SNP, single nucleotide polymorphism; SV, structural variant; TE, transposable element.

Davide Scaglione and Angelo Ciacciulli contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *The Plant Genome* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

DM19527/7303/2016;
 NOVARANCIA—Innovazioni tecnologiche
 (genetiche, fitosanitarie ed agronomiche)
 per la valorizzazione e tracciabilità
 dell'Arancia Rossa di Sicilia, Grant/Award
 Number: PSR Misura 16.1. D.D.S.
 n.215/2022; IT Citrus
 genomics—Genomica funzionale
 miglioramento genetico ed innovazione per
 la valorizzazione dei prodotti della filiera
 agrumicola, Grant/Award Numbers: PON,
 —, D.M. 01/Ric del 18 gennaio 2010

orange accessions, allowing them to be clustered into varietal groups. Furthermore, genotyping of SNPs and SVs was extended to leaf and juice samples of commercial varieties belonging to two varietal groups (Moro and Tarocco) collected from 26 sites in Southern Italy, confirming the usefulness of the identified markers for the identification of specific clones. Interestingly, we found that the insertion of the transposable element VANDAL in the gene exons significantly affected the level of allelic-specific expression. Finally, the markers developed in the present work contribute to unraveling the origin and diversification of sweet oranges, representing a reliable and efficient molecular tool for the unambiguous fingerprint of somatic mutants and an asset for the traceability of orange plant material and fruit juice.

Plain Language Summary

The large variability (i.e., ripening time, fruit shape, color, acidity, and sugar content) of the sweet oranges arose from spontaneous mutations, which genetic basis is largely uncharacterized and unexploitable using traditional markers. The sequencing of the DNA of 20 Italian varieties allowed the identification of short and large mutations, specific to each or shared among accessions. More than 200 short mutations can separate more than 220 different accessions, clustering into varietal groups. A subset of short and large mutations was used to trace leaf and juice samples of commercial varieties collected from 26 sites in Southern Italy. Moreover, the insertion of a mobile element in a functional portion of genes modifies the expression of the gene. The genetic modifications that we identified contribute to elucidating the origin and variability of sweet oranges, representing a reliable and efficient molecular tool for the traceability of orange plant material and fruit juice.

1 | INTRODUCTION

Sweet orange [*Citrus sinensis* (L.) Osbeck] is the most important citrus species for the fresh market and the production of juice, accounting for more than 48 of the 104 million tonnes of global citrus production (orange, tangerine/mandarin, grapefruit, and lemon/lime) forecasted for 2023/2024 (<https://apps.fas.usda.gov/psdonline/circulars/citrus.pdf>). A genome-wide study (Q. Xu et al., 2013) suggested that *C. sinensis* originated from a backcross hybrid between pummelo (P) and mandarin (M). However, the citrus genome consortium challenged this (P × M) × M backcross origin scenario since clear P/P segments were found in the sweet orange genome, which would require both parents to have some P ancestry (Velasco & Licciardello, 2014; Wu et al., 2014). More recently, Wu et al. (2018) proposed that the species' genomic composition could be explained by frequent pummelo introgressions into type-2 (early-admixture) mandarins (i.e., a mandarin having a small percentage of pummelo alleles). Although the sweet orange was first introduced into Europe in the 15th century (Langgut, 2017), it seems certain that the Portuguese contributed to the spread of the species by introducing a superior variety in

the 16th century, favoring the establishment of its commercial importance in Europe (Scora, 1975). Ferrari (1646) in his *Hesperides* described 16 varieties but singled out the Portuguese orange as something new and good. The Portuguese orange is likely the ancestor of all the cultivars that originated over the last centuries in the different growing areas worldwide (Deng et al., 2020).

Some citrus genotypes are clonally propagated apomictically (X. Wang et al., 2017) through nucellar embryony (the development of non-sexual embryos originating in the maternal nucellar tissue of the ovule), and this natural process may have been co-opted during domestication; otherwise, grafting is a relatively recent phenomenon (Mudge et al., 2009). Both modes of clonal propagation have led to the domestication of fixed (desirable) genotypes, including interspecific hybrids, such as oranges, limes, lemons, grapefruits, and other types (Wu et al., 2018). Notably, all sweet orange varieties arose exclusively from somatic mutations, so they differ for one or very few characters and share the same genetic background with minimal sequence variation (Caruso et al., 2016). Spontaneous mutations led to the generation of hundreds of cultivated selections worldwide

(Barry et al., 2020). Currently, sweet orange varieties can be divided into varietal groups based on fruit characteristics (common, acidless, Navel, and blood). Each varietal group is also divided into subgroups (i.e., “Valencia” types among the common ones and “Tarocco,” “Moro,” and “Sanguigno + Sanguinello” among the blood oranges). Common and Navel oranges are the most widespread group and have been cultivated worldwide for centuries with many different mutations identified around the world. On the other hand, blood oranges have been traditionally cultivated only in the Mediterranean area, particularly in Italy, where most of the phenotypic diversity was described (Caruso et al., 2016). Finally, some acidless varieties are also cultivated for niche markets and are known by different names (Barry et al., 2020).

In somatic mutants, mutations initially affect single cells of the shoot apical meristem. Later, the mutation can spread to one or more cell layers or in specific portions of them and can remain as mosaic or become solid (Pelsy, 2010). Among the woody, vegetatively propagated crops, the effect of somatic mutations on phenotypic traits has been particularly studied in grapevine through genetics (Cardone et al., 2016; Hocquigny et al., 2004; Pelsy et al., 2015). The interest in such effects led to the development of X-scan, a bioinformatics tool dedicated to the detection of somatic and mosaic structural variants (Marroni et al., 2017). In citrus, the majority of the mutations responsible for specific phenotypic changes, such as fruit color, ripening period, acidity, and fruit size, remain unknown. However, recent findings shed light on the molecular basis of some aspects of sweet orange fruit pigmentation and flavor: All anthocyanins-rich orange selections differ from common sweet oranges for the presence of a *Copia*-like retrotransposon upstream of an *MYB* gene regulating anthocyanin biosynthesis (Butelli et al., 2012), while loss of acidity is due to mutations in the sequence of the *Noemi* gene encoding a basic helix-loop-helix transcription factor, as demonstrated in different studies comparing acidic and acidless citrus varieties (Butelli et al., 2019; Strazzer et al., 2019). The mosaic nature of many somatic mutations distinguishing varieties makes their detection particularly challenging because the new mutation will always appear in a heterozygous state but with a frequency always lower than that expected for a germinal mutation (0.5) that will depend on the cell layer composition of the analyzed sample.

The narrow genetic base of the sweet orange germplasm has also implications for cultivar identification using molecular markers. Traditional markers such as single-sequence repeats or single nucleotide polymorphisms (SNPs) retrieved from public databases and used for the genotyping of other citrus species have been ineffective for sweet orange clonal fingerprinting. Older molecular approaches based on dominant random markers identified some polymorphisms, but those markers suffered from a lack of reproducibility (Jones et al.,

Core Ideas

- Clonal mutations and genetic traceability in sweet orange are reported.
- 221 sweet orange accessions have been successfully genotyped through a panel of 239 single nucleotide polymorphisms (SNPs).
- Leaf and juice samples of commercial varieties collected in 26 Italian sites have been genetically traced using SNPs and structural variants.
- Significant allelic-specific expression of transposable element VANDAL have been detected into gene exons.

1997). Whole genome resequencing projects have helped identify causative mutations of specific phenotypes in fruit crops, such as peach (Falchi et al., 2013; Vendramin et al., 2014), date palm (Hazzouri et al., 2015), apple (Zhang et al., 2014), and grapevine (Carbonell-Bejerano et al., 2017; Y. Xu et al., 2016), and provide reliable information to detect SNPs, InDels, and structural variations (SVs) responsible for cultivar diversification. In the case of citrus, resequencing projects identified different kinds of mutations, such as chromosomal rearrangements and deletions, as well as mosaicism, leading to cultivar diversification in clementine (Terol et al., 2015), and to different patterns of resistance genes to Huanglongbing in *Poncirus* (Rawat et al., 2017). Analysis of resequencing data from 100 citrus accessions allowed the identification of the locus for nucellar polyembryony (X. Wang et al., 2017). Furthermore, a large subset of sweet orange varieties ranging in acidity content has been resequenced to identify mutations, SVs, and transposable elements responsible for the acidless trait (L. Wang et al., 2021). Recently, allele-specific expression (ASE) and hidden somatic mutations from 87 sweet orange somatic varieties were collected to generate a phased somatic variation map and combined to demonstrate that somatic mutations influence fruit development in sweet orange (N. Wang et al., 2024).

Here, we report the identification of SNPs, SVs, and InDels from the resequencing of 20 sweet orange genomes, successfully differentiating varietal groups, subgroups, and specific clonal selections. The resequenced sweet orange varieties and mutants were chosen from diverse varietal groups differing for important phenotypic traits, such as ripening period, fruit firmness and acidity content, presence/absence of anthocyanins or lycopene in the pulp, and anthocyanins in the peel. Moreover, the comparison among the 20 genomes provided evidence that the mobile element families present in the genome widely differed in their transpositional activity over the time of clonal selection of sweet orange accessions.

2 | MATERIALS AND METHODS

2.1 | Plant material

Sweet orange trees were grown at the CREA experimental orchard located in Palazzelli (Siracusa, Italy). Table 1 lists the 20 accessions selected for resequencing by specifying their origin and main features and includes representatives of the common, blood, and navel varietal groups. The list contains old lines and nucellar selections, old varieties, such as Shamouti (Barry et al., 2020), widespread cultivars (Lane Late, Cara Cara, and Campbell), and some blood oranges that represent typical Italian cultivars (Caruso et al., 2016). Some selections are characterized by the acidless or low acid trait (Vaniglia and Ferreri) or loss of pigmentation (Moro NP). Selections also differ for the ripening period, ranging from early to late (Table 1).

Illumina GoldenGate genotyping analyses were conducted on 221 leaf samples and 18 juice samples collected from 221 accessions, consisting of commercial selections, nucellar clones, and old germplasm lines sampled at the CREA experimental farm located in Palazzelli (Siracusa, Italy) (Table S1). Competitive allele-specific PCR (KASP) validations were conducted on 52 leaf and 33 juice samples collected from the certification and germplasm collection of CREA, four nurseries, and 22 orchards, all located in Southern Italy (Table S2).

2.2 | DNA extraction

Around 200 mg of freeze-dried leaf tissue was used for each extraction of genomic DNA (gDNA) using a NucleoSpin Plant II kit (Macherey-Nagel). Orange juice gDNA samples were obtained starting from 500 μ L of filtered juice. Briefly, the samples were supplemented with 3.5 mL TES buffer (Tris 0.2 M pH 8, ethylenediaminetetraacetic acid [EDTA] 1 mM, and sodium dodecyl sulfate 1%) and incubated on ice for 20 min. After four 1:1 phenol extractions, the gDNA was ethanol precipitated and resuspended in 10 mg/ μ L RNase (Invitrogen). After 30 min of incubation at 37°C, the gDNA was column purified using a NucleoSpin Plant II kit. The gDNA of each sample was quantified using Picogreen (Invitrogen) and normalized to 50 ng/ μ L with 10 mM Tris-HCl pH 8.0, and 1 mM EDTA.

2.3 | Whole genome library construction and sequencing

Illumina libraries of sweet orange samples were constructed following the instructions of the manufacturer's protocol of the Nextera DNA Sample Preparation kit (Illumina Inc.). The purification of the libraries was conducted with magnetic beads AMPure XP (Agencourt), then quantified on a Caliper GX (Perkin Elmer). The libraries were validated using a Bio-

analyzer 2100 (Agilent), ensuring that fragments average was in the range of 300–400 bp, quantified using Qubit (Invitrogen), and then sequenced using an Illumina HiSeq 2500 (Illumina), generating 125 bp paired ends.

