bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Research Report

Genome-wide analysis of long terminal repeat retrotransposons from the cranberry *Vaccinium macrocarpon*

Nusrat Sultana[a,b]*, Gerhard Menzel[b], Kathrin M. Seibt[b], Sònia Garcia[c], Beatrice Weber[b], Sedat Serçe[d], and Tony Heitkam[b]*

[a]Department of Botany, Faculty of Life and Earth Sciences, Jagannath University, Dhaka 1100, Bangladesh

[b]Institute of Botany, Technische Universität Dresden, D-01062 Dresden, Germany (present address)

[c]Institut Botànic de Barcelona (IBB-CSIC), 08038 Barcelona, Catalonia, Spain

[d]Department of Agricultural Genetic Engineering, Ayhan Şahenk Faculty of Agricultural Sciences and Technologies, Niğde Ömer Halisdemir University, 51240, Niğde, Turkey

Running title: Retrotransposons in cranberry

Words: 7057 (main text) + 2650 (references) + 555 (figure legends)

***Corresponding authors:**

nusrat.sultana1@mailbox.tu-dresden.de, tony.heitkam@tu-dresden.de

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

ABSTRACT

BACKGROUND: Long terminal repeat (LTR) retrotransposons are widespread in plant genomes and play a large role in the generation of genomic variation. Despite this, their identification and characterization remains challenging, especially for non-model genomes. Hence, LTR retrotransposons remain undercharacterized in *Vaccinium* genomes, although they may be beneficial for current berry breeding efforts.

OBJECTIVE: Exemplarily focusing on the genome of American cranberry (*Vaccinium macrocarpon* Aiton), we aim to generate an overview of the LTR retrotransposon landscape, highlighting the abundance, transcriptional activity, sequence, and structure of the major retrotransposon lineages.

METHODS: Graph-based clustering of whole genome shotgun Illumina reads was performed to identify the most abundant LTR retrotransposons and to reconstruct representative *in silico* full-length elements. To generate insights into the LTR retrotransposon diversity in *V. macrocarpon*, we also queried the genome assembly for presence of reverse transcriptases (RTs), the key domain of LTR retrotransposons. Using transcriptomic data, transcriptional activity of retrotransposons corresponding to the consensuses was analyzed.

RESULTS: We provide an in-depth characterization of the LTR retrotransposon landscape in the *V. macrocarpon* genome. Based on 475 RTs harvested from the genome assembly, we detect a high retrotransposon variety, with all major lineages present. To better understand their structural hallmarks, we reconstructed 26 Ty1-*copia* and 28 Ty3-*gypsy in silico* consensuses that capture the detected diversity. Accordingly, we frequently identify association with tandemly repeated motifs, extra open reading frames, and specialized, lineage-typical domains. Based on the overall high genomic abundance and transcriptional

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

activity, we suggest that retrotransposons of the Ale and Athila lineages are most promising to monitor retrotransposon-derived polymorphisms across accessions.

CONCLUSIONS: We conclude that LTR retrotransposons are major components of the *V. macrocarpon* genome. The representative consensuses provide an entry point for further *Vaccinium* genome analyses and may be applied to derive molecular markers for enhancing cranberry selection and breeding.

Keywords: cranberry, *Vaccinium macrocarpon*, repetitive DNA, LTR retrotransposon, Ty1-*copia*, Ty3-*gypsy*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

INTRODUCTION

The genus *Vaccinium* L. belongs to the family Ericaceae Juss. comprising approx. 450 species distributed all over the world [1]. Available molecular phylogenies suggest that the genus is monophyletic, whereas intrageneric relationships are more difficult to resolve [2, 3]. Dependent on different systematic treatments, the genus is divided into approximately 30 sections and subgenera [1-4]. *Vaccinium macrocarpon* Aiton, also known as the American cranberry, is native to North America [5, 6] and is one of the most commonly cultivated *Vaccinium* species, along with the highbush blueberry (*V. corymbosum* L.) [5]. The small berry fruits produced from wild and cultivated species of the genus are rich in nutritious secondary plant metabolites, including some beneficial for human health with anticancer, antioxidant, antidiabetic, and many other properties [7, 8]. The American cranberry and the European cranberry (a close relative, *V. oxycoccus* L.) belong to the same section or subgenus, named *Oxycoccus* [1, 4]. Although these two species differ in geographical distribution, they can produce fertile offspring upon hybridization [5, 6].

Vaccinium species have ploidy levels ranging from diploid to hexaploid, with a base chromosome number x = 12 [5]. As cranberry (*V. macrocarpon*) is a diploid and self-fertile species with a relatively small genome size (approximately 470 Mb), it has become a valuable reference to study the genetics and genomics of the genus [6, 9-11]. So far, whole genome sequence data, a reference genome assembly, and transcriptome datasets have been published for the diploid cranberry and for both the diploid and the tetraploid blueberries [11-13]. In addition, several high-throughput genetics, genomics, and transcriptomics datasets as well as other important breeding information of different species of the genus *Vaccinium* are either published or underway, accessible through the web platform https://www.vaccinium.org.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Recent technological advancements allow tracing of genome dynamics on a fine-scale level. Especially transposable elements (TEs) play a major role in genome evolution and contribute largely to the generation of polymorphisms between genotypes [14-18]. Hence, TEs can also be applied as molecular markers for the differentiation of accessions and to inform plant breeding programs [19-23]. Thus, the correct annotation and characterization of repetitive sequences, particularly TEs, are a prerequisite for an enriched genome assembly as well as for subsequent genome characterization and valorization [10-12]. Nevertheless, as many different TE classification hierarchies have been suggested [24-29], as an integration of tools, databases, and pipelines is still discussed [30], and as there is still no gold standard for TE identification and classification available, TE annotation and classification for non-model species is still a considerable undertaking.

In plants, the most abundant TEs belong to the group of long terminal repeat (LTR) retrotransposons contributing up to 80 % of nuclear DNA [31-33]. They consist of two LTRs flanking the main protein-encoding genes *gag* and *pol* [25, 33, 34]. Whereas *gag* (encoding the capsid and nucleocapsid) is considered as a single gene, the *pol* gene generally encodes several domains, including a protease (PR), an RNase H (RNH), a reverse transcriptase (RT), and an integrase (INT) [25, 33, 34]. In addition to the protein domains, two conserved motifs are characteristic for these full-length elements: a primer binding site (PBS) and a polypurine tract (PPT) [25, 31, 33, 34]. Based on the order of the protein domains and their sequence, the LTR retrotransposons are further divided into the superfamilies Ty3-*gypsy* and Ty1-*copia* [24, 28].

In *Vaccinium* genomes, we have already gained first insights into the repetitive DNA composition, especially into that of satellite DNAs [35, 36]. For dispersed repeats, first estimates were also brought forward, suggesting that nearly 40 % of the cranberry genome

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

was composed of TEs [10]. Very recently, a TE annotation was also made available for the genome of tetraploid blueberry (*V. corymbosum*) [13]. However, the contribution of individual TE families to these genomes is still not well-studied [10-12].

Here, we aim to complement the broad overview studies with a deep analysis of the most abundant, individual TE families in the cranberry genome. The precise knowledge of the LTR retrotransposon composition, as well as the underlying sequences and structures can inform annotation of berry genomes and may provide support for berry breeding and selection.

MATERIALS AND METHODS

Graph-based clustering of repeat sequences of the *V. macrocarpon* genome

To representatively survey the cranberry LTR retrotransposon landscape, we used publicly available whole genome paired-end Illumina reads of *Vaccinium macrocarpon* (NCBI BioProject PRJNA245813) of cranberry cultivar 'Ben Lear' (accession number CNJ99-125-1) [10]. Before clustering, we pre-treated the reads by quality filtering, adapter trimming, and processing of read length: Illumina Truseq adapters were removed using Trimmomatic with the parameters ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 [37]. Fastx_trimmer and FASTx quality filter from FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) were used to trim all reads to 150 bp length. Shorter sequences remaining after quality filtering and trimming, were removed with seqtk (http://github.com/lh3/seqtk). This was followed by interlacing of paired-ends with the FASTX-Toolkit. We then randomly selected 2× 1M, 2× 5M, and 2× 10M reads representing different genome coverages ranging from 0.02× to 2.04×. For read clustering, we used the Galaxy web interface of the RepeatExplorer pipeline2 [38, 39] with the parameters -l 39 (minimal overlap of clustering) and -o 30

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(minimum overlap for assembly), considering the different genome coverages. The generated read clusters were further assigned to the retrotransposon lineages according to their similarity to reference elements in REXdb [28] and by RepeatMasker (http://www.repeatmasker.org, [40]). Based on the cluster graph shape, protein domains and repeat masker hits, each cluster was manually annotated and assigned to superclusters from the RepeatExplorer run of the highest genome coverage, which was expected to generate the highest number of full-length elements.

Construction and classification of full-length LTR retrotransposons

To reconstruct full-length Ty1-*copia* and Ty3-*gypsy* retrotransposon sequences, several steps were followed: First, clusters and superclusters belonging to the individual retrotransposon types were identified based on their graph shape and protein domain hits. Second, contigs from the identified clusters were imported into Geneious Prime 2019, hereafter called Geneious (http://www.geneious.com, [41]). Third, we used the longest contig sequence as reference and mapped the remaining contigs against it using the Geneious mapper tool (with interation parameter = 10). We extracted the consensus sequence from the mapped contigs. Afterwards, short reads from the respective cluster were mapped against the consensus contig sequence to derive the final consensus sequence. If the identified LTR retrotransposon sequence was represented by only a single RepeatExplorer cluster, the consensus sequence was directly used for the reconstruction of the full-length element. However, if the LTR retrotransposon was split across multiple clusters, additional steps were considered for the reconstruction of the full-length element: These steps included iterative multiple sequence alignments of the consensus sequences of all clusters belonging to a single supercluster in Geneious. The parameters for the Geneious alignment tools were: automatically determine

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

direction, alignment type = global alignment, cost matrix = 65 % similarity (5.0/-4.0), gap open penalty = 12, gap extension penalty = 3, refinement iterations = 2. Then, short read sequences from the respective supercluster were mapped against the reconstructed consensus sequence to derive the final full-length sequences using the Geneious mapper tool (with iteration parameter = 10) within Geneious.

