Genome relationships and LTR-retrotransposon diversity in three cultivated Capsicum L. (Solanaceae) species

Rafael de Assis
Universidade Estadual de Londrina
Viviane Yumi Baba
Universidade Estadual de Londrina
Leonardo Adabo Cintra
Universidade Estadual de Londrina
Leandro Simões Azeredo Gonçalves
Universidade Estadual de Londrina
Rosana Rodrigues
Universidade Estadual do Norte Fluminense Darcy Ribeiro
Andre Luís Laforga Vanzela (andrevanzela@uel.br)
https://orcid.org/0000-0002-2442-2211

Research article

Keywords: chili peppers, FISH, LTR retrotransposons, plant genome, transposable elements

Posted Date: February 26th, 2020

DOI: https://doi.org/10.21203/rs.2.24581/v1

Version of Record: A version of this preprint was published on March 17th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6618-9.

Abstract

Background: Plant genomes are rich in repetitive sequences, and transposable elements (TEs) are the most accumulated of them. This mobile fraction can be distinguished as Class I (retrotransposons) and Class II (transposons). Retrotransposons that are transposed using an intermediate RNA and that accumulate in a “copy-and-paste” manner have been screened in three completely sequenced genomes of peppers (Solanaceae family). The goal of this study was to understand the genome relationships among Capsicum annuum , C. chinense and C. baccatum , based on a comparative analysis of the function, diversity and the chromosome distribution of TE lineages in the Capsicum karyotypes. Due to the great commercial importance of pepper in natura , as a spice or as an ornamental plant, these genomes have been widely sequenced, and all of the assemblies are available in the SolGenomics group. These sequences have been used to compare all repetitive fractions from a cytogenomic point of view. Results: The qualification and quantification of LTR-retrotransposons (LTR-RT) families are contrasted with molecular cytogenetics data, and the results show a strong genome similarity between C. annuum and C. chinense as compared to C. baccatum . The Gypsy superfamily is more abundant than Copia , especially for Tekay/Del family members, including a high representation in C. annuum and C. chinense . On the other hand, C. baccatum accumulates more Athila/Tat sequences. The FISH results show retrotransposons differentially scattered along chromosomes, with the exception of CRM family sequences, which mainly have a proximal accumulation associated with heterochromatin bands. Conclusions: The results confirm a close genomic relationship between C. annuum and C. chinense in comparison to C. baccatum . Centromeric GC-rich bands appeared to be associated with the accumulation regions of CRM elements, whereas terminal and subterminal AT- and GC-rich bands do not correspond to the accumulation of the retrotransposons in the three Capsicum species tested here.

Background

Plant genomes are composed of repetitive and non-repetitive portions, organized in families according to the nature of the sequences, mobility throughout genomes and localization in the chromosomes [1, 2]. Transposable elements are virus-like sequences, and they are grouped into two classes, namely Class I or retrotransposon and Class II or transposon-like. The retrotransposons are the most common elements in plant genomes [1], and they use polygenic chain enzymes, such as reverse transcriptase, for retrotransposition via an intermediate RNA molecule [3-5]. Transposons, on the other hand, use different enzymes, such as transposase (transposons), helicase/replicase (helitrons), polymerase B (polintons) and tyrosine replicase for cryptons [2, 4], to reposition themselves along the genomes.

The LTR-retrotransposons are classified into *Copia*, *Gypsy,* Bel-Pao, retrovirus, and endogenous retrovirus superfamilies, and they may be differentiated by the protein domain organization on the polygenic sequence [2, 5, 6]. The superfamilies *Copia* and *Gypsy* differ from each other in terms of the location of integrase in the polygenic chain: when the integrase is positioned upstream of the reverse transcriptase the element is recognized as a *Copia* retrotransposon, whereas a downstream position identifies the *Gypsy* members [5, 7]. Both superfamilies are subdivided into many lineages [4, 5], with the *Gypsy* taxon encompassing, for example, clades named of Athila/Tat and chromoviruses with CRM and Del lineages [5, 7]. Similarly, *Copia* retrotransposons are grouped into other lineages, such as the Ivana/Oryco, Tork and others [5, 7]. Because LTR-RTs present independent activity in the chromosomes and different fates in the genomes, closely related species may exhibit variability in their occurrence, amount and chromosome distribution, which influence “fluctuations” in DNA C-values (amount DNA in a haploid nucleus, in picograms) [8-10].

The presence of TEs within genomes may influence gene expression and function. Depending on their insertion region, TEs may change the transcript splicing/processing and coding regions (see [11]). In plants, the LTR-RT lineages may be clustered in different chromosome regions, regardless of the superfamily to which they belong. This may be instanced in the accumulation of *Copia* elements in the proximal chromosomal regions in many plants [12, 13]. Depending on the *Gypsy* lineage, their positioning along the chromosomes can be much more diverse. Whereas CRM elements often localize at centromeric regions [14, 37], Athila/Tat and Del elements were found in heterochromatic and euchromatic regions, often scattered across the chromosomes [14, 15]. However, when the retrotransposons are localized using FISH probes made for superfamilies and not specifically for each *Copia* or *Gypsy* lineage, the FISH signals may appear with a scattered profile, as observed in the chromosomes of *Copaifera* [16]. Studies in samples of Solanaceae have shown that *Gypsy* retrotransposons may be associated with heterochromatin at pericentromeric regions of *Solanum* chromosomes [17]. The accumulation of *Gypsy*-Del elements has been reported in both heterochromatic and euchromatic regions of *Capsicum annuum* [18].

Capsicum species are an excellent model to investigate the dynamics and distribution of LTR-RTs, because of their relatively large genomes, grouped within 2*n* = 24 and 26 chromosomes, and DNA C-values range from 3.16 to 5.77 pg, in which the repetitive DNA families may represent more than 70% of the genomes [19]. *Capsicum* species present a great diversity of repetitive DNA families, especially regarding the number and distribution of ribosomal sequences and heterochromatin sites [20-26]. In this context, the present study aims to know and compare the occurrence and distribution of LTR-RTs on the chromosomes, focusing on *Gypsy* superfamily that has been predominant in the *Capsicum* genomes, as well as their association with the diversity of other repetitive sequences. Peppers are among the most important vegetables in the world due to their high versatility and wide range of applications in cooking, industry, and decoration [27]. Therefore, large investments have been made to obtain high throughout sequences of *C. annuum*, *C. chinense* and *C. baccatum* [19, 28]. This large data volume has increased the possibility of studying and comparing the genomic organization of these species from the cyto-molecular point of view.

Given the gaps in knowledge regarding the repetitive fractions of *Capsicum* species, some questions about the dynamics and distribution of LTR-RTs remain unanswered, such as: Are the heterochromatin rich regions collocated with TE rich regions? Do closely related genomes share the same LTR-RT families concerning quantity and chromosome localization? Our discussion focused on the characterization of retrotransposons based on a broad cytogenomic comparison using the repetitive fraction available in the large *C. annuum*, *C. chinense,* and *C. baccatum* datasets.

Results

Comparative analyses based on the conserved domains of transposable elements

The high coverage sequencing scaffolds of *Capsicum annuum* (3.07Gb)*, C. chinense* (3.22Gb) and *C. baccatum* (3.01Gb) from the *Pepper Genome Platform* were used for the analysis. The search based on conserved coding domains of polygenic chain (POL) of retrotransposons showed that fractions related with conserved protein domain of reverse transcriptase, integrase and RNase H represents 2.75%, 17.14% and 2.47% in the *C. annuum, C. chinense* and *C. baccatum* datasets, respectively (Tables 2 and S1). When Class I and II elements were compared, conserved sequences of Class I elements were more abundant in the three datasets. The Class I was more accumulated in *C. annuum* and *C. chinense* (>90%) and less accumulated in *C. baccatum* (~70%). The conserved sequences similar to Class II elements were less represented with <10% in these datasets (Tables 2 and S1).

After organizing the Class I sequences as LTR, non-LTR and endogenous retrovirus, the percentages of LTR elements showed more similarity between *C. annuum* and *C. chinense* (89.32 and 98.78% respectively) than in *C. baccatum* with 70.55%. Both non-LTR and ERVs sequences were less accumulated, and they were not equally distributed among species: 2.21% and 0.15% in *C. annuum*, 0.06% and 0.01% in *C. chinense,* and 1.51% and 0.23% in *C. baccatum* (Table 2). Sequences recognized as *Copia* superfamily members were low representativeness in all three datasets (<10%), while *Gypsy* members were the most accumulated. *C. annuum* and *C. chinense* exhibited 85.77% and 98.43% of the *Gypsy*-related sequences respectively, while in *C. baccatum*, they were 63.27% (Table S1). In the *Gypsy* superfamily, the Tekay/Del elements were predominant in *C. annuum* and *C. chinense*, with 67.68 and 95.43% respectively, while in *C. baccatum* the Tekay/Del elements presented only 15.7%. The Athila/Tat clade was more representative in *C. baccatum* (43.25%) than in *C. annuum* and *C. chinense* at 16.36 and 2.72%, respectively (Tables 2 and S1). Other *Gypsy* lineages, such as CRM and Galadriel, had lower representation (see Tables 2, S1 and Figure 1C).

The other elements belonging to the non-LTR groups, transposons, for example, were less accumulated in these three datasets, such as in LINE and SINE (<5%), ERVs (<1%), CACTA (<0.1%), hAT (<2%), MuDR (<0.1%) and Helitron (<0.1%). Nevertheless, Sola elements presented an interesting contrast, representing 5.75% and 7.31% in the *C. annuum* and *C. baccatum* datasets respectively, and only 0.94% in *C. chinense* dataset. Ribosomal DNA was also estimated, once literature shows a large variation in the number of sites between these species. Although 5S rDNA showed no great variation among the three species, the 35S rDNA sequences exhibited a contrasting accumulation, with 18.68% in *C. baccatum,* 1.29% in *C. annuum* and 0.12% in *C. chinense* (Tables 2 and S1).

