

DATABASE

Open Access



Construction of a comprehensive library of repeated sequences for the annotation of *Citrus* genomes

Delphine Giraud^{1*} , Nathalie Choisne² , Marilyne Summo^{3,4} , Stéphanie Sidibe-Bocs^{3,4} , Héléna Vassilieff² , Gilles Costantino¹, Gaetan Droc^{3,4} , Pierre-Yves Teycheney^{5,6} , Florian Maumus² , Patrick Ollitrault^{3,4}  and François Luro¹ 

Abstract

Background The comprehensive annotation of repeated sequences in genomes is an essential prerequisite for studying the dynamics of these sequences over time and their involvement in gene regulation. Currently, the diversity of repeated sequences in *Citrus* genomes is only partially characterized because the annotations have been performed using heterogeneous bioinformatics tools, each with its specificity and dedicated only to the annotation of transposable elements.

Results We combined complementary repeat-finding programs including REPET, CAULIFINDER, and TAREAN, to enable the identification of all types of repetitive sequences found in plant genomes, including transposable elements, endogenous caulimovirids, and satellite DNAs. A fine-grained annotation method was first developed to create a consensus sequence library of repeated sequences identified in the genome assemblies of *C. medica*, *C. micrantha*, *C. reticulata*, and *C. maxima*, the four ancestral parental species involved in the formation of economically valuable cultivated *Citrus* varieties. A second, faster annotation method was developed to enrich the dataset by adding new repeated sequences retrieved from genome assemblies of other *Citrus* species and closely related species belonging to the *Aurantioideae* subfamily. The final reference library contains 3,091 consensus sequences, of which 94.5% are transposable elements. The diversity of endogenous caulimovirids was characterized for the first time within the genus *Citrus*, contributing 160 consensus sequences to the final reference library. Finally, 10 satellite DNAs were also identified.

Conclusion Combining multiple repeat detection methods enables the comprehensive annotation of all repeated sequences in *Citrus* genomes. Using the final reference library reported in this work will improve our understanding of the dynamics of repeated sequences during *Citrus* speciation, particularly following the genome duplication and hybridization events that led to modern cultivars. The exploration of repeat position insertions along chromosomes using the developed web interface, RepeatLoc Citrus, will also make it possible to further investigate the role of transposable elements and endogenous caulimovirids in genome structure and gene regulation in *Citrus* species.

*Correspondence:
Delphine Giraud
delphine.giraud79@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2025. **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/>.

Keywords Citrus, Transposable elements, Endogenous caulimovirids, Satellite DNAs, Genome annotation

Background

Over the last two decades, cost reductions and improvements in DNA sequencing technologies have led to an exponential increase in plant genome assemblies [1]. This trend has stimulated the development of several bioinformatics tools dedicated to the annotation of repetitive sequences, also known as the “repeatome”, which represent a significant fraction of most eukaryotic genomes [2]. The characterization of repeated sequences has revealed their ubiquity in plant genomes [3–6]. Transposable elements (TEs) are genetic sequences that can replicate in large numbers and insert into chromosomes [7–9]. Initially dismissed as “junk DNA”, TEs are now recognized as key players in genome dynamics and species evolution [10]. In many plants, they have contributed to the emergence of new phenotypes by acting as *cis*-regulatory elements, modifying the chromatin landscape around genes, mediating gene duplication, and disrupting coding sequences [6, 11–13]. For example, TE insertions have been linked to variation in the fruit color of Chardonnay grape clones [14] or to sex determination in melon [15]. In tomato, the elongated fruit phenotype has been linked to a TE-mediated duplication of the SUN locus [16], and the yellow color of fruit flesh in several genotypes has been linked to a TE insertion in the PSY1 locus [17]. Recent studies have demonstrated the involvement of TEs in the variation of key agronomic traits in maize and rice species and confirmed the need to take them into account when breeding new varieties [18–21]. Satellite DNAs, or long tandem repeats, are also found in plant genomes, particularly at centromeres, but are sometimes concentrated in other specific chromosomal regions [22–24]. Unlike TEs, they cannot self-replicate, but unequal recombination during meiosis and polymerase replication splicing contribute to their accumulation in large numbers in genomes [25]. They are mainly involved in genome structure and chromosome stability by participating in the formation of heterochromatin. In addition to TEs and satellite DNAs, plant genomes often contain endogenous caulimovirids (ECVs), another class of repetitive DNA sequences [26]. ECVs result from the integration of genomic sequences from members of the viral family *Caulimoviridae* during infections [27–30]. Their integration into their host genomes is not an obligatory step in their replication cycle but is thought to occur during the repair of DNA breaks [26]. To date, their role in the structure and evolution of plant genomes is poorly understood.

It is well established that repeated sequences such as TEs play a role in genomic structural variations causing phenotypic variations observed in vegetatively

propagated crops [19, 31, 32]. The phenotypic diversity observed in sweet orange cultivars has been directly linked to specific TE insertions near host genes and shown to be involved in traits such as blood color, apomixis, and acidity changes [33–39]. Given the high economic value of citrus fruits, with an annual production of over 100 million tons in 2023, there is a particular relevance in integrating repeatome annotation data into breeding schemes for new citrus varieties [40, 41]. Whole genome annotations of repetitive sequences have been achieved for several *Citrus* spp., but further analyses showed that only a fraction of the repeated sequences were identified. Indeed, retrotransposons, LINEs (long interspersed nuclear elements), and MITEs (miniature inverted repeat transposable elements) were best characterized, but their estimated abundances in the analyzed genomes depended on the annotation methods and the quality of genome assemblies. For example, in sweet orange (*C. x aurantium* var. *sinensis*), the relative amount of *Copia* and *Gypsy* elements varies between 15.3% and 24.2% of the genome according to the different annotation methods used [38, 39, 42–44]. Other types of TEs, such as Helitrons and SINEs, remain to be fully identified, resulting in an underestimation of their abundance and diversity. Furthermore, the diversity of ECVs within *Citrus* genomes has not been comprehensively determined. However, several studies have indicated their relatively high density in sweet oranges, at 2.3 copies per Mb, compared with an average of 0.2 copies per Mb in the genomes of other seed plants [29, 45].

In this context, our main objective is to determine the comprehensive diversity of repeated sequences in the genomes of *Citrus* spp. and related species from the *Aurantioideae* subfamily through the annotation of TEs, ECVs, and satellite DNAs (especially macrosatellites). Using several pipelines, two complementary annotation methods were developed to construct a reference consensus sequence library that is representative of the repeated sequences found in the genomes of primitive, wild, and cultivated *Citrus* spp. [46–48]. It is anticipated that this library will serve as a gold standard for the annotation of *Citrus* genomes, which will facilitate the study of the insertion dynamics of repeated sequences and their roles in genome structure and clonal variation in this botanical genus.

Development of a *Citrus* reference library

Fine-grained annotation of transposable elements (TEs)