The reconstructed full-length sequences were searched for the typical structural features of LTR retrotransposons, namely protein domain sequences, LTR regions, PBS, PPT, and internal repetitions. Different software and databases were used for these purposes: the LTR_Finder web server ([42]; http://tlife.fudan.edu.cn/tlife/ltr_finder/) was applied to predict common structural features along the reconstructed LTR retrotransposons, using the default parameters except "predict PBS by using tRNA database", which was variable for different elements. Dotplot analyses were performed using the default parameter of the EMBOSS dottup tool integrated in Geneious [41, 43]. These dotplots showed the positions of LTRs as short diagonals in the upper right and lower left corners. Internal tandemly repeated structures lead to regular patterns of short lines in the dotplot. Furthermore, the REXdb database underlying RepeatExplorer's protein domain finder tool (default parameters) was used to identify TE protein domains along the query elements and to predict their position and direction [38, 39]. Finally, the initial results from the protein domain finder tool were quality-filtered using the default parameters. For the Tandem Repeat Finder (TRF) tool the following parameters were used: alignment (match, mismatch, indel=2,7,7; 2,5,7; 2,5,5; 2,3,5), minimum alignment score to report repeats (20-60), and minimum period size (10-300) [44]. The TRF output was manually inspected, compared to the dotplot findings, and recorded for different features like period size, consensus sequence, number of repeat units, and position within the full-length LTR retrotransposon.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The online databases used to characterize and classify each full-length sequence were The Gypsy Database (GyDB) [25], Repbase [27], and REXdb database [28]. Repbase was used from the web-based software censor (https://www.girinst.org/censor/index.php) and GyDB from the online server (http://gydb.org/index.php/Main_Page). The REXdb database is integrated in RepeatExlorer2 (https://galaxy-elixir.cerit-sc.cz/).

The reconstructed full-length sequences were named according to the conventions and deposited in the Repbase database. Detailed information is summarized in Supplementary File 1 and Tables S1.

Assignment of retrotransposon lineages and clades

A comparison against a database of well-known protein domains serves as the basis of a detailed retrotransposon classification. For this, protein sequences of reverse transcriptases (RTs) and, if applicable, chromodomains (CHD) were extracted from the publicly available genome assembly for *V. macrocarpon* (GenBank assembly accession: GCA_000775335.1) using the protein domain finder tool DANTE implemented in the RepeatExplorer Galaxy web-interface [38, 39]. DANTE was used with the quality filtering (minimum identity = 0.3; minimum similarity: = 0.4; minimum alignment length = 0.8; interruptions = 3). The program output 2149 sequences which were then clustered with a 90 % similarity threshold using CD-HIT ([45] http://weizhongli-lab.org/cd-hit). RT and CHD amino acid sequences were aligned with MAFFT [46] within Geneious (parameters: algorithm = auto [selects appropriate strategy from L-INS-I, FFT-NS-i and FFT-NS-2 according to data size], scoring matrix = “200PAM/K=2”, gap open penalty = 1.53, offsetvalue = 0.123). Pairwise distance matrix from the RT alignment was calculated in Geneious for boxplot visualizations, which were created with R graphics [47].

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The resulting multiple sequence alignments of RT and CHD sequences were visualized as a dendrogram using the approximate maximum likelihood estimation algorithm of the FastTree tool [48] and the Randomized Axelerated Maximum Likelihood RAxML method [49] implemented in Geneious (default parameters). This allowed assigning the detected LTR retrotransposons to published lineages and clades according to their encoded RT and CHD sequences.

Analysis of *cis*-regulatory element associated with the 5' LTR of the full-length LTR retrotransposons

We extracted the 5' LTRs of the identified Ty1-*copia* and Ty3-*gypsy* full-length elements. Ambiguous nucleotides from each of the 5' LTR sequences were replaced by an N using the software "Sequence Manipulation Suite" [50] (http://www.bioinformatics.org/sms2/filter_dna.html). For the identification of regulatory sequence motifs, 5' LTR sequences were individually searched against the Plant Care database [51] (http://bioinformatics.psb.ugent.be/webtools/plantcare/html), a database on *cis*-acting regulatory elements in plants.

All-against-all comparison of full-length LTR retrotransposon sequences

To detect the sequence similarity within lineages, and also for fine-scale classification of the phylogenetically closely related sequences, all reconstructed *V. macrocarpon* full-length LTR retrotransposons were subjected to an all-against-all comparison together with publicly available reference sequences (Table S1). Representative reference sequences for each LTR retrotransposon lineage from different plant genomes were downloaded from GyDB 2.0 [25], (Table S2). The 5' LTRs from these lineages were extracted using Geneious [41]. LTR

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

regions were aligned in Geneious and a distance matrix of pairwise LTR identities was exported. For each lineage, an all-against-all sequence comparison was created using the Python-based software FlexiDot [52]. The FlexiDot parameters were: word size of 26 (-k 26) with 5 mismatches allowed (-S 5). In the dotplot, pairwise sequence identities of the 5'LTRs were printed and shaded.

Transcriptome analysis

To assess retrotransposon transcription, publicly available *V. macrocarpon* paired-end mRNA-seq data (Illumina GAIIx; 63.6 million 100 bp reads) from leaves and shoot tips of cranberry cultivar 'Ben Lear' (accession number CNJ99-125-1) of BioProject PRJNA246586 were used [10]. All reads were quality-filtered using the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) within Galaxy [53] with the parameters "quality cut-off" = 10, "percent above cutoff" = 95. Adapter trimming was performed using the BBDuk Geneious plugin [41, 54] with default parameters. About 36.1 million quality-filtered paired-end mRNA reads were used for read mapping against the reconstructed *V. macrocarpon* full-length elements of the Ty1-*copia* and the Ty3-*gypsy* superfamily. Transcriptome reads were mapped with the 'Map to reference' tool in Geneious [41] with the parameters: "mapper" = Geneious for RNA-Seq; "sensitivity" = medium-low sensitivity; "span annotated mRNA introns". The graph coverage of the mapped reads was exported and the transcriptome proportion was calculated the counted reads in Excel. A scatter plot visualization was created to compare the genome proportion (derived from the RepeatExplorer analysis) and the transcriptome proportion of the representative full-length Ty1-*copia* and Ty3-*gypsy* LTR retrotransposons [47].

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Data availability

Full-length retrotransposon sequences and clustered core RT and CHD sequences from this analysis were deposited in the GDV database with the accession number GDV19002 (https://www.vaccinium.org/publication_datasets).

RESULTS

Composition of the *V. macrocarpon* retrotransposon landscape

The graph-based clustering with the RepeatExplorer2 pipeline reveals that about 91% of cranberry genome is repetitive (Fig. S1, [35]). According to the automated read cluster classification, only 49% of total repeat reads were assigned to known repeat types. Out of those, LTR retrotransposons constitute the major portion (44%), followed by DNA transposons (11%), non-LTR retrotransposons (6%), rDNAs (3%), satellite DNAs (0.06%) and other repeat fractions (Fig. S1, [35]). The *in silico* analysis shows that full-length LTR retrotransposons constitute a considerable portion of cranberry genome of 4.25 % Ty1-*copia* and 8.41 % Ty3-*gypsy* retrotransposons (Fig. S1; Table S1).

For a comprehensive overview of the LTR retrotransposon population, we identified the reverse transcriptase (RT) domains in the *V. macrocarpon* reference genome assembly. We queried the *V. macrocarpon* assembly and identified 181 Ty1-*copia* and 294 Ty3-*gypsy* RT instances. The highest number of RT sequences belonged to the Ale- (n = 110, Ty1-*copia*) and Tat- (n = 184, Ty3-*gypsy*) types (Table S3).

Maximum-likelihood analyses allowed us to verify the classification of the reference full-length elements for both superfamilies. For the Ty1-*copia* retrotransposons, we noticed a close relationship between five well-defined lineages: Ikeros, Bianca, Tork, TAR, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Angela. A close relationship of Ivana and SIRE sequences was also observed. Although Alesia only harbors two members, it forms a single, well-separated branch. Finally, the 110 Ale RT sequences are polyphyletic and strongly diversified (Fig. 1A). In the superfamily Ty3-*gypsy* (Fig. 2), the non-chromoviruses and chromoviruses are similarly organized. The non-chromovirus retrotransposons of the Tat type were the most abundant and the most diversified. The chromovirus lineage produced four branches representing the clades CRM, Reina, Galadriel, and Tekay (Fig. 2A), with a clade of non-chromovirus Athila as sister.

The separation of the chromoviruses into clades could be verified by reconstructing a dendrogram based on their signature chromodomain (CHD) sequences (Fig. S2). Although full-length Reina elements were not found in the RepeatExplorer-based reconstructed retrotransposon sequences (Table S4), the CHD search along the genome assembly of the same *V. macrocarpon* accession, reveals that CHD protein sequences are indeed present from all four chromovirus clades, i.e., CRM, Galadriel, Tekay, and Reina (Fig. S2).

Although branch lengths in the dendrogram provide information on the relative divergence between retrotransposon groups, the boxplot visualizations allow a direct comparison of the respective RT diversities (Figs. 1B, 2B). For this, pairwise identities from the RT amino acid multiple sequence alignments were utilized. In Ty1-*copia* lineages, Ale RT sequences have the highest diversity, as they exhibit the widest spread with a maximum of 90 %, a minimum of 28 %, and a median pairwise identity of 59 %. The lowest median identity (57 %) is observed for Ivana, whereas the highest conservation is observed for Bianca with a median pairwise identity of 84 % (Fig. 1B). Generally, Ty1-*copia* RT sequences are more diverse than those from Ty3-*gypsy*. Regarding the median RT identity, the most diverse clade in the Ty3-*gypsy* superfamily is Tat (61%), whereas Reina is the least diversified (73 %). The remaining clades of the Ty3-*gypsy* superfamily (Athila, CRM, Galadriel, and Tekay) have

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

similarly diverse RT sequences, with median RT identities ranging between 65 % and 74 % (Fig. 2B).

Structural diversity of Ty1-*copia* and Ty3-*gypsy* LTR retrotransposons

Based on the clustered short reads, we reconstructed and characterized 26 Ty1-*copia* and 28 Ty3-*gypsy* full-length retrotransposons, representative for 54 repeat families in the genome of *V. macrocarpon* (Table 1). The majority of these full-length elements (40 out of 54) was represented by a single, circular RepeatExplorer cluster in our analysis (Fig. S3). Reads from the remaining Ty1-*copia* and Ty3-*gypsy* elements were split into smaller clusters. Due to overlapping read pairs, we were able to connect these smaller clusters to superclusters spanning retrotransposons of full length (Table S1). Full-length consensuses were reconstructed from eight Ty1-*copia* lineages (Ale, Alesia, Bianca, Ivana/Oryco, TAR, Angela, Tork, and Sireviruses/Maximus, the last being incomplete) and three Ty3-*gypsy* lineages (non-chromoviruses Tat and Athila, and chromoviruses). Typical structural features, including sequence lengths, coding domains and conserved sequence motifs, were analyzed in detail to understand their structural diversity (Figs. 3, 4; Fig. S4; Table 1; Table S5).