2.4 | Variant calling, somatic SNPs, and InDels

Reads were cleaned from any adapter residuals on both 5' and 3' termini with cutadapt 1.11 (Martin, 2011) and successively removed low-quality bases with the use of *erne-filter* (Del Fabbro et al., 2013). Reads were then aligned using BWA-MEM v 0.7.3 (Heng, 2013) using as reference genome the *Csi_valencia_1.0*, while variant calling was performed using GATK 4 after removal of reads with mapping quality below 10 and filtering duplicated read pairs. Only variant sites covered by at least 10 reads were retained. To exclude most false positives, several classes of genomic intervals were not considered: repetitive regions and N gaps longer than 1000 bp, microsatellite sites extended by 10 bp on both sides as identified with Sputnik algorithm, GATK re-alignment intervals extended by 10 bp on both sides. Candidate somatic alleles were only retained with a minimum of three supporting reads, while homozygous calls were only considered with a coverage of eight reads. The following filters were also applied to GATK parameters: $-2.5 > \text{ReadPosRankSum} > 2.5$, $\text{ReadPosRankSum} > \text{StrandBias} \leq 0.0$, $\text{Dels} \leq 0.9$, $\text{QUAL} \geq 100$. Control over coverage parameters was also applied by not considering sites where more than 30% of samples fall either below 0.50 or above 1.75 of the sample median coverage. A minimum of 15% for the minor allele frequency was required to assess the presence of a candidate mutant allele call, while a maximum of 2% was considered as acceptable contamination of either reference or alternative alleles to keep a homozygous (i.e., not mutant) call. Calls not assessed as confident homozygous or heterozygous were labeled as ambiguous. Sites reporting more than 5% of samples with ambiguous calls were dropped. Code for detection of somatic SNP/INDEL sites is available at <https://bitbucket.org/dscaglione-igatech/shortread-somatic-variants-scripts> as *somatic_hunter_V03_alpha.py* and the simplified procedure is depicted in Figure S1. Regions with a match on RepBase18.08 via RepeatMasker or simple tandem repeats identified with Tandem Repeat Finder were extended by 5 bp and excluded from the search space. Other parameters set to the script were $\text{skip_indel} \Rightarrow \text{False}$, $\text{max_jux} \Rightarrow 500$, $\text{slop} \Rightarrow 0$, $\text{privateness} \Rightarrow 0.8$, $\text{min_call_ratio} \Rightarrow 1.0$, $\text{window} \Rightarrow 5$, $\text{only_alt} \Rightarrow \text{False}$, $\text{gatk} \Rightarrow -2.5, 2.5, 0.0, 0.9, 100$, $\text{max_ambiguity} \Rightarrow 0.1$, $\text{max_disgregation} \Rightarrow 1$, $\text{hcov_ratio} \Rightarrow 0.5$, $\text{max_alt_freq} \Rightarrow 0.02$, $\text{cov_coeff} \Rightarrow 3.0$, $\text{groups} \Rightarrow [\text{'NAV= ar_Navel_Cara_Cara, ar_Navel_Nuc_Lane_Late_C2611, ar_Navel_Cara_Cara_Lindcove'}, \text{'VAN= ar_Vaniglia_Biondo, ar_Vaniglia_}$

TABLE 1 List of resequenced orange clones including a description of the main phenotypic traits.

Accession name	Abbreviation	Varietal group and subgroup	Origin	Ripening period ^a	Anthocyanins in the rind	Flesh anthocyanin content (mg/L) ^{b,c}	TSS (%) ^c	Acidity ^{c,d}	Fruit size	Additional description and remarks
Navel Cara Cara	Cara Cara	Navel	Old line	Medium	Undetected	Undetected	11.6	0.9	Large	Presence of lycopene in the flesh. No seeds.
Navel Cara Cara Lindcove nuc. F8187	F8187	Navel	Nucellar line	Medium	Undetected	Undetected	11.7	0.9	Medium to large	Nucellar line of Cara Cara without lycopene pigmentation.
Navel Lane Late nucellare C2611	Lane Late	Navel	Nucellar line	Late	Undetected	Undetected	11.5	0.9	Large	Nucellar selection of one of the most cultivated navel selections. No seeds.
Shamouti	Shamouti	Common	Old line	Medium	Undetected	Undetected	11.8	1.3	Medium to large	Old variety discovered in Palestine area.
Ovale NuCELLARE	Ovale	Common	Nucellar line	Late	Undetected	Undetected	9.5	1.4	Medium to large	Nucellar selection of a late variety of common orange of Italian origin; subjected to the production of out of season bloom and fruit.
Valencia Campbell S2G 18 19 Nucellare	Campbell	Common (Valencia)	Nucellar line	Late	Undetected	Undetected	9.2	1.5	Medium to large	Nucellar selection of a Valencia late-ripening variety, widespread in many citrus areas for its productivity. Few seeds.
Vaniglia Biondo	Van Biondo	Acidless	Old line	nd	Undetected	Undetected	12.1	0.1	Small to medium	Acidless variety with seeds (>5).
Vaniglia Sanguigno	Van Sang	Acidless	Old line	nd	Undetected	Undetected	12.3	0.1	Small to medium	Acidless variety presenting lycopene in the flesh. Seedy.
Sanguinello Comune	Sanguinello	Pigmented (Sanguinello)	Old line	Medium	Weak	13.9	11.3	1.5	Small to medium	Blood orange variety now disappearing due to the small fruit size and lower fruit quality compared to the new mid- to late-season Tarocco selections.
Sanguinello Moscato R I	Moscato	Pigmented (Sanguinello)	Old line	Medium	Weak	4	10.8	1.2	Small to medium	Medium ripening blood orange variety with better flavor compared with Sanguinello Comune.
Doppio Sanguigno	Doppio Sang	Pigmented (Sanguigno)	Old line	Medium to late	Strong	10.1	11.7	1.4	Small to medium	Old blood variety disappearing for poor characteristics; highly pigmented in the peel. Seedy.
Moro non pigmentato Nuc.	Moro NP	Pigmented (Moro)	Nucellar line	Early	no	0.3	11.2	1.3	Small to medium	Nucellar line of Moro without anthocyanin pigmentation. Seeds number per fruit above five.

(Continues)

TABLE 1 (Continued)

Accession name	Abbreviation	Varietal group and subgroup	Origin	Ripening period ^a	Anthocyanins in the rind	Flesh anthocyanin content (mg/L) ^{b,c}	TSS (%) ^c	Acidity ^{c,d}	Fruit size	Additional description and remarks
Moro Nucleare 58-8D-1	Moro nuc.	Pigmented (Moro)	Nucleare line	Early	Strong	66.3	10.5	1.3	Small to medium	One of the most cultivated Moro clones. Zero to two seeds per fruit.
Moro VCR	Moro VCR	Pigmented (Moro)	Old line	Early	Strong	49	11.1	1.2	Small to medium	Moro shoot-tip grafted clone; known as Moro M45. Zero to two seeds per fruit.
Tarocco TDV	TDV	Pigmented (Tarocco)	Nucleare line	Early	Weak	71.4	11.2	1	Medium to large	Nucleare selection isolated from a degenerative mutation. One of the earliest Tarocco clone. Soft fruit, highly pigmented in the pulp.
Tarocco Lempso C Nucleare	Lempso	Pigmented (Tarocco)	Nucleare line	Medium	Strong	17.9	11.5	1.1	Medium to large	Cultivated selection with the highest rind anthocyanin pigmentation within Tarocco varietal group
Tarocco Meli C 8158 Nucleare	Meli	Pigmented (Tarocco)	Nucleare line	Late	Weak	11	11.1	1.7	Large	One of the most important late varieties with high acidity content.
Tarocco Ippolito M507	Ippolito	Pigmented (Tarocco)	Old line	Medium	Medium	85.1	11.2	1	Large	One of the most pigmented Tarocco varieties, both in the peel and in the pulp
Tarocco dal muso contrada Bernaldo	Dal Muso	Pigmented (Tarocco)	Old line	Medium	Weak	19.9	11.5	1.3	Medium to large	Slightly necked fruit; mutation discovered in an old Sicilian orchard.
Tarocco Ferreri acidless	Ferreri	Pigmented (Tarocco)	Old line	nd	Strong	19.9	12.6	0.1	Small to medium	Acidless mutation of Tarocco discovered in an orchard.

Abbreviation: TSS, transcription start site.

^aRipening period refers to the usual harvest period in the CREA experimental farm: early: December; medium: January–February; late: March–April.

^bTotal anthocyanin was measured using a spectrophotometer and expressed as cyanidin 3-glucoside equivalents (mg/L).

^cFlesh anthocyanin content, TSS, and acidity were measured as reported by Caruso et al. (2016).

^dTitrate acidity is expressed as a percentage of anhydrous citric acid.

Sanguigno', 'TAR= ar_Tarocco_Ferreri_acidless, ar_Tarocco_Nuc_Lempso_C, ar_Tarocco_TDV, ar_Tarocco_Dal_Muso, ar_Tarocco_Ippolito_VCR, ar_Tarocco_Meli_C_8158_Nuc', 'SAN=ar_Sanguinello_Comune, ar_Sanguinello_Moscato_RI', 'DSAN=ar_Doppio_Sanguigno', 'MOR= ar_Moro_Nuc_non_pigmentato, ar_Moro_Nuc_58-8D-1_Russo, ar_Moro_VCR'], lowglob_coeff=>0.35. The filtering process is based on calling parameters, allelic coverage, and grouping coherence (e.g., a true somatic SNP is expected to be fixed in one or more established clonal types, yet impossible to be variable across multiple distant groups such as common oranges and pigmented ones).

Phylogenetic trees were built using a presence/absence matrix (1: call for a candidate somatic SNP; 0: absence of somatic allele), as for dominant markers, considering the somatic nature of the mutations in the analysis. Then a distance matrix was calculated using the dominant marker model in GenAlex. The tree was built using NeighborNet.

2.5 | Somatic structural variants identification

For each of the 20 resequenced sweet oranges, the mobile element insertions were detected by a custom script based on Abyss assembler and blast procedure. Briefly, clusters of discordant read pairs were identified from the alignment file through Abyss assembly of them and reciprocal positioning of two contigs on both sides of a breakpoint with a minimum alignment of 100 bp provided by blastn search. After, paired reads linked to the former ones were subjected to local Abyss assembly, and resulting contigs were searched against a mobile element database RepBase 18.08. Mobile element insertions were confirmed on successful blast hits on both ends of a mobile element as present in the reference database. Insertions were then recalled across the cohort by counting the fraction of discordant reads at each identified locus over a number of reads that spanned the insertion breakpoint (absence of insertion) to determine the genotypic state of the insertion. The python code used for this analysis is available at <https://bitbucket.org/dscaglione-igatech/shortread-somatic-variants-scripts>, and the methodology is described in detail by Pinosio et al. (2016). Insertions have been detected using the insertion_analysis_db.py script and further recalled using MAQuanti-MEI_dbg.py.

Deletions were identified with Delly (Rausch et al., 2012) and re-called across the cohort with a similar approach as above, by counting the proportion of read pairs spanning the deletions (presence of deletion) over the single reads that are reading through the breakpoints (absence of deletion). Recalling and genotyping of deletions were carried out with the script MAQuanti-DEL_refine.py.

Large somatic copy number variation (CNV) was detected with X-scan software (Marroni et al., 2017) using as input

the same variant call format file as used in somatic SNP detection after filtration of repetitive regions. Default settings have been used: each local test was carried out on a dynamic sliding window of 200 SNP sites, with 180 SNPs overlapping from window to window. In brief, the large SV calls were manually revised, taking into consideration coverage profiles and allelic imbalance. Heterozygous deletions are the same as hemizygous deletions; they are marked by a 50% reduction in coverage coupled with an alternative allele fraction that peaks at 100% or 0% depending on in which haplotype the deletion occurred. Chimeric deletions, on the contrary, show a reduction in coverage below 50%, and the two alleles are still present at any degree of reciprocal prevalence. CNVs are identified with a significant deviation from standard allelic balance coupled with increased coverage in multiples of 50% (one extra copy, CN3) of the original coverage.

Phylogenetic trees were built using the same procedure as for SNP sites. VANDAL-like insertion occurrences were identified by a basic local alignment search tool (BLAST) alignment of the inserted sequences with a representative set of complete elements that have been identified in the *Citrus clementine* genome (GCF_000493195.1) and *C. sinensis* genome (GCF_000317415.1) by identification of complete elements with target duplication sites of 9 bp on the structure of the presence of common sequence at 5' and 3' ends (Table S3). VANDAL insertions were identified as candidate somatic events due to the zygosity states as retrieved for all insertion loci via MAQuanti-MEI_dbg.py script. By comparing the coverage of short read pairs supporting the insertion of VANDAL-like sequences versus the number of reads supporting the absence of insertion in the same locus, we were able to determine the genotypic state.

2.6 | Validation of SNPs, InDels, and transposable element (TE) insertions

Sanger sequencing was performed to validate 76 SNPs, of which 65 were specific to single accessions and 11 were common to two or more accessions. Accession-specific SNPs were chosen from 15 of the 20 resequenced accessions; one or two clonal selections have been used as a negative control for each SNP validation. Primers were designed in regions without other polymorphic sites and to produce amplicons without putative InDels to avoid the generation of potentially low-quality sequences (Table S4). Table S4 lists the primers that were used for polymerase chain reaction (PCR) amplification of the regions flanking each SNP. PCR products were purified using Wizard SV Gel and PCR Clean-Up System (Promega) and sequenced using an ABI3130 Genetic Analyzer. GeneStudio software version 2.2 (GeneStudio Inc.) was used to inspect sequence alignments.

Thirteen putative accession-specific insertions–deletions (1–10 bp) were selected for validation using high-resolution melting (HRM) analysis (Table S5). HRM genotyping was performed on three accessions (one carrying the mutation and two as controls) using a Rotor-GeneQ real-time PCR (Qiagen) and PCR conditions as previously reported (Caruso et al., 2014).

A subset of 42 putative TE insertions, of which 24 were accession-specific and 18 shared among two or more accessions, were selected for PCR validation (Table S6). Two sets of primers for each putative insertion were designed: the first set amplified the region on the left side of the putative breakpoint, with the forward primer located in the reference genome and the reverse in the TE (sx); the second set amplified the right region using a forward primer designed in the TE and the reverse primer located in the genome (dx) (Figure S2). PCR products were amplified using the manufacturer conditions of Platinum Taq DNA Polymerase (Invitrogen) and visualized on 1% agarose gel. The number of accessions used for the validation of each insertion is reported in Table S6.

