Genomic re-assessment of the transposable element landscape of the potato genome

Diego Zavallo[*1], Juan Manuel Crescente[*2,4], Magdalena Gantuz[3,4], Melisa Leone[1,5], Leonardo Sebastian Vanzetti[2,4], Ricardo Williams Masuelli[3,4], and Sebastian Asurmendi[1]

[1]Instituto de Agrobiotecnología y Biología Molecular (IABIMO), Instituto Nacional de Tecnología Agropecuaria (INTA), Consejo Nacional de investigaciones Científicas y Tecnológicas (CONICET), Los Reseros y Nicolas Repeto, Hurlingham, Argentina

[2]Grupo Biotecnologia y Recursos Genéticos, EEA INTA Marcos Juárez, Ruta 12 km 3, 2580 Marcos Juárez, Argentina

[3]Instituto de Biología Agrícola de Mendoza (IBAM), Facultad de Ciencias Agrarias (FCA), CONICET-UNCuyo, Almirante Brown 500, M5528AHB, Chacras de Coria, Mendoza, Argentina

[4]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

[5]Agencia Nacional de Promocion Científica y Tecnológica (ANPCyT), Argentina.

Corresponding autor: Diego Zavallo;
E-mail: zavallo.inta@inta.gob.ar
Phone: (5411) 4621 447/1676, int. 3526

Summary

Transposable elements are DNA sequences with the ability to auto-replicate and move that occupy large proportions of the host genome. TEs are responsible for shaping the genome size and are major drivers of gene regulation, epigenetics, stress regulation and genome evolution. Given their significance, the development of clear and efficient TEs annotation pipelines has become essential for many species. The latest TE discovery *de novo*-based methods tools, along with available TEs from Repbase and sRNA-seq data, allowed us to perform an improved detection, classification and annotation of reliable potato TEs throughout an open-source and freely available pipeline (https://github.com/DiegoZavallo/TE_Discovery). Previous TEs annotation for this genome varied from 20 to 34% genome coverage. However, our pipeline revealed that only ca. 16% of the genome can clearly be annotated as TEs. Since we used a variety of tool, approaches and rules, our results suggest that several TEs annotation

projects might overestimate annotations. Additionally, we described the distribution of the different types of TEs across the genome, where LTRs and MITEs present a clear clustering pattern in pericentromeric and subtelomeric/telomeric regions respectively. Finally, we analyzed the insertion age and distribution of LTRs families which display a distinct pattern between the two major families. Thus, older Gypsy elements concentrated around heterochromatic regions and younger Copia elements located predominantly on euchromatic regions. Overall, we delivered not only a reliable, ready-to-use potato TEs annotation files, but also all the necessary steps to performed de novo detection for other species.

Significance Statement

Transposable elements have been historically underprivileged in genome assembly projects because of their repetitive nature as well as their heterogeneity, distribution and mutation rates. All these characteristics make their accurate detection harder to achive. Here we provide a comprehensive potato TEs landscape, based on a wide variety of identification tools and approaches, clustering methods, copies detection, filtering rules and clear outputs, which the scientific community will likely use for potato research.

Introduction

Transposable elements (TEs) are DNA sequences with the ability to auto-replicate and move throughout the host genome. In plants, TE copies can occupy large proportions of the DNA complement, representing more than half of the genome in some cases. For example, they comprise about 85% and 88% of wheat and maize genomes, respectively (Schnable et al., 2009; Appels et al., 2018) which indicates the relevance of these elements in genome architecture and size (Roessler et al., 2018). Moreover, in the past few years an increasing numbers of studies shed light on their importance in gene regulation (Hirsch and Springer, 2017; Judd and Feschotte, 2018), epigenetics, stress response and genome evolution (Hosaka and Kakutani, 2018) among others.

TEs are usually divided into two major classes based on their transposition "multiplication" method. Class I elements (or retrotransposons) copy themselves via an RNA intermediate that is reverse transcribed, whereas class II elements (or DNA transposons) move through a "cut and paste" mechanism. Another type of classification is by their self-ability to transpose (i.e. autonomous), characteristic shared by some TEs from both classes. The non-autonomous elements can move but rely on autonomous TEs for their mobility (Wicker et al.,2007).

Given their significance, the identification, classification and annotation of TEs has become a new field in science which involves both "wet lab" biology and bioinformatics. As Hoen et al. (2015) describes in their review on TE annotation benchmarking, precisely detection and annotation of TEs is a difficult task due to its great diversity, both within and among genomes. TEs differ across multiple attributes, including transposition mechanism, sequence, length, repetitiveness, and chromosomal distribution. In addition, whereas recently inserted TEs have a relatively low variability within the family, over time they accumulate mutations and diverge; which makes them harder to detect (Hoen et al., 2015). Furthermore, depending on the type of element, specific well-conserved motif may remain invariant, whereas other features are highly diverse.

There are basically two main strategies for TE annotation: the homology-based and the *de novo*-based methods, which can also be referred to as library-based and signature-based, respectively. The homology-based strategy uses libraries of known TEs, such as the Repbase repository (Jurka et al., 2005) to screen genomes in order to identify similar sequences, most commonly by using RepeatMasker (Smit et al., 1996). On the other hand, the "*de novo*" approach uses the characteristic structural features such as LTRs (Long Terminal Repeats) for retrotransposons and TIRs (Terminal Inverted Repeats) for DNA transposons to identify new elements.