We performed a fine-grained annotation of the repeated sequences found in the genome assemblies of the four ancestral taxa *C. medica*, *C. micrantha*, *C. reticulata*,

and *C. maxima*, which are the progenitors of the major cultivated citrus varieties [46, 49] (Fig. 1; Table 1). This annotation was first implemented on genome assemblies for each progenitor using the REPET v3.0 package [50–52]. According to statistics of assembly data, only a subgenome was analyzed for each species. These subgenomes are made up of “virtual contigs” or “chunks” obtained by removing stretches of > 11 undefined bases (Ns) to exclude gaps in these sequences. Only chunks longer than 500 kbp in *C. medica* and chunks longer than 100 kbp in *C. micrantha*, *C. reticulata*, and *C. maxima* were conserved, covering 93–96% of the genome assemblies. The REPET TEdenovo pipeline was used to detect all repeated sequences in the subgenomes and to generate libraries of consensus sequences representative of each TE family (setting parameters: at least 3 sequences per group, Grouper, Recon, and Piler clustering, no remove redundancy). The consensus sequences were then classified with the PASTEC classifier (included in the TEdenovo pipeline [52]) according to structural and functional features based on characterized TEs from the RepBase23.12 database [53, 54], the library of profiles from Pfam32.0 [55] and the GyDB2.0 database [56] specially formatted for REPET. All consensus sequences classified as simple sequence repeats (SSR), rDNA, potential host genes (PHG; containing host gene Pfam domains), and unclassified sequences built with less than 10 copies per cluster were removed from the TE libraries. After filtering, TE libraries of 13,483, 10,957, 11,451, and 12,958 consensus sequences were obtained for *C. medica*, *C. micrantha*, *C. reticulata*, and *C. maxima*, respectively (Fig. 1). Each TE library was used to annotate the genome assembly from which it was generated, using the TEannot pipeline included in REPET (default parameters). This step resulted in an estimated TE coverage of 56% in *C. medica*, 51% in *C. micrantha*, 52% in *C. reticulata*, and 55% in *C. maxima*. To refine the annotation of TEs, TE libraries were again filtered to retain only consensus sequences with at least one full-length fragment (FLF) in the genome (i.e. a fragment covering more than 95% of the consensus sequences), and then a second TEannot run was performed. This iterative process resulted in the filtered TE libraries of 3,579, 4,469, 4,427, and 3,834 consensus sequences for *C. medica*, *C. micrantha*, *C. reticulata*, and *C. maxima*, respectively, with a reduced TE coverage of 1–3% depending on the species (Fig. 1). All consensus sequences were manually curated using copy coverage plots, structural and functional features (ORFs, tandem repeats, polyA tail, protein domain HMM profiles, SSR, and BLAST results), and MCL clustering. This manual curation allowed us to classify some TEs at the family level, identify ECVs that would otherwise be

automatically classified as TEs, and reclassify some inaccurately classified consensus sequences (Additional files 2 and 3). To reduce intra-species redundancy, consensus sequences that fully aligned to a longer other consensus sequence with an identity greater than 80% were removed from each TE library following the Wickers’ rule [8] and using Cd-hit v. 4.8.1 [57].

In addition to REPET, several other TE detection programs were used. First, the EDTA pipeline [65] was launched on the four genomes to validate the structure of the REPET predictions and to solve annotation conflicts encountered with the PASTEC classifier. Consensus sequences generated using EDTA and qualified as Helitron were specifically studied. Their structures were confirmed using DANTE, the TE protein domain identification tool included in the RepeatExplorer pipeline [66], and novel Helitrons were added to the TE libraries. In parallel, the four *Citrus* assemblies were analyzed using a collection of TE structure-based tools, including LTR_STRUC searching for LTR retrotransposons [67], MUSTv2, MITE-Hunter searching for MITEs [68, 69], and SINE-Finder searching for short interspersed nuclear elements (SINEs) [70]. For each *Citrus* species, these newly detected consensus sequences were added to the respective TE libraries, and redundant sequences were filtered out using Cd-hit with the same parameters as above.

Finally, the four TE libraries were combined, and inter-specific redundancy was reduced using Cd-hit clustering. This resulted in a fine-grained annotation library of 2,875 consensus sequences representative of the diversity of TEs (2,720 consensus sequences) and ECVs (155 consensus sequences) found in the four ancestral *Citrus* taxa (Fig. 1, Additional file 1).

Annotation of endogenous caulimovirid sequences (ECVs)

Special attention was dedicated to the identification of ECVs. The 155 ECVs consensus sequences identified in the four ancestral taxa using the TEdenovo pipeline were included in the fine-grained annotation library (Fig. 1). During consensus sequence quality checks, sequences containing a movement protein domain (MP), which is a hallmark of *Caulimoviridae* genomes, were temporarily classified as ECVs. To confirm their classification as ECVs, the 155 consensus sequences were compared with the custom library of reverse transcriptase (RT) protein sequences from known caulimovirids and LTR retrotransposons used by CAULIFINDER, a pipeline specifically designed for the specific annotation of ECVs [71]. Consensus sequences displaying a best hit against caulimovirid RTs with e-values < 1e-06 following BLASTx analyses were retained as ECVs in the fine-grained annotation library.

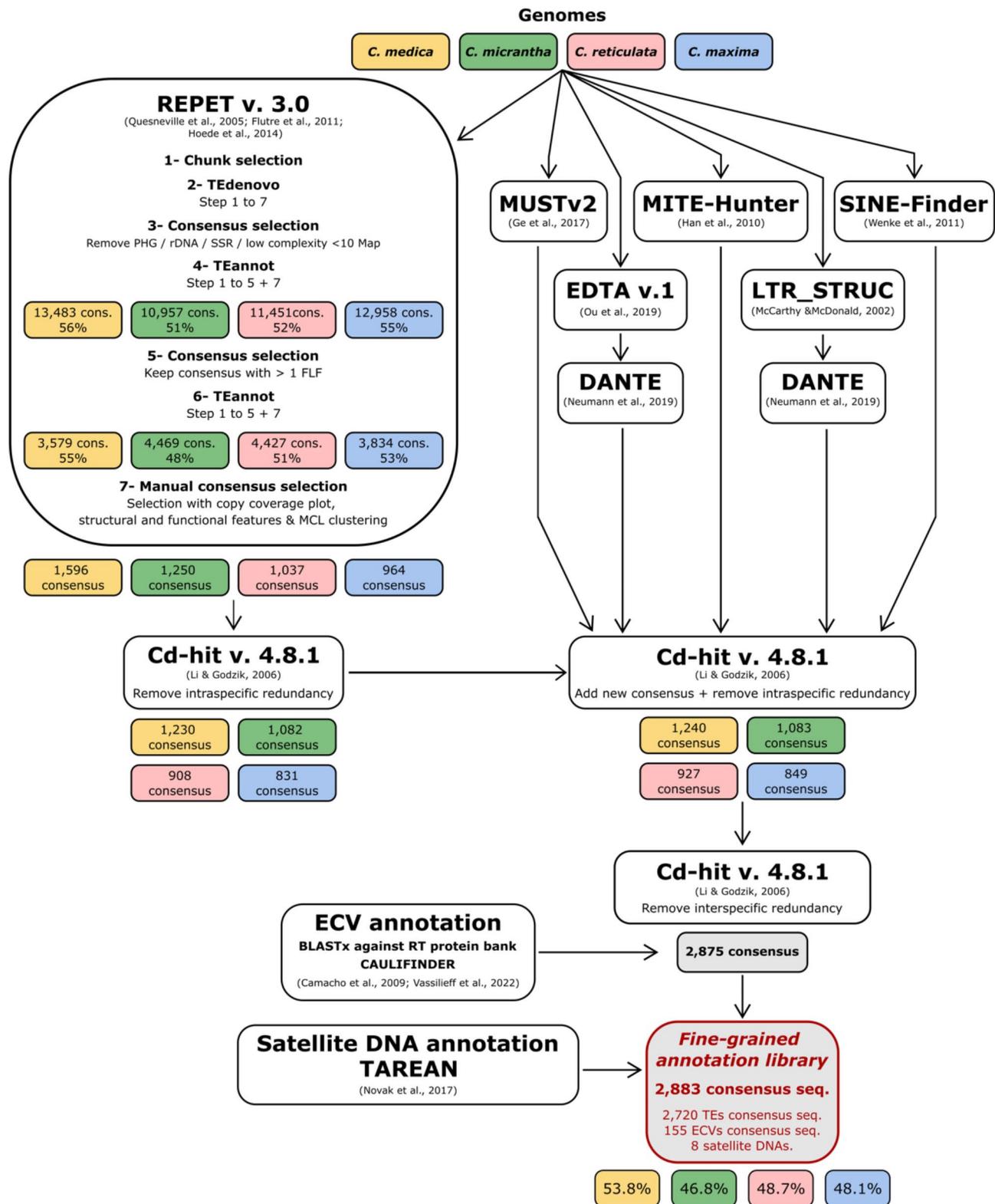


Fig. 1 Schematic representation of the fine-grained method employed for *de novo* annotation of repeated sequences in the four ancestral *Citrus* taxa *C. medica*, *C. micrantha*, *C. reticulata*, and *C. maxima*. Results from the different bioinformatics tools REPET, EDTA, MUST, TAREAN, and CAULIFINDER were combined to produce a reference fine-grained annotation library of 2,883 consensus sequences representative of the diversity of all repeated sequences retrieved in *Citrus*. The numbers shown in the species-specific colored boxes indicate the number of consensus sequences and their genome coverage obtained using the TEannot pipeline. ECV: Endogenous Caulimovirid element; FLF: Full Length Fragment; PHG: Potential Host Gene; rDNA: ribosomal DNA; SSR: Simple Sequence Repeat; TE: Transposable Element