2.7 | Sanger sequencing for Ruby validation

The DNA extracted from leaves of Tarocco nuc 57-1E-1, Vaccaro, Tunnuliddu, Dolce Demmi, and Valencia using the protocol (Qiagen), was amplified with primers designed in the allele RD-2/RD-1 region used as a marker for anthocyanins pigmentation. The primers TCS1FW 5'-ACCAAGCCGATAAATACTGAT-3' and PMC-2ESRev 5'-CTTCACATCGTTCGCTGTTTC-3' were used to obtain the amplicon. The amplification consisted of the denaturation at 95°C for 6 min, 35 cycles of 95°C for 1 min, 58°C for 1 min, 72°C for 1.5 min, and the final extension at 72°C for 15 min. PCR was performed using the manufacturer's instructions of Platinum Taq DNA Polymerase (Life Science). Primers GAL-POL1-Fw 5'-gccctggagcttaggctaa-3' (reported in Butelli et al. [2012]), ltr_fw2 5'-CACCCACCAATTTCCTAACATTAAC-3', PMC-47 5'-TCCTCTCCTGTCCATGCACCTTACGAAC-3' (Butelli et al., 2012), and Ruby1REV 5'-TCAGCCACCGCAGTCTACAGCT-3' were used for Sanger sequencing. The following primers were used to amplify and sequence R and r-2 alleles: Fw_RUBmicp3 5'-ATTTGCGGTTGGGTGGGTAA-3', PMC-2ESRev 5'-CTTCACATCGTTCGCTGTTTC-3', and RTPMC-1ESRev 5'-TCTCCTCGTTTGATATTCGGGT-3'. Amplification conditions are the same as used above, using an annealing temperature of 56°C. Amplicons were verified on agarose gel 1.5% in TAE 1x. Eurofins Genomics Service performed the Sanger sequencing. The ".ab1 files" were aligned against Ruby (JN402330; Butelli et al., 2012) using Benchling (<https://www.benchling.com>) (Supporting Information Dataset 1; Figure S3).

2.8 | Illumina GoldenGate genotyping analyses

Using the 1169 sets of filtered polymorphic SNPs identified, 768 were selected to design two custom-made 384-plex GoldenGate VeraCode oligo pool assay sets for the BeadXpress Reader (Illumina), named Orange1 and Orange2 (Table S7). SNP selection was based on the absence of neighboring polymorphisms and Illumina Functionality Scores. The 384-plex assays were used to genotype 221 orange gDNA leaf samples, while the 18 gDNA samples obtained from juice were genotyped with Orange1. The results were analyzed using GenomeStudio software (Illumina) for automatic genotype clustering and calling based on the cy3/cy5 dye signal intensities ratio.

Due to the highly clonal origin of sweet oranges, accessions did not separate into three clusters but formed a large cluster and, when the SNP was validated, a much smaller cluster, which could in some cases include only the genotype on which that specific SNP was discovered. Therefore, the cluster positions could not be correctly determined automatically with GenomeStudio and a manual revision of the genotype calls was carried out for each SNP, using the position of the sequenced genotype to adjust the corresponding cluster position.

Genotyping data were used to construct a neighbor-joining tree using the functions `dist.gene`, `nj`, and `compute.brln` of `ape` package v 5.0 (Paradis & Schliep, 2019) of R software, version 3.2.3, using the default parameters.

To identify the minimum number of SNPs to assign each accession to the proper varietal group, we used a pipeline starting with the R-cran package "adegenet" (Jombart & Ahmed, 2011) to validate the varietal groups calculating the `Fst` by "genet.dist"; the validated groups were imposed to the discriminant analysis of principal components (DAPC) (function `dapc`) to produce the subset of the SNP panels, keeping the most informative ones. To get the private alleles of the validated groups, the function "private_alleles" of the R-cran Package `poppr` (Kamvar et al., 2015) was applied to the dataset. To thin out the most informative SNPs from DAPC, one private allele for Navel, Moro, Tarocco, and Vaniglia groups was selected. The correlation with the groups selected the most informative ones for Sanguigno-Sanguinello and common-Valencia from DAPC analysis. The `assignplot` of the DAPC validated the SNPs pruning performed on the subsetted SNPs.

To define the unique combination of alleles across all loci of the multi-locus genotypes (MLGs) and identify possible accessions having the same genotype at all loci, the MLGs were computed by the `mlg.id` function of "poppr" package. Moreover, the minimum coverage network (MSN) was used to visualize the genetic relationships between accessions of oranges by the `msn()` function of "poppr" package. To cluster

the accessions without a priori assumption, the `find.clusters` function with Akaike information criterion (AIC) option of package `adegenet` was used.

2.9 | KASP genotyping on blood orange commercial varieties

A subset of 14 SNP sites, of which five were exclusive to the GoldenGate chip (chr5_27963361, chr2_27663272, chr4_2468242, chr5_14646862, and chr8_14495489) and nine were retrieved from the original source of SNPs (chr6_18850775, chr1_14405110, chr3_25755721, chr2_1340230, chr3_28292108, chr2_1098559, chr2_14040730, chr8_8297233, and chr3_10588995), were selected for KASP genotyping of 85 samples, of which 52 were from leaf and 33 from fresh juice (Table S4; Table S8). The KASP assays were designed following LGC Genomics KASP assay design, considering 70 bp surrounding the site of interest at the 5' and at the 3', marking surrounding polymorphisms with IUPAC codes, to optimize the primer design. Samples were collected in different nursery foundation blocks (NFBs), nurseries, and orchards to verify the reliability of SNP genotyping for plant material and fresh juice traceability. Detailed information regarding the leaf and juice samples and the sampling sites is reported in Table S2. Among the selected SNPs, 10 were accession-specific of the following varieties: Ippolito, Lempso, Meli, TDV; while four SNPs were common to Moro nuc. and Moro VCR (Table S4).

2.10 | RNA-Seq library preparation and sequencing

Total RNA was extracted from 3 mL of filtered juice from three biological replicates of Moro nuc., TDV, Cara Cara, and Van Biondo, as previously reported (Catalano et al., 2020). Total RNA was resuspended in 50 μ L of RNase-free water and quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific). The qualities and the quantities were evaluated using a Nanodrop 1000 spectrophotometer (Thermo Scientific) and gel electrophoresis (agarose 0.8% in TAE 1x). The quality was considered optimal for values of 260/280 between 1.80 and 2.0.

DNase treatment was carried out by adding to 40 mL of RNA 1x of RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), 0.1 M of dithiothreitol (Invitrogen), 5 m buffer, and 1x of DNase in a final volume of 50 mL. Samples were incubated at 37°C for 30 min and purified using the RNA Cleanup protocol (Qiagen), according to the manufacturer's protocol.

Libraries were prepared with 100 ng of total RNA using the Universal Plus mRNA protocol (Tecan) and sequenced on Illumina NovaSeq 6000 with paired reads of 150 bp.

2.11 | Allele-specific expression analysis

ASE was performed on the four varieties described in the previous paragraph (Moro nuc., TDV, Cara Cara, and Van Biondo), which will be named for brevity: Moro, Tarocco, Navel, and Vaniglia, respectively.

All the cultivars have the same MLG for each gene, composed of one haplotype corresponding to the sweet orange reference and of an alternative one. This facilitates the task of obtaining the phases of the two haplotypes, which is required for assessing the ASE (León-Novelo et al., 2018). Briefly, the alternative haplotype was obtained by substituting, in each polymorphic position of the cultivar, the alternative allele of heterozygous SNPs to the reference allele. Using this approach, two haplotypes were obtained for each sample, one corresponding to the reference and one to the alternative haplotype.

ASE of the two haplotypes was assessed using allelic imbalance metre (ALLIM) (Pandey et al., 2013) as previously described (Magris et al., 2021), briefly summarized below. The allelic imbalance ratio was expressed as the mean of three replicates (except Moro for which only two libraries were successfully analyzed) and the replicates were used for assessing the statistical significance using a Stouffer's meta-analysis with weight and direction effect, and a Benjamini-Hochberg correction was used to correct for multiple testing. Genes were considered as showing allelic imbalance when the adjusted *p*-value was lower than 0.05.

Genes with low coverage (less than 100 reads per cultivar) were marked as uninformative. Genes were considered as carrying a heterozygous VANDAL insertion if the element was located inside the gene or at a distance lower than 50 kb from the gene. Variation of ASE as a function of VANDAL insertion was tested by comparing genes with no insertion to genes carrying a VANDAL element.

As a further test, each gene in each clone was classified as "No VANDAL" if no VANDAL element was located inside the gene or outside at a distance lower than 50 kb, as "Out of exon" if the insertion of a VANDAL element occurred at a distance lower than 50 kb without affecting exons, and as "Exon" if the VANDAL insertion disrupted the gene exon. Differences in the distributions were tested using the Wilcoxon–Mann–Whitney test.

3 | RESULTS AND DISCUSSION

3.1 | Diversity among somatic mutants of sweet orange varieties

In fruit crops, somatic mutants originate from bud mutations and represent one of the most important resources for citrus breeding. The sweet orange collection of CREA counts

more than 200 somatic mutations belonging to varietal groups named common, Valencia, navel, blond, Vaniglia, Moro, Tarocco, Sanguigno, and Sanguinello (Caruso et al., 2016). We decided to perform the resequencing of 20 sweet orange accessions, choosing them among a subset of varieties that can be considered representative of the main traits (Table 1). These accessions belong to the typical varietal groups separated for the presence (Tarocco, Sanguigno, and Sanguinello) and the absence (all the remaining varieties including those containing lycopene) of anthocyanins in the fruit; we considered acidless (Vaniglia group) and low acidic (Tarocco Ferreri) accessions, differentiating also for the presence of lycopene or anthocyanins in the fruit; we also included varieties different for their origin, separating between old and nucellar lines; we selected samples among early, medium, and late varieties; we also considered the presence and absence of seeds in the fruit; both local (such as Tarocco clones, Moro clones, Sanguigno, Navel) and of foreign origin (Shamouti) varieties were represented. Most local accessions have a commercial interest and have a fair diffusion in some areas. It is noteworthy that the subset included Navel Cara Lindcove nuc. F8187, a nucellar line of a Navel Cara Cara that has lost the lycopene pigmentation of the flesh, and Moro non pigmentato Nuc., which is a nucellar line of Moro Nucellare 58-8D-1 lacking anthocyanins. Resequencing of both nucellar lines could represent a valuable resource for the understanding of the genetic control of pulp pigmentation.

3.2 | Sequencing and coverage analysis of 20 accessions of sweet oranges

More than 316 Gbp of raw data were generated from 20 accessions of sweet oranges, ranging from 9.74 Gbp for Ovale to 22.43 Gbp for Ferreri with an average of 15.84 Gbp/accession. After filtering adapters read-through and low-quality bases, an average of 13.58 Gbp per sample was retained for alignment on Csi_valencia_1.0, corresponding to an average coverage of 37x. After removing ambiguously mapped reads, unpaired or improperly paired mapping reads, and performing local re-alignment around InDels, median coverages of samples ranged from 22x for Ovale to 35x for Moro nuc. (Table S8).

3.3 | Somatic SNPs discriminate single and multiple varieties

Following filtering of Illumina reads, alignment, and SNP/INDEL calling via GATK Unified Genotyper, a total set of 7,244,768 uncharacterized polymorphic sites has been used as input to the filtering and re-calling pipeline described in Section 2, which implemented the retention of candidate somatic variants given a series of parameters

including coverage, quality, privateness, allelic ratio, and whitelisted genomic regions. This latter yielded a final set of 1169 SNPs (Table 2; Table S7). Of these, 837 were found in a single accession (accession-specific) out of the 20 analyzed, whereas the remainder were shared between multiple accessions or common to all representatives of one or more varietal types. Note that 586 SNPs were identified within the eight non-pigmented samples, including navel and common oranges, and 583 in the 12 analyzed pigmented oranges. Common orange Shamouti (also known as Jaffa orange) had the highest number of 114 accession-specific SNPs. InDels were not considered in this analysis to avoid the presence of abundant false calls due to the misalignment problems that often occur in the flanking regions of small repetitive elements (microsatellites and homopolymers).

A total of 230 SNPs are found in genic regions (introns and exons), of which 140 were distributed in the coding regions of 49 genes, including 82 non-synonymous, one splice site acceptor, one splice site_donor, five start gains, 11 stop gains, and 40 synonymous (Table S9). Overall, 771 were reported as transitions and 398 as transversions, with a Ts/Tv ratio of 1.94.

Figure 1(A) shows a neighbor-joining tree constructed with SNP data as recorded by cohort-wise genotyping (Euclidean distances are calculated with somatic SNP as dominant markers).

Pigmented and common oranges are separated with 14 sites shared across the eight common oranges and 39 for the pigmented oranges. The finding of seven shared mutations suggests an early common selection lineage for Shamouti and Vaniglia oranges. Similarly, Ovale shared a loss-of-heterozygosity (LOH) region on chromosome 7 with Navel types. It should be noted that three mutations shared between Navel, Vaniglia, Campbell, and Shamouti were not found in Ovale. Since Ovale was the accession with the lowest median coverage, we consider it likely that we failed to sequence the mutated allele on these sites, where the coverage in Ovale ranged from five to eight reads only; it is, therefore, reasonable to consider such sites as shared across the eight common oranges. In pigmented oranges, four SNP sites made it possible to highlight a common precursor for Tarocco and Sanguigno + Sanguinello types, while Moro types are depicted as an older selection lineage, with Moro NP sharing only two mutations with the other two Moro accessions (Moro VCR and Moro nuc.). Similarly, only one site was shared by Doppio Sang with the two Sanguinello (Moscato and Comune) types (Table S7). While specific varietal groups we defined a priori (Navels; Sanguigno + Sanguinello; Tarocco; Vaniglia) correspond to monophyletic groups, our analysis confirmed the polyphyletic origin of the clones classified as common (Campbell, Ovale, and Shamouti). This result is in agreement with previous bibliographic information (Hodgson, 1967) and resequencing data (L. Wang et al., 2021).