Moreover, autonomous TEs have well-defined motifs like RT (reverse-transcriptase) or TR (transposase) that can also be used for accurately TE identification (Wicker et al., 2007). Several tools based on structural features are available like LTRharvest (Ellinghaus et al., 2008) and TIRvish command which are part of the GenomeTools suite (Gremme et al., 2013). Other tools based on this criterium are specifically designed to discover different types of TEs families (e.g. SINE Scan (Mao and Wang, 2016), HelitronScanner (Xiong et al., 2014) and MITETracker (Crescente et al., 2018). Another "*de novo*" based strategy relies on the most important biological mechanism that silences TEs - RNA directed DNA Methylation (RdRM) - in which double stranded RNA (dsRNAs) are processed to 21-24 nt small interfering RNAs (siRNAs) and guide methylation on homologous DNA loci. For instance, TASR (for Transposon Annotation using Small RNAs) tool uses sRNAs Illumina data as guide to annotate TEs (El Baidouri et al., 2015). However, the most accurate assessment on TEs annotation requires a combination of both strategies with an additional manual curation step (Platt et al., 2016).

Nevertheless, all these tools (both class specific and holistic) often give a representative sequence for each family (usually a full-length TE sequence), but they lack on presenting their copies across the genome, not only for other potentially autonomous copies, but also for members that are partially or entirely deficient in one or more domain. Furthermore, in the case of non-autonomous TEs (e.g. SINEs, TRIMs and MITEs), the amount of copies across the genome is highly elevated due to their repetitive content and their short length; however this information is usually not presented.

For the scientific community, especially in genomics, the need of reliable annotation is becoming fundamental. Generally, for each genome sequencing project, annotations are made, consisting mainly of coding protein genes (structural and functional) and miRNAs genes, but lacking on comprehensive and accurately TE annotations. As an example, the Gramene Database (http://www.gramene.org) is a curated, open-source database for comparative functional genomics in crops and model plant species that has 57 plant genomes with information for more than two million genes, including biological pathway genes. By contrast, the PGSB Repeat Element Database (Nussbaumer et al., 2012) which compiles publicly available repeat sequences from TREP (Wicker et al., 2002), TIGR repeats (Ouyang and Buell, 2004), PlantSat (Macas et al., 2002), Genbank and "*de-novo*" detected LTR retrotransposon sequences from LTR STRUC (McCarthy and McDonald, 2003) only comprise 62.000 sequences (less than 3% with respect to genes). Nonetheless, in the last version of sunflower genome, Badouin et al. (2017) performed a comprehensive search for repeat elements by

developing a new tool called Tephra (https://github.com/sestaton/tephra) (Staton, 2018), which discovers and annotates all types of transposons. Tephra combines existing specific transposon discovery tools for all types of TEs, classifies and annotates them, but it still lacks information of copy numbers across the genome. A recent study addressed this issue by applying a method called “Russian doll” due to its nesting strategy. This method builds nested libraries establishing different searching rules for each one of them. The first one includes only “potentially autonomous TEs”, the second one contains “all TEs”, including non-autonomous and a third one that also includes uncategorized “repeat elements” (Berthelier et al., 2018).

Solanum tuberosum, the cultivated potato, is the third most important food crop after rice and wheat, and the main horticultural crop Devaux et al. (2014). The sequencing of the genome of S. tuberosum resulted in an assembly of 727 Mb of 810.6 Mb sequenced. Because most potato cultivars are autotetraploid (2n=4x=48) and highly heterozygous, the sequencing was performed on a homozygous doubled-monoploid potato clone. The most current potato genome assembly (4.03) contains 39.031 annotated genes and 62.2% of the genome corresponds to repetitive elements at scaffold level (Consortium et al., 2011).

Many attempts were made to discover repeats elements in Solanaceae families. Most of these studies were mainly focused on tandem repetitive elements, whereas studies of complex repetitive elements were mostly performed on limited groups of TEs (Mehra et al., 2015).The researchers assessed the complex repetitive elements on potato *(S. tuberosum*) and tomato *(S. lycopersicum)* genomes, identifying 629,713 and 589,561 repetitive elements, respectively. Mehra et al. used RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), which uses a repetitiveness-based strategy, leaving aside other approaches as mentioned above, but still enriching the amount of repeat families previously identified in the 4.03 version of the *S. tuberosum* genome released by the Potato Genome Sequencing Consortium (http://solanaceae.plantbiology.msu.edu/pgsc download.html) (PGSC DM v4.03) performed with RepeatMasker (Smit et al., 1996).

In this study we present an improved version of transposable elements detection and annotation from *S. tuberosum*. Our approach relies on the combination of the latest TE discovery de novo-based methods tools, with available TEs from Repbase and Illumina sRNA-seq data to obtain TEs from a broad spectrum of approaches. We then clustered the redundant elements from the different sources, blasted them against the genome to find copies and apply a series of filters depending on the TE family to obtain a comprehensive and curated whole-genome atlas of potato transposable elements. Furthermore, we provide to the research community our pipeline, annotation results and

files, which are publicly available at https://github.com/DiegoZavallo/TE_Discovery to encourage reproducibility and the eventual implementation of our framework in diverse organisms.

Results and Discussion

The goal of the present work was to inventory a comprehensive TE repertoire of the potato genome and provide legible, ready-to-use files for the scientific community. To address this issue, we searched TEs from sources that represent different approaches according to their detection method, such as publicly available data, similarity-based, structure-based and mapping-based strategies. Moreover, we introduce a set of scripts to gather, detect, filter and ultimately annotate TE copies from all classes across the potato genome and present *gff3* files of TE features. Additionally, we display the different types of TEs accumulation and distribution across the genome, a table resuming the data and metrics, such as distances to nearest gene and LTRs insertion ages.

