scientific data

Article in Press

Check for updates

OPEN

DATA DESCRIPTOR

Digitalization of the genotypic and phenotypic information retained within the FEM grapevine germplasm bank

Silvia Vezzulli[1,5] ✉, Paola Bettinelli[1,5], Daniela Nicolini[1], Daniele Migliaro[2], Giulia Betta[1], Geovani Luciano de Oliveira[3], Carlotta Pirrello[1], Silvano Clementi[1], Cinzia Dorigatti[1], Alessandra Zatelli[1], Monica Dalla Serra[1], Raffaele Filippi[1], Pietro Antonini[1], Siria Bassetti[1], Marco Calovi[1], Monica Visentin[1], Andrea Campestrin[1], Tiziano Tomasi[1], Lorenzo Spina[1,4], Luca Zulini[1], Laura Costantini[1], Diego Micheletti[1] & Marco Stefanini[1]

The Fondazione Edmund Mach (FEM) maintains a grapevine germplasm bank consisting of 3,065 grapevine accessions that supports long-term conservation and utilisation of crop genetic diversity. In addition to acquisitions, the germplasm includes advanced crossbreeding selections developed to improve grape quality and introgress tolerance and resistance traits. This objective highlights the importance of verifying accession identity through standardized molecular profiles, complemented by ampelographic characterization when necessary. A systematic inventory of the germplasm was undertaken and a comprehensive ontologization of all accession-related data and metadata was carried out in the context of FAIR (Findability, Accessibility, Interoperability, and Reusability) compliance. This overarching survey led to the published FEMVitisDataset encompassing 2,905 accessions, resulting in the identification of 1,931 unique profiles. Of these, 897 were classified as novel, as no exact or near-exact genetic match was found across the consulted reference databases. By providing accession-level passport data, validated SSR profiles, and associated phenotypic information, the FEMVitisDataset expands the documentation of grapevine diversity and represents a valuable resource for germplasm management, genetic studies, and modern breeding programs integrating molecular and phenotypic information.

Background & Summary

The grapevine germplasm bank of the Fondazione Edmund Mach (FEM) is primarily an *ex situ* collection of grapevine (*Vitis* spp.) genetic resources originating from approximately 50 geographic origins worldwide. In addition to conserved acquisitions, the germplasm includes advanced crossbreeding selections developed at FEM over decades to improve grape quality and introgress tolerance/resistance traits against abiotic/biotic stresses, under a climate crisis scenario. The FEM germplasm bank aims to the long-term conservation and utilization of grapevine genetic diversity to potentially support base and applied research activities—ranging from conventional crossbreeding and clonal selection to New Genomic Techniques (NGTs) like cisgenesis and genome editing—and culminates in the systematic accession inventory of the FEMVitisDataset[1] presented herein. Previous publications relied on part on this germplasm, prior to the curation and standardization process of the dataset, focusing on specific phenotypic traits such as berry skin colour somatic variants[2,3], flower sex[4], phenology and fertility traits[5], aromatic somatic variants[6], and seedlessness somatic mutations[7]. Lately, in the context of Findability, Accessibility, Interoperability, and Reusability (FAIR) compliance, a comprehensive

[1]Research and Innovation Centre, Fondazione Edmund Mach, 38098, San Michele all'Adige, Italy. [2]CREA – Research Centre for Viticulture and Enology, 31015, Conegliano, Italy. [3]Molecular Biology and Genetic Engineering Center, Universidade Estadual de Campinas, 13083-875, Campinas, Brazil. [4]University School for Advanced Studies IUSS Pavia, 27100, Pavia, Italy. [5]These authors contributed equally: Silvia Vezzulli, Paola Bettinelli. ✉e-mail: silvia.vezzulli@fmach.it

Dataset Component	Status		Counts	%
ACCESSIONS	Maintained		3,065	100.0
	FEM Crossbreeding*		168	5.5
	Unpublished		160	5.2
	Published		2,905	94.8
GENETIC PROFILES^	Low Quality		30	1.0
	High Quality		2,875	99.0
	Unique	Known	1,034	36.0
		Novel	897	31.2
	Redundant	Known	881	30.6
		Novel	63	2.2

Table 1. Data curation workflow of the FEM grapevine molecular germplasm dataset. This flow matrix summarizes the progressive curation steps, tracking the dataset from the initial total physical accessions down to the final categorized unique genetic profiles. It reports the breakdown of public accessions and genotyped samples, as well as the low-quality profiles that were excluded from subsequent classifications. Consequently, the final distribution into unique, redundant, known, and novel categories is calculated exclusively using high-quality genetic profiles (n = 2,875). *Eight FEM crossbreed genotypes have been patented and therefore included in the published dataset. ^The total number of genetic profiles is calculated at the accession level (2,905): a single profile was considered for each accession, regardless of the 2,956 total genetic profiles recorded in the marker file due to the replicated sampling of accessions maintaining more than one physical position in the field.

ontologization of all accession-related data and metadata was carried out. The resulting structured dataset, specifically marker allele calls, was partially exploited in the genetic structure and resistance loci analyses of a specific subset of *Vitis* hybrids[8].