Table 1 List of the genome assemblies used for annotating repeated sequences

Scientific Name according to Ollitrault classification [58]	Scientific Name according to Tanaka classification [59]	Organism Common Name	Abbreviation in library	Assembly Accession	Source	Reference
<i>Citrus medica</i>	<i>Citrus medica</i>	Citron	CITME	PRJEB78498	EBI	[49]
<i>Citrus micrantha</i>	<i>Citrus micrantha</i>	Papeda micrantha	CITMI	PRJEB78515	EBI	[49]
<i>Citrus reticulata</i>	<i>Citrus reticulata</i>	Wild mandarin Cleopatra	CITRE	PRJEB78537	EBI	[49]
<i>Citrus maxima</i>	<i>Citrus maxima</i>	Pummelo	CITMA	GCA_002006925.1	GenBank	[46]
<i>Citrus x aurantium</i> var. <i>clementina</i>	<i>C. clementina</i>	Clementine	CITCL	GCA_000493195.1	GenBank	International Citrus Genome Consortium
<i>Citrus x aurantium</i> var. <i>sinensis</i> haplotype A1	<i>C. sinensis</i>	Valencia Sweet orange	CITSL_A1	GCA_022201045.1	GenBank	[60]
<i>Citrus x aurantium</i> var. <i>sinensis</i> haplotype B1	<i>C. sinensis</i>	Valencia Sweet orange	CITSL_B1	GCA_022201065.1	GenBank	[60]
<i>Citrus x limon</i> var. <i>limon</i> haplotype Primary	<i>C. limon</i>	Lemon	CITLL_Pri	CGD21001	CGD	[61]
<i>Citrus x limon</i> var. <i>limon</i> haplotype Alternative	<i>C. limon</i>	Lemon	CITLL_Alt	CGD21001	CGD	[61]
<i>Citrus cavaleriei</i>	-	Ichang papeda	CCAV	GCA_002013975.2	GenBank	[46]
<i>Citrus trifoliata</i>	<i>Citrus trifoliata</i>	Hardy orange	CTRI	GCA_018350135.1	GenBank	[62]
<i>Citrus x aurantium</i> var. <i>unshiu</i>	<i>Citrus unshiu</i>	Satsuma mandarin	CUNS	GCA_002897195.1	GenBank	[63]
<i>Citrus hindsii</i>	<i>Citrus hindsii</i>	Hong Kong Kumquat	CHIN	PRJNA735863	GenBank	[39]
<i>Citrus maxima</i> (L.) Osbeck.	<i>Citrus maxima</i> (L.) Osbeck.	Pummelo cv.'Cupi Majjayou'	CMAX	PRJNA318855	GenBank	[64]
<i>Murraya koenigii</i>	-	Curry leaf	MURKO	GCA_025617385.1	GenBank	-
<i>Atalantia buxifolia</i>	-	-	ABUX	GCA_002013935.1	GenBank	[46]

Annotation of satellite DNAs

Finally, satellite DNAs (tandem repeats) were identified using TAREAN [72] (Fig. 1). Briefly, this pipeline performs a graph-based clustering from libraries of paired-end Illumina reads. Clusters with circular topology are isolated, and monomer units of the satellite DNAs are characterized. For each ancestral taxon, a sample of 1 million paired-end reads was randomly extracted from Illumina libraries used for genome assemblies (Table 1) and analyzed by TAREAN. To avoid redundancy, satellite DNAs detected in the four genome assemblies were aligned with each other using BLASTn [73] and only those with sequence identity < 80% and e-values < 1e-06 were conserved. In total, eight different satellite DNAs were identified and added to the fine-grained annotation library (Table 2, Additional file 1).

Table 2 Classification and number of repeated sequences included in the final reference library

		Number of consensus sequences
Retrotransposons (Class I elements)		
LTR <i>Copia</i> (RLC)	<i>Tork</i>	302
	<i>Sire</i>	191
	<i>Retrofit</i>	379
	<i>Oryco</i>	47
	Unclassified RLCs	28
LTR <i>Gypsy</i> (RLG)	<i>Athila</i>	216
	<i>Tat</i>	168
	<i>Crm</i>	50
	<i>Reina</i>	118
	<i>Del</i>	48
	<i>Galadriel</i>	24
	Unclassified RLGs	5
LINEs (RIX)		320
SINEs (RSX)		5
Unclassified Class I		4
DNA transposons (Class II elements)		
TIRs	<i>Mutator</i> (DTM)	279
	<i>hAT</i> (DTA)	223
	<i>CACTA</i> (DTC)	107
	<i>Pif-Harbinger</i> (DTH)	74
	<i>Tc1-Mariner</i> (DTT)	20
	Unclassified TIRs (DTX)	1
MITEs	Unclassified MITEs (DXX-MITE)	175
	<i>hAT</i> (DTA-MITE)	11
	<i>Mutator</i> (DTM-MITE)	30
	<i>Pif-Harbinger</i> (DTH-MITE)	3
Helitron (DHX)		26
Unclassified TEs		67
TOTAL Transposable Elements		2 921
ECVs (Endogenous Caulimovirids)		160
Satellite DNAs		10
TOTAL Repeated sequences		3 091

Enrichment of the fine-grained annotation library with repeated sequences from other *Citrus* and related species

To be more inclusive of the diversity of the repeated sequences found across the *Citrus* genus, we extended our search to the assembled genomes of 8 other *Citrus* species and two related species *Atalantia buxifolia* and *Murraya koenigii* (Table 1). A second, faster annotation method was developed to reduce the execution time compared to the fine-grained method (Fig. 2). This method uses the same combination of repeat-finding programs but skips some steps and uses the fine-grained annotation library as a reference. Firstly, the TEdenovo pipeline was used on DNA chunks longer than 20kbp, deriving from the sampling of the genome assemblies. The resulting consensus sequence libraries were filtered by removing PHG, rDNA, SSR sequences, and unclassified sequences built with < 10 copies per cluster as described above. In addition to the REPET pipeline, TAREAN, and MITE-Hunter were also used to annotate the satellite DNAs and the MITEs in each genome assembly, respectively. Using BLASTn, consensus sequences without sequence homology to TEs, ECVs, or satellite DNAs already present in the fine-grained annotation library were retained (setting parameters: < 80% of identity, e-value < 1e-06). They were manually curated and classified as described above and then added to the fine-grained annotation library.

The annotation of the 10 extra genomes using the faster method led to the identification of 208 additional new consensus sequences, corresponding to 201 TEs, 5 ECVs, and 2 satellite DNAs (Additional file 1). In total, the final reference library contains 3,091 consensus sequences representative of the diversity of repeated sequences, including 2,921 TEs (94.5%), 160 ECVs (5.2%), and 10 satellite DNAs ranging in size from 141 to 181 bp (0.3%) (Fig. 3; Table 2) [74]. Among annotated TEs, 1,905 consensus sequences are retrotransposons (Class I elements) and 949 are DNA transposons (Class II elements).