Gypsy autonomous elements dominated the datasets

The search for putative autonomous retrotransposons was focused on *Gypsy* superfamily members (Tekay/Del, CRM, and Athila/Tat lineages), once their sequences were the most accumulated in the three datasets.

The characterization of retrotransposons was first based on LTR_STRUC [29] output file of *Capsicum annuum* (considered here as “reference”), which resulted in 254 sequences. From these, only four sequences from CRM, three from Tekay/Del and two from

Athila/Tat lineages were characterized. Other 267 sequences that have been identified using the BLAST tool came from *C. chinense* and *C. baccatum* datasets, and that were contrasted against *C. annuum* dataset. The BLAST identified four sequences from CRM, and seven from Tekay/Del lineages as putative autonomous elements.

This characterization of these sequences was based on LTRs' presence (Dotter l/b [30]), annotation of GAG and polygenic chain (Artemis [31]) and minimal size of elements (>3.500 base pairs long). The pseudomolecules that were identified as non-autonomous retrotransposons were those that they did not exhibit one or more LTRs, lost genes from the polygenic chain or ORFs, or had large inverted stretches. After two rounds of alignments (ClustalW [32] followed by Mauve [33]) with putative autonomous sequences, they were organized in groups within each lineage. The Tekay/Del sequences were clustered in five groups, with sequences varying from 8,105 to 8,902 bp length. The sequences of group 2 were shared between *C. annuum* and *C. chinense*, while those of groups 1 and 5 were exclusive for *C. annuum*, and those of groups 3 and 4 were exclusive for *C. baccatum* (Figure S1). The CRM sequences were clustered in three groups, varying from 5,259 to 7,328 bp in length. Except for groups 1 and 2 that appeared in *C. annuum* and *C. baccatum,* the others were shared among the three species (Figure S2). The two sequences of the Athila/Tat lineage were distinct from each other, making the two groups exclusive of *C. annuum* (Figure S3).

The bootstrapped maximum likelihood phylogenetic tree was performed with complete sequences and organized them in three clades: CRM, Athila, and Del (Figure 2). The CRM sequences were clustered into four groups: A) CRM_2, CRM_4 and CRM_7; B) CRM_5; C) CRM_3 and CRM_6; and D) CRM_1. The Athila sequences presented two groups within *C. annuum*. The Del elements were more diverse, being organized into three well-supported groups: A) Del_3; B) Del_1 and Del_4; and C) Del_7 and Del_9. The remaining Del sequences did not form well-supported groups (Figure 2). One sequence was selected from each cluster for a graphical representation (Figure 3B).

The clusters of putative autonomous elements were evaluated by their accumulation and distribution in each dataset, in order to compare the probable dynamic of these retrotransposons in the genomes differentiation. In panel 3A, the color scale represents how accumulated these elements are in each dataset (blue represents less accumulation and yellow, more accumulation). In general, *C. annuum* and *C. chinense* shared similar sequences in relation to *C. baccatum*, following the phylogeny proposal for the genus. An exception was observed for Del I (Del_1) group, which on the heatmap (Figure 3A) exhibited greater accumulation in the *C. baccatum* genome. The same was observed for the group Del V (Del_8 and Del_10), which despite being composed of sequences from *C. baccatum,* was more accumulated in *C. annuum*. When the reverse transcriptase regions of these putative elements were analyzed, this tendency, although not the same as in the complete elements, was retained, with some exceptions such as the accumulation of the group Del II, which was composed only of sequences from *C. baccatum* (Del_5, Del_6, Del_7, and Del_9), exhibiting more accumulation in *C. chinense*.

Comparative cytogenetics

The sequences of the reverse transcriptase of the most representative LTR-RT lineages were aligned, and four consensus sequences were used for the primer design: three for the *Gypsy* superfamily (CRM, Tekay/Del, and Athila) and one for the *Copia* superfamily (Ivana/Oryco). The data is presented in Tables 2 and S1. Fluorescence *in situ* hybridization assays revealed either scattered or clustered signals, depending on the LTR-RT lineage analyzed. The probe for Ivana/Oryco lineage (*Copia*) showed a hybridization profile with a few signals scattered along chromosomes, with clear differences between the pairs and, sometimes, with small interstitial and proximal dots (see arrowheads in Figure 4A-B and Figure S6A-F).

The Tekay/Del probe showed scattered signals along the chromosomes in three species (Supplementary figure 4). Although the signals observed in *C. annuum* (Figure 4C) and *C. chinense* (Supplementary figure 4G) were more evident in the interstitial and proximal regions of all the chromosomes, in *C. baccatum*, the signals accumulated in half of the chromosomes and were very weak in the others (Figure 4D). These results confirm the observations from bioinformatic analysis, which shows a greater accumulation of Tekay/Del sequences in *C. annuum* and *C. chinense* than in *C. baccatum*.

The Athila/Tat probe also exhibited scattered signals along the chromosomes, being much more accumulated in the interstitial regions. In *C. annuum*, except for a pair with less intense signals, the remainder exhibited brighter signals at the proximal to interstitial (close to proximal) regions, without any signals in terminal ones (Figure 4E), similar to those found in *C. chinense* (Supplementary figures 6G-I). *Capsicum baccatum* exhibited four chromosomes with less intense signals and brighter FISH signals in the remaining chromosomes. However, four chromosomes showed stronger signals than those observed in *C. annuum* and *C. chinense* (Figure 4H). The CRM probes showed FISH signals accumulated in the proximal regions of all chromosomes in the three species (see Figures 4F-G and

Supplementary figure 5). Nevertheless, there was a clear difference in signal intensity among chromosome pairs, with a minor signal in a pair of *C. baccatum* and another in *C. chinense* (see arrowheads).

The C-CMA/DAPI banding was performed to verify that LTR-RTs' accumulation areas corresponded to the AT- and GC-rich band regions as well as to check if the diversity in the distribution profiles of repetitive sequences is equivalent in these species. *Capsicum annuum* showed two pairs without bands, three pairs with terminal C-DAPI$^+$ dots co-located with more intense C-CMA$^+$ bands and terminal C-CMA$^+$ bands in 11 chromosomes, including three pairs with stronger terminal C-CMA$^+$ bands (Figures 5A-B). *Capsicum chinense* exhibits a larger number of intense C-CMA$^+$ bands (seven pairs), and of these, two pairs exhibited intense adjacent C-DAPI$^+$ bands. Smaller terminal C-CMA$^+$ bands were observed in 11 chromosomes, and proximal bands were observed as centromeric dots in nine pairs. Some of these bands were evidenced with DAPI and CMA_3 staining. Thinner interstitial bands were stained with C-DAPI but not with C-CMA (Figures 5C-D). *Capsicum baccatum* showed six pairs with terminal dots and six with centromeric and/or interstitial C-DAPI bands. In three pairs, these bands appeared as CMA$^+$/DAPI$^+$. The strongest terminal C-CMA bands were detected in four pairs in addition to thinner terminal bands in at least one arm of all chromosomes. Proximal centromeric dot bands were observed in 11 pairs, of which only three also appeared as DAPI$^+$ (Figure 5E-F).

Discussion

Differential accumulation of repetitive DNA families on Capsicum genomes

TEs can move through genomes, representing an evolutionary force that modifies genome structure via mechanisms, such as illegitimate recombination, gene capture, shuffling of regulatory motifs and the generation of new functionality or silencing (see [34]). In the last instance, TEs may cause a change in the genomes' global structure and fluctuations in genome size [10]. TEs have been useful to compare genomes and karyotypes in evolutionary studies as well as other applicable approaches, such as the study in grapes and blood orange that showed the origin of alterations in the expression of some genes after the insertion of TEs next to them [35].

The 'Mobilome' occupies the largest portion of plants' genomes and play an important role in the physical and functional aspects of chromosomal structures, such as those of the CR lineage (centromeric retrotransposons) that are associated with chromosomal kinetics [36, 37]. In some monocotyledons, for instance, the Mobilome represents around 75% of the genome, such as 80% in maize [6], and LTR-RTs are the most dynamic elements found within the genome [7, 38]. In *Nicotiana attenuata* and *N. obtusifolia*, for example, LTR retrotransposons reach up to 81% and 64% of their respective genomes [39]. The study of LTR-RTs in a chromosomal landscape may assist the understanding of some extent of the regulatory potential of TEs along chromosomes, and may also hold the prospect of its possible application in crop breeding programs, such as peppers. Maize is a good example because several TE families described near the genes have been identified as enhancers or repressors under stressful conditions [40, 41].

Previous studies addressing TEs in the Solanaceae family have shown that the *Gypsy* elements of *Solanum lycopersicum* are more abundant than the *Copia* superfamily members [42], although an approach using only autonomous elements showed a predominance of the Tekay/Del (*Gypsy*) and Tork (*Copia*) lineages in this species [43]. This suggests that estimates may vary when considering only autonomous or include sequences of non-autonomous elements. The present results in *Capsicum* indicate that only 0.002% of the *C. annuum* dataset corresponds to putative autonomous elements, followed by 0.001% in *C. chinense* and 0.004% in *C. baccatum*. The remaining sequences correspond to non-autonomous elements. It is important to highlight that the three genomic datasets were obtained by high-covered sequencing [19, 28], which may support the assembly of the pseudochromosomes as complete elements. Even though the *Capsicum* sequencing does not cover the entire genome (the sequencing part comprises 87% of the *C. annuum*, 94% of *C. chinense*, and 83% of *C. baccatum* genomes) [19, 44] it can be stated that TEs occupy an important fraction of *Capsicum* genomes. According to Lisch [35], the major part of coding repetitive fractions relates to fragments of non-autonomous elements, which may be amplified by the activity of the autonomous. This might explain the high percentage of LTR-RT fragments in these three datasets.