The systematic curation and progressive filtering of the FEMVitisDataset[1], tracking the workflow from the total physical field accessions to the final validated unique genetic profiles and their corresponding records across the shared files, are summarized in Table 1. Nine reference Simple Sequence Repeat (SSR) markers[9], recommended by the Grapevine International Community (former IGGP, International Grapevine Genome Program), were used to characterize the FEM germplasm bank of 3,065 grapevine accessions, including 168 crossbreeding selections. Consistently with institutional confidentiality policies, except for eight FEM-registered varieties[10], the related data are not publicly released because these materials are part of ongoing proprietary crossbreeding programs. The published FEMVitisDataset[1] therefore included 2,905 accessions. The authenticity (Trueness-To-Type, TTT) evaluation of the cultivars was assessed through a multi-step workflow combining SSR profiles, reference-database comparisons, and supporting pedigree or phenotypic evidence. First, SSR allele lengths were called using a standardized allele binning procedure. Second, the resulting genetic profiles were compared with the information available in public and private databases (*V*IVC[11], GRIN[12], EU-*Vitis*[13], CREA-VE[14]). Regarding the specific case of 32 detected families (*cépages*) of somatic variants (e.g. for berry colour, seedlessness, leaf hairiness), ontologized ampelographic information and phenotypic traits were deployed to support their distinction that SSR markers alone cannot resolve. This workflow resulted in the identification of 1,931 unique profiles, while for 30 accessions the excess of missing data prevented TTT validation. Eventually, within the unique genetic profiles, 897 were not previously reported and were therefore classified as novel. The release of these validated profiles also provides the basis for their submission and integration into the *Vitis* International Variety Catalogue (*V*IVC[11]), contributing to the expansion and enrichment of the reference genetic resources of the Grapevine International Community.

Being one of the historically and commercially most important crops in temperate regions, *Vitis* collections are maintained across many Countries in the Mediterranean area and all over the world. According to the last FAO report, globally, 85,886 grapevine accessions (acc.) are conserved in 112 genebanks (FAO, 2025[15]). Besides grapevine, nowadays several genebanks and databases are dedicated to crop species to preserve and maintain the treasure of genetic diversity (Supplementary Text 1). Genetic diversity is described by genetic indicators; several initiatives and international efforts have been carried out also in wild plants and other kingdoms (e.g. Biodiversa + – GINAMO). In this context, the dataset presented here aims to bridge the gap between genetic preservation and actionable science. By providing a comprehensive and standardized data repository, this work is designed to support long-term conservation strategies and accelerate (molecular) breeding programs. Ultimately, maximizing the utility and accessibility of these genomic resources is crucial to foster research addressing crop resilience and adaptation in the face of the ongoing climate crisis.

Methods

FAIR data modelling and semantic annotation. To ensure the FAIRification of the generated dataset, a multi-layered standardization and semantic harmonisation strategy was implemented. At the experimental level, all phenotypic and observational metadata were structured in compliance with the Minimum Information About a Plant Phenotyping Experiment standard (MIAPPE v1.1[16]), enabling comprehensive and unambiguous documentation of experimental design and treatment factors required for valid data interpretation and reuse. Semantic interoperability of phenotypic observations was achieved through systematic annotation using

controlled vocabularies and ontologies. General plant anatomical and morphological terms were mapped to the Plant Ontology[17], while grapevine-specific phenotypic traits were aligned with the VITIS Ontology (CO_356[18]), a crop-specific implementation derived from the Crop Ontology[19] framework. This approach ensures consistent trait definition across datasets and supports cross-study comparison and machine-readable interpretation.

The dataset was structured to be compatible with the Breeding API[20] specifications, which provide a standardised framework for programmatic data exchange in plant breeding and genebank information systems. BrAPI operationalises and extends several concepts originally defined in the Multi-Crop Passport Descriptors[21] (MCPD) standard developed by the Food and Agriculture Organization of the United Nations. To protect the integrity of the private germplasm bank, all fields referring to plant position were withheld from the public release.

Plant material and germplasm bank structure. The FEM grapevine germplasm is primarily maintained in San Michele all'Adige, Italy (46°11′N, 11°08′E, 223 m asl, FAO holding institution ITA362). Over the past 30 years, a total of 3,065 accessions have been collected and are maintained *ex situ*. The germplasm bank also includes 168 FEM advanced crossbreeding selections, consisting of 109 selected disease-resistant and 59 selected *V. vinifera* × *V. vinifera* genotypes. As the institutional crossbreeding programmes are subject to confidentiality constraints, detailed information for these advanced selections was not publicly released, with the exception of eight FEM-registered varieties (Iasma Eco 1, 2, 3, 4; Termantis, Nermantis, Charvir, Valnosia[10]) officially released at national level (all listed) and also protected at European level (the last four, disease-resistant). Consequently, the publicly released germplasm dataset comprised 2,905 accessions.

The germplasm is described through 40 fields (Supplementary Table 1), a subset of which is detailed as follows. Each accession is assigned to a FEM class (germplasm_attribute_value.fem_class) that encodes the mode of acquisition or generation of the material. These include 'collection' (material sourced from external repositories or collected in the wild), 'crossbreeding' (material derived from FEM breeding programmes), and 'mutation' (material generated through FEM-induced physical, chemical, or biological mutagenesis, including lines developed via NGTs). As no accessions belonging to the 'mutation' class are included in the publicly released dataset, this class is not further described here. Each accession is assigned a FEM accession number (germplasm.accessionnumber). For accessions belonging to the 'collection' class, this corresponds to a progressive acquisition number, whereas for 'crossbreeding' selections the identifier follows the structure YYYY.000.000, indicating year of crossing, cross number, and progeny individual.

A Persistent Unique IDentifier (PUID), compliant with MIAPPE v1.1 Biological Material ID (DM-41), is assigned to each accession. The PUID consists of the FEM accession number prefixed with FEM:GC for germplasm collection individuals, and FEM:GX for FEM crossbreeding selections. The accession name (germplasm.defaultdisplayname) corresponds to the *V*IVC prime name (uppercase) when TTT status was confirmed using nine universal SSR markers[9] based on comparison with reference profiles (see Section Marker profile structure and fingerprinting). When no correspondence with reference profiles was identified (germplasm.germplasmname 'unknown'), the historical FEM name (lowercase, germplasm_attribute_value.fem_historicalname) is used. In cases where two or more accessions of unknown TTT shared identical marker profiles, a common historical name was selected for display and submission to the *V*IVC database (ongoing work). The validation method for TTT calls is defined in the germplasm field germplasm_attribute_value.ttt_validation, indicating either the database used for TTT profile comparison ('crea', 'euvitis', 'grin', 'vivc'), or the application of alternative or additional identification strategies ('pedigree', 'mutant (somatic)', 'mutant (polyploid)').