A benchmark for the annotation of repeated sequences in *Citrus* genomes and their role in genome dynamics

There is currently no standard method for the annotation of all repeated sequences in assembled genomes. Multiple bioinformatic programs can be used, but each has its own specificity, resulting in variable precision levels in the annotation, depending on the nature of the repeats. Here, we propose two complementary methods for the annotation of all repeated sequences in *Citrus* genomes, including TEs, ECVs, and satellite DNAs (Figs. 1 and 2). These annotation methods allowed us to construct the first comprehensive reference consensus sequence library of *Citrus* repeats. This final reference library paves the way to the accurate, reproducible, and standardized quantification of repeated sequences in *Citrus* genomes. Just as the fine-grained method can be reused to annotate repeated sequences in

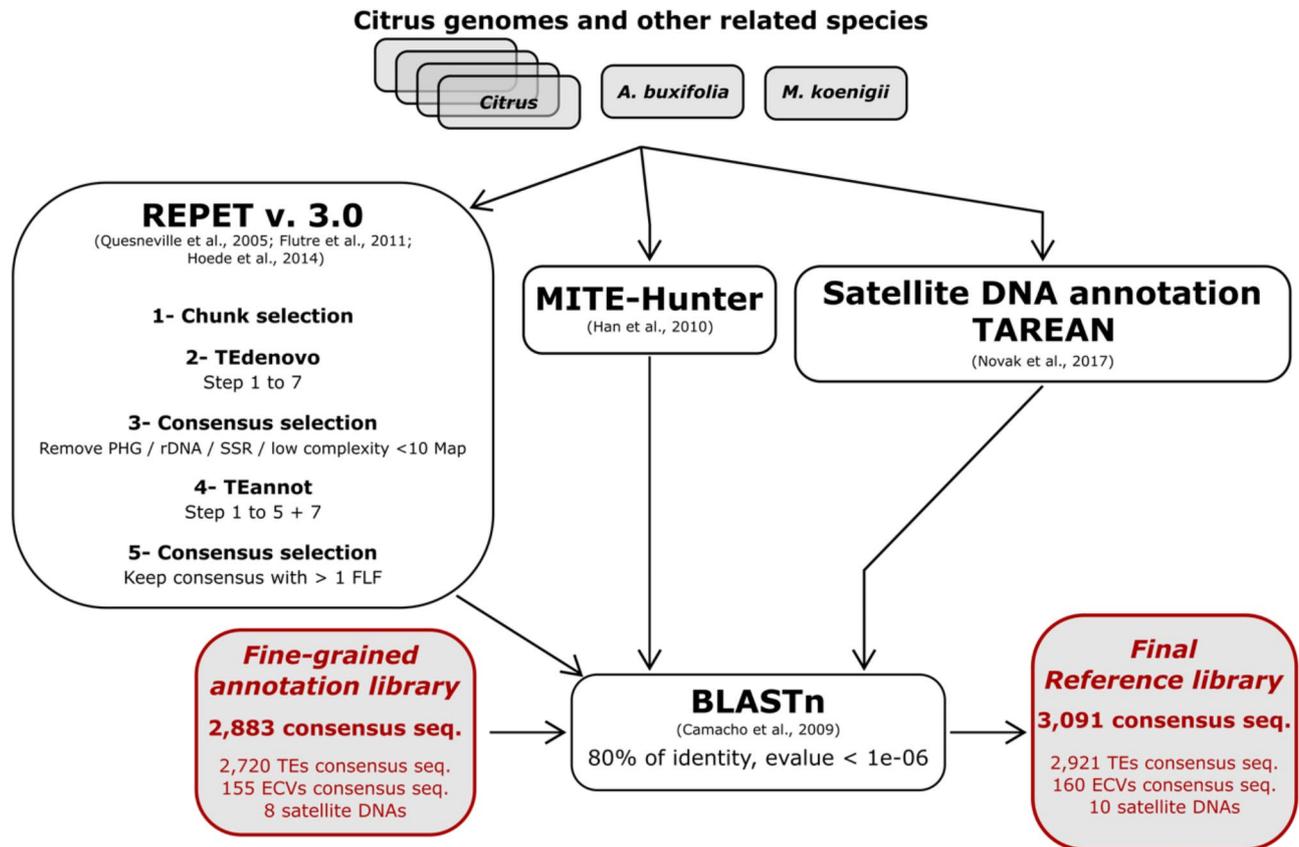


Fig. 2 Schematic representation of the method used to annotate repeated sequences in other *Citrus* species and related species. Compared to the fine-grained method, multiple steps were skipped to reduce execution time, and the fine-grained annotation library was used as a reference. A final reference library of 3,116 consensus sequences was built containing the diversity of TEs, ECVs, and satellite DNAs identified in *Citrus* and related species

other plant genomes (Fig. 1), the final reference library can be enriched by annotating newly assembled *Citrus* genomes using the second, faster method described in this paper (Fig. 2). The final reference library contains the diversity of repeated sequences identified in twelve citrus species either from the divergent speciation including the four ancestral species (citron, papeda micrantha, wild mandarin, and pummelo), or from the reticulate speciation with the four cultivated species, sweet orange, clementine satsuma mandarin and limon. In future, the quick annotation of new assembled citrus genomes will facilitate repeated sequences detection and improved their study in citrus species.

In 2019, Liu et al. [75] annotated five different *Citrus* genomes (*C. x aurantium* var. *sinensis*, *C. x aurantium* var. *clementina*, *C. medica*, *C. ichangensis* and *C. maxima*) and built a library of 450 consensus sequences, including 340 L retrotransposons and 110 MITEs. We compared this library with our final reference library using BLASTn (e-value < 1e-06). This comparison revealed that 423 of the 450 consensus sequences were common to both libraries, representing only 13.7% of the repeat diversity found in our final reference library (Fig. 4). The comparison showed that 45 consensus sequences from the study of Liu et al. were

incorrectly classified (e.g. 7 MITEs consensus sequences corresponded to LTR retrotransposons, 6 L retrotransposon sequences corresponded to satellite DNAs, and 27 consensus sequences had no homology with either sequences of our final reference library or sequences in the RepBase23.12 database (BLASTn; e-value < 1e-06). More recently, Wu et al. [76] developed a pan-genome TE library of 21,680 TEs consensus sequences based on the genome annotation of 14 *Citrus* and 6 related species, compared with our final reference library of 3,091 consensus repeated sequences from 14 *Citrus* and 2 related species. The sequence library of Wu et al. is unpublished, precluding its comparison with our final reference library. However, using seven times fewer consensus sequences, our library and method revealed the diversity of TEs, ECVs, and satellite DNAs found in *Citrus* spp.

The *Citrus* consensus sequence library developed in this study will significantly improve the study of the TEs, ECVs, and satellite DNAs dynamics in *Citrus*. It will enable the comparison of repeated sequence counts and locations in *Citrus* species that diverged following the allopatric speciation that occurred in the last 6–8 million years [48, 77]. The comparison of ancestral parental taxa and admixed species will also enable a better