According to Qin et al. [28], the *Gypsy* members were the most abundant LTR retrotransposons in *Capsicum*, with the highest insertion activity among Solanaceae species. When comparing *Gypsy* and *Copia* lineages in this plant group, the percentage of LTR-RTs for *C. annuum, C. chinense,* and *C. baccatum* was 89%, 98%, and 70%, respectively. This result pointing out a predominance of *Gypsy* (>70%) over *Copia* superfamily (<10%). A contrasting accumulation of Tekay/Del, Athila/Tat and CRM lineages of *Gypsy* were noted in these three genomes. These data point toward the importance of LTR-RTs' fate in the process of genome organization and differential accumulation between related species, even after considering that *C. annuum* and *C. chinense* (Annuum clade) are closer when compared to *C. baccatum* of the Baccatum clade (see supplementary figure S7) [45]. Kim et al. [44] reported that *Solanum* is a closely

related genus of *Capsicum* and shares a common ancestor 19.6 MYA, and more recently, the authors demonstrated *C. annuum* and *C. chinense* share a common ancestor 1.14 MYA, while these species shared 1.74 MYA with *C. baccatum*. These three species also differ in the accumulation of 35S rDNA, with about 20% more sequences in *C. baccatum* than in *C. annuum* and *C. chinense*, besides the number of rDNA sites in the chromosomes [24]. These genomic differences may be responsible for certain difficulties in performing interspecific crosses between these species of distinct clades because of the pre- and post-zygotic barriers, as reported by Manzur et al. [46] and Cremona et al. [47].

The differential activity of retrotransposons among close genomes was also reported in *Helianthus* [48] and *Solanum* [49, 50], and these results corroborate with those obtained in the present study regarding *Capsicum*. The present data have also shown differences among *C. annuum* and *C. chinense*, especially in Tekay/Del, ERVs, and Line-RTE accumulation, suggesting that other elements, besides LTR-RT ones, evolve independently. The differential accumulation of Tnt1 retrotransposons in *Nicotiana* may be a good example to understand and to support the idea of an independent fate of TEs on genome differentiation [51, 52].

Recovered Del, CRM, and Athila/Tat autonomous elements support the Gypsy LTR-RTs' predominance

The ability of retrotransposons to activate and invade plant genomes may be associated with some internal and external factors, such as biotic and abiotic stresses, breeding processes, injuries, climatic changes, polyploidization, hybridization, and other events (see [34, 53]). However, the activation and proliferation of TEs may be influenced by the ability to cheat cellular silencing controls [53]; and the autonomous elements containing a complete polygenic chain, regulators and both LTRs can do that [54]. The absence of some regions may make these elements non-autonomous. In the present study, more non-autonomous sequences (8-folds) were found in the three *Capsicum* datasets than potentially autonomous ones. This result suggests that these repetitive element classes may have undergone different events of degeneration along with genomes differentiation.

The putative autonomous elements recovered in the present analysis, i.e., ten sequences of Tekay/Del, two sequences of Athila/Tat of *C. annuum* and seven of CRM, varied between the datasets. This idea follows the proposal of the independent fate of TEs among genomes [55], and it can be exemplified by the occurrence of some Tekay/Del elements exclusive in *C. baccatum* compared to three others found in *C. annuum* and *C. chinense*. Also, we can mention the thirty-fold difference in CRM amount in *C. baccatum* in relation to *C. annuum* and *C. chinense*. This result is in accordance with Hawkins et al. [56] report, which suggests that in *Gossypium* species, different lineages of LTR-RTs evolved at different moments along with genome evolutionary history, generating a threefold difference in DNA content among diploid species. In another example, De Castro Nunes et al. [37] observed also a greater accumulation of CRM copies in the diploid *Coffea* species in comparison to the hybrid tetraploid *C. arabica*.

Not all LTR-RT rich regions in Capsicum chromosomes are heterochromatin hotspots

It is well established in the literature that TEs, especially LTR-RT superfamilies, occupy "specific" chromosomal regions, with the consensus that *Copia* elements are distributed preferentially along the chromosomes associated with euchromatin, while *Gypsy* elements are resident in heterochromatin-rich regions (see [7]). In *Coffea*, *Brachiaria,* and *Secale*, for example, *Gypsy* probes were located in proximal heterochromatin-rich chromosome regions [16, 57, 58], but in *Gossypium* species, *Gypsy* probes are hybridized along chromosomes [59]. However, when the elements are considered according to their phylogenetic positions, i.e., lineages of *Copia* and *Gypsy* [5, 7, 60], it becomes evident that there are many differences in the TE distribution profiles, in both plants (see [10, 61) and animals [62, 63]. Thus, it seems wiser to believe that each element has its characteristics, including chromosomal positioning, genome impact, epigenetic influence, diversification rate, and other features.

Previous studies using FISH in *Capsicum* spp. have been restricted to rDNA probes [24], which demonstrated that *C. baccatum* accumulates more in terminal 35S rDNA sites compared to *C. annuum* and *C. chinense*, which exhibited just two to four pairs. Moscone et al. [20, 64], Scaldaferro et al. [23] and Martins et al. [26] reports have shown wide variability in the presence of terminal, interstitial and proximal heterochromatic bands in *Capsicum* species, such as the large and minor heterochromatic terminal bands in *C. annuum*, *C. chinense,* and *C. baccatum* observed in the present study. FISH results using different LTR-RT probes showed hybridization signals accumulated from proximal (CRM) to interstitial region (Athila/Tat and Tekay/Del), scattered, or minor dots along chromosomes (Tekay/Del, Oryco and Tork), such as in *Brachiaria* [14]. However, no preferential accumulation or strong signals were found in terminal chromosome regions, suggesting that the LTR-RT families analyzed have no accumulation at regions containing rDNA or terminal heterochromatin in *Capsicum* chromosomes.

FISH using the Athila/Tat probe strongly hybridized at proximal to interstitial regions in almost all the chromosomes of *Capsicum*. There was also no evident co-location with heterochromatic regions, although there were small AT- and GC-rich bands in few chromosomes, i.e., without evident correlation with heterochromatin hotspots. In this case, as well, the scattered signals (or dots) observed after FISH with Tekay/Del and Oryco probes are in agreement with the concept of dispersed localization of retroelements within plant genomes, but without dependence on co-localization with heterochromatin blocks. This Athila/Tat dispersion pattern, such as interstitial dots, has been described by Park et al. [65] in *C. annuum*. Using the *Passiflora edulis* for comparison, members of *Ty3*/*Gypsy* superfamily were the most accumulated, and their sequences appeared scattered along chromosomes, including at the pericentromeric regions [66]. In some Solanaceae species, such as tomato and peppers, elements of the Tekay/Del *Gypsy* superfamily had a scattered accumulation profile as reported by Park et al. [65], with hybridization in the chromosomes of *Solanum lycopersicum* (tomato) and *Capsicum annuum* (pepper), in which pepper had a higher number of and more intense signals than those observed in tomato.

Different from the other LTR-RT probes, the CRM probe exhibited intense signals in the proximal regions, associated with centromeres, and in *Capsicum* chromosomes, these regions were rich in CMA^+ and $DAPI^+$ signals. One notable exception is the centromeric retrotransposon lineage of chromovirus (also called centromeric retrotransposon of maize or CRM), which occurs preferentially in proximal chromosome regions. CRMs carry particular domains called chromodomain (CHRomatin Organization MOdifer DOMAIN) and CR motifs that have the potential to interact with the CENH3 centromeric protein and to participate in the centromere function [36, 67]. FISH centromere signals using CRM lineage probes have been described in several plant species, for example in some monocotyledon groups [58, 68, 69], suggesting that besides the association with specific centromeric proteins, this accumulation may also be associated with recombination-poor regions.

Conclusions

This comparative cytogenomic analysis using the three most economically important *Capsicum* species showed great diversity in genome composition, although there is a closer approximation between *C. annuum* and *C. chinense* as compared to *C. baccatum* as it is suggested for the Annuum and Baccatum complex phylogeny. The dataset screening of these three species showed that there was a differential accumulation of transposable elements, especially those from the lineages Tekay/Del, CRM and Athila/Tat from *Gypsy*, while those of the *Copia* superfamily were underrepresented. From a chromosomal point of view, these transposition elements were dispersed along the chromosomes (*Copia*) as well as in blocks (*Gypsy*), highlighting those of the CRM lineage that predominated the centromeric region. Another aspect is that LTR-RTs are not always associated with heterochromatin-rich regions. These data support the idea of the independent fate of LTR-RTs. Such genomic and chromosomal differences between closely related species should be taken into account in breeding programs, as they may interfere with the success of interspecific crosses and the introgression of agronomic traits of interest. *Capsicum* spp. proved to be a good model for most studies on the repetitive fraction, from both genomic and chromosomal points of view, considering the diversity in the accumulation and genomic distribution of LTR-RTs.

Methods

Plant materials

Seeds of *Capsicum annuum* cv. Criollo de Morelos (accession GBUEL145), *C. chinense* (accession GBUEL27) and *C. baccatum* (accession GBUEL118), identified by Dr. Leandro S. A. Gonçalves, were obtained from the gene bank of Londrina State University. The samples were sowed in 128-cell polystyrene trays containing the substrate Vivatto®. Ten seedlings of each species were grown in the Cytogenetics and Plant Diversity Laboratory greenhouse.