The *V*IVC accession number is reported (germplasm_attribute_value.vivc_accessionnumber) when the accession is present in *V*IVC catalogue. When TTT was validated within the CREA Viticulture and Enology private molecular database, the corresponding CREA accession number is also reported (germplasm_attribute_value.crea_accessionnumber).

Additional recorded attributes include FEM synonyms, and geographical origin and designated use (Table, Wine, Rootstock, Dry grape for raisins, or Juice) derived from *V*IVC. Each distinct genetic profile is assigned a unique profile identifier (germplasm_attribute_value.unique_ssrprofile_code). Recurrent profiles are encoded using letter-based identifiers (e.g. AA, AAA, AB, DU), whereas non-redundant profiles are assigned sequential numeric identifiers (e.g. 1, 2, 3). The code 'LOW_Q' was used to identify genetic profiles with an excess of missing data that prevented TTT validation, as described in the section Marker profile structure and fingerprinting.

Taxonomic nomenclature related fields follow the *V*IVC database to ensure consistency in grapevine germplasm documentation. This includes the designation of *Vitis vinifera* subspecies (subsp.) as *sylvestris* and *sativa*, acknowledging that for the latest the most recent taxonomic review is subsp. *vinifera*[22]. For the same reason of consistency and traceability, the designation *Muscadinia rotundifolia* was adopted instead of *V. rotundifolia*, reflecting its distinct chromosome number and reproductive incompatibility with *Vitis* species[23,24]. The taxonomic basis (germplasm_attribute_value.taxonomy_basis) is classified as 'established', 'literature-derived (with citation)', or 'presumed', based on pedigree information of TTT profiles, published sources, or ampelographic and historical evidence, respectively. Each accession is further assigned to an FEM category (germplasm_attribute_value.fem_category) defined using the combination of taxonomic fields and FEM class. Seven categories were identified, comprising five collection categories ('*V. vinifera* subsp. *sativa*'; '*V. vinifera* subsp. *sylvestris*'; 'Selected hybrids'; '*V*. spp. non-*vinifera*'; 'Naturalized hybrids') and two crossbreeding categories ('*V. vinifera* subsp. *sativa*, FEM crossbreeding'; 'Selected hybrids, FEM crossbreeding').

Observation unit structure. The FEM grapevine germplasm is currently cultivated and conserved across five locations: Giaroni, San Donà, Inferno and Girelli, Vigalzano. The first four sites are located within the municipality of San Michele all'Adige and are in close spatial proximity, with a maximum distance of approximately 700 m among them. Vigalzano is located approximately 26 km away in the Valsugana valley and therefore

represents a geographically distinct growing environment. Two sites comprise multiple fields: Giaroni includes six fields (A–F), while San Donà consists of two fields (A and B). The Vigalzano site is specifically dedicated to the conservation of traditional landraces originating from the Trentino region. The sites Inferno and Girelli are managed without chemical treatments, as they are reserved for disease-resistant material.

For internal management purposes, the position of each individual plant was recorded using a hierarchical spatial reference system including block (row for trees), plot (pole) and plant (planting hole) identifiers. Different events affecting plants, ranging from non-lethal (e.g. 'winter frost', 'late spring frost', 'hail', 'heat/sunburn', 'emasculation' etc.) to lethal causes leading to death and eradication (e.g. 'natural', 'Flavescence Dorée', 'viroses', 'Esca disease' etc.), were included (Supplementary Table 2). Based on FEM policy and analogously to other international germplasm-holding institutions, the dataset fields reporting the physical position of individual accessions (observation_unit.observationunitpui, place.name, observation_unit.observationunitposition) were considered sensitive and therefore not released as part of the publicly available dataset. Data linkability is preserved internally via an identical, master version of the dataset maintained by FEM, which maps the public accession identifiers back to their withheld plant-position data.

Marker profile structure and fingerprinting. For genotyping, approximately 1 cm^2 of young leaf tissue was collected in 96-well Collection Microtubes plates and homogenized using a TissueLyser (Qiagen). Genomic DNA was extracted following the cetyltrimethylammonium bromide (CTAB) protocol[25] optimized for grapevine young leaves and adapted to 96-well plate volumes (Supplementary Text 2). Quality and quantification of the DNA was assessed using a NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific), and the concentrations were normalized to 10 $ng\,\mu L^{-1}$ prior to amplification. The sample name (sample.name), referring to the plant position, was considered sensitive and therefore not released as part of the publicly available dataset.

All accessions were genotyped using the nine universal microsatellites for grapevine fingerprinting, according to the OIV resolution OIV-VITI 609-2019: VVS2[26]; VVMD5 and VVMD7[27]; VVMD25, VVMD27, VVMD28 and VVMD32[28]; VrZAG62, and VrZAG79[29]. The markers were amplified in three multiplex PCR reactions using a GeneAmp 9700 thermal cycler (Thermo Fisher Scientific) following a previously defined protocol[30]. Fragment separation was performed by capillary electrophoresis at the Sequencing and Genotyping Platform of FEM (San Michele all'Adige, Italy) using an ABI 3130xl Genetic Analyzer (Life Technologies), with the GeneScan 500LIZ size standard (Life Technologies) as internal ladder. Allele calling was carried out using GeneMapper V4.0 (Applied Biosystems, 2020), and with GeneMarker V3.0.0 (SoftGenetics, 2022). A minimum peak height threshold of 1,000 relative fluorescence units (RFU) was applied for allele calling. More details about data harmonization and robust allele binning across different historical batches and software transitions are given in the Technical Validation section. Genotypes showing a single detectable allele (putative hemizygous calls) were conservatively coded as homozygous by duplicating the observed allele. This curation strategy was adopted to ensure full compliance with the international *V*IVC standard, thereby facilitating direct interoperability with major global grapevine databases. Furthermore, this convention prevents an artificial inflation of null allele frequencies (excess of null alleles) and minimizes the impact of potential allele dropout, ensuring robust and unbiased performance in subsequent population genetics and identity validation analyses. The structure of the microsatellite marker dataset is detailed in Supplementary Table 3.