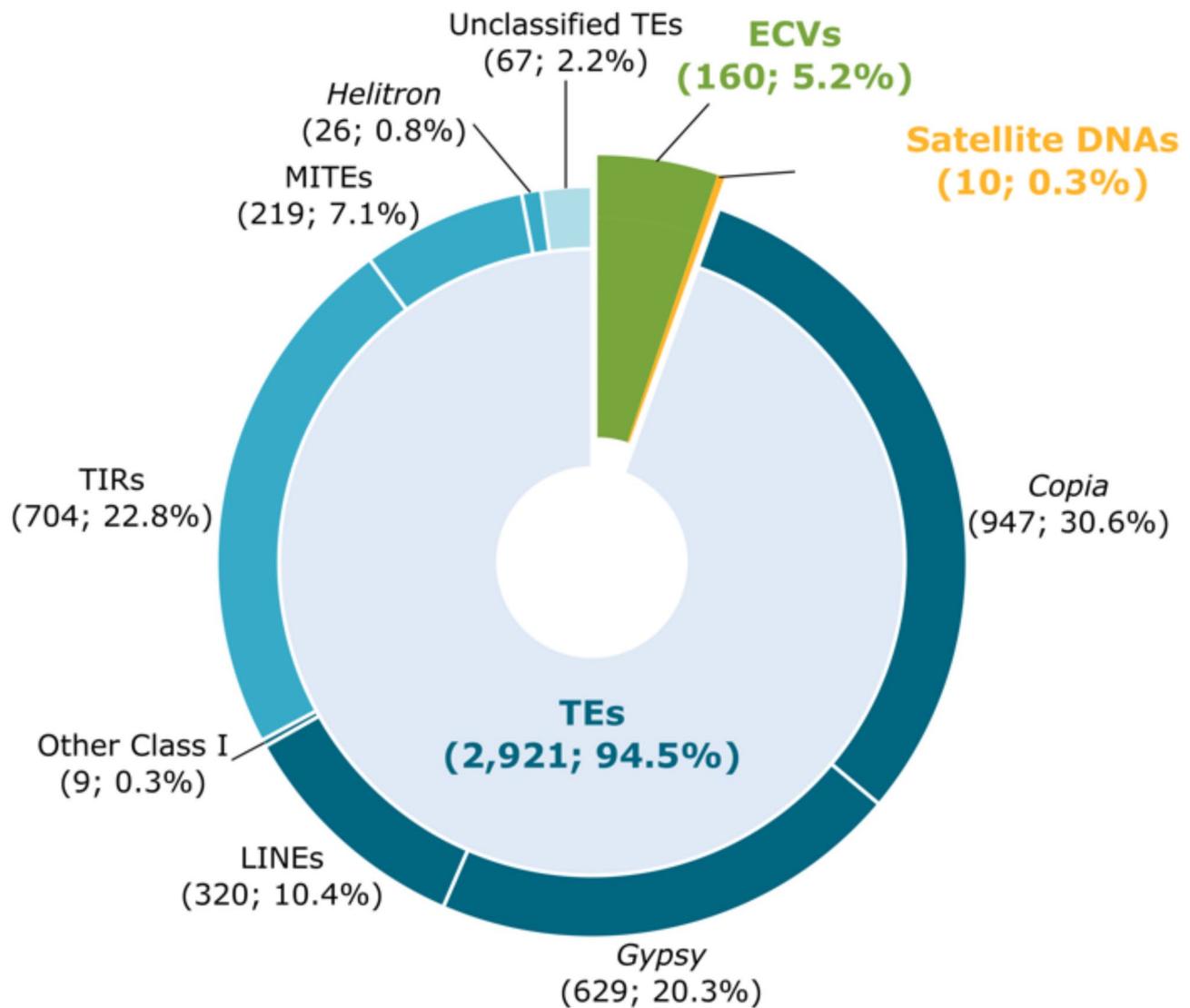


Fig. 3 Classification of the repeated sequences contained in the final reference library. Numbers in parentheses refer to the number of consensus sequences and the associated percentage

characterization of the consequences of hybridizations and whole genome duplication events on the dynamics of TEs [19, 78–81]. The tools and methods described in this paper could also help unravel the role of TEs and ECVs in some phenotypic variations observed in vegetatively propagated crops such as sweet oranges [34–37, 39]. Finally, the comprehensive annotation of ECVs in *Citrus* genomes will allow a better understanding of their putative role in genome architecture, gene regulation, and their integration dynamics in plant genomes.

RepeatLoc Citrus: visualize and explore the repeatome of the *Citrus* genomes

To enhance the study of repeated sequences in *Citrus* genomes, we developed an interface to visualize and analyze TEs, ECVs, and satellite DNAs along the nine

chromosomes of the four *Citrus* ancestor species. This interface, called RepeatLoc Citrus, has been implemented through the Citrus Genome Hub (<https://citrus-tools-genome-hub.southgreen.fr/repeatloc/>; accessed 4 April 2025), a publicly available web portal that offers the possibility to study and compare the genomes of *Citrus* species. Following the selection of the query *Citrus* species, users can visualize the density and positions of repeated sequences along the chromosomes (Fig. 5). For each type of repeated sequence, the number of copies and their coverage on each chromosome is available from arborescent entries corresponding to the different levels of TEs, ECVs, and satellite DNA classification. Users can also explore specific positions on chromosomes and extract detailed information about genes and repeated sequences by following the link to the Genome Browser

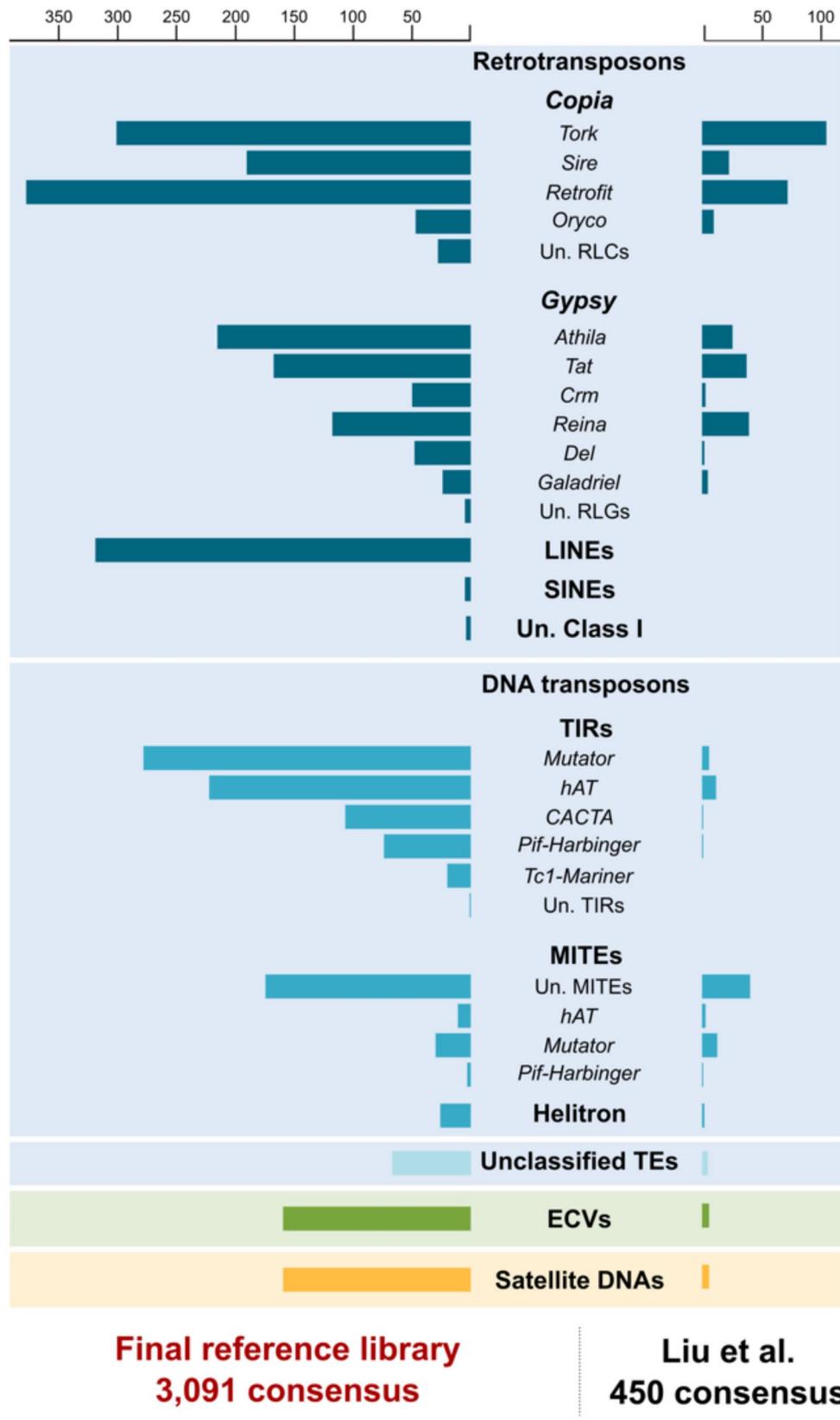


Fig. 4 Comparison of repeated sequences composing the final reference library of *Citrus* and the library generated by Liu et al. [75]. Horizontal bars represent the number of consensus sequences identified for each TE family, ECVs, or satellite DNAs in the final reference library (left) and in the Liu et al. library (right, classified according to best-hit match with the final reference library)

a.