TEs in the potato genome

Figure 1 shows an overview of the pipeline. First, *S. tuberosum* public TEs selected from the Repbase (Jurka et al., 2005), the potato reference genome and Illumina small RNA-seq data were used as **input data**. Then the **TEs identification** was performed by applying two “All TEs” tools: Tephra and TASR, which discover *de novo* TEs with structure-based and map-based approaches, respectively. To complement the search, in order to have at least two sources for each type of TEs we applied “Specific TEs” tools: MITE Tracker and SINE Scan. The obtained sequences were merged into multi-fasta files for each type of TEs and headers were renamed. **A clustering** step was performed with Vsearch to reduce redundancy (Data S1). Next, the **detection of copies** on the genome of the different TE families was achieved by conducting a BLAST search with specific parameters according to the type of TEs (see Experimental procedures section). A **copy filter** step was performed by establishing several rules to detect “potential autonomous TEs” with very astringent criteria and a second set of rules for “All TEs” copies content with more relax parameters to account for autonomous and non-autonomous TEs of all types (Table 1). Finally, an **annotation** step was executed to generate *gff3* files containing TE types and tool sources for each element on the description column of the files. The variability in TE content is huge in plants and usually

displays a correlation between genome size and TE coverage. For instance, as much as 85 % of maize genome or 70 % of Norway spruce genome (Nystedt et al., 2013) has been annotated as transposons, whereas only 21 % of the more compact *Arabidopsis thaliana* genome has been annotated as transposons (Ahmed et al., 2011).

The data presented by Mehra et al. (2015) comprised an annotation file of all the 1,061,377 repetitive elements, including the rRNA, tRNA, simple repeats and low complexity elements which represents almost 50% of the genome. To compare the percentage coverage of the more complex TE elements with our results, we parsed their annotation files by extracting 629,713 elements annotated as some class of TE and ran an *in-house* script, which extracts the coverage from the annotation file. The obtained file covered ~34% of the genome. Our pipeline revealed a TE content of ~16% TEs of potato genome (excluding the unanchored ChrUn). This result represents only half of the genomic coverage according to the data presented by Mehra et al. (2015). One explanation is that we applied a not very commonly used filtering step in TE discovery projects. This filter discarded too short, overlapping sequences and therefore retrieved reliable and precise TEs.

LTRs comprised over 13% of the potato genome, which corresponds to over 80% of the total TEs. The most abundant superfamily was Gypsy, whereas the other types of TEs barely made up 1% of the genome coverage. For instance, each DNA TEs (TIRs, MITEs and Helitrons) covered almost the same percentage of the genome with 0.51, 0.72 and 0.72 %, respectively. These coverage patterns are in accordance with most plant genomes that had TEs identification projects (Du et al., 2010; Andorf et al., 2015; Badouin et al., 2017; Alaux et al., 2018).

Table 2 resumes the amount and diversity of all identified TEs, filtered copies and proportion in the genome of all TE families. Beside the contrast on TE genome coverage compared with the one reported by Mehra et al. (2015), we can point out a few other differences.

First, we described all types of TEs families reported by Wicker et al. (2007), including LARDs, TRIMs and MITEs. Even with the extra family types of TEs, we annotated less than half of the sequences that they reported. We scored a total of 243,010 elements against the 629,713 complex elements. When the two sets were compared, a total of 198,025 (81%) of our sequences overlapped with their set, which evidences the effectiveness of both annotation pipelines.

However as already mentioned, they only used one tool relied on repetitiveness-based strategy to discover *de novo* TEs, leaving aside a wide variety of detection methods that could have been combined, which is recognized to improve TE detection efficiency (Kamoun et al., 2013; Hoen et al., 2015; Arensburger et al., 2016).

Moreover, we applied a clustering method to decrease redundancy and established a set of filters selected specifically for each type of TEs, which reduced the amount of TEs copies.

We believe that all these steps included in our pipeline retrieve more reliable and precise set of potato TEs sequences.

Finally, we deliver ready-to-use annotation files in *gff3* format of all TEs annotated with our pipeline with detailed descriptions in the ninth column including *TE id*, *source name*, *type*, *source* and *unique id* (Data S2).

Distribution of TEs across the potato genome

It is well know that the distribution, amount and coverage in the genome of TEs vary greatly, particularly between plants and animals, where LTRs and non-LTRs (LINEs and SINEs) are the predominant type of TEs, respectively Chalopin et al. (2015). Moreover, the distribution of TEs is highly dependent on the family type even within related species. Some TEs are more prone to concentrate near protein coding genes, while others are more equally distributed.

To assess the type-specific landscape of the diverse TEs categories, we performed circos ideograms for each TE type separated by class, as well as gene density as reference. Each concentric circle represents the coverage percentage of one type of TEs. Since they have different coverage ranges, each type has its own color pallet to appreciate the distribution across the chromosomes (Figure 2).

Left panel of Figure 2 shows the class I elements, where LTRs stand out, not only for their clear pattern of clustering around centromeric and pericentromeric regions, but also for their high coverage in some areas of the chromosomes. For instance, each dark red line of the second circle represent up to 36% of LTRs coverage per Mb, which explains the 13% of genome-wide coverage (Table 2). This behavior has already been reported for other plants (Baucom et al., 2009; Paterson et al., 2009; Badouin et al., 2017) and is like the negative image of the protein coding gene distribution (Figure 2, left panel).