Genomic analysis

The following three genomes used for the bioinformatics analysis were obtained from Pepper Genome Platform (http://peppergenome.snu.ac.kr/): scaffolds from *Capsicum annuum* v.1.6, C. *chinense* v.1.2, and *C. baccatum* v.1.2. Files were used to search for autonomous and non-autonomous LTR-RTs. For genomic comparisons, a database was built to run a local BLAST, which contained all the TEs conserved protein sequences from REXdb [5], GypsyDB [61], RepBase [70] and NCBI (http://www.ncbi.nlm.nih.gov/), containing 283,676 protein sequences. To identify the rDNA sequences, a second database was compiled comprising nucleotides 35S and 5S rDNA sequences from different organisms, obtained from NCBI, containing 1,652 sequences.

The repetitive fraction (transposable elements, 35 and 5S rDNA) evaluation was conducted by the Blast version 2.2.28+, comparing the genomic dataset against the local databases. The parameters used were E-value 10e-4, max target seqs 1 and the remaining were set by program default. The results obtained are plotted in Table S1, which includes quantitative and qualitative estimates according to the phylogeny proposal by Neumann et al. [5] as well as 35S and 5S rDNA fractions. The aim was to recognize and to differentiate retrotransposons in different lineages, and produce a more refined in situ hybridization, and to support the idea of the independent fate of these elements. The sequences were grouped into the superfamilies *Copia* (Ale, Alesia, Angela, Bianca, Bryco, Lyco, Gymco, Ikeros, Ivana, Osser, SIRE, TAR and Tork lineages) and *Gypsy* (CRM, Chlamyvir, Galadriel, Tcn1, Reina, Tekay/Del, Athila, Tat, Ogre, Retand, Phygy, and Selgy lineages). Due to their similarity, Athila, Tat, Ogre, and Retand members have been referred as Athila/Tat clade.

Autonomous and non-autonomous estimates and primer design

LTR_STRUC [29] has been used to compare and search for LTR-retrotransposons in the reference genome of *Capsicum annuum*. Putative retrotransposons sequences were then classified into *Gypsy* and *Copia* superfamilies according to their similarity measured against the Gypsy Database protein domains (http://www.gydb.org/index.php/Main_Page) by Genewise alignment [71] and annotated with Artemis [31]. The putative elements were then used as a database for BLAST rounds against the other datasets.

The BLAST output files were organized with sequences greater than 3,500 bp and more than 70% identity, and they were used to search for sequences with both LTRs with the Dotter l/b using program default parameters [30]. Subsequently, sequences that carried both LTRs were submitted to the online BLAST at NCBI to search for the presence of conserved domains. The sequences were also screened by the presence of stop codon by the Pfam online tool [72]. The complete nucleotide sequences with the correct protein order were aligned with the MAUVE [33] program and grouped as per the graphic similarity. These data were then validated by aligning them with CLUSTALW [32]. From each group, a representative sequence was chosen for the graphic annotation using the IBS [73] program. To understand the relationships among these sequences, the alignment was used to construct a bootstrapped maximum likelihood (ML) phylogenetic tree (1,000 bootstraps) in MegaX [74]. The ML tree was generated with GTR (general time-reversible) mutation model, gamma-distributed and invariant sites (G+I) rates among sites, and the heuristic method was nearest-neighbor-interchange (NNI). To understand how these putative autonomous elements are dispersed and accumulated along with the datasets, HeatMapper tool [75] was used in the three datasets through two approaches, one being with a representative sequence from each group of putative autonomous elements and the other with the reverse transcriptase region of these groups.

The most conserved stretch of the reverse transcriptase of each LTR-RT lineage was used for primer designing with the custom primers, OligoPerfect TM Designer tool of Thermo Fisher Scientific (http://tools.thermofisher.com, see Table 1). The primers' viability was assessed with the PCR primer stats tool (http://www.bioinformatics.org).

DNA extraction, PCR and LTR-RT probes

DNA was isolated from young leaves of each species using the cetyltrimethylammonium bromide (CTAB) method [76], purified with phenol:chloroform (1:1, v/v), chloroform:isoamyl alcohol (24:1, v/v) and RNase (1 mg mL^{-1}) and precipitated in 100% absolute ethanol. The samples were eluted in 10 mM Tris–HCl pH 8 and the concentrations were estimated using a NanoDrop 2000 Spectrophotometer (Thermo Scientific).

The LTR-RT probes were obtained by PCR using specific primers for each lineage, with *C. annuum* as a DNA template. A standard PCR [5 U μL^{-1} *Taq* polymerase (0.5 µL), 10× buffer (2.5 µL), 50 mM $MgCl_2$ (1.5 µL), 10 mM dNTP (1 µL), 5 mM primers (2 µL each) and H_2O up to a final volume of 25 µL] was used under the following conditions: 94 ºC for 2 min, 30 cycles of 94 ºC for 40 s, 59 ºC for 40 s and 72 ºC for 1 min, and a final extension of 72 ºC for 10 min. The reactions were tested via electrophoresis in an agarose gel at 3 V cm^{-1} and stained with ethidium bromide. The probes for each LTR-RT lineage were obtained through the re-amplification of PCR products that involved labeling with biotin-11-dUTP (*Gypsy* families) or Cy3-dUTP (*Copia* families).

Cytogenetic analysis

The root tips were pretreated with 0.5% colchicine (1h 30min) and fixed in ethanol-acetic acid (3:1, v:v). The fixed material was treated in a solution of 2% cellulase and 20% pectinase and squashed in a drop of 60% acetic acid. After liquid nitrogen freezing, the coverslips were removed, and the slides were air-dried.

For fluorescence *in situ* hybridization, slides received a mix containing a solution (30 µL) composed of 100% formamide (15 µL), 50% polyethylene glycol (6 µL), 20× SSC (3 µL), 100 ng of calf thymus DNA (1 µL), 10% SDS (1 µL) and 100 ng of probes (4 µL). The mix was denatured at 90 °C for 10 min, and hybridization was performed at 37 °C for 24 h in a humid chamber. Post-hybridization washes were carried out with 70% stringency, using an SSC buffer, with pH 7.0. After the probe detection with an avidin-fluorescein isothiocyanate (FITC) conjugate, washes were performed in 4× SSC/0.2% Tween-20 at room temperature. The slides were mounted with 25 µL of DABCO, a solution composed of glycerol (90%), 1,4-diaza-bicyclo (2.2.2)-octane (2.3%), 20 mM Tris–HCl, pH 8.0 (2%), 2.5 mM $MgCl_2$ (4%) and distilled water (1.7%) in addition to 1 µL of 2 µg mL^{-1} 4,6'-diamidino-2-phenylindole (DAPI).

For C-CMA/DAPI banding, the samples were incubated in a sequence of 45% acetic acid (8 min), saturated solution of $Ba(OH)_2$ at room temperature (8 min) and in 2× saline sodium citrate at 60 °C (1h and 30min). After this, the slides were stained with a CMA_3 for 90 min and DAPI for 30 min [77] and mounted in a solution of McIlvaine buffer plus glycerol (1:1, v:v) with 2.5 mM $MgCl_2$.

The chromosome images were acquired in greyscale with a Leica DM4500 B microscope coupled with a DFC300FX camera, pseudo-colored (blue for DAPI, greenish-yellow for FITC and red for Cy3) and contrasted using GIMP 2.8 Linux.

Abbreviations

Centromeric Retrotransposon of Maize: CRM

Endogenous Retrovirus: ERV

Fluorescence in situ hybridization: FISH

LTR-retrotransposons: LTR-RT

Long Terminal Repeats: LTRs

Long Interspersed Elements: LINEs

Picograms: pg

Ribosomal DNA: rDNA

Saline-Sodium Citrate: SSC

Short interspersed nuclear elements: SINEs

Sodium Dodecyl Sulfate: SDS

Transposable elements: TEs

Declarations

Ethics approval and consent to participate: not applicable

Consent for publication: not applicable

Availability of data and material: Original sequences of these three genomes can be accessed at Pepper Genome Platform (http://peppergenome.snu.ac.kr/)

Competing interests: The authors declare no competing interests

Funding: This study was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors thank the Brazilian agencies FINEP, Fundação Araucária, CNPq, CAPES, and ProPPG-UEL for their financial support.

Authors' contributions: RA performed all the experiments, wrote and corrected the manuscript. VYB and LAC performed bioinformatic analysis. LSA G and RR were responsible for varieties of supply, plant maintenance, analyses, and manuscript corrections. ALLV

designed the study, checked the data analyses, and organized, wrote and corrected the manuscript, as the head of the group. All authors have read and approved the manuscript.

Acknowledgments: ALLV, LSAG, and RR thank CNPq for the productivity scholarship (numbers 309902/2018-5, 307911/2018-7 and 307569/2017-9, respectively) and RR for the scholarship FAPERJ E-26/202.985/2017.

References

1 Heslop-Harrison JS, Schwarzacher T. Organisation of the plant genome in chromosomes. The Plant Journal. 2011, 66.1: 18-33.

2 Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annual review of plant biology. 2014, 65:505-530.

3 Kazazian HH. Mobile elements: drivers of genome evolution. Science. 2004, 303(5664):1626-1632.

4 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature reviews genetics. 2007, 8(12): 973.

5 Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA. 2019, 10(1): 1.

6 Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009, 326(5956):1112-1115.

7 Orozco-Arias S, Isaza G, Guyot R. Retrotransposons in plant genomes: Structure, identification, and classification through bioinformatics and machine learning. International journal of molecular sciences. 2019, 20(15):3837.

8 Bennetzen JL. Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica. 2002, 115(1):29-36.

9 Dodsworth S, Jang TS, Struebig M, Chase MW, Weiss-Schneeweiss H, Leitch AR. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). Plant systematics and evolution. 2017, 303(8):1013-1020.

10 De Souza TB, Chaluvadi SR, Johnen L, Marques A, González-Elizondo MS, Bennetzen JL, Vanzela ALL. Analysis of retrotransposon abundance, diversity and distribution in holocentric *Eleocharis* (Cyperaceae) genomes. Annals of botany. 2018, 122(2): 279-290.

11 Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2017. 1860(1), 157-165.

12 Heslop-Harrison JS, Brandes A, Schwarzacher T. Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. Chromosome research, 2003. 11, 241-253.

13 Underwood CJ, Henderson IR, Martienssen RA. Genetic and epigenetic variation of transposable elements in *Arabidopsis*. Current opinion in plant biology. 2017 Apr 1;36:135-41.

14 Santos FC, Guyot R, Do Valle CB, Chiari L, Techio VH, Heslop-Harrison P, Vanzela ALL. Chromosomal distribution and evolution of abundant retrotransposons in plants: gypsy elements in diploid and polyploid *Brachiaria* forage grasses. Chromosome research. 2015, 23(3):571-582.

15 Mlinarec J, Franjević D, Harapin J, Besendorfer V. The impact of the Tekay chromoviral elements on genome organisation and evolution of *Anemone* sl (Ranunculaceae). Plant Biology. 2016 Mar;18(2):332-47.

16 Gaeta ML, Yuyama PM, Sartori D, Fungaro MHP, Vanzela ALL. Occurrence and chromosome distribution of retroelementsand NUPT sequences in *Copaifera langsdorffii* Desf. (Caesalpinioideae). Chromosome research. 2010, 18:515–524.

17 Yang TJ, Lee S, Chang SB, Yu Y, de Jong JH, Wing RA. Indepth sequence analysis of the centromeric region of tomatochromosome 12: Identification of a large CAA block and char-acterization of centromeric retrotranposons. Chromosoma. 2005, 114:103-117.

18 Park M, Park J, Kim S, Kwon J-K, Park HM, Bae IH, Yang T-J, Lee Y-H, Kang B-C, Choi D. Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. The plant journal. 2012, 69:1018-1029.

19 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. Nature genetics. 2014, 46(3):270.

20 Moscone EA, Lambrou M, Hunziker AT, Ehrendorfer F. Giemsa C-banded karyotypes in *Capsicum* (Solanaceae). Plant systematics and evolution. 1993, 186(3-4):213-229.

21 Park YK, Kim BD, Kim BS, Armstrong KC, Kim NS. Karyotyping of the chromosomes and physical mapping of the 5S rRNA and 18S-26S rRNA gene families in five different species in *Capsicum*. Genes & genetic systems. 1999, *74*(4):149-157.

22 Park YK, Park KC, Park CH, Kim NS. Chromosomal localization and sequence variation of 5S rRNA gene in five *Capsicum* species. Molecules and cells. 2000, 10(1):18-24.

23 Scaldaferro MA, Grabiele M, Moscone EA. Heterochromatin type, amount and distribution in wild species of chili peppers (*Capsicum*, Solanaceae). Genetic resources and crop evolution. 2013, 60(2):693-709.

24 Scaldaferro MA, da Cruz MVR, Cecchini NM, Moscone EA. FISH and AgNor mapping of the 45S and 5S rRNA genes in wild and cultivated species of *Capsicum* (Solananceae). Genome. 2015, 59(2):95-113.

25 Aguilera PM, Debat HJ, Grabiele M. An integrated physical map of the cultivated hot chili pepper, *Capsicum baccatum* var. Pendulum. International Journal of Agriculture & Biology. 2017, doi: 10.17957/IJAB/15.0303

26 Martins LDV, Peron AP, Lopes ÂCDA, Gomes RLF., Carvalho RD, Feitoza LDL. Heterochromatin distribution and histone modification patterns of H4K5 acetylation and H3S10 phosphorylation in *Capsicum* L. Crop Breeding and applied biotechnology. 2018, 18(2):161-168.

27 Moreira AFP, Ruas PM., De Fátima Ruas, C, Baba VY, Giordani W, Arruda IM, et al. Genetic diversity, population structure and genetic parameters of fruit traits in *Capsicum chinense*. Scientia Horticulturae. 2018, 236, 1-9.

28 Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, et al. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proceedings of the national academy of sciences. 2014, 111(14):5135-5140.

29 McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003 Feb 12;19(3):362-7.

30 Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 1995, 167(1-2):GC1-GC10.

31 Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2011 Dec 22;28(4):464-9.

32 Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994 Nov 11;22(22):4673-80.

33 Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research. 2004 Jul 1;14(7):1394-403.

34 Galindo-González L, Mhiri C, Deyholos MK, Grandbastien MA. LTR-retrotransposons in plants: Engines of evolution. Gene. 2017, 626:14-25.

35 Lisch D. How important are transposons for plant evolution?. Nature reviews genetics. 2013, 14(1):49.

36 Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, Macas J. Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA. 2011, 2(1):4.

37 De Castro Nunes R, Orozco-Arias S, Crouzillat D, Mueller LA, Strickler SR, Descombes P, Vanzela AL. Structure and distribution of centromeric retrotransposons at diploid and allotetraploid *Coffea* centromeric and pericentromeric regions. Frontiers in plant science. 2018, 9:175.

38 Negi P, Rai AN, Suprasanna P. Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. Frontiers in plant science. 2016, 7:1448

39 Xu S, Brockmöller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, et al. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proceedings of the National Academy of Sciences. 2017, 114(23), 6133-6138.

40 Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS genetics. 2015, 11(1), e1004915.

41 Mao H, Wang H, Liu S, Li Z, Yang X, Yan J et al. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nature Communications. 2015, 6, 8326.

42 Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC genomics. 2016, 17(1):624.

43 Paz RC, Kozaczek ME, Rosli HG, Andino NP, Sanchez-Puerta MV. Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in *Solanum lycopersicum*. Genetica. 2017, 145(4-5):417-430.

44 Kim S, Park J, Yeom SI, Kim YM, Seo E, Kim KT, Kim MS, Lee JM, Cheong K, Shin HS, Kim SB. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome biology. 2017 Dec;18(1):210.

45 Carrizo García C, Barfuss MH, Sehr EM, Barboza GE, Samuel R, Moscone EA, Ehrendorfer F. Phylogenetic relationships, diversification and expansion of chili peppers (*Capsicum*, Solanaceae). Annals of botany. 2016, 118(1):35-51.

46 Manzur JP, Fita A, Prohens J, Rodríguez-Burruezo A. Successful wide hybridization and introgression breeding in a diverse set of common peppers (*Capsicum annuum*) using different cultivated Ají (*C. baccatum*) accessions as donor parents. PLoS One. 2015, 10(12), e0144142.

47 Cremona G, Iovene M, Festa G, Conicella C, Parisi M. Production of embryo rescued hybrids between the landrace “Friariello” (*Capsicum annuum* var. *annuum*) and *C. baccatum* var. *pendulum*: phenotypic and cytological characterization. Euphytica. 2018, 214(8), 129.

48 Staton SE, Bakken BH, Blackman BK, Chapman MA, Kane NC, Tang S, Burke JM. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. The plant journal. 2012, 72(1):142-153.

49 Di Filippo M, Traini A, D'Agostino N, Frusciante L, Chiusano ML. Euchromatic and heterochromatic compositional properties emerging from the analysis of *Solanum lycopersicum* BAC sequences. Gene. 2012, 499(1):176-181.

50 Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Fich EA. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nature genetics. 2014, 46(9):1034.

51 Vernhettes S, Grandbastien MA, Casacuberta JM. The evolutionary analysis of the Tnt1 retrotransposon in *Nicotiana* species reveals the high variability of its regulatory sequences. Molecular biology and evolution. 1998, 15(7):827-836.

52 Melayah D, Lim KY, Bonnivard E, Chalhoub B, Borne FDD, Mhiri C, Grandbastien, MA. Distribution of the Tnt1 retrotransposon family in the amphidiploid tobacco (*Nicotiana tabacum*) and its wild *Nicotiana* relatives. Biological journal of the linnean society. 2004, 82(4):639-649.

53 Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Molecular ecology. 2013, 22(6):1503-1517.

54 Kumar A, Bennetzen, JL. Plant retrotransposons. Annual review of genetics. 1999, 33(1):479-532.

55 Jurka J, Bao W, Kojima KK. Families of transposable elements, population structure and the origin of species. Biology direct. 2011, 6(1):44.

56 Hawkins JS, Proulx SR, Rapp RA, Wendel JF. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proceedings of the national academy of sciences. 2009, 106(42):17811-17816.

57 Yuyama PM, Pereira LFP, Dos Santos TB, Sera T, Vilas-Boas LA, Lopes FR, et al. FISH using a gag-like fragment probe reveals a common Ty 3-gypsy-like retrotransposon in genome of *Coffea* species. Genome. 2012, 55(12), 825-833.

58 Zhang H, Koblížková A, Wang K, Gong Z, Oliveira L, Torres GA, et al. Boom-bust turnovers of megabase-sized centromeric DNA in Solanum species: rapid evolution of DNA sequences associated with centromeres. The Plant Cell. 2014, 26(4), 1436-1447.

59 Lu H, Cui X, Liu Z, Liu Y, Wang X, Zhou Z. Discovery and annotation of a novel transposable element family in *Gossypium*. BMC plant biology. 2018, 18(1):307.

60 Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic acids research. 2010 Oct 29;39(suppl_1):D70-4.

61 Ma B, Xin Y, Kuang L, He N. Distribution and characteristics of transposable elements in the mulberry genome. The plant genome. 2019; doi:10.3835/plantgenome2018.12.0094

62 Schemberger MO, Nogaroto V, Almeida MC, Artoni RF, Valente GT, Martins C,et al. Sequence analyses and chromosomal distribution of the Tc1/Mariner element in Parodontidae fish (Teleostei: Characiformes). Gene. 2016, 593(2), 308-314.

63 Schemberger MO, Nascimento VD, Coan R, Ramos É, Nogaroto V, Ziemniczak K, et al. DNA transposon invasion and microsatellite accumulation guide W chromosome differentiation in a Neotropical fish genome. Chromosoma. 2019, 1-14.

64 Moscone EA, Lambrou M, Ehrendorfer F. Fluorescent chromosome banding in the cultivated species of *Capsicum* (Solanaceae). Plant systematics and evolution. 1996, 202(1-2):37-63.

65 Park M, Jo S, Kwon JK, Park J, Ahn JH, Kim S, Kim BD. Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC genomics. 2011, 12(1):85.

66 Pamponét VCC, Souza MM, Silva GS, Micheli F, de Melo CAF, de Oliveira S G, Corrêa RX. Low coverage sequencing for repetitive DNA analysis in *Passiflora edulis* Sims: cytogenomic characterization of transposable elements and satellite DNA. BMC genomics. 2019, 20(1):262.

67 Houben A, Schroeder-Reiter E, Nagaki K, Nasuda S, Wanner G, Murata M, et al. CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley. Chromosoma. 2007, 116(3), 275-283.

68 Nagaki K, Neumann P, Zhang D, Ouyang S, Buell CR, Cheng Z, Jiang J. Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. Molecular biology evolution. 2005, 22:845-855.

69 Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon.* Nature. 2010, 463:763–768. doi: 10.1038/nature08747

70 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research. 2005;110(1-4):462-7.

71 Birney E, Clamp M, Durbin R. GeneWise and genomewise. Genome research. 2004 May 1;14(5):988-95.

72 El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer EL. The Pfam protein families database in 2019. Nucleic acids research. 2018 Oct 24;47(D1):D427-32.

73 Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, Lahrmann U, Zhao Q, Zheng Y, Zhao Y, Xue Y. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics. 2015 Jun 10;31(20):3359-61.

74 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution. 2018, *35*(6):1547-1549.

75 Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, Wishart DS. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res. 2016. (epub ahead of print). doi:10.1093/nar/gkw419

76 Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bullettin. 1987, 19:11-15.

77 Schwarzacher T, Ambros P, Schweizer D. Application of Giemsa banding to orchid karyotype analysis. Plant systematics and evolution. 1980, 134(3-4):293-297.

Tables

Table 1: List of primers of reverse transcriptase sequences, that have been designed for probes obtaining and FISH.

Elements	Primers
Athila/Tat - RT	F 5’ GGGTGGTATTGCTTCTTGGA 3’
	R 5’ GAATCACCTACCACAGAG 3’
CRM - RT	F 5’ CCACCAACAAGATAACGG 3’
	R 5’ CCATCCATTCATAGAGACC 3’
Tekay/Del - RT	F 5’ GTTCAGGGTGCCAAGTGT 3’
	R 5’GGGCGTTAGTCAACCTGAAG 3’
Ivana/Oryco - RT	F 5' GGTTCAAGGATCGGTTGATAG 3’
	R 5’ GTTGAGCTTGCACACCATGT3’

RT = reverse transcriptase

Table 2. Frequency and relative values of repetitive fraction in the datasets of the three *Capsicum* genomes.

Species/Lineages		*C. annuum*				*C. chinense*				*C. baccatum*			
		num seq	NR (%)	size bp	size CR%	num seq	NR (%)	size bp	size CR%	num seq	NR (%)	size bp	size CR%
Copia	Sirevirus	7421	1.69	1339493	1.59	9176	2.29	1361598	0.26	3463	0.94	546478	0.69
	Osser	45	0.01	1335	0.00	65	0.02	2534	0.00	6	0.00	241	0.00
	Tork	5380	1.22	1645586	1.95	2378	0.59	357575	0.07	8773	2.37	5231667	6.58
Gypsy	Chromovirus	321932	73.28	58580214	69.41	278519	69.43	49370284	7 95.71	78757	21.28	15931607	20.03
	Non-chrom.	15	0.00	935	0.00	37	0.01	1242	0.00	13	0.00	970	0.00
	OTA	70160	15.97	13807983	16.36	74530	18.58	14025454	2.72	22614	3 61.11	34401828	43.24
Other LTRs		12836	2.92	5071268	6.01	886	0.22	44305	0.01	3845	1.04	1439107	1.81
Non-LTR-RTs		0	0.00	1866004	2.21	3305	0.82	296857	0.06	3648	0.99	1201048	1.51
Endogenous Virus		1371	0.31	127214	0.15	1392	0.35	71067	0.01	1171	0.32	185499	0.23
Transposons		30738	7.00	5840435	6.92	28627	7.14	5135873	1.00	29582	7.99	7154489	8.99
5S rDNA		717	0.16	89820	0.11	1370	0.34	172715	0.03	199	0.05	30499	0.04
45S rDNA		1267	0.29	1086206	1.29	849	0.21	637555	0.12	18075	4.88	14858366	18.68

Repetitive fraction comparison in *Capsicum* genomes. num seq – number of sequences found after the Blast rounds. NR (%) – relative value. Size bp – total length of a repetitive class. Size CR (%) – percentage that the class represents in the scaffolds.

Figures

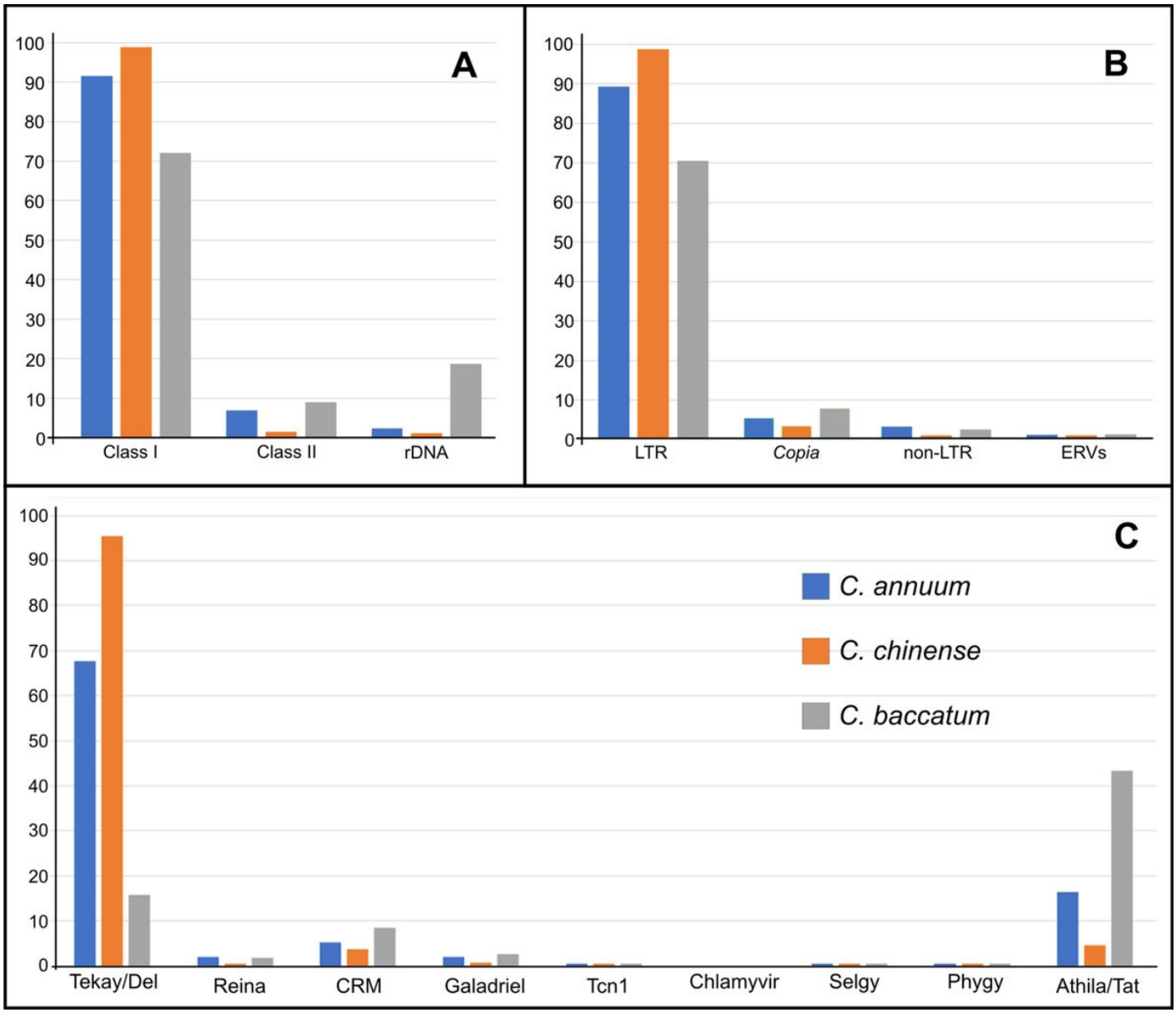

Figure 1

Comparison of the relative distribution (%) of repetitive DNA families among three Capsicum datasets. (A) Note that the Class I elements are more representative than Class II and rDNA sequences in the three cases. (B) LTR-RT elements predominated over non-LTR and ERVs, but note that the C. baccatum genome exhibited 30% fewer sequences than other two datasets. Observe also that Copia superfamily elements were less accumulated (<10%) than Gypsy ones. (C) Observed that the most accumulated elements were Tekay/Del, Athila/Tat, and CRM, but, except for CRM lineage, there was a big difference in the quantity of the elements in each dataset.