Repeat_{loc} Citrus

Citrus medica

Search keyword... Ex: RE123 (1)

Search by location... Ex: CITME_001:1000000..2000000

(2) Level 0: root
Satellite_DNA
TE
EVE

Level 1: root > TE
LTR
MITE
TIR
LINE
Unclass

Level 2: root > TE > LTR
Copia
Gypsy

Level 3: root > TE > LTR > Copia
Tork
Retrofit
Sire
Oryco
Unknown

Download Tork annotations

(3) CITME_001 Count: 2992, Coverage: 3018765 bp 8.29%
CITME_002 Count: 2882, Coverage: 2402965 bp 6.19%
CITME_003 Count: 4798, Coverage: 4234690 bp 8.20%
CITME_004 Count: 2786, Coverage: 2367446 bp 6.78%
CITME_005 Count: 3394, Coverage: 3068695 bp 8.47%
CITME_006 Count: 2541, Coverage: 2423079 bp 7.59%
CITME_007 Count: 3109, Coverage: 2683951 bp 7.91%
CITME_008 Count: 3278, Coverage: 3345223 bp 9.00%
CITME_009 Count: 3027, Coverage: 3050812 bp 9.14%

(4) Total Count: 28807, Coverage: 26595626 bp 7.95%

(6) Show selected area on Jbrowse: CITME_009 24194791 28844324

(5) Selected area count: 411, Coverage: 453055 bp, 1.36%

Name	Start	Stop	Length
ms36139_CITRE_RE877_RLC_tork	24223277	24230013	6736
ms36212_CITMI_RE726_RLC_tork	24230060	24230160	100
ms36122_CITMI_RE111_RLC_tork	24232467	24232642	175
ms36104_CITME_RE521_RLC_tork	24239740	24245234	5494
ms36128_CITMI_RE1074_RLC_tork	24245580	24245757	177
ms36126_CITMI_RE847_RLC_tork	24290784	24291123	339
ms36108_CITME_RE622_RLC_tork	24320790	24320989	199

b.

Citrus Genome Hub

Home Genomes Browse Tools Download About Us

Citrus medica

Available Tracks

- 1) Reference sequence
- 2) Reference Gene annotation
- 3) Comparative Genomics

Genome Track View Help

0 5,000,000 10,000,000 15,000,000 20,000,000 25,000,000 30,000,000

CITME_009 CITME_009:24684801..24870400 (185.6 Kb) Go

24,700,000 24,750,000 24,800,000

(7) Locus

(8) Tork

Fig. 5 (See legend on next page.)

(See figure on previous page.)

Fig. 5 Examples of *Citrus* repeatome exploration available on the RepeatLoc Citrus interface. **a.** Users can select the *Citrus* species (1) and the repeated sequences (2) they wish to localize on chromosomes, number of copies, and genome coverage is displayed (3). A specific region can be studied by clicking and dragging with the mouse on the chromosome (4) or by querying its positions (1). The results are displayed below (5), showing the size and position of the copies in the selected window. Data on the selected window can be downloaded using the link to the JBrowse (6). **b.** Screenshot of the JBrowse obtained following the link (6). Users can compare the position of the genes (7) and the repeated sequences (8) found in the region of interest

(JBrowse). The addition of new genome assemblies in future updates of RepeatLoc Citrus will strengthen the analysis of the repeatome within the *Citrus* genus.

Conclusion

In this study, we combined several dedicated pipelines to build a final reference library of 3,091 consensus sequences representing the diversity of *Citrus* transposable elements, endogenous caulimovirids, and satellite DNAs. This library provides important resources for studying the dynamics of repeated sequences in *Citrus* genomes and their role in phenotypic variations and speciation. To complement the existing citrus databases on genomic, expression and variation data (<https://www.citrusgenomedb.org/>; <http://citrus.hzau.edu.cn/>; <http://citgvd.cric.cn/home>; <http://www.orangeexpdb.com/>, accessed 29 March 2025), the development of the online interface, RepeatLoc Citrus (tool included in the Citrus Genome Hub), will improve the localisation of repeated sequences along the chromosomes and highlight their putative involvement in altering gene regulation.

Abbreviations

ECV	Endogenous Caulimovirid
EVE	Endogenous Viral Element
FLF	Full Length Fragment
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MITE	Miniature Inverted-repeat Transposable Element
PHG	Potential Host Gene
rDNA	Ribosomal DNA
SINE	Short Interspersed Nuclear Element
SSR	Simple Sequence Repeat
TE	Transposable Element
TIR	Terminal Inverted Repeat

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12863-025-01321-6>.

Additional file 1. Details about the nature and the origin of repeated sequences composing the final reference library.

Additional file 2. Script developed to the automatic curation of retransposons RLC, RLG and RLX.

Additional file 3. Script developed to the automatic curation of retransposons RIX.

Acknowledgements

Thanks to Carole Quinton and Cointreau for their involvement in this project. Thanks to Raphaël Morillon and Claire Billot for their help in realizing this project. Thanks to the Citrus BRC for allowing us to study the genetic diversity of the genus *Citrus* by giving us access to different varieties of citrus fruit [82].

Authors' contributions

DG: Conceptualization, Methodology, Data curation, Writing – original draft. NC: Methodology, Data curation, Writing – review & editing. SSB, GC, GD, and HV: Software, Validation, Data management. MS and GD: Development and deployment of the visualization tool RepeatLoc Citrus. PYT and FM: Writing – review & editing. PO: Conceptualization, Writing – review & editing. FL: Conceptualization, Funding acquisition, Writing – review & editing. All authors read and approved the final manuscript.

Funding

DG was financed by the BAP (Biologie et amélioration des Plantes) department of INRAE and the Rémy-Cointreau Company (Hesperides project). PYT was funded by the European Regional Development Fund (ERDF, contract REU005756) and Région Réunion through the project "Dispositif de partenariat en santé et biodiversité".

This work has been realized with the support of MESO@LR-Platform at the University of Montpellier, with Cirad financial support and technical support from the bioinformatics group of the UMR AGAP Institute, member of the French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform. The CIRAD has subscribed an institutional license from <https://phoenixbioinformatics.org/repbase> (ref: _00Do0J6b5_5001J10Do2W;ref: 07/23/2021-07/22/2024, access extended to 08/11/2024) to use the Repbase database v. 23.12 (<https://www.girinst.org/repbase>).

Data availability

The accession numbers of the genome assemblies used in this study are listed in Table 1. The final reference library generated in this study is available in the UR1390-AGAP Corse Dataverse (Data INRAE <https://entrepot.recherche.data.ovu.fr/dataverse/inrae>) at <https://doi.org/10.57745/ZVYEBU>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹UR AGAP Corse, INRAE, Institut Agro, CIRAD, University of Montpellier, San Giuliano F-20230, France

²URGI, INRAE, Université Paris-Saclay, Versailles F-78026, France

³UMR AGAP, CIRAD, Institut Agro, INRAE, University of Montpellier, Montpellier F-34060, France

⁴UMR AGAP, CIRAD, Montpellier F-34398, France

⁵CIRAD, UMR PVBMT, Saint Pierre, La Réunion F-97410, France

⁶UMR PVBMT, Université de la Réunion, Saint-Pierre de La Réunion F-97410, France

Received: 3 March 2025 / Accepted: 9 April 2025

Published online: 18 April 2025

References

1. Kersey PJ. Plant genome sequences: past, present, future. *Curr Opin Plant Biol*. 2019;48:1–8. <https://doi.org/10.1016/j.pbi.2018.11.001>.