At the other end, SINEs, which are the least represented type (besides LARDs which are not shown since we discovered only 18 elements) seem to have an evenly distribution pattern with a slight tendency towards telomeric and subtelomeric regions in some chromosomes. LINEs and TRIMs also display homogeneous distribution patterns with some coverage hotspots, mainly near telomeric regions. Due to their length and the copies detection filtering parameters chosen, only 248 LINEs elements were found in the genome, contrasting with the more than 50,000 elements found by Mehra et al.

(2015). As a reference, Heitkam et al. (2014) extracted 59,390 intact LINE sequences were extracted from 23 plant genomes in order to classify them into families. It is somehow unlikely that potato alone have 50,000 LINEs. A look at the data presented by Mehra et al. (2015) shows that some elements annotated as LINEs are small fragments with some identity to LINEs given by the RepeatMasker tool.

On a recent work, Gao et al. (2016) performed a comprehensive analysis of TRIMs in 48 plant genomes, including *S. tuberosum*. They observed that TRIMs are generally enriched in genic regions and they likely play a role in gene evolution Gao et al. (2016). They discovered 12,473 copies in the potato genome representing 0.46% of the genome, which is consistent with our results (Table 2).

Right panel of Figure 2 shows class II TEs distribution in which MITEs have a clear pattern of concentration around gene-rich subtelomeric regions. In accordance with previous works, MITEs are often found close to or within genes, where they affect gene expression (Bureau and Wessler, 1994b). Indeed, MITEs may affect gene regulation via small RNA pathways, but also have the capacity to alter gene structure in the traditional role of TEs in the evolution of new genes and transcriptome diversity (Yan et al., 2004; Kuang et al., 2009).

TIRs and Helitrons displayed an unbiased distribution across the chromosomes (Figure 2, right panel). Helitron chromosome distribution seems to be ambiguous of the plant species analyzed. While in *Arabidopsis* Helitrons are enriched in gene-poor pericentromeric region, in maize they are more abundant in gene-rich regions (Yang and Bennetzen, 2009). Rice, on the other hand, exhibited a less ordered pattern of Helitron distribution, with some pericentromeric regions rich in these elements and others less so (Yang and Bennetzen, 2009), more similar to our results.

In sum, this kind of analysis allowed us to have a holistic view of the different distribution patterns of TEs across the genome and to elucidate their potential role in transcriptional gene regulation.

TEs and genes

The importance of accumulation of TEs near genes, where these elements could influence gene expression has been extensively reported (Bureau and Wessler 1994b,a; Wang et al., 2013). As we described above, we found a strong correlation between LTRs with pericentromeric regions as well as a correlation between MITEs and rich-gene subtelomeric regions. In that sense, we interrogated the distances of the different types of elements to the closest gene in the genome to assess how many are likely within or close to genes. For this purpose, we extracted metrics such as median

distance to the closest gene, number of TEs overlapping gene transcripts as well as in the vicinity. LTRs are typically far from genes with a median distance of 10.25 kb, 9.01% overlapped to the coding sequence genes and only 5.11% are in the vicinity between 0-1 kb from the nearest gene, totaling a 32.23% within the first 5 kb (including elements within transcripts) (Figure 3 and Table S1). TRIM, as LTR-like TEs, show a similar behavior, with a median distance of 8.12kb, 6.43% within a gene and 6.19% in the range of 0-1 kb to the nearest transcript, comprising a 36.62% in the first 5 kb. In contrast, LINEs and SINEs are close to genes (median of 4.16 kb and 3.86 kb, respectively) with more than 20% within a transcript and rising up to >50% in the 0-5 Kb range (52.44% and 56.20% respectively).

Class II TEs also have more than 50% of their elements within the first 5 kb. However, the distribution varies between TIRs and MITEs with respect to Helitrons. TIRs and MITEs are the only types of TEs that appear to have less elements within a gene than in the 0-1 kb range (TIRs: 10.05% against 15.5% and a median distance of 3.32 kb; MITEs: 7.07% against 14.21% and a median distance of 3.53 kb). On the other hand, Helitrons display a similar pattern to LINEs and SINEs with a median of 4.56 kb and 23.65% within a transcript. These differences can be observed in the less pronounced curve of TIR and MITE histograms showing that a significant proportion of elements are indeed on the proximity of genes but not necessarily within them. This finding suggests that they might have an impact on their expression (Figure 3, right panel and Table S1).

Overall, DNA TEs and nonLTRs retrotransposons have more than 50% of the elements inserted within transcripts or in a range of 0-5 kb from the nearest gene. Several studies have reported examples of transcriptional impact due to TEs insertion near genes are known in tomato (Xiao et al., 2008; Quadrana et al., 2014), potato (Momose et al.,2010; Kloosterman et al., 2013), melon (Martin et al., 2009) and orange (Butelli et al., 2012) among others via several mechanisms including disruption of promoter or reduction of transcription through the spread of epigenetic silencing often suppressing expression. However, TE can also introduce new promoter sequences, leading to up-regulation of proximal gene (Yan et al., 2004; Cowley and Oakey, 2013; Du- bin et al., 2018). Insertion of a TE into the coding sequence can disrupt gene function, generally resulting in loss-of-function mutations, particularly if located in an exon. Intronic TEs can also be harmful, for instance by altering splicing patterns (Saze et al., 2013; Ong-Abdullah et al., 2015). In contrast, LTRs and TRIMs barely overcome 30% of the elements in the same range and, with some exceptions, they appear in intergenic, heterochromatic and gene-poor-regions (Kumar and Bennetzen, 1999). In *Arabidopsis*, Wang et al. (2013) reported that gene expression is positively correlated

with the distance of the gene to the nearest TE, and negatively correlated with the number of proximal TEs. Whether LTR-like TEs specifically target these regions or if they are simply not selected to be inserted nearby genes remain unclear (Sigman and Slotkin, 2016).