TABLE 2 Catalog of somatic single nucleotide mutations.

Private SNPs		Shared SNPs	Total SNPs
Common, navel, and acidless oranges			586
Van Biondo	40	Navel + Acidless + Common ^a	14
Van Sang	34	Acidless	42
Cara Cara	18	Acidless + Shamouti	7
Lane Late	39	Cara Cara (Navel Cara Cara + Navel Cara Cara Lindcove nuc. F8187)	34
Ovale	56	Acidless + Cara Cara (Navel Cara Cara + Navel Cara Cara Lindcove nuc. F8187) ^b	1
Shamouti	114	Navel + acidless + Campbell + Shamouti ^c	3
Campbell	91	Navel	90
		LOH ^d in Navel + Ovale (chr7: 25,148,001–25,188,014)	3
Pigmented oranges			583
Sanguinello	41	Tarocco + Moro + Sanguigno + Sanguinello	39
Moscato	23	Tarocco	20
Dal Muso	34	LOH in Tarocco ^d (chr5: 33,563,926–33,687,243)	17
Ferreri	33	Tarocco (w/o Dal Muso)	3
Ippolito	37	Tarocco + Sanguigno + Sanguinello	4
Meli	37	Sanguigno + Sanguinello	76
Lempso	40	Sanguinello	20
TDV	31	Moro	2
Moro nuc.	9	Moro (Moro nuc. + Moro VCR)	44
Moro NP	62		
Moro VCR	11		
Total	750	Total	419
			1169

Note: Shared SNPs are polymorphisms found in more than one orange clone.

Abbreviation: SNP, single nucleotide polymorphism.

^aOne mutation here counted (chr3_6587269) is absent in F8187 because originally bear by the cell layer lost during its nucellar propagation, while another mutation (chr3_6620528) is absent in Van Biondo;

^bThis mutation (chr3_6561168) has been lost in F8187 by its nucellar propagation.

^cThree sites of < Navel + Acidless + Campbell + Shamouti > (chr3_6603818, chr5_23287669, chrUn_44601499) are missing the somatic mutation for the Ovale accession.

^dLoss-of-heterozygosity (LOH) detected as the absence of an allele for a contiguous set of SNPs.

Sanger sequencing was performed to validate 76 accession-specific and group-specific SNPs. Considering the importance of clonal fingerprinting within the species, the SNPs were mostly selected among the accession-specific ones. We were able to validate 66 out of the 76 polymorphic loci (Table S4) chosen among heterozygous and homozygous mutations (Figure S4). The validation rate of 86.8% was slightly higher than those reported by L. Wang et al. (2021), who were able to validate 84% and 78% of SNPs in two independent validation assays. This confirms the reliability of the custom pipeline used for somatic SNP identification.

3.4 | Accession-specific and varietal-group InDels

A total of 59 accession-specific small InDels of 1–18 bp has been identified, generally between 1 and 8 mutations for each variety, except TDV; the other 16 mutations are shared among

accessions belonging to the same or different varietal groups (Table S10). Note that 58.6% of all the 59 InDels were located within intergenic regions, while the remaining 41.4% fell in the gene bodies. Of the latter, 11 InDels were found in the coding sequence and only one reported an insertion of three nucleotides; all the others are deemed to cause a frameshift mutation and potentially loss of function. Thirteen accession-specific InDels were analyzed through HRM analysis. Eleven of the 13 were correctly validated in the specific accession by comparing it with the other two varieties used as negative controls (Figure S5).

3.5 | Analysis of somatic transposable element-related insertions

The analysis of mobile element insertions across the cohort of 20 accessions allowed us to identify and genotype 234 sites transposable element-related insertions, while 36 calls

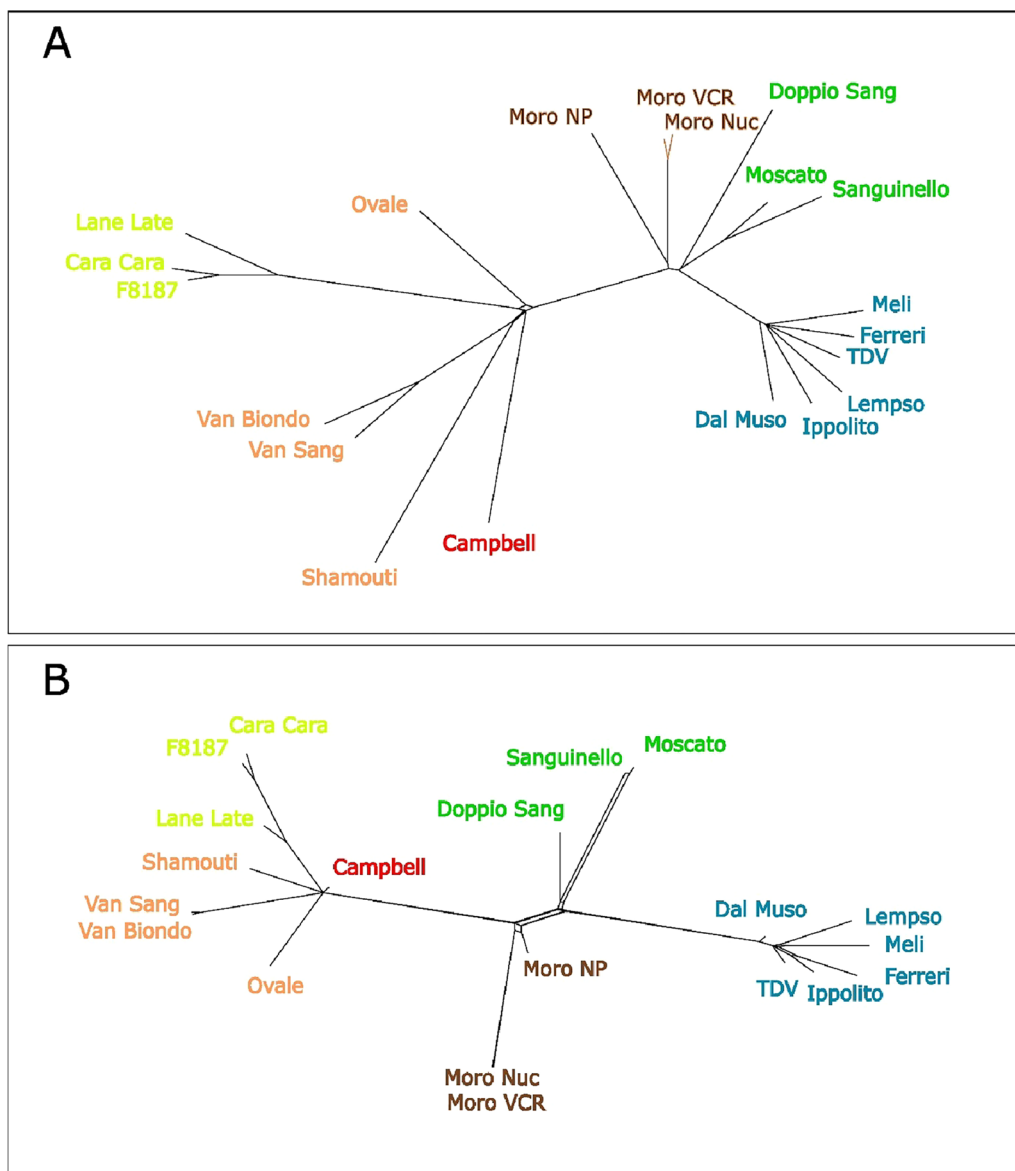


FIGURE 1 Neighbor-joining tree constructed with (A) single nucleotide polymorphism (SNP) data and (B) insertion of structural variants data. Different colors indicate the varietal groups: Navel types (yellow), Valencia (red), Sanguigno types (green), Tarocco types (light blue), Moro types (brown), and other common oranges (orange).

identified by the bioinformatics pipeline were excluded due to missing data or inconsistent patterns based on established grouping of sweet oranges by SNP analysis and field information (Table S11). Of 234 single-nucleotide variants, 21 were common to all pigmented varietal groups (Tarocco, Moro, and Sanguigno + Sanguinello). Tarocco and Sanguigno + Sanguinello types shared six additional events of transposable element-related insertions (Figure 1B) indicating the presence of a common ancestor, in agreement with the SNP analysis (Figure 1A). Moreover, Tarocco accessions exclusively shared 35 insertions, two of which were not present in Dal Muso, confirming (as suggested by SNP data) that this is an older selection compared to the other Tarocco accessions in this panel. Ippolito and Ferreri shared three mobile element

insertions, suggesting a common origin, and similarly, one insertion was shared between Lempso and Meli. For Moro, most insertions were shared between Moro nuc. and Moro VCR, which confirmed the same selection topology, as suggested by the SNPs analysis. Similarly, Doppio Sang shared two insertions with the two Sanguigno + Sanguinello types, indicating a common distant ancestry as evident in SNP data (Figure 2).

Unlike in pigmented oranges, in common ones, any insertion is shared across all sub-groups. In the group of “non-Vaniglia” and “non-Navel” pigmented oranges, the absence of shared TE insertions (i.e., all are private to one or some specific accession) precluded any hint about ancestry and confirmed the observation based on SNPs. In Navel types, Cara

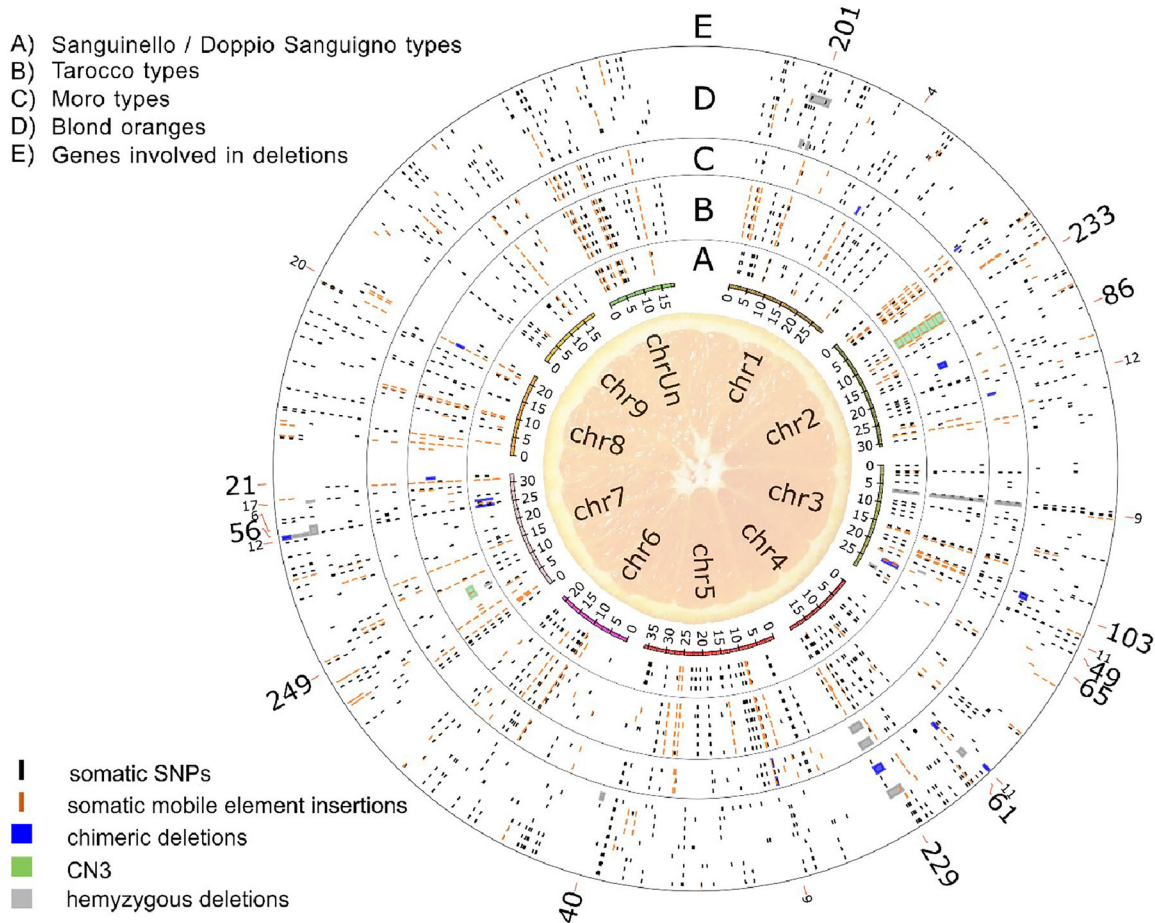


FIGURE 2 Circos representation of the somatic mutation landscape found through sequencing of 20 representative Italian sweet oranges. “A–D” circles contain the accession of the four major groups (Sanguinello/Sanguigno, Tarocco, Moro, and all blond oranges, respectively). Circle “E” reports the number of genes involved in large somatic deletions. SNP, single nucleotide polymorphism.

Cara and F8187 had the highest shared number of insertions as did the two Vaniglia accessions. Lane Late with nine insertions in common with the other two Navel types confirmed its selection divergence (Figure 1B), as shown by the SNP data (Figure 1A). Overall, the two topologies presented by SNP data and insertion analysis independently corroborate each other, confirming the reliability of the somatic mutation discovery pipeline for lineage tracking purposes.