Microsatellite data were analysed using GenAlEx[31] V6.5 and Cervus[32] V3.0.7 to support multilocus genotype comparisons, identity and parentage analysis, TTT validation, and for the identification of unique genetic profiles. Preliminary analyses were performed to identify potential genotyping errors, detect name inconsistencies arising from sampling or planting errors, and identify erroneously duplicated accessions. Mutant alleles, defined as alleles that deviate from the expected number of repeat units of the corresponding microsatellite marker, were validated based on pedigree information and comparison with the expected TTT reference profile. Accessions with five or more missing SSR loci, which precluded reliable TTT assessment, were classified as low-quality profiles. Allelic discrepancies relative to the corresponding TTT reference profiles were retained in the released dataset to preserve traceability of the original technical allele calls. The identity analysis and profile validation followed a multi-layer approach that is described in detail in the **Technical Validation** section. Briefly, a profile was classified as "novel" when, after purging low-quality data and collapsing internal duplicates, its 9-locus SSR fingerprint yielded zero exact or near-exact matches across all four reference databases (*V*IVC[11], EU-*Vitis*[13], GRIN[12], CREA-VE[14]) within our calibrated mismatch tolerance. These databases are continuously updated and do not have static version numbers; they were last accessed in October 2025.

Observation phenotype structure. The phenotypic data included in the dataset were *ad hoc* recorded (compiled) from 2014 to 2024 seasons. Where possible, the descriptors developed by the OIV[33] were used. The phenotypic traits with compiled fields were divided into four categories: i) 53 OIV ampelography descriptors (four assessments per descriptor per genotype) requested for the registration of a new variety (Supplementary Table 4); ii) the phenology stages from woolly bud to ripening (one measure per week per accession) using the Eichhorn-Lorenz (E-L) scale modified by Coombe[34], together with the ripening curve – from véraison to 18 °Brix – and berry skin colour (Supplementary Table 5); iii) seven fertility parameters (one measure per year per accession): e.g. flower sexual organs, presence of inflorescences per plant (all kind of shoots), presence of bunches per plant (all kind of shoots) (Supplementary Table 6); iv) 14 biotic stress resistance parameters: e.g. downy mildew, powdery mildew, black rot, erinosis, gray mold resistance at leaf and bunch level (the minimum disease resistance score recorded worst assessment per year per accession was retained), and viroses (ELISA test) (Supplementary Table 7). Further fields were listed (foreseen) into the mentioned phenotype structures to accommodate phenotypes of interest in subsequent assessments.

File Name	Worksheets	Years	Samples	Unit of aggregation	Missing-Value Codes	Data Type	Records
01_Germplasm	1	-	2,905	Accession	Empty cell	Metadata	2,905
03_Gnt_call.SSRMarker	1	6	2,956	Plant	0	SSR allele calls	26,604
04_Observation. Ampelography	2	2	4	Plot	Empty cell	Ampelographic descriptors	424
05_Observation. BerryColor.PhenoStage.Brix	13*	8	7,091	Plot	Empty cell	Berry color descriptor	6,543
						Phenological stage descriptor	94,357
						°Brix measurement	13,061
06_Observation. Fertility.FlowerSex	1	9	11,923	Plant	Empty cell	Fertility descriptors	114,497
						Flower sex descriptor	355
07_Observation. BioStress	4	4	4	Plot	Empty cell	Disease resistance descriptors or disease detection	140

Table 2. Data manifest and structural blueprint of the FEMVitisDataset. This manifest summarizes the organization of the deposited repository files on Zenodo. *by year and sampling block e.g., Y2022a, Y2022b

Besides the phenotypes belonging to the described categories, other relevant traits were ontologized (e.g. abiotic stress tolerance parameters, such as chlorophyll content of the leaves, N-P-K concentration in leaves). In the future, oenological features will be also included and this phenotype structure will face continuous expansion.

All the phenotypic observations were introduced by the first 12 fields reported in the observation unit structure (Supplementary Table 2). Those fields, together with the sample name (sample.name), since refer to the plant position, were considered sensitive and therefore not released as part of the publicly available dataset.

Data Records

The FEMVitisDataset[1] is deposited at Zenodo (https://doi.org/10.5281/zenodo.18001007). It is structured into six Excel (.xlsx) files corresponding to six of the seven data structure supplementary tables. The spatial and positional mapping described in the Observation unit dataset (detailing the internal physical hierarchy, field coordinates, and block arrangements as described in Supplementary Table 2) is withheld from the public release in compliance with institutional security policies regarding the physical location of the germplasm collection. Nevertheless, all phenotypic data associated with these observation units are interconnected via the unique FEM accession numbers. To facilitate dataset navigation, quality auditing, and downstream reuse, a comprehensive summary of the FEMVitisDataset[1] structure—detailing file names, record counts, primary keys, withheld fields, and missing-value identifiers—is provided as a data manifest in Table 2.

01_Germplasm: contains the published germplasm metadata. It includes the FEM accession number, which serves as the primary key for all other released tables (dataset structure described in Supplementary Table 1).

03_Gnt_call.SSRMarker: reports allele calls for nine SSR loci used for grapevine fingerprinting (dataset structure described in Supplementary Table 3). The raw molecular dataset comprised 26,604 data points, defined as samples × SSR locus combinations (2,956 × 9). Multiple samples were analyzed for a subset of accessions due to their presence in different locations, resulting in a higher number of samples than accessions. During manual quality control, 895 data points (3.4%) were identified as missing values (two null alleles).