LTR elements age and evolution

LTRs are the most represented TEs in the majority of plants genome, encompassing more than 75% of the nuclear genome of some species (Kumar and Bennetzen, 1999; Paz et al., 2017). For this reason, we decided to explore the evolutionary history of LTRs during potato genome evolution.

We analyzed the age and distribution of the elements discovered by Tephra tool (Staton, 2018) since it has a command that allows the analysis of the rapidly evolving LTR sequences. *tephra ltrage* allowed the characterization of phylogenetic substructure within families of LTR retrotransposons. A total of 8,034 full-length LTRs were analyzed in terms of chromosomal distribution and insertion age by plotting all LTRs elements together (Figure 4, upper panel), and then grouped into superfamiles (Figure 4, bottom panel).

Younger LTRs are enriched in euchromatic subtelomeric regions and correspond mainly to Copia (RLC) family of TEs (average insertion age of 2.54 mya), whereas older LTRs are more abundant in heterochromatic pericentromeric regions where Gypsy (RLG) elements are mostly located (average insertion age of 4.14 mya). These distributions are in agreement with previous findings in maize Sun et al. (2018), wheat Luo et al. (2017) and tomato Paz et al. (2017).

Unclassified LTRs (RLX) display a more evenly distribution, although slightly towards to pericentromeric regions, as well as an intermediate insertion age (average insertion age of 3.78 mya) (Figure 4, bottom panel).

To determine whether these aging differences were a family trait or a different mutation/substitution rate dependent by region, we divided each chromosome in euchromatic and heterochromatic regions. In order to do so, we compared the pachytene karyotype published elsewhere (Consortium et al., 2011) with the chromosomal ideograms and plot the frequency of each LTR family by insertion age according to the region (Figure 5, upper panel). Gypsy family (RLG) display a Gaussian distribution centered around four millions years both in euchromatic and heterochromatic regions, whereas Copia family (RLC) have a chi-square distribution with a peak between two and three millions years on both regions. These results

suggest that the insertion age of LTRs retrotransposons depends on the superfamily and not on differential mutation/substitution rate by region.

Furthermore, we plot ten random independent families encompassing different member sizes from RLG, RLC and RLX to assess the age distribution of individual elements by region (Figure 5, bottom panel). Only four out of the ten evaluated Gypsy families showed significant differences (T-test $p<0.05$) in aging by region. In all the cases heterochromatic elements were older than euchromatic elements; which reflects a mild age difference within elements of the same family towing to their insertion sites. This suggests that insertions into heterochromatic regions are more likely to persist for longer periods of time.

Overall, these variances rely mildly on differential mutation/substitution rate by region but strongly on superfamily traits. A hypothesis presented by Pereira (2004) is based on differential insertion mechanism, where elements of the youngest family (Copia) insert randomly into the genome, but there is negative selection in the euchromatic regions and elements of the older families (Gypsy) preferentially insert into the pericentromeric regions Pereira (2004).

Conclusion

TEs have been historically neglected in genome assembly projects, partially due to their repetitive nature but also because for their heterogeneity in sequence, size, number of copies, distribution and mutation rates both inter and intra species, which make them very difficult to detect accurately (Bourque et al., 2018). However, in the past years, TEs discovery and annotation for the main crops have been emerging with diverse results of coverage, complexity and accuracy. For instance, by using the CLARITE software on the IWGSC RefSeq v1.0 genome assembly, Alaux et al. (2018) found over 5 million elements from all types of TEs in wheat. On the other hand, maize repeat elements were annotated with the Ensembl Genomes repeat feature pipeline which comprised the discovery of low complexity regions (Dust) (Morgulis et al., 2006), tandem repeat (TRF) (Benson et al., 1999) and complex repeats (RepeatMasker) (Smit et al., 1996) but lacking on TEs classes and types categorization (Andorf et al., 2015).

Soybean represents an example of a thoroughly annotated TEs database. The SoyTEdb provides resources and information by using a combination of structure-based and homology-based approaches, where TEs were structurally annotated and categorized (Du et al., 2010).

As we described above, potato has some TE sequences annotated on the RepBase, which is a list of repeat families based on RepeatMasker provided by the Potato Genome Sequencing Consortium and the work from Mehra et al. (2015), which in turn is based on RepeatModeler. However, this important worldwide crop still lacked a comprehensive, multiple-based discovery approach oriented to transposon elements annotation.

This work arises as a need of a reliable potato TE atlas for ongoing projects that involved epigenetic and transcriptional regulation for different *Solamun tuberosum* hybrids in contrasting environments.

Plants are known for their phenotypic plasticity and good adaptation to environmental changes due to their sessile condition. There is an increasing evidences of the impact that TEs have on the transcriptome on a given condition. For instance, TEs may have direct effects on genes, by providing them with new coding or regulatory sequences, an effect on the epigenetic status of the chromatin close to genes, and more subtle effects by imposing diverse evolutionary constraints to different chromosomal regions (Makarevitch et al., 2015; Vicient and Casacuberta, 2017; Cambiagno et al., 2018). Finding common patterns or specific alterations on these features (TEs) and linking them with sRNAs profiles and DNA methylation patterns could help researchers to elucidate the underlying mechanisms of transcriptional changes during different conditions.