The annotation of TEs by sequence similarity provided evidence that the integration landscape consists of two main families of elements. Forty-seven events of transposable element-related insertions determined by Class I long terminal repeat (LTR) retrotransposon of the Gypsy (34) and Copia (13) superfamily were mostly common to either all pigmented oranges or Navel type or Acidless types and Shamouti. In contrast, 181 insertions were due to the activity of members of the Class II DNA transposon VANDAL (mutator-like) family, for which an anti-silencing mechanism operated by the VANC protein family has been shown in *Arabidopsis* to operate a demethylation of such elements and their activation (Hosaka et al., 2017). VANDAL elements are peculiar because they

are autonomous but apparently lack entirely or have short and degenerate inverted terminal repeats (Kapitonov & Jurka, 2000). We discovered a higher proportion of integrations near transcription start sites (TSSs) for VANDAL-like sequences compared with all the other TE families detected (Figure 3). This agrees with previous reports of a strong insertional preference for VANDAL elements in *Arabidopsis* near TSS (Fu et al., 2013). Also, VANDAL insertions are overrepresented in occurrences identified in a single accession (private) or covering only partially a varietal group. Only the Tarocco group, represented by six samples, shows a higher number of VANDAL than Class I retrotransposons (Copia and Gypsy). On the contrary, VANDAL occurrences shared across all pigmented oranges do not show the same level of enrichment but Gypsy and Copia all together make most of the somatic insertions defining this macro-group.

On a time scale, since most VANDAL insertion events are reported in single accessions, partial groups, and Tarocco group, whereas common oranges groups and the aggregation of all pigmented oranges show a prevalence of Gypsy/Copia insertions, this can indicate a more intense activity of these

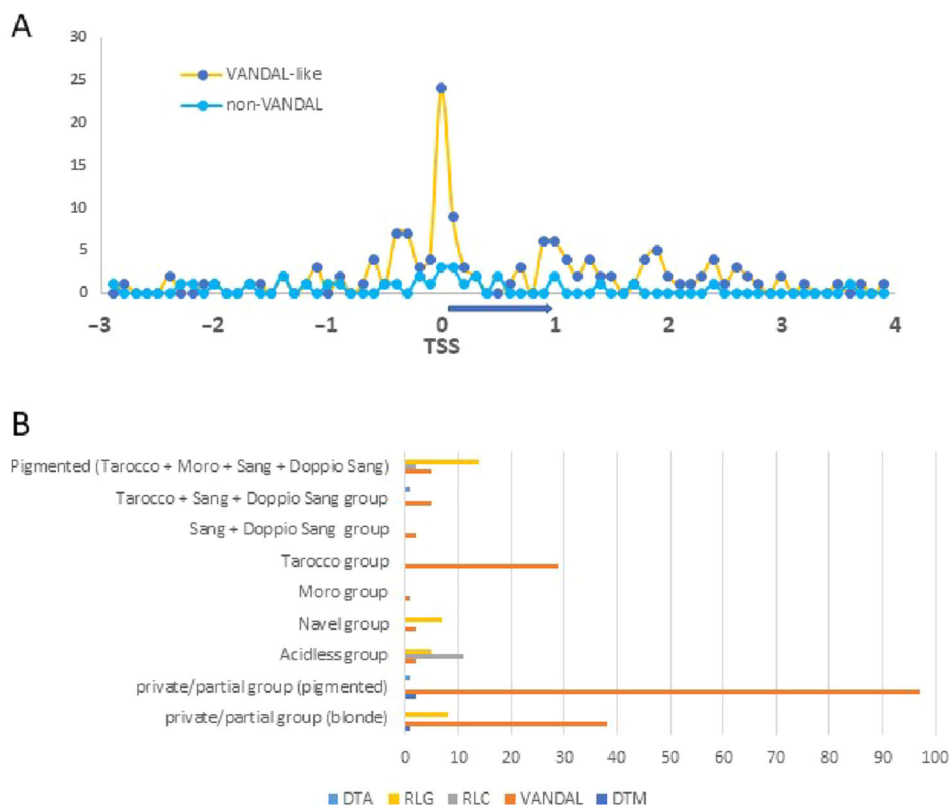


FIGURE 3 Transposable element-related insertions. (A) Relative position of somatic integration sites of VANDAL elements versus non-VANDAL elements. Gene body lengths are scaled from 0 (transcription start site [TSS]) to 1 on the x -axis, while the upstream and downstream regions are represented as 1 kbp per unit. (B) Distribution of mobile insertions by group-specific, private, or present only on a fraction of a varietal group. DTA, hAT; DTM, mutator-like; RLC, Copsya-like; RLG, gypsy-like.

Class I elements in the past; DNA transposons, and specifically, VANDAL-like sequences showed a more recent activity in more recent clonal selections.

A total of 42 insertions, of which 24 were accession-specific, were selected for validation by PCR. Each PCR was carried out by two independent amplicons (one for each side of the mobile element). The validation rate was 80.9%. L. Wang et al. (2021) reported a higher rate of PCR validation of TE insertions, although they validated shared insertions. We failed to amplify both amplicons in three putatively specific TE of Ippolito, two TE specific to Meli, and one specific to TDV; we could not validate two putative group-specific insertions shared among all Tarocco selections (Figure S6). The remainder PCRs fully validated the specificity of the in silico insertion genotyping method with at least one primer combination (Table S6).

3.6 | VANDAL insertions affect the expression of flanking genes

Insertions of mobile elements proximal to genes may be responsible for ASE (Marroni et al., 2014). We thus investigated if VANDAL elements in or near genes elicited this phenomenon. We performed ASE using RNA sequencing

(RNA-Seq) data from four accessions (Moro nuc., Tarocco TDV, Navel Cara Cara, and Vaniglia Biondo), and we analyzed a total of 80 VANDAL insertions, of which 23 were in exons. Genes located at a distance lower than 50 kb from a VANDAL element showed higher levels of ASE in terms of statistical significance and of magnitude than genes without VANDAL insertions (Figure S7). We also hypothesized that genes affected by a VANDAL insertion in the exonic regions may exhibit higher levels of ASE if the TE insertion results in frameshift mutations that in turn trigger nonsense-mediated mRNA decay (Mendell et al., 2004; Raxwal & Riha, 2023). Our results suggest that the insertion of VANDAL in the gene exons significantly affected the level of ASE, both measured as p -value and as \log_2 ratio (Figure 4). Somatic insertions of VANDAL elements that occur preferentially near the TSS do appear to be able to affect gene expression in an allele-specific manner.

3.7 | Analysis of somatic large structural variants

Deviations from the expected equal frequency for the two alleles in heterozygous SNP loci over relatively large

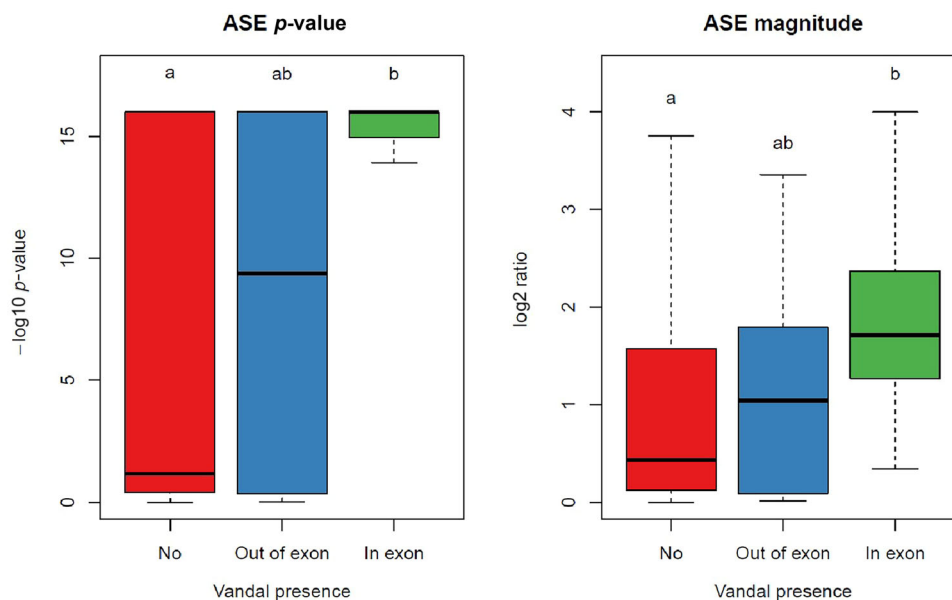


FIGURE 4 Effect of the presence of somatic VANDAL insertions on allele-specific expression in terms of statistical significance, measured as $-\log_{10}$ of the p -value (left), and in terms of magnitude, measured as the absolute value of the \log_2 ratio of the relative expression of the two haplotypes (right). “No”: genes not interested by a VANDAL insertion. “Out of exon”: genes closer than 50 kb to a VANDAL insertion, but for which the insertion is not located in an exon. “In exon”: genes for which a VANDAL insertion occurred in an exon. Different letters indicate statistical significance according to Wilcoxon–Mann–Whitney test. ASE, allele-specific expression.

chromosomal regions can be used to detect mosaic structural variants using the X-scan software (Marroni et al., 2017), with a complete LOH indicating a fixed structural variant. We used this approach to identify chimeric deletions and explore a third type of somaclonal mutation in the 20 sweet oranges. Table 3 summarizes the large structural variants we found along with an indication of their nature (hemizygous deletion, chimeric deletion, or copy number variant) derived by combining the reduction of heterozygosity analysis results with the depth-of-coverage analysis of the regions involved. All the private deletions found as hemizygous were identified in clones derived by nucellar embryogenesis, as expected when those variants occurred before the nucellar propagation.

The double hemizygous deletion of Valencia Campbell in chromosome 1 has been checked to verify if it could be due to a misassembly issue in the reference sequence; an independent analysis using as reference the *C. clementine* genome confirmed that the two deletions are not juxtaposed. A 330 kb hemizygous deletion on chromosome 3 appears to be shared among all the 11 pigmented oranges (Table 3), which further consolidates the hypothesis of a completely independent selection lineage for the anthocyanin-pigmented oranges. Instead, a 1.7 Mb copy number variant (three copies) was found only in Tarocco clones. Only one event, shared between Moro VCR and Moro nuc., was not classified due to its short physical size (depth of coverage analysis and manual inspection of alignments were also unsuccessful in categorizing such signal). However, the latter (together with

another large hemizygous deletion shared between the same two clones [1.9 Mb]) corroborates the diversity patterns also found in SNP and insertion data. More interestingly, the same genomic region of chromosome 4, where we found the hemizygous deletion in Moro VCR and Moro nuc., was involved in two other independent deletions in unrelated clonal selections with different starting points (left-most coordinate to the reference) and identical terminal breakpoint (peritelomeric). The starting point of a hemizygous deletion in common orange Shamouti was some 65 kb upstream of Moro’s one, while Van Biondo showed a chimeric deletion with a starting breakpoint some 600 kb closer to the telomere (Figure 5). This could point to the existence of fragile sites along the chromosomes where recurrent deletions occur.

3.8 | Illumina GoldenGate, an approach to genotyping of sweet orange germplasm

Out of 768 SNP positions tested, 498 were left after the removal of genotyping failures (137) and monomorphic sites (133). Wherever possible, manual correction of clusters for calls has been applied considering the potential chimeric nature of some of these mutations, which would generate unbalanced intensities for the two alleles (i.e., chimeric SNPs are closer to homozygous calls than true heterozygous ones) (Figure S8). Likely, the higher validation rate we observed

TABLE 3 List of large deletions/CNV found by allele frequency imbalance.

Chromosome	Start	Stop	Size	Type	Clone(s)
chr1	6,223,238	8,765,559	2,542,322	H	Ovale
chr1	6,705,607	6,969,702	264,096	H	Campbell
chr1	7,780,439	8,004,214	223,776	H	Campbell
chr1	19,310,486	19,365,970	55,485	C	Moro VCR
chr2	1,328,704	1,507,656	178,953	C	Campbell
chr2	7,819,940	9,606,679	1,786,740	CNV = 3	All Tarocco analyzed (6)
chr2	16,212,221	17,075,956	863,736	C	Ippolito
chr2	24,178,466	24,268,944	90,479	C	Moro VCR
chr3	6,409,623	6,739,792	330,170	H	All Tarocco, Moro, and Sanguigno + Sanguinello analyzed (11)
chr3	19,359,467	20,051,614	692,148	C	Van Biondo
chr3	22,697,634	22,760,398	62,765	H	Lempso
chr3	23,790,150	24,204,815	414,666	C	Sanguinello and Moscato
chr3	26,213,440	26,658,955	445,516	H	Doppio Sanguigno
chr4	7,606,699	7,788,263	181,565	C	Cara Cara
chr4	8,257,455	8,906,853	649,399	H	Ovale
chr4	8,711,780	8,932,861	221,082	C	Van Biondo
chr4	17,697,569	19,592,000	1,894,432	H ^a	Shamouti
chr4	17,763,020	19,592,000	1,828,981	H ^a	Moro VCR and Moro nuc.
chr4	18,372,427	19,592,000	1,219,574	C ^a	Van Biondo
chr5	8,861,746	8,927,616	65,871	? ^b	Moro VCR and Moro nuc.
chr5	35,606,772	36,049,614	442,843	H	Campbell
chr7	7,248,051	9,369,668	2,121,618	CNV = 3	Meli
chr7	8,050,210	8,110,014	59,805	H	Tarocco TDV
chr7	24,840,261	24,970,488	130,228	C	Sanguinello and Moscato
chr7	25,063,719	26,166,735	1,103,017	H	Ovale
chr7	25,063,781	25,352,288	288,508	C/H	Cara Cara: C F8187 and Lane Late: H
chr7	26,261,242	26,329,987	68,746	C	Sanguinello and Moscato
chr7	29,297,202	29,387,740	90,539	H	Ovale
chr7	31,698,783	31,854,240	155,458	C	Ippolito
chr8	20,442,860	20,605,831	162,972	C	Ippolito

Abbreviations: C, chimeric deletion; H, hemizygous deletion.