To illustrate the structure of the individual Ty1-*copia* lineages, we selected eight representative elements (Fig. 3; Fig. S4). Among all the Ty1-*copia* lineages, we reconstructed seven full-length Tork, five Ale, two Alesia, one Bianca, three Ivana/Oryco, and one Sireviruses/Maximus (Table 1; Table S5): The longest consensus element belongs to the latter group and is about 7 kb, it is partially incomplete, containing only the 5' LTR (1545 bp). Other lineages have retrotransposon lengths ranging between 4.8-6.8 kb with LTR lengths between 127-994 bp (Table 1; Table S5). While the highest genome proportion among the reconstructed Ty1-*copia* elements belongs to the elements SCL22_Ale (0.57 %),

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the lowest genome proportion is observed for SCL344_Alesia (0.01 %) (Table S5). Although all full-length *in silico* Ty1-*copia* elements contain uninterrupted *gag* and *pol* protein domains, some characteristic retrotransposon features were absent. For example, while most of the *in silico* elements that represent Ale, Angela and Tork retrotransposons harbored the PBS, PPT, and both 5' and 3' LTRs, one of these structural motifs was missing in the representative *in silico* sequences of Bianca, Ivana/Oryco, TAR, and Sireviruses/Maximus (Table 1; Table S5). The most common PBS-type detected in Ty1-*copia* lineages corresponds to the Met-type. For Ale and TAR retrotransposons, however, a PBS of the Asn/Met-type (Ale) or Glu/Ser/Met/Tyr/Trp-type (TAR) was more common (Table 1; Table S5).

Six reference elements were selected to represent the structural variability of each Ty3-*gypsy* lineage (Fig. 4). The most diversified and abundant Ty3-*gypsy* elements in the *V. macrocarpon* genome belong to the Tat group (Table 1). A total of 13 full-length elements have been characterized, with genome proportions ranging between 0.12 and 1.34 % (Table 1). In contrast, the other two groups only contribute 0.01-1.92 % (Athila) and 0.02-0.23 % (chromoviruses) to the genome (Table 1; Table S5). Although *gag* and *pol* genes were detected for all full-length elements, structurally complete elements with all motifs typical for LTR retrotransposon (PBS, PPT and both 3' and 5' LTR regions) were mostly present in the Galadriel clades (chromoviruses). Most elements of the other groups (chromovirus Tekay, CRM and non-chromoviruses Tat and Athila) lack at least one of these structural motifs. Moreover, some Ty3-*gypsy* elements have additional protein domains as compared to Ty1-*copia* elements. For instance, the *pol* genes of the non-chromovirus Tat clade have dual ribonuclease H domains (gRNH and aRNH), whereas the chromoviruses harbor an additional chromodomain (CHD) region (Fig. 4; Table S5). A canonical CHD was only found for the elements of the Tekay and the Galadriel clades, but surprisingly not within the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

elements of the chromovirus CRM clade (Fig. 4; Fig. S4; Table S5). SCL16_Ogre is the longest Ty3-*gypsy* Tat element, with a total length of 19 kb and 1251-1326 bp long LTRs. In contrast, SCL80_CRM is the shortest element (5.2 kb) with LTRs ranging between 361-556 bp (Table S5). A Met-type PBS was only found in the CRM and Galadriel chromoviruses, whereas other Ty3-*gypsy* elements were more variable regarding their PBS (Table 1).

Cranberry retrotransposons are often associated with short tandem repeats

Although tandem repeats, i. e. satellite or ribosomal DNAs, represent major repetitive genome fractions that are distinct from the more dispersed transposable elements, tandemly repeated motifs can also be associated with retrotransposons. In *V. macrocarpon*, we detected tandem repeats in the reconstructed full-length Ty1-*copia* and Ty3-*gypsy* retrotransposon sequences (Table 2; Table S6). These tandem repeats were either localized within the 3' or the 5' LTRs, within the untranslated regions (UTRs) adjacent to either LTR, or in some cases within the UTRs adjacent to the *gag* gene. Tandem repeats found in different retrotransposons vary in sequence composition, period size, and number of repetitions (Table 2; Table S6). For the Ty1-*copia* superfamily, we identified tandem repeats in all lineages, embedded in either the LTRs or UTRs, with consensus period lengths ranging from 2 to 49 bp with about 2 to 8 repetitions (Table 2; Table S6).

In Ty3-*gypsy* LTR retrotransposons, tandem repeats were detected both in non-chromoviruses and chromoviruses, yet to a different extent. Non-chromovirus Tat and chromoviruses harbored tandem repeats both within the LTRs and UTRs, whereas non-chromovirus Athila had tandem repeats only embedded within the UTR adjacent to the *gag* gene (Table 2). Among all Ty3-*gypsy* LTR retrotransposons, Tat (including Ogre) harbor tandem repeats with the largest monomer size (up to 93 bp) and the highest number of

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

repetitions (up to 15 times) (Table 2; Table S6).

Cis-regulatory elements are often associated with LTR sequences

As LTRs frequently harbor *cis*-regulatory elements, our targeted search revealed that almost all reconstructed retrotransposons include typical promoter motifs, including TATA and CAAT boxes within their LTRs (Fig. S5A-B; Table 2; Table S7). However, as the LTR composition varies, number and position of these promoter sequences are variable. For instance, the LTRs of the Ty1-*copia* lineages Ale, Angela, TAR, Tork, Sireviruses/Maximus, and all Ty3-*gypsy* lineages contained multiple promoter motifs in different positions. However, the Bianca LTR only contains a CAAT box, lacking a TATA box. For Ivana, two out of three *in silico* elements (SCL42_Ivana and SCL310_Ivana) contain both promoter motifs in their LTRs (Fig. S5A-B; Table S7).

In addition to the promoter motifs, we detected other *cis*-regulatory elements for different LTR retrotransposons (Table S7). These regions are mostly related to light, hormone, and cell cycle responsiveness, as well as abiotic stress tolerance (Fig. S5A-B; Table S7). Among the eight representative sequences of the Ty1-*copia* lineages, most regulatory motif types were predicted for TAR (19), Tork (19), and Ale (10) (Table 2), whereas among the Ty3-*gypsy* elements, the Tat retrotransposons (including Ogre) harbor most variability with 38 distinct motifs.

Sequence homogeneity between closely related elements

In an attempt to fine-scale the classification of the full-length retrotransposon elements of *V. macrocarpon*, closely related elements were subjected to an all-against-all sequence

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

comparison with publicly available references to reveal sequence similarities among them (Table S2).

Overall, the *V. macrocarpon* full-length elements have higher similarity to other elements within their protein coding regions, contrasting the variability within their UTRs and LTRs (Fig. S6A-E). However, the levels of similarity for coding and non-coding regions vary depending on the lineage (Fig. S4). For instance, the RT similarities within the Ty1-*copia* and within the Ty3-*gypsy* groups ranged between 57-74 % and 55-88 %, respectively. In contrast, LTR similarities ranged between 17-47 % for Ty1-*copia* groups and between 11-55 % for Ty3-*gypsy* groups, respectively (Fig. 6A-E; Fig. S4; Table 1). Ale and Tat sequences are the most diversified, both within their RTs and their LTRs (Table 1).

Transcriptome analysis reveals transcriptional activity of LTR retrotransposons

We investigated whether the identified full-length LTR retrotransposons are transcribed into RNA by analyzing the reference transcriptome database of *V. macrocarpon* from leaves and shoot tips (Fig. 5; Fig. S7; Table S8). We have mapped a total of 36.1 million transcript sequences against the 54 reconstructed Ty3-*gypsy* and Ty1-*copia* sequences of *V. macrocarpon.* Subsequently, we compared the genome proportion derived from the RepeatExplorer analysis against the transcriptome proportion for each individual element (Fig. 5).

All the identified full-length elements are transcribed to a different extent, thereby contributing a total of 0.1 % by Ty1-*copia* retrotransposons and 0.03 % by Ty3-*gypsy* retrotransposons to the used *V. macrocarpon* transcriptome reads (Fig. 5; Fig. S7; Table S8).

Among all LTR retrotransposon lineages, Ale has the highest proportion of transcript sequences (Fig. 5; Table S8). In fact, the two reference elements SCL5_Ale and SCL22_Ale

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

have both the highest genomic proportions of all 26 Ty1-*copia* representatives and the highest percentage of mapped RNA reads of all 54 reference LTR retrotransposons in *V. macrocarpon* (Fig. 5; Table S8). Nevertheless, the results also show that the percentage of mapped transcripts is dependent on the individual full-length elements and cannot be generalized across and within lineages. Nevertheless, the number of mapped transcript sequences is not proportional to their genomic abundance: most lineages, including Angela, Bianca, Ivana, TAR, Tork, and SIRE, have less representative transcript sequences as compared to their proportional genomic abundance (Fig. 5; Table S8).

Despite their higher genomic abundance, the transcription level of Ty3-*gypsy* retrotransposons is much lower (0.037%) than the transcription of Ty1-*copia* retrotransposons (0.105%): SCL15_Ogre, SCL27_Ogre, SCL28_Tekay, and SCL31_CRM generate most transcripts with similar counts (0.003-0.004 % of the total transcript sequences; Fig. 5; Table S8). Noteworthily, SC31_CRM has only a low genomic abundance (0.22%), but the highest number of transcripts (0.004%) as compared to other Ty3-*gypsy* elements.

We have also investigated the distribution of mapped transcript reads along the most transcribed full-length Ty1-*copia* and Ty3-*gypsy* elements to gain insight into the coverage of the different structural features. In Fig. S7, the most transcribed elements from each group and the distribution of mapped RNA reads are shown. The distribution of mapped transcript sequences along the different structural components of the LTR retrotransposons shows either a homogenous profile along the full-length element or an enrichment in the LTRs and UTRs. SCL5_Ale and SCL49_Tork have homogeneously mapped transcript sequences along their *gag-pol* genes and UTRs (Fig. S7). In contrast, SCL42_Ivana, SCL40_SIRE, SCL49_Ogre, and SCL31_CRM have the highest number of mapped transcript reads in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

two LTRs. Moreover, SCL79_Bianca and SCL6_Athila have most transcripts mapped to the internal UTR located between the *gag* and *pol* genes as well as to the 3' LTR (Fig. S7).