As mentioned above, the current data regarding TEs on potato genome is either scarce or unclear for transcriptional analysis. Given the importance of transposon elements, we provide here a comprehensive potato TEs landscape, based on a wide variety of identification tools and approaches, clustering methods, copies detection, filtering rules and clear outputs, that the scientific community will likely use for metadata analysis.

Experimental procedures

Input data

Latest assembly version of *Solanum tuberosum* genome sequence was downloaded from the Potato Genome Sequencing Consortium (PGSC v4.03 of the doubled monoploid *S. tuberosum* Group Phureja DM1-3). Potato TEs sequences from Repbase Giri (Jurka et al., 2005) were downloaded prior registration. Note that LTRs transposons are divided in LTR (Long Terminal Repeat) and I (Internal) sequences, hence must be concatenated. We gathered 126 LTRs, 18 LINEs, 2 SINEs and 42 TIRs family sequences (https://www.girinst.org/repbase/update/search.php?query=tuberosum\&querytype=Taxonoy)

Illumina sRNA-seq data for TASR run was generated by our lab (data unpublished); however, any available data from public repositories such as SRA (https://www.ncbi.nlm.nih.gov/sra) could be used as input.

Transposable elements identification

TEs obtained from different sources were merged together according to each element classification. Tephra, which uses several structure-based tools, was applied to harvest different kind of TEs in the potato genome. Tephra all command (which run all subcommands) was executed using teprhra config.yml file with default configuration parameters with the exception of S. tuberosum TEs from Repbase for the repeatdb parameter. A total of 1,325 Helitrons, 7,694 LTRs, 2,994 MITEs, 2,414 TIRs and 7,011 TRIMs families were found with this program. No non-LTR TEs (LINEs and SINEs) were found by Tephra. TASR El Baidouri et al. (2015), a tool for de novo discovery of TEs using small RNA data, was also used in this work. 21, 22 and 24nt lenght sRNAs were parsed from files belonging to all treatments and replicates from our sRNA-Seq data and subsequently concatenated them together to be used as input. TASR.v.1.1.pl perl script was run with default parameters except for: *-cpu 14, -nsirna 10, and -cnumber 5.* A total of 1,916 families were found, presented as multifasta files comprising all the elements for each family. A perl script provided by TASR developer was run in order to generate a consensus sequence for each family. Then we created a single multifasta file with the entire consensus. The PASTEC tool (Hoede et al., 2014) was used for this purpose, since TASR does not classify TEs into categories. From the 1,916 families discovered, a total of 891 LTRs, 49 LINEs, 15 SINEs, 9 TRIMs, 2 LARDs, 84 TIRs, 35

MITEs and 5 Helitrons were classified. For MITEs discovery, MITE Tracker (Crescente et al., 2018) was used using default parameters. A total of 1,045 MITEs elements were detected. SINE (Scan Mao and Wang, 2016), an efficient de novo tool to discover SINEs was also used in this work with default parameters. A total of 13 SINEs families were detected. As a result, we obtained 8,711 LTRs elements from Tephra and TASR y Repbase, 67 LINEs elements from TASR y Repbase, 30 SINEs elements from TASR, Repbase and SINE Scan, 540 TIRs elements from Tephra, TASR and Repbase, 4,074 MITEs elements from Tephra and TASR y MITE Tracker and 1,330 Helitrons elements from Tephra and TASR.

Pipeline description

The obtained TEs were subjected to an in house pipeline to detect, filter and annotate TEs copies across the potato genome. The pipeline was developed using bash scripts and Jupyter notebooks.

1.1 add annotation.ipynb uses TEs sequences retrieved from each program and merge them together into one multi-fasta file per studied TE type (LTRs, LINEs, SINEs, TRIMs, LARDs, TIRs, MITEs and Helitrons). This script adds to each sequence a unique identifier containing an auto-incremented number, TE classification and the program source from which it was obtained.

2.1 vsearch.sh uses VSearch program to cluster similar sequences that share 80% identity, according to the 80/80/80 rule in the study of Wicker et al. (2007). It is executed once per TE type, thus obtaining as a result TEs clustered by type.

2.2 vsearch_merge.ipynb This script uses vsearch outputs to create fasta files containing one TE per family. It also adds family description indicating from which program the members came from.

3.1 blast.sh performes a genome-wide BLASTn search using the files from the previous step and searches for TE elements in the potato genome. For this task, this script uses the following parameters: *-perc identity 80, -evalue 10e-3 and -task blastn* (except for LTRs).- *qcov hsp perc 80* was used for SINEs, TRIMs and MITEs, whereas -*qcov hsp perc 50* was set for the rest.

3.2 blast filter.ipynb filters BLASTn results with parameters according to Table 1. First, each file is filtered by a length range defined by min len and max len. Afterwards, a length threshold range is calculated by multiplying the query length and a min threshold and max threshold. The subject sequence has to be inside this range to be considered valid. Later, minimum identity percentage and query coverage are required

for the sequences to be valid. Ultimately, duplicated hits are removed by searching those whose start and end positions overlaps within a margin of plus or minus 5 nt.

4.1 annotate.ipynb transforms the BLAST tab-delimited results to a *gff3* format file, adding a detailed description for each TE. The description includes TE id (a numeric identifier of the element after clustering) source name (original id name of the element before clustering) type (family type of the element), source (program or tool from which the TEs was detected) and unique id (unique identifier for copy element).