^aCoordinates have been manually revised by inspection of alignments, terminal breakpoint is indicative and it is not excluded a scaffold misplacement on chromosome end masking a complete terminal deletion or chromosomal replacement.

^bToo short span did not allow to classify this mutational event (either hemizygous deletion or CNV).

in Sanger compared to GoldenGate could be due to the selection of SNPs with higher quality scores. A total of 223 leaves and 18 juice samples belonging to 223 accessions were successfully genotyped with the SNP panel across two 384 SNP plates. The 241 samples of leaves and juice were represented by 10 Valencia, four Vaniglia, 41 Navel, and 34 other common oranges along with 15 Moro, 28 Sanguigno/Sanguinello, 106 Tarocco, and three other accessions for the blood oranges (Table S1; Table S12).

The neighbor-joining tree discriminated the accessions into groups and subgroups, reflecting their ancestry and origin (Figure 6). In particular, by using the set of SNPs represented in the GoldenGate, we were able to discriminate the follow-

ing groups and subgroups: Navel, Moro, Tarocco, Vaniglia, Common, Valencia, and Sanguigno + Sanguinello.

The population differentiation measured as *F*_{st} correctly assigned the varieties to the proper varietal groups (Figure S9). In one case, it was not possible to discriminate varietal subgroups. Specifically, Sanguigno and Sanguinello accessions did not form different clusters (Figure S9A,B). The two subgroups share fruits with similar pomological traits and differ in the presence of seeds (Barry et al., 2020). The lack of diversification between these subgroups might be due to the few resequenced accessions in our dataset.

By combining the groups obtained on the DAPC and private_alleles analysis, we identified the following 10 SNPs as

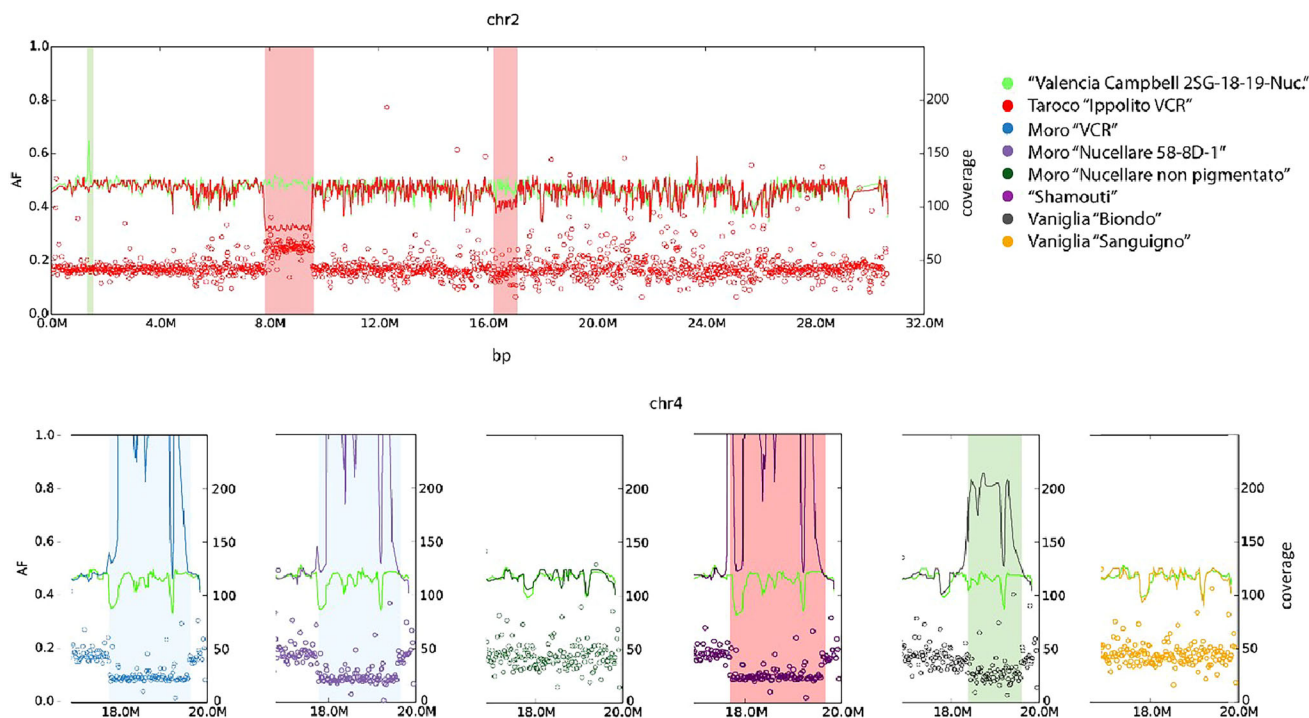


FIGURE 5 X-scan analyses: solid lines indicate the alternative allele frequency (Campbell represents the control sample); dots indicate the coverage. Upper panel: example of the coexistence of both hemizygous and chimeric deletion on chromosome 2 of Ippolito. Lower panel: a series of X-scan analyses in the same segment of chromosome 4 of different accessions indicating independent deletions. A shared hemizygous deletion in Moro VCR and Moro nuc., yet absent in “Moro non pigmentato,” an extended version of the same deletion in Shamouti and a chimeric—shorter—deletion in Van Biondo.

the minimum number of analyzed loci capable to discriminate varietal groups and subgroups (Table S13): 106_GP_IGA_va, 065_GP_IGA_tabd, 001_GP_IGA_mo, 069_GP_IGA_tabd, 067_GP_IGA_tabd, 033_GP_IGA_na, 1024_GP_IGA_SAN, 1032_GP_IGA_SAN, 1029_GP_IGA_SAN, and 1327_SP_IGA_valCAMP (Table S14). This minimum SNP subset could be useful to assign unknown clones, leaves (Figure S10A–D), or juice samples (Figure S10E) to specific varietal groups. The low value of F_{st} (Figure S9B) between the Tarocco and Sanguigno + Sanguinello groups (0.31) could reinforce the idea that these two groups differentiated recently (Casella, 1935).

By using both plates of SNPs, not all the genotypes could be differentiated; Table S14 shows the eight nonunique profiles obtained for 24 genotypes. AIC clustering function generated 15 groups or clusters (Table S14). Ten of the 15 groups are characterized by private alleles shared among all accessions within each group. These clusters reflect the varietal and clonal groups. The presence of small clusters, including few accessions, is influenced by ascertainment bias arising from the limited SNP calling to the 20 resequenced genomes. As reported in Table S15, “C1” is synonymous with Sanguigno + Sanguinello, and all the accessions of this group shared a common private allele, the T of 1029_GP_IGA_SAN; “C3” groups together all the TDV clones (old line and different

nucellar selections), sharing 12 private alleles; “C5” is made of a group of eight Moro accessions that shared 12 private alleles. Interestingly, “C7” grouped four Moro nucellar lines and one Moro mutant, all showing very low or absent anthocyanin accumulation (Table S16), and shared 15 private alleles; “C8” is synonymous with Vaniglia containing four accessions and 29 private alleles; “C10” groups together the two Dal Muso sharing 17 private alleles; “C11” contains all the VA, sharing 15 private alleles; “C12” contains all the Lane Late clones (from nucellar selection and shoot-tip grafting) plus Chislett and Powell Navels, with 8 private alleles; “C13” groups the Cara Cara clones (two old lines with lycopene-pigmented pulp and four nucellar lines without lycopene), all characterized by 14 private alleles; “C14” groups the Meli clones with five private alleles; C15 groups the Lempsso clones with 14 private alleles. Table S14 lists the numbers and names of the private alleles for each group.

The MSN analysis (Figure 7) helps to build the mutational history of sweet orange varietal groups, supporting previous information regarding many of the analyzed accessions and revealing similarities that were not known. The Sanguigno + Sanguinello, Moro, Vaniglia, and Navel originate from the common oranges. While Sanguigno + Sanguinello and Moro seem to derive from different ancestral clones, it seems clear that Tarocco branched from Sanguigno + Sanguinello, as

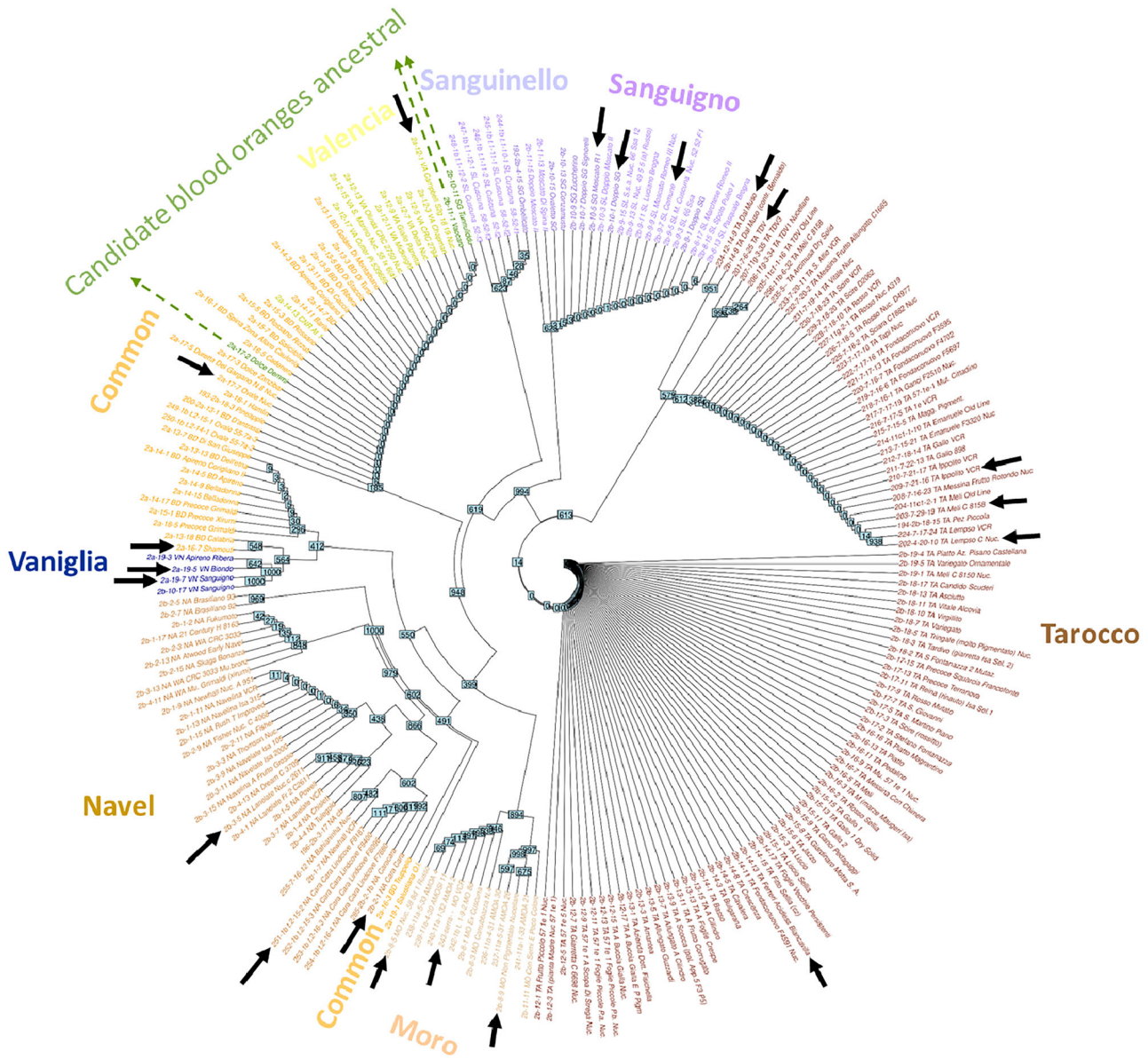


FIGURE 6 Circular tree of all the genotypes analyzed through the Illumina GoldenGate showing the bootstrap for each node, and color represents the varietal groups and subgroups: ● Tarocco, ● Moro, ● Common, ● Navel, ● Vaniglia, ● Candidate blood oranges ancestral (interrupted arrows), ● Valencia, ● Sanguinello, and ● Sanguigno. Arrows indicate the accessions whose genome has been sequenced in the present manuscript.

previously reported (Casella, 1935). The link of Tarocco to Sanguigno + Sanguinello appears to be due to Tarocco Liscio Sellia and Ovaletto Sanguigno. The Vaniglia group shows the closest similarity to the common selection Biondo dell’Etna. The origin of Navel oranges is still debated (Barry et al., 2020); from our data, they could derive from the Salustiana or a very close accession not included in the CREA collection. Within the Navel group, we observed a subcluster made of late-maturing clones. Although Chislett and Powell were described as bud sports of Washington Navel (patent numbers USPP8212P and USPP6733P, respectively; Edwards, 1993, 1995), our data reveal a common origin with Lane late.