04_Observation.Ampelography: comprises ampelographic observations recorded for officially released and patented accessions. The file is organised into two worksheets corresponding to two assessment periods, reflecting the documentation requirements for varietal registration dossiers. (dataset structure described in Supplementary Table 4).

05_Observation.BerryColor.PhenoStage.Brix: comprises phenological observations, berry colour assessments and ripening measurements. Multiple worksheets are provided, each corresponding to a specific year and phenotyping tranche. Worksheet identifiers encode the year and field of observation (e.g. Y2022a and Y2022b), reflecting distinct but internally consistent observation campaigns conducted on different collection tranches (dataset structure described in Supplementary Table 5).

06_Observation.Fertility.FlowerSex: reports multi-year fertility-related observations, including flower sex determination and associated reproductive traits. All available records are compiled into a single worksheet spanning multiple seasons (dataset structure described in Supplementary Table 6).

07_Observation.BioStress: includes biotic stress evaluations recorded over multiple years for a limited set of disease-resistant accessions that have been officially released. Separate worksheets correspond to different assessment years and disease observations (dataset structure described in Supplementary Table 7).The BioStress (File 07) and Ampelography (File 04) datasets contain phenotypic evaluations for a restricted subset of four registered accessions and are therefore not intended to provide representative coverage of the entire FEM germplasm collection. Unavailable metadata and unobserved phenotypic traits are represented as empty cells, whereas the absence of an SSR allele peak was recorded as a null allele and coded as "0" (Table 2).

Data Overview

The *ex situ* grapevine repository conserved at FEM, net of genotypes lost due to environmental incompatibility (i.e., cold winters for hot climate background as *M. rotundifolia*), currently holds 3,065 accessions. This represents a substantial expansion compared with the historical collection[35] which encompassed 2,272 accessions. The latter was renewed starting from 2013 – due to plant age, presence of trunk diseases, several redundancies

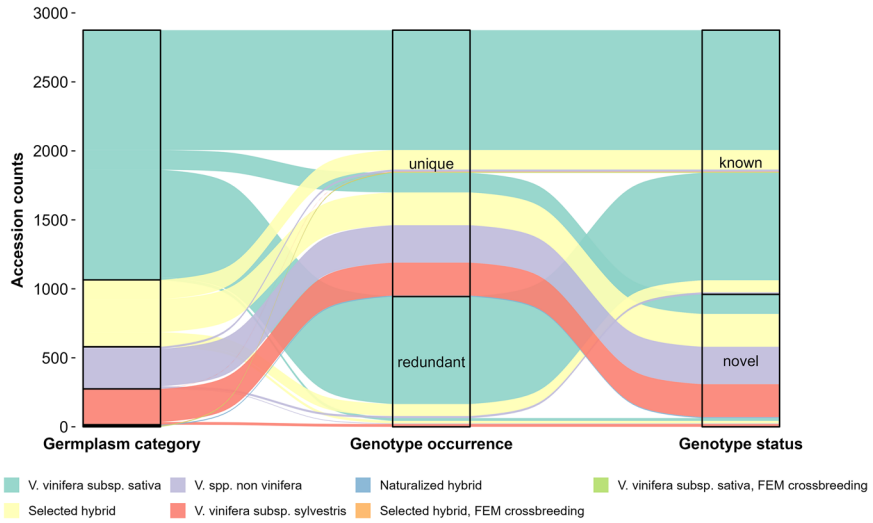


Fig. 1 Integrated overview of the FEMVitisDataset based on accession counts. The left axis summarises the taxonomic and origin-based composition of the collection. The central axis reports genotype occurrence derived from multilocus microsatellite profiling, distinguishing unique genetic profiles from redundant profiles. The right axis describes genotype status relative to external reference databases, separating accessions matching previously documented profiles (known) from those not yet associated with a reference identity (novel).

– and on this a comprehensive re-genotyping effort was motivated by the expansion and refinement of grapevine reference databases, enabling the re-evaluation of the TTT status (authenticity) as well as the identification of novel, previously unassigned, TTT profiles.

The publicly deposited FEMVitisDataset[1] comprises a curated subset of 2,905 grapevine accessions. This difference reflects the exclusion of 163 proprietary breeding selections subject to intellectual property restrictions. The published dataset represents a wide spectrum of grapevine genetic diversity, namely *V. vinifera* cultivars, wild relatives, interspecific and naturalized hybrids, and eight genotypes released by FEM as new varieties. Germplasm classification fields described taxonomic assignment, origin and mode of acquisition, providing a structured framework for stratified exploration of both genotypic and phenotypic data.