2. Maumus F, Quesneville H. Deep investigation of *Arabidopsis thaliana* junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS ONE. 2014;9(4):e94101 Jordan IK, editor. <https://doi.org/10.1371/journal.pone.0094101>.
3. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65(1):505–30. <https://doi.org/10.1146/annurev-arplant-050213-035811>.
4. Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genom. 2016;17(1):624. <https://doi.org/10.1186/s12864-016-2980-z>.
5. Nicolas J, Peterlongo P, Tempel S. Finding and characterizing repeats in plant genomes. In: Edwards D, editor. Plant Bioinformatics, vol. 1374. New York: Springer New York; 2016. p. 293–337. https://doi.org/10.1007/978-1-4939-3167-5_17.
6. Mhiri C, Borges F, Grandbastien MA. Specificities and dynamics of transposable elements in land plants. Biology. 2022;11(4):488. <https://doi.org/10.3390/biology11040488>.
7. Kumar A, Bennetzen JL. Plant retrotransposons. Annu Rev Genet. 1999;33:479–532. <https://doi.org/10.1146/annurev.genet.33.1.479>.
8. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhou B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82. <https://doi.org/10.1038/nrg2165>.
9. Lee SI, Kim NS. Transposable elements and genome size variations in plants. Genomics Inf. 2014;12(3):87. <https://doi.org/10.5808/GI.2014.12.3.87>.
10. Huang Y, Lee YCG. Blessing or curse: how the epigenetic resolution of host-transposable element conflicts shapes their evolutionary dynamics. Proc R Soc B Biol Sci. 2024;291(2020):20232775. <https://doi.org/10.1098/rspb.2023.2775>.
11. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9(5):397–405. <https://doi.org/10.1038/nrg2337>.
12. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49–61. <https://doi.org/10.1038/nrg3374>.
13. Stefanov BA, Nowacki M. The roles of transposable elements in transgenerational inheritance and genome evolution. In: Witzany G, editor. Epigenetics in Biological Communication. Cham: Springer Nature Switzerland; 2024. p. 369–85. https://doi.org/10.1007/978-3-031-59286-7_18.
14. Kobayashi S. Retrotransposon-induced mutations in grape skin color. Science. 2004;304(5673):982–982. <https://doi.org/10.1126/science.1095011>.
15. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon-induced epigenetic change leads to sex determination in melon. Nature. 2009;461(7267):1135–8. <https://doi.org/10.1038/nature08498>.
16. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 2008;319(5869):1527–30. <https://doi.org/10.1126/science.1153040>.
17. Fray RG, Grierson D. Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression. Plant Mol Biol. 1993;22(4):589–602. <https://doi.org/10.1007/BF00047400>.
18. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nat Genet. 2011;43(11):1160–3. <https://doi.org/10.1038/ng.942>.
19. Vitte C, Fustier MA, Alix K, Tenaillon M. The bright side of transposons in crop evolution. Brief Funct Genomics. 2014;13(4):276–95. <https://doi.org/10.1093/bfpg/elu002>.
20. Bubb KL, Hamm MO, Min JK, Ramirez-Corona B, Mueth NA, Ranchalis J, et al. The regulatory potential of transposable elements in maize. 2024. Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.07.10.602892>. [cited 2024 Sep 12].
21. Li X, Dai X, He H, Lv Y, Yang L, He W, et al. A pan-TE map highlights transposable elements underlying domestication and agronomic traits in Asian rice. Natl Sci Rev. 2024;11(6). <https://academic.oup.com/nsr/article/doi/10.1093/nsr/nwae188/7687832>.
22. Plohl M, Luchetti A, Meštrović N, Mantovani B. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene. 2008;409(1–2):72–82. <https://doi.org/10.1016/j.gene.2007.11.013>.
23. Plohl M. Those mysterious sequences of satellite DNAs. Period Biol. 2010;112(4):403–10.
24. Garrido-Ramos MA. Satellite DNA in plants: more than just rubbish. Cytogenet Genome Res. 2015;146(2):153–70. <https://doi.org/10.1159/000437008>.
25. Plohl M, Mestrovic N, Mravinac B. Satellite DNA evolution. In: Garrido-Ramos MA, editor. Genome Dynamics. Basel: S. KARGER AG; 2012. p. 126–52. <https://doi.org/10.1159/000337122>.
26. Vassilief H, Geering ADW, Choise N, Teycheney PY, Maumus F. Endogenous Caulimovirids: fossils, zombies, and living in plant genomes. Biomolecules. 2023;13(7):1069. <https://doi.org/10.3390/biom13071069>.
27. Geering ADW, Maumus F, Copetti D, Choise N, Zwickl DJ, Zytynicki M, et al. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat Commun. 2014;5(1):5269. <https://doi.org/10.1038/ncomms6269>.
28. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet. 2014;46(3):270–8. <https://doi.org/10.1038/ng.2877>.
29. Diop SI, Geering ADW, Alfama-Depauw F, Loac M, Teycheney PY, Maumus F. Tracheophyte genomes keep track of the deep evolution of the *Caulimoviridae*. Sci Rep. 2018;8(1):572. <https://doi.org/10.1038/s41598-017-16399-x>.
30. Schmidt M, Seibt KM, Weber B, Schwarzacher T, Schmidt T, Heikam T. Broken, silent, and in hiding: tamed endogenous pararetroviruses escape elimination from the genome of sugar beet (*Beta vulgaris*). Ann Bot. 2021;128(3):281–99. <https://doi.org/10.1093/aob/mcab042>.
31. Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, et al. Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. PLoS ONE. 2012;7(3):e32973 Schönbach C, editor. <https://doi.org/10.1371/journal.pone.0032973>.
32. Munasinghe M, Read A, Stitzer MC, Song B, Menard CC, Ma KY, et al. Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. PLoS Genet. 2023;19(12):e1011086 Vitte C, editor. <https://doi.org/10.1371/journal.pgen.1011086>.
33. De Felice B. Transposable sequences in citrus genome: role of mobile elements in the adaptation to stressful environments. Tree Sci Biotechnol. 2009;3:79–86.
34. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. 2012;24(3):1242–55. <https://doi.org/10.1105/tpc.111.095232>.
35. Butelli E, Garcia-Lor A, Licciardello C, Las Casas G, Hill L, Recupero GR, et al. Changes in anthocyanin production during domestication of *Citrus*. Plant Physiol. 2017;173(4):2225–42. <https://doi.org/10.1104/pp.16.01701>.
36. Shimada T, Endo T, Fujii H, Nakano M, Sugiyama A, Daido G, et al. MITE insertion-dependent expression of CitRKD1 with a RWP-RK domain regulates somatic embryogenesis in citrus nucellar tissues. BMC Plant Biol. 2018;18(1):166. <https://doi.org/10.1186/s12870-018-1369-3>.
37. Borredá C, Pérez-Román E, Ibanez V, Terol J, Talon M. Reprogramming of retrotransposon activity during speciation of the genus *Citrus*. Genome Biol Evol. 2019;11(12):3478–95. <https://doi.org/10.1093/gbe/evz246>.
38. Wang L, Huang Y, Liu Z, He J, Jiang X, He F, et al. Somatic variations led to the selection of acidic and acidless orange cultivars. Nat Plants. 2021;7(7):954–65. <https://doi.org/10.1038/s41477-021-00941-x>.
39. Wang N, Song X, Ye J, Zhang S, Cao Z, Zhu C, et al. Structural variation and parallel evolution of apomixis in citrus during domestication and diversification. Natl Sci Rev. 2022;9(10):nwac114. <https://doi.org/10.1093/nsr/nwac114>.
40. Pereira Gonzatto M, Scherer Santos J. In: Pereira Gonzatto M, Scherer Santos J, editors. Introductory Chapter: world citrus production and research. Citrus Research - Horticultural and Human Health Aspects. IntechOpen(Rijeka); 2023. <https://doi.org/10.5772/intechopen.110519>.
41. USDA. Citrus: world markets and trade. United States Department of Agriculture; 2024; p. 12. Available from: <https://apps.fas.usda.gov/psdonline/circulars/citrus.pdf>.
42. Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, et al. The draft genome of sweet orange (*Citrus sinensis*). Nat Genet. 2013;45(1):59–66. <https://doi.org/10.1038/ng.2472>.
43. Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014;32(7):656–62. <https://doi.org/10.1038/nbt.2906>.
44. Yang L, Deng H, Wang M, Li S, Wang W, Yang H, et al. A high-quality chromosome-scale genome assembly of blood orange, an important pigmented sweet orange variety. Sci Data. 2024;11(1):460. <https://doi.org/10.1038/s41597-024-03313-0>.