3.9 | Sanger sequencing of the Ruby D allele elucidated the ancestry of pigmented oranges

To try to elucidate the clustering of pigmented oranges Vaccaro, Tunnuliddu, and Dolce Demmi into the branch of common oranges (Figure 6), Sanger sequencing of these accessions, in addition to Moro and Tarocco, was performed for the Ruby gene, confirming the presence of the RD alleles (Supporting Information Dataset 1). The RD alleles consist of RD-1, which includes the entire Tcs1 located upstream to the Ruby gene, and RD-2, having the solo-3’ LTR of Tcs1 and Ruby (Butelli et al., 2012). The RD-2 is a consequence

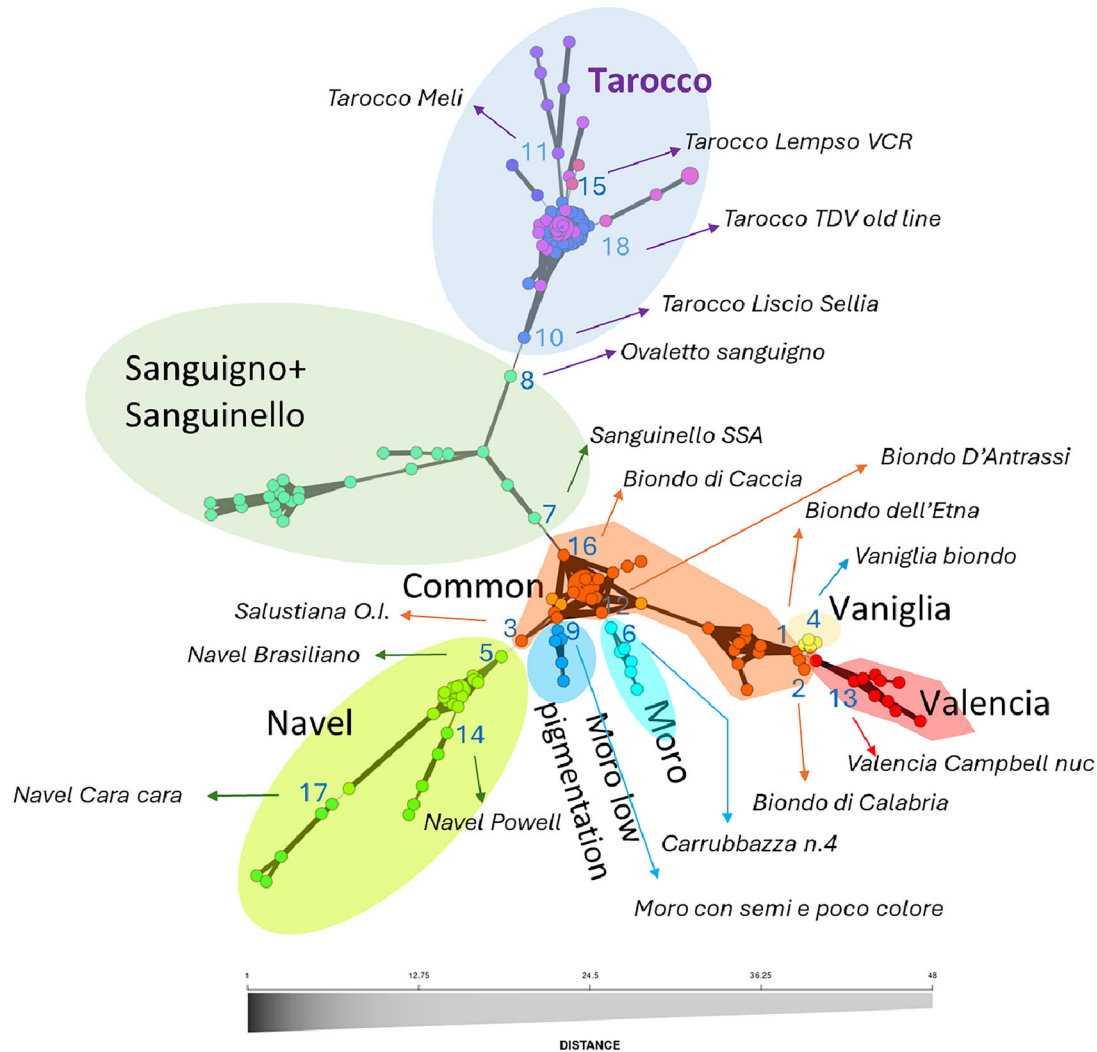


FIGURE 7 Minimum coverage networks (MSN) to visualize the relationships among accessions of oranges. It will represent the clonal evolution of the current membership groups of oranges.

of RD-1 as the excision of Tcs1. The presence of the RD-1 allele, the ancestral insertional event, in Vaccaro, Tunnuliddu, and Dolce Demmi is consistent with what has been observed in Tarocco (Butelli et al., 2012), demonstrating that these accessions can be considered the candidate ancestral accessions for the Mediterranean pigmented oranges (Figure S11). Moreover, the retrotransposon insertion represents the first event that can discriminate common from pigmented oranges. Consequently, while Tunnuliddu and Vaccaro are still characterized by a common orange genetic background (Figure 6; Figure S11), the presence of Tcs1 insertion can explain the pigmentation and indicate these accessions as early selection events in the pigmented oranges lineage. We can hypothesize that the lack of SNPs specific to pigmented oranges is due to the time in which this selection was generated and maintained.

3.10 | KASP assay for plant and juice traceability

A KASP assay was used to assess the genotyping consistency of blood oranges from different sites of cultivation, nurseries, and NFBs of four different Italian regions; it was performed to verify the reliability of a set of cultivar-specific or group-specific SNPs to implement traceability protocols for plants and juices. For this purpose, we selected nine SNPs specific for four of the most recent and highly propagated Tarocco cultivars, namely, Ippolito (two SNPs), Lempso (three SNPs), Meli (two SNPs), and TDV (two SNPs). We also selected five SNPs shared by Moro nuc. and Moro VCR and a specific SNP for Moro nuc. (Table S2).

The nine Tarocco accession-specific polymorphic loci were validated in all leaf and juice samples of the four cultivars;

these markers are adequate for traceability of specific commercial clones of relatively recent origin (Table S4; Table S2). In the case of Moro, genotyping was not always successful: Chr2_14040730 was correctly validated in all the 22 Moro samples, while the three other SNPs (chr5_14646862, chr8_8297233, and chr8_14495489) were validated in samples coming from NFBs, nurseries, and the CREA germplasm collection, where plant material traceability is easier. The validation failed for most samples labeled as Moro nuc., Moro VCR, or generically as Moro, coming from different orchards. It could be due to the presence of cultivar populations rather than single clones in cultivation, so the SNPs identified in the genomes of Moro nuc. and Moro VCR are not shared with other Moro cultivated clones. The low validation rate might also be due to mislabeling or wrong assignments of some samples to specific cultivated clones. Interestingly, the Moro nuc.-specific SNP (chr3_105889959) helped genotype the CREA sample only, but not any other Moro samples. This likely indicates that this specific mutation occurred on the reference plant of the CREA germplasm.

4 | CONCLUSIONS

The study offers new perspectives in the understanding of sweet orange diversity. The identified somatic variants efficiently separated the resequenced and genotyped cultivars into the known cultivar groups and subgroups; they also revealed the lineage of many cultivars, including Italian accessions. An excellent agreement was observed in the grouping inferences obtained using variants differing in mutational mechanisms and rates such as SNPs, TE-related insertions, and large SVs. More importantly, the present study allowed cultivar fingerprinting of leaf and juice samples using different techniques and strategies, from high-throughput genotyping to PCR, HRM, and Sanger sequencing, which require conventional laboratory equipment. The developed markers could be used for juice traceability, to unambiguously identify commercial clones in the framework of the plant certification program, and they might be included as Supporting Information for plant variety protection. In perspective, the developed tools might be useful in associating the genotypes with interesting pomological traits.

AUTHOR CONTRIBUTIONS

Davide Scaglione: Data curation; formal analysis; methodology; resources; software; writing—original draft; writing—review and editing. **Angelo Ciacciulli:** Data curation; formal analysis; methodology; writing—review and editing. **Stefano Gattolin:** Methodology; writing—original draft; writing—review and editing. **Marco Caruso:** Investigation; resources; validation; writing—original draft; writing—review and editing. **Fabio Marroni:** Formal analysis; method-

ology; software; writing—review and editing. **Giuseppina Las Casas:** Validation; writing—review and editing. **Irena Jurman:** Resources; writing—review and editing. **Grazia Licciardello:** Writing—review and editing. **Antonino Felice Catara:** Funding acquisition; writing—review and editing. **Laura Rossini:** Investigation; resources; supervision; writing—original draft; writing—review and editing. **Concetta Licciardello:** Investigation; resources; supervision; validation; visualization; writing—original draft; writing—review and editing. **Michele Morgante:** Conceptualization; investigation; project administration; supervision; writing—review and editing.

ACKNOWLEDGMENTS

We thank Giuseppe Reforgiato for his essential support in suggesting the choice of plant material and his helpful discussions on this manuscript.

This work has been funded by IT Citrus genomics—Genomica funzionale miglioramento genetico ed innovazione per la valorizzazione dei prodotti della filiera agricola (PON–D.M. 01/Ric del 18 gennaio 2010), “Qualitrace—Messa a punto e validazione di tool genetici e chimici per la tracciabilità integrata e la valorizzazione della qualità dell’Arancia Rossa di Sicilia IGP” (MIPAAF DM19527/7303/2016), and “NOVARANCIA—Innovazioni tecnologiche (genetiche, fitosanitarie ed agronomiche) per la valorizzazione e tracciabilità dell’Arancia Rossa di Sicilia” (PSR Misura 16.1. D.D.S. n.215/2022).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the BioProject PRJNA1120371.

ORCID

Davide Scaglione  <https://orcid.org/0000-0002-8930-2392>

Angelo Ciacciulli  <https://orcid.org/0000-0002-5410-437X>

Stefano Gattolin  <https://orcid.org/0000-0003-3855-9160>

Marco Caruso  <https://orcid.org/0000-0001-9068-4490>

Fabio Marroni  <https://orcid.org/0000-0002-1556-5907>

Giuseppina Las Casas  <https://orcid.org/0000-0003-2632-3820>

Irena Jurman  <https://orcid.org/0009-0008-8583-2119>

Grazia Licciardello  <https://orcid.org/0000-0002-2846-9009>

Antonino Felice Catara  <https://orcid.org/0000-0003-0605-8206>

Laura Rossini  <https://orcid.org/0000-0001-6509-9177>

Concetta Licciardello  <https://orcid.org/0000-0003-1380-6540>

Michele Morgante  <https://orcid.org/0000-0003-2591-2156>

REFERENCES

- Barry, G. H., Caruso, M., & Gmitter, F. G. (2020). Commercial scion varieties. In M. Talon, C. Caruso, & F. G. Gmitter, Jr. (Eds.), *The genus Citrus* (pp. 83–104). Elsevier. <https://doi.org/10.1016/B978-0-12-812163-4.00005-X>
- Butelli, E., Licciardello, C., Ramadugu, C., Durand-Hulak, M., Celant, A., Reforgiato Recupero, G., Froelicher, Y., & Martin, C. (2019). Noemi controls production of flavonoid pigments and fruit acidity and illustrates the domestication routes of modern citrus varieties. *Current Biology*, *29*, 158–164. <https://doi.org/10.1016/j.cub.2018.11.040>
- Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato Recupero, G., & Martin, C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell*, *24*, 1242–125. <https://doi.org/10.1105/tpc.111.095232>
- Carbonell-Bejerano, P., Royo, C., Torres-Pérez, R., Grimplet, J., Fernandez, L., Franco-Zorrilla, J. M., Lijavetzky, D., Baroja, E., Martínez, J., García-Escudero, E., Ibáñez, J., & Martínez-Zapatera, J. M. (2017). Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiology*, *175*, 786–801. <https://doi.org/10.1104/pp.17.00715>
- Cardone, M. F., D'Addabbo, P., Alkan, C., Bergamini, C., Catacchio, C. R., Anaclerio, F., Chiatante, G., Marra, A., Giannuzzi, G., Perniola, R., Ventura, M., & Antonacci, D. (2016). Inter-varietal structural variation in grapevine genomes. *The Plant Journal*, *88*, 648–661. <https://doi.org/10.1111/tpj.13274>
- Caruso, M., Distefano, G., Pietro Paolo, D., La Malfa, S., Russo, G., Gentile, A., & Recupero, G. R. (2014). High resolution melting analysis for early identification of citrus hybrids: A reliable tool to overcome the limitations of morphological markers and assist rootstock breeding. *Scientia Horticulturae*, *180*, 199–206. <https://doi.org/10.1016/j.scienta.2014.10.024>
- Caruso, M., Ferlito, F., Licciardello, C., Allegra, M., Strano, M. C., Di Silvestro, S., Russo, M. P., Pietro Paolo, D., Caruso, P., Las Casas, G., Stagno, F., Torrissi, B., Rocuzzo, G., Reforgiato Recupero, G., & Russo, G. (2016). Pomological diversity of the Italian blood orange germplasm. *Scientia Horticulturae*, *213*, 331–339. <https://doi.org/10.1016/j.scienta.2016.10.044>
- Casella, D. (1935). L'agrumicoltura siciliana. *Acireale: R. Stazione Sperimentale di Frutticoltura e di Agrumicoltura*, *2*, 1–147.
- Catalano, C., Ciacciulli, A., Salonia, F., Russo, M. P., Caruso, P., Caruso, M., Russo, G., Distefano, G., & Licciardello, C. (2020). Target-genes reveal species and genotypic specificity of anthocyanin pigmentation in citrus and related genera. *Genes*, *11*(7), 807. <https://doi.org/10.3390/genes11070807>
- Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS ONE*, *8*(12), e85024. <https://doi.org/10.1371/journal.pone.0085024>
- Deng, X., Yang, X., Yamamoto, M., & Kumar Biswas, M. (2020). Domestication and history. In M. Talon, C. Caruso, & F. G. Gmitter (Eds.), *The genus Citrus* (pp. 33–55). Elsevier. <https://doi.org/10.1016/C2016-0-02375-6>
- Edwards, M. (1993). *Plant Varieties Journal*, *6*(2), 6. Plant Breeders Rights Australia.
- Edwards, M. (1995). *Plant Varieties Journal*, *8*(2), 27–28. Plant Breeders Rights Australia.
- Falchi, R., Vendramin, E., Zanon, L., Scalabrin, S., Cipriani, G., Verde, I., Vizzotto, G., & Morgante, M. (2013). Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. *Plant Journal*, *76*(2), 175–187. <https://doi.org/10.1111/tpj.12283>
- Ferrari, G. B. (1646). *Hesperides sive Malorum Aureorum cultura et usu*. Sumptibus Hermannii Scheus.
- Fu, Y., Kawabe, A., Etcheverry, M., Ito, T., Toyoda, A., Fujiyama, A., Colot, V., Tarutani, Y., & Kakutani, T. (2013). Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. *The EMBO Journal*, *32*, 2407–2417. <https://doi.org/10.1038/emboj.2013.169>
- Hazzouri, K. M., Flowers, J. M., Visser, H., Khierallah Hussam, S. M., Rosas, U., Pham, G. M., Meyer, R. S., Johansen Caryn, K., Fresquez Zoë, A., Masmoudi, K., Haider, N., El Kadri, N., Idaghdour, Y., Malek Joel, A., Thirkhill, D., Markhand Ghulam, S., Krueger, R. R., Zaid, A., & Purugganan, M. D. (2015). Whole genome resequencing of date palms yields insights into diversification of a fruit tree crop. *Nature Communications*, *6*, Article 8824. <https://doi.org/10.1038/ncomms9824>
- Heng, L. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. <https://doi.org/10.48550/arXiv.1303.3997>
- Hocquigny, S., Pelsy, F., Dumas, V., Kindt, S., Heloir, M.-C., & Merdinoglu, D. (2004). Diversification within grapevine cultivars goes through chimeric states. *Genome*, *47*(3), 579–589. <https://doi.org/10.1139/g04-006>
- Hodgson, R. W. (1967). Horticultural varieties of citrus. In W. Reuther, H. J. Webber, & L. D. Batchelor (Eds.), *The citrus industry* (Vol. 1, pp. 431–591). University of California Press.
- Hosaka, A., Saito, R., Takashima, K., Sasaki, T., Fu, Y., Kawabe, A., Ito, T., Toyoda, A., Fujiyama, A., Tarutani, Y., & Kakutani, T. (2017). Evolution of sequence-specific anti-silencing systems in *Arabidopsis*. *Nature Communications*, *8*(1), Article 2161. <https://doi.org/10.1038/s41467-017-02150-7>
- Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. <https://doi.org/10.1093/bioinformatics/btr521>
- Jones, C. J., Edwards, K. J., Castaglione, S., Winfield, M. O., Van Wiel Sale, C., Bredemeijer, C., Buianti, M., Maestri, E., Malcevshi, A., Marmioli, N., Aert, R., Volckaet, G., Rueda, J., Linacero, R., Vazquez, A., & Karp, A. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding*, *3*, 381–390. <https://doi.org/10.1023/a:1009612517139>
- Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, *6*, Article 208. <https://doi.org/10.3389/fgene.2015.00208>
- Kapitonov, V. V., & Jurka, J. (2000). Molecular paleontology of transposable elements from *Arabidopsis thaliana*. In J. F. McDonald (Eds.), *Transposable elements and genome evolution* (Vol. 1, pp. 27–37). Georgia genetics review I. Springer. https://doi.org/10.1007/978-94-011-4156-7_4
- Langgut, D. (2017). The citrus route revealed: From Southeast Asia into the Mediterranean. *HortScience*, *52*(6), 814–822. <https://doi.org/10.21273/HORTSCII1023-16>
- León-Novelo, L., Gerken, A. R., Graze, R. M., McIntyre, L. M., & Marroni, F. (2018). Direct testing for allele-specific expression differences between conditions. *G3 Genes|Genomes|Genetics*, *8*(2), 447–460. <https://doi.org/10.1534/g3.117.300139>
- Magris, G., Jurman, I., Fornasiero, A., Paparelli, E., Schwoppe, R., Marroni, F., Di Gaspero, G., & Morgante, M. (2021). The genomes

- of 204 *Vitis vinifera* accessions reveal the origin of European wine grapes. *Nature Communications*, 12, Article 7240. <https://doi.org/10.1038/s41467-021-27487-y>
- Marroni, F., Pinosio, S., & Morgante, M. (2014). Structural variation and genome complexity: Is dispensable really dispensable? *Current Opinion in Plant Biology*, 18, 31–36. <https://doi.org/10.1016/j.pbi.2014.01.003>
- Marroni, F., Scaglione, D., Pinosio, S., Policriti, A., Miculan, M., Di Gaspero, G., & Morgante, M. (2017). Reduction of heterozygosity (ROH) as a method to detect mosaic structural variation. *Plant Biotechnology Journal*, 15, 791–793. <https://doi.org/10.1111/pbi.12691>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17, 10–12. <https://doi.org/10.14806/ej.17.1.200>
- Mendell, J., Sharifi, N., Meyers, J., Martinez-Murillo, F., & Dietz, H. C. (2004). Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nature Genetics*, 36, 1073–1078. <https://doi.org/10.1038/ng1429>
- Mudge, K., Janick, J., Steven, S., & Goldschmidt, E. E. (2009). A history of grafting. *Horticultural Reviews*, 35, 437–493. <https://doi.org/10.1002/9780470593776.ch9>
- Pandey, R. V., Franssen, S. U., Futschik, A., & Schlötterer, C. (2013). Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular Ecology Resources*, 13(4), 740–745. <https://doi.org/10.1111/1755-0998.12110>
- Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Pelsy, F. (2010). Molecular and cellular mechanisms of diversity within grapevine varieties. *Heredity*, 104, 331–340. <https://doi.org/10.1038/hdy.2009.161>
- Pelsy, F., Dumas, V., Bévillacqua, L., Hocquigny, S., & Merdinoglu, D. (2015). Chromosome replacement and deletion lead to clonal polymorphism of berry color in grapevine. *PLOS Genetics*, 11(4), e1005081. <https://doi.org/10.1371/journal.pgen.1005081>
- Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., & Morgante, M. (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology*, 33, 2706–2719. <https://doi.org/10.1093/molbev/msw161>
- Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), i333–i339. <https://doi.org/10.1093/bioinformatics/bts378>
- Rawat, N., Kumar, B., Albrecht, U., Du, D., Huang, M., Yu, Q., Zhang, Y., Duan, Y.-P., Bowman, K. D., Gmitter, F. G., Jr., & Deng, Z. (2017). Genome resequencing and transcriptome profiling reveal structural diversity and expression patterns of constitutive disease resistance genes in Huanglongbing-tolerant *Poncirus trifoliata* and its hybrids. *Horticulture Research*, 4, 17064. <https://doi.org/10.1038/hortres.2017.64>
- Raxwal, V. K., & Riha, K. (2023). The biological functions of nonsense-mediated mRNA decay in plants: RNA quality control and beyond. *Biochemical Society Transactions*, 51(1), 31–39. <https://doi.org/10.1042/BST20211231>
- Scora, R. W. (1975). On the history and origin of citrus. *Bulletin of the Torrey Botanical Club*, 102(6), 369–375. <https://doi.org/10.2307/2484763>
- Strazzer, P., Spelt, C. E., Li, S., Blied, M., Federici, C. T., Roose, M. L., Koes, R., & Quattrocchio, F. M. (2019). Hyperacidification of *Citrus* fruits by a vacuolar proton-pumping P-ATPase complex. *Nature Communications*, 10, Article 744. <https://doi.org/10.1038/s41467-019-08516-3>
- Terol, J., Ibanez, V., Carbonell, J., Alonso, R., Estornell, L. H., Licciardello, C., Gut, I. G., Dopazo, J., & Talon, M. (2015). Involvement of a citrus meiotic recombination TTC-repeat motif in the formation of gross deletions generated by ionizing radiation and MULE activation. *BMC Genomics*, 16, Article 69. <https://doi.org/10.1186/s12864-015-1280-3>
- Velasco, R., & Licciardello, C. (2014). A genealogy of the citrus family. *Nature Biotechnology*, 32, 640–642. <https://doi.org/10.1038/nbt.2954>
- Vendramin, E., Pea, G., Dondini, L., Pacheco, I., Dettori, M. T., Gazza, L., Scalabrin, S., Strozzi, F., Tartarini, S., Bassi, D., Verde, I., & Rossini, L. (2014). A unique mutation in a MYB gene cosegregates with the nectarine phenotype in peach. *PLoS ONE*, 9, e90574. <https://doi.org/10.1371/journal.pone.0090574>
- Wang, L., Huang, Y., Liu, Z., He, J., Jiang, X., He, F., Lu, Z., Yang, S., Chen, P., Yu, H., Zeng, B., Ke, L., Xie, Z., Larkin, R. M., Jiang, D., Ming, R., Buckler, E. S., Deng, X., & Xu, Q. (2021). Somatic variations led to the selection of acidic and acidless orange cultivars. *Nature Plants*, 7, 954–965. <https://doi.org/10.1038/s41477-021-00941-x>
- Wang, N., Chen, P., Xu, Y., Guo, L., Li, X., Yi, H., Larkin, R. M., Zhou, Y., Deng, X., & Xu, Q. (2024). Phased genomics reveals hidden somatic mutations and provides insight into fruit development in sweet orange. *Horticulture Research*, 11, uhad268. <https://doi.org/10.1093/hr/uhad268>
- Wang, X., Xu, Y., Zhang, S., Cao, L., Huang, Y., Cheng, J., Wu, G., Tian, S., Chen, C., Liu, Y., Yu, H., Yang, X., Lan, H., Wang, N., Wang, L., Xu, J., Jiang, X., Xie, Z., Tan, M., ... Xu, Q. (2017). Genomic analyses of primitive, wild, and cultivated citrus provide insights into asexual reproduction. *Nature Genetics*, 49(5), 765–772. <https://doi.org/10.1038/ng.3839>
- Wu, G. A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., Perrier, X., Ruiz, M., Scalabrin, S., Terol, J., Takita, M. A., Labadie, K., Poulain, J., Couloux, A., Jabbari, K., Cattonaro, F., Del Fabbro, C., Pinosio, S., ... Rokhsar, D. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology*, 32, 597–698. <https://doi.org/10.1038/nbt.2906>
- Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., Domingo, C., Tadeo, F. R., Carbonell-Caballero, J., Alonso, R., Curk, F., Du, D., Ollitrault, P., Roose, M. L., Dopazo, J., Gmitter, F. G., Jr., Rokhsar, D. S., & Talon, M. (2018). Genomics of the origin and evolution of *Citrus*. *Nature*, 554, 311–316. <https://doi.org/10.1038/nature25447>
- Xu, Q., Chen, L. L., Ruan, X., Chen, D., Zhu, A., Chen, C., Bertrand, D., Jiao, W. B., Hao, B. H., Lyon, M. P., Chen, J., Gao, S., Xing, F., Lan, H., Chang, J. W., Ge, X., Lei, Y., Hu, Q., Miao, Y., ... Ruan, Y. (2013). The draft genome of sweet orange (*Citrus sinensis*). *Nature Genetics*, 45, 59–66. <https://doi.org/10.1038/ng.2472>
- Xu, Y., Gao, Z. T. J., Jiang, W., Zhang, S., Wang, Q., & Qu, S. (2016). Genome-wide detection of SNP and SV variations to reveal early ripening-related genes in grape. *PLoS ONE*, 11(2), e0147749. <https://doi.org/10.1371/journal.pone.0147749>
- Zhang, S., Chen, W., Xin, L., Gao, Z., Hou, Y., Yu, X., Zhang, Z., & Qu, S. (2014). Genomic variants of genes associated with three horticultural

tural traits in apple revealed by genome re-sequencing. *Horticulture Research*, 1, 45045. <https://doi.org/10.1038/hortres.2014.45>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Scaglione, D., Ciacciulli, A., Gattolin, S., Caruso, M., Marroni, F., Casas, G. L., Jurman, I., Licciardello, G., Catara, A. F., Rossini, L., Licciardello, C., & Morgante, M. (2025). Deep resequencing unveils novel SNPs, InDels, and large structural variants for the clonal fingerprinting of sweet orange [*Citrus sinensis* (L.) Osbeck]. *The Plant Genome*, 18, e20544. <https://doi.org/10.1002/tpg2.20544>