A structural overview of the publicly released FEMVitisDataset[1] is given in Fig. 1. Thirty accessions with low-quality SSR profiles were excluded from the visualization; these comprised 13 selected hybrids, nine *V. vinifera* subsp. *sativa*, four *Vitis* spp. non-*vinifera*, and four *V. vinifera* subsp. *sylvestris* accessions. Consequently, the visualization includes only accessions with high-quality SSR profiles (n = 2,875; Table 1). The dataset is structured according to taxonomic origin (Germplasm category, seven as described in Plant material and germplasm bank structure), genetic redundancy (Genotype occurrence, unique or redundant), and whether each SSR profile was previously reported in public reference databases (Genotype status, known or novel). Detailed counts and percentages based on 2,875 high-quality SSR profiles are provided in Supplementary Table 8. The collection is dominated by *V. vinifera* subsp. *sativa* (63.0%), followed by selected interspecific hybrids originating from international crossbreeding programmes (16.9%), non-*vinifera Vitis* species (10.6%), and *V. vinifera* subsp. *sylvestris* accessions (9.0%). Minor components include a small number of naturalized hybrids collected during a field campaign in New Jersey (USA), as well as FEM crossbreeding selections, each representing less than 1% of the publicly released dataset. Genotype status derived from multilocus microsatellite profiling distinguished unique genetic profiles (67.2%) from redundant profiles (32.8%), which reflects the presence of replicated genotypes within the collection. The distribution of unique and redundant profiles is not uniform across germplasm classes: *V. vinifera* subsp. *sativa* includes comparable proportions of unique and redundant genotypes, whereas *V. vinifera* subsp. *sylvestris*, non-*vinifera Vitis* species, and selected hybrids are predominantly represented by unique genetic profiles. Comparison with external reference databases further distinguishes accessions matching previously documented profiles (known, 66.6%) from those not yet associated with a reference identity (novel, 33.4%). The majority of known profiles correspond to cultivated *V. vinifera* accessions, with a smaller contribution from selected hybrids, while most non-*vinifera* materials, *sylvestris* accessions, and naturalized hybrids fall within the novel profiles. Together, this integrated representation highlights the compositional structure of the FEMVitisDataset[1], providing a concise overview of the diversity captured within the FEM grapevine germplasm and supporting targeted reuse of genotypic and phenotypic data. Finally, 84 accessions corresponded to 51 somatic variants of 32 varieties (*cépages*) that are listed in the Supplementary Table 9; further evidence supporting their identification is expanded in the Technical Validation section. These natural mutants for berry colour, seedlessness, leaf hairiness, véraison timing, and bunch compactness offer significant potential for future forward genetics or functional genomics studies.

Technical Validation

Due to the complexity and historical depth of the germplasm repository, TTT identity analysis was conducted through an iterative, evidence-driven curation workflow that integrated SSR profiles, passport information,

phenotypic records, historical documentation, and comparisons with multiple public reference databases. Rather than relying solely on fixed computational thresholds, candidate identity assignments were progressively evaluated using increasingly permissive similarity criteria and subsequently verified through comprehensive manual inspection. This iterative process enabled the resolution of ambiguous cases and historical inconsistencies while distinguishing true differences from technical artifacts.

Allele Calling. Allele lengths were called by an experienced laboratory technician using a standardized allele binning procedure. To account for potential run-to-run sizing variation associated with the GeneScan 500LIZ size standard and to ensure consistent allele binning across historical batches, instrument runs, and transition from GeneMapper to GeneMarker, a strict ratiometric recalibration protocol was applied. Two reference cultivars, 'Teroldego' and 'Pinot noir' ENTAV115, were systematically included as internal controls in all capillary electrophoresis runs. Their sizing behaviour provided a common reference for allele assignment across both software platforms, and reference panel files were recalibrated when necessary. This procedure minimized technical sizing shifts and maintained consistent allele binning across time and instrument runs, providing a harmonized framework for subsequent identity validation.

Database Precedence and Complementary Sources of Evidence. When matching external profiles to assign international nomenclature, precedence was given hierarchically to the following databases: i) the *Vitis* International Variety Catalogue[11]; ii) the European *Vitis* Database[13] (EU-*Vitis*); iii) the Germplasm Resources Information[12] (GRIN) Network National Plant Germplasm System (NPGS) of the United States Department of Agriculture - Agricultural Research Service (USDA-ARS); iv) the CREA Viticulture and Enology private molecular database[14] (CREA-VE); v) lastly, historical FEM denomination when no external confirmation could be obtained. In case of identical profiles among previously unnamed accessions, a harmonized historical denomination was assigned and proposed for integration into *V*IVC. Additional identifications were obtained by integrating complementary sources of evidence: vi) consistency between the SSR profile of the unidentified accession and those of declared parents of a synonymous accession; vii) validation through supporting data (historical name, phenotypic traits, genotypic data) of the accession as a mutant variant (somatic or polyploid) that SSR markers alone cannot resolve. The source database of the reference SSR profile used for TTT validation is reported in the germplasm field germplasm_attribute_value.ttt_validation (Supplementary Table 1).

Exact-Match Rules and Allowed Mismatches. While a strict baseline of 100% allele identity across all successfully amplified loci was used as the ideal target for identical genotypes, the final acceptance of mismatches was not bound to a rigid mathematical threshold but was evaluated case-by-case by a research expert. Pairwise comparisons were screened using different combinations of parameters, including minimum number of matching loci, allowance of fuzzy matches and maximum number allowed mismatches. Historical and passport nomenclature was used as supporting information to resolve ambiguous cases, particularly when an accession carried a documented historical or passport name corresponding to a known cultivar in the consulted databases but showed unexpected allelic discrepancies. Such cases were subjected to further investigation to account for potential somatic mutations, chimerisms, or historical electrophoresis artifacts without compromising true varietal discrimination. Ultimately, a profile was classified as 'novel' only when no exact or near-exact genetic match was found across any of the consulted reference databases within the calibrated mismatch tolerance.

Low-quality Profile Threshold Rationale. Rather than adopting an arbitrary cutoff, the low-quality threshold (minimum number of successfully amplified SSR loci) was determined ad hoc after an empirical sensitivity analysis. By inspecting the number of discarded genotypes against various thresholds, we established the optimal cut-off that effectively purged unreliable or highly fragmented profiles without inadvertently sacrificing rare or unique historical accessions.

Somatic Variant Validation. Because standard 9-locus SSR fingerprinting cannot discriminate somatic variants sharing the same genetic profile, the 51 somatic variants identified within 32 cultivar groups were distinguished and validated using complementary ampelographic and phenotypic evidence, as detailed in Supplementary Table 9. Supporting evidence comprised: (i) traits available in the published phenotypic datasets, such as berry colour and phenology shifts (05_Observation. File), and fertility (06_Observation. File); and (ii) direct field observations of specific traits, such as berry pulp colour and taste, leaf blade division and pubescence, flowering defects and seedlessness, assessed *in situ* by expert curators. These field observations correspond to targeted ampelographic descriptors that will be incorporated into the structured ampelographic dataset module.