LTR age

To determine the tentative LTRs insertion age we uses *tephra ltrage* command implemented by the Tephra package. This command uses the LTR-RT element of full-lenght LTRs discovered by Tephra, aligns the LTR sequences and generates a neighbor-joining guide tree with MUSCLE Edgar (2004). The alignment and guide tree are used to generate an alignment in PHYLIP format. A likelihood divergence estimate was calculated with baseml from PAML Yang (2007) by specifying the K80 substitution model. This divergence value (hereafter d) was used to calculate LTRRT age with the formula T = d/2r, where r= 1e8 is the default substitution rate.

Data Statement

Scripts from this work, including all pipeline steps as well as circos ideogram, distance histogram and LTR age plot scripts are available at (https://github.com/DiegoZavallo/TE Discovery). Annotation and fasta files are available as supporting information.

Acknowledgements

The authors would like to thank Dr. Julia Sabio y Garcia for the assistance with English-language editing, and Humberto Julio Debat for his critical comments on the manuscript. This work was supported by the Instituto Nacional de Tecnología Agropecuaria (INTA) and by ANPCyT PICT 2015-1532, and PICT 2016-0429. The funders had no role in this study design, data collection and analysis, decision to publish or preparation of the manuscript.

Supporting Information

Data S1. Fasta files containing family sequences for each type transposon elements.

Data S2. gff3 file containing all TEs copy coordinates including Chr00.

Table S1. Table resuming metrics of TEs to the nearest genes.

References

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., and Quesneville, H. (2011). Genome-wide evidence for local dna methylation spreading from small rna- targeted sequences in arabidopsis. Nucleic Acids Research, 39(16):6919–6931.

Alaux, M., Rogers, J., Letellier, T., Flores, R., Alfama, F., Pommier, C., Mo- hellibi, N., Durand, S., Kimmel, E., Michotey, C., et al. (2018). Linking the international wheat genome sequencing consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome biology, 19(1):111.

Andorf, C. M., Cannon, E. K., Portwood, J. L., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Braun, B. L., Campbell, D. A., Vinnakota, A. G., Sribalusu, V. V., et al. (2015). Maizegdb update: new tools, data and interface for the maize model organism database. Nucleic acids research, 44(D1):D1195– D1201.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 361(6403):eaar7191.

Arensburger, P., Pi´egu, B., and Bigot, Y. (2016). The future of transposable element annotation and their classification in the light of functional genomics- what we can learn from the fables of jean de la fontaine? Mobile genetic elements, 6(6):e1256852.

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Bri`ere, C., Owens, G. L., Carr`ere, S., Mayjonade, B., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. Nature, 546(7656):148.

Baucom, R. S., Estill, J. C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., Westerman, R. P., SanMiguel, P. J., and Bennetzen, J. L. (2009). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the b73 maize genome. PLoS Genetics, 5(11):e1000732.

Benson, G. et al. (1999). Tandem repeats finder: a program to analyze dna sequences. Nucleic acids research, 27(2):573–580.

Berthelier, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., and Car- rier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga tisochrysis lutea. BMC genomics, 19(1):378.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvak, Z., Levin, H. L., Macfarlan, T. S., et al. (2018). Ten things you should know about transposable elements. Genome biology, 19(1):199.

Bureau, T. E. and Wessler, S. R. (1994a). Mobile inverted-repeat elements of the tourist family are associated with the genes of many cereal grasses. Proceedings of the National Academy of Sciences, 91(4):1411–1415.

Bureau, T. E. and Wessler, S. R. (1994b). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. The Plant Cell, 6(6):907–916.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., and Martin, C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. The Plant Cell, 24(3):1242–1255.

Cambiagno, D. A., Nota, F., Zavallo, D., Rius, S., Casati, P., Asurmendi, S., and Alvarez, M. E. (2018). Immune receptor genes and pericentromeric trans- posons as targets of common epigenetic regulatory elements. The Plant Jour- nal, 96(6):1178–1190.

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N. (2015). Com- parative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome biology and evolution, 7(2):567–580.

Consortium, P. G. S. et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature, 475(7355):189.

Cowley, M. and Oakey, R. J. (2013). Transposable elements re-wire and fine- tune the transcriptome. PLoS genetics, 9(1):e1003234.

Crescente, J. M., Zavallo, D., Helguera, M., and Vanzetti, L. S. (2018). Mite tracker: an accurate approach to identify miniature inverted-repeat transpos- able elements in large genomes. BMC bioinformatics, 19(1):348.

Devaux, A., Kromann, P., and Ortiz, O. (2014). Potatoes for sustainable global food security. Potato Research, 57(3-4):185–199.

Du, J., Grant, D., Tian, Z., Nelson, R. T., Zhu, L., Shoemaker, R. C., and Ma, J. (2010). Soytedb: a comprehensive database of transposable elements in the soybean genome. Bmc Genomics, 11(1):113.

Dubin, M. J., Scheid, O. M., and Becker, C. (2018). Transposons: a blessing curse. Current opinion in plant biology, 42:23–29.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5):1792–1797.

El Baidouri, M., Kim, K. D., Abernathy, B., Arikit, S., Maumus, F., Panaud, O., Meyers, B. C., and Jackson, S. A. (2015). A new approach for annotation of transposable elements using small rna mapping. Nucleic acids research, 43(13):e84–e84.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. BMC bioinformatics, 9(1):18.

Gao, D., Li, Y., Do Kim, K., Abernathy, B., and Jackson, S. A. (2016). Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome biology, 17(1):7.