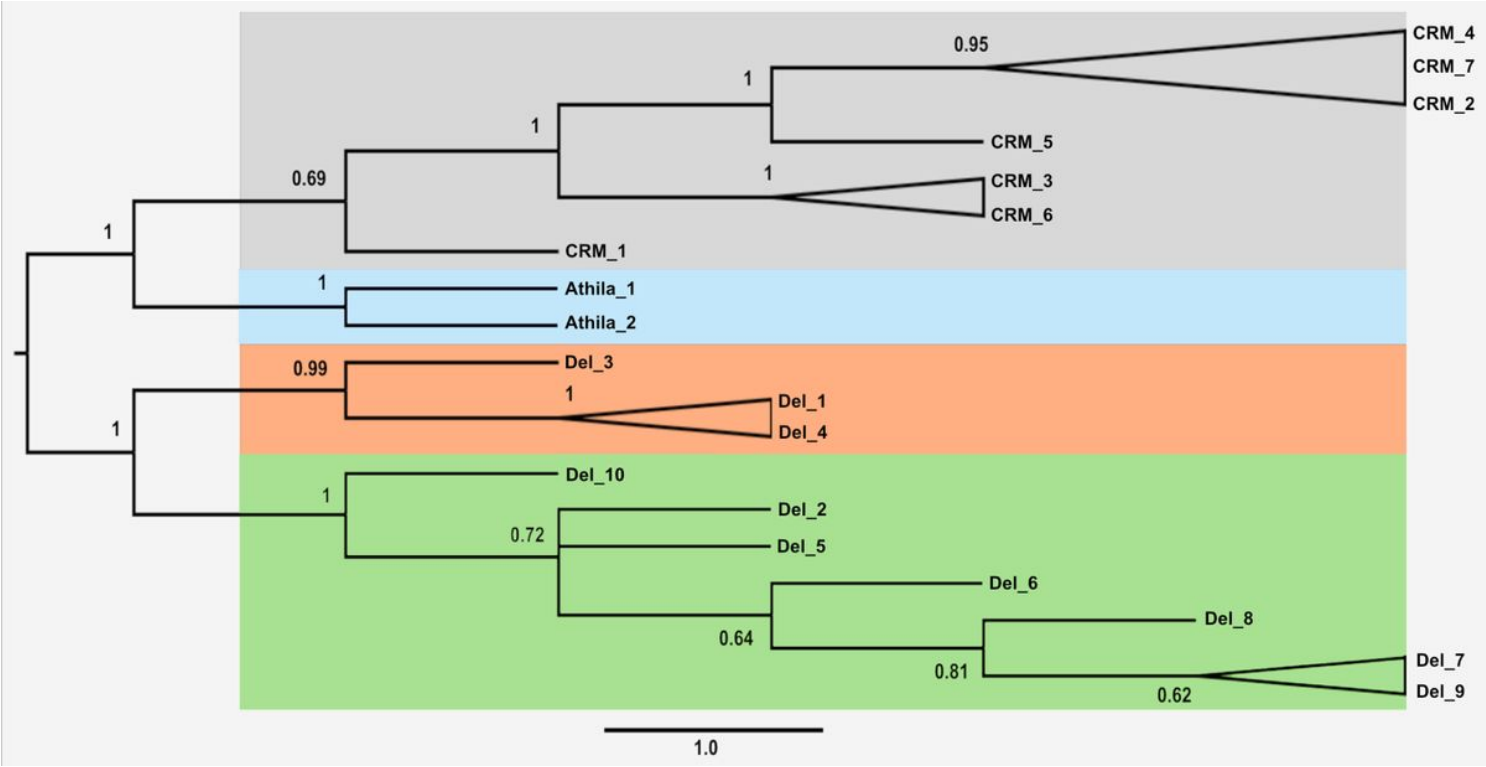

Figure 2

Phylogenetic tree based on putative autonomous sequences of Capsicum using the maximum likelihood method with bootstrap 1000. CRM sequences organize four groups (grey, see also the figure S2) and Athila sequences two groups (blue, see also the figure S3). The Del sequences were organized in two well-supported groups (orange), corroborating with the MAUVE alignment (Figure S1), and the third group of sequences without well-supported values (green), except for the sequences Del_7 and Del_9 (Figure S1) corroborating with the MAUVE alignment.

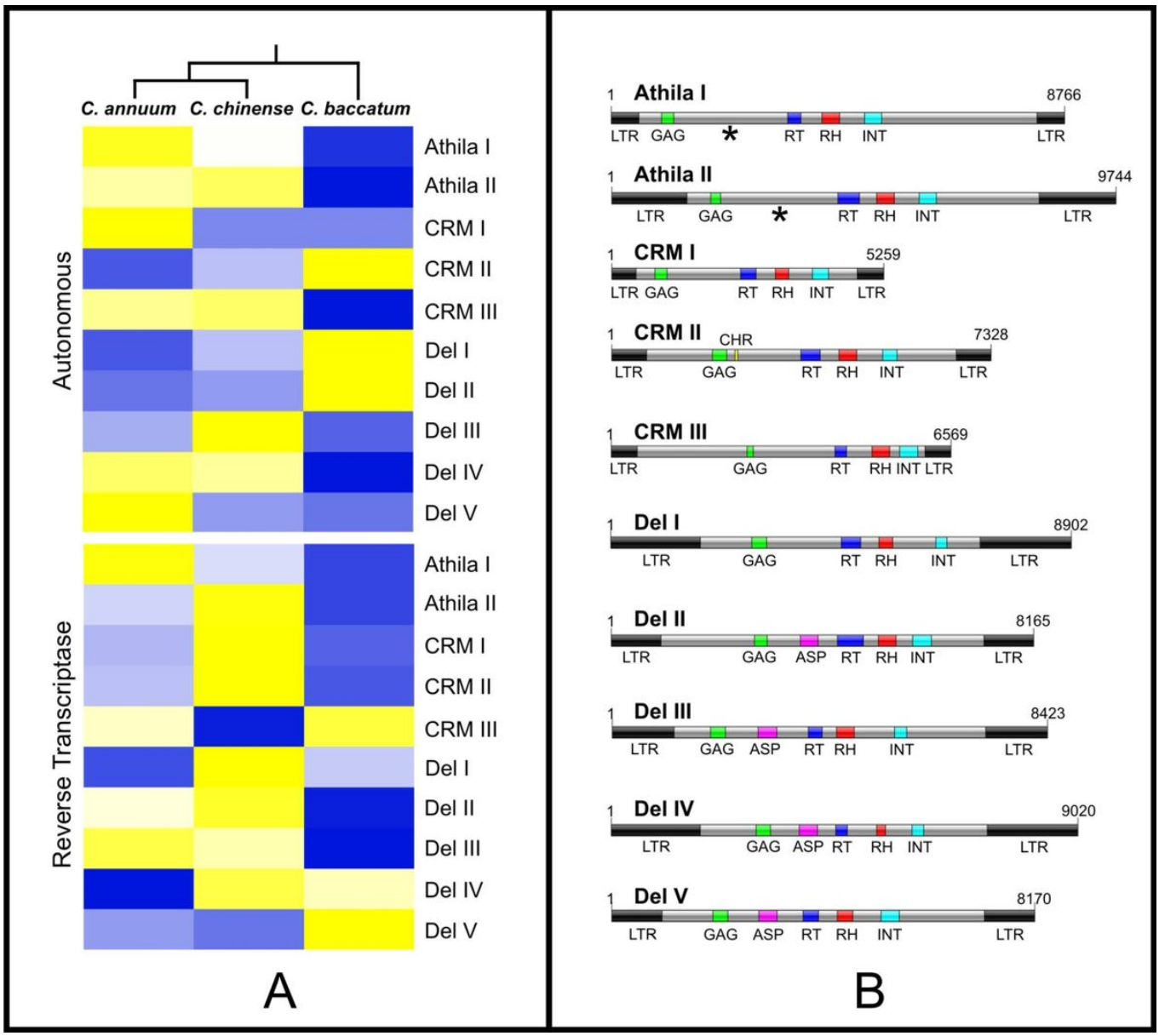

Figure 3

Comparative distribution of the putative autonomous LTR-RTs and the reverse transcriptase sequences along the C. annuum, C. chinense, and C. baccatum datasets. (A) The Annuum clade, composed by C. annuum and C. chinense and the Baccatum clade (C. baccatum) can be distinguished by the dendrogram on top of the heatmap. In the Heatmap, lower and higher accumulation (blue, intermediate and yellow, respectively) represent the amount of conserved sequences found in each dataset. Image shows that Athila/Tat elements accumulated more in C. annuum and C. chinense (marked in yellow and light-yellow), the CRM groups were differentially accumulated, highlighting the absence of CRM III and the predominance of Del I and II in C. baccatum, and bigger accumulation of Del III, IV, and V in C. annuum and C. chinense. The reverse transcriptase of these elements exhibits a similar pattern of distribution than the one observed for the complete elements, Athila/Tat I and II were more accumulate in C. annuum and C. chinense, respectively. CRM I and II were more accumulated in C. chinense, while the group CRM III was more accumulated in C. baccatum. Capsicum annuum and C. chinense had a bigger accumulation of Del I, II and II than C. baccatum, while the groups Del III and IV exhibited more accumulation in C. baccatum. (B) Graphical representation of LTR-RTs groups. LTR – long terminal repeat, GAG – nucleocapsid, RT – reverse transcriptase, RH – RNAse H, INT – integrase, ASP – aspartase. Asterisks present in Athila illustrations refer to the hallmark ORF for this lineage. Note a difference in extension (bp length) among elements, including GAG and POL positioning, and LTR sizes. Note also that only CRM II exhibits a chromodomain sequence and that all the Del elements present an additional aspartase locus.

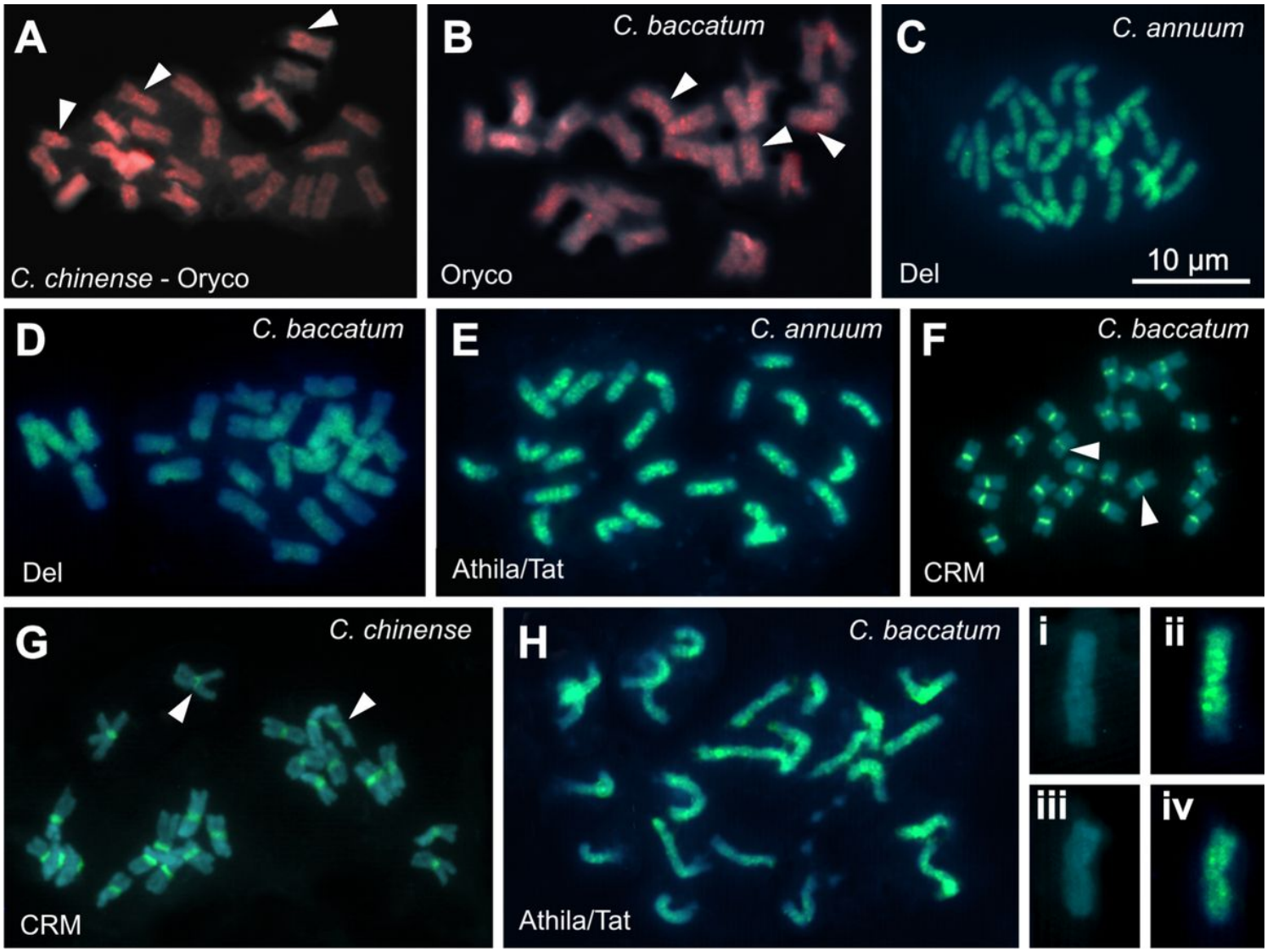

Figure 4

FISH using LTR-RTs probes against metaphases and prometaphases of Capsicum species. Chromosomes were counter-stained with DAPI (blue), Copia probes with Cy3-11-dUTP (red) and Gypsy probes labelled with biotin-11-dUTP / avidin–FITC conjugate (green). The Copia Ivana/Oryco probe showed few hybridization signals scattered along chromosomes, with a low accumulated profile in both C. chinense (A) and C. baccatum (B). The Gypsy Tekay/Del probe exhibited hybridization signals dispersed along the chromosomes in the three species, but with a larger accumulation in C. annuum chromosomes (C) than the other two species, such as in C. baccatum (D). The Gypsy Athila/Tat probe showed brighter hybridization signals than Tekay/Del, accumulating in the pericentromeric to interstitial regions of all C. annuum chromosomes (E), differently of C. baccatum because some chromosomes accumulated many signals and others very few (H). The boxes i, ii, iii and iv are highlighting differences in the pericentromeric and interstitial Athila/Tat signals in two C. baccatum chromosomes. The Gypsy CRM probe showed FISH signals accumulated in the centromeric regions, but with two pairs in each species with much less intense signals. Note the arrows in C. baccatum (F) and C. chinense (G). The bar represents 10 µm.

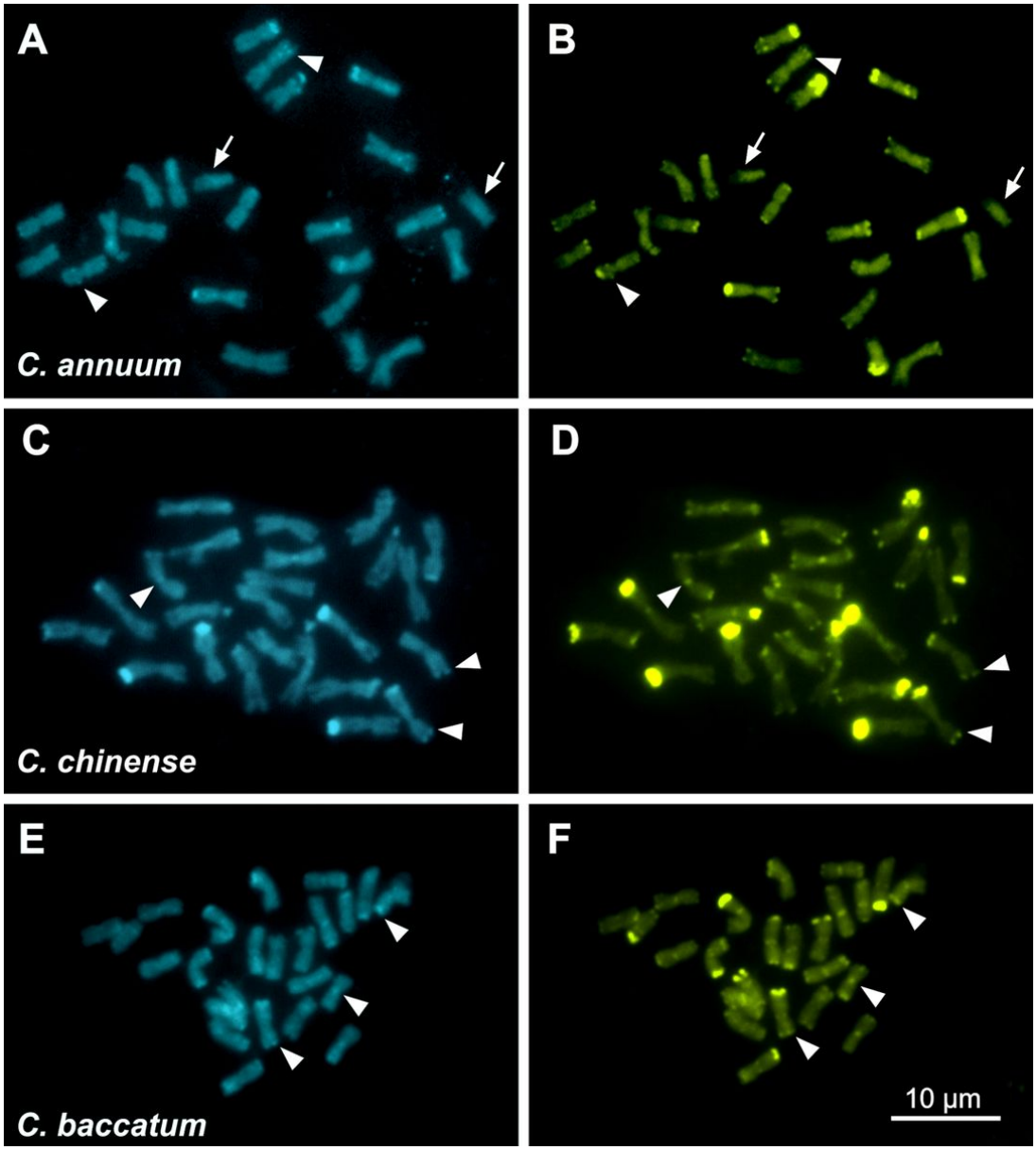

Figure 5

Capsicum species show considerable diversity in the C-CMA/DAPI banding profiles. Observe that C. annuum presents four chromosomes without fluorescent signals. C-DAPI interstitial dots was detected in three chromosome pairs and terminal bands two pairs (A), all of them were co-located with C-CMA bands. Ten pairs exhibit C-CMA signals, being three pairs with stronger terminal and the other as the terminal to subterminal small signals (B). Capsicum chinense showed four pairs with strongest DAPI signals, besides minor centromeric, interstitial and terminal signals on a few chromosomes (C), while C-CMA bands were observed in all the chromosomes, varying as strongest terminal bands in seven pairs, minor centromeric bands in six pairs and as terminal and interstitial dots in ten pairs (D). Some of these bands have been evidenced by DAPI and CMA3 (C-D). Observe that C. baccatum exhibits six pairs with minor terminal and six with centromeric and/or interstitial C-DAPI (E). C-CMA signals were detected in four pairs as strongest terminal bands, but minor terminal bands were observed in all chromosomes, as well as minor centromeric in almost all the chromosomes (F). The bar represents 10 µm.

Supplementary Files

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

- Supplementarymaterialsrev2.pdf