DISCUSSION

Although a draft genome assembly for *V. macrocarpon* is already available [10], we are only starting to decipher its repetitive fraction. As we showed in our previous work that focused on satellite DNAs in *Vaccinium* genomes [36], repetitive DNAs are still the last genomic portions to be assembled completely. Nevertheless, as knowledge of repetitive DNAs is needed to understand *Vaccinium* genomes in their entirety, we extend our previous work to characterize the contribution of long terminal repeat (LTR) retrotransposons to the cranberry genome. As we base our analysis on a random sampling of Illumina reads, we avoid biases introduced by a potentially erroneous genome assembly [39]. Nevertheless, as this approach may neglect less abundant and more diverged repeats [55], all quantifications are likely underestimations.

We showed that nearly a quarter of the *V. macrocarpon* genome is made up of highly abundant, repetitive TE families, with over 91% of genome considered as repetitive [35]. This nearly doubles the estimate retrieved for the cranberry genome assembly (40%) [10, 11], and likely is a result of a systematic exclusion of repetitive regions from the genome assembly. Within the repetitive genome fraction, LTR retrotransposons are the most abundant TEs (44 % of the repetitive fraction), followed by DNA transposons (11 %), and non-LTR retrotransposons (6 %). The observed high abundance of LTR retrotransposons is typical for plants and compares well to other TEs in other species such as flax [56], apple [57] maize [58], and oats [59].

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Cranberry LTR retrotransposons are diverse in sequence and structure

Overall, in the *V. macrocarpon* genome, Ty3-*gypsy* elements are on average longer and contribute a higher genome proportion than Ty1-*copia* elements. This was also reported in other plants, such as *Avena sativa* [59], *Triticum* sp. [60], *Festuca pratensis* [61], and *Malus domestica* [57]. Nevertheless, opposite proportions are also possible, as in *Musa acuminata* and *Pyrus bretschneideri*, in which Ty1-*copia* retrotransposons outweighed Ty3-*gypsy* elements [62]. Variation in sequence lengths between these two groups is thought to reflect length differences in the non-coding rather than in the coding element regions, whereas the genome proportion depends on their copy number rather than on the mere retrotransposon length [63-65]. However, elements from Errantivirus/Athila and Ogre/Tat lineages from Ty3-*gypsy* were found to have both the highest genome proportions and the largest sizes in the *V. macrocarpon* genome. Besides the generally large lengths of Ogre/Tat retrotransposons in *V. macrocarpon*, the high number of RT sequences detected in this study also indicates a high copy number – both contributing to their high genomic abundance. Similarly, for the Ale lineage, we found the most RT instances and a higher genome fraction as compared to the other Ty1-*copia* lineages.

Characterization of *V. macrocarpon*'s RT sequences revealed that Tat and Ale are the most diversified groups among the Ty3-*gypsy* and Ty1-*copia* retrotransposons, respectively. By studying the pairwise RT identities, we observed that individual lineages diversify differently. This is in accordance with our previous report on the related diploid species *Vaccinium corymbosum* [35]. Species- and genus-specific lineage diversification and heterogeneity is reported by many authors for several species. For example, within Ty1-*copia*, Sirevirus/Maximus retrotransposons were highly diversified in sugarcane [64], in the Fabeae [65], and in maize [66], whereas Ale retrotransposons were found in highest diversity

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

in *Chenopodium quinoa* [67] and *Populus trichocarpa* [68]. Within Ty3-*gypsy*, the Tat lineage was found to proliferate in sugarcane [64] and in the Fabeae [65], whereas the Athila clade in *Populus trichocarpa* [68], and the Tekay chromoviruses was for example more abundant in *Beta vulgaris* [69]. The overall abundance and diversification of the various LTR retrotransposon types also reflects the evolutionary relationships of the host species and may be similar within a group of closely related species [70, 71]. The environmental adaptation of a particular species may also be affected by the insertion and molecular diversification of these retrotransposons [16, 72].

Structurally, regarding the encoded protein domains, the reverse transcriptase (RT) is the most ancient domain and key to the dispersion of retrotransposons [17, 73]. The RT is present in all autonomous LTR retrotransposons, whereas the modular acquisition of other polyprotein domains is assumed to result from independent lineage-specific evolutionary histories [17, 74]. Therefore, structural differences among the lineages are considerable. This is particularly evident in the Tat and chromovirus lineages that harbor additional protein-coding domains. All members of the Tat lineage in *V. macrocarpon* harbor two ribonuclease H domains. Ustyantsev *et al.* (2015) [75] speculated that this duality of ribonucleases H might benefit the successful strand transfer, a property common to retroviruses that suffer natural selection pressure during proliferation.

Chromoviral integrases are typically marked by a chromodomain at their C-terminus, presumably with the ability to bind specific histone variants [69 76-79]. Therefore, chromoviruses may directly impact chromatin structure and function [79]. For the chromovirus clade Galadriel and Tekay, a chromodomain was detected both in the *V. macrocarpon* genome assembly and in the representative full-length elements from the read cluster analysis. In contrast, for the CRM clade, chromodomain regions were extracted from

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the genome assembly only, although both genome assembly and read cluster analysis were based on the same *V. macrocarpon* accession. Compared to the CHDs of the Tekay, Galadriel, and Reina clades, their identification within centromeric retrotransposons is more difficult, as they typically do not show structural similarities to type I and II cellular chromodomains and are highly variable even between families within a species [69, 78]. Nevertheless, this domain, also referred to as the CR motif or CHDCR, correlates with various sequence and structural features [28]. For example, the RTs of CRM elements tend to form a separate branch in the analysis of sequence relationship. Another feature is the position of the CHDCR, which is defined by its extension into the 3' LTR. Although some of the reference elements of the CRM clade have incomplete LTR sequences, an extended *gag-pol* ORF into the 3' LTR is recognizable. CR chromodomains of CRM elements were found to be mostly associated with the centromeric heterochromatin. These domains are likely responsible for the diversification, integrity, and stability in the functional centromeric heterochromatin through an RNA-mediated mechanism in different beet and grass species [69, 79-81].

The PBS is an important structural feature for LTR retrotransposon transcription located downstream of the 5’ LTR, with PBS type and length characteristic for a given plant LTR retrotransposon group [28]. The PBS is present in most of the *in silico* retrotransposons constructed for *V. macrocarpon*. Nevertheless, some consensuses lack a designated PBS, likely either an assembly or identification error, or the nature of the PBS itself in the studied species. In chromoviruses as well as in Ty1-*copia* retrotransposons, a PBS corresponding to the initiator methionine tRNA (Met-tRNA) was most common, similar to other plant genomes [28, 82]. In contrast, the PBS regions in other Ty3-*gypsy* retrotransposons were more diverse. The specific PBS site is not only significant for the transcription of LTR retrotransposons, but also for their post-transcriptional regulation through tRNA-derived

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

small RNAs [82, 83]. Therefore, the acquisition of distinct PBS sequences in different retrotransposon lineages may point to individual evolutionary benefits.

The LTRs of the identified full-length Ty1-*copia* and Ty3-*gypsy* retrotransposons were enriched for different *cis*-regulatory elements, including putative promoters and sequences related to hormone responsiveness, cell development, and stress response. This indicates the relevance of these motifs for retrotransposon transcription and activity. In addition, the *V. macrocarpon* full-length elements harbored tandem repeats not only in their LTR region but also in their UTRs. Yet, none were found within coding sequences. The presence of *cis*-regulatory regions and tandem repeats in the LTR is common for LTR retrotransposons and can have significant impacts on their reverse transcription and proliferation [25, 55]. Moreover, *cis*-regulatory sequences could act as enhancers or repressors and thereby affect the expression of downstream genes as part of regulatory networks and pathways [55, 84]. Although still a matter of investigation, the origin of tandem repeats from different parts of the retrotransposons (i.e. LTRs or UTRs) with specific functions are reported for many organisms. An example are the centromeric tandem repeats that may have originated from different retrotransposons in maize [85], wheat [86], and potato [87], suggesting their importance for the formation of functional centromeres in these species. Moreover, the existence of short tandem repeats varying in length and sequence in UTR sequences near the PPT was also reported for many other plant genomes, such as legumes [63] and *Silene latifolia* [88]. Although the specific function of these short repeats is still debated, it is hypothesized that they could serve as a junction, connecting the *gag-pol* gene to the 3' end. Thus, in evolutionary time-spans, they may serve as a seed for the generation of longer satellite DNA arrays, with the potential to acquire a structural function for the organization of chromosomes [88].

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We found that the overall similarity among the full-length elements within each lineage varied. For example, Tat retrotransposons are diversified and have little overall sequence similarity within the UTR regions as was also reported for other organisms [63, 65]. The highly diverse UTRs and LTRs on clade level (see Fig. S6) may indicate their ancient origin and diversification in the *Vaccinium* ancestors [28]. LTRs and UTRs appear to evolve faster than coding regions due to different mechanisms, like unequal recombination or illegitimate recombination [63, 89-92].

Cranberry LTR retrotransposons of the Ale-type are strongly transcribed in *V. macrocarpon* leaves and shoot tips

As expected, all *V. macrocarpon* LTR retrotransposons investigated showed at least a basal level of transcription, together making up about 0.13 % of the total transcriptome. According to [10], transcription of TEs in *V. macrocarpon* is quite low (about 4.3%) and in contrast to the highly transcribed protein-coding genes (about 83.59%). This is in line with our study, in which only the most abundant LTR retrotransposons were queried. A low transcription of retrotransposons compared to conventional protein-coding genes is a common finding in most plant species, including grasses [93], flax [56], and sugarcane [64]. The low transcription of full-length LTR retrotransposons indicates out that these sequences are rather strictly regulated in cranberry [10].

Transcriptional activity of full-length LTR retrotransposons is not correlated with their genome proportion in *V. macrocarpon*. Higher levels of expression than those found here for less abundant elements (such as Ale elements) have been reported in sugarcane [64], maize [58], poplar [94], *Arabidopsis arenosa* [95], and sunflower [96]. Overall, Ty1-*copia* sequences appeared to be more transcribed than Ty3-*gypsy* elements in *V. macrocarpon*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

genome being the highest transcription profiles (~0.08%) those of the Ale full-length elements. In contrast, Athila elements have a higher genomic percentage (2.44%), but have fewer transcript reads assigned (0.004 %), even not covering the full reference element length. In fact, it is considered that Ale and Sireviruses are some of the most transcriptionally active Ty1-*copia* lineages in plants [58, 64, 66, 95-97]. In cranberry, for Ale retrotransposons, this also seems to hold true, whereas for Sireviruses, fewer genomic and transcriptomic proportions were detected. Nevertheless, both Ale and Sireviruses are preferentially accumulated in the euchromatic region of the genome in different species [58, 95, 96] and hence could affect gene regulatory pathways.

In contrast, retrotransposon heterochromatization and fragmentation may decrease element transcription. We know from other plant genomes that some lineages are more prone to truncation and heterochromatic burial, e.g. Athila/Errantiviruses or even the related pararetroviruses in sugar beet [98, 99]. Similar tendencies may impact the TE landscape in *V. macrocarpon*, yielding lineages that are less transcribed than others, e.g Athila/Errantiviruses and Sireviruses. Transcriptional activity of TEs is generally dependent on several factors and may be correlated with development [100], (epi)genetic regulation [15, 16, 64, 101], and environmental adaptation [15, 95, 100]. Therefore, TE transcription can depend on the genomic neighborhood, the presence of *cis*-regulatory motifs, tissue types, developmental states, and environmental effects [100, 101]. Here we observed considerable differences regarding the genome and transcriptome abundances of repetitive sequences in *V. macrocarpon*, which imply that the mechanisms of transcriptional regulation vary depending on the LTR retrotransposon.

Regarding their applicability, the retrotransposons identified here may provide suitable targets for the development of molecular markers in assisted breeding programs [19-23]. We

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

suggest targeting retrotransposons that are most abundant and for which high transcription implies closeness to genes. Hence, we would suggest Ale retrotransposons within the Ty1-*copia* superfamily, as well as Athila retrotransposons within the Ty3-*gypsy*. Both are abundant in genomic and transcriptomic databases, and thus may provide many primer binding sites and likely are situated in the more openly packed euchromatin.

Conclusion

Using short reads and the sequence assembly, we have provided an exhaustive overview of the LTR retrotransposon landscape in the cranberry genome. We have detected all major LTR retrotransposon lineages, with some showing association to short tandemly repeated motifs. Considering their high genomic abundance and transcriptional activity, we suggest that Ale and Athila TEs likely represent the most useful targets to survey TE-derived polymorphisms across genotypes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Conflict of interest

The authors have no conflict of interest to report.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) TUBITAK-2215 PhD scholarship and the Scientific Research Projects Unit (BAP) of Niğde Ömer Halisdemir University (FEB 2017/18 DOKTEP). This work also acknowledges the financial support by the Jagannath University, Bangladesh, in the form of study grants as well as the Georg Forster fellowship of the Alexander von Humboldt Foundation awarded to NS.

Abbreviations

LTR - long terminal repeat; TEs - transposable elements; PR - protease; gRNH - RNase H; RT- reverse transcriptase; INT - integrase, PBS - primer binding site; PPT - polypurine tract; CHD – chromodomain; UTR – untranslated region; DNA - deoxyribonucleic acid; RNA - ribonucleic acid.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

References

[1] Trehane J. Blueberries, cranberries, and other vacciniums. Timber Press; 2004.

[2] Kron KA, Powell EA, Luteyn JL. Phylogenetic relationships within the blueberry tribe (Vaccinieae, Ericaceae) based on sequence data from matK and nuclear ribosomal ITS regions, with comments on the placement of Satyria. Am J Bot. 2002;89(2):327-36.

[3] Powell EA, Kron KA. Hawaiian blueberries and their relatives—a phylogenetic analysis of *Vaccinium* sections *Macropelma*, *Myrtillus*, and *Hemimyrtillus* (Ericaceae). Syst Bot. 2002;27(4):768-79.

[4] Vander Kloet SP, Dickinson TA. A subgeneric classification of the genus *Vaccinium* and the metamorphosis of V. section *Bracteata Nakai*: more terrestrial and less epiphytic in habit, more continental and less insular in distribution. J Plant Res. 200;122(3):253-68.

[5] Hancock JF, Lyrene P, Finn CE, Vorsa N, Lobos GA. Blueberries and cranberries. In: Hancock JF, editor. Temperate fruit crop breeding. Springer; 2008, p. 115-50.

[6] Česoniene L, Daubaras R, Paulauskas A, Zukauskiene J, Zych M. Morphological and genetic diversity of European cranberry (*Vaccinium oxycoccos* L., Ericaceae) clones in Lithuanian reserves. Acta Soc Bot Pol. 2013;82(3):211-217.

[7] Moyer RA, Hummer KE, Finn CE, Frei B, Wrolstad RE. Anthocyanins, phenolics, and antioxidant capacity in diverse small fruits: *Vaccinium*, *Rubu*s, and *Ribes*. J Agric Food Chem. 2002;50(3):519-25.

[8] Nickavar B, Amin G. Anthocyanins from *Vaccinium arctostaphylos* berries. Pharm Biol. 2004;42(4-5):289-91.

[9] Zdepski, A., Debnath, S.C., Howell, A., Polashock, J., Oudemans, P., Vorsa, N. and

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Michael, T.P. Cranberry. In Genetics, genomics and breeding of berries; Folta, K., Kole, C., Ed.; Boca Raton, F.L., p. 200 ISBN 9781578087075 - CAT# N10335, 2011.

[10] Polashock J, Zelzion E, Fajardo D, Zalapa J, Georgi L, Bhattacharya D, Vorsa N. The American cranberry: first insights into the whole genome of a species adapted to bog habitat. BMC Plant Biol. 2014;14(1):1-8.

[11] Diaz-Garcia L, Garcia-Ortega LF, González-Rodríguez M, Delaye L, Iorizzo M, Zalapa J. Chromosome-level genome assembly of the American cranberry (*Vaccinium macrocarpon* Ait.) and its wild relative *Vaccinium microcarpum*. Front Plant Sci. 2021;12:137.

[12] Gupta V, Estrada AD, Blakley I, Reid R, Patel K, Meyer MD, Andersen SU, Brown AF, Lila MA, Loraine AE. RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing. Gigascience. 2015;4(1):s13742-015.

[13] Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, Wisecaver JH, Yocca AE, Alger EI, Tang H, Xiong Z. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. GigaScience. 2019;8(3):giz012.

[14] Biémont C, Vieira C. Junk DNA as an evolutionary force. Nature. 2006;443(7111):521-4.

[15] Oliver KR, McComb JA, Greene WK. Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biol Evol. 2013;5(10):1886-901.

[16] Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016;17(1):1-4.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[17] Ustyantsev K, Blinov A, Smyshlyaev G. Convergence of retrotransposons in oomycetes and plants. Mob DNA. 2017;8(1):1-1.

[18] Niu XM, Xu YC, Li ZW, Bian YT, Hou XH, Chen JF, Zou YP, Jiang J, Wu Q, Ge S, Balasubramanian S. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. Proc Natl Acad Sci. 2019;116(14):6908-13.

[19] Kalendar R, Schulman AH. Transposon-based tagging: IRAP, REMAP, and iPBS. In Molecular Plant Taxonomy 2014 (pp. 233-255). Humana Press, Totowa, NJ.

[20] Jiang S, Zong Y, Yue X, Postman J, Teng Y, Cai D. Prediction of retrotransposons and assessment of genetic variability based on developed retrotransposon-based insertion polymorphism (RBIP) markers in *Pyrus* L. Mol Genet Genom. 2015;290(1):225-37.

[21] Zong Y, Kang H, Fang Q, Chen X, Zhou M, Ni J, Zhang Y, Wang L, Zhu Y, Guo W. Phylogenetic relationship and genetic background of blueberry (*Vaccinium* spp.) based on retrotransposon-based SSAP molecular markers. Scientia Horticulturae. 2019;247:116-22.

[22] Seibt KM, Wenke T, Wollrab C, Junghans H, Muders K, Dehmer KJ, Diekmann K, Schmidt T. Development and application of SINE-based markers for genotyping of potato varieties. Theor Appl Genet. 2012;125(1):185-96.

[23] Reiche B, Koegler A, Morgenstern K, Brueckner M, Weber B, Heitkam T, Seibt KM, Troeber U, Meyer M, Wolf H, Schmidt T. Application of retrotransposon-based Inter-SINE Amplified Polymorphism (ISAP) markers for the differentiation of common poplar genotypes. Can J For Res. 2021 (ja).

[24] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E. A unified classification system for eukaryotic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

transposable elements. Nat Rev Genet. 2007;8(12):973-82.

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

[26] Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol. 2015;86:90-109.

[27] Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6(1):1-6.

[28] Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA. 2019;10(1):1-7.

[29] Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):1-4.

[30] Elliott TA, Heitkam T, Hubley R, Quesneville H, Suh A, Wheeler TJ. TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. Mob DNA. 2021;12(1):1-5.

[31] Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. 2007;17(7):1072-81.

[32] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Fulton L, Graves TA, Minx P. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326(5956):1112-5.

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

[34] Marín I, Lloréns C. Ty3/Gypsy retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Mol Biol Evol. 2000;17(7):1040-9.

[35] Sultana N, Serçe S, Menzel G, Heitkam T, Schmidt T. Comparative analysis of repetitive sequences reveals genome differences between two common cultivated *Vaccinium* Species (*V. corymbosum* and *V. macrocarpon*). J Mol Biol Biotech. 2017;2:24.

[36] Sultana N, Menzel G, Heitkam T, Kojima KK, Bao W, Serçe S. Bioinformatic and Molecular Analysis of Satellite Repeat Diversity in *Vaccinium* Genomes. Genes. 2020;11(5):527.

[37] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

[38] Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics. 2013;29(6):792-3.

[39] Novák P, Neumann P, Macas J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc. 2020;15(11):3745-76.

[40] Smit AF, Hubley R, Green P. RepeatMasker Open-4.0. 2015.

[41] Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Cooper A, Markowitz S, Duran C, Thierer T. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647-9.

[42] Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(suppl_2):W265-8.

[43] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276-7.

[44] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res.1999;27(2):573-80.

[45] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.

[46] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80.

[47] Team RC. R: A language and environment for statistical computing.

[48] Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010;5(3):e9490.

[49] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.

[50] Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 2000;28(6):1102-4.

[51] Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

2002;30(1):325-7.

[52] Seibt KM, Schmidt T, Heitkam T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. Bioinformatics. 2018;34(20):3575-7.

[53] Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit. 2010;20;5.

[54] Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014 Mar 17.

[55] Orozco-Arias S, Isaza G, Guyot R. Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. Int J Mol Sci. 2019;20(15):3837.

[56] González LG, Deyholos MK. Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. BMC Genomics. 2012;13(1):1-7.

[57] Sun HY, Dai HY, Zhao GL, Ma Y, Ou CQ, Li H, Li LG, Zhang ZH. Genome-wide characterization of long terminal repeat-retrotransposons in apple reveals the differences in heterogeneity and copy number between Ty1-*copia* and Ty3-*gypsy* retrotransposons. J Integr Plant Biol. 2008;50(9):1130-9.

[58] Meyers BC, Tingey SV, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. 2001;11(10):1660-76.

[59] Liu Q, Li X, Zhou X, Li M, Zhang F, Schwarzacher T, Heslop-Harrison JS. The repetitive DNA landscape in *Avena* (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. BMC Plant Biol.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

2019;19(1):1-7.

[60] Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, De Oliveira R, Mayer KF, Paux E, Choulet F. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol. 2018;19(1):1-8.

[61] Ebrahimzadegan R, Houben A, Mirzaghaderi G. Repetitive DNA landscape in essential A and supernumerary B chromosomes of *Festuca pratensis* Huds. Sci Rep. 2019;9(1):1-1.

[62] Vitte C, Fustier MA, Alix K, Tenaillon MI. The bright side of transposons in crop evolution. Brief Funct Genom. 2014;13: 276–95.

[63] Macas J, Neumann P. Ogre elements—a distinct group of plant Ty3/*gypsy*-like retrotransposons. Gene. 2007;390(1-2):108-16.

[64] Domingues DS, Cruz GM, Metcalfe CJ, Nogueira FT, Vicentini R, de S Alves C, Van Sluys MA. Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics. 2012;13(1):1-4.

[65] Macas J, Novak P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fukova I, Doležel J, Kelly LJ, Leitch IJ. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PloS one. 2015;10(11):e0143424.

[66] Bousios A, Kourmpetis YA, Pavlidis P, Minga E, Tsaftaris A, Darzentas N. The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. Plant J 2012;69(3):475-88.

[67] Kolano B, Bednara E, Weiss-Schneeweiss H. Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of *Chenopodium*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

quinoa (Amaranthaceae). Plant Cell Rep. 2013;32(10):1575-88.

[68] Natali L, Cossu RM, Mascagni F, Giordani T, Cavallini A. A survey of Gypsy and Copia LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the *Populus trichocarpa* (L.) genome. Tree Genet Genomes. 2015;11(5):1-3.

[69] Weber B, Heitkam T, Holtgräwe D, Weisshaar B, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T. Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. Mob DNA 2013;4(1):1-6.

[70] Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. Genomic repeat abundances contain phylogenetic signal. Syst Biol. 2015;64(1):112-26.

[71] Vitales D, Garcia S, Dodsworth S. Reconstructing phylogenetic relationships based on repeat sequence similarities. Mol Phylogenetics Evol. 2020;147:106766.

[72] Benabdelmouna A, Darmency H. Copia-like retrotransposons in the genus *Setaria*: Sequence heterogeneity, species distribution and chromosomal organization. Plant Syst Evol. 2003;237(3):127-36.

[73] Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 1990;9(10):3353-62.

[74] Capy P, Maisonhaute C. Acquisition/loss of modules: the construction set of transposable elements. Russ J Genet. 2002;38(6):594-601.

[75] Ustyantsev K, Novikova O, Blinov A, Smyshlyaev G. Convergent evolution of ribonuclease H in LTR retrotransposons and retroviruses. Mol Biol Evol. 2015;32(5):1197-207.

[76] Gorinšek B, Gubenšek F, Kordiš D. Evolutionary genomics of chromoviruses in

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

eukaryotes. Mol Biol Evol. 2004;21(5):781-98.

[77] Kordiš D. A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. Gene. 2005;347(2):161-73.

[78] Novikova O. Chromodomains and LTR retrotransposons in plants. Commun Integr Biol. 2009;2(2):158-62.

[79] Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. 2008;18(3):359-69.

[80] Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR. Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics. 2003;163(2):759-70.

[81] Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, Widmer A, Doležel J, Macas J. Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA. 2011;2(1):1-6.

[82] Martinez G. tRNAs as primers and inhibitors of retrotransposons. Mob Genet Element. 2017;7(5):1-6.

[83] Cullen H, Schorn AJ. Endogenous Retroviruses Walk a Fine Line between Priming and Silencing. Viruses. 2020;12(8):792.

[84] Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants. 2019;5(12):1250-9.

[85] Sharma A, Wolfgruber TK, Presting GG. Tandem repeats derived from centromeric retrotransposons. BMC Genomics. 2013;14(1):1-1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[86] Cheng ZJ, Murata M. A centromeric tandem repeat family originating from a part of Ty3/*gypsy*-retroelement in wheat and its relatives. Genetics. 2003;164(2):665-72.

[87] Tek AL, Song J, Macas J, Jiang J. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. Genetics. 2005;170(3):1231-8.

[88] Kejnovsky E, Kubat Z, Macas J, Hobza R, Mracek J, Vyskot B. Retand: a novel family of *gypsy*-like retrotransposons harboring an amplified tandem repeat. Mol Genet Genom. 2006;276(3):254-63.

[89] Devos KM, Brown JK, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 2002;12(7):1075-9.

[90] Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol Biol Evol. 2003;20(4):528-40.

[91] Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci. 2004 Aug 24;101(34):12404-10.

[92] Sanchez DH, Gaubert H, Drost HG, Zabet NR, Paszkowski J. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat Commun. 2017;8(1):1-7.

[93] Vicient CM, Jääskeläinen MJ, Kalendar R, Schulman AH. Active retrotransposons are a common feature of grass genomes. Plant Physiol. 2001;125(3):1283-92.

[94] Cossu RM, Buti M, Giordani T, Natali L, Cavallini A. A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. Tree Genet

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Genomes. 2012;8(1):61-75.

[95] Wos G, Choudhury RR, Kolář F, Parisod C. Transcriptional activity of transposable elements along an elevational gradient in *Arabidopsis arenosa*. Mob DNA. 2021;12(1):1-2.

[96] Qiu F, Ungerer MC. Genomic abundance and transcriptional activity of diverse gypsy and copia long terminal repeat retrotransposons in three wild sunflower species. BMC Plant Biol. 2018;18(1):1-8.

[97] Bousios A, Darzentas N, Tsaftaris A, Pearce SR. Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants?. BMC Genomics. 2010;11(1):1-4.

[98] Wollrab C, Heitkam T, Holtgräwe D, Weisshaar B, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T. Evolutionary reshuffling in the Errantivirus lineage Elbe within the *Beta vulgaris* genome. Plant J. 2012;72(4):636-51.

[99] Schmidt N, Seibt KM, Weber B, Schwarzacher T, Schmidt T, Heitkam T. Broken, silent, and in hiding: Tamed endogenous pararetroviruses escape elimination from the genome of sugar beet (*Beta vulgaris*). Ann Bot. 2021; mcab042.

[100] Meagher TR, Vassiliadis C. Phenotypic impacts of repetitive DNA in flowering plants. New Phytol. 2005;168(1):71-80.

[101] Ito H. Small RNAs and transposon silencing in plants. Dev. Growth Differ. 2012;54(1):100-7.

Table 1. General structural features and repeat compositions of full-length LTR retrotransposons. Classification of the in silico retrotransposons was performed according to Neumann *et al.* 2019 [25].

Class / Superfamily	**Lineage (group)|superclade|clade|subclade**	**Total full-length elements**	**Genome proportion [1] (%)**	**Total element size (kb)**	**LTR size (bp)**	**PBS (tRNA types)[2]**	**Range of pairwise similarity (%)**	
							LTR	**RT**
Ty1-*copia*	Ale	5	0.02-0.6	5.2-4.9	401-127	Asn, Met	17-41	58-72
	Alesia	2	0.01-0.1	5-5.4	289-562	n.d.	54	63
	Bianca	1	0.1	7	313-326	n.d.		
	Ivana	3	0.02-0.2	4.9-5.5	417-438	Met	41.5-47.4	62-69
	TAR	6	0.1-0.3	5.5-6.5	585-994	Glu, Ser, Met, Tyr, Trp	21-41	63-74
	Angela	1	0.21	6.8	796	Met		
	Tork	7	0.1-0.3	4.8-5.6	329-773	Met	17-43	57-70
	SIRE	1	0.2	7	1,545	Met		
	Total	26	4.3					
Ty3-*gypsy*	non-chromovirus|OTA|Tat|Ogre	7	0.1-1.3	11.5-19.9	761-4560	Asn, Arg, Thr	10.9-28.9	55.4-73.7
	non-chromovirus|OTA|Tat|TatV	6	0.1-1.0	10.5-11.7	646-1646	Lys, Arg	20.3-35.6	62.3-74.7
	non-chromovirus|OTA|Athila	5	0.01-1.9	5.7-9.8	854-1403	Asp	8.7-49.6	69-81.8
	chromovirus|Tekay	4	0.03-0.2	5.4-8.6	1149-1925	Gln	49.4	87.5
	chromovirus|CRM	4	0.02-0.2	5.3-5.5	361-556	Met	47.6-54.3	67.1-75.9
	chromovirus|Galadriel	2	0.1	5.3-6.4	465-1061	Met	54.5	67.6
	Total	28	8.4					

[1]Genome proportion = range of genome proportion for each specific reconstructed full-length elements
[2]n. d. = not detected.

Table 2: Summary of tandem repeats and *cis*-regulatory elements associated with the reconstructed full-length LTR-retrotransposon sequences.

Lineage /clade /subclade	**Location of tandem repeats**	**Consensus period size**	**Number of repetitions**	**Total number of different *cis*-regulatory motifs identified in 5' LTR region**
Ale	5' and 3' LTR, 5' UTR	2-48	1.9-7.9	10
Alesia	5 LTR, 5' UTR	30	2.2-4.9	5
Bianca	5'UTR	25	2.8	2
Ivana	5'LTR	27	2.9	8
TAR	5'LTR	10-17	1.9-4.4	19
Angela	5'LTR	18	5.4	9
Tork	5'and 3' LTR, 5' UTR	11-49	2-6.8	19
SIRE	5' LTR	15	2	12
Ogre	5' and 3' LTR, UTR adjacent to the *gag* gene	6-91	2-14.2	38
TatV	5' and 3' LTR, UTR	15-93	2.1-15.1	24
Athila	3' LTR, UTR	21-42	2-2.5	26
Tekay	5' and 3' LTR, UTR	2-29	1.9-18.5	10
CRM	5' and 3' LTR, UTR	17-48	1.9-4.5	8
Galadriel	n.d.	n.d.	n.d.	13

n.d. = no tandem repeat detected

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

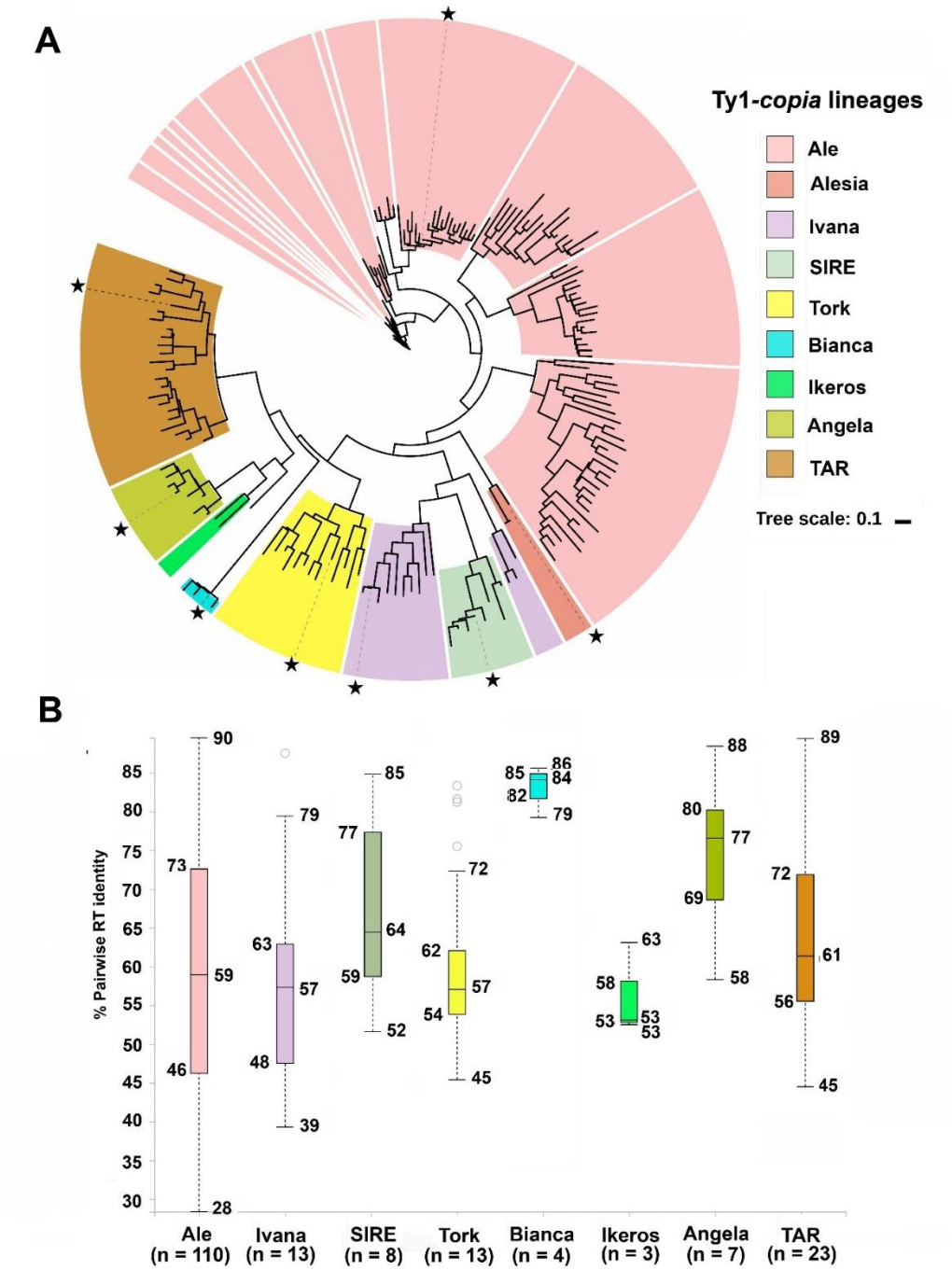

Fig. 1. Diversity of the different Ty1-*copia* lineages in the *V. macrocarpon* genome. A) The dendrogram was derived from 181 Ty1-*copia* RT amino acid sequences calculated with the approximate Maximum-Likelihood method using the FastTree tool (Price *et al*. 2010) [45]. Lineages are color-coded (see internal legend). The stars indicate the position of the representative reconstructed full-length elements from *V. macrocarpon* genome shown in Fig. 3A. B) The boxplots illustrate the pairwise sequence identities of amino acid sequences of all Ty1-*copia* RTs. The lineage Alesia was excluded from the boxplots as only a single copy was detected in the *V. macrocarpon* genome. Colors in the boxplots correspond to panel (A). For each lineage, the total number of RT sequences (n) is provided, which includes the reference sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

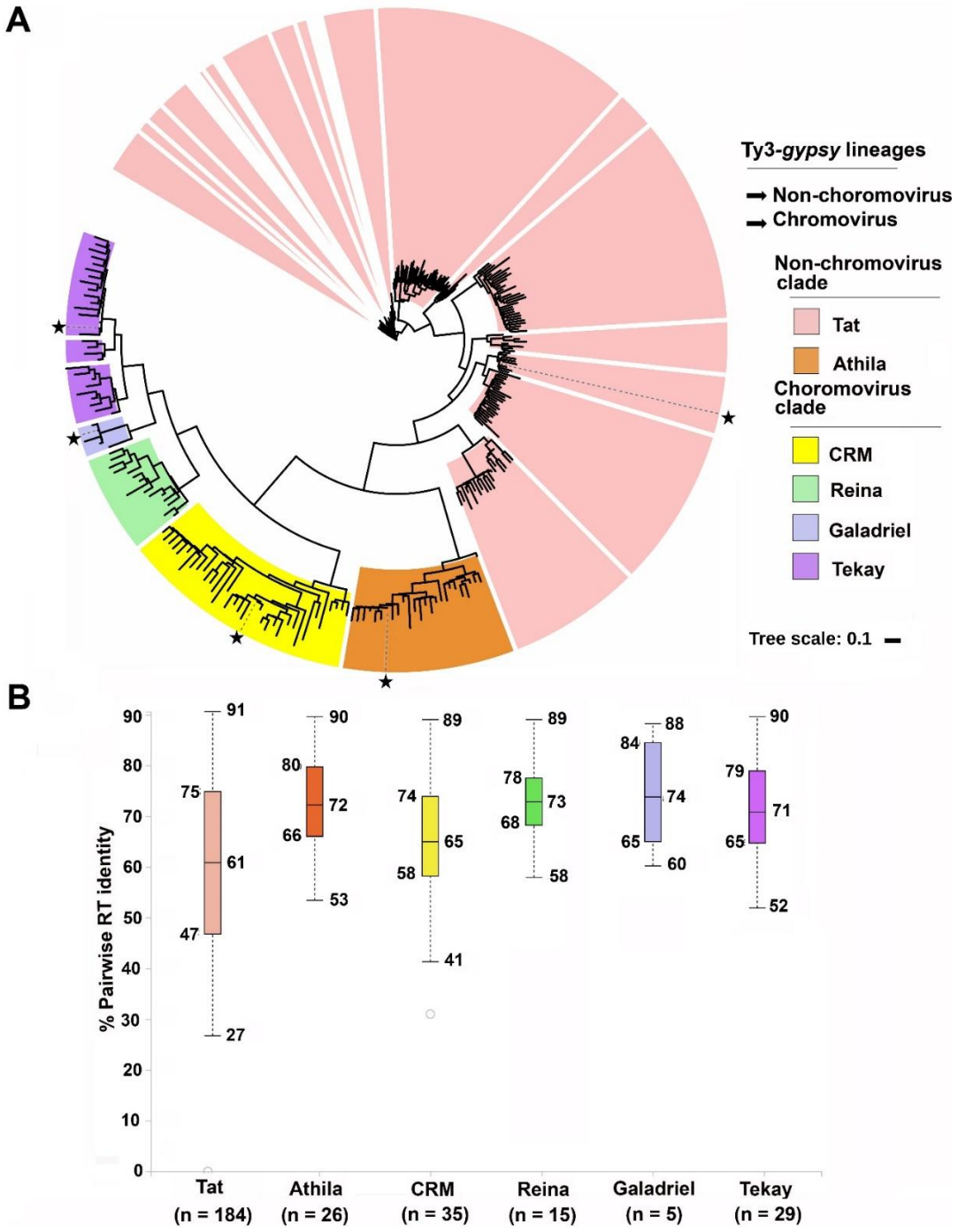

Fig. 2. Diversity of the different Ty3-*gypsy* lineages in the *V. macrocarpon* genome. A) The dendrogram was derived from 294 Ty3-*gypsy* RT amino acid sequences calculated with the approximate Maximum-Likelihood method with the FastTree tool [45]. Clades are color-coded (see internal legend). The stars indicate the position of the representative for the reconstructed full-length elements from *V. macrocarpon* genome showed in Fig. 4A. B) The boxplots illustrate the pairwise sequence identities of amino acid sequences of all Ty3-*gypsy* RTs. Colors in the boxplots correspond to panel (A). For each clade, the total number of RT sequences (n) is provided, which includes the reference sequence