Gremme, G., Steinbiss, S., and Kurtz, S. (2013). Genometools: a compre- hensive software library for efficient processing of structured genome annota- tions. IEEE/ACM Transactions on Computational Biology and Bioinformat- ics (TCBB), 10(3):645–656.

Heitkam, T., Holtgrawe, D., Dohm, J. C., Minoche, A. E., Himmelbauer, H., Weisshaar, B., and Schmidt, T. (2014). Profiling of extensively diversified plant line s reveals distinct plant-specific subclades. The Plant Journal, 79(3):385–397.

Hirsch, C. D. and Springer, N. M. (2017). Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1860(1):157–165.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). Pastec: an automatic transposable element clas- sification tool. PloS one, 9(5):e91929.

Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., et al. (2015). A call for benchmarking transposable element annotation methods. Mobile DNA, 6(1):13.

Hosaka, A. and Kakutani, T. (2018). Transposable elements, genome evolution and transgenerational epigenetic variation. Current opinion in genetics & development, 49:43–48.

Judd, J. and Feschotte, C. (2018). Gene expression: Transposons take remote control. eLife, 7:e40921.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and genome research, 110(1-4):462–467.

Kamoun, C., Payen, T., Hua-Van, A., and Fil´ee, J. (2013). Improving prokary- otic transposable elements identification using a combination of de novo and profile hmm methods. BMC genomics, 14(1):700.

Kloosterman, B., Abelenda, J. A., Gomez, M. d. M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., Horvath, B. M., van Eck, H. J., Smaczniak, C., Prat, S., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature, 495(7440):246.

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P. B., Ouyang, S., Jiang, J., Buell, C. R., and Baker, B. (2009). Identification of miniature inverted-repeat transposable elements (mites) and biogenesis of their sirnas in the solanaceae: new functional implications for mites. Genome research, 19(1):42–56.

Kumar, A. and Bennetzen, J. L. (1999). Plant retrotransposons. Annual review of genetics, 33(1):479–532.

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y., et al. (2017). Genome sequence of the progenitor of the wheat d genome aegilops tauschii. Nature, 551(7681):498.

Macas, J., Meszaros, T., and Nouzova, M. (2002). Plantsat: a specialized database for plant satellite repeats. Bioinformatics, 18(1):28–35.

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross- Ibarra, J., and Springer, N. M. (2015). Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS genetics, 11(1):e1004915.

Mao, H. and Wang, H. (2016). Sine scan: an efficient tool to discover short interspersed nuclear elements (sines) in large-scale genomic datasets. Bioin- formatics, 33(5):743–745.

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., Pitrat, M., Dogimont, C., and Bendahmane, A. (2009). A transposon- induced epigenetic change leads to sex determination in melon. Nature, 461(7267):1135.

McCarthy, E. M. and McDonald, J. F. (2003). Ltr struc: a novel search and identification program for ltr retrotransposons. Bioinformatics, 19(3):362–367.

Mehra, M., Gangwar, I., and Shankar, R. (2015). A deluge of complex repeats: the solanum genome. PloS one, 10(8):e0133962.

Momose, M., Abe, Y., and Ozeki, Y. (2010). Miniature inverted-repeat trans- posable elements of stowaway are active in potato. Genetics, 186(1):59–66.

Morgulis, A., Gertz, E. M., Schaffer, A. A., and Agarwala, R. (2006). A fast and symmetric dust implementation to mask low-complexity dna sequences. Journal of Computational Biology, 13(5):1028–1040.

Nussbaumer, T., Martis, M. M., Roessner, S. K., Pfeifer, M., Bader, K. C., Sharma, S., Gundlach, H., and Spannagl, M. (2012). Mips plantsdb: a database framework for comparative plant genome research. Nucleic acids research, 41(D1):D1144–D1151.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The norway spruce genome sequence and conifer genome evolution. Nature, 497(7451):579.

Ong-Abdullah, M., Ordway, J. M., Jiang, N., Ooi, S.-E., Kok, S.-Y., Sarpan, N., Azimi, N., Hashim, A. T., Ishak, Z., Rosli, S. K., et al. (2015). Loss of karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature, 525(7570):533.

Ouyang, S. and Buell, C. R. (2004). The tigr plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic acids research, 32(suppl 1):D360–D363.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The sorghum bicolor genome and the diversification of grasses. Nature, 457(7229):551.

Paz, R. C., Kozaczek, M. E., Rosli, H. G., Andino, N. P., and Sanchez-Puerta, M. V. (2017). Diversity, distribution and dynamics of full-length copia and gypsy ltr retroelements in solanum lycopersicum. Genetica, 145(4-5):417–430.

Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the *arabidopsis thaliana* genome. Genome biology, 5(10):R79.

Platt, R. N., Blanco-Berdugo, L., and Ray, D. A. (2016). Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome biology and evolution, 8(2):403–410.

Quadrana, L., Almeida, J., As´ıs, R., Duffy, T., Dominguez, P. G., Bermu´dez, L., Conti, G., Da Silva, J. V. C., Peralta, I. E., Colot, V., et al. (2014). Natural occurring epialleles determine vitamin e accumulation in tomato fruits. Nature Communications, 5:4027.

Roessler, K., Bousios, A., Meca, E., and Gaut, B. S. (2018). Modeling interactions between transposable elements and the plant epigenetic response: a surprising reliance on element retention. Genome biology and evolution, 10(3):803–815.

Saze, H., Kitayama, J., Takