RAFAEL DE ASSIS

DNAS REPETITIVOS EM GENOMAS DE SOLANACEAS:

Capsicum L. E *Solanum* L. COMO MODELOS DE ESTUDO

Londrina
2023

RAFAEL DE ASSIS

DNAS REPETITIVOS EM GENOMAS DE SOLANACEAS:

Capsicum L. E *Solanum* L. COMO MODELOS DE ESTUDO

Londrina
2023

RAFAEL DE ASSIS

DNAS REPETITIVOS EM GENOMAS DE SOLANACEAS:

Capsicum L. E *Solanum* L. COMO MODELOS DE ESTUDO

Tese apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de Doutor.

Orientador: Dr. André Luis Laforga Vanzela.

Londrina
2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Assis, Rafael de.

DNAS repetitivos em genomas de solanaceas: Capsicum L. E Solanum L. como modelos de estudo / Rafael de Assis. - Londrina, 2023.
119 f.

Orientador: André Luís Laforga Vanzela.
Tese (Doutorado em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2023.
Inclui bibliografia.

1. Centrômero - Tese. 2. FISH - Tese. 3. DNA satélite - Tese. 4. Retrotransposons - Tese. 1. Laforga Vanzela, André Luís. Ii. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. Iii. Título.

CDU 575.1

RAFAEL DE ASSIS

DNAS REPETITIVOS EM GENOMAS DE SOLANACEAS:

Capsicum L. E *Solanum* L. COMO MODELOS DE ESTUDO

Tese apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de Doutor.

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. André Luis Laforga Vanzela
Universidade Estadual de Londrina - UEL

_____

Profª. Drª. Renata da Rosa
Universidade Estadual de Londrina

_____

Drª. Thaíssa Boldieri de Souza
Universidade Estadual de Londrina - UEL

_____

Prof. Dr. Paulo Roberto da Silva
Universidade Estadual do Centro-Oeste - UNICENTRO

_____

Dr. Romain Guyot
Institut de Recherche pour le Développement - IRD

Londrina, 27 de janeiro de 2023.

AGRADECIMENTOS

A CAPES pela concessão da bolsa durante o período de doutorado e pela bolsa do Programa de Doutorado Sanduíche no Exterior.

Ao Programa de Pós-graduação em Genética e Biologia Molecular da Universidade Estadual de Londrina.

Ao meu orientador, Prof. Dr. André Luis Laforga Vanzela, por todos esses anos em que fiz parte de seu laboratório, por todas as oportunidades e crescimento e que tive ao longo dos anos de mestrado e doutorado.

Ao Dr. Romain Guyot, muito obrigado pelo acolhimento e conhecimentos compartilhados durante o período em que estive em seu laboratório no IRD em Montpellier.

Ao Prof. Dr. Paulo Roberto da Silva, que foi meu orientador durante toda a graduação e foi em seu laboratório onde comecei como IC.

A todos os professores com quem tive aula ao longo de toda a minha formação.

A minha família, vocês foram ponto essencial para que esse objetivo fosse alcançado, amo vocês.

Aos amigos do Laboratório de Citogenética e Diversidade Vegetal (LCDV), por todos os perrengues e momentos de descontração que dividimos e mais do que isso, por todo o conhecimento que foi compartilhado.

A minha querida amiga Renata Giacomin, minha primeira parceira de laboratório e amiga para a vida.

Aos meus amigos, que sempre me apoiaram e entenderam minhas ausências, vocês são extremamente importantes e deixam a vida mais leve.

Aos membros da banca de avaliação deste trabalho, certamente suas sugestões farão grande diferença.

Muito obrigado!

ASSIS, Rafael de. **DNAs repetitivos em genomas de Solanaceas:** *Capsicum* e *Solanum* como modelos de estudo. 2023. 119 f. Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina, 2023.

RESUMO

Os genomas vegetais são compostos principalmente por sequências repetitivas. Elementos transponíveis (ETs), que formam a parte móvel dos genomas, são classificados de acordo com seu mecanismo de transposição (Classes I e II), estes podem acumular diferencialmente dependendo do grupo vegetal. A fração repetitiva não codificante, representada pelas sequências satélite tendem a se acumular em blocos ao longo dos cromossomos. Algumas regiões cromossômicas tendem a ser *hotsposts* para o acúmulo de sequências repetitivas, como por exemplo os centrômeros e as regiões terminais. A família Solanaceae apresenta distribuição cosmopolita, e espécies com grande valor econômico, como por exemplo as batatas, tomates, berinjelas, pimentas, tabaco, entre outras. Devido a esse grande impacto econômico, muitas das espécies já possuem seu sequenciamento genômico disponível em repositórios online. Os gêneros *Capsicum* L e *Solanum* L. são próximas filogeneticamente. As pimentas e pimentões pertencem ao gênero *Capsicum* L, enquanto o gênero *Solanum* L. possui a maior diversidade, com mais de 1.500 espécies descritas, as quais inclui o tomate cultivado e as espécies selvagens. Mesmo sendo espécies com grande valor comercial, alguns pontos ainda permanecem poucos explorados, como por exemplo as sequências de DNA satélite que compõe esses genomas e sua localização em espécies próximas, bem como a diversidade de sequências repetitivas presente nos centrômeros. Diante disso, o objetivo geral deste trabalho foi avaliar frações repetitivas em grupos distintos dentro da família Solanaceae, utilizando como organismos modelo as pimentas e os tomates. Para isso, foram utilizadas ferramentas de bioinformática para identificar essa fração em sequenciamento de alta cobertura e métodos de citogenética molecular para a localização física. Com base na mineração de sequências altamente repetitivas, foi possível identificar duas famílias de DNA satélite que se encontram acumuladas em regiões distais dos cromossomos de *Capsicum*, que previamente eram descritas como sendo compostas por sequencias derivadas de DNAr que formavam um megassatelite. Essas famílias de DNA satélite diferem entre si quanto ao número de sítios hibridizados nos cromossomos de *Capsicum*, bem como na sua provável origem. Foi possível também identificar os elementos de transposição e as sequências de DNA satélite que compõem as prováveis regiões centroméricas de espécies de *Solanum* que formam o clado das espécies de tomate. As análises mostraram que espécies de tomate apresentam poucas cópias da linhagem de retrotransposon CRM, que esta frequentemente associada a região centromérica, associada a essa região está o retrotransposon-like TGR4 e o elemento *Jinling*.

Palavras-chave: Centrômero; FISH; genomas; Retrotransposons; SatDNA.

ASSIS, Rafael de. **Repetitive DNA in Solanaceae genomes:** *Capsicum* and *Solanum* as research models. 2023. 119 p. Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina, 2023.

ABSTRACT

Plant genomes are mainly composed of repetitive sequences. Transposable elements (TEs), which form the mobile part of the genomes, are classified according to their transposition mechanism (Classes I and II), these can accumulate differentially depending on the plant group. The non-coding repetitive fraction, represented by satellite sequences tend to accumulate in blocks along the chromosomes. Some chromosomal regions tend to be hotspots for the accumulation of repetitive sequences, for example, the centromeres and terminal regions. The Solanaceae family has a cosmopolitan distribution, and species with great economic value, such as potatoes, tomatoes, eggplants, peppers, tobacco, among others. Due to this great economic impact, many of the species already have their genome sequencing available in online repositories. The genus *Capsicum* L. and *Solanum* L. are phylogenetically close. The peppers belong to the genus *Capsicum* L., and *Solanum* L. is considered to have the greatest diversity, with more than 1,500 described species, which includes the commercial tomato and the wild species. Even being species with great commercial value, some points remain unexplored, such as the satellite DNA sequences that compose these genomes and their location in close species, as well as the diversity of repetitive sequences present in the centromeres. Given this, the overall objective of this work was to evaluate repetitive fractions in distinct groups within the Solanaceae family, using peppers and tomatoes as model organisms. To do so, bioinformatics tools were used to identify this fraction in high coverage sequencing and molecular cytogenetic methods for physical localization. Based on the mining of highly repetitive sequences, it was possible to identify two families of satellite DNA that are accumulated in distal regions of *Capsicum* chromosomes, which were previously described as being composed of rDNA-derived sequences forming a megasatellite. These families of satellite DNA differed in the number of hybridized sites on *Capsicum* chromosomes, as well as in their probable origin. It was also possible to identify the transposon elements and satellite DNA sequences that make up the probable centromeric regions of *Solanum* species that form the tomato species clade. The analyses showed that tomato species have few copies of the retrotransposon lineage CRM, which is frequently associated with the centromeric region, associated with this region is the retrotransposon-like TGR4 and the Jinling element.

Keywords: SatDNA; Retrotransposons; Centromere; FISH; genomes.

LISTA DE FIGURAS

FUNDAMENTAÇÃO TEÓRICA

Figura 1 - Principais clados da família Solanaceae. Seta indica o início do clado das espécies que possuem número básico x=12. Adaptado de Solanaceae Source 15

Figura 2 - Filogenia proposta para o gênero Capsicum adaptada de Carrizo-Garcia e colaboradores (2016). Árvore obtida por análise de máxima parcimônia com base nos genes matK, psbA-trnH e waxy.....17

Figura 3 - Distribuição das espécies quem compõem o clado dos tomates e de espécies próximas filogeneticamente na região andina.....19

Figura 4 - Sistema de classificação proposto por Wicker e colaboradores (2007) para os elementos de transposição de Classe I. Adaptado de Wicker e colaborados (2007).....21

Figura 5 - Sistema de classificação proposto por Wicker e colaboradores (2007) para os elementos de transposição de Classe II. Adaptado de Wicker e colaborados (2007)..22

CAPÍTULO 1

Figure 1 - Dotplot of distinct contigs from *C. baccatum* containing repeats of CDR-1 and CDR-2 satellites. Repetitive profile of the satellites analyzed; for better visualization, only a fraction of the scaffold was used. The differences that can be observed in the repeat profile between the two satDNAs are due to monomer length. Note that the two satellite sequences are not interspaced in the contigs. Additionally, CDR-1 appears to have more degeneration along the repeats, while CDR-2 seems to be more conserved 63

Figure 2 - Dotplot of CDR-1 and CDR-2 against sequences of rDNA from *Capsicum* species. Note that there was no similarity between the satellites and the rDNA sequences, indicating that the satellites are not part of the cistron. 64

Figure 3 - The presence of the CDR-2 monomer in the retrotransposon sequences. (A) Dotplot of the CDR-2 monomer repeated 10 times

against one sequence that carries the monomer. (B) A zoomed-in box highlighting the number of copies of CDR-2 present within the LTR sequence; there are at least 10 copies of the monomer. (C) The organization of the LTR sequence that carries the CDR-2 monomer. The sequence belongs to the Gypsy superfamily, and by alignment of the RT domain, it was classified as belonging to the TAT/Ogre lineage. Note that even though the retrotransposon sequence belongs to the TAT/Ogre lineage, it lost both LTRs. 65

Figure 4 - FISH assay using CDR-1, CDR-2, and 35S rDNA probes against metaphases and prometaphases of *Capsicum baccatum* and *C. chinense*. The sample was counterstained with DAPI (blue), and probes were counterstained with Cy3 (red) and avidin–FITC conjugate (green). (A and B) Double FISH in *C. baccatum* with CDR-1 (red) and CDR-2 (green) probes; boxes indicate chromosomes that have a colocalization of the satellites. Observe that CDR-1 localizes immediately below the secondary constriction (arrowheads). (C) CDR-1 and (D) CDR-2 probes in *C. chinense*. (E) Merged images in *C. chinense* highlighting the colocalization of the two satellites. (F) FISH in *C. baccatum* with the CDR-1 probe; the arrows show some signals that were not colocalized with the 35S rDNA probe. (G) FISH in *C. baccatum* with 35S rDNA. (H) Merged images of CDR-1 and 35S rDNA (F and G). A yellowish signal indicates the colocalization of signals. The bar represents 10 μm. 66

Figure 5 - Distribution of 5S rDNA, CDR-2, 35S rDNA, and CDR-1 sequences in pseudochromosomes from *Capsicum annuum*. The 5S coding sequence exhibited only one peak of accumulation on chromosome 7. The CDR-2 monomer exhibited two accumulation peaks, the higher one on chromosome 5 and the minor one on chromosome 9. Sequences of 35S rDNA appeared more accumulated in pseudochromosomes 2 and 6, but small peaks can be seen in the sub-terminal regions of chromosomes 8, 9, and 10 and more interstitial in chromosome 5. All the pseudomolecules exhibited accumulation of CDR-1 monomers in the sub- terminal regions, but with some variation in their concentration among pseudochromosomes. 67

CAPÍTULO 2

Figure 1 - Phylogenetic tree and repetitive elements comparative analysis. The left panel is the chloroplast phylogenetic tree, the number are the bootstrap values and after each specie the genome size (1C) in Mbp, *S. tuberosum*, *Capsicum annuum*, and *Vitis vinifera* were used as outgroups. The right panel is the RepeatExplorer comparative analysis. Tekay were the most representative followed by Athila (both from Gypsy superfamily) and LINEs. Note that even with a well annotated database, some elements were only addressed to Gypsy superfamily 107

Figure 2 - RT domains phylogenetic tree from the full-length elements retrieved in *Solanum lycopersicum* with the EDTA pipeline and annotated with Inpactor. For this tree the branch length was considered. The majority of full-length elements were grouped with Tekay references, the red circle in Tekay branch indicate the clade with TGRIV copies. The blue circle refers to the CRM clade, only the references compose this clade due the lack of full-length sequences from this lineage. Highlight among the Copia elements is the Rider clade, which was grouped with Tork elements.....108

Figure 3 - Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum lycopersicum* cv. Moneyberg. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS exhibited three major peaks (chromosomes 6, 8, and 11) colocalized with TGRIV peaks, and four minor peaks (chromosomes 1, 2, 3, and 12). The sequences from 35s rDNA exhibited a major peak in the chromosome 2 and two minors in chromosomes 6 and 11.....109

Figure 4 - Dot-plot of the TGRIV sequence extracted from the centromeric region of *Solanum lycopersicum.* Note that this element does not carry any sequence similar to satellite.....110

Figure 5 - Dot-plot of the centromeric Retrotransposon sequence extracted from the centromeric region of *Solanum lycopersicum.* Note that this element carries a satellite sequence at the LTRs 111

Figure 6 - Dot-plot of the centromeric Retrotransposon sequence extracted from

the centromeric region of *Solanum lycopersicum* versus the *Jinling* element. Note that even the high similarity between the sequences the sequence, there are some stretches with deletions 112

Figure 7 - FISH assay using satTCS, TGRIV, *Jinling,* and 35S rDNA probes against metaphases and prometaphases of *Solanum lycopersicum* **(A-F)**, *S. pimpinellifolium* **(G-J)**, and *S. cheesmaniae* **(K-L)**. The sample was counterstained with DAPI (blue), and probes were counterstained with Cy3 (red) and avidin–FITC conjugate (green). **(A)** Double FISH in interphasic nucleus of *S. lycopersicum* with satTCS (red) and 35S rDNA (green) probes, **(B)** the same probes in metaphasic chromosomes, and **(C-D)** only the satTCS probe in metaphasic chromosomes. **(E)** TGRIV (green), the boxes highlight the signals present in the centromeric region of chromosomes from *S. lycopersicum*. **(F)** *Jinling* probe in metaphase from *S. lycopersicum*. Note the scattered profile from this probe, typical from the dispersed retrotransposons. **(G, I and J)** FISH in *S. pimpinellifolium* with the satTCS probe, and **(H)** the same species in a double FISH with 35S rDNA probe and the satellite. Note that in this specie the number of chromosomes with signal is bigger the *S. lycopersicum*, and is possible to see minor signals, always in the proximal region. **(K-M)** FISH in *S. cheesmaniae* with the satTCS, and 35S rDNA probes.....113

SUMÁRIO

1 INTRODUÇÃO 12
1 FUNDAMENTAÇÃO TEÓRICA 14
1.1 Família Solanaceae 14
1.2 O gênero Capsicum 15
1.3 O gênero Solanum 18
1.4 DNAs repetitivos 19
1.5 Elementos de transposição 20
1.6 DNAs satélite 22
2 OBJETIVOS 24
2.1 Objetivo geral 24
2.2 Objetivos específicos 24
3 REFERÊNCIAS BIBLIOGRÁFICAS 25
4 CAPÍTULO 1 – Abundance of of distal repetitive DNA sequences in CAPSICUM L. (Solanaceae) chromosomes 31
INTRODUCTION *..... 34*
MATERIALS AND METHODS *..... 36*
Plant material 36
Genomic analysis 36
Identification of the repetitive fraction in the 25S-18S rDNA IGS sequences 38
DNA extraction, PCR, and oligo probes 39
Fluorescent in situ hybridization (FISH) 40
RESULTS *..... 41*
Characterization of satDNA sequences from Capsicum genomes 41
CDR-1 and CDR-2 occur in the terminal chromosome regions 44
DISCUSSION *..... 44*
rDNA and satellite sequences are colocalized at the Capsicum chromosome ends 44
CDR-1 and CDR-2 belong to different repeating families with distinct origins and fates 47
REFERENCES *..... 53*
IMAGES *..... 63*
SUPPLEMENTARY MATERIAL *..... 68*

5 CAPÍTULO 2 - Comparative analysis of retrotransposons among tomato species and the characterization of centromeric elements 82
INTRODUCTION *84*
MATERIAL AND METHODS *87*
Chloroplast genome assembly and phylogenetic reconstruction 87
Repeats annotation 87
De novo genome assembly and search for RT domains 88
Transposable elements annotation 88
Annotation of centromeric sequences 88
Plant material 89
Genome size estimation 89
DNA extraction, PCR, and probe design 90
Chromosome preparation for FISH 91
RESULTS *92*
Phylogenetic inferences and genome size 92
Repeats diversity annotation 92
Centromeric composition among tomato species 93
DISCUSSION *95*
REFERENCES *100*
IMAGES *106*
SUPPLEMENTARY MATERIAL *114*
6 Conclusões 118

1 INTRODUÇÃO

A família Solanaceae possui aproximadamente 100 gêneros com mais de 2700 espécies descritas as quais apresentam distribuição cosmopolita, com ampla variedade de hábitats, morfologia e ecologia (OLMESTEAD *et al.*, 2008). Entre os gêneros que compõe essa família, destacam-se o gênero *Capsicum* por ser o único que possui pungência e o gênero *Solanum* L. como o mais diverso (Knapp 2008; AZA-GONZÁLEZ *et al.*, 2011). Espécies de *Capsicum* L. exibem grande variedade de formas, tamanhos e cores de frutos, a pungência característica do grupo é devido ao acúmulo dos capsaicinóides, e tais características fazem com que o grupo possua grande importância econômica (AZA-GONZÁLEZ *et al.*, 2011; QIN *et al.*, 2014; KIM *et al.*, 2014). O gênero possui 43 espécies, das quais cinco são consideradas domesticadas e as demais silvestres (MOSCONE *et al.*, 2006; CARRIZO- GARCÍA *et al.*, 2013; THE PLANT LIST, 2013). Os estudos filogenéticos sugerem três principais clados Annuum, Baccatum e Pubescens (CARRIZO-GARCÍA *et al.*, 2016). O gênero *Solanum* possui grande impacto na economia devido a suas espécies amplamente utilizadas na alimentação como *Solanum melongena* L. (berinjela), *S. tuberosum* L. (batata), *S. lycopersicum* L. (tomate). O tomate cultivado (*S. lycopersicum* L.) pertence a seção *Lycopersicum* do gênero *Solanum*, juntamente com mais 12 espécies selvagens (KNAPP, PERALTA, 2016). Espécies desse gênero possuem grande potencial como modelo de estudos genéticos e de processos biológicos, tal qual *Arabidopsis thaliana*, devido ao seu genoma relativamente pequeno (2C = ~2 pg), grande quantidade de metabolitos secundários e mapas cromossômicos bem estruturados (marcadores clássicos e moleculares) para a espécie cultivada (TANKSLEY 1993; RICK, YODER 1998).

Entre as frações que compõe os genomas vegetais, a fração repetitiva pode ser organizada em grupos variados de acordo com a natureza e o modo de repetição das sequências (HESLOP-HARRISON E SCHWARZACHER, 2011; BENNETZEN E WANG, 2014). Esta fração repetitiva pode alcançar até 90% nos genomas de algumas angiospermas (AMBROŽOVÁ *et al.*, 2011). De maneira geral, os DNA repetitivos podem ser encontrados dispersos (elementos de transposição) ou organizados em *tandem* (DNA ribossômico, microssatélites, minissatélites, satélites, sequências teloméricas e centroméricas e elementos de transposição – TEs), e podem ser classificados de acordo com o tamanho dos motivos repetidos, composição de bases e localização nos cromossomos.

Tendo em vista que a livre disponibilidade dos genomas completamente sequenciados e montados, características morfológicas contrastantes e diferenças na composição genômica entre as espécies próximas, o foco desse estudo foi identificar e caracterizar sequências repetitivas em espécies de dois gêneros da família Solanaceae, *Capsicum* e *Solanum*, e localizá-las fisicamente procurando associar a localização cromossômica dessas sequências com sequências previamente descritas.

1 FUNDAMENTAÇÃO TEÓRICA

1.1 FAMÍLIA SOLANACEAE

A família Solanaceae pertence a ordem Solanales e é formada por cerca 100 gêneros com mais de 2700 espécies descritas até o momento e que apresentam distribuição cosmopolita. Esse grupo vegetal possui ampla variedade de hábitats, morfologia e ecologia (OLMESTEAD *et al.*, 2008). Os principais representantes dessa família são os tomates, batatas e berinjela (*Solanum* L.), pimentas (*Capsicum* L.), tabaco (*Nicotiana* L.), petúnias (*Petunia* Juss.) e as damas-da-noite (*Cestrum nocturnum.* L.), tais espécies possuem grande importância econômica e abrangem cerca de 60% de todas as espécies. Destacando-se como o maior gênero, *Solanum* abrange cerca de 1500 espécies descritas e muitas delas com grande importância comercial (OLMESTEAD *et al.*, 2008, STEHMANN *et al.*, 2015).

Em relação as características morfológicas, os representantes da família Solanaceae podem ser herbáceos, arbustivos, arbóreos, escandentes e epifíticos dos mais variados ambientes terrícolas (HUNZIKER, 2001). Possui folhas alternas, simples ou compostas, inteiras ou lobadas, sem estípulas, glabras, pubescentes ou tomentosas, com indumento constituído de diferentes tipos de tricomas, desde simples, glandulares até os estrelados, dendríticos, lepidotos, dentre outros, ocorrendo principalmente em *Solanum*. Em relação a morfologia floral, suas flores são perfeitas, de pequenas a grandes e vistosas, pentâmeras, actinomorfas ou zigomorfas, diclamídeas, solitárias ou em diferentes tipos de inflorescências terminais ou laterais, cálice gamossépalo, podendo ser acrescente e ampliado no fruto, corola gamopétala de diferentes formas (tubulosa, campanulada, rotácea ou estrelada) com estames adnatos ao tubo da corola. Os frutos podem ser bagas, drupas ou cápsulas (NEE, 1999; HUNZIKER, 2001; KNAPP, 2002).

As descrições citogenéticas sobre a família Solanaceae sugerem a predominância de cariótipos com $2n = 24$ (FIGURA 1), e variações quanto a esse valor podem estar relacionadas com eventos de poliploidia, aneuploidia e disploidia (BADR *et al.*, 1997; MOSCONE *et al.*, 2007; PADILHA *et al.*, 2016). Algumas espécies pertencentes ao gênero *Solanum* e *Lycianthes*, por exemplo, apresentam $2n= 24$ e cariótipos simétricos, onde há predominância de cromossomos meta- e submetacêntricos (ACOSTA *et al.*, 2005; ACOSTA, GUERRA E MOSCONE, 2012). O gênero *Solanum* possui ainda espécies poliplóides que podem apresentar até $2n = 8x$

= 96 (HUNZIKER, 2001). No gênero *Capsicum*, predominam os números haploides *n*= 12 e 13 e cariótipos simétricos com cromossomos meta- e submetacêntricos, de modo geral, cariótipos com *n*=12 são compostos por 11 pares metacêntrico e 1 par submetacêntrico, sendo essa característica mais observada nas espécies domesticadas. Tais características foram descritas por MOSCONE *et al.* (1993; 2007), SCALDAFERRO, GABRIELE E MOSCONE (2013) e SCALDAFERRO *et al.* (2016).

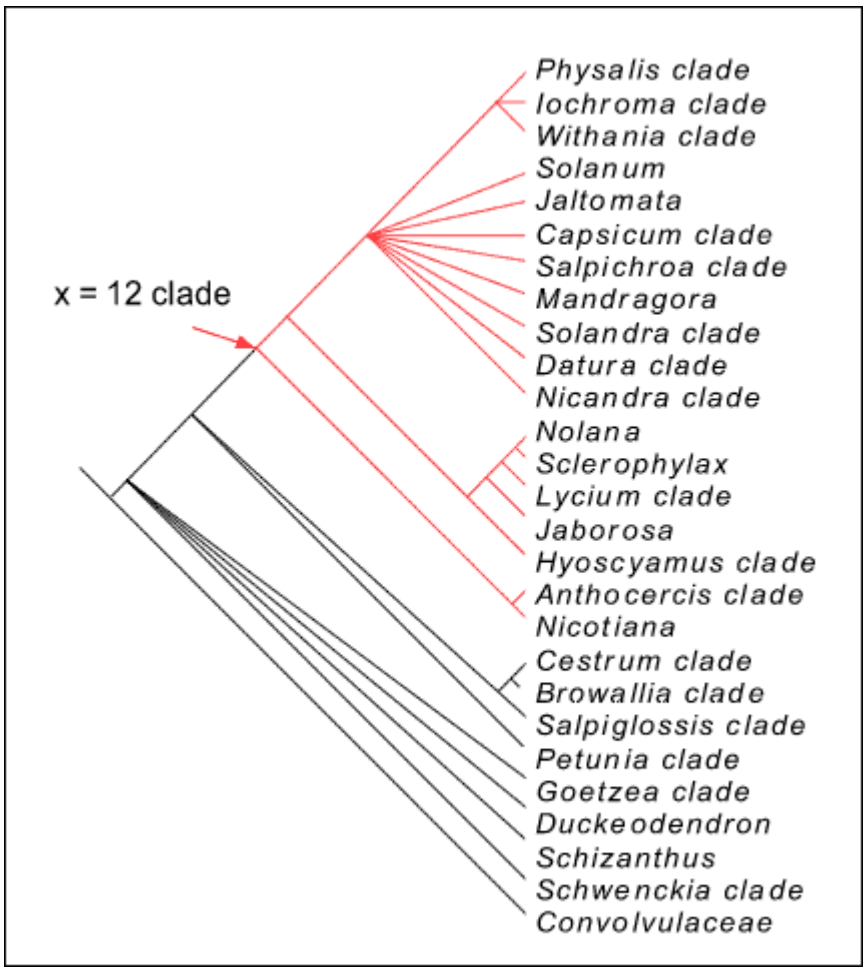


Figura 1: *Principais clados da família Solanaceae. Seta indica o início do clado das espécies que possuem número básico* x*=12. Adaptado de Solanaceae Source.*

1.2 O GÊNERO *CAPSICUM*

As pimentas e pimentões que pertencem ao gênero *Capsicum,* possuem como centro de origem as Américas Central e do Sul, esse gênero é amplamente cultivado nas regiões tropicais e subtropicais e apresenta grande variedade de formas, tamanhos e cores dos frutos (DJIAN-CAPORALINO *et al.*, 2007; QIN *et al.*, 2014). As pimentas são utilizadas na ornamentação, produção de compostos medicinais

e na culinária (MOSCONE et al, 2003; 2007). Devido ao alto teor de nutrientes e a pungência decorrente do acúmulo de capsaicinóides, este gênero é altamente apreciado na culinária e o que leva a um grande valor de mercado, sendo cultivado em larga e pequena escala em diferentes regiões do globo (AZA-GONZÁLEZ, NÚÑEZ-PALENIUS E OCHOA-ALEJO, 2011; KIM *et al.*, 2014). No Brasil, o cultivo de pimentas e pimentões constitui uma importante atividade socioeconômica para o setor agrícola, e ocorre em praticamente todas as regiões do país, descantando-se a produção familiar (REIFSCHNEIDER *et al.*, 2000; HENZ, 2004). Das 43 espécies descritas para o gênero (FIGURA 2), cinco são consideradas domesticadas (*C. annuum* L., *C. baccatum* L., *C. chinense* Jacq., *C. frutescens* L. e *C. pubescens* R E P) e as demais silvestres (MOSCONE *et al.*, 2007; CARRIZO GARCÍA *et al.*, 2013, THE PLANT LIST. 2013). A filogenia do gênero apresenta 11 clados, entre esses pode-se destacar *Capsicum annuum*, *C. chinense* e *C. frutescens* pertencem ao clado Annuum, *C. baccatum* ao clado Baccatum e *C. pubescens* ao clado Pubescens (CARRIZO GARCÍA *et al.*, 2016).

Com relação às características citogenômicas, o gênero *Capsicum* apresenta variação quanto a quantidade de DNA nuclear entre as espécies. Os valores C de DNA foram estimados por MOSCONE *et al.* (2003) em 25 amostras de nove espécies diploides de *Capsicum*, os quais variaram de 3,32 pg em *C. annuum* até 5,77 pg em *C. parvifolium*. Do mesmo modo que são observadas variações no valor C de DNA entre as espécies, os perfis de bandas de heterocromatina também são variáveis. Os primeiros trabalhos de bandeamento envolvendo espécies do gênero *Capsicum*, demonstraram diversidade quanto a presença de bandas terminais (bandas mais intensas ou tênues) e quanto a presença de bandas intersticiais e proximais (MOSCONE *et al.*, 1993; 1996), assim como em trabalhos mais recentes (SCALDAFERRO *et al.*, 2013; MARTINS *et al.*, 2018). A quantidade e localização de sítios de DNAr também uma característica que apresenta variação nas espécies do gênero *Capsicum*, SCALDAFERRO *et al.* (2016) evidenciaram a posição de DNAr por meio de hibridizações com sondas de 35S e 5S. Em seu trabalho, puderam mostrar que *C. baccatum* possui maior quantidade de sítios de hibridização de 35S, enquanto *C. annuum* e *C. chinense* a quantidade foi menor. A literatura também evidencia a presença de poucos sítios, muitas vezes únicos, de DNAr 5S em espécies de *Capsicum* (PARK *et al.*, 1999a; PARK *et al.*, 1999b; SCALDAFERRO, GABRIELE E MOSCONE, 2013; AGUILERA, DEBAT E GRABIELE, 2017).

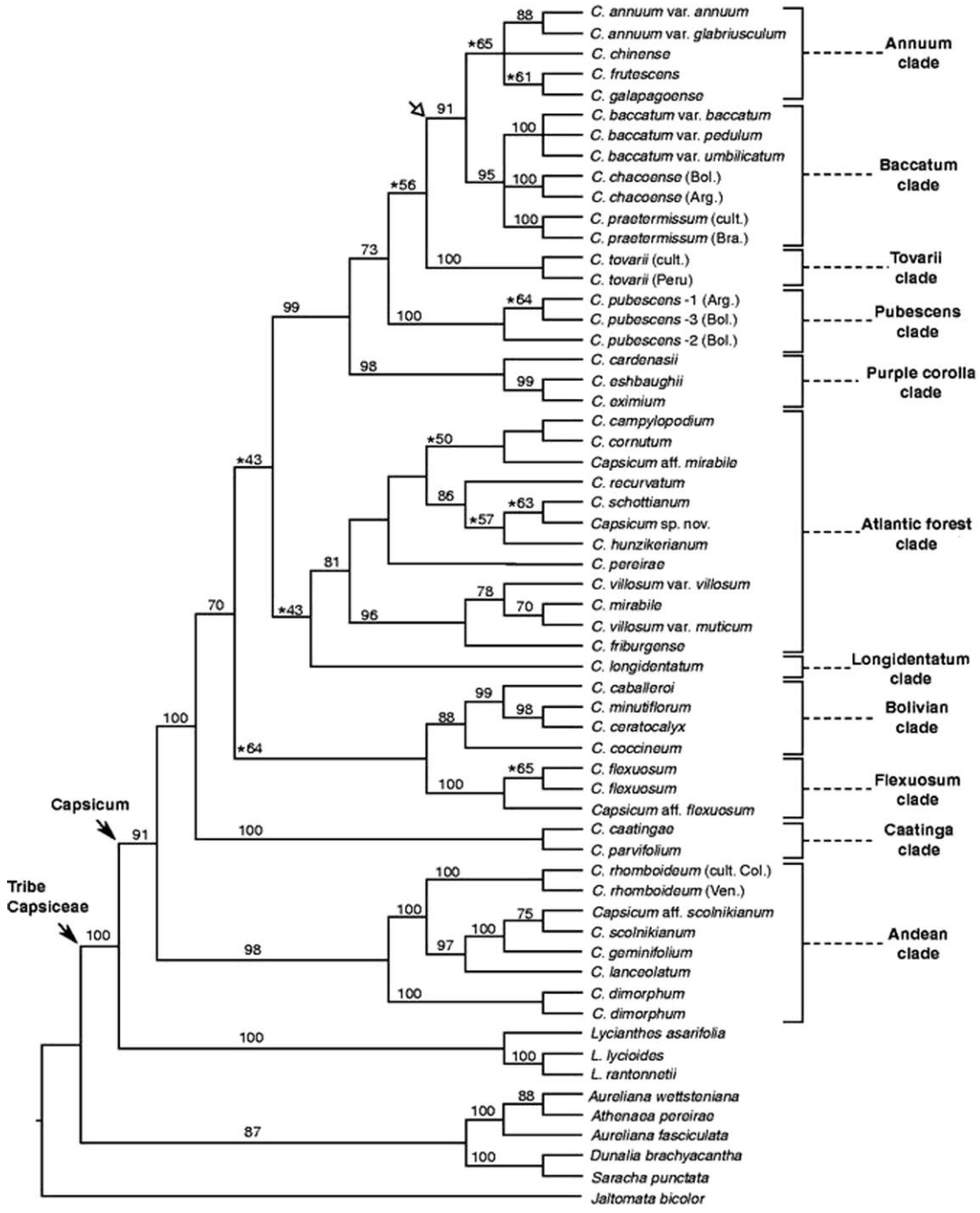


Figura 2: *Filogenia proposta para o gênero* Capsicum *adaptada de Carrizo-Garcia e colaboradores (2016). Árvore obtida por análise de máxima parcimônia com base nos genes* matK, psbA-trnH *e* waxy.

1.3 O GÊNERO *SOLANUM*

Dentro da família Solanaceae, *Solanum* é o gênero mais diversificado taxonomicamente e apresenta cerca de 1.500 espécies distribuídas na América Central e do Sul, Austrália e África, sendo a América do Sul o centro primário de diversidade e endemismo (NEE 1999; KNAPP 2008). No Brasil, de aproximadamente 250 espécies descritas, 100 são endêmicas e ocorrem desde florestas, como a Floresta Atlântica, até regiões áridas como a Caatinga (AGRA 2007). O gênero *Solanum* é diferenciado da maioria dos demais a ele morfologicamente relacionados por apresentar suas anteras com deiscência poricida, característica compartilhada apenas com o gênero *Lycianthes*, cujas flores possuem cálice diferenciado morfologicamente (WEESE & BOHS 2007). Com relação à classificação, vários tratamentos foram propostos para o gênero, a maioria deles não congruentes, como observados nas classificações infragenéricas realizados por SENDTNER (1846), DUNAL (1852), BITTER (1919), SEITHE (1962), D'ARCY (1972; 1991), WHALEN (1984) e NEE (1999). Destes, os mais utilizados até hoje são os de D'ARCY (1972) e NEE (1999). D'ARCY (1972) formaliza no gênero *Solanum* as divisões infragenéricas, reconhecendo sete subgêneros e 52 seções para as espécies do mundo. Entretanto, NEE (1999) no seu tratamento para as espécies do Novo Mundo formaliza três subgêneros, 24 seções e várias séries

Com grande valor econômico, *Solanum* possui espécies amplamente utilizadas na alimentação como *Solanum melongena* L. (berinjela), *S. tuberosum* L. (batata), *S. lycopersicum* L. (tomate), além de apresentar espécies medicinais como *S. paniculatum* L., incluída na Relação Nacional de Plantas Medicinais de interesse ao SUS (RENISUS 2009). Outras espécies do gênero são de interesse na indústria farmacêutica por apresentar compostos esteroidais como a solasodina, um alcaloide esteroidal importante para a síntese de hormônios (SILVA, AGRA & BHATTACHARYYA 2005).

O tomate (*Solanum lycopersicum* L.) pertence a seção *Lycopersicum*, juntamente com mais 12 espécies selvagens (KNAPP, PERALTA, 2016). Seu centro de origem é a América do Sul, na região andina, do Equador ao Chile (FIGURA 3) (LIANG *et al.,* 2017). Como centros de domesticação propostos estão o Peru, como centro primário, e o México como centro secundário (PERALTA; KNAPP; SPOONER, 2005; KNAPP; PERALTA, 2016). Entra as hortaliças, a cultura do tomate é a segunda maior no mundo, atrás apenas da batata (FAOSTAT, 2022), cultivado em quase todos

os estados do Brasil é dividida entre produção para mesa e para indústria, onde a produção para consumo *in natura* é significativamente maior. O tomate comercial teve sua diversidade genética reduzida devido à domesticação fora do seu centro de origem, e pelo melhoramento genético artificial realizado ao longo dos anos, com base em um número limitado de genótipos (SAAVEDRA, SPOOR, HARRIER, 2001; BOITEUX; FONSECA, GONZÁLEZ-ARCOS, 2016). Em consequência da sua importância econômica e nutricional, os recursos genéticos do tomateiro têm sido amplamente explorados em todo o mundo.

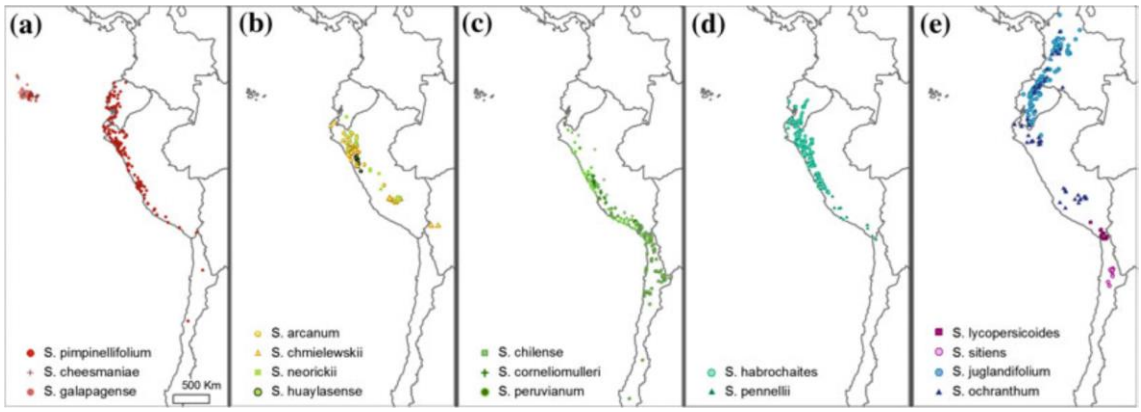


Figura 3: *Distribuição geográfica de espécies de Solanum spp. que compõem o clado Lycopersicum. (a) Espécies com frutos vermelho e laranja. (b-e) Espécies com frutos verdes.*

Solanum lycopersicum possui grande potencial como modelo de estudos genéticos e de processos biológicos, tal qual *Arabidopsis*, devido ao seu genoma relativamente pequeno (2C = ~2 pg), grande quantidade de metabolitos secundários e mapas cromossômicos bem estruturados (marcadores clássicos e moleculares) (TANKSLEY 1993; RICK, YODER 1998). Outra característica que reforça seu potencial como modelo de estudos é ampla riqueza de germoplasma presente nas espécies selvagens que podem ser cruzadas com o tomateiro cultivado (STEVENS, RICK 1986).

1.4 DNAs REPETITIVOS

O DNA nuclear nos vegetais é composto por uma fração não repetitiva (uma ou poucas cópias de sequências codificantes, íntrons, promotores e sequências de DNA regulatórias) e outra fração repetitiva, com centenas ou até milhares de cópias. Essa última fração pode ser organizada em grupos variados de acordo com a

natureza e o modo de repetição das sequências (HESLOP-HARRISON E SCHMIDT, 1998; HESLOP-HARRISON E SCHWARZACHER, 2011; BENNETZEN E WANG, 2014) e chegam a representar, por exemplo, 50 a 60% dos genomas de *S. tuberosum* e *S. lycopersicum*, respectivamente (TOMATO GENOME CONSORTIUM, 2012; MEHRA, GANGWAR, SHANKAR, 2015). Em um contexto geral, a fração repetitiva pode ser pouco representativa em alguns genomas, como 3% em *Utricularia gibba*, Lentibulariaceae (IBARRA-LACLETTE *et al.*, 2013), ou bem mais representativa, como 76,4% em *Capsicum annuum* e 79,6% em *C. chinense* (KIM *et al.*, 2014), ou até uma fração ainda maior, como 90% do genoma de algumas espécies do gênero *Fritillaria* (AMBROZOVA *et al.*, 2011).

1.5 Elementos de transposição

A fração repetitiva pode ser encontrada dispersa nos genomas (elementos de transposição) ou organizadas em *tandem* (DNA ribossômico, microssatélites, minissatélites, satélites, sequências teloméricas e centroméricas e TEs). A fração repetitiva codificante engloba, por exemplo, elementos de transposição, sequências medianamente repetidas de DNAr 35S e 5S, genes que codificam para histonas e proteínas do citoesqueleto (KUBIS, SCHMIDT E HESLOP-HARRISON, 1998; CONTENTO HESLOP-HARRISON E SCHWARZACHER, 2005). A fração repetitiva não codificante pode ser encontrada em *tandem*, representando sequências menores como os microssatélites com 1-6 pb, minissatélites com 10-30 pb e DNAsat com mais de 30 pb (GARRIDO-RAMOS, 2017).

Os elementos transponíveis são divididos em duas classes, I e II (LISCH, 2013). Os de Classe I (FIGURA 4), também chamados retrotransposons, são os mais comuns nos vegetais. A transposição ocorre pelo mecanismo de “copia e cola”, onde o RNAm transcrito é convertido em DNA complementar, por meio da enzima transcriptase reversa, e integrado em outra região do genoma por meio da enzima integrase (KAZAZIAN, 2004; WICKER *et al.*, 2007). Elementos de Classe II – Tranposons (FIGURA 5), são transpostos por meio um mecanismo de “corta e cola”, no qual o elemento é excisado e reintegrado em outra região do genoma utilizando a enzima transposase. Muitos transposons de DNA são flanqueados por sequências terminais repetidas invertidas (TIR – terminal inverted repeat), a transposase reconhece as TIRs, permitindo assim a excisão do transposon e sua reintegração em outra região

do genoma (MUÑOZ-LÓPEZ E GARCÍA-PÉREZ, 2010; LEVIN E MORAN, 2011). O baixo acúmulo de elementos de Classe II, quando comparados aos de Classe I, se deve ao fato que o elemento é excisado de uma região e reintegrado em outra, enquanto os de Classe I são copiados e sua cópia é reintegrada.

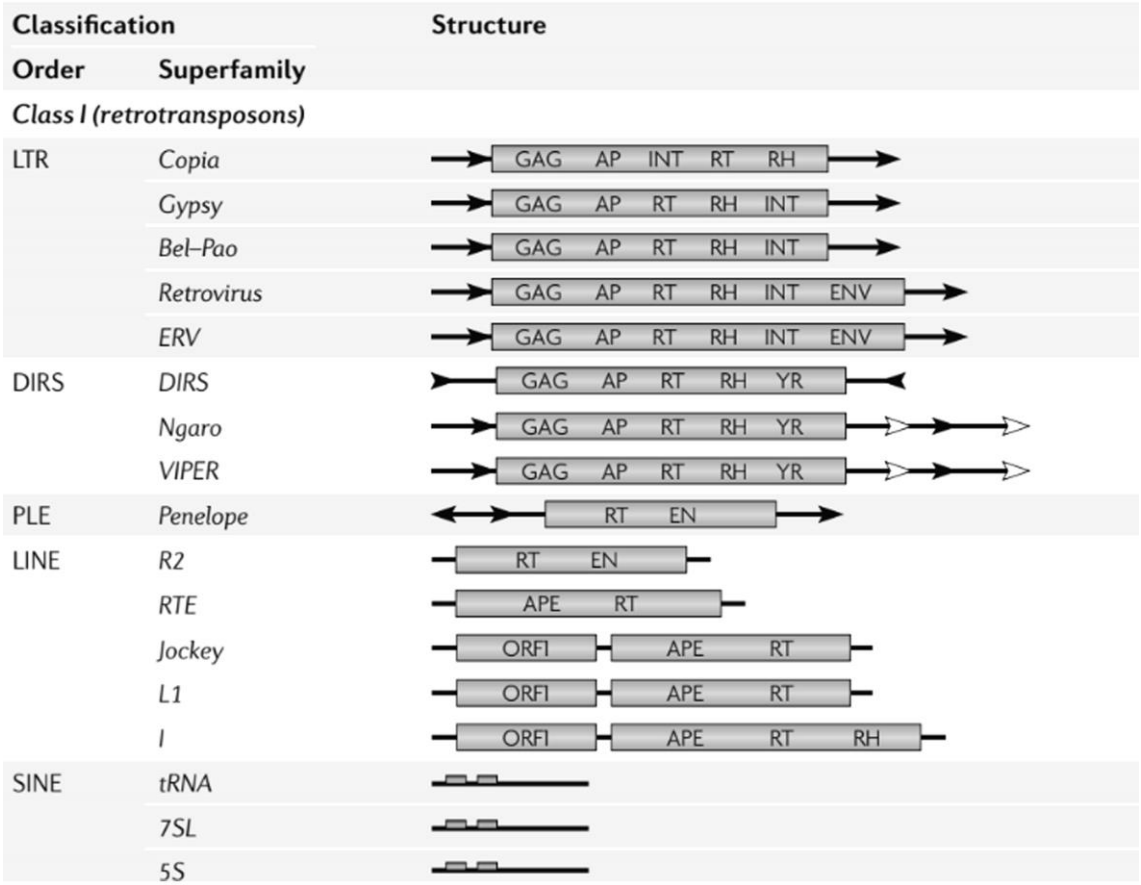


Figura 4: *Sistema de classificação proposto por Wicker e colaboradores (2007) para os elementos de transposição de Classe I. Adaptado de Wicker e colaborados (2007).*

Há também dois outros grupos que utilizam mecanismos diferentes de movimentação e acumulação como os Helitrons, que se espalham pelo genoma por "círculo rolante" via helicase, e os Polintons, que se transpõem por auto síntese, mediado pela polimerase B (KAPITONOV E JURKA, 2008; LISCH, 2013). O sucesso da integração é alcançado em função do mecanismo de reparo na dupla fita de DNA, após a excisão do fragmento. Elementos pertencentes a essa classe diferenciam-se dos da Classe I pela ausência de um RNA intermediário (WICKER *et al.*, 2007; LISCH, 2013). Em geral, os elementos de Classe II se acumulam pouco nos genomas vegetais, contudo podem ser responsáveis por diversas mutações (BENNETZEN E WANG, 2014).

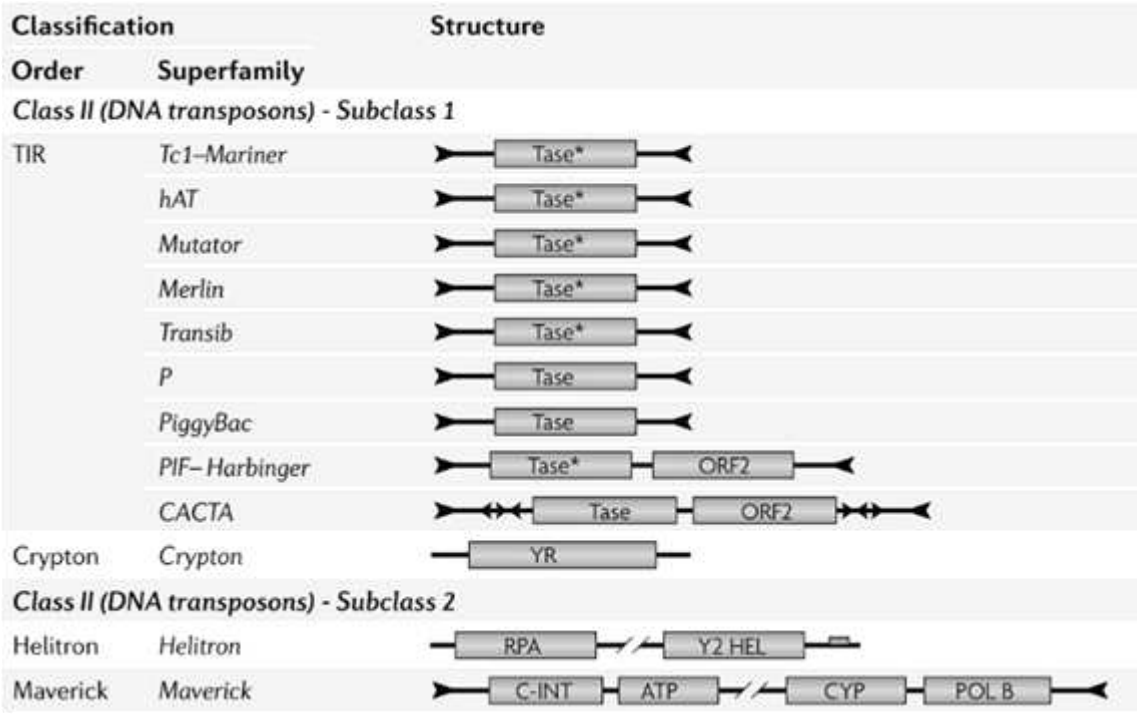


Figura 5: *Sistema de classificação proposto por Wicker e colaboradores (2007) para os elementos de transposição de Classe II. Adaptado de Wicker e colaborados (2007).*

1.6 DNAs SATÉLITE

Os DNA satélite desempenham um importante papel na composição e dinâmica dos genomas vegetais, contribuindo com a estabilidade estrutural e funcional dos cromossomos. Essa classe de DNA repetitivo é composta por uma diversidade de sequências organizadas em *tandem*, e podem alcançar um número muito alto de cópias nos genomas, a depender da espécie (GARRIDO-RAMOS, 2015). Sequências de DNA satélite correspondem à fração repetitiva não codificante com mais de 30pb e são considerados os principais componentes da heterocromatina (GARRIDO-RAMOS, 2015; 2017). A variação na composição dos monômeros tende a ser maior entre as espécies mais distantes filogeneticamente, e o contrário ocorre entre espécies mais próximas, onde os monômeros tendem a ser mais similares (DOSWORTH *et al.,* 2016). No geral, os DNAsat ocupam as regiões de heterocromatina, preferencialmente localizadas nos centrômeros, pericentrômeros e subtelômeros (GARRIDO-RAMOS, 2015; PLOHL et al., 2008; 2014; THAKUR; PACKIARAJ; HENIKOFF, 2021). Apesar de não haver um consenso sobre o tamanho dos monômeros dos DNAsat, os satélites

mais comuns possuem entre 150-180 pb e 300-360 pb, com macrosatélites podendo alcançar poucos kilobases de tamanho. Sequências repetitivas mais curtas, como os microssatélites (*Simple sequences Repeat* - SSR – 2-6 pb) e os minissatélites (10-100 pb), diferenciam-se também em função das regiões cromossômicas em que ocorrem (GARRIDO-RAMOS, 2017; THAKIR; PACKIARAJ; HENIKOFF, 2021).

Os DNAsat foram descritos inicialmente como bandas adicionais de DNA após centrifugação do DNA em gradiente de densidade (HEMLEBEN, 1990). Desde então, *repeats* em *tandem* de diferentes composições, tamanhos e abundâncias vêm sendo identificados (GARRIDO-RAMOS, 2015). Eventos de *crossing-over* desigual são importantes para a formação de satélites e para o aumento na variabilidade dos monômeros, contudo, o *crossing-over* por si só não é o único mecanismo que leva a formação e expansão dos DNA satélites. Amplificações gênicas ou de pedaços de DNA repetitivo codificante, seguido de duplicações e deleções também desempenham papel importante para o surgimento e evolução dessas sequências (SMITH, 1976; MEHROTRA E GOYAL, 2014). A distribuição dos DNAsat nos genomas pode seguir o princípio da distribuição equilocal e equidistante da heterocromatina, onde blocos de heterocromatina tendem a ocupar regiões similares em cromossomos não homólogos, sejam elas pericentroméricas, intersticiais ou subteloméricas. Essa disposição equilocal seria independentemente do tamanho dos cromossomos, contudo, ao contrário do posicionamento equidistante, pois levaria em conta a posição no braço a partir do centrômero (GARRIDO-RAMOS, 2017; GUERRA, 2000).

Com o avanço das plataformas de sequenciamento, montagem, análise de genomas e de técnicas cito-moleculares, tanto a busca por DNAsat quanto a localização física dessas sequências nos cromossomos, vem sendo explorados em diferentes grupos de plantas. Diferentes DNAsat associados a região centromérica foram reportados em batata (GONG *et al*., 2012), assim como outros associados à região pericentromérica foram identificados em *Vicia faba* (ROBLEDILLO *et al.*, 2018), ou ocupando regiões subterminais e intersticiais como em *Allium* (PEŠKA *et al.,* 2019). Esses são apenas alguns trabalhos com foco em DNAsat, mas com a utilização de ferramentas de bioinformática, e novos trabalhos sobre a origem, organização dessas sequências nos genomas e localização dessa fração repetitiva, têm sido cada vez mais explorados em plantas.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Avaliar frações repetitivas em grupos distintos dentro da família Solanaceae, utilizando como organismos modelo as pimentas e os tomates.

2.2 OBJETIVOS ESPECÍFICOS

a) Buscar e caracterizar sequências de DNA satélite nos genomas de espécies de *Capsicum* e *Solanum*;
b) Identificar as sequencias que compõem as regiões centroméricas em espécies de *Solanum*;
c) Selecionar as sequências de melhor qualidade e distribuição no genoma para construção de *primers* e oligos;
d) Mapear as sequências de DNAs repetitivos nos cromossomos das espécies de *Capsicum* e *Solanum*;
e) Comparar a distribuição das sequências nos cromossomos e nas montagens das pseudo-moléculas de *Capsicum* e *Solanum*;

3 REFERÊNCIAS BIBLIOGRÁFICAS

ACOSTA, C. M. *et al.* Karyotype analysis in several South American species of Solanum and Lycianthes rantonnei (Solanaceae). **Taxon**, v. 54, n. 3, p. 713-723, 2005.

ACOSTA, C. M.; GUERRA, M.; MOSCONE, E. A. Karyological relationships among some South American species of Solanum (Solanaceae) based on fluorochrome banding and nuclear DNA amount. **Plant systematics and evolution**, v. 298, n. 8, p. 1547-1556, 2012.

AGRA, M.F. Diversity and Distribution of *Solanum* subgenus *Leptostemonum* in Brazil. In: SPOONER, D.M.; BOHS, L.; GIOVANNONI, J.; OLMSTEAD, R.G.; SHIBATA, D. (orgs.). **Acta Horticulturae - VI International Solanaceae Conference: Genomics Meets Biodiversity**. Madison, Wisconsin, International Society for Horticultural Science, v. 745. p. 31-43, 2007.

AGUILERA, P. M.; DEBAT, H.; GRABIELE, M. An Integrated Physical Map of the Cultivated Hot Chili Pepper, Capsicum baccatum var. Pendulum. **International Journal of Agriculture and Biology**, v. 19, n. 3, 2017.

AMBROŽOVÁ, K. *et al.* Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. **Annals of Botany**, v. 107, n. 2, p. 255-268, 2010.

ARABIDOPSIS GENOME INITIATIVE *et al.* Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. **nature**, v. 408, n. 6814, p. 796, 2000.

AZA-GONZÁLEZ, C.; NÚÑEZ-PALENIUS, H. G.; OCHOA-ALEJO, Neftalí. Molecular biology of capsaicinoid biosynthesis in chili pepper (Capsicum spp.). **Plant cell reports**, v. 30, n. 5, p. 695-706, 2011.

BADR, A.; SF, Khalifa; AI, Aboel-Atta. Chromosomal criteria and taxonomic relationships in the Solanaceae. **Cytologia**, v. 62, n. 2, p. 103-113, 1997.

BENNETZEN, J. L.; WANG, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. **Annual review of plant biology**, v. 65, p. 505-530, 2014.

CARRIZO GARCÍA, C. *et al.* Phylogenetic relationships, diversification and expansion of chili peppers (*Capsicum*, Solanaceae). **Annals of Botany**, v. 118, n. 1, p. 35-51, 2016.

CARRIZO GARCÍA, C. *et al.* Wild Capsicums: identification and in situ analysis of Brazilian species. **Breakthroughs in the genetics and breeding of Capsicum and eggplant. Edited by S. Lanteri, and GL Rotino**, p. 205-213, 2013.

CONTENTO, A.; HESLOP-HARRISON, J. S.; SCHWARZACHER, T. Diversity of a major repetitive DNA sequence in diploid and polyploid Triticeae. **Cytogenetic and Genome Research**, v. 109, n. 1-3, p. 34-42, 2005.

D'ARCY, W.G. Solanaceae II: typification of subdivisions of *Solanum* **Annals of the Missouri Botanical Garden** 59: 262-278, 1972.

D'ARCY, W.G. The Solanaceae since 1976, with a review of its biogeography. In: Solanaceae III – Taxonomy, chemistry, evolution. Ed. Hawkes, J.G.; Lester, R.N.;Nee, M.;Eschad. N. London: Royal Botanic Gardens, 1991. p.75-137.

DE CASTRO NUNES, R. *et al.* Structure and distribution of centromeric retrotransposons at diploid and allotetraploid Coffea centromeric and pericentromeric regions. **Frontiers in Plant Science**, v. 9, p. 175, 2018.

DJIAN-CAPORALINO, C. *et al.* Root-knot nematode (Meloidogyne spp.) Me resistance genes in pepper (*Capsicum annuum* L.) are clustered on the P9 chromosome. **Theoretical and Applied Genetics**, v. 114, n. 3, p. 473-486, 2007.

DODSWORTH, Steven et al. Using genomic repeats for phylogenomics: a case study in wild tomatoes (Solanum section Lycopersicon: Solanaceae). **Biological Journal of the Linnean Society**, v. 117, n. 1, p. 96-105, 2016.

DUNAL, M.F. Solanaceae. In: **Prodromus systematis naturalis regni vegetabilis.** Ed. Candolle, A.P. 1-690. Paris: Victoris Masson, 1852.

FAO - Food and Agriculture Organization of the United Nations. **Statistics: FAOSTAT** Domains/Production/Crops. Disponível em: <http://faostat3.fao.org/faostat-gateway/go/to/download/Q/QC/E>. Acessado em: 28 de agosto de 2022.

GARRIDO-RAMOS, M. A. Satellite DNA in plants: more than just rubbish. **Cytogenetic and genome research**, v. 146, n. 2, p. 153-170, 2015.

GARRIDO-RAMOS, M. Satellite DNA: An evolving topic. **Genes**, v. 8, n. 9, p. 230, 2017.

GONG, Zhiyun et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. **The Plant Cell**, v. 24, n. 9, p. 3559-3574, 2012.

grasses. **Chromosome research**, v. 23, n. 3, p. 571-582, 2015.

GUERRA, M. Patterns of heterochromatin distribution in plant chromosomes. **Genetics and Molecular Biology**, v. 23, n. 4, p. 1029-1041, 2000.

HEMLEBEN, V. **Molekularbiologie der Pflanzen: mit 28 Tabellen**. Fischer, 1990.

HESLOP-HARRISON, J. S. SCHMIDT, T. Plant Nuclear Genome Composition. **Encyclopedia of Life Sciences**, p. 1–8, 2007.

HESLOP-HARRISON, J. S.; SCHMIDT, T. Genomes, genes and junk: the large-scale organization of plant chromosomes. **Trends in Plant Science**, v. 3, n. 5, p. 195-199, 1998.

HESLOP-HARRISON, J. S.; SCHWARZACHER, T. Organisation of the plant genome in chromosomes. **The Plant Journal**, v. 66, n. 1, p. 18-33, 2011.

HUNZIKER, A. T. **Genera Solanacearum: the genera of Solanaceae illustrated, arranged according to a new system**. ARG Gantner, 2001.

IBARRA-LACLETTE, E. *et al.* Architecture and evolution of a minute plant genome. **Nature**, v. 498, n. 7452, p. 94, 2013.

KAPITONOV, V. V.; JURKA, J. A universal classification of eukaryotic transposable elements implemented in Repbase. **Nature Reviews Genetics**, v. 9, n. 5, p. 411, 2008.

KAZAZIAN, H. H. Mobile elements: drivers of genome evolution. **science**, v. 303, n. 5664, p. 1626-1632, 2004.

KIM, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. **Nature genetics**, v. 46, n. 3, p. 270, 2014.

KNAPP, S. Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. **Journal of Experimental Botany**, v. 53, n°. 377, p. 2001-2022, 2002.

KNAPP, S. Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. **Journal of Experimental Botany**, v. 53, n°. 377, p. 2001-2022, 2002.

KNAPP, Sandra. A revision of the Solanum havanense species group and new taxonomic additions to the Geminata clade (Solanum, Solanaceae). **Annals of the Missouri Botanical Garden**, p. 405-458, 2008.

KNAPP, Sandra; PERALTA, Iris Edith. The tomato (Solanum lycopersicum L., Solanaceae) and its botanical relatives. In: **The tomato genome**. Springer, Berlin, Heidelberg, 2016. p. 7-21.

KUBIS, S.; SCHMIDT, T.; HESLOP-HARRISON, J. Repetitive DNA elements as a major component of plant genomes. **Annals of Botany**, v. 82, n. suppl_1, p. 45-55, 1998.

LEVIN, H. L.; MORAN, J. V. Dynamic interactions between transposable elements and their hosts. **Nature Reviews Genetics**, v. 12, n. 9, p. 615, 2011.

LIANG, Sun et al. Origin of the domesticated horticultural species and molecular bases of fruit shape and size changes during the domestication, taking tomato as an example. **Horticultural Plant Journal**, v. 3, n. 3, p. 125-132, 2017.

LISCH, D. How important are transposons for plant evolution? **Nature Reviews Genetics**, v. 14, n. 1, p. 49, 2013.

MARTINS, L. V. *et al.* Heterochromatin distribution and histone modification patterns of H4K5 acetylation and H3S10 phosphorylation in Capsicum L. **Crop Breeding and Applied Biotechnology**, v. 18, n. 2, p. 161-168, 2018.

MEHRA, M.; GANGWAR, I.; SHANKAR, R. A deluge of complex *repeats*: the Solanum genome. **PloS one**, v. 10, n. 8, p. e0133962, 2015.

MEHROTRA, Shweta; GOYAL, Vinod. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. **Genomics, proteomics E bioinformatics**, v. 12, n. 4, p. 164-171, 2014.

MOSCONE, E. A. *et al.* Analysis of nuclear DNA content in *Capsicum* (Solanaceae) by flow cytometry and Feulgen densitometry. **Annals of Botany**, v. 92, n. 1, p. 21-29, 2003.

MOSCONE, E. A. *et al.* Giemsa C-banded karyotypes in *Capsicum* (Solanaceae). **Plant Systematics and Evolution**, v. 186, n. 3-4, p. 213-229, 1993.

MOSCONE, E. A. *et al.* The evolution of chili peppers (*Capsicum*-Solanaceae): a cytogenetic perspective. In: **VI International Solanaceae Conference: Genomics Meets Biodiversity 745**. 2006. p. 137-170.

MUÑOZ-LÓPEZ, M.; GARCÍA-PÉREZ, J. L. DNA transposons: nature and applications in genomics. **Current genomics**, v. 11, n. 2, p. 115-128, 2010.

NEE, M. Synopsis of *Solanum* in the world. In: NEE, M.; SYMON, D.E.; LESTER, R.N.; JESSOP, J.P. (Eds.). **Solanaceae IV: Advances in Biology & Utilization.** Kew: Royal Botanic Gardens. 1999, 285-333.

NEUMANN, P. *et al.* Plant centromeric retrotransposons: a structural and cytogenetic perspective. **Mobile DNA**, v. 2, n. 1, p. 4, 2011.

OLMSTEAD, R.G.; BOHS, L.; MIGID, H.A.; SANTIAGO-VALENTÍN, E.; GARCIA, V.F.N.; COLLIER, S.M. A molecular phylogeny of the Solanaceae. **Taxon** 57 (4): 1159-1181, 2008.

PADILHA, H. K. M. *et al.* Agronomic evaluation and morphological characterization of chili peppers (*Capsicum annuum*, Solanaceae) from Brazil. **Embrapa Clima Temperado-Artigo em periódico indexado (ALICE)**, 2016.

PARK, Y. *et al.* Karyotyping of the chromosomes and physical mapping of the 5S rRNA and 18S-26S rRNA gene families in five different species in Capsicum. **Genes E genetic systems**, v. 74, n. 4, p. 149-157, 1999a.

PARK, Y. *et al.*, Chromosomal localization and sequence variation of 5S rRNA gene in five Capsicum species. **Molecules and cells**, v. 10, n. 1, p. 18-24, 1999b.

PERALTA, Iris E.; KNAPP, Sandra; SPOONER, David M. New species of wild tomatoes (Solanum section Lycopersicon: Solanaceae) from Northern Peru. **Systematic Botany**, v. 30, n. 2, p. 424-434, 2005.

PEŠKA, V. *et al*. Comparative dissection of three giant genomes: *Allium cepa*, *Allium sativum*, and *Allium ursinum*. **International Journal of Molecular Sciences**, v. 20, n. 3, p. 733, 2019.

PLOHL, M. *et al*. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem *repeats* in centromeric (hetero) chromatin. **Gene**, v. 409, n. 1, p. 72- 82, 2008.

PLOHL, M.; MEŠTROVIĆ, N.; MRAVINAC, B. Centromere identity from the DNA point of view. **Chromosoma**, v. 123, n. 4, p. 313-325, 2014.

PLOHL, M.; MEŠTROVIĆ, N.; MRAVINAC, B. Satellite DNA evolution. In: **Repetitive DNA**. Karger Publishers, 2012. p. 126-152.

PRINGLE, C. R. The universal system of virus taxonomy of the International Committee on Virus Taxonomy (ICTV), including new proposals ratified since publication of the Sixth ICTV Report in 1995. **Archives of virology**, v. 143, n. 1, p. 203-210, 1998.

QIN, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. **Proceedings of the National Academy of Sciences**, v. 111, n. 14, p. 5135-5140, 2014.

REIFSCHNEIDER, F.J.B. (Org.). ***Capsicum*: pimentas e pimentões no Brasil**. Brasília: Embrapa Comunicação para Transferência de Tecnologia. Embrapa Hortaliças, 2000. 113p.

RENISUS. Relação Nacional de Plantas Medicinais de Interesse ao SUS. 2009. Disponível em: < http://portal.saude.gov.br> Acesso em: 20 de novembro 2022.

RICK, Charles M.; YODER, John I. Classical and molecular genetics of tomato: highlights and perspectives. **Annual review of genetics**, v. 22, n. 1, p. 281-300, 1988.

ROBLEDILLO, L. Á. *et al*. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. **Scientific Reports**, v. 8, n. 1, p. 1-11, 2018.

SAAVEDRA, G.; SPOOR, W.; HARRIER, L. Molecular markers and genetic base broadening in Lycopersicon spp. **Acta horticulturae**, p. 503-507, 2001.

SCALDAFERRO, M. A. *et al.* FISH and AgNor mapping of the 45S and 5S rRNA genes in wild and cultivated species of *Capsicum* (Solananceae). **Genome**, v. 59, n. 2, p. 95-113, 2016.

SCALDAFERRO, M. A.; GRABIELE, M.; MOSCONE, E. A. Heterochromatin type, amount and distribution in wild species of chili peppers (Capsicum, Solanaceae). **Genetic resources and crop evolution**, v. 60, n. 2, p. 693-709, 2013.

SEITHE, A. Die haarrarten der Gattung *Solanum* L. und ihre taxonomische Verwertung. **Bot. Jahrb. Syst. Pflanzeng.** 81(3): 261-336, 1962.

SENDTNER, O. Solanaceae et Cestrinneae. In: Von Martius, C.F.P. (Ed.). **Flora Brasiliensis** 6 (10): 1-338, 1846.

SILVA, T.M.S.; AGRA, M.F.; BHATTACHARYYA, J. Studies on the alkaloids of *Solanum* of Northeastern Brazil. **Revista Brasileira de Farmacognosia** v. 15, n.4, 292-293, 2005.

SMITH, George P. Evolution of repeated DNA sequences by unequal crossover. **Science**, v. 191, n. 4227, p. 528-535, 1976.

STEHMANN, J. R. et al. Solanaceae in Lista de espécies da flora do Brasil. **Jardim Botânico do Rio de Janeiro. Disponível em:< Disponível em: http://floradobrasil. jbrj. gov. br/jabot/floradobrasil/FB14636>. Acesso em**, v. 18, 2022.

STEVENS, M. Allen; RICK, Ch M. Genetics and breeding. In: **The tomato crop**. Springer, Dordrecht, 1986. p. 35-109.

TANKSLEY, S. Linkage map of tomato (*Lycopersicon esculentum*)(2N= 24). **Gengetic Maps: Locus Maps of Complex Genomes**, 1993.

THAKUR, J.; PACKIARAJ, J.; HENIKOFF, S. Sequence, chromatin and evolution of satellite DNA. **International Journal of Molecular Sciences**, v. 22, n. 9, p. 4309, 2021.

THE PLANT LIST**. Version 1.1. Published on the Internet;** http://www.theplantlist.org/ (acesso em 14 de janeiro de 2019). 2013.

TOMATO GENOME CONSORTIUM *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. **Nature**, v. 485, n. 7400, p. 635, 2012.

WEESE, T.L.; BOHS, L. A three-gene phylogeny of the genus *Solanum* (Solanaceae). **Systematic Botany** 32: 445-463, 2007.

WHALEN MD. Conspectus of species groups in *Solanum* subgenus *Leptostemonum*. **Gentes Herbarum** 12: 179-292, 1984.

WICKER, T. *et al.* A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973, 2007.

WILLARD, H. F.; WAYE, J. S. Hierarchical order in chromosome-specific human alpha satellite DNA. **Trends in Genetics**, v. 3, p. 192-198, 1987.

4 CAPÍTULO 1 – ABUNDANCE OF OF DISTAL REPETITIVE DNA SEQUENCES IN CAPSICUM L. (SOLANACEAE) CHROMOSOMES

O artigo a seguir foi submetido a revista *Genome*

Gmail **Rafael de Assis <rafaeldeassiis@gmail.com>**

gen-2022-0083.R1 - Confirmation of Manuscript Submission

1 mensagem

Genome <onbehalfof@manuscriptcentral.com> 15 de janeiro de 2023 às 19:17

Responder a: genome@cdnsciencepub.com
Para: rafaeldeassiis@gmail.com, leandrosag@uel.br, Romain.guyot@ird.fr, Romain.guyot@gmail.com, andrevanzela@uel.br

Dear André Vanzela and co-authors,

This is an automated confirmation that your manuscript submission has been received.

Manuscript ID: gen-2022-0083.R1
Title: Abundance of distal repetitive DNA sequences in Capsicum L. (Solanaceae) chromosomes
Contributing Authors: de Assis, Rafael; Gonçalves, Leandro S.A.; Guyot, Romain; Vanzela, André Laforga Sr.
Contact author during peer review: Dr. André Vanzela

If there are any concerns with the submission, the editorial office staff will be in contact with the contact author. Otherwise, the manuscript will be sent to the Editor-in-Chief for consideration.

Statements endorsed on submission are listed below my signature block. Please contact the editorial office at genome@cdnsciencepub.com if you believe any of them to be inaccurate or untrue.

We are committed to combatting plagiarism. New manuscript submissions will be processed using Crossref Similarity Check to identify duplication of previously published work (i.e. plagiarism). To find out more about Crossref visit http://www.crossref.org/crosscheck.html

You may view the status of your manuscript at any time by checking your Author Center after logging in to https://mc06.manuscriptcentral.com/genome-pubs.
Please mention the above manuscript ID in all future correspondence with the editorial office.

Thank you for submitting your manuscript to Genome.

Sincerely,

André Vanzela
Editorial Office
Genome

Abundance of distal repetitive DNA sequences in *Capsicum* L. (Solanaceae) chromosomes

Rafael de Assis[1], Leandro Simões Azeredo Gonçalves[2], Romain Guyot[3], André Luis Laforga Vanzela[1]*

[1]*Laboratório de Citogenética e Diversidade Vegetal, Departamento de Biologia Geral, Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina, 86097-570, Paraná, Brazil.*
[2]*Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Londrina, 86057-970, Paraná, Brazil.*
[3]*Institute de Recherche pour le Développement, CIRAD, Université de Montpellier, UMR DIADE, Montpellier, France.*
[4] *Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales 170001, Colombia.*

*Corresponding author: E-mail: andrevanzela@uel.br, ORCID: 0000-0002-2442-2211
Rafael de Asis, ORCID: 0000-0002-4420-4588
Leandro S.A. Gonçalves, ORCID: 0000-0001-9700-9375
Romain Guyot, ORCID: 0000-0002-7016-7485

Running title: Distal repetitive sequences in *Capsicum* chromosomes

Abstract Chili peppers belong to the Solanaceae family and have great commercial value. They are commercialized *in natura* and used as spices and for ornamental and medicinal purposes. Although their genome sequencing has already been published, some points can still be explored regarding satellite DNA sequences. In this work, we exploited the non-coding repetitive fraction, represented by satellite sequences, that tends to accumulate in blocks along chromosomes, especially near the chromosome ends of peppers. The objectives were to characterize and understand the organization and distribution of two satellite DNA sequences in three *Capsicum* genomes and to physically locate them in two others, totaling five species. To do this, bioinformatics tools were used to identify this fraction from three genomes with high-coverage sequencing. We also used molecular cytogenetic methods for the physical localization of these repeats. Data showed two satellites with different origins, i.e., CDR-1 of uncertain origin and CDR-2 from a TAT/Ogre lineage, present in species of different phylogenetic clades. Satellites occupied sub-terminal chromosomal regions, sometimes collocated with rDNA sequences. Our results expand knowledge about the diversity of sub-terminal regions of *Capsicum* chromosomes, especially concerning the presence and size of satellite sites within and between karyotypes.

Keywords: *In situ* hybridization, peppers, satDNA, sub-terminal regions.

INTRODUCTION

Capsicum (Solanaceae) comprises approximately 41 species native to Central and South America, and they are grouped into eleven clades (Djian-Caporalino et al. 2007; Carrizo-García et al. 2016). The most consumed species worldwide belong to three clades, Annuum (*C. annuum* L., *C. chinense* Jacq., *C. frutescens* L.), Baccatum (*C. baccatum* L.), and Pubescens (*C. pubescens* R. & P.). *Capsicum annuum* is the main cultivated species, including paprika, jalapeño, cayenne, and cherry peppers. It has been widely used as a spice because of the pungency derived from an alkaloid group called capsaicinoids and as medicine and ornamental plant (Moscone et al. 2003). Among wild species, *C. chacoense* stands out by the loss of pungency due to a mutation in the *pun2* regulation locus (Stellari et al. 2010). This species is called "sweet pepper" and has been of great commercial interest. The production of *C. annuum* and *C. baccatum* (dedo-de-moça, cumari, and cambuci peppers) has significant importance for family and small-scale agriculture in Brazil (Leite et al. 2016).

Cytogenetic studies in *Capsicum* species have shown considerable diversity in chromosome morphology, distribution profiles of heterochromatin bands, ribosomal DNA (rDNA), and transposable elements (TEs) (Moscone et al. 1993; Park et al. 1999; Scaldaferro et al. 2013; Aguilera et al. 2017; de Assis et al. 2020; Yañez-Santos et al. 2021). Three *Capsicum* species have been sequenced, and the genomes are publicly available at the Pepper Genome Platform (Kim et al. 2014; Qin et al. 2014). These datasets have opened new opportunities to explore karyotype composition using bioinformatics and genomics tools, including the analysis of the satellite DNA (satDNA) portion. Satellite repeats are an important part of the *Capsicum* repeatome (Grabiele et al. 2018). Based on bioinformatic and cytological analyses, Grabiele and co-authors stated that chromosomes of *Capsicum* are rich in a satDNA family derived

from 18S-25S rDNA, in which rDNA units have been amplified and tandemly dispersed, organizing a main GC-rich component of heterochromatin.

Theoretically, satDNA is organized into long arrays, corresponding to a non-coding DNA fraction with more than 30 base pairs (bp) per monomer, and it may be considered the main portion that organizes heterochromatin regions (Garrido-Ramos 2015; 2017). Events involving satellite sequences, such as unequal crossing-over, insertion, recombination, and amplification/reduction, may produce variability in the number and distribution of repeats along chromosomes (Mehrotra and Goyal 2014; Garrido-Ramos 2015; Ribeiro et al. 2020). Satellite sequences can accumulate in important chromosomal regions, such as centromeres (Plohl et al. 2014; Robledillo et al. 2018; Torres and Oliveira 2018), pericentromeres (Vondrak et al. 2020; Cintra et al. 2021), and distal regions (Urdampilleta et al. 2009), contributing to the functional and structural stability of these regions. Distal chromosome regions are hotspots of repetitive families, in addition to telomeres (Baird 2018). The best example may be the distal location of 35S rDNA sites found in most plant groups, including those belonging *Solanum* (Dong et al. 2000), *Capsicum* (Moscone et al. 1995; Scaldaferro et al. 2016), *Nicotiana* (Lim et al. 2000), and *Cestrum* (Vanzela et al. 2017). Distally located satellite sequences were also observed in *Allium*, *Deschampsia*, and *Passiflora* (González et al. 2018; Pamponét et al. 2019; Ahmad et al. 2020). This pattern of distribution can also be found in Solanaceae, such as in *Cestrum* species (Fregonezi et al. 2006; Souza et al. 2021) and *Solanum bulbocastanum* (Tek et al. 2005).

Our main motivation was to test the hypothesis that there is a mega satDNA derived from 18S-25S rDNA, which predominates at the ends of the *Capsicum* chromosomes, as well as to verify whether repetitive sequences of other natures also contribute to the organization of the distal chromosome regions in pepper species. Here,

we define “distal” as those regions or sites located away from centromeres. To that end, we used complete pseudochromosome sequences of *C. annuum*, *C. chinense,* and *C. baccatum* to screen candidates for satDNA sequences, with subsequent physical mapping by *in situ* hybridization and robust bioinformatic analysis. The two most accumulated repetitive satDNA sequences in all three datasets were in distal chromosome regions of five *Capsicum* species, with many signals colocalized with secondary constrictions and 35S rDNA sites. Therefore, we evaluated the relationships between these satDNA sequences and rDNA and transposable element sequences to infer the probable mechanisms associated with repetitive DNA amplification and dispersion at the chromosome ends.

MATERIALS AND METHODS

Plant material

Seeds of *Capsicum annuum* cv. Criollo de Morelos (accession GBUEL145), *C. chinense* (accession GBUEL27), *C. baccatum* (accession GBUEL118), *C. frutescens* (accession GBUEL230), and *C. chacoense* (accession GBUEL276) were sown in 128-cell polystyrene trays containing the substrate Vivatto®. Ten seedlings of each species were grown in the Laboratório de Citogenética e Diversidade Vegetal, Universidade Estadual de Londrina, Brazil.

Genomic analysis

For the genomic analysis, public data were obtained from the Pepper Genome Platform (http://peppergenome.snu.ac.kr/): pseudochromosomes from *Capsicum annuum* v.1.6, *C. chinense* v.1.2, and *C. baccatum* v.1.2.

To identify satellite sequences, the Tandem Repeats Finder (TRF) version 4.09 (Benson 1999) script was used in the three genomic datasets. Each output file was filtered separately using TRF-filter.pl (Guyot unpublished), G-numeric-1.12.35, and filtering commands based on bash scripts of the Linux platform. To identify tandem repeat families in datasets, repeats > 30 bp in length and with copy number > 100 were used as criteria. Repetitive sequences were evaluated by a local blastn version 2.5.0 in three situations: i) TRF output against each dataset to predict cluster formation, ii) TRF output against RepBase (Censor-GIRI - https://www.girinst.org/repbase) to eliminate any sequence associated with transposable elements, and iii) TRF output against a rDNA database with 35S, 5S, and ribosomal intergenic spacer (IGS) sequences. After this evaluation, the sequences were then assumed to be predicted satellites. To group these different presumed satellites, a dotplot was generated, and the sequences that exhibited similarities were grouped into a cluster. From each cluster, a consensus sequence was extracted after alignment in MegaX using Muscle (Tamura et al. 2021). All the consensus sequences obtained were used as a database on a local blast against each genome to define satellite sequences that could be more informative during the *in situ* hybridizations. The sequences that compose the two satellites analyzed here were compared with the described sequences of 5S and 35S rDNA by dotplot. To test the relationship between the putative satDNA sequences and transposable elements, the *Capsicum* genomes were screened by the EDTA pipeline (Ou et al. 2019), aiming to retrieve transposon- and retrotransposon-like sequences. The EDTA output was checked for the presence of CDR-1 (*Capsicum* Distal Repeat 1) and CDR-2 (*Capsicum* Distal Repeat 2), and the sequences that carried any putative satellite were evaluated regarding their lineage. RepeatMasker (Smit 2013) was employed to map the two satellites using default parameters, and then it was used to extract the mapped monomers. As an

identification criterion, monomers sharing > 80% sequence homology and over 80% alignment extension were considered to belong to the same satDNA family (Ruiz-Ruano et al. 2016). The extract monomers were concatenated and re-analyzed with TRF to verify the presence of subtypes. The *C. baccatum* dataset was screened for the presence of any sequence from rDNA that could be related to a satellite sequence using RepeatMasker against the rDNA database described above.

For the two satellites used for *in situ* hybridization and the 35S rDNA, the richness of these sequences along the pseudochromosomes was obtained using RepeatMasker with default parameters. The DensityMap tool (Guizard et al. 2016) was used to obtain the sequence density within the 1 kbp successive window along the pseudochromosomes. The distribution in the assembled chromosomes was plotted using ShinyCircos (Yu et al. 2018).

Identification of the repetitive fraction in the 25S-18S rDNA IGS sequences

IGS sequences of each *Capsicum* genome dataset were analyzed independently. To this end, a local blastn using the default option was run for each genome, with the 25S-18S rDNA IGS of *C. pubescens* (GenBank FJ460247.1) as the database. The output was manually cleaned to select alignments longer than 1000 bp. Sequences were checked using Dotplot. The IGS region containing a repeated stretch was identified and extracted to be used as input in the TRF tool. The TRF output has been used to characterize monomers, as well as to create a weblogo (Crooks et al. 2004), for each dataset.

DNA extraction, PCR, and oligo probes

DNA was isolated from young leaves of each species using 2% cetyltrimethylammonium bromide (CTAB) extraction buffer, according to Doyle and Doyle (1990). Samples were 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. Ethanol-precipitated DNA samples were resuspended in 10 mM Tris–HCl pH 8, and the concentrations were estimated using a NanoDrop 2000 Spectrophotometer (Thermo Scientific).

To amplify the first satDNA (a 179-bp long sequence called CDR-1) via PCR, the specific primers CDR-1-F 5′GGGCGGTTTGGATGGTCAA3′ and CDR-1-R 5′TTGACTAAAATCATGCCCGGAC3′ were used. The second satDNA was named CDR-2 (60 bp in length), and a full-length sequence was synthesized and conjugated with 5' biotin (Thermo Fisher Scientific) (Supplementary Table 1). For the 35S rDNA probe, the p*Ta*71 probe containing a 9-kb EcoRI fragment of 18S+5.8S+26S isolated from *Triticum aestivum* (Gerlach and Bedbrook 1979) was labeled with digoxigenin-11-dUTP via nick translation. For the other probes, the genomic DNA of *C. annuum* was used as the template. A first PCR was conducted using a mix containing 50 mM $MgCl_2$ (1.5 µL), 10 mM dNTP (1 µL), 5 mM primers (2 µL each), ~30 ng template DNA, 1.25 U of Taq polymerase, and ultrapure water to a final volume of 25 µL. Amplicons were used in a second PCR with conjugated nucleotides using 0.2 mM dNTP containing dGTP (25%), dCTP (25%), dATP (25%), dTTP (17.5%), and biotin-dUTP or Cy3-dUTP (7.5%). Standard PCR was used under the following conditions: 94 °C for 2 min, 30 cycles of 94 °C for 40 s, from 54 to 56 °C for 40 s, and 72 °C for 1 min, and a final extension of 72 °C for 10 min. After that, probes were precipitated by centrifugation at 4 °C and resuspended in 10 mM Tris-HCl buffer, pH 8. The reactions

were tested via electrophoresis in an agarose gel at 3 V cm^{-1} and stained with ethidium bromide.

Fluorescent in situ hybridization (FISH)

FISH assays were conducted in five species of *Capsicum* (*C. annuum*, *C. chinense*, *C. chacoense*, *C. baccatum*, and *C. frutescens*) using probes of 35S rDNA and CDR-1 and CDR-2 satDNA families.

The root tips were pretreated with 0.5% colchicine (1 h 30 min) and fixed in ethanol-acetic acid (3:1, v:v). The fixed material was treated in a solution of 2% (w:v) cellulase from *Aspergillus niger* (Sigma Aldrich, 0.3 U/mg) and 20% pectinase from *Aspergillus niger* (Sigma Aldrich, 1 U/mg) and squashed in a drop of 60% acetic acid. After freezing in liquid nitrogen, the coverslips were removed, and the slides were air-dried. Fluorescence *in situ* hybridization was performed as described by Heslop-Harrison et al. (1991), with modifications. For that, the slides received a mix (30 µL) containing a solution 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 SSC buffer at pH 7.0. After probe detection with an avidin-fluorescein isothiocyanate (FITC) conjugate and anti-digoxygenin-rhodamine (anti-DIG), 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).

Images were acquired in grayscale with a Leica DM4500 B microscope coupled with a DFC300FX camera, pseudocolored (blue for DAPI, greenish-yellow for FITC, and red for Cy3 and digoxigenin), and contrasted using GIMP 2.8 Linux.

RESULTS

Characterization of satDNA sequences from Capsicum genomes

The search for satellite sequences using TRF was chosen because we used long scaffolds and complete pseudochromosome assemblies. The TRF output datasets were screened, and sequences such as rDNA were excluded, making a total of ten likely satellites (Table S1). After a search made by blastn, it was possible to obtain an overview of the richness of the putative satellite sequences in each genome. Only the two most repeated monomers in all three datasets were used for FISH: CDR-1, represented by a predicted 179-bp monomer, and CDR-2, with a length of 60 bp. The repetitive profiles of CDR-1 and CDR-2 were evaluated for each genome using the dotplot, which showed the following variation in the genomes: from 0.43 to 0.56% for CDR-1 and from 0.01 to 0.03% for CDR-2 (Table S2), as represented in two contigs of *C. baccatum* (Figure 1 and Figure S1). In addition, to test whether these two satDNAs had any relationships with rDNA sequences, dotplot alignments were performed against 5S and 35S sequences from *Capsicum* and other Solanaceae species. This analysis showed that the satellite sequences are not part of the ribosomal sequences (Figure 2 and Figure S1). In addition, we characterized the repetitive portion present in the IGS sequences from *C. pubescens* (reference) and the other three species. In all cases, the extracted repeated region inserted in the middle of IGS was composed of a microsatellite monomer of 8 bp (Figure S2). The monomers were relatively conserved among species, with 53% GC and 47% AT on average, and the number of copies ranged

from 79 to 205 (Table S3 and Figure S3). We also evaluated whether there was any sequence related to rDNA that could be stated as a satellite DNA sequence. To this end, we mapped the three genomic datasets with an rDNA database using RepeatMasker. The output showed that rDNA sequences mostly accumulated in *C. baccatum* (Table S4), but no satDNA sequences were found.

To test whether the satellite sequences originated from transposable elements, the *Capsicum annuum* genome assembly was screened by the EDTA pipeline. The output was used as a query in blastn searches against the two satellites. CDR-2 was found in a few sequences of putative retrotransposons, while CDR-1 had no hits with the sequences retrieved with EDTA. The sequences that carried CDR-2 were subjected to another round of local blast using a database of protein domains from transposable elements in an attempt to address the sequences to some lineage already described. This blast search identified 25 long elements (from 12 to 15 kb in length), with all regulatory and protein domains representing the TAT/Ogre lineage from the Gypsy LTR retrotransposon superfamily. However, one or both LTRs were lost or mischaracterized. For this reason, the LTRs were not present in the alignment with the dotplot (Figures 3a and c). CDR-2 satDNA monomers (10 copies) appeared after the polygenic chain downstream of integrase as a tandem repeat family of 3' non-coding end typical of TAT/Ogre retrotransposons (Figures 3a and c). The analysis of the CDR-2 monomers within these 25 elements made it possible to identify five groups of sequences representing lineages that share a conserved 30-bp-long fragment with upstream and downstream deletions and 16 substitutions (Figure S4).

The search of CDR-1 monomers in the online blastn showed exhibited a significant similarity with the predicted BTB/POZ domain of the At5g41330 ankyrin (ANK) gene family (Supplementary Figure 5) of *Capsicum annuum* (GenBank

accession number XM_047403318.1). A set of 400 ANK genes of *Capsicum* (Lopez-Ortiz 2020) and other Solanaceae species obtained from NCBI was used to search for similarities with CDR-1, but no matches were found. These results suggest that CRD-1 sequences may have invaded the ANK-containing regions rather than having emerged from them.

Each predicted DNA satellite retrieved showed subtypes. CDR-1 exhibited a greater number of subtypes (Figure S6), possibly because it is more widely distributed along the chromosome tips, compared to the CDR-2 family, which exhibited a more restricted localization to one chromosome pair in each species (Figures S7 and S8). CDR-1 and CDR-2 sequences appeared, in general, in distal regions. CDR-1 accumulated on almost all pseudochromosomes, while CDR-2 accumulated on a few pseudochromosomes. A main peak of CDR-2 on pseudochromosome 5, in addition to two additional peaks on chromosomes 2 and 9 (not revealed in FISH assays), was observed in *C. annuum* (Figure 5). Two peaks were observed in *C. baccatum*, one on pseudochromosome 7 and another on 11, which was not evident after FISH (Supplementary Figure 7). *Capsicum chinense* exhibited only a peak on pseudochromosome 5 (Figure S8). To evaluate whether the peak regions rich in CDR-1 and CDR-2 were in the same pseudochromosomes of rDNA sites, 5S and 35S rDNA were added to the plot (Figures 5 and Figures S7 and S8). The predicted monomers, identified in the first rounds of analysis and used for the design of primers and oligomers, were able to cover all subtypes in both cases. Furthermore, these probes generated reliable marks on chromosomes after FISH assays.

CDR-1 and CDR-2 occur in the terminal chromosome regions

The FISH assays with the CDR-1 probe exhibited a predominance of sub-terminal signals but with a small variation in pairs carrying hybridization signals in both arms. *Capsicum annuum* and *C. frutescens* presented a greater number of signals, with only six distal regions without FISH signals (Figures 4, Figure S9, and Table 1). However, it is important to emphasize that in some of them, the signs were tiny and often difficult to perceive. In addition, to support the bioinformatics data showing that the CDR-1 satellite sequence is not part of 35S rDNA, we performed double-FISH with the corresponding probes. The data show that although the signals were colocalized in 18 distal regions, the FISH signals with the CDR-1 probe were independent in 10 positions, and the signals with the 35S probe were also independent in two positions (Figure 4f-h). FISH with the CDR-2 probe exhibited a distal signal in the short arm of a submetacentric pair in *C. baccatum*, *C. chinense*, and *C. chacoense*, while in *C. annuum* and *C. frutescens*, the signals were distal in the long arm (Figures 4, Figure S10, and Table 1). Double FISH using CDR-2 and 35S rDNA probes in *C. annuum* showed a colocalized signal in one pair, probably chromosome 5, indicating that this chromosome end has an accumulation of CDR-1, CDR-2, and 35S rDNA (Figures S9, S10 and S11). However, colocalization between CDR-2 and 35S rDNA was not observed in *C. baccatum* and *C. chinense*, as shown in the images of Figures S7 and S8.

DISCUSSION

rDNA and satellite sequences are colocalized at the Capsicum chromosome ends

Pseudochromosomes of *Capsicum* species assembled by Qin et al. (2014) and Kim et al. (2014) were used to search for satellite DNA sequences using the TRF tool. This strategy identified two satDNA families, which were compared *in silico* and with *in situ* hybridization against rDNA sequences located at the distal chromosome regions.

Previous studies have reported repetitive DNA richness at the chromosome ends of *Capsicum* species, including heterochromatin and rDNA sites (Moscone et al. 1993; Aguilera et al. 2016; Scaldaferro et al. 2013 and 2016). According to Grabiele et al. (2018), karyotypes of *Capsicum* are rich in a satDNA family derived from 18S-25S rDNA, in which rDNA units have been amplified and tandemly dispersed, organizing the main GC-rich component of heterochromatin. This hypothesis caught our attention because Jo et al. (2011) found few FISH signals (two-four) in the secondary constrictions using a 25S rDNA probe in *C. annuum*, *C. frutescens,* and *C. baccatum*, while the IGS probe tested by Grabiele et al. (2018) revealed a greater number of terminal signals. These authors support the idea of amplification and dispersion of rDNAs, including fragments of these sequences, which should be recognized as the major GC-rich fraction of heterochromatin. However, the assumption of Grabiele and co-authors (2018) leads us to the main question: Are there requirements to define satellite DNA monomers, or can any repeat sequence moderately amplified be named satellite DNA? Comparing this assumption with our findings, we reach another interpretation of the hyper-amplified rDNA sequences.

Sequences of 35S rDNA occur as tandemly arranged repetitive units that vary greatly in copy number between species, from a few to several hundred or even thousands of copies (Garcia et al. 2017). This repeated family is maintained through concerted evolution. However, events of sequence dispersion may occur when concerted evolution is changed. This could explain the wide distribution of rDNA sites at the chromosome ends in *Capsicum*. In our interpretation, given the paralogous nature of rDNA sequences, they could generate other copies, including non-functional fragments or pseudogenes. According to Garcia et al. (2017), many rDNA arrays could carry pseudogenes, and many species could carry variants of rDNA sequences or part of

sequences. The number of rDNA sites may vary significantly in angiosperms (Roa and Guerra 2012), and the number of 35S rDNA sites tends to be higher in large genomes, such as those observed in *Capsicum* (Kwon and Kim 2009). An increase in rDNA sites has been observed even in species with different genome sizes, such as in the holocentric genera *Rhynchospora* and *Eleocharis* of the Cyperaceae family (Vanzela et al. 1998; Da Silva et al. 2010, respectively) and other groups of plants with monocentric chromosomes, including *Alstromeria*, Alstroemeriaceae (Chacón et al. 2012; Ribeiro et al. 2021).

Gains and losses of rDNA sequences may be associated with different mechanisms, such as i) amplification of sequences linked to satDNA sequences, ii) structural rearrangements changing chromosome morphology that culminate in changes in the other sequences, and iii) deletion, amplification, or dispersion of repeats without significant changes in chromosome morphology (see Pedrosa-Harand et al. 2006). This last mechanism could be a good explanation for the multiple rDNA sites in *Capsicum* because few rDNA sequences (or fragments of the cistron, for example) could be moved to heterologous chromosomes. Subsequently, amplifications without large structural chromosome rearrangements could generate the multiple FISH signals reported by Grabiele et al. (2018). In our interpretation, even if rDNA sequences could accumulate at the chromosome ends, often appearing hypercondensed (visually similar to or associated with heterochromatic regions), such regions should still be considered rDNA and not satellite DNA. We support this idea because satDNA CDR-1, with 179 bp in length, appeared on almost all distal chromosome regions and was often colocalized with 35S rDNA sites in the five tested species. However, CDR-1 does not belong to the regular sequence of the rDNA cistron, as shown in our dotplot graphs and FISH assays.

We performed a robust analysis and found no typical sequence of satDNA that could have been derived from rDNA.

Aguilera et al. (2016) reported a study of physical mapping of the 5S rDNA locus in six wild and five cultivated taxa of *Capsicum*, and they showed interstitial/distal FISH signals in the short arm in different chromosome pairs. These data corroborate the report of Kwon and Kim (2009) concerning the 5S rDNA FISH signal in one chromosome pair. Nevertheless, both studies present a variation in relation to terminal-interstitial positioning among species. Roa and Guerra (2015) suggested that most angiosperms bear only one pair carrying 5S rDNA sites and that a greater number of 5S rDNA sites is atypical. According to Aguilera et al. (2016), 5S rDNA sites are colocalized with a CMA^{+} heterochromatic band, as previously described by Scaldaferro et al. (2013).

CDR-1 and CDR-2 belong to different repeating families with distinct origins and fates

The CDR-1 satellite family has a length of 179 bp, with 53% GC bases, while CDR-2 is 60 bp long and has 45% GC bases. Despite these two satellites occupying the sub-terminal regions of the *Capsicum* chromosomes and being colocalized with 35S rDNA, the number of sites was immensely contrasting between these two satDNA families in the five tested species. Satellite DNAs are grouped into families that differ in location, nucleotide sequences, complexities, and monomer length and abundance since they arise from sequences of different evolutionary origins (Garrido-Ramos 2017). For instance, there is evidence that conserved satDNA monomers occupying the centromeres of rice and maize represent a class of functional elements that regulate chromosome dynamics (Cheng et al. 2002). Therefore, in our opinion, the most important consideration would be to reflect on the reasons that make sequences

accumulate more in one region than others, such as those observed in the chromosome ends of *Capsicum* species. It is common to find heterochromatin sequences attached to the nuclear envelope (NE) in eukaryote cells, organizing a gene repression region (Capella and Braun 2020). In addition, it is common for telomeres to position themselves at the periphery of the nucleus, and this localization seems to be important to heterochromatin replication late in the S phase, as well as to maintain the stability of the chromosomes in the nuclei (Oko et al. 2020). However, depending on the DNA family set that organizes the distal regions, we could think of new possibilities or effects of the attachment telomeres-NE in facilitating rearrangements involving the sub-terminal chromosomal portions, such as the amplification and dispersion of repetitive DNA families. CDR-1 and CDR-2 satDNA sequences occupy the sub-terminal regions, and together with 35S rDNA sequences and Ca167TR repeats (Zhou et al. 2019), they could be close to the point of NE attachment. In the same way, the fixation of these sequences in nearby nuclear regions could favor the amplification and equilocal expansion of distal heterochromatin for heterologous chromosomes, carrying parts of 35S rDNA sequences juxtaposed to heterochromatin. This could also explain the silencing of several additional rDNA loci found by Grabiele et al. (2018) in *Capsicum* genomes. In terms of chromosome location, some plant groups have satellite DNA families predominating at the chromosome distal regions, as described for *Passiflora* (Sader et al. 2021), *Vigna* (Ribeiro et al. 2020), *Camellia* (Heitkam et al. 2015), and *Citrus* (He et al. 2020). In the genus *Deschampsia* P. Beauv. (Poaceae), for example, sub-terminal chromosome regions were rich in several satDNA families (González et al. 2020). González and co-authors (2020) suggested that this chromosome region could favor the accumulation of repetitive DNA due to low recombination rates. This could be

a reasonable explanation for the large accumulation of repetitive families at the *Capsicum* chromosome ends.

Apparently, both satellite families had different origins and fates. CDR-1 satDNA, with an uncertain origin, exhibited high amplification and dispersion capacity. Most likely, the association of these satellite sequences in the BTB/ANK gene family occurred by an invasion of CDR-1, since these associations rarely occur and both sequences are common in the distal chromosome regions in *Capsicum* (Lopez-Ortiz et al. 2020). On the other hand, CDR-2 satDNA, which originated from TAT/Ogre retrotransposons, exhibited a reduced amplification rate and was poorly dispersed. The phylogeny suggests that the last common ancestor between *C. annuum* and *C. chinense* existed 4 MYA. It also indicates that these two species shared an ancestor with *C. baccatum* at ~1.74 MYA and that the common ancestor with *C. chacoense* existed at ~2.39 MYA (Carrizo-Garcia et al. 2016). Although both satellite families appeared in five species belonging to three different phylogenetic clades (Carrizo-Garcıa et al. 2016), they occurred in similar chromosomal positions. If we take CDR-2 as an example, the origin from 3' untranslated regions of TAT/Ogre elements could lead us to think that this satDNA could have a high amplification and expansion capacity due to the nature and ability of retrotransposon movement, such as observed in *Coffea* species (Cintra et al. 2021). Furthermore, the 25 TAT/Ogre sequences containing repeats of CDR-2 are non-autonomous due to loss of LTR, which is essential for retrotransposon movement (Wicker 2007). This could explain the limited expansion of these satellites to a chromosomal pair, especially if we compare them with other Athila/Tat and CRM retrotransposons accumulated in the proximal chromosome regions in these species (de Assis et al. 2020). Despite evidence pointing to this origin, we cannot rule out the possibility that CDR-2 also invaded the TAT/Ogre untranslated 3' region.

It was believed that distal regions of *Capsicum* chromosomes were formed by a megasatellite derived from rDNA. However, despite the rDNA richness in the sub-terminal regions of some species, based on our findings using *in situ* (FISH) and *in silico* (ShinyCircos) hybridization and by characterizing the repetitive fraction of 25S-18S-IGS in *Capsicum* (composed of an 8-bp long SSR), we suggest that these regions have satellite sequences colocalized with rDNA but not derived from them. The pool of distinct similar repetitive families (rDNA, satDNA, and TEs) between these five species could be explained by the "library hypothesis", in which sequences are inherited from a common ancestor (Garrido-Ramos, 2015). Although the present study provides new information about these chromosome regions, few *Capsicum* genomes have been sequenced to date. It would be very important for us to know if this abundance of repetitive sequences is maintained in the phylogenetic clades, as well as all the diversity of sequences that make up the distal chromosome regions in other species of this genus.

Supplementary Information - The online version contains supplementary material available at <link>

Supplementary Table 1: Consensus satellite sequences retrieved from *Capsicum* genomes.

Supplementary Table 2: Number and proportion of CDR-1 and CDR-2 satellite units in the *C. annuum*, *C. baccatum*, and *C. chinense* genomes.

Supplementary Table 3: Analysis of repetitive and tandemly accumulated regions downstream of the 25S-18S rDNA IGS. As these are moderately repetitive rDNA sequences, the 25S-18S IGS sequence of each species was randomly chosen from the studied genomes, including *C. pubescens* as a reference. Note that these regions are composed of microsatellites with 8-bp motifs, relatively conserved between *Capsicum* species, with 53% GC and 47% AT on average.

Supplementary Table 4: Number and proportion of rDNA sequence units in the *C. annuum*, *C. baccatum*, and *C. chinense* genomes.

Supplementary Figure 1: Dotplot of CDR-1 and CDR-2 against sequences of rDNA from *O. sativa* and *N. tabacum.* Note that the two satellites do not have any similarity with the rDNA sequences.

Supplementary Figure 2: (A) Dotplot from the stretch of the scaffold containing the putative IGS sequence identified in *Capsicum baccatum*. **(B)** Expanded region with tandem repeats.

Supplementary Figure 3: WebLogo representing DNA motifs that appear tandemly accumulated downstream of IGS regions of 25S-18S rDNA. According to the sizes of motifs (8 bp), they could be considered microsatellites and not satDNA. These regions are also relatively conserved among *Capsicum* species. 25S-18S rDNA of *C. pubescens* was used as a reference.

Supplementary Figure 4: Alignment of the CDR-2 subtypes found within the non-autonomous LTR-RT. The sequences present within the LTR-RT elements were extracted, concatenated, and analyzed with the TRF tool. From the TRF output, the most accumulated sequences were presumed subtypes of the predicted CDR-2 monomer. Only the conserved regions identified in the alignment are shown in the figure.

Supplementary Figure 5: Dotplot of CDR-1 against the predicted gene from the *C. annuum* ANK gene family. Note that the CDR-1 monomer appears amplified in tandem at the beginning of the predicted gene from *C. annuum*.

Supplementary Figure 6: Repetition profile of CDR-1 monomers. (A) Dotplot of the predicted CDR-1 monomer was repeated 10 times against a fragment of chromosome 1 from *Capsicum annuum*. Note the fragmentations and inversions present in the fragment of chromosome 1 from *C. annuum*, which indicate the presence of CDR-1 monomer subtypes. (B) Alignment of the subtypes found in *Capsicum* genomic datasets. Note that even though there are gaps and indels, some streches are well aligned with the predicted monomer.

Supplementary Figure 7: Distribution of 5S rDNA, CDR-2, 35S rDNA, and CDR-1 sequences in pseudochromosomes from *Capsicum baccatum*. The 5S coding sequence exhibited only one peak of accumulation on chromosome 7. The CDR-2 monomer exhibited one peak of accumulation on chromosomes 7 and 11. The 35S rDNA sequence accumulated on pseudochromosomes 2, 3, 4, 6, 7, 10, and 12. All the pseudomolecules exhibited an accumulation of CDR-1 monomers in the distal regions. Note that there was an accumulation of CDR-2 and 5S sequences on pseudochromosome 7.

Supplementary Figure 8: Distribution of 5S rDNA, CDR-2, 35S rDNA, and CDR-1 sequences on pseudochromosomes from *Capsicum chinense*. The 5S coding sequence exhibited only one peak of accumulation on chromosome 7. The CDR-2 monomer exhibited one peak of accumulation on chromosome 5. The 35S rDNA sequence accumulated in two pseudomolecules, chromosomes 2, 5, 6, and 12. All the pseudomolecules exhibited an accumulation of CDR-1 monomers in the distal regions. Note that CDR-2, 5S, and 35S rDNA do not accumulate in the same pseudochromosomes, except for a small interstitial peak of 35S on chromosome 5.

Supplementary Figure 9: FISH using the CDR-1 probe against metaphases and prometaphases of *Capsicum annuum* (A), *C. chacoense* (B), and *C. frutescens* (D), all with $2n$=24. Chromosomes were counterstained with DAPI (blue), and the CDR-1 probe was counterstained with Cy3-11-dUTP (red). Hybridization signals are observed in almost all chromosomes, accumulating mainly in the distal regions. Although

different factors can influence the FISH signal intensity, such as chromosome condensation, the presence of cytoplasm, and image capture conditions, CDR-1 signals seem to be less intense in *C. chinense* (C) than in *C. annuum* (A). The arrowheads in *C. chacoense* (B) indicate the satellite region adjacent to secondary constriction, and the arrows in *C. frutescens* (D) indicate proximal FISH signals. The bar represents 10 μm.

Supplementary Figure 10: FISH using the CDR-2 probe against metaphases and prometaphases of *C. annuum* (A), *C. chinense* (B), *C. chacoense* (C), *C. baccatum* (D), and *C. frutescens* (E), all with 2*n*=24. Chromosomes were counterstained with DAPI (blue), and CDR-2 was labeled with biotin-11-dUTP/avidin–FITC conjugate (green). Note hybridization signals in the distal region of only one pair in all species. The bar represents 10 μm.

Supplementary Figure 11: FISH using 35S rDNA and CDR-2 probes in prometaphase of *Capsicum annuum* (A and B, respectively). On the right, a chromosome pair containing FISH signals co-located with both probes (merged image in the box).

ACKNOWLEDGMENTS

R.A. is grateful to CAPES (Finance Code 001) and CNPq for awarding the scholarships. A.L.L.V. is grateful for financial support from the Brazilian Agency CNPq (processes 407194/2018-5 and 309902/2018-5). The authors also thank ProPPG-UEL, PPG-GBM, FINEP, and Fundação Araucária for other types of support.

ARTICLE INFORMATION

Availability of data and materials

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

AUTHOR INFORMATION

Author ORCIDs

André Luis Laforga Vanzela https://orcid.org/0000-0002-2442-2211

Author Contributions

A.L.L.V. and R.A. conceived the study. R.A. conducted the experiments and analyzed the data. R.A. and R.G. provided bioinformatics support to the team. L.S.A.G. provided the plants and collected the leaves and roots. A.L.L.V. and R.A. interpreted the data and wrote the manuscript. All the authors read and approved the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Funding

This study was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação Araucária, Paraná (FA), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Ethical statement

This article does not contain any studies with human participants or animals performed by any of the authors.

Competing interests
The authors declare no competing interests.

REFERENCES

Aguilera, P.M., Debat, H.J., Scaldaferro, M.A., Marti, D.A., and Grabiele, M. 2016. FISH mapping of the 5S rDNA locus in chili peppers (*Capsicum*-Solanaceae). An Acad Bras Ciênc. **88**(1): 117-125. http://dx.doi.org/10.1590/0001-37652301620140616

Aguilera, P.M., Debat, H.J., and Grabiele, M. 2017. An integrated physical map of the cultivated hot chili pepper, *Capsicum baccatum* var. *pendulum*. Int J Agric Biol. **19**: 465-469. http://dx.doi.org/10.17957/IJAB/15.0303

Baird, D.M. 2018. Telomeres and genomic evolution. Phil Trans R Soc B. **373**: 20160437. http://dx.doi.org/10.1098/rstb.2016.0437

Benson, G. 1999. Tandem Repeats Finder: a program to analyze DNA sequences. Nucl Acids Res. **27**(2): 573-580. https://doi.org/10.1093/nar/27.2.573

Capella, M., and Braun, S. 2020. ESCRTing heterochromatin out of the nuclear periphery. Dev Cell. **53**(1): 3-5. https://doi.org/10.1016/j.devcel.2020.03.013

Carrizo-García, C., Barfuss, M.H., Sehr, E.M., Barboza, G.E., Samuel, R., Moscone, E. A., et al. 2016. Phylogenetic relationships, diversification and expansion of chili peppers (*Capsicum*, Solanaceae). Ann Bot. **118**(1): 35-51. https://doi.org/10.1093/aob/mcw079

Chacón, J., Sousa, A., Baeza, C.M., and Renner, S.S. 2012. Ribosomal DNA distribution and a genus-wide phylogeny reveal patterns of chromosomal evolution

in *Alstroemeria* (Alstroemeriaceae). Am J Bot. **99**(9): 1501-1512. https://doi:10.3732/ajb.1200104

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., et al. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell. **14**: 1691-1704. https://doi.org/10.1105/tpc.003079

Cintra, L.A., Souza, T.B.D., Parteka, L.M., Barreto, L.M., Pereira, L.F.P., Gaeta, M.L., et al. 2022. An 82 bp tandem repeat family typical of 3′ non-coding end of Gypsy/TAT LTR retrotransposons is conserved in *Coffea* spp. pericentromeres. Genome. **65**(3): 137-151. https://doi.org/10.1139/gen-2021-0045

Da Silva, C.R.M., Quintas, C.C., and Vanzela, A.L.L. (2010). Distribution of 45S and 5S rDNA sites in 23 species of *Eleocharis* (Cyperaceae). Genetica. **138**(9): 951-957. https://doi.org/10.1007/s10709-010-9477-5

De Assis, R., Baba, V.Y., Cintra, L.A., Gonçalves, L.S.A., Rodrigues, R., and Vanzela, A.L.L. 2020. Genome relationships and LTR-retrotransposon diversity in three cultivated *Capsicum* L. (Solanaceae) species. BMC Genom. **21**:237. https://doi.org/10.1186/s12864-020-6618-9

Djian-Caporalino, C., Fazari, A., Arguel, M.J., Vernie, T., VandeCasteele, C., Faure, I., et al. 2007. Root-knot nematode (*Meloidogyne* spp.) Me resistance genes in pepper (*Capsicum annuum* L.) are clustered on the P9 chromosome. Theor Appl Genet. **114**(3): 473-486. https://doi.org/10.1007/s00122-006-0447-3

Dong, F., Song, J., Naess, S.K., Helgeson, J. P., Gebhardt, C., and Jiang, J. 2000. Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. Theor Appl Genet. **101**(7): 1001-1007. https://doi.org/10.1007/s001220051573

Doyle, J.J., and Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. Focus. **12**: 13-15.

Fregonezi, J.N., Fernandes, T., Torezan, J.M.D., Vieira, A.O.S., and Vanzela, A.L.L. 2006. Karyotype differentiation of four *Cestrum* species (Solanaceae) based on the physical mapping of repetitive DNA. Genet Mol Biol **29**(1): 97-104. https://doi.org/10.1590/S1415-47572006000100019

Garcia, S., Kovařík, A., Leitch, A.R., and Garnatje, T. 2017. Cytogenetic features of rRNA genes across land plants: analysis of the Plant rDNA database. Plant J. **89**: 1020-1030. https://doi.org/10.1111/tpj.13442

Garrido-Ramos, M.A. 2015. Satellite DNA in plants: more than just rubbish. Cytogenet Genome Res. **146**(2): 153-170. https://doi.org/10.1159/000437008

Garrido-Ramos, M.A. 2017. Satellite DNA: an evolving topic. Genes. **8**(9): 230. https://doi.org/10.3390/genes8090230

Gerlach, W.L., and Bedbrook, J.R. 1979. Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res. **7**(7): 1869-1885. https://doi.org/10.1093/nar/7.7.1869

González, M.L., Chiapella, J.O., and Urdampilleta, J.D. 2018. Characterization of some satellite DNA families in *Deschampsia* antarctica (Poaceae). Polar Biol. **41**(3): 457-468. https://doi.org/10.1007/s00300-017-2205-1

Grabiele, M., Debat, H.J., Scaldaferro, M.A., Aguilera, P.M., Moscone, E.A., Seijo, J.G., et al. 2018. Highly GC-rich heterochromatin in chili peppers (*Capsicum*-Solanaceae): A cytogenetic and molecular characterization. Sci Hortic. **238**: 391-399. https://doi.org/10.1016/j.scienta.2018.04.060

Guizard, S., Piégu, B., and Bigot, Y. 2016. DensityMap: a genome viewer for illustrating the densities of features. BMC Bioinform. **17**(1): 1-6. https://doi.org/10.1186/s12859-016-1055-0

Heslop-Harrison, J.S., Schwarzacherm T., Anamthawat-JoÂnssonm K., Leitch, A.R., Shi, M., Leitch, I.J. 1991. *In situ* hybridization with automated chromosome denaturation. Technique 3: 109-115.

Heitkam, T., Petrasch, S., Zakrzewski, F., Kögler, A., Wenke, T., Wanke, S., Schmidt, T. 2015. Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest Camellia japonica. Chromosome Res. **23**(4), 791-806. https://doi.org/10.1007/s10577-015-9500-x

Jo, S.H., Park, H.M., Kim, S.M., Kim, H.H., Hur, C.G., and Choi, D. 2011. Unraveling the sequence dynamics of the formation of genus-specific satellite DNAs in the family solanaceae. Heredity. **106**(5): 876-885. https://doi.org/10.1038/hdy.2010.131

Kim, S., Park, M., Yeom, S.I., Kim, Y-M., Lee, J.M., Lee, H-A., et al. 2014. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicu*m species. Nat Genet. **46**: 270-278. https://doi.org/10.1038/ng.2877

Kwon, J.K., and Kim, B.D. 2009. Localization of 5S and 25S rRNA genes on somatic and meiotic chromosomes in *Capsicum* species of chili pepper. Mol Cells. **27**: 205. https://doi.org/10.1007/s10059-009-0025-z

Leite, P.S.S., Rodrigues, R., Silva, R.N.O., Pimenta, S., Medeiros, A.M., Bento, C.S.et al. 2016. Molecular and agronomic analysis of intraspecific variability in *Capsicum baccatum* var. *pendulum* accessions. Genet Mol Res. **15**(4): 1-16. http://dx.doi.org/10.4238/gmr.15048482

Lim, K.Y., Matyášek, R., Lichtenstein, C.P., and Leitch, A.R. 2000. Molecular cytogenetic analyses and phylogenetic studies in the *Nicotiana* section Tomentosae. Chromosoma. **109**: 245-258. https://doi.org/10.1007/s004120000074

Lopez-Ortiz, C., Peña-Garcia, Y., Natarajan, P., Bhandari, M., Abburi, V., Dutta, S.K., et al. 2020 The ankyrin repeat gene family in *Capsicum* spp: Genome-wide survey, characterization and gene expression profile. Sci Rep. **10**: 4044. https://doi.org/10.1038/s41598-020-61057-4

Mehrotra, S., and Goyal, V. 2014. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genom Proteom Bioinform. **12**(4): 164-171. https://doi.org/10.1016/j.gpb.2014.07.003

Melo, N.F., and Guerra, M. 2003. Variability of the 5S and 45S rDNA sites in *Passiflora* L. species with distinct base chromosome numbers. Ann Bot. **92**: 309-316. https://doi.org/10.1093/aob/mcg138

Moscone, E.A., Lambrou, M., Hunziker, A.T., and Ehrendorfer, F. 1993. Giemsa C-banded karyotypes in *Capsicum* (Solanaceae). Pl Syst Evol. **186**: 213-229. https://doi.org/10.1007/BF00940799

Moscone, E.A., Loidl, J., Ehrendorfer, F., and Hunziker, A.T. 1995. Analysis of active nucleolus organizing regions in *Capsicum* (Solanaceae) by silver staining. Am J Bot. **82**(2): 276-287. https://doi.org/10.1002/j.1537-2197.1995.tb11495.x

Moscone, E.A., Baranyi, M., Ebert, I., Greilhuber, J., Ehrendorfer, F., and Hunziker, A.T. 2003. Analysis of nuclear DNA content in *Capsicum* (Solanaceae) by flow cytometry and Feulgen densitometry. Ann Bot. **92**(1): 21-29. https://doi.org/10.1093/aob/mcg105

Oliveira, L.C., and Torres, G.A. 2018. Plant centromeres: genetics, epigenetics and evolution. Mol Biol Rep. **45**: 1491-1497. https://doi.org/10.1007/s11033-018-4284-7

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. **20**: 275. https://doi.org/10.1186/s13059-019-1905-y

Pamponét, V.C.C., Souza, M.M., Silva, G.S., Micheli, F., Melo, C.A.F., Oliveira, S.G., et al. 2019. Low coverage sequencing for repetitive DNA analysis in *Passiflora edulis* Sims: citogenomic characterization of transposable elements and satellite DNA. BMC Genom. **20**: 262. https://doi.org/10.1186/s12864-019-5576-6

Park, Y.K., Park, K.C., Park, C.H., and Kim, N-S. 2000. Chromosomal localization and sequence variation of 5S rRNA gene in five *Capsicum* species. Mol Cells. **10**: 18-24. https://doi.org/10.1007/s10059-000-0018-4

Pedrosa-Harand, A., de Almeida, C.C.S., Mosiolek, M., Blair, M.W., Schweizer, D., and Guerra, M. 2006. Extensive ribosomal DNA amplification during Andean common bean (*Phaseolus vulgaris* L.) evolution. Theor Appl Genet. **112**: 924-933. https://doi.org/10.1007/s00122-005-0196-8

Plohl, M., Meštrović, N., and Mravinac, B. 2014. Centromere identity from the DNA point of view. Chromosoma. **123**: 313-325. https://doi.org/10.1007/s00412-014-0462-0

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., et al. 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. PNAS. **111**(14): 5135-5140. **Erro! A referência de hiperlink não é válida.** https://doi.org/10.1073/pnas.1400975111

Ribeiro, T., Nascimento, J., Santos, A., Félix, L.P., and Guerra, M. 2021. Origin and evolution of highly polymorphic rDNA sites in *Alstroemeria longistaminea* (Alstroemeriaceae) and related species. Genome. **64**: 833-845. https://doi.org/10.1139/gen-2020-0159

Roa, F., and Guerra, M. 2012. Distribution of 45S rDNA sites in chromosomes of plants: Structural and evolutionary implications. BMC Evol Biol. **12**: 225. https://doi.org/10.1186/1471-2148-12-225

Robledillo, L.A., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., et al. 2018. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. Sci Rep. **8**: 5838. https://doi.org/10.1038/s41598-018-24196-3

Ruiz-Ruano, F.J., López-León, M.D., Cabrero, J., and Camacho, J.P.M. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep. **6**, 28333. https://doi.org/10.1038/srep28333

Scaldaferro, M.A., Grabiele, M., and Moscone, E.A. 2013. Heterochromatin type, amount and distribution in wild species of chili peppers (*Capsicum*, Solanaceae). Genet Resour Crop Evol. **60**: 693-709. https://doi.org/10.1007/s10722-012-9867-x

Scaldaferro, M.A., da Cruz, M.V.R., Cecchini, N.M., and Moscone, E.A. 2016. FISH and AgNor mapping of the 45S and 5S rRNA genes in wild and cultivated species of *Capsicum* (Solananceae). Genome. **59**(2): 95-113. https://doi.org/10.1139/gen-2015-0099

Smit, A., Hubley, R., and Green, P. 2013. RepeatMasker 4.0. *Seattle, WA: Institute for Systems Biology*. https://www.repeatmasker.org/

Souza, T.B., Parteka, L.M., De Assis, R., and Vanzela, A.L.L. 2022. Diversity of satDNA fraction in *Cestrum*, the Genus with the Largest Genomes within Solanaceae. Mol Biol Rep. online: 1-15. https://doi.org/10.1007/s11033-022-07728-z

Stellari, G.M., Mazourek, M., and Jahn, M.M. 2010. Contrasting modes for loss of pungency between cultivated and wild species of *Capsicum*. Heredity. **104**(5): 460-471. https://doi.org/10.1038/hdy.2009.131

Tek, A.L., Song, J., Macas, J., and Jiang, J. 2005. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. Genetics. **170**(3): 1231-1238. https://doi.org/10.1534/genetics.105.041087

Urdampilleta, J.D., De Souza, A.P., Schneider, D.R., Vanzela, A.L.L., Ferrucci, M.S., and Martins, E.R.F. 2009. Molecular and cytogenetic characterization of an AT-rich satellite DNA family in *Urvillea chacoensis* Hunz. (Paullinieae, Sapindaceae). Genetica. **136**: 171-177. https://doi.org/10.1007/s10709-008-9332-0

Vanzela, A.L.L., de Paula, A.A., Quintas, C.C., Fernandes, T., Baldissera, J.N.C., and de Souza, T.B. 2017. *Cestrum strigilatum* (Ruiz & Pavon, 1799) B chromosome shares repetitive DNA sequences with A chromosomes of different *Cestrum* (Linnaeus, 1753) species. Comp Cytogenet. **11**(3): 511-524. https://doi.org/10.3897/CompCytogen.v11i3.13418

Vanzela, A.L.L., Cuadrado, A., Jouve, N., Luceño, M., and Guerra, M. 1998. Multiple locations of the rDNA sites in holocentric chromosomes of *Rhynchospora* (Cyperaceae). Chromosome Res. **6**: 345-350. https://doi.org/10.1023/A:1009279912631

Vondrak, T., Robledillo, L.A., Novák, P., Koblížková, A., Neumann, P., and Macas, J. 2020. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. Plant J 101(2): 484-500. https://doi.org/10.1111/tpj.14546

Yañez-Santos, A.M., Paz, R.C., Paz-Sepúlveda, P.B., and Urdampilleta, J.D. 2021. Full-length LTR retroelements in *Capsicum annuum* revealed a few species-specific family bursts with insertional preferences. Chromosome Res. **29**: 261-284. https://doi.org/10.1007/s10577-021-09663-4

Yu, Y., Ouyang, Y., and Yao, W. 2018. shinyCircos: an R/Shiny application for interactive creation of Circos plot. Bioinform .**34**(7): 1229-1231. https://doi.org/10.1093/bioinformatics/btx763

Table 1. Distribution of CDR-1 and CDR-2 satDNA families in the *Capsicum* karyotypes.

Species	CDR-1 FISH signals				CDR-2 FISH signals
	both ends	one of the ends	interstitial*	no signals	
C. annuum	7 pairs	4 pairs	4 pairs	1 pair	1 sm pair
C. baccatum	6 pairs	5 pairs	1 pair	1 pair	1 sm pair
C. chinense	6 pairs	5 pairs	4 pairs	1 pair	1 sm pair
C. chacoense	6 pairs	5 pairs	3 pairs	1 pair	1 sm pair
C. frutescens	8 pairs	3 pairs	2 pairs	1 pair	1 sm pair

sm = submetacentric pair; * = point out for pairs that present interstitial FISH signals independent of other signals

IMAGES

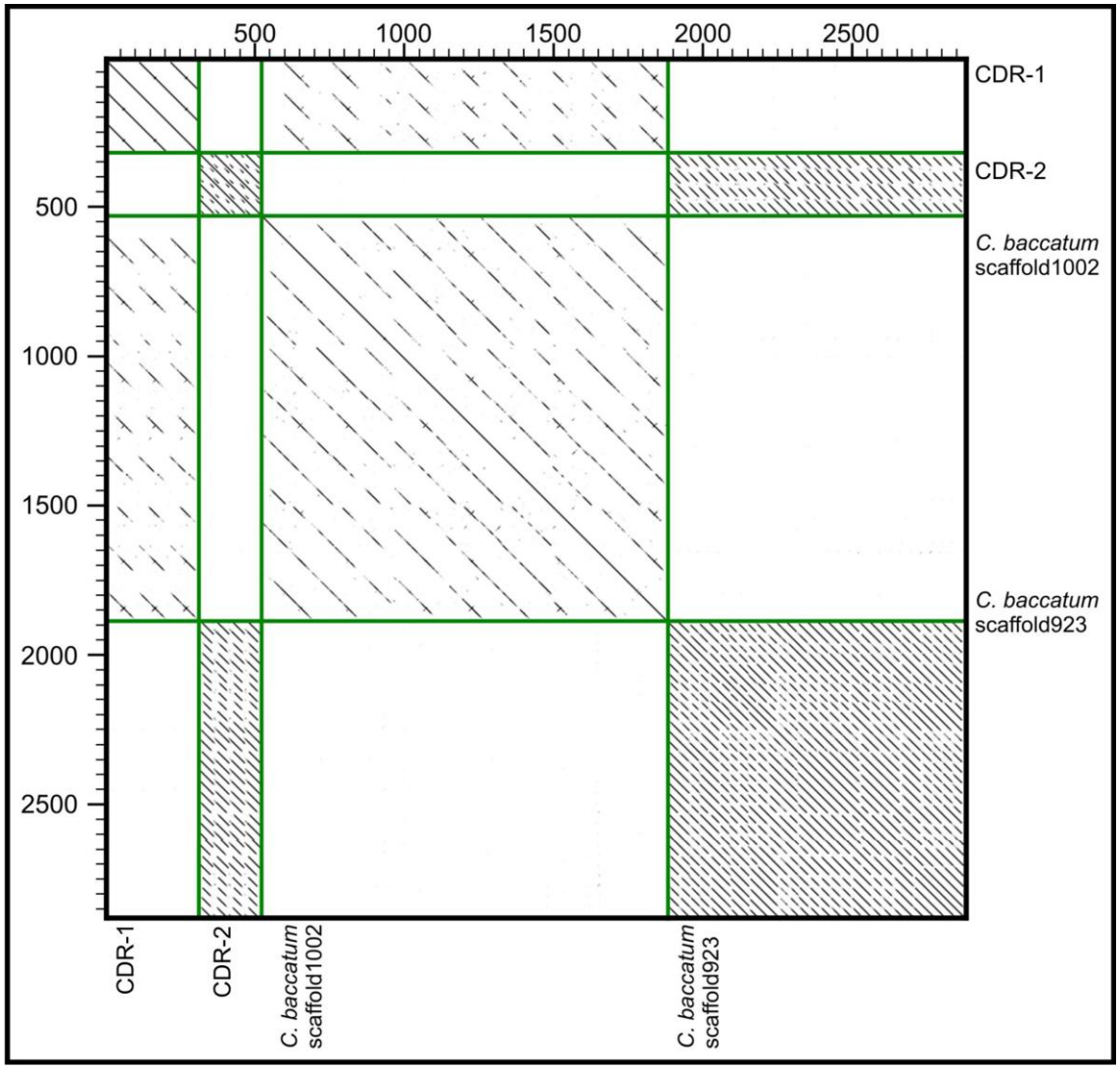


Figure 1 Dotplot of distinct contigs from *C. baccatum* containing repeats of CDR-1 and CDR-2 satellites. Repetitive profile of the satellites analyzed; for better visualization, only a fraction of the scaffold was used. The differences that can be observed in the repeat profile between the two satDNAs are due to monomer length. Note that the two satellite sequences are not interspaced in the contigs. Additionally, CDR-1 appears to have more degeneration along the repeats, while CDR-2 seems to be more conserved

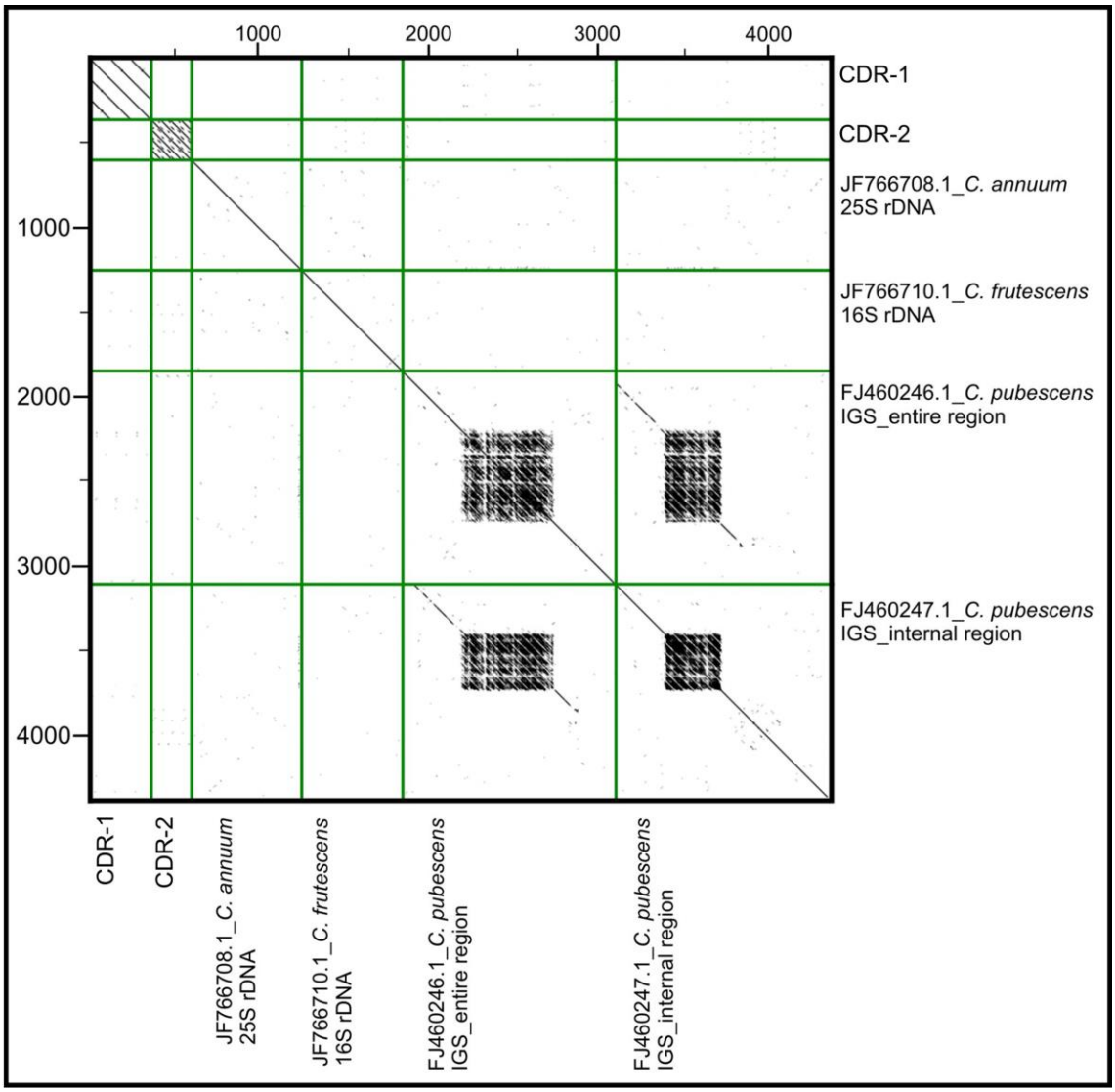


Figure 2: Dotplot of CDR-1 and CDR-2 against sequences of rDNA from *Capsicum* species. Note that there was no similarity between the satellites and the rDNA sequences, indicating that the satellites are not part of the cistron.

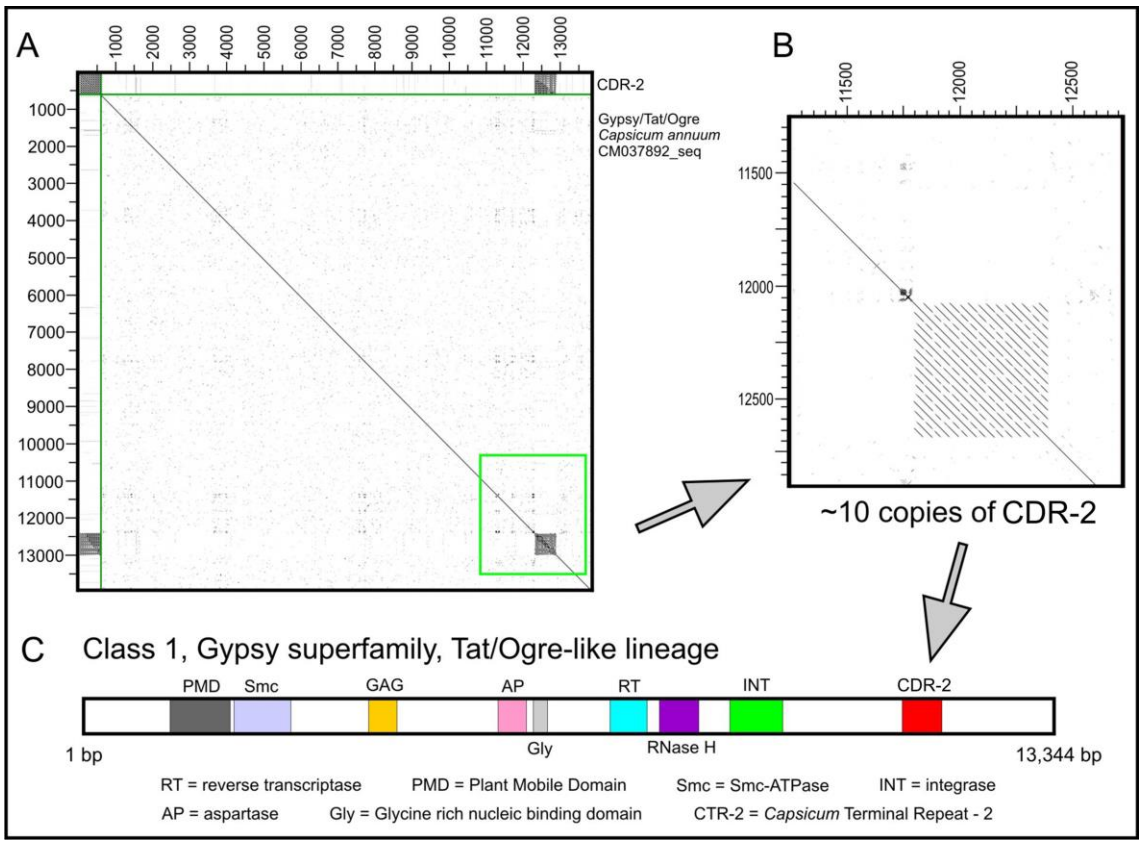


Figure 3: The presence of the CDR-2 monomer in the retrotransposon sequences. (A) Dotplot of the CDR-2 monomer repeated 10 times against one sequence that carries the monomer. (B) A zoomed-in box highlighting the number of copies of CDR-2 present within the LTR sequence; there are at least 10 copies of the monomer. (C) The organization of the LTR sequence that carries the CDR-2 monomer. The sequence belongs to the Gypsy superfamily, and by alignment of the RT domain, it was classified as belonging to the TAT/Ogre lineage. Note that even though the retrotransposon sequence belongs to the TAT/Ogre lineage, it lost both LTRs.

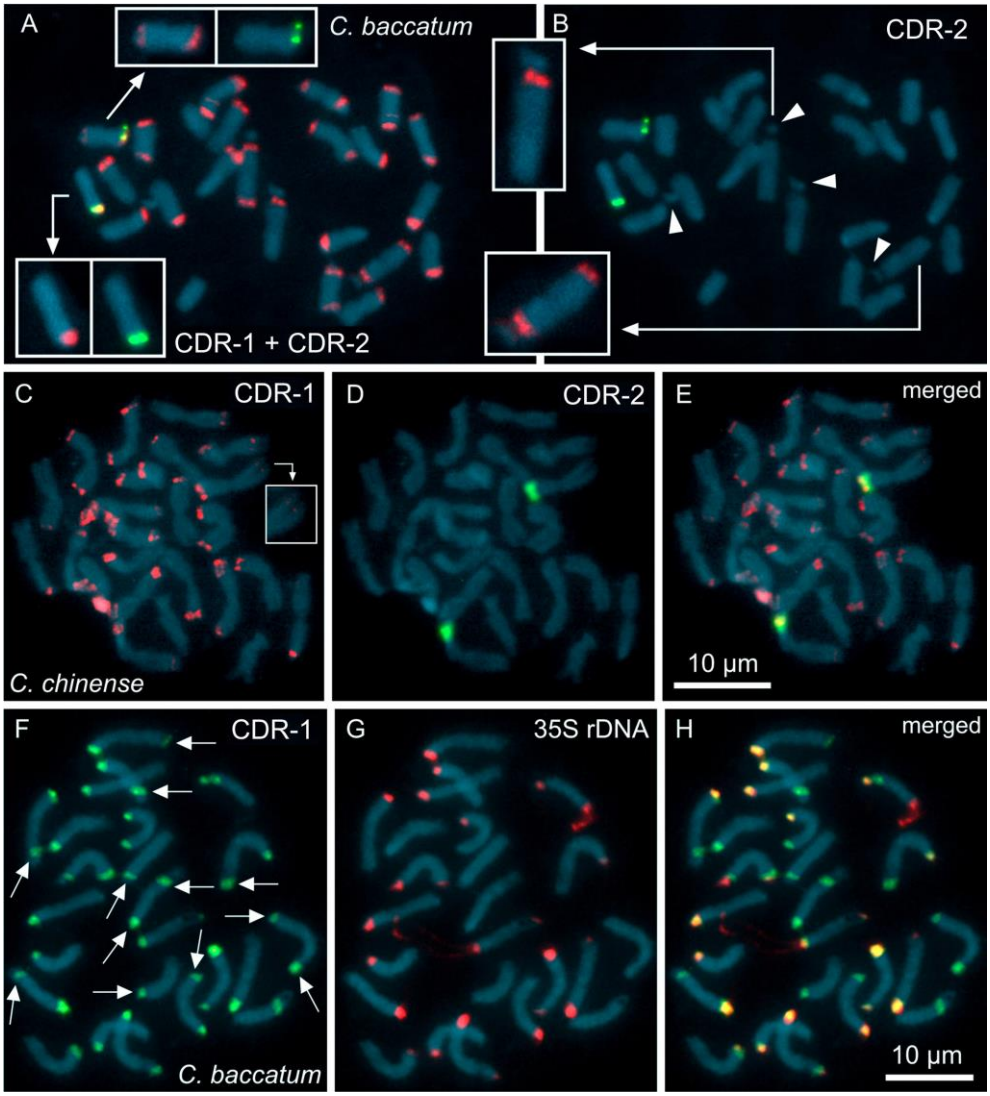


Figure 4: FISH assay using CDR-1, CDR-2, and 35S rDNA probes against metaphases and prometaphases of *Capsicum baccatum* and *C. chinense*. The sample was counterstained with DAPI (blue), and probes were counterstained with Cy3 (red) and avidin–FITC conjugate (green). (A and B) Double FISH in *C. baccatum* with CDR-1 (red) and CDR-2 (green) probes; boxes indicate chromosomes that have a colocalization of the satellites. Observe that CDR-1 localizes immediately below the secondary constriction (arrowheads). (C) CDR-1 and (D) CDR-2 probes in *C. chinense*. (E) Merged images in *C. chinense* highlighting the colocalization of the two satellites. (F) FISH in *C. baccatum* with the CDR-1 probe; the arrows show some signals that were not colocalized with the 35S rDNA probe. (G) FISH in *C. baccatum* with 35S rDNA. (H) Merged images of CDR-1 and 35S rDNA (F and G). A yellowish signal indicates the colocalization of signals. The bar represents 10 µm.

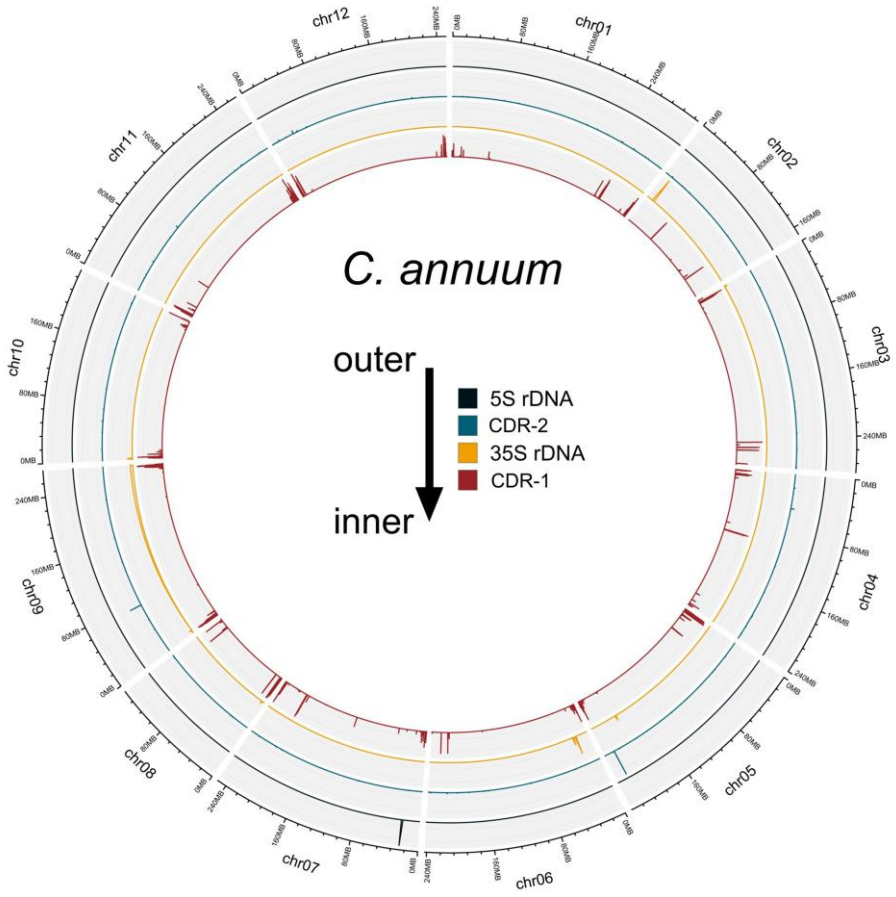


Figure 5: Distribution of 5S rDNA, CDR-2, 35S rDNA, and CDR-1 sequences in pseudochromosomes from *Capsicum annuum*. The 5S coding sequence exhibited only one peak of accumulation on chromosome 7. The CDR-2 monomer exhibited two accumulation peaks, the higher one on chromosome 5 and the minor one on chromosome 9. Sequences of 35S rDNA appeared more accumulated in pseudochromosomes 2 and 6, but small peaks can be seen in the sub-terminal regions of chromosomes 8, 9, and 10 and more interstitial in chromosome 5. All the pseudomolecules exhibited accumulation of CDR-1 monomers in the sub-terminal regions, but with some variation in their concentration among pseudochromosomes.

SUPPLEMENTARY MATERIAL

Abundance of distal repetitive DNA sequences in *Capsicum* L. (Solanaceae) chromosomes

Rafael de Assis[1], Leandro Simões Azeredo Gonçalves[2], Romain Guyot[3], André Luis Laforga Vanzela[1]*

[1]*Laboratório de Citogenética e Diversidade Vegetal, Departamento de Biologia Geral, Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina, 86097-570, Paraná, Brazil.*
[2]*Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Londrina, 86057-970, Paraná, Brazil.*
[3]*Institute de Recherche pour le Développement, UMR DIADE, Montpellier, France.*

*Correspondent author: E-mail: andrevanzela@uel.br, ORCID: 0000-0002-2442-2211
Rafael de Asis, ORCID: 0000-0002-4420-4588
Leandro S.A. Gonçalves, ORCID: 0000-0001-9700-9375
Romain Guyot, ORCID: 0000-0002-7016-7485

Supplementary Table 1: Consensus satellite sequences retrieved from *Capsicum* genomes.

Names	Sizes	Sequences
CDR-1	179 bp	CCTAGTATGGGCCATTGGGGTGGGCGGGGCGGTTTGGATGGTCAA ACAGGCGAAACGGCACAAATGGGCCATTTTCGGGCCAAATCAGT GTGCTATAGCCCACAATTTTTAGGGTGGGCCCAGGGTCCGGGCAT GATTTTAGTCAAAAATTGGTCGGGCTACCACAGGCAGGCTGGGGA
CDR-2	60 bp	CAGACATCTTATTTCGTCTTCATAGGGCGCCATCCCCTAGTTGACA TCTTATCTCGTCTT
SAT-101	60 bp	CACATTCAAAACAAAAAAATCCCCTTTAAAGTCCCTACACTTAAC CATTTTCAACCAAC
SAT-103	60 bp	GCTGACGACCTTCGTGATAGGGCATCACAAATATGGTAAGCTACC ACGAATGTCGTCAC
SAT-104	60 bp	ACTTTAAATAGAGTGATGGTCCAACGTGTCTCATTGAAGGGCGGA CGACCCAATTTAAA
SAT-106	42 bp	TCATTCCCATTGATTCCTATAAGACGTACCTTTTAGTCGAG
SAT-107	38 bp	CCGTACCGGCGTGGTACCCGATCCAACATATACATAT
SAT-108	41 bp	TGACTTTCAATTTCATCACCCTATTCTTCAGGCGGGCTCC
SAT-109	42 bp	GGAATCAAAGGGAATGACTCGACTAAAAGGTACGTCTTATA
SAT-110	38 bp	TTTTTCAGGAGGGTCCTGAACAGAAAGTAAAATCCCA

Supplementary Table 2: Number and proportion of units of CDR-1 and CDR-2 satellites in the genomes of *C. annuum*, *C. baccatum* and *C. chinense*.

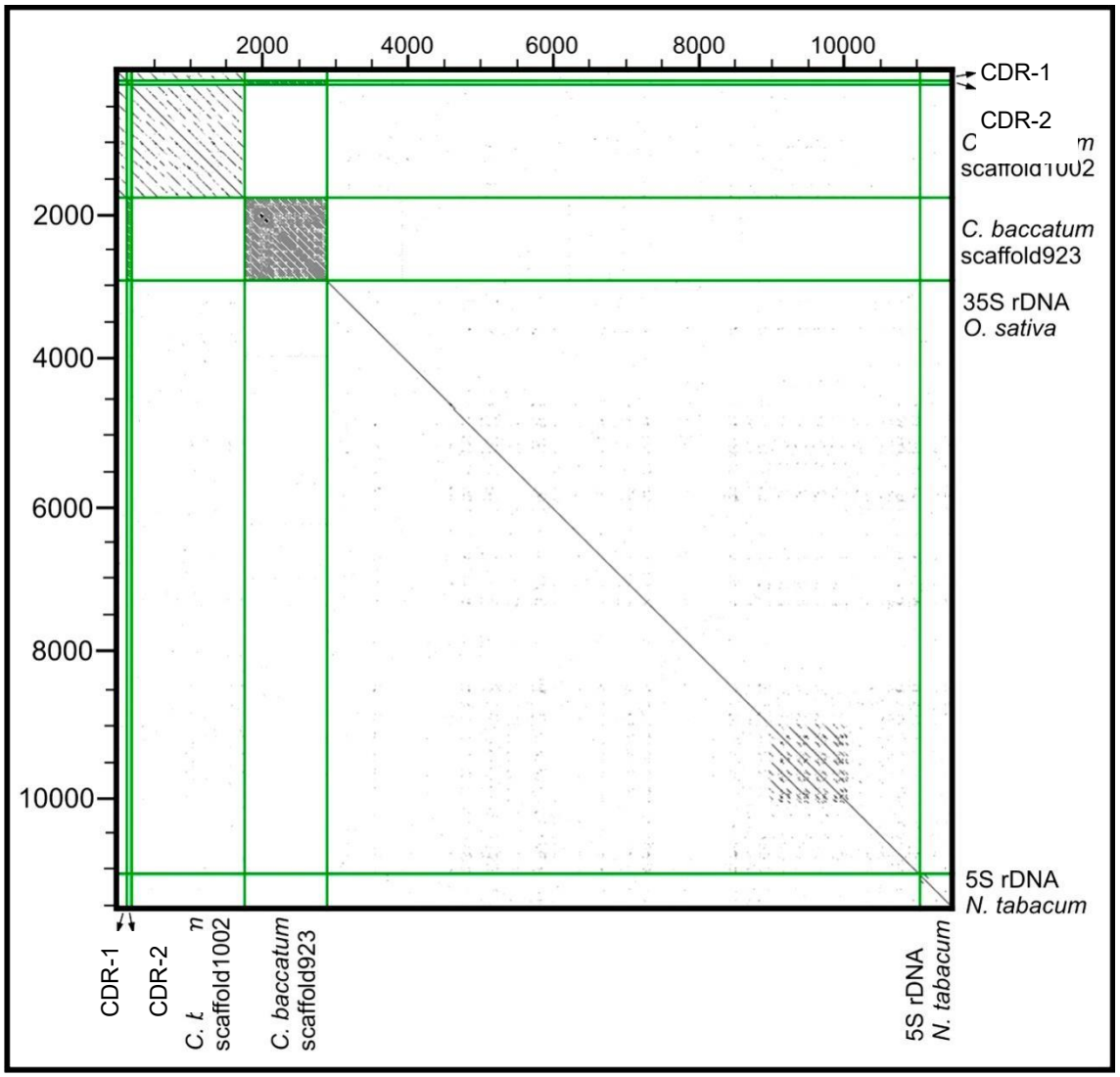
Names	Sizes	*C. annuum*		*C. baccatum*		*C. chinense*	
		Nº monomers	%	Nº monomers	%	Nº monomers	%
CDR-1	179 bp	121472	0.56%	128783	0.43%	129239	0.46%
CDR-2	60 bp	10318	0.01%	25999	0.03%	12791	0.02%

Supplementary Table 3: Analysis of repetitive regions and tandemly accumulated downstream of IGS regions of 25S-18S rDNA. As these are moderately repetitive rDNA sequences, the 25S-18S-IGS sequence of each species was randomly chosen from the studied genomes, including *C. pubescens* as reference. Note that these regions are composed of microsatellites with motifs of 8 bp length, relatively conserved between *Capsicum* species, with on average 53% GC and 47% AT.

Species	Predominant monomer	base pairs	Copy Number	Length	ACGTcount (%)				(%) bp matches
					G	C	A	T	
C. pubescens	GCACCATG	8	127	~1020	26	33	32	9	76
C. annuum	GCACCATG	8	116	~930	21	26	28	24	78
C. chinense	GCACCATG	8	79	~630	26	33	31	10	76
C. baccatum	GCACCATG	8	205	~1640	27	26	22	25	78

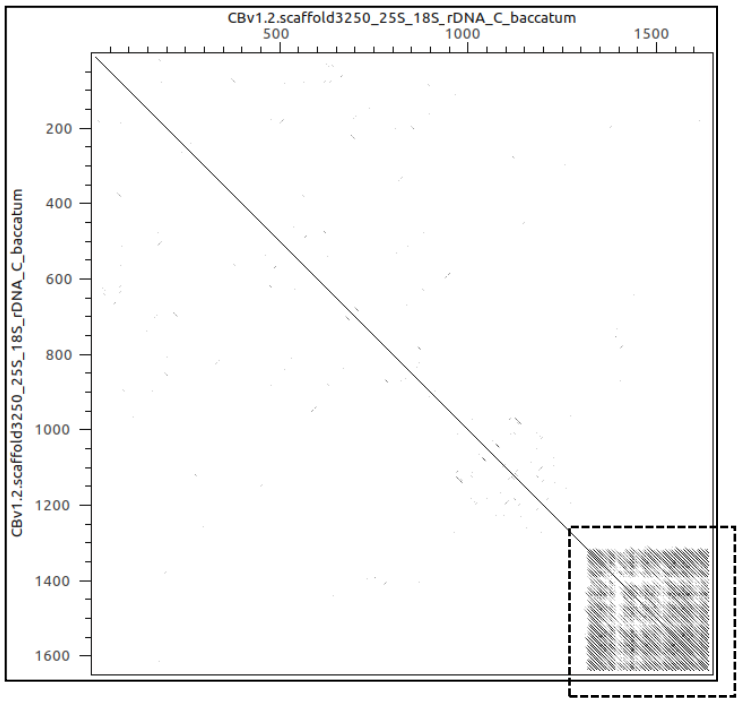
Supplementary Table 4: Number and proportion of units of rDNA sequences in the genomes of *C. annuum*, *C. baccatum* and *C. chinense*.

Names	*C. annuum*		*C. baccatum*		*C. chinense*	
	Nº monomers	%	Nº monomers	%	Nº monomers	%
rDNA	6961	0.06 %	23647	0.55%	9620	0.05 %

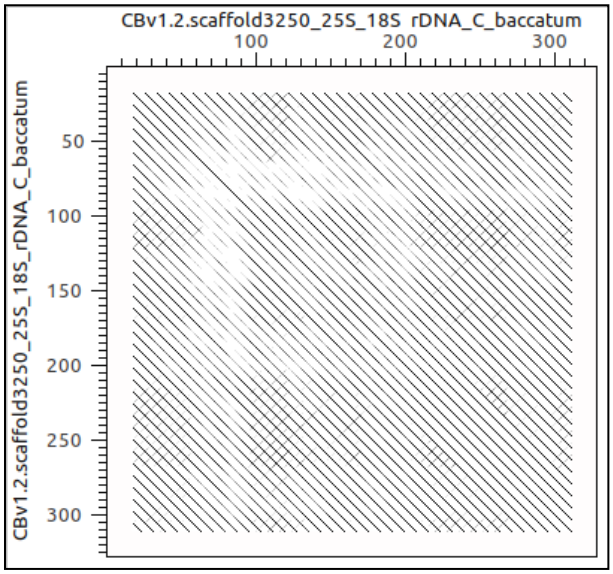


Supplementary Figure 1: Dotplot of CDR-1 and CDR-2 against sequences of rDNA from *O. sativa* and *N. tabacum.* Note that the two satellites don't have any similarity with the rDNA sequences.

(A)

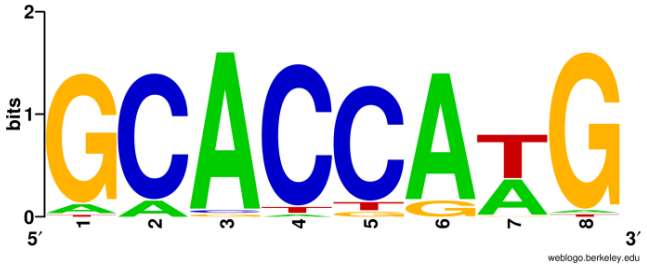


(B)

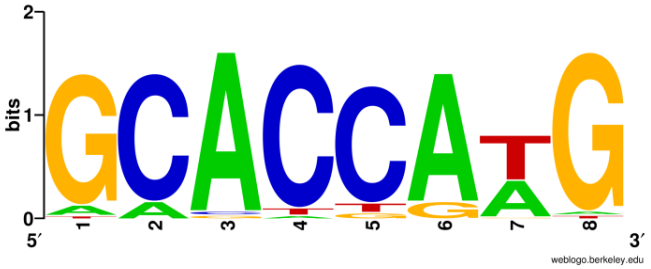


Supplementary Figure 2: (A) Dotplot from the stretch of the scaffold containing the putative IGS sequence identified in *Capsicum baccatum*. **(B)** Expanded region with the tandem repeats.

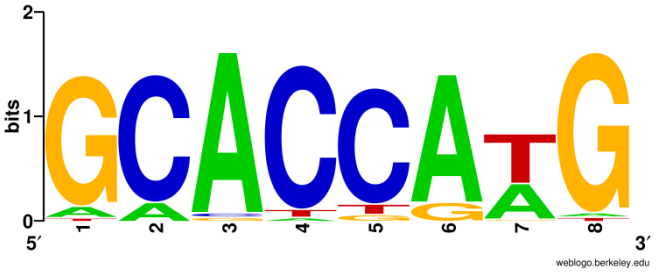
C. pubescens



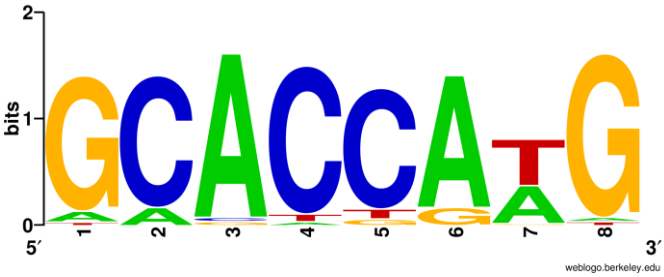
C. annuum



C. chinense



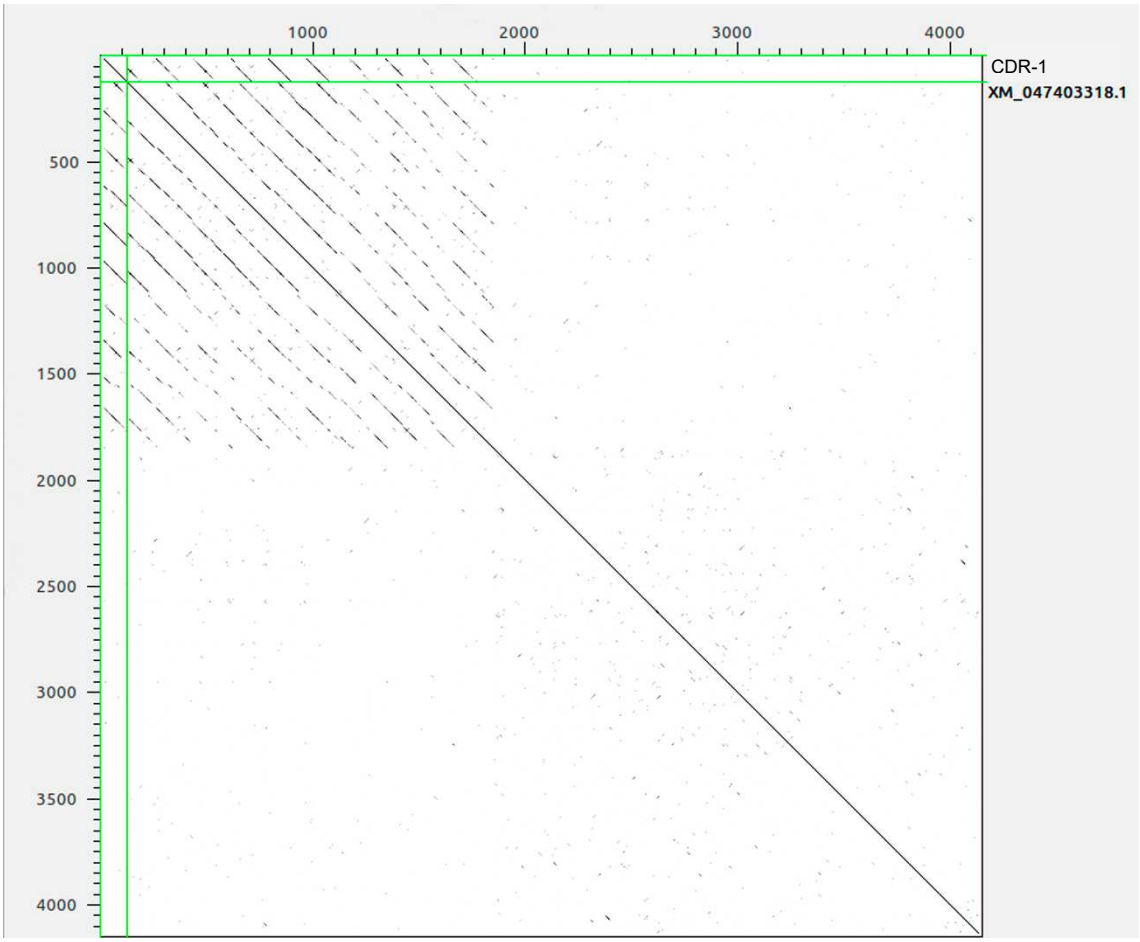
C. baccatum



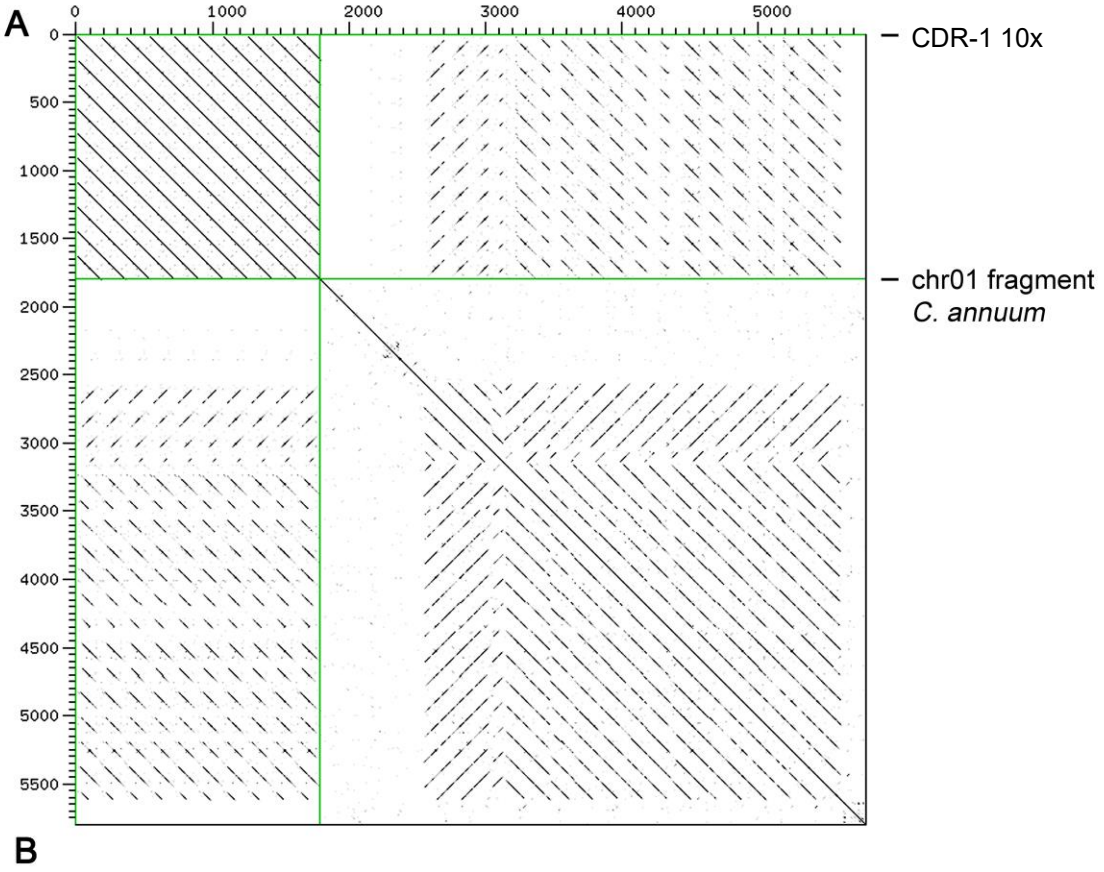
Supplementary Figure 3: WebLogo representing DNA motifs that appear tandemly accumulated downstream of IGS regions of 25S-18S rDNA. According to the sizes of motifs (8 bp), these ones could be considered as microsatellites, and not satDNA. Observe also these regions are relatively conserved between *Capsicum* species. 25S-18S rDNA of *C. pubescens* was used as reference.

```
                          .........| .........| .........| .........| .........| .........| .........70
CDR-2_60bp                CA-------- ---------- ---------- -----GACAT CTTATTTCGT CTTCATAGGG CGCCATCCCC
CDR-2_sub-type_1_40bp     T--------- ---------- --------C- ---------- -TTATTTCGT CTTCATAGGG CGCCATCCCC
CDR-2_sub-type_2_78bp     T--------- ---------- --------C- -----GACAT CTTATT-CGT CTTCATAGGG CGCCATCCCC
CDR-2_sub-type_3_106bp    T--------- ---------- --------C- -----GACAT -TTATTTCGT CTTCATAGGG CGCCATCCCC
CDR-2_sub-type_4_106bp    T--------- ---------- --------C- -----GACAT -TTATTTCGT C-TCATAGGG CG-CATCCCC
CDR-2_sub-type_5_39bp     ---------- ---------- ---------- ---------T CTTATT-CGT CTTCATAGGG CGCCATCCCC

                          .........| .........| .........| .........110
CDR-2_60bp                TAGTTGACAT CTTA--T--- ---------- --CTCGTCTT
CDR-2_sub-type_1_40bp     TAGTTGACA- ---------- ---------- ----------
CDR-2_sub-type_2_78bp     TAGTTGACA- ---------- ---------- ----------
CDR-2_sub-type_3_106bp    TAGTTGACA- ---------- ---------- ----------
CDR-2_sub-type_4_106bp    TAGTTGACA- ---------- ---------- ----------
CDR-2_sub-type_5_39bp     TAGTTGAC-- ---------- ---------- ----------
```

Supplementary Figure 4: Alignment of the sub-types of CDR-2 found within the non-autonomous LTR-RT. The sequences present within the LTR-RT elements were extract, concatenate, and analyzed with the TRF tool. From the TRF output, the most accumulated sequences were predicted as sub-types of the predicted CDR-2 monomer. Only the conserved regions identified in the alignment are shown in the figure.

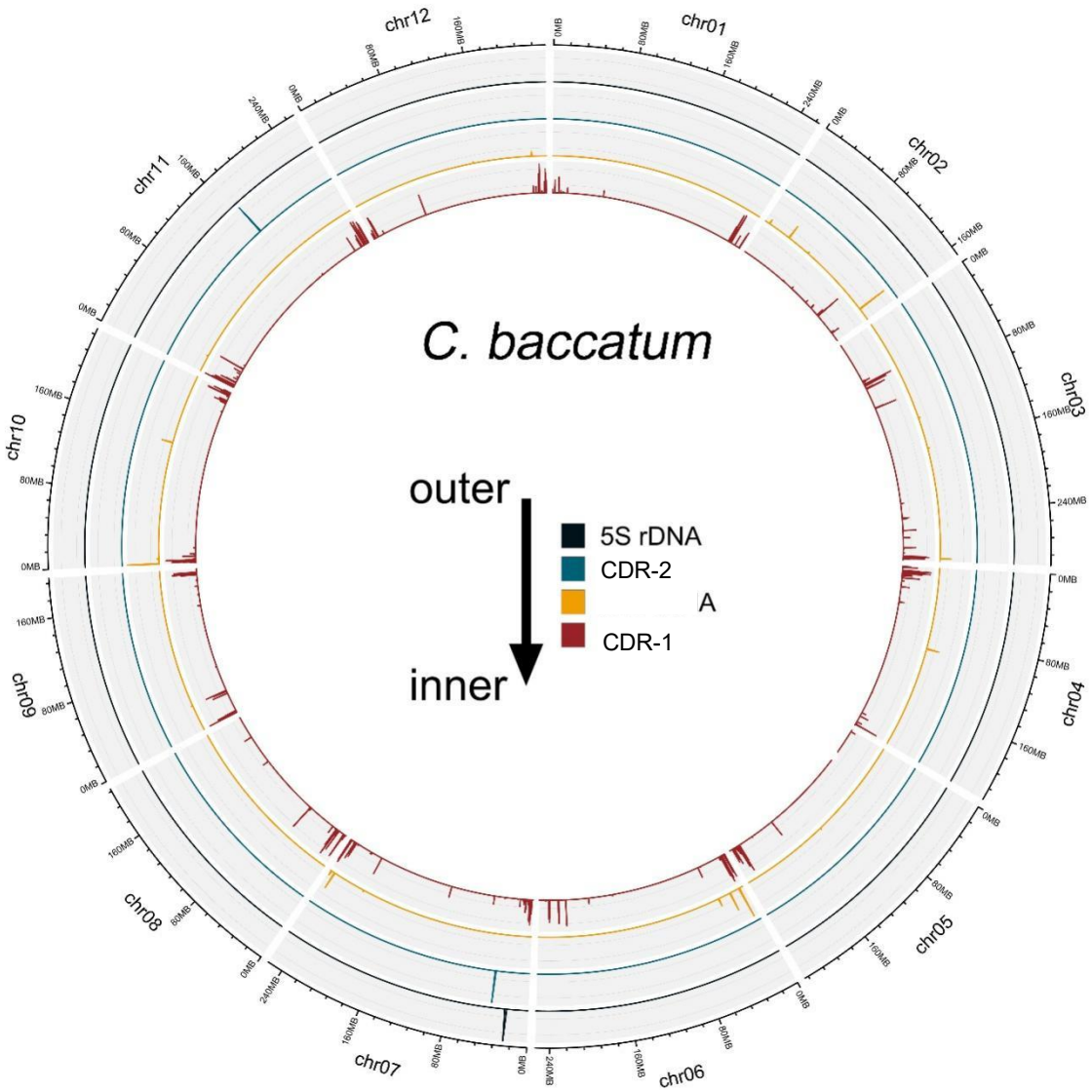


Supplementary Figure 5: Dotplot of CDR-1 against the predicted gene from *C. annuum* from the ANK gene family. Note that the CDR-1 monomer appears amplified in tandem at the beginning of the predicted gene from *C. annuum*.

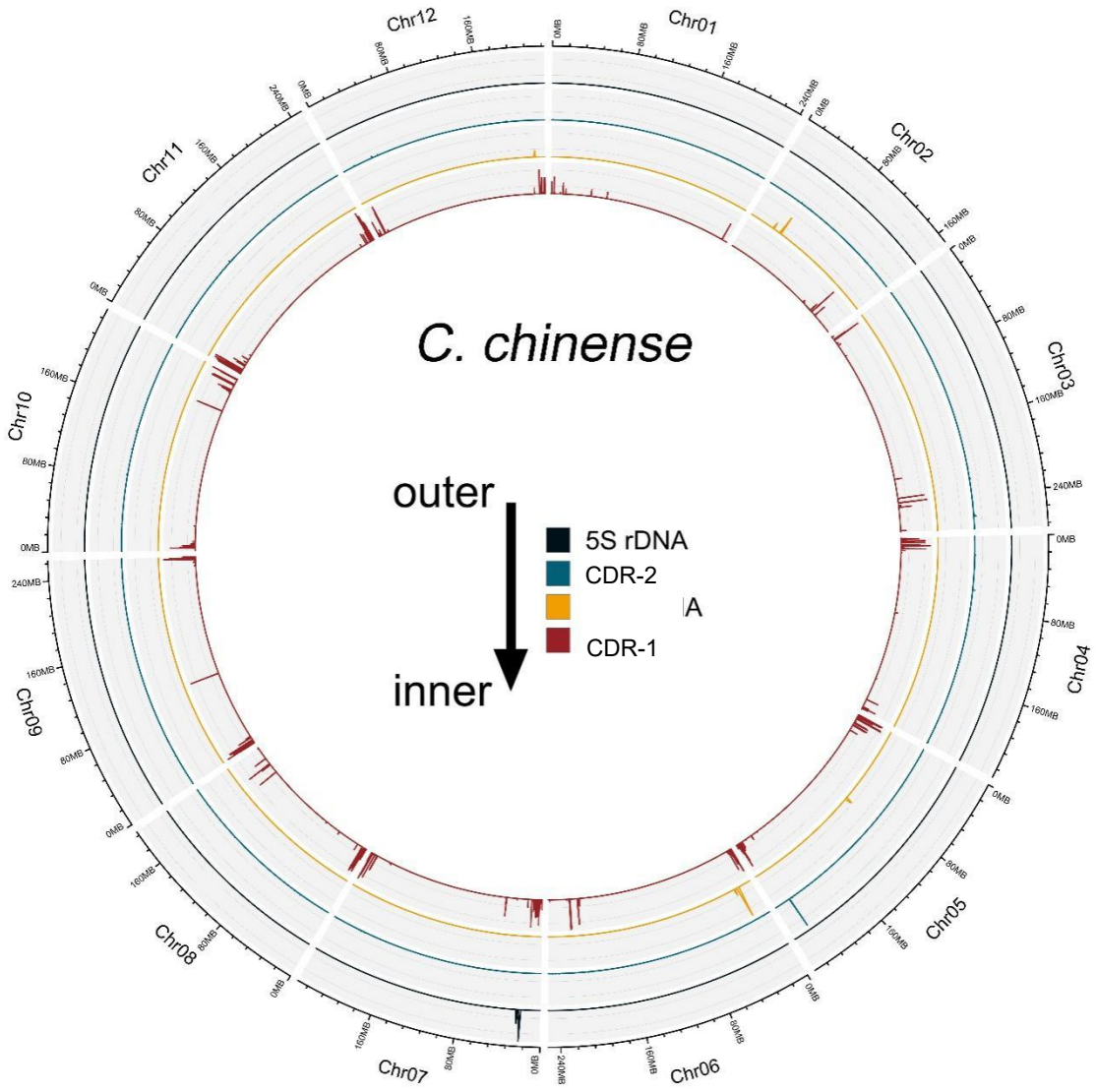



```
                  .........|.........|.........|.........|.........|.........|.........|.........|.........|.........100
CDR1-predict-monomer CCTAGTATGG GCCATTGGGG TGGGCGGGGC GGTTTGGATG GTCAAACAGG CGAAACGGCA CAAATGGGCC ATTTTCGGGC CAAATCAGTG TGCTATAGCC
annuum_sub-type_1    ---------- ---------- ---------- -ATTTGGGTG G-CAAACG-G -GAAAC-G-G -GAAC-GG-C ATTTT-AGGC CAAATC-GTG TGCTATAGCC
annuum_sub-type_2    ---------- ---------- ---------- ---------- --CAAACGGG C-AAAC---G -GAAC--G-- ATTTT--GG- CAAATC-GTG TGCTATAGCC
annuum_sub-type_3    ---------- ---------- ---------- ---------- ---------- --AAAC--GG C-AAC-GGCC -TTTT-AGGC CA----CGTG T-C--T-GCC
baccatum_sub-type_1  ---------- ---------- ---------- ---------- --CAAACGGG CGAAACG-GG CGAAC-GGCC ATTTTCAGGC ------CGTG T-C-A--GCC
baccatum_sub-type_2  ---------- ---------- ---------- ---------- --CAAACGGG C-AAACG-GG --AAC--GCC ATTTTCAGGC ------CGTG T-C-A--GCC
baccatum_sub-type_3  ---------- ---------- ---------- ---------- ---------- --AAACGGGG C-AAC-GGCC ATTTTCA-GC ---A--CGTG T-C----GCC
chinense_sub-type_1  ---------- ---------- --------GC CATTTGGGTG G-CAAACGGG CGAAAC-G-- CGAAC-GG-C ATTTT--GGC CAAATC-GTG TGCTATAGCC
chinense_sub-type_2  ---------- ---------- --------GC CATTTGGGTG G-CAAACGGG CGAAACGG-G CGAAC-GG-C ATTTTCAGGC CAAATC--TG TGCTATA-CC
chinense_sub-type_3  ---------- ---------- ---------- ---------- ---------- -------GGG C--AC----- ---------C CAAA-CCGTG -GCTATAGC-

                  .........|.........|.........|.........|.........|.........|.........|.........|.........|.........200
CDR1-predict-monomer CACAATTTTT AGGGTGGGCC CAGGGTCCGG GCATGATTTA ---------- ---------- AGTCCAA--A ATTGGTCGGG CTACCACAGG CAGGCTGGGG
annuum_sub-type_1    CACG-TTT-G GGGGTGGGCC -G-GGT-CGG G--TG-T-TG ---------- ---------- GG---AA--A AT----CGGG GCATGCCAGG GAGG-TGGGG
annuum_sub-type_2    CA---TTTTG GGGGT-GGC- CG-GGT-CG- GC-TGTT-TG ---------- ---------- GG-C-AA--A AT----C-GG --ATGCCAGG GAGG-TGGGG
annuum_sub-type_3    -ACGA--TTG GGGGTGGGCC C--GG--C-G G--T-TTTT- ---------- ---------- GG-C-AA--A AT-----G-G GCATGCCAGG GAGGCT-GGG
baccatum_sub-type_1  -A-GA----G GGGGTGGGCC CG-GG--CG- GC-T-TTTT- ---------- ---------- GG-------- ---------- ---------- ----------
baccatum_sub-type_2  -A-GA----G GGGGTGGGCC CG-GG--CGG GC-T-TTTTG ---------- ---------- GGTC-A---A AT-----G-- ---------- ----------
baccatum_sub-type_3  -ACGA----- GGGGT-G-CC C--GG---GG G--T-TTTT- ---------- ---------- GGT--AA--- ---------- ---------- ----------
chinense_sub-type_1  CACG-TTTTG GGGGTGGGCC CG-GGT-CG- GC-TGT-T-- ---------- ---------- ---------- ---------- ---------- ----------
chinense_sub-type_2  CACGATTTTG GGGGTGGGCC CG-GGT-CGG GC-TG----- ---------- ---------- ---------- ---------- ---------- ----------
chinense_sub-type_3  C-C-A-TTT- G--------- -----T-C-- ---TG----G ---------- ---------- -GTC-A---A A-----C--G GC-T-C---G -A-GCT-GG-

                  .........210
CDR1-predict-monomer A--------- -----
annuum_sub-type_1    A-CC------ -----
annuum_sub-type_2    A-CCA---G- -TGGA
annuum_sub-type_3    A---ACAGGC -----
baccatum_sub-type_1  ---------- -----
baccatum_sub-type_2  ---------- -----
baccatum_sub-type_3  ----ACAGGC -----
chinense_sub-type_1  ---------- -----
chinense_sub-type_2  ---------- -----
chinense_sub-type_3  --CC------ -----
```

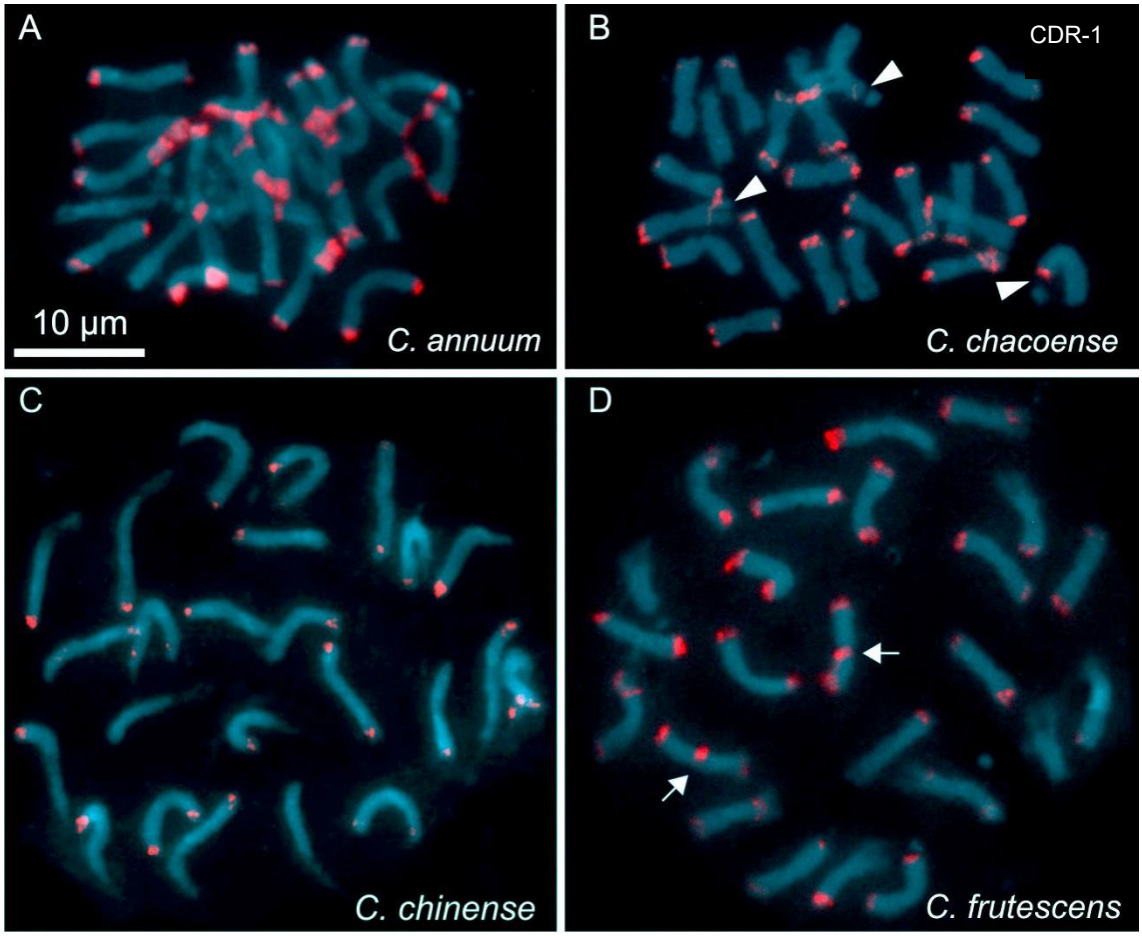
Supplementary Figure 6: Repetition profile of CDR-1 monomers. (A) Dotplot of the predicted CDR-1 monomer repeated 10 times against a fragment of chromosome 1 from *Capsicum annuum*. Note that the fragmentations and inversions present in the fragment of chromosome 1 from *C. annuum*, which indicates the presence of CDR-1 monomers sub-types. (B) Alignment of the subtypes found in *Capsicum* genomic datasets. Note that even there are gaps and indels some scratches are well aligned with the predicted monomer. Only the conserved regions identified in the alignment are shown in the figure.



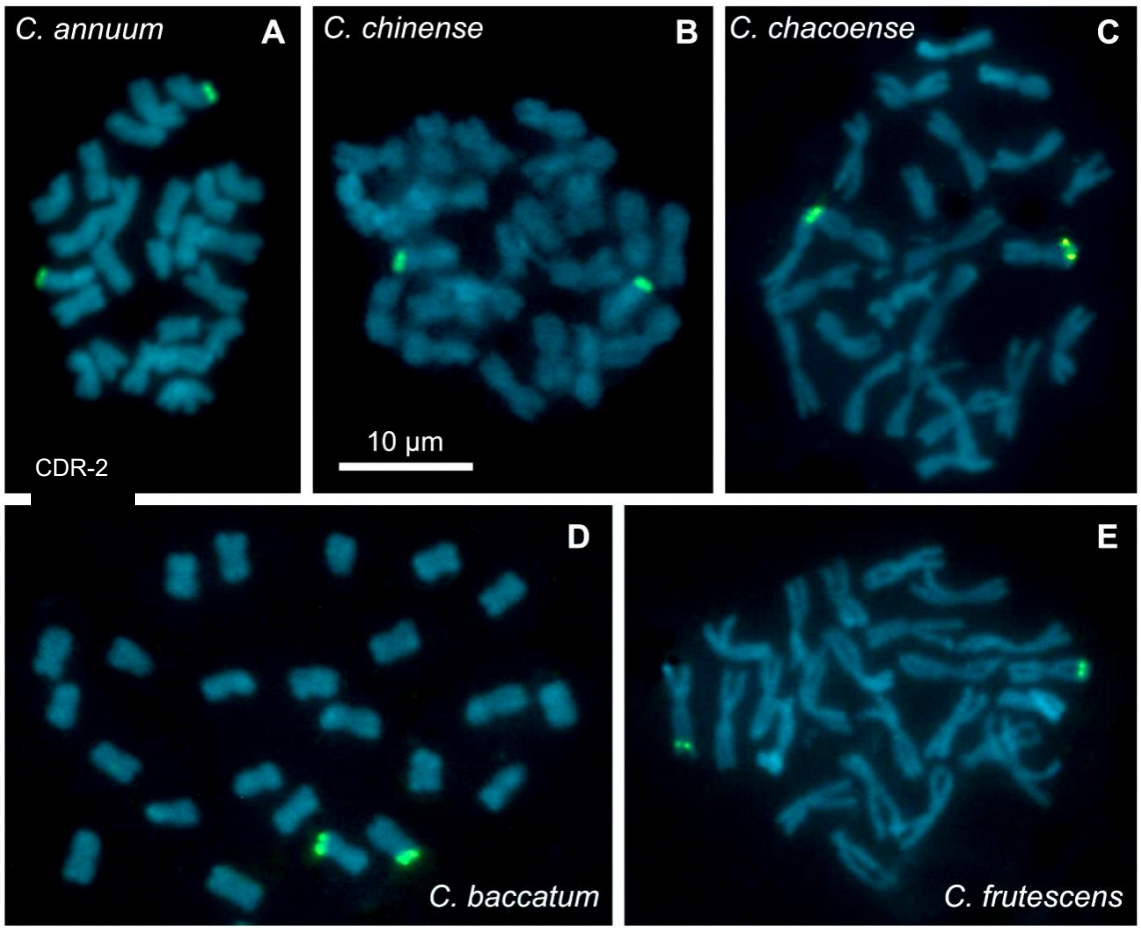
Supplementary Figure 7: Distribution of 5S rDNA, CDR-2, 35S rDNA and CDR-1 sequences in the pseudochromosomes from *Capsicum baccatum*. The 5S coding sequence exhibited only on peak of accumulation in chromosome 7. The CDR-2 monomer exhibited one peak of accumulation in each chromosome 7 and 11. The 35S rDNA sequence was accumulated in pseudochromosomes 2, 3, 4, 6, 7, 10, and 12. All the pseudomolecules exhibited accumulation of CDR-1 monomer in the distal regions. Note that there was accumulation of CDR-2 and 5S sequences in the pseudochromosome 7.



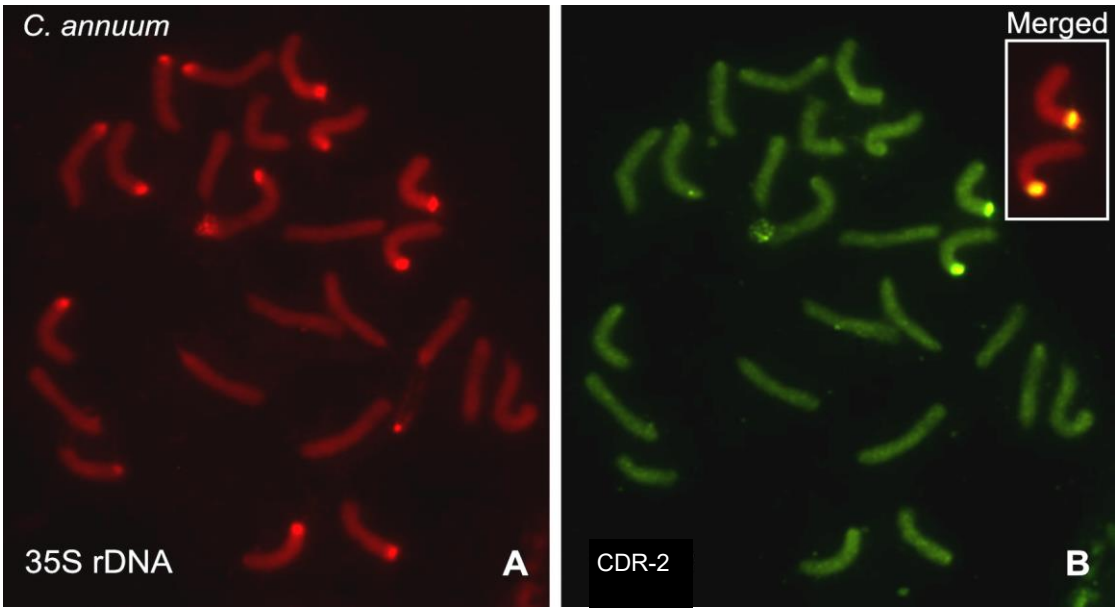
Supplementary Figure 8: Distribution of 5S rDNA, CDR-2, 35S rDNA and CDR-1 sequences in the pseudochromosomes from *Capsicum chinense*. The 5S coding sequence exhibited only on peak of accumulation in chromosome 7. The CDR-2 monomer exhibited one peak of accumulation in chromosome 5. The 35S rDNA sequence was accumulated in two pseudomolecules, the chromosomes 2, 5, 6 and 12. All the pseudomolecules exhibited accumulation of CDR-1 monomer in the distal regions. Note that CDR-2, 5S, and 35S rDNA do not accumulate in the same pseudochromosomes, except for a small interstitial peak of 35S on chromosome 5.



Supplementary Figure 9: FISH using CDR-1 probe against metaphases and prometaphases of *Capsicum annuum* (A), *C. chacoense* (B), and *C. frutescens* (D), all with 2*n*=24. Chromosomes were counter-stained with DAPI (blue), CDR-1 probe with Cy3–11-dUTP (red). Observe hybridization signals in almost all chromosomes, accumulating mainly in the distal regions. Although different factors can influence the FISH signal intensity, such as chromosome condensation, presence of cytoplasm, and image capture conditions, CDR-1 signals seems to be less intense in *C. chinense* (C) in relation to *C. annuum* (A). The arrowheads in *C. chacoense* (B) indicates the satellite region adjacent to secondary constriction, and the arrows in *C. frutescens* (D) indicates proximal FISH signals. The bar represents 10 µm.



Supplementary Figure 10: FISH using CDR-2 probe against metaphases and prometaphases of *Capsicum annuum* (A), *C. chinense* (B), *C. chacoense* (C), *C. baccatum* (D), and *C. frutescens* (E), all with 2*n*=24. Chromosomes were counter-stained with DAPI (blue), and CDR-2 labeled with biotin-11-dUTP / avidin–FITC conjugate (green). Note hybridization signals in the distal region of only one pair in all species. The bar represents 10 µm.



Supplementary Figure 11: FISH using 35S rDNA and CDR-2 probes in a prometaphase of *Capsicum annuum* (A and B, respectively). Note on the right a chromosome pair containing FISH signals co-located with both probes (merged image in the box).

5 CAPÍTULO 2 - COMPARATIVE ANALYSIS OF RETROTRANSPOSONS AMONG TOMATO SPECIES AND THE CHARACTERIZATION OF CENTROMERIC ELEMENTS

O artigo a seguir será submetido a revista *Frontiers in Plant Science*

Comparative analysis of retrotransposons among tomato species and the characterization of centromeric elements

Rafael de Assis[1], Leandro Simões Azeredo Gonçalves[2], Willem M. J. van Rengs[3], Charles Underwood[3], Romain Guyot[4], André Luis Laforga Vanzela[1]*

[1]*Laboratório de Citogenética e Diversidade Vegetal, Departamento de Biologia Geral, Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina, 86097-570, Paraná, Brazil.*
[2]*Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Londrina, 86057-970, Paraná, Brazil.*
[3]*Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany.*
[4]*Institute de Recherche pour le Développement, UMR DIADE, Montpellier, France.*

*Correspondent author: E-mail: andrevanzela@uel.br, ORCID: 0000-0002-2442-2211
Rafael de Asis, ORCID: 0000-0002-4420-4588
Leandro S.A. Gonçalves, ORCID: 0000-0001-9700-9375
Romain Guyot, ORCID: 0000-0002-7016-7485
Willem van Rengs, ORCID:
Charles J. Underwood, ORCID: 0000-0001-5730-6279

ABSTRACT: The cultivated tomato (*Solanum lycopersicum*) is a well-described model organism, and its wild relatives a great source of genetic diversity. Several tomato genomes have been published in recent years, including long reads with more fidelity, which include more information about the repetitive fraction, like retrotransposons and satellite DNA. Some retrotransposons can accumulate at the centromeric and pericentromeric regions, like CRM and, in some cases, they are often accompanied by arrays of satellite DNA. Species from the *Solanum* genus seem to have the CRM lineage underrepresented in the genomes, this information can rise questions about the composition of the centromeric and pericentromeric regions. The goals of this work were to characterize the centromeric region of the cultivated tomato and to explore the diversity of this region in closely related species from the tomato clade. To do this, bioinformatic tools were used to retrieve and map repetitive elements in pseudochromosomes and molecular cytogenetic methods for their physical localization of them. We used species from the tomato clade as a model organism focusing on the centromeric and pericentromeric regions. Here we show that centromeric regions of tomato species are composed of sequences from the Tekay lineage in arranges interspaced with satellite sequences. We also demonstrate that the *Jinling* retrotransposon is dispersed through the entire chromosome, not only at the pericentromeric regions as the previous report stated. This diversity of the arrange of centromeric sequences could be observed in all wild relatives.

Keywords: centromere, FISH, *Solanum*, transposons.

INTRODUCTION

Differences in DNA C-values can reach 2,400-fold among angiosperms (Pellicer & Leitch, 2020) due to whole-genome duplication (WGD) events, often associated with chromosome rearrangements (Hofstatter et al., 2022). However, the main portion of plant genomes is composed of repetitive DNA sequences, and changes in this fraction are related to expression or recombination rates and transposable elements (TEs) activity, which can lead to fluctuations in the genome sizes (Bennetzen et al., 2005; Wendel et al., 2016). In angiosperms, TEs are the most representative repetitive fraction, these elements can be classified according to their mechanism of transposition (Orozco-Arias et al., 2029). The Class I elements comprise the retrotransposons, which are transposed by an intermediate RNA, while the Class II elements, represented mainly by transposons, are those excised and inserted in different regions of the genome. (Bennetzen et al., 2014; Makałowski et al., 2019). Most of these elements are Long Terminal Repeat Retrotransposons (LTR-RT), and they can be found either dispersed or clustered through the chromosomes (Bennetzen et al., 2014). Some retrotransposons can accumulate at the centromeric and pericentromeric regions, like CRM (Centromeric Retrotransposon of Maize) and, in some cases, they are often accompanied by arrays of satellite DNA (satDNA). Centromeric retrotransposons belong to the Chromoviruses lineage and share common domains (chromodomain or CHRomatin Organization MOdifer domain and a targeting domain called CR motif) at the C-terminus of the integrase that may be linked to their chromosomal distribution (Neumann et al., 2011; Presting, 2018). This lineage was first described in maize (Nagaki et al., 2003), but it is widespread through plant genomes, like in *Coffea* (De Castro Nunes et al., 2018) and *Capsicum* (de Assis et al., 2020). However, members of this lineage seem to be underrepresented in the genomes of the *Solanum* clade (Yang et al., 2005; Benoit et al., 2019; Domínguez et al., 2020). To get an idea of the impact of TEs on Solanaceae

genomes structure just compare the DNA C-value expansion during the evolution of *Capsicum* and *Solanum* (about 3-fold), associated with activity and accumulation of Tekay (Park et al., 2014) and CRM retrotransposons (de Assis et al., 2020).

Satellite DNA families are also part of the repetitive fraction and are equally important to the organization and differentiation of genomes in Solanaceae, such as *Cestrum* (Souza et al., 2022). Satellites are organized according to repeat size, behavior in genomes, amplification, and dispersion mode. Microsatellites (SSR or Simple Sequences Repeats) are found in arrays up to 1kb, clustered in one region or dispersed throughout the chromosomes, and can be accumulated by slippage replication (Walsh, 1987; Garrido-Ramos, 2015). The minisatellites are those with repeats bigger than 10bp up to 30bp in length with an amplification mechanism like SSR. Differently from micro- and minisatellites, the satellite DNA presents monomers greater than 30bp, but with variable length (i.e., monomers of 150bp or 180bp or bigger), possessing long arrays reaching several kb in length (Garrido-Ramos, 2017; Thakur et al., 2021), and the amplification mechanism is given by unequal crossing-over, replication slippage, rolling circle replication of extrachromosomal circular DNAs (Ruiz-Ruano et al., 2016; Garrido-Ramos, 2017). Satellite DNA families are usually found in centromeric/pericentromeric (proximal regions) and sub-telomeric or distal regions and they played important roles in maintaining the heterochromatin architecture (Garrido-Ramos, 2017). The diversity of heterochromatin regions has been described for many species, such as *Rhynchospora* (Costa et al., 2021), *Alstroemeria* (Ribeiro et al., 2021) *Cestrum* (de Souza et al., 2022), *Vicia* (Robledillo et al., 2018). For the *Solanum,* section *Lycopersicum*, which comprises the tomato clade, a large portion of heterochromatin is Giemsa C-bands/CMA^{+} located at proximal and distal regions (Brasileiro-Vidal et al., 2009).

Monocentric chromosomes are those that have a well-defined primary constriction and that in anaphase migrate with the centromeres facing the poles of the cell and the telomeres facing the cell interior (McKinley et al., 2016). An active centromere is defined as the region where the functional CenH3 proteins and where the spindle microtubes bind during cell division (Comai et al., 2017). Many of the centromeres studied so far are composed of repeats ranging from 100 to 1000bp, as well as other classes like retrotransposons (Neumann et al., 2011; Melters et al., 2013). Centromeres can also contain different arrays of satDNA, and an example in Solanaceae is *Solanum tuberosum*, which has five chromosomes with no satellite repeats and the remaining with satDNA occupying megabases of the proximal region (Gong et al., 2012). Another classic example is in *Arabidopsis thaliana*, where the centromeric set presents a 180bp tandem repeat (pAt360) intermingled with Athila LTR-RT elements. (Pelissier et al., 1995).

Due to the great commercial importance of tomatoes, the diversity and impact of repetitive sequences have been explored. One important retrotransposon is the Rider element, which activity led to the reshaping of tomato fruit due to the duplication of the *SUN* locus (Xiao et al., 2008; Jiang et al., 2008). Four major classes of repetitive elements are described up to date in tomatoes, among them, the centromeric region has a transposon-like called TGRIV (Chang et al., 2008). High-quality genome sequencing is now available for some tomato species, which makes it possible to investigate shared features among them, like the composition of the repeatome focusing on the centromeric and pericentromeric regions. Due to the functional conservation of the centromeric region and genomic data available for tomatoes, some doubts have arisen about the tomato centromere sequences conservation and composition of them among different species from the same clade, and about lack of CRM lineages in the proximal

regions. Therefore, our interest was to compare the centromeric region of closely related species from the tomato clade, using bioinformatics tools and cyto-molecular techniques to investigate the presence of repetitive elements that are in the centromere of the cultivated tomato and wild species.

MATERIAL AND METHODS

Chloroplast genome assembly and phylogenetic reconstruction

Illumina reads from all described tomato species (Table 1)were used to reconstruct the chloroplast genome (cp) with the NOVOPlasty software (Dierckxsens et al., 2016). At the end of the assembly process, the cp genomes were compared with the reference *Solanum lycopersicum* (Genbank n. DQ347959) cp genome using a dot-plot, to check for incongruences of the assembly. The assembled cp genome was aligned with MAFFT 7.305 (Katoh et al., 2019) with the following parameters BLOSUM62 and 200PAM/ k = 2. The chloroplast genomes from *Capsicum annuum* L. (Genbank n. KJ619462) and *Solanum tuberosum* (Genbank n. DQ231562) were used as outgroups. The tree was constructed from the alignment with FastTtree software (Price et al., 2010) and edited in iTol (Letunic and Bork, 2021).

Repeats annotation

For the tomato repeatome analysis, Illumina reads were evaluated by FastQ report (Andrew, 2010) and then randomly reduced to 0.1× coverage. The reduced genomic dataset was used for RepeatExplorer comparative pipeline (Novák et al., 2010, 2013) with default parameters. The accessions used for this are listed in Table 1.

De novo genome assembly and search for RT domains

The Illumina reads were assembled for the species that neither the scaffolds nor the contigs were available at NCBI. For this MaSuRCA (Zimin et al., 2013) was used to assemble genomes and the output was used to identify the RT domains using CENSOR (Jurka et al., 1996). A minimum of 80% of nucleotide identity and 80% of sequence coverage against a reference database composed of 2,676 protein domains obtained from GypsyDatabase have been used. The CENSOR output was used to extract the reverse transcriptase sequences (RT) using a custom script. The RT domains retrieved were aligned with the references and a tree was built as described for the chloroplast tree.

Transposable elements annotation

The library of transposable elements was constructed with an EDTA pipeline (Ou et al., 2019) using long reads of genomic sequencing from *S. lycopersicum*, *S. pimpinelifolium*, *S. pennellii,* and *S. cheesmaniae* (Genbank accessions available in sup Table 1). The retrieved LTR-RT elements were then annotated with the pipeline implemented in Inpactor (Orozco-Arias et al., 2018). The phylogenetic tree was constructed the same as described for cp genomes but with RT domains that were excreted from the annotated elements. For the tree, reference sequences were added, including the centromeric retrotransposon-like TRGIV (Genbank n. EU526907) and the Rider element (Genbank n. EU195798).

Annotation of centromeric sequences

For the commercial tomato (*Solanum lycopersicum*), the centromeric region has a repetition transposon-like (TGRIV). Aiming to better understand the composition of centromeric regions, the RT domain from TGRIV was extracted and aligned with all the

other RT domains found in the genome. The sequences belonging to the same clade as the TGRIV, based on tree topology, were considered TGRIV-like and used to map the pseudochromosomes. The dataset of TGRIV-like sequences was used as a custom database in RepeatMasker (Smit et al., 2013). The gff output was used as an entry in DensityMap (Guizard et al., 2016) with a window (-sc) of 1Mb, and finally, the output of DensityMap was used in ShinyCircos (Yu et al., 2018). Regions that exhibited a pic of accumulation of TGRIV-like were extracted from the assembly and compared using by dot-plot, extracted sequences were then annotated using CDD NCBI to identify the domains and Artemis (Carver et al., 2012) to infer the ORFs and do the annotation as EMBL format.

Plant material

Seeds of *Solanum lycopersicum* cv. Moneyberg, *S. pimpinellifollium* (accession LAJ614), *S. galapagense* (accession LA1401), *S. pennellii* (accession LA0716), and *S. cheesmaniae* (LA1039) were sowed in 128-cell polystyrene trays containing the substrate Vivatto®. Ten seedlings of each species were grown in the Laboratório de Citogenética e Diversidade Vegetal, Universidade Estadual de Londrina, Brazil.

Genome size estimation

Nuclear DNA amount measurements (DNA C-values) were done with young leaves using 1 mL of cold LB01 buffer plus 1 mg/mL propidium iodide (Doležel et al., 2007). For this, leaves were chopped in 250 µL of buffer and RNAse (1 mg/mL^{-1}). Samples were filtered in a 30-µm nylon mesh and centrifuged at 500× *g*. Subsequently, the nuclei were stained with propidium iodide (1 mg/mL^{-1}). Samples were filtered again in a 20 µm nylon mesh. A BD ACCURI C6 flow cytometer (Becton, Dickinson, and Company) was used, according to equipment specifications. Three independent

estimations were performed on different days, using at least 30,000 nuclei in each assay. *Raphanus sativus* L. (2C = 1.26pg), was used as standard (Doležel et al., 2007). The 2C values were calculated as sample peak mean / standard peak means × 2C DNA amount of standard (pg).

DNA extraction, PCR, and probe design

To confirm the distribution of the putative centromeric elements, different probes were designed for in situ hybridizations. Two oligomers labeled with digoxigenin-11-dUTP, one for the putative centromeric satellite (satTCS) and one for the 3'LTR from the putative centromeric LTR (LTR-cent), and a primer for 3' LTR from the centromeric retrotransposon-like TRGIV, all the sequences are available in Table 2.

The genomic DNA of *Solanum lycopersicum* was used as a template. For that, it was isolated from young leaves using the cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1987), 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).

PCR was conducted using a mix containing 50 mM $MgCl_2$ (1.5 µL), 10 mM dNTP (1 µL), 5 mM primers (2 µL each), ~30 ng DNA template, 1.25 U of Taq polymerase, and ultrapure water to a final volume of 25 µL. Probes were labeled using 0.2 mM dNTP, containing dGTP (25%), dCTP (25%), dATP (25%), dTTP (17.5%) and biotin-dUTP or Cy3-dUTP (7.5%). A standard PCR was used under the following conditions: 94 °C for 2 min, 30 cycles of 94 °C for 40 s, from 54 to 56 °C (depending on the primer set) for 40 s, and 72 °C for 1 min, and a final extension of 72 °C for 10 min.

The reactions were tested via electrophoresis in an agarose gel at 3 V cm^{-1} and stained with ethidium bromide. The p*Ta*71 clone containing an insert of ~9 kb with a complete 35S rDNA isolated from *Triticum aestivum* (Gerlach and Bedbrook, 1979) was labeled with digoxigenin-11-dUTP via nick translation and also used as a probe.

Chromosome preparation for FISH

The fluorescence *in situ* hybridization (FISH) assays were conducted in four species of *Solanum* (*S. lycopersicum, S. pimpinelifollium, S. pennelli* and *S. cheesmaniae*) using probes of 5S and 35S rDNA, TGRIV, and CTR-2 satDNA families. The root tips were pretreated with 0.5% colchicine (1 h 30 min) and fixed in ethanol-acetic acid (3:1, v:v). Samples were 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 FISH, slides received a mix containing a solution (30 µL) composed of 100% formamide (15 µL), 50% dextran sulfate (6 µL), 20× SSC (3 µL), 100ng of calf thymus DNA (1 µL), 10% SDS (1 µL) and 100ng of probes (4 µL). The mix was denatured at 90 °C for 10 min, and hybridization was performed at 37 °C for 24h in a humid chamber. Post-hybridization washes were carried out with 70% stringency, using an SSC buffer, with a pH 7.0. After the probe detection with an avidin-fluorescein isothiocyanate (FITC) conjugate and anti-digoxygenin-rhodamine (anti-DIG), 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). 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 digoxigenin) and contrasted using GIMP 2.8 Linux.

RESULTS

Phylogenetic inferences and genome size

The reconstruction of the phylogeny of tomato species with a CP three evidenced three major groups (Figure 1). These groups didn't exhibit a relationship between the DNA C-values and three positioning, with genome sizes ranged from 831 Mbp to 1125 Mbp (Table 1).

From the tree obtained, was possible to observe the presence of three major groups. The first group was composed of *Solanum lycopersicum*, *S. pimpinellifolium* (red fruits), *S. cheesmaniae*, *S. galapagense* (orange fruits), and *Solanum habroichaites* (green fruits) (Figure 1). The second was composed of species with green fruits, and *S. lycopersicoides* was the most divergent forming an isolated group. The DNA content followed this pattern of distribution, species with the lowest amount of DNA were grouped in the same clade while the bigger ones were in another. One specie stands out for being the most divergent among the tomato species, *Solanum lycopersicoides* was the more basal specie in the maternal tree, but with DNA content value similar to the most recent species (Figure 1).

Repeats diversity annotation

Transposable elements belonging to the Gypsy superfamily were the most common in all datasets (Figure 1). No drastic difference in accumulation rates were observed among all species, Tekay and Athila were the most representative lineages, both from the Gypsy superfamily. Although no drastic difference, the LINE elements

stand out for the great accumulation among seven green fruit species (Figure 1). As observed on the CP tree, *Solanum lycopersicum* and *S. pimpinellifolium* had highly similar accumulation patterns, the same could be observed for the other groups (Figure 1). *Solanum habrochaites* stands out for not being grouped with the green fruit species, even though it has green fruits, and showed an accumulation of LINE elements as well as the green fruits. Overall, Tekay elements were the most accumulated among all species, except for *S. lycopersicoides*, which almost didn't accumulate, based on the RepeatExplorer comparative analysis (Figure 1). Even with a robust library of well-annotated elements, some fractions of transposable elements were not addressed to any lineage or only to the superfamily level (Figure 1).

Centromeric composition among tomato species

The tomato centromere is described as having a retrotransposon-like sequence called TGRIV, the RT domain from this sequence was extracted and added with all the RTs from the full-length elements retrieved for the RT phylogenetic tree construction (Figure 2). The RT domain from this repetitive element was grouped in the Tekay lineage (Figure 2), this approach made it possible to identify the putative copies of TGRIV in four genomic datasets, totalling 319 full copies. We individually map each set of sequences on pseudochromosomes of *S. lycopersicum* cv. Moneyberg (Figure 3), *S. pimpinellifolium* cv. LA2093 (Supplementary figure 1), *S. cheesmaniae* cv. LA1039 (Supplementary figure 2), and *S. pennelli* (Supplementary figure 3) aiming to identify the putative centromeric region. The search using RepeatMasker, followed by Density map, showed that in *S. cheesmaniae*, the TGRIV-like sequences are not clustered in one region, even though it is possible to see a clear peak in 4 pseudochromosomes, for the remaining species the peaks were very clear (Supplementary figure 2).

Once stated where was the putative centromeric region, they were extracted and analyzed with a dot-plot, and a set of sequences that composes the centromere of tomato species was retrieved. Firstly, we have identified a TGRIV sequence containing 9366bp length (Figure 4), followed by a retrotransposon-like with 8898bp (Figure 5), which contains satellite sequences in both LTRs and, finally, a satellite family containing monomers with 75bp. These sequences appear to have a differential accumulation among chromosomes from the same species and among species (Figure 1 and Supplementary figures 1, 2, and 3).

We also used a satellite sequence called TCS-1 (Tomato Centromeric Satellite – 1) described as residing in centromeres, to virtually map the pseudochromosomes with shinyCircos and physically the chromosomes with FISH. Results of the bioinformatic analysis showed that *Solanum lycopersicum*, *S. pimpinellifolium*, and *S. cheesmaniae* exhibited three pseudochromosomes with a bigger peak, always in the same chromosome pair (Figure 1 and Supplementary figures 1 and 2). *Solanum pennelli*, differently, exhibited an additional chromosome with a bigger peak, besides differences in which chromosome has the peak (Supplementary figure 3).

Fluorescence in situ hybridization assays revealed either scattered or clustered signals, depending on the probe analyzed. The putative centromeric satellite exhibited strong signals in 3 pairs in *Solanum lycopersicum* and *S. cheesmaniae* (Figure6), while in *S. pimpinellifolium* it was possible to observe 4 pairs with stronger signals and 1 pair with weak signals, this species was the only one that exhibited signals for 35S and the satellite in the same chromosome pair (Figure 6), they remain the signals observed were in distinct chromosomes. Both TGRIV and *Jinling* belong to the Tekay lineage, usually probes made for this lineage show a scattered profile, this pattern was observed for the *Jinling*, but some signals were present at the centromeric region (Figure 6). The probe

designed for TGRIV exhibited pericentromeric signals, but not in all chromosomes, some chromosomes had a more intense signal as shown in the boxes in Figure 6. A generalist probe made for the RT domain of Tekay lineage was also tested, which made it possible to observe the scattered profile as well as a few centromeric signals (Figure 6). All the strong signals observed corroborate with the data observed by bioinformatic analysis.

DISCUSSION

The tomato clade of *Solanum* sect. *Lycopersicum* exhibits, in general, diploid species ($2n = 2x = 24$), with small acro- to metacentric chromosomes and accumulation of heterochromatin at proximal and distal chromosome regions (Brasileiro-Vidal et al., 2009; Anderson et al., 2010). This repetitive fraction has been extensively explored in conventional and molecular cytogenetics, which improved the understanding of the diversity of proximal and distal chromosomal regions. In addition to the differences in the amount of pericentric heterochromatin between *S. lycopersicum* and wild species mentioned by Anderson et al. (2010), these authors reported significant rearrangements in the pericentric/heterochromatic regions. Proximal chromosome regions of plants are rich in LTR-RTs, and these elements may be responsible for large- to small-scale rearrangements (Bennetzen and Wang, 2014), and this was also demonstrated in a comparison between the genomes of the tomato and potato clades (Gaiero et al., 2019). In fact, our results showed that tomato clade species have few copies or fragments of CRM retrotransposons, and an accumulation of Tekay retrotransposons, both Chromoviruses clade of the Gypsy superfamily. Considering that CRM retrotransposons predominate in the centromeric regions of several plant groups, such as in *Capsicum* (de Assis et al., 2020; Yañez-Santos et al., 2021) and other plant genera (Neumann et al.,

2011; De Castro Nunes et., al 2018), it is possible that rearrangements are involved in the drastic reduction of CRM elements in *Capsicum* genomes. This can be supported by the reports that tomato genomes contain more degraded or truncated elements, such as an example of *Jinling* retrotransposon commonly found in the pericentromeric heterochromatin (Wang et al., 2006; Gaiero et al., 2019).

The centromere plays important roles in sister chromatid cohesion and regular segregation during cell divisions (Oliveira and Torres, 2018), functional centromeres are, in general, associated with the presence of a centromere-specific H3 variant, CenH3 (Henikoff et al., 2001). For most species described, the centromeres occur in the primary constriction, and they are rich in repetitive DNA sequences (Musacchio et al., 2017). Many of these sequences have a direct association with CenH3 domains, which are fundamental for chromatid and chromosome kinetics (Oliveira and Torres, 2018). However, not always retrotransposons accumulated in the centromere region present chromodomains responsible for recognizing CenH3. This may be the case for the two retrotransposons mentioned by Nagaki et al. (2011) in *Nicotiana*, as well as the two Tekay members (TGRIV and *Jinling*) accumulated in the centromeric regions of species of tomato clade. TGRIV elements were described by Chang and colleagues (2008), as present in all centromere regions, including a small amount in the pericentromeres. Our comparative analysis across species of the tomato clade showed that TGRIV predominates in the centromeres even though they don't have typical chromodomains (like CRM ones), but that autonomous elements seem also to be widespread along chromosomes from *S. lycopersicum*, and in wild species. We have observed *Jinling* elements in all centromeric and interstitial regions, varying in quantity within and between chromosomes. However, this finding is contrary to the previous distribution pattern description of *Jinling* made by Wang and colleagues (2006), that state that this

element is present only in the pericentromeric heterochromatin. *Jinling* has been found in all 13 genomes of the tomato clade, and it is related to TGRIV elements, except by the presence of satDNA monomers in the LTRs. Even though these two elements are phylogenetically close, we still cannot define whether the satellites present in the LTR of *Jinling* were lost in TGRIV, or if the opposite occurred. Ty3/Gypsy elements were the most frequent in the tomato species (Gaiero et al. 2019), especially those from the Tekay lineage. The FISH assays exhibited two distinct distribution patterns: TGRIV was accumulated in the proximal region, while the *Jinling* had a scattered profile, with clear signals in the proximal region, like the TGRIV.

Chromosomes of species of the tomato clade exhibit a variety of repetitive sequences, such as those called “telomeric” (TR), interstitial/telomeric repeats (ITR), sub-telomeric to proximal (TGRI), other centromeric/pericentromeric retrotransposons (Zhong et al. 1998), as well as satellite DNA families. Some satellites are important from the standpoint of karyotypic diversity in the tomato clade. An example is the repetitive sequence St49 present in the functional centromeres of *S. tuberosum*, which has 77% of identity with the telomeric-like centromeric repeats described in *S. bulbocastanum* (Gong et al., 2012). This satellite appears associated with ITR sequences at centromeres of *S. chacoense*, *S. pinnatisectum* and *S. bulbocastanum* (He et al., 2013). It is possible that satDNA/ITR sequences may have occupied these regions after illegitimate recombination, rolling cycle replication or other events related to the "de novo" formation of repeats. We feature a new satDNA family in some centromeres of the 4 species, named satTCS. This satellite was composed of monomers with 75bp length, and it is probably related to the centromeric repeats mentioned by Su and colleagues (2021), but not characterized by them. The characterization and physical mapping of satTCS showed a differential accumulation of this sequence in the

centromeres of three or four chromosome pairs, depending on the species. These data increase the evidence about the high variability in the centromere regions of the *Solanum* sect. *Lycopersicum* clade. The diversity in centromeric regions of *Solanum* has been described by Gong and colleagues (2012), where the potato centromeres were not equally composed, six centromeres present megabases-sized satellite repeat arrays that are unique to each centromere and the remaining centromeres are composed of single- or low-copy DNA sequences. For the genus *Cestrum*, the centromere is enriched by satellite sequences (de Souza et al., 2022).

The tomato is an important cultivated species, and it considered a model organism for genetic assays, such as MicroTom. In this work we explore the proximal chromosome regions of 13 species of tomato clade, and we could observe a great diversity of sequences that resides in these regions. We access this diversity using long reads genome sequencing (PacBio and Nanopore) in 4 species. The retrotransposon *Jinling*, described as commonly accumulated in the pericentromeric heterochromatin, seems to be an important component of centromeres, along with the TGRIV, although we found it dispersed though the chromosomes. In addition, a centromeric satellite named of TCS was also common in some centromeres and it was accumulated differentially, signal intensity and number of pairs carrying it. Our results reinforce the idea of a great diversity of sequences residing in the centromeric region. These data is important for future studies about the centromeric architecture and the phylogenetic relationships and evolutionary approaches.

Supplementary Information -

Supplementary Information - The online version contains supplementary material available at <link>

Supplementary Figure 1: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum pimpinellifolium*. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS exhibited three major peaks (chromosomes 6, 8, and 11) colocalized with TGRIV peaks, and four minor peaks (chromosomes 1, 2, 3, and 12). The sequences from 35s rDNA exhibited a major peak in the chromosome 11 and three minors in chromosomes 2, 3, and 6.

Supplementary Figure 2: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum cheesmaniae*. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS exhibited three major peaks (chromosomes 6, 8, and 11) colocalized with TGRIV peaks, and four minor peaks (chromosomes 1, 2, 3, and 12). The sequences from 35s rDNA exhibited a major peak in the chromosome 2 and two minors in chromosomes 6 and 11.

Supplementary Figure 3: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum pennellii*. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS was accumulated in chromosomes 2, 6, 8, and 12 colocalized with TGRIV peaks, and two chromosomes with minor peaks (chromosomes 3, and 4). The sequences from 35s rDNA exhibited a major peak in the chromosome 7.

ACKNOWLEDGMENTS

R.A. is grateful to CAPES (Finance Code 001) and CNPq for awarding the scholarships. A.L.L.V. is grateful for financial support from the Brazilian Agency CNPq (processes 407194/2018-5 and 309902/2018-5). The authors also thank ProPPG-UEL, PPG-GBM, FINEP, and Fundação Araucária for other types of support.

ARTICLE INFORMATION

Availability of data and materials

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

AUTHOR INFORMATION

Author ORCIDs

André Luis Laforga Vanzela https://orcid.org/0000-0002-2442-2211

Author Contributions

A.L.L.V. and R.A. conceived the study. R.A. conducted the experiments and analyzed the data. R.A. and R.G. provided bioinformatics support to the team. L.S.A.G. provided the plants and collected the leaves and roots. A.L.L.V. and R.A. interpreted the data and wrote the manuscript. All the authors read and approved the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Funding

This study was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação Araucária, Paraná (FA), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Ethical statement

This article does not contain any studies with human participants or animals performed by any of the authors.

Competing interests

The authors declare no competing interests.

REFERENCES

Anderson L.K., Covey P.A., Larsen L.R., Bedinger P., Stack S.M. (2010). Structural differences in chromosomes distinguish species in the Tomato Clade. Cytogenet Genome Res; 129:24-34.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., ... & Macas, J. (2018). Satellite DNA in Vicia faba is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific reports*, *8*(1), 1-11.

Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 2005;95:127–32.

Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65:505–30.

Benoit, M., Drost, H. G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D., & Paszkowski, J. (2019). Environmental and epigenetic regulation of Rider retrotransposons in tomato. *PLoS genetics*, *15*(9), e1008370.

Brasileiro-Vidal, A.C., Melo-Oliveira, M.B., Carvalheira, G.M.G. Guerra, M. (2009). Different chromatin fractions of tomato (*Solanum lycopersicum* L.) and related species. *Micron*, 40(8): 851-859.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics, 28(4), 464-469.

Chang, S. B., Yang, T. J., Datema, E., Van Vugt, J., Vosman, B., Kuipers, A., ... & De Jong, H. (2008). FISH mapping and molecular organization of the major repetitive sequences of tomato. *Chromosome research*, *16*(7), 919-933.

de Souza, T. B., Parteka, L. M., de Assis, R., & Vanzela, A. L. L. (2022). Diversity of the repetitive DNA fraction in Cestrum, the genus with the largest genomes within Solanaceae. Molecular Biology Reports, 49(9), 8785-8799.

Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc. 2:2233-2244. https://doi.org/10.1038/nprot.2007.310

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J. M., Colot, V., & Quadrana, L. (2020). The impact of transposable elements on tomato diversity. *Nature communications*, *11*(1), 1-11.

Doyle, J.J., and Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. Focus. **12**: 13-15.

Gaiero P, Vaio M, Peters SA, Schranz ME, Jong H, Speranza PR (2019) Comparative analysis of repetitive sequences among species from the potato and the tomato clades. Annals of Botany 123: 521–532.

Gerlach, W.L., and Bedbrook, J.R. 1979. Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res. **7**(7): 1869-1885. https://doi.org/10.1093/nar/7.7.1869

Gong, Z., Wu, Y., Koblížková, A., Torres, G. A., Wang, K., Iovene, M., Pavel Neumann, Wenli Zhang, Petr Novák, C. Robin Buell, Jiří Macas, Jiang, J. (2012). Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *The Plant Cell*, *24*(9), 3559-3574.

Henikoff S, Ahmad K, Malik HS (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, 293(5532): 1098-1102.

Hofstatter et al., 2022, Repeat-based holocentromeres influence genome architecture and karyotype evolution. Cell 185, 3153–3168

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

Jiang, N., Gao, D., Xiao, H., & Van Der Knaap, E. (2009). Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. *The Plant Journal*, *60*(1), 181-193.

Jurka, J., Klonowski, P., Dagman, V., Pelton, P. (1996) CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. Computers and Chemistry Vol. 20 (No. 1): 119-122. (ftp://ftp.ncbi.nlm.nih.gov/repository/repbase/SOFTWARE/)

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, *20*(4), 1160-1166.

Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., ... & Choi, D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics*, *46*(3), 270-278.

Kursel, L.E. and Malik, H.S. (2016) Centromeres.Curr. Biol.26, R487–R490.

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*, *49*(W1), W293-W296.

Makałowski, W., Gotea, V., Pande, A., & Makałowska, I. (2019). Transposable elements: Classification, identification, and their use as a tool for comparative genomics. In *Evolutionary Genomics* (pp. 177-207). Humana, New York, NY.

Marques, A., Schubert, V., Houben, A., & Pedrosa-Harand, A. (2016). Restructuring of holocentric centromeres during meiosis in the plant *Rhynchospora pubera*. Genetics, 204(2), 555-568.

Melo, N.F., and Guerra, M. 2003. Variability of the 5S and 45S rDNA sites in *Passiflora* L. species with distinct base chromosome numbers. Ann Bot. **92**: 309-316. https://doi.org/10.1093/aob/mcg138

Murata, M., Ogura, Y. & Motoyoshi, F. Centromeric repetitive sequences in *Arabidopsis-Thaliana*. *Jpn J. Genet.* **69**, 361–370 (1994).

Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR, Jiang J: Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 2003, 163:759-770.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., ... & Macas, J. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA*, *2*(1), 1-16.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinform 29,792-793. https://doi.org/10.1093/bioinformatics/btt054

Novak, P., Robledillo, L.A., Koblizkova, A., Vrbova, I., Neumann, P., Macas, J. 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res. 45,111. https://doi.org/10.1093/nar/gkx257

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

Orozco-Arias, S.; Liu, J.; Tabares-Soto, R.; Ceballos, D.; Silva Domingues, D.; Garavito, A.; Ming, R.; Guyot, R. Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. Biology 2018, 7, 32.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. **20**: 275. https://doi.org/10.1186/s13059-019-1905-y

Pelissier T, Tutois S, Deragon JM, Tourmente S, Genestier S,Picard G: Athila, a new retroelement from Arabidopsis thaliana. Plant Mol Biol 1995, 29:441-452

Pellicer, J., & Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. New Phytologist, 226, 301–305. https://doi.org/10.1111/nph.16261

Presting, G. G. (2018). Centromeric retrotransposons and centromere function. *Current Opinion in Genetics & Development*, *49*, 79-84.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), e9490.

Ribeiro, T., Vaio, M., Félix, L. P., & Guerra, M. (2022). Satellite DNA probes of Alstroemeria longistaminea (Alstroemeriaceae) paint the heterochromatin and the B chromosome, reveal a G-like banding pattern, and point to a strong structural karyotype conservation. *Protoplasma*, *259*(2), 413-426.

Sharma S, Raina SN. Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. Cytogenet Genome Res. 2005;109:15–26.

Shibata, F., and Murata, M. (2004). Differential localization of the centromere-specific proteins in the major centromeric satellite of Arabidopsis thaliana. J. Cell Sci. 117: 2963–2970.

Smit, A., Hubley, R., and Green, P. 2013. RepeatMasker 4.0. *Seattle, WA: Institute for Systems Biology*. https://www.repeatmasker.org/

Walsh, J. B. (1987). Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics*, *115*(3), 553-567.

Wendel J F, Jackson SA, Meyers BC, Wing RA (2016) Evolution of plant genome architecture. Genome Biol 17:1-14. https://doi.org/10.1186/s13059-016-0908-1

Willard, H. F. & Waye, J. S. Hierarchical order in chromosome-specific human alpha-satellite DNA. *Trends Genet.* **3**, 192–198 (1987).

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & Van Der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *science*, *319*(5869), 1527-1530.

Yañez-Santos, A. M., Paz, R. C., Paz-Sepúlveda, P. B., & Urdampilleta, J. D. (2021). Full-length LTR retroelements in Capsicum annuum revealed a few species-specific family bursts with insertional preferences. *Chromosome Research*, *29*(3), 261-284.

Yang, T. J., Lee, S., Chang, S. B., Yu, Y., de Jong, H., & Wing, R. A. (2005). In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a large CAA block and characterization of pericentromere retrotranposons. *Chromosoma*, *114*(2), 103-117.

Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Yorke JA, Dvorak J, Salzberg S. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm. Genome Research. 2017 Jan 1:066100.

IMAGES

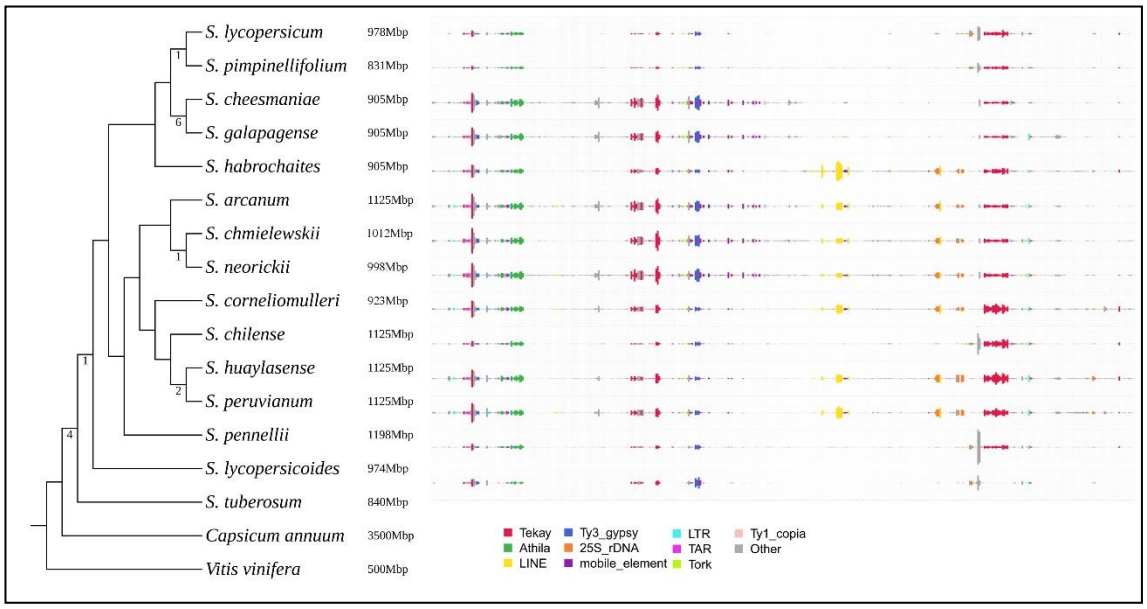


Figure 1: Phylogenetic tree and repetitive elements comparative analysis. The left panel is the chloroplast phylogenetic tree, the number are the bootstrap values and after each specie the genome size (1C) in Mbp, *S. tuberosum*, *Capsicum annuum*, and *Vitis vinifera* were used as outgroups. The right panel is the RepeatExplorer comparative analysis. Tekay were the most representative followed by Athila (both from Gypsy superfamily) and LINEs. Note that even with a well annotated database, some elements were only addressed to Gypsy superfamily.

1

2

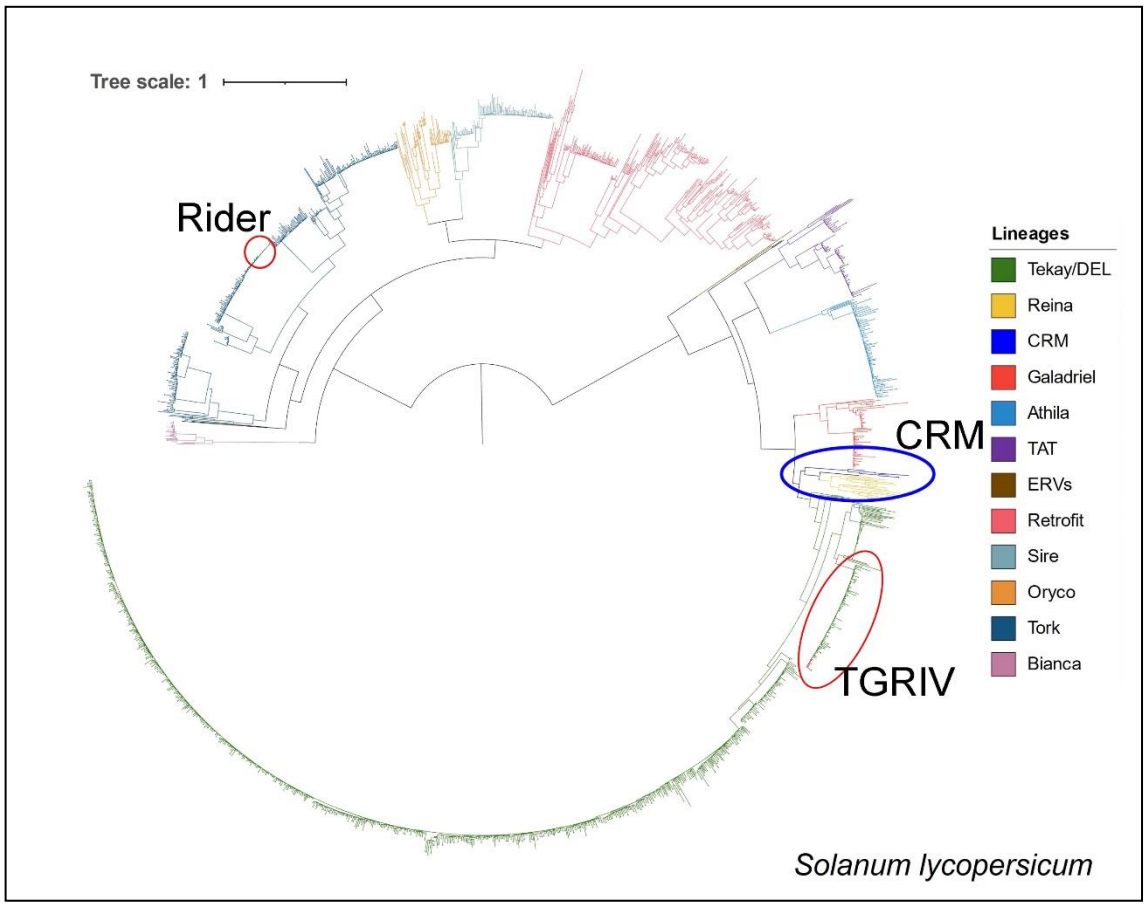


Figure 2: RT domains phylogenetic tree from the full-length elements retrieved in *Solanum lycopersicum* with the EDTA pipeline and annotated with Inpactor. For this tree the branch length was considered. The majority of full-length elements were grouped with Tekay references, the red circle in Tekay branch indicate the clade with TGRIV copies. The blue circle refers to the CRM clade, only the references compose this clade due the lack of full-length sequences from this lineage. Highlight among the Copia elements is the Rider clade, which was grouped with Tork elements.

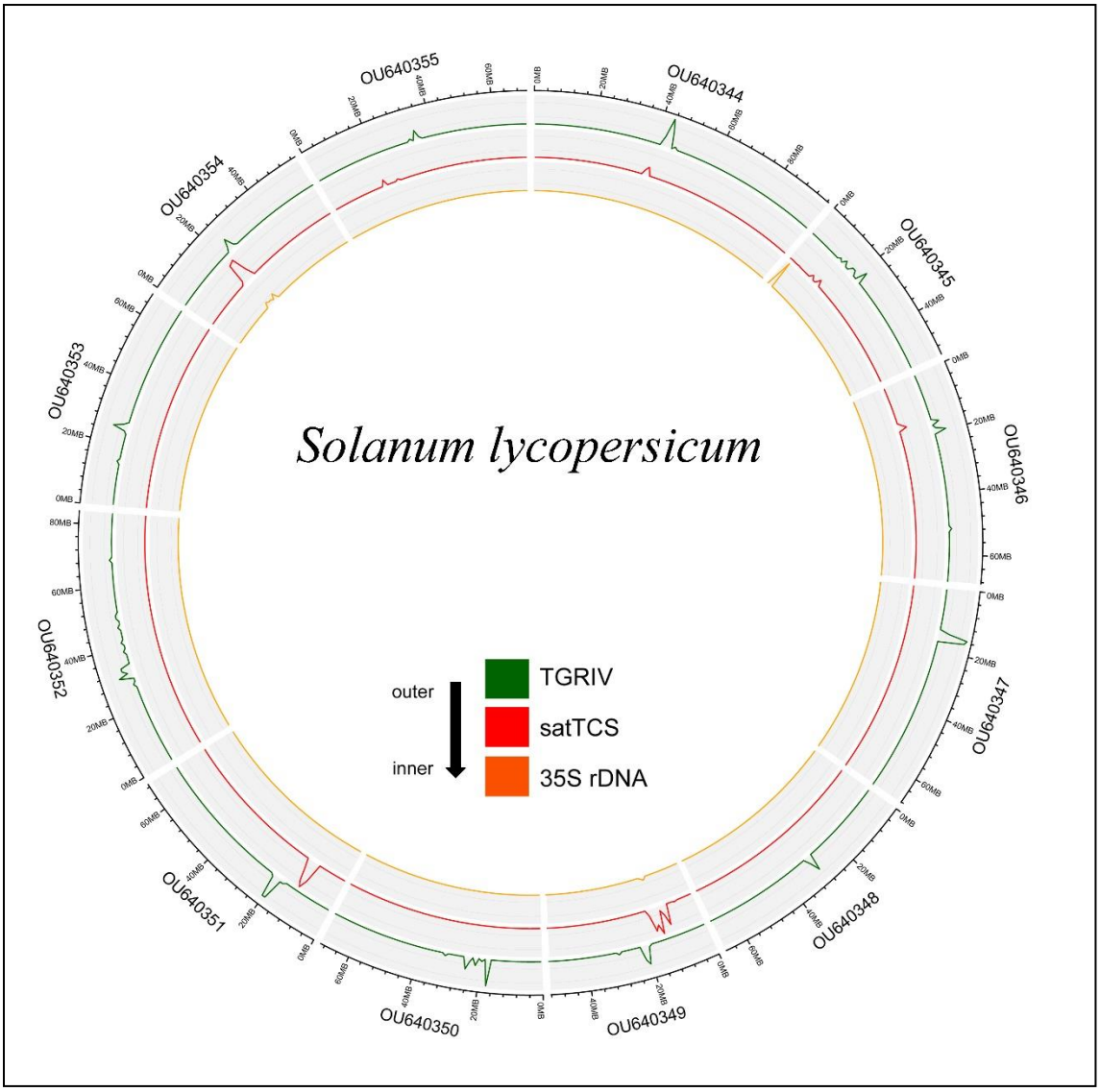


Figure 3: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum lycopersicum* cv. Moneyberg. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS exhibited three major peaks (chromosomes 6, 8, and 11) colocalized with TGRIV peaks, and four minor peaks (chromosomes 1, 2, 3, and 12). The sequences from 35s rDNA exhibited a major peak in the chromosome 2 and two minors in chromosomes 6 and 11.

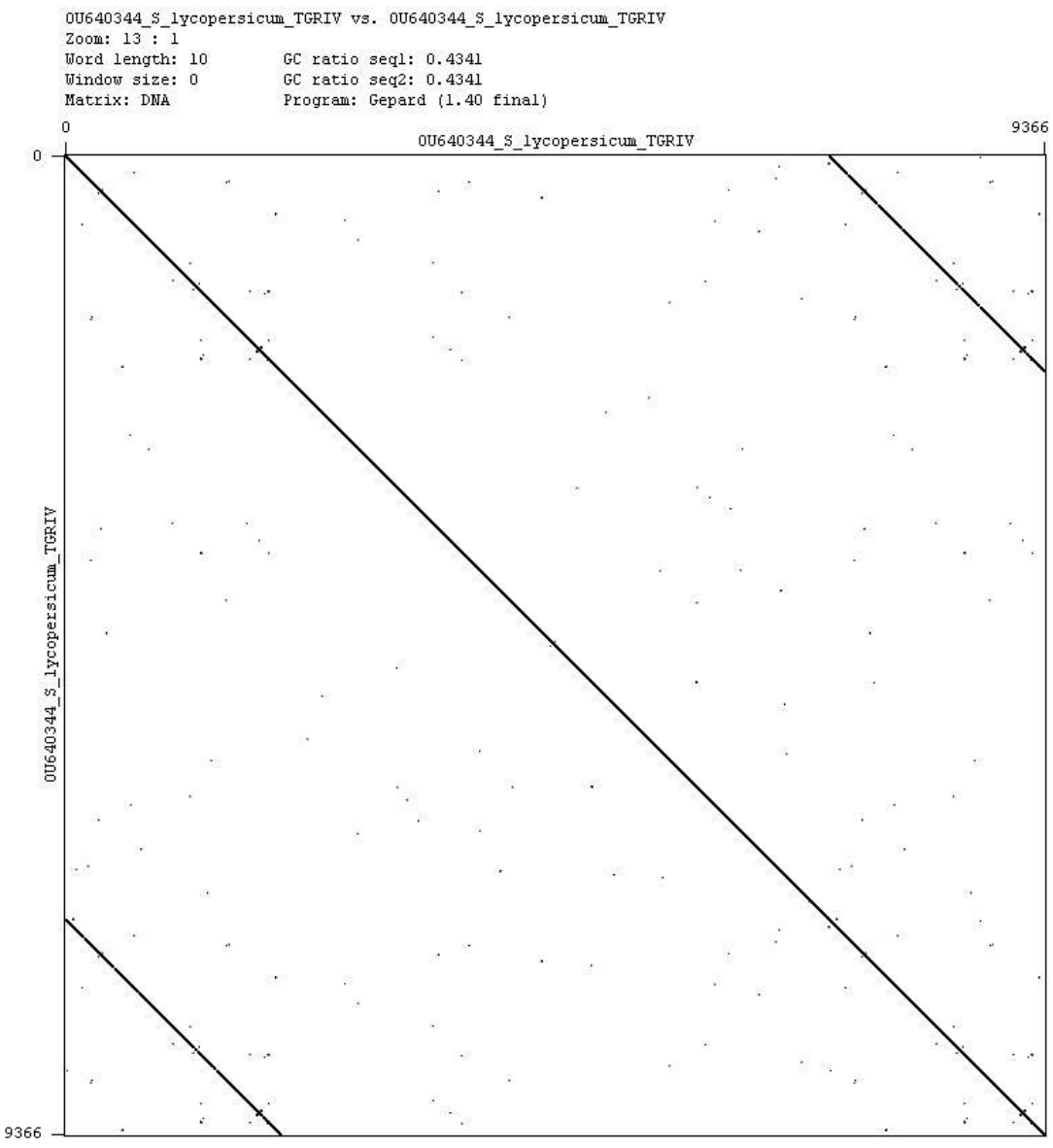


Figure 4: Dot-plot of the TGRIV sequence extracted from the centromeric region of *Solanum lycopersicum.* Note that this element does not carry any sequence similar to satellite.

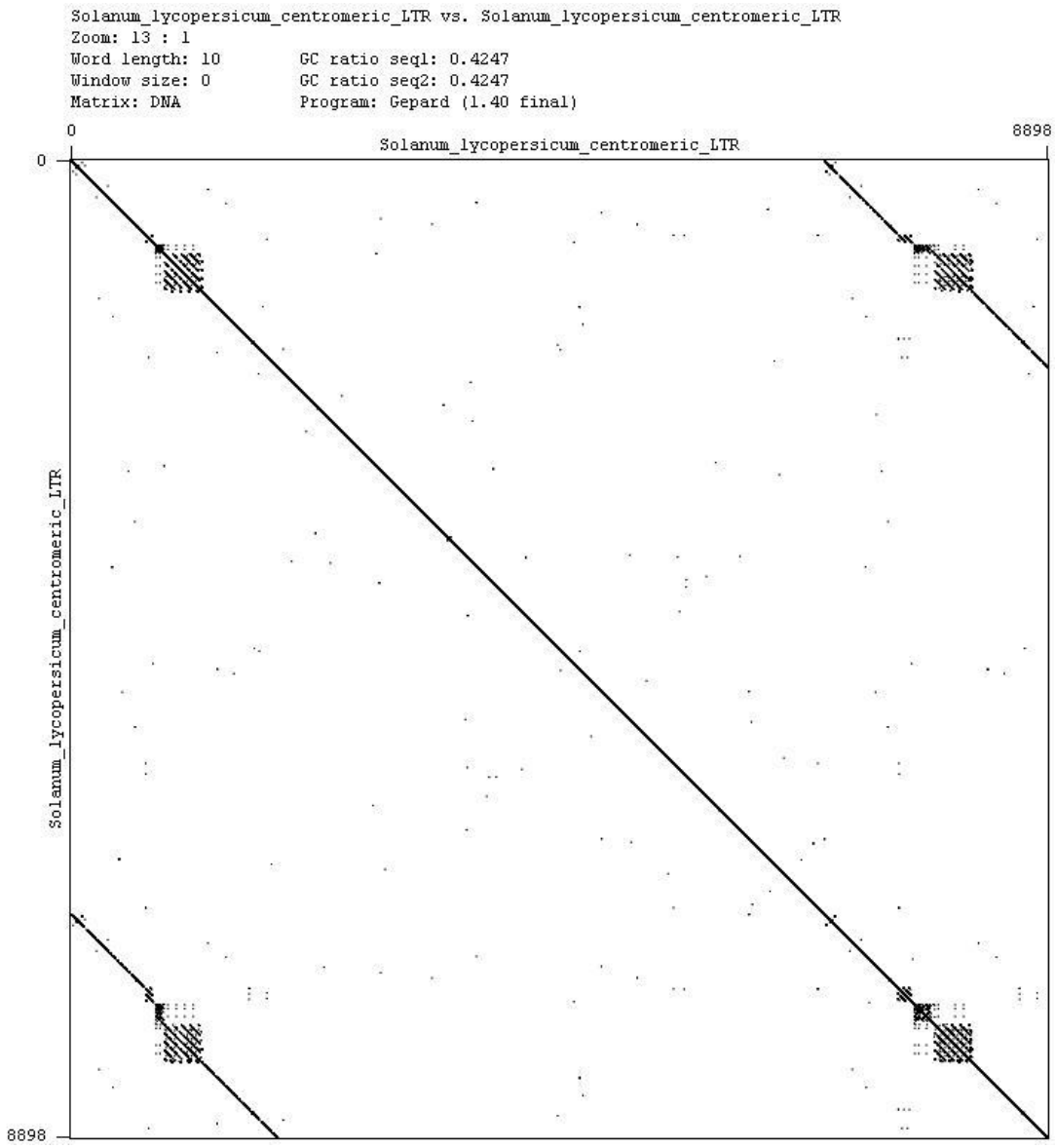


Figure 5: Dot-plot of the centromeric Retrotransposon sequence extracted from the centromeric region of *Solanum lycopersicum.* Note that this element carries a satellite sequence at the LTRs.

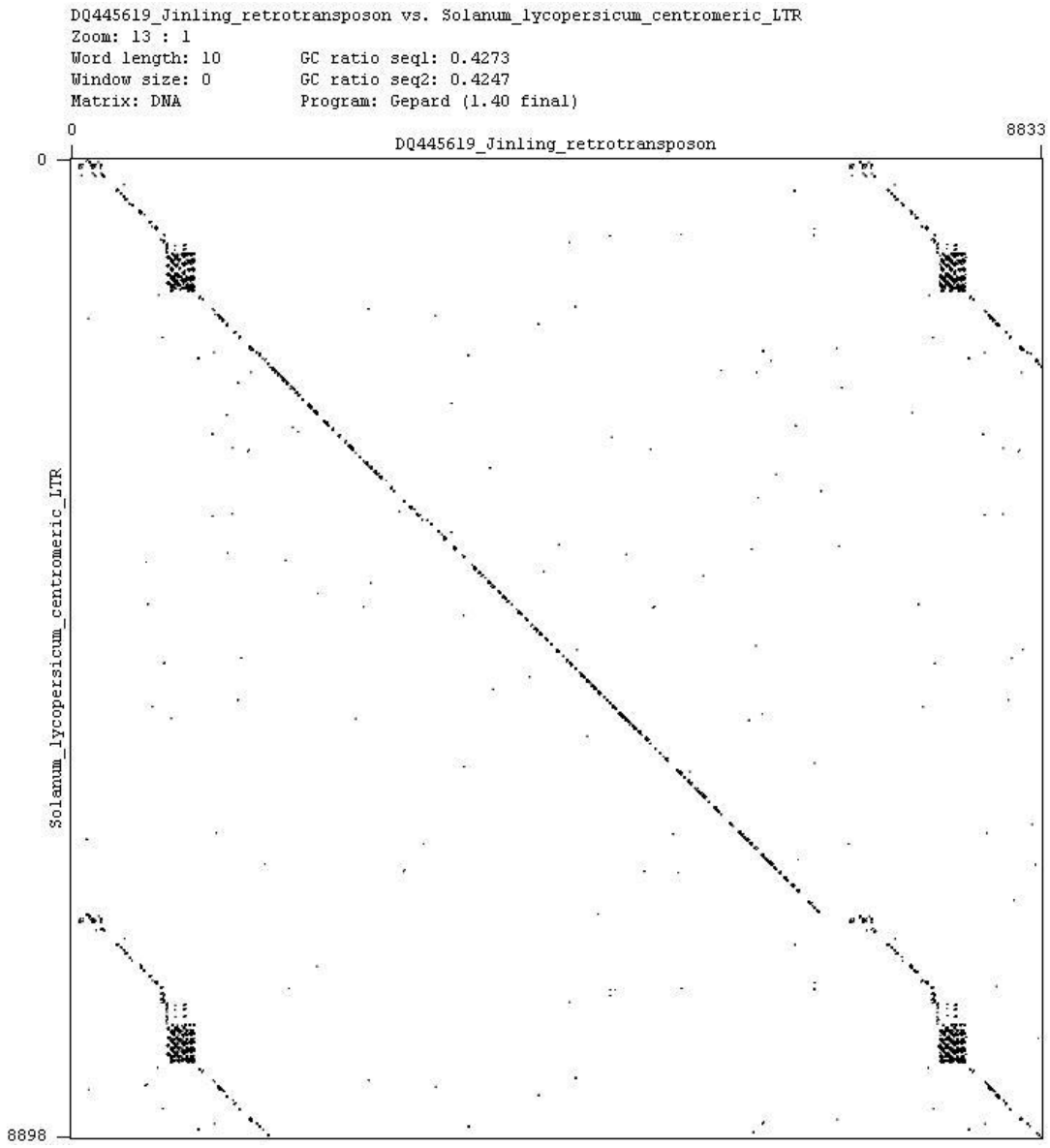


Figure 6: Dot-plot of the centromeric Retrotransposon sequence extracted from the centromeric region of *Solanum lycopersicum* versus the *Jinling* element. Note that even the high similarity between the sequences the sequence, there are some stretches with deletions.

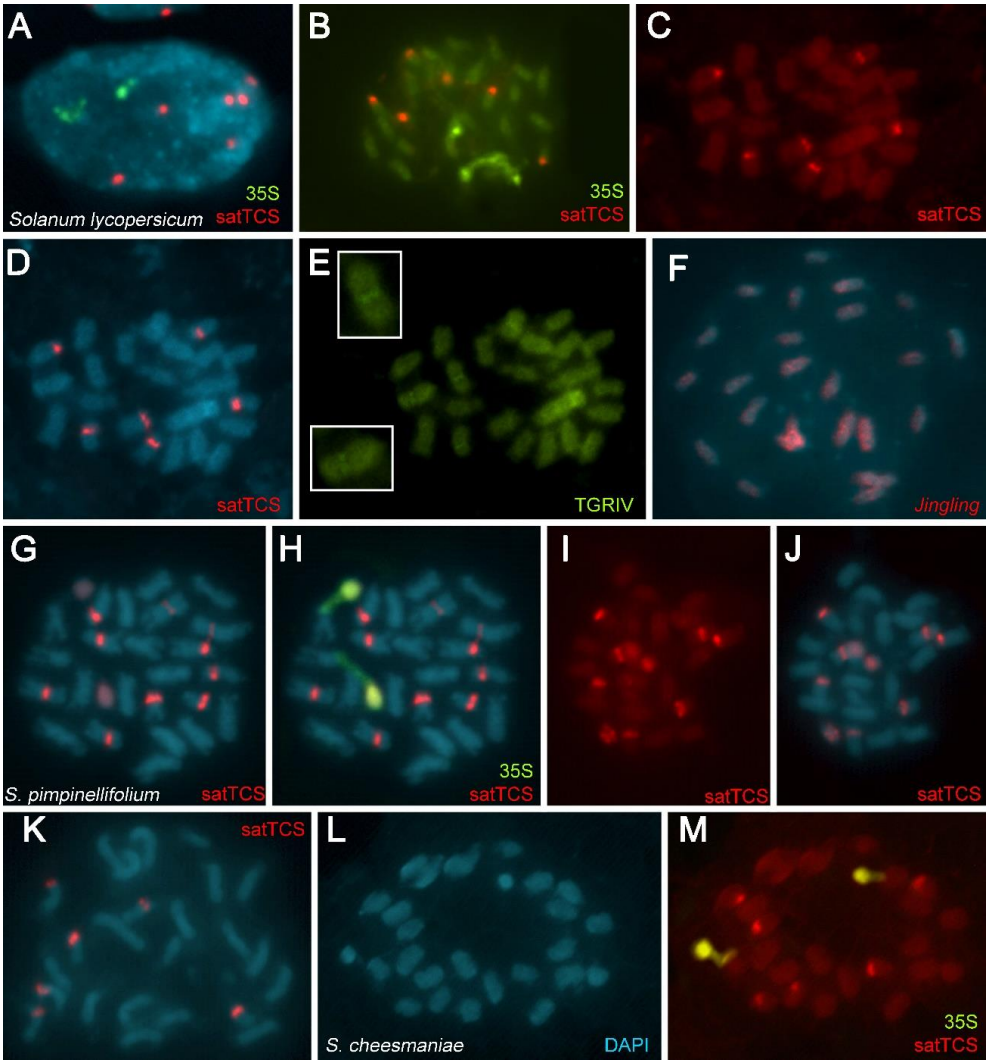


FIGURE 7: FISH assay using satTCS, TGRIV, *Jinling,* and 35S rDNA probes against metaphases and prometaphases of *Solanum lycopersicum* **(A-F)**, *S. pimpinellifolium* **(G-J)**, and *S. cheesmaniae* **(K-L)**. The sample was counterstained with DAPI (blue), and probes were counterstained with Cy3 (red) and avidin–FITC conjugate (green). **(A)** Double FISH in interphasic nucleus of *S. lycopersicum* with satTCS (red) and 35S rDNA (green) probes, **(B)** the same probes in metaphasic chromosomes, and **(C-D)** only the satTCS probe in metaphasic chromosomes. **(E)** TGRIV (green), the boxes highlight the signals present in the centromeric region of chromosomes from *S. lycopersicum*. **(F)** *Jinling* probe in metaphase from *S. lycopersicum*. Note the scattered profile from this probe, typical from the dispersed retrotransposons. **(G, I and J)** FISH in *S. pimpinellifolium* with the satTCS probe, and **(H)** the same species in a double FISH with 35S rDNA probe and the satellite. Note that in this specie the number of chromosomes with signal is bigger the *S. lycopersicum*, and is possible to see minor signals, always in the proximal region. **(K-M)** FISH in *S. cheesmaniae* with the satTCS, and 35S rDNA probes.

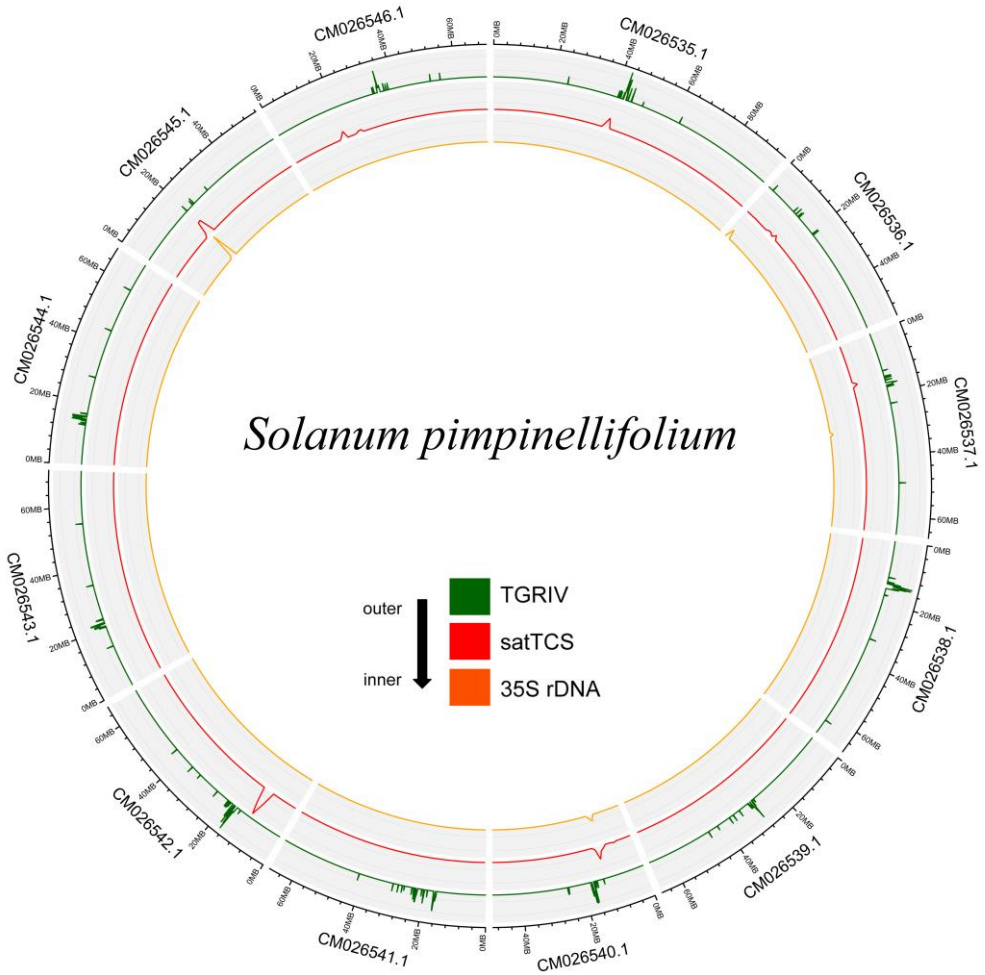
SUPPLEMENTARY MATERIAL

Comparative analysis of retrotransposons among tomato species and the characterization of centromeric elements

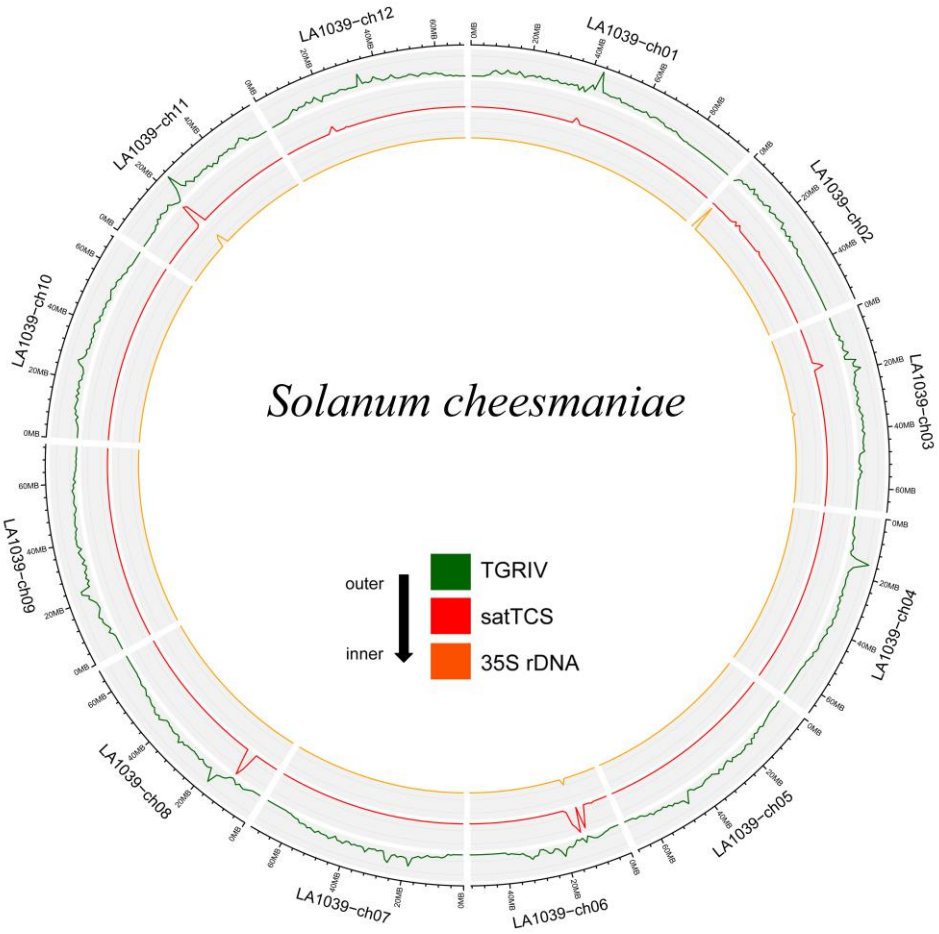
Rafael de Assis[1], Leandro Simões Azeredo Gonçalves[2], Willem M. J. van Rengs[3], Charles Underwood[3], Romain Guyot[4], André Luis Laforga Vanzela[1]*

[1]*Laboratório de Citogenética e Diversidade Vegetal, Departamento de Biologia Geral, Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina, 86097-570, Paraná, Brazil.*
[2]*Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Londrina, 86057-970, Paraná, Brazil.*
[3]*Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany.*
[4]*Institute de Recherche pour le Développement, UMR DIADE, Montpellier, France.*

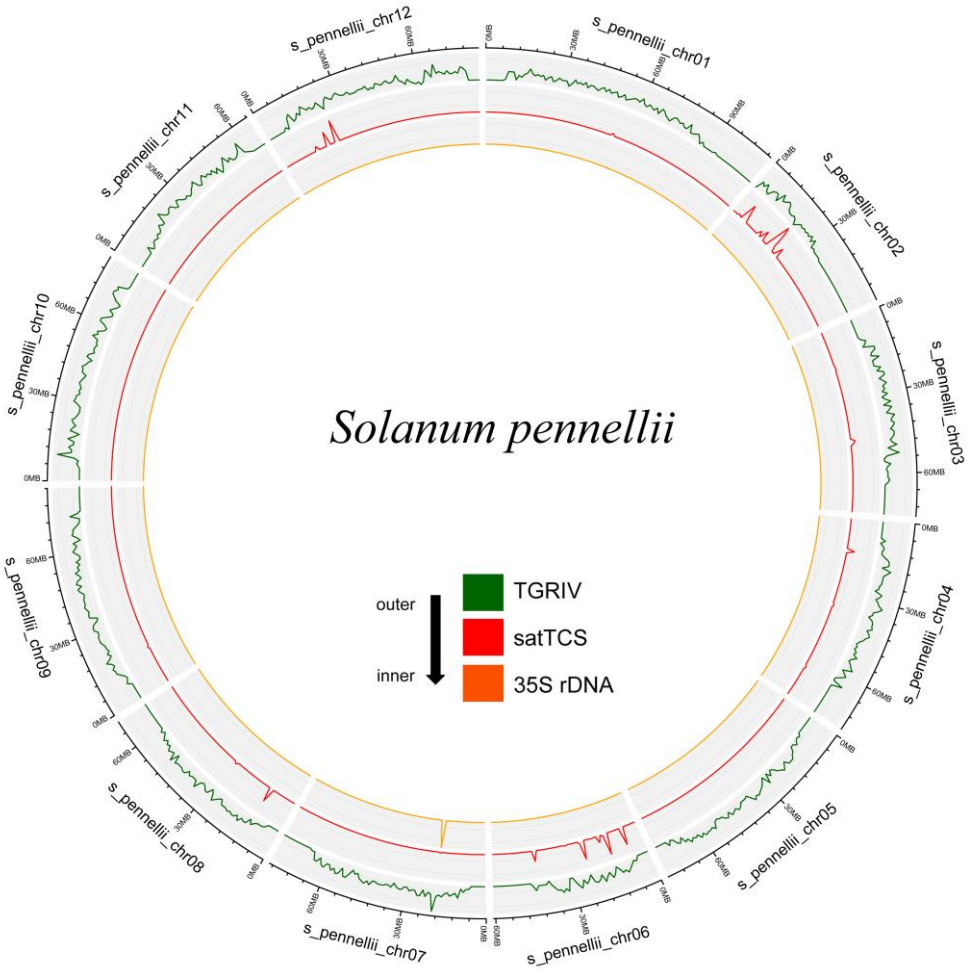
*Correspondent author: E-mail: andrevanzela@uel.br, ORCID: 0000-0002-2442-2211
Rafael de Asis, ORCID: 0000-0002-4420-4588
Leandro S.A. Gonçalves, ORCID: 0000-0001-9700-9375
Romain Guyot, ORCID: 0000-0002-7016-7485
Willem van Rengs, ORCID:
Charles J. Underwood, ORCID: 0000-0001-5730-6279



SUPPLEMENTARY FIGURE 1: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum pimpinellifolium*. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS exhibited three major peaks (chromosomes 6, 8, and 11) colocalized with TGRIV peaks, and four minor peaks (chromosomes 1, 2, 3, and 12). The sequences from 35s rDNA exhibited a major peak in the chromosome 11 and three minors in chromosomes 2, 3, and 6.



SUPPLEMENTARY FIGURE 2: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum cheesmaniae*. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS exhibited three major peaks (chromosomes 6, 8, and 11) colocalized with TGRIV peaks, and four minor peaks (chromosomes 1, 2, 3, and 12). The sequences from 35s rDNA exhibited a major peak in the chromosome 2 and two minors in chromosomes 6 and 11.



SUPPLEMENTARY FIGURE 3: Distribution of TGRIV, satTCS and 35S rDNA sequences in the pseudochromosomes of *Solanum pennellii*. The set of TGRIV sequences exhibited peaks of accumulation in all chromosomes. The satTCS was accumulated in chromosomes 2, 6, 8, and 12 colocalized with TGRIV peaks, and two chromosomes with minor peaks (chromosomes 3, and 4). The sequences from 35s rDNA exhibited a major peak in the chromosome 7.

6 CONCLUSÕES

A família Solanaceae possui espécies com grande importância econômica, entre elas destacam-se as pimentas e os tomates. Ferramentas de bioinformática aliadas a metodologias de citogenética molecular são boas abordagens para a análise da fração de DNA repetitiva presente nos genomas vegetais. Neste trabalho foram utilizadas espécies de pimentas (*Capsicum* L.) e de tomate (*Solanum* L.) como modelos para análise da fração repetitiva das regiões distais bem como das sequências que compõem a região centromérica.

As regiões distais dos cromossomos de *Capsicum* se mostraram ricas em sequências repetitivas, neste trabalho identificamos duas famílias de DNA satélite as quais diferiram quanto a composição e a provável origem das sequências. Estudos prévios relataram a existência de um megasatélite derivado de sequencias de DNAr ocupando as regiões distais de quase todos os cromossomos das espécies de Capsicum, contudo pudemos observar que o satélite aqui descrito como CDR-1 não possui relação com nenhuma sequência já descrita de DNAr. O que traz uma nova perspectiva de interpretação quanto a região distal dos cromossomos das pimentas. O satélite CDR-2 foi encontrado presente em algumas sequências de retrotransposons da linhagem Athila nas três espécies analisadas. Tais elementos apresentaram todos os domínios proteicos necessários para a transposição, contudo essas sequências perderam uma ou ambas as LTRs as quais são partes essenciais do mecanismo de transposição. Elementos que perderam a capacidade de transposição são classificados como elementos não-autônomos.

Regiões cromossômicas proximais são ricas em sequencias repetitivas de diferentes naturezas, o centrômero pode ser composto tanto por sequências com poucas cópias como por sequências altamente repetitivas. Algumas dessas sequências são residentes exclusivamente da região centromérica, como por exemplo o retrotransposon-like TGR4, descrito e mapeado inicialmente em *Solanum lycopersicum*. Tendo como base a localização das sequências TGR4 é possível explorar a diversidade de outras sequências que compõem a região centromérica de espécies próximas ao tomate comercial (*S. lycopersicum*). Utilizando buscas manuais em genomas de alta cobertura, foi possível encontrar outras sequências, compondo a região centromérica. O retrotransposon Jinling, descrito previamente como presente apenas na região pericentromérica, foi encontrado compondo o *set* de sequências centroméricas. Assim como em outras espécies, em espécies de *Solanum* pertencentes ao clado Lycopersicum,

o centrômero apresentou sequências satélite, uma das sequências foi caracterizada aqui neste trabalho, o DNA satélite TCS-1, que apresentou diferença quanto ao número de sinais em diferentes espécies. As sequências centroméricas aqui avaliadas foram mineradas primeiramente no genoma da espécie comercial de tomate (*S. lycopersicum*) e foram caracterizadas em genomas de espécies próximas que compõem o clado dos tomates dentro do gênero Solanum. Tais sequencias demonstram que a região centromérica, mesmo podendo diferir muito entre espécies próximas, mantém um grau de conservação quanto as sequencias que o compõem.

Sequências repetitivas, tais como DNA satélite e retrotransposons, são ferramentas importantes na análise da diversidade genômica, tanto de maneira intraespecífica como interespecífica. Essas sequências repetitivas podem ser usadas como marcadores para inferir a divergência entre espécies próximas, evidenciando as diferenças acumuladas ao longo dos processos evolutivos. Aqui utilizamos espécies de *Capsicum* e *Solanum* como modelos de estudo para analisar a fração repetitiva. Espécies filogeneticamente próximas apresentaram padrões diferentes quanto ao acúmulo de sinais de hibridização das sequencias analisadas. Tais marcas cromossômicas podem ser utilizadas para futuros estudos de evolução cromossômica, assim como estudo de divergência e evolução das sequências repetitivas em espécies próximas.