45. Yu H, Wang X, Lu Z, Xu Y, Deng X, Xu Q. Endogenous pararetrovirus sequences are widely present in Citrinae genomes. *Virus Res.* 2019;262:48–53. <https://doi.org/10.1016/j.virusres.2018.05.018>.
46. Wang X, Xu Y, Zhang S, Cao L, Huang Y, Cheng J, et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat Genet.* 2017;49(5):765–72. <https://doi.org/10.1038/ng.3839>.
47. Talon M, Caruso M, Gmitter FG, editors. *The genus citrus*. Duxford: Woodhead Publishing, an imprint of Elsevier; 2020. p. 521.
48. Wu GA, Terol J, Ibanez V, López-García A, Pérez-Román E, Borredá C, et al. Genomics of the origin and evolution of *Citrus*. *Nature.* 2018;554(7692):311–6. <https://doi.org/10.1038/nature25447>.
49. Droc G, Giraud D, Belsler C, Labadie K, Duprat S, Cruaud C, et al. A super-genome for cultivated citrus reveals evolutionary features during the allopatric phase of their reticulate evolution. 2024. Available from: <https://bioRxiv.org/cgi/content/short/2024.10.17.618847v1>.
50. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol.* 2005;1(2):166–75. <https://doi.org/10.1371/journal.pcbi.0010022>.
51. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in *de novo* annotation approaches. *PLoS ONE.* 2011;6(1):e16526. <https://doi.org/10.1371/journal.pone.0016526>.
52. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V et al. PASTEC: an automatic transposable element classification tool. *PLoS ONE.* 2014;9(5):e91929. <https://doi.org/10.1371/journal.pone.0091929>.
53. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet.* 2008;9(5):411–2. <https://doi.org/10.1038/nrg2165-c1>.
54. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 2015;6(1):11. <https://doi.org/10.1186/s13100-015-0041-9>.
55. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res.* 2014;42:222–30. <https://doi.org/10.1093/nar/gkt1223>.
56. Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 2011;39(Database):70–4. <https://doi.org/10.1093/nar/gkq1061>.
57. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22(13):1658–9. <https://doi.org/10.1093/bioinformatics/btl158>.
58. Ollitrault P, Curk F, Krueger RR. *Citrus taxonomy*. In: Gentile A, La Malfa S, Deng Z, editors. *The genus citrus*. Elsevier (United Kingdom); 2020. p. 57–81. <https://doi.org/10.1016/b978-0-12-812163-4.00004-8>.
59. Tanaka T. *Citologia: semi-centennial commemoration papers on Citrus studies*. Osaka Citol Support Found; 1961.
60. Wu B, Yu Q, Deng Z, Duan Y, Luo F, Gmitter F Jr. A chromosome-level phased genome enabling allele-level studies in sweet orange: a case study on citrus Huanglongbing tolerance. *Hortic Res.* 2022;10(1):uhac247. <https://doi.org/10.1093/hr/uhac247>.
61. Di Guardo M, Moretto M, Moser M, Catalano C, Troglio M, Deng Z, et al. The haplotype-resolved reference genome of lemon (*Citrus limon* L. Burm f). *Tree Genet Genomes.* 2021;17(6):46. <https://doi.org/10.1007/s11295-021-01528-5>.
62. Huang Y, Xu Y, Jiang X, Yu H, Jia H, Tan C, et al. Genome of a citrus rootstock and global DNA demethylation caused by heterografting. *Hortic Res.* 2021;8(1):69. <https://doi.org/10.1038/s41438-021-00505-2>.
63. Shimizu T, Tanizawa Y, Mochizuki T, Nagasaki H, Yoshioka T, Toyoda A, et al. Draft sequencing of the heterozygous diploid genome of satsuma (*Citrus unshiu* Marc.) using a hybrid assembly approach. *Front Genet.* 2017;8:180. <https://doi.org/10.3389/fgene.2017.00180>.
64. Lu Z, Huang Y, Mao S, Wu F, Liu Y, Mao X, et al. The high-quality genome of pummelo provides insights into the tissue-specific regulation of citric acid and anthocyanin during domestication. *Hortic Res.* 2022;9:uhac175. <https://doi.org/10.1093/hr/uhac175>.
65. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellings AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 2019;20(1):275. <https://doi.org/10.1186/s13059-019-1905-y>.
66. Neumann P, Novák P, Hošťáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob DNA.* 2019;10(1):1. <https://doi.org/10.1186/s13100-018-0144-1>.
67. McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics.* 2003;19(3):362–7. <https://doi.org/10.1093/bioinformatics/btf878>.
68. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 2010;38(22):e199–199. <https://doi.org/10.1093/nar/gkq862>.
69. Ge R, Mai G, Zhang R, Wu X, Wu Q, Zhou F. MUSTV2: an improved *de novo* detection program for recently active miniature inverted repeat transposable elements (MITES). *J Integr Bioinforma.* 2017;14(3):20170029. <https://doi.org/10.1515/jib-2017-0029>.
70. Wenke T, Döbel T, Sörensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell.* 2011;23(9):3117–28. <https://doi.org/10.1105/tpc.111.088682>.
71. Vassilief H, Haddad S, Jamilloux V, Choisine N, Sharma V, Giraud D, et al. CAU-LIFINDER: a pipeline for the automated detection and annotation of caulimovirid endogenous viral elements in plant genomes. *Mob DNA.* 2022;13(1):31. <https://doi.org/10.1186/s13100-022-00288-w>.
72. Novák P, Ávila Robledillo L, Kobližková A, Vrbová I, Neumann P, Macas J. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* 2017;45(12):e111–111. <https://doi.org/10.1093/nar/gkx257>.
73. Camacho C, Coulouris S, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10(1):421. <https://doi.org/10.1186/1471-2105-10-421>.
74. Giraud D. Library of repeated sequences annotated in Citrus genome assemblies. Recherche Data Gouv; 2025. Available from: <https://entrepot.recherche.data.gouv.fr/citation?persistentId=doi:10.57745/ZVYEBU>. [cited 2025 Apr 4].
75. Liu Y, Tahir ul Qamar M, Feng JW, Ding Y, Wang S, Wu G, et al. Comparative analysis of miniature inverted-repeat transposable elements (MITES) and long terminal repeat (LTR) retrotransposons in six *Citrus* species. *BMC Plant Biol.* 2019;19(1):140. <https://doi.org/10.1186/s12870-019-1757-3>.
76. Wu Y, Wang F, Lyu K, Liu R. Comparative analysis of transposable elements in the genomes of *Citrus* and *Citrus*-related genera. *Plants.* 2024;13:17. <https://doi.org/10.3390/plants13172462>.
77. Swingle WT, Reece PC. *The botany of Citrus and its wild relatives*. In: Reuther W, Webber HJ, Batchelor DL, editors. *The Citrus industry*. Berkeley: University of California press; 1967. p. 190–430.
78. Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, et al. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.* 2010;186(1):37–45. <https://doi.org/10.1111/j.1469-8137.2009.03096.x>.
79. Madlung A, Wendel JF. Genetic and epigenetic aspects of polyploid evolution in plants. *Cytogenet Genome Res.* 2013;140(2–4):270–85. <https://doi.org/10.1159/000351430>.
80. Song Q, Chen ZJ. Epigenetic and developmental regulation in plant polyploids. *Curr Opin Plant Biol.* 2015;24:101–9. <https://doi.org/10.1016/j.pbi.2015.02.007>.
81. Vicent CM, Casacuberta JM. Impact of transposable elements on polyploid plant genomes. *Ann Bot.* 2017;120(2):195–207. <https://doi.org/10.1093/aob/mcx078>.
82. Luro F, Bloquel E, Tomu B, Costantino G, Tur I, Riolacci S, et al. The INRA-CIRAD citrus germplasm collection of San Giuliano, Corsica. In: Zech-Matterne V, Fiorentino G, editors. *AGRUMED: archaeology and history of citrus fruit in the Mediterranean*. Naples: Publications du Centre Jean Bérard; 2017. p. 243–61. Available from: <https://doi.org/10.4000/books.pcbj.2232>.

Publisher's Note

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