Usage Notes

This study provided a curated and FAIR-compliant dataset for grapevine germplasm conserved at the FEM repository, integrating molecular, taxonomic, historical, and multi-year phenotypic information. Data standardization and interoperability were central objectives, with field definitions aligned to BrAPI and nomenclature harmonized with *V*IVC, facilitating direct integration and future updates. The FEM accession number represents the primary and persistent identifier across the entire resource and serves as the key linking element among all released datasheets. This common identifier ensures traceability across data layers and supports integrative reuse of genetic and phenotypic information. Due to the omission of the physical positional indicators, data integration is restricted exclusively to this accession level. Consequently, for accessions evaluated across multiple field plots, intra-varietal spatial resolution is collapsed into a single accession identity. Nevertheless, temporal and longitudinal linkages remain fully functional, enabling seamless tracking of phenotypic traits across multiple years for each validated accession.

The marker dataset enabled reliable alignment and comparison of microsatellite profiles for accession identification, TTT validation, and detection of synonyms and homonyms. Furthermore, when combined with appropriate higher-density marker systems, these nine SSR profiles provide a valuable starting point for complex downstream analyses such as parentage inference, genetic structure assessment, and phylogenetic investigation[8,36–41]. All curation steps, including the exclusion of low-quality profiles and the assignment of unique genetic profiles, were explicitly documented to support transparency and reproducibility. Users should note that genotypes displaying a single detectable allele at individual loci were conservatively recoded as homozygous following *V*IVC allele-calling conventions.

In addition to germplasm metadata and marker data, the resource included complementary datasets describing phenology and fertility for a subset of accessions, enabling integrated analyses across molecular, phenotypic, and reproductive dimensions and enhancing reuse for germplasm management, (molecular) breeding, and comparative studies. Phenological and fertility traits are sensitive to annual climatic variation and local environmental conditions. Most field sites are located in close proximity within the same geographical area, while two sites are located at a greater, yet limited, distance. Nevertheless, observations collected across different years, field locations, and sampling campaigns should therefore not be interpreted as directly comparable measurements obtained under uniform environmental conditions. The multi-year structure of the dataset preserves temporal information and supports longitudinal analyses and the investigation of responses to interannual climatic variation, provided that observations are interpreted in the context of the corresponding survey year. Because precise plot coordinates are masked, fine-scale spatial and microclimatic effects cannot be reconstructed directly from the released accession-level records.

Finally, the ontological definition of the data structure for additional traits, including ampelography and biotic stress resistance, provided a framework for foreseen future data recording, compilation, and integration. The incorporation of visual assets is also recognized as a priority for future expansions of the repository. In this context, users should note that the BioStress (File 07) and Ampelography (File 04) datasets cover only four registered accessions and are not representative of the entire FEM germplasm collection. Rather, they are provided primarily as standardized templates for the future integration of additional phenotypic data. Beyond the scope of the present dataset, this data structure may also provide a reference framework for the development of interoperable data repositories for other grapevine germplasm collections. Serving as an open-access template, its downstream applications extend to helping the viticultural science community standardize phenotypic and genotypic data. This architectural standard will enable seamless data cross-referencing and integration among different grapevine collections, fostering large-scale research to address global viticultural challenges.

Data availability

The FEMVitisDataset[1] is deposited at Zenodo (https://doi.org/10.5281/zenodo.18001007).

Code availability

No custom code was developed for this study. All analyses were conducted using established software tools in accordance with the documentation and standard protocols provided by the respective developers. Relevant software settings and parameters are described.

Received: 17 April 2026; Accepted: 30 September 2026;
Published: xx xx xxxx

References

1. Vezzulli, S. *et al*. FEMVitisDataset. *Zenodo* https://doi.org/10.5281/zenodo.18001007 (2025).
2. Vezzulli, S. *et al*. Pinot blanc and Pinot gris arose as independent somatic mutations of Pinot noir. *J. Exp. Bot.* **63**, 6359 (2012).
3. Migliaro, D. *et al*. Structural dynamics at the berry colour locus in *Vitis vinifera* L. somatic variants. *Aust. J. Grape Wine Res.* **20**, 485–495 (2014).
4. Battilana, J. *et al*. Linkage mapping and molecular diversity at the flower sex locus in wild and cultivated grapevine reveal a prominent SSR haplotype in hermaphrodite plants. *Mol. Biotechnol.* **54**, 1031–1037 (2013).
5. Grzeskowiak, L., Costantini, L., Lorenzi, S. & Grando, M. S. Candidate loci for phenology and fruitfulness contributing to the phenotypic variability observed in grapevine. *Theor. Appl. Genet.* **126**, 2763–2776 (2013).
6. Emanuelli, F., Sordo, M., Lorenzi, S., Battilana, J. & Grando, M. S. Development of user-friendly functional molecular markers for *VvDXS* gene conferring muscat flavor in grapevine. *Mol. Breed.* **33**, 235–241 (2014).
7. Costantini, L. *et al*. Somatic variants for seed and fruit set in grapevine. *BMC Plant Biol.* **21**, 1–33 (2021).
8. de Oliveira, G. L. *et al*. Exploring genetic diversity in North American *Vitis* species: screening for resistance loci and implications for utilization in molecular breeding programs. *Hortic. Environ. Biotechnol.* **66**, 1719–1740 (2025).
9. Lacombe, T. *et al*. Publication of the 3rd edition of the OIV ampelographic descriptors. In IVES (International Viticulture and Enology Society) Conference Series, *OIV* https://doi.org/10.58233/Xk2jsdGQ (2024).
10. Consorzio innovazione vite (CIVIT). https://www.civit.tn.it/.
11. Roeckel, F. *et al*. *Vitis* International Variety Catalogue (*V*IVC). www.vivc.de (2025).
12. Search Accessions GRIN-Global. https://npgsweb.ars-grin.gov/gringlobal/search (2025).
13. The European Vitis Database (EU-Vitis). http://www.eu-vitis.de/index.php (2025).
14. CREA Viticulture and Enology (CREA-VE) Attività di analisi per la caratterizzazione varietale della vite. https://www.crea.gov.it/en/web/viticoltura-e-enologia/-/attivit%C3%A0-di-analisi-per-la-caratterizzazione-varietale-della-vite (2025).
15. FAO. Status of germplasm characterization, by crop group for selected genera. In *The Third Report on the State of the World's Plant Genetic Resources for Food and Agriculture* https://doi.org/10.4060/cd4711en (FAO, Rome, Italy, 2025).
16. Minimum Information About Plant Phenotyping Experiments (MIAPPE). https://www.miappe.org/ (2025).
17. Planteome 2024 Update: Reference Ontologies and Knowledgebase for Plant Biology | *Nucleic Acids Research* | Oxford Academic. https://academic.oup.com/nar/article/52/D1/D1548/7459474.
18. Vitis Ontology CO_356. https://agroportal.lirmm.fr/ontologies/CO_356.
19. Crop Ontology. https://cropontology.org/.

20. Selby, P. *et al*. BrAPI—an application programming interface for plant breeding applications. *Bioinformatics* **35**, 4147–4155 (2019).
21. FAO/IPGRI Multi-Crop Passport Descriptors. *Alliance Bioversity International - CIAT* https://alliancebioversityciat.org/publications-data/fao-ipgri-multi-crop-passport-descriptors-mcpd.
22. *International Code of Nomenclature for Algae, Fungi, and Plants*. vol. 159 (Koeltz Botanical Books, 2018).
23. Wen, J. *et al*. Chloroplast phylogenomics of the New World grape species (*Vitis*, Vitaceae). *J. Syst. Evol.* **56**, 297–308 (2018).
24. Weakley, A. S. *Flora of the Southeastern United States - 2020 Edition*.
25. Doyle, J. DNA Protocols for Plants. In *Molecular Techniques in Taxonomy* (eds Hewitt, G. M., Johnston, A. W. B. & Young, J. P. W.) 283–293 https://doi.org/10.1007/978-3-642-83962-7_18.(Springer Berlin Heidelberg, Berlin, Heidelberg, 1991).
26. Thomas, M. R. & Scott, N. S. Microsatellite repeats in grapevine reveal DNA polymorphisms when analysed as sequence-tagged sites (STSs). *Theor. Appl. Genet.* **86**, 985–90 (1993).
27. Bowers, J. E., Dangl, G. S., Vignani, R. & Meredith, C. P. Isolation and characterization of new polymorphic simple sequence repeat loci in grape (*Vitis vinifera* L.). *Genome* **39**, 628–33 (1996).
28. Bowers, J. E., Dangl, G. S. & Meredith, C. P. Development and Characterization of Additional Microsatellite DNA Markers for Grape. *Am. J. Enol. Vitic.* **50**, 243–246 (1999).
29. Sefc, K. M., Regner, F., Turetschek, E., Glössl, J. & Steinkellner, H. Identification of microsatellite sequences in *Vitis riparia* and their applicability for genotyping of different *Vitis* species. *Genome* **42**, 367–73 (1999).
30. Zini, E. *et al*. R-Loci Arrangement Versus Downy and Powdery Mildew Resistance Level: A Vitis Hybrid Survey. *Int. J. Mol. Sci.* **20**, 3526 (2019).
31. Peakall, R. & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* **28**, 2537–2539 (2012).
32. Kalinowski, S. T., Taper, M. L. & Marshall, T. C. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* **16**, 1099–1106 (2007).
33. OIV. *2nd Edition of the OIV Descriptor List for Grape Varieties and Vitis Species*. (2009).
34. Coombe, B. G. Growth Stages of the Grapevine: Adoption of a system for identifying grapevine growth stages. *Aust. J. Grape Wine Res.* **1**, 104–110 (1995).
35. Emanuelli, F. *et al*. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol.* **13**, 39 (2013).
36. Migliaro, D. *et al*. Grapevine non-*vinifera* genetic diversity assessed by simple sequence repeat markers as a starting point for new rootstock breeding programs. *Am. J. Enol. Vitic.* **70**, 390–397 (2019).
37. Augusto, D. *et al*. Grapevine Diversity and Genetic Relationships in Northeast Portugal Old Vineyards. *Plants* **10**, 2755 (2021).
38. Cretazzo, E. *et al*. Genetic Characterization by SSR Markers of a Comprehensive Wine Grape Collection Conserved at Rancho de la Merced (Andalusia, Spain). *Plants* **11**, 1088 (2022).
39. Arslan, N. *et al*. Genetic Diversity and Population Structure Analysis of Anatolian Kara Grapevine (*Vitis vinifera* L.) Germplasm Using Simple Sequence Repeats. *Horticulturae* **9**, 743 (2023).
40. de Oliveira, G. L. *et al*. Genetic diversity, population structure and parentage analysis of Brazilian grapevine hybrids after half a century of genetic breeding. *Sci. Hortic.* **311**, 111825 (2023).
41. Perko, A. *et al*. Monitoring and Genotyping of Wild Grapevine (*Vitis vinifera* L. subsp. *sylvestris*) in Slovenia. *Plants* **13**, 1234 (2024).

Acknowledgements

The authors are grateful to Paolo Fontana and Luca Bianco (FEM) for the initial contribution to dataset structure and FAIRification process.

Funding

This overarching study was initially (2020-2021) supported by the Autonomous Province of Trento (AdP-UGMGV) and mainly (2022-2025) supported by the Agritech National Research Center and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D.D. 1032 17/06/2022, CN00000022).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-08482-8.

Correspondence and requests for materials should be addressed to S.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.