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

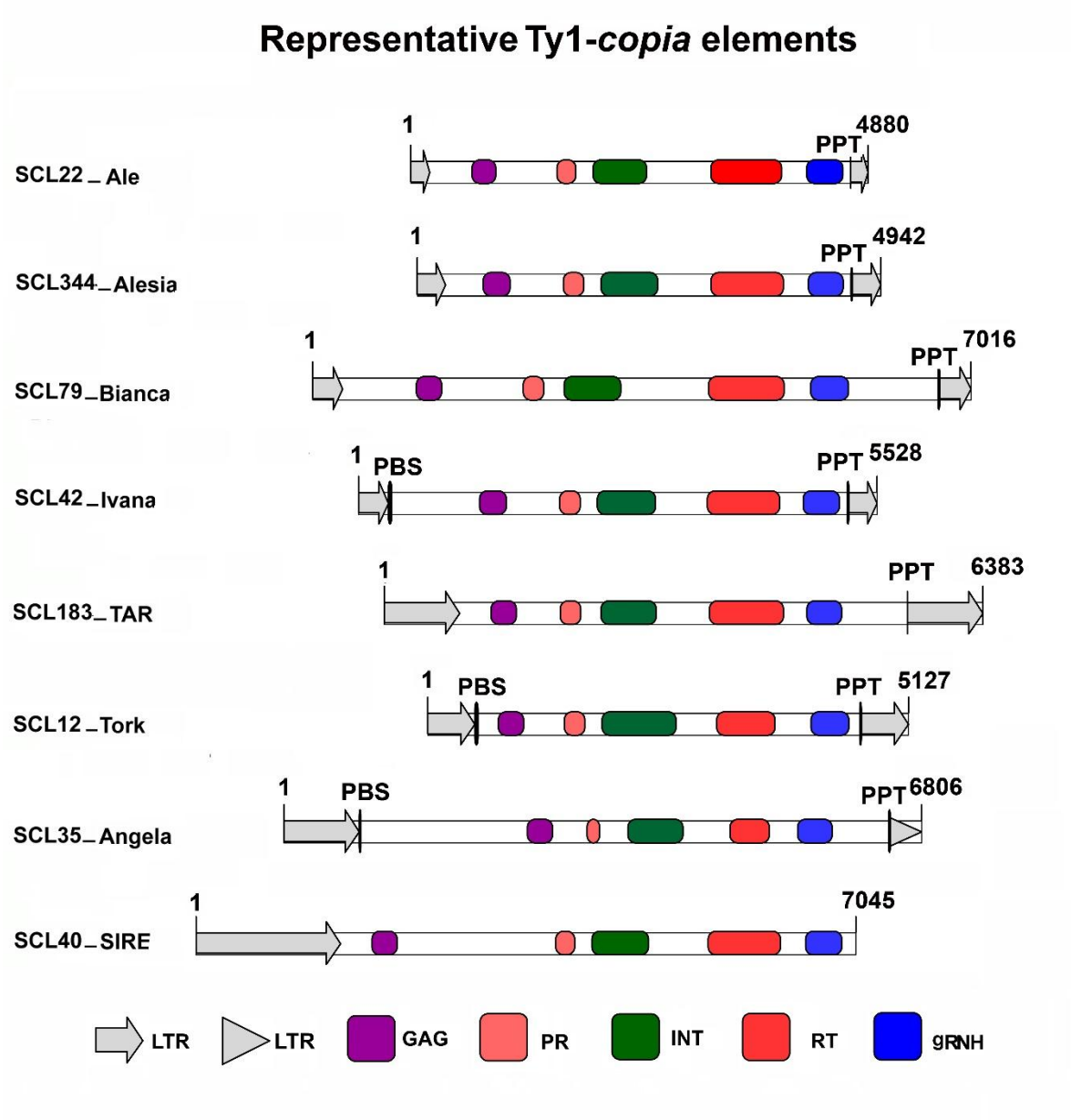

Fig. 3. Structural features of representative *in silico* elements of Ty1-*copia* LTR retrotransposons in the *V. macrocarpon* genome. Long terminal repeats (LTRs) are represented as grey open arrows (intact LTR) and grey arrowheads (truncated LTR). Only the SCL40_SIRE consensus is incomplete and lacks the 3' LTR. Black vertical lines adjacent to the LTRs represent primer binding site (PBS) and polypurine tracts (PPT), while thickness is according to length of this region (Table S2). Structural features of *gag* domain and the four genes of the *pol* domain are depicted as color-coded boxes: GAG = *gag* domain, PR = protease, RT = reverse transcriptase, gRNH = *gypsy*-type RNase H, INT = integrase.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

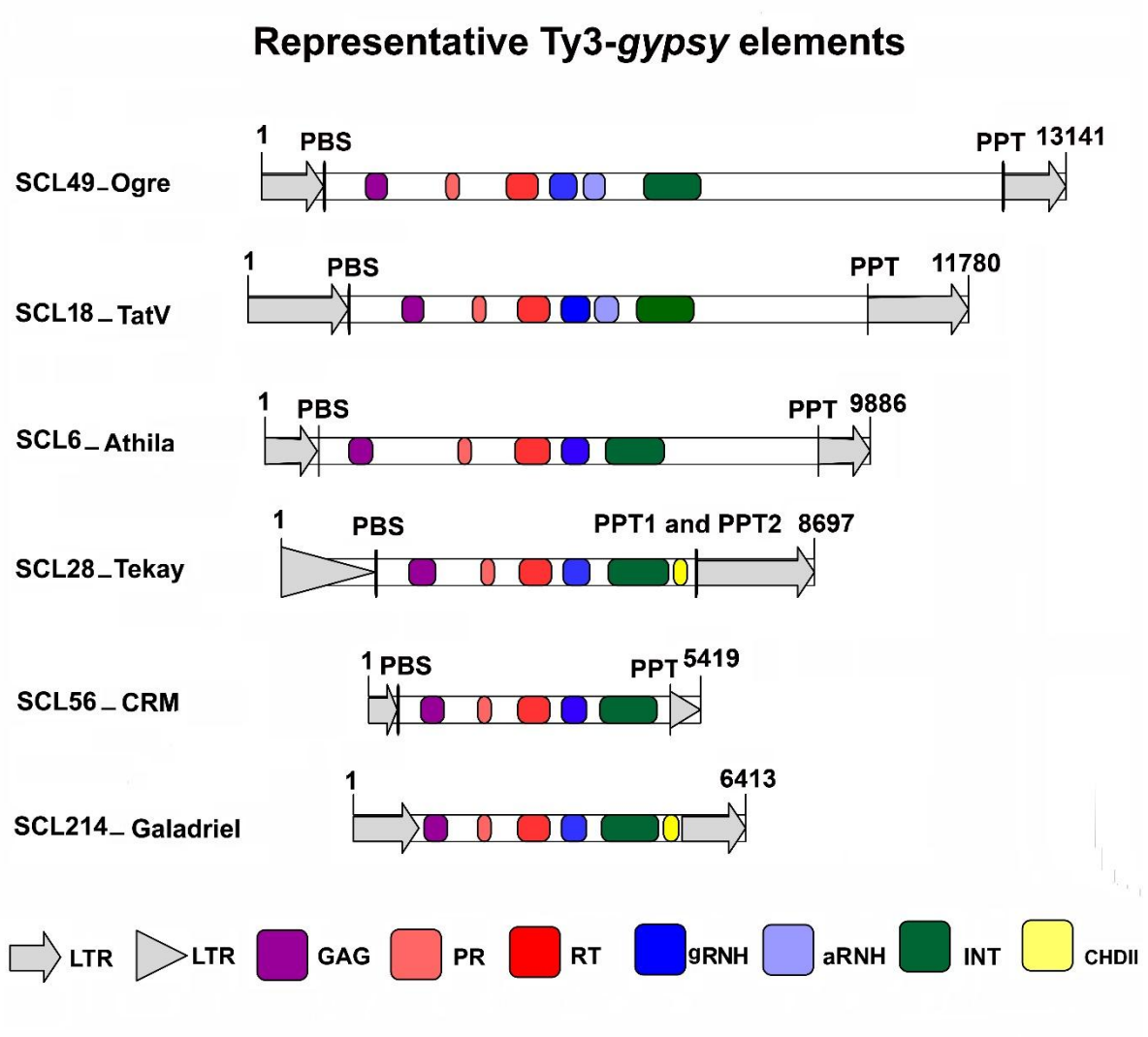

Fig. 4. Structural features of full-length representative consensus elements of Ty3-*gypsy* LTR retrotransposons in the *V. macrocarpon* genome. Long terminal repeats (LTRs) are represented as grey open arrows (intact LTR) and grey arrowheads (truncated LTR). Black vertical lines adjacent to the LTRs represent primer binding site (PBS) and polypurine tracts (PPT), while thickness is according to length of this region (Table S2). Protein domains encoded in *gag* and *pol* are shown within the LTRs border. Structural features of the *gag* and *pol* genes are depicted as color-coded boxes: GAG = *gag* gene, PR = protease, RT = reverse transcriptase, gRNH = gypsy RNase H, aRNH = archaeal RNase H, INT = integrase, CHDII = chromodomain II.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452536; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

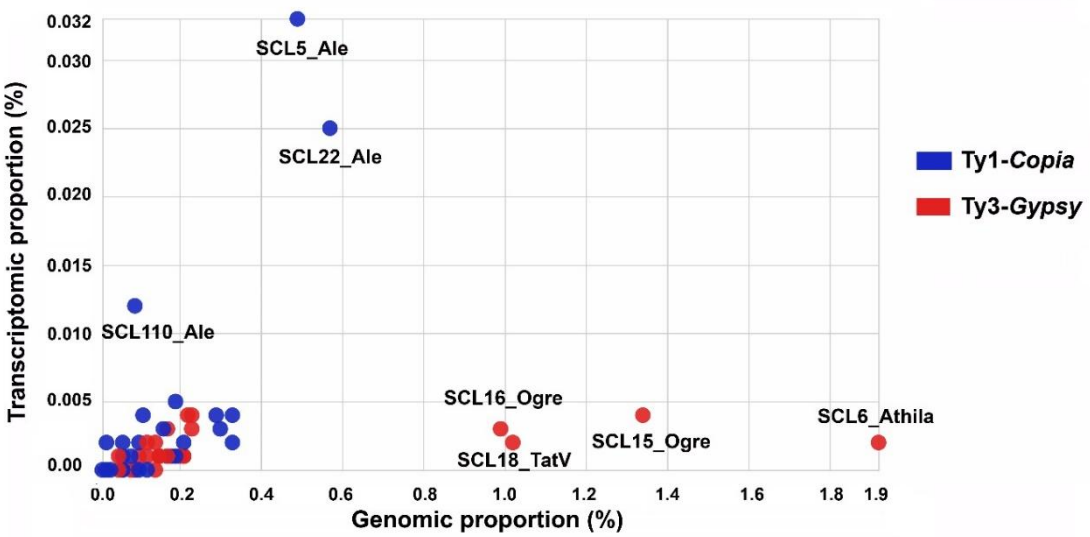

Fig. 5. Comparison of genome and transcriptome proportions for the 26 Ty1-*copia* and 28 Ty3-*gypsy* full-length consensus retrotransposons from *V. macrocarpon*. The genome proportions of the reconstructed full-length Ty1-*copia* and Ty3-*gypsy* elements are plotted against their transcriptome proportions. Genome proportions were calculated from the RepeatExplorer output (Table S2). Transcriptome proportions were calculated from read mapping of the publicly available Illumina cDNA sequence reads of *V. macrocarpon* (accession number PRJNA246586). Major groups of the 54 elements are color-coded (see internal legend). Only the names of elements of the highest genome and/or transcriptome proportions are annotated. For details, see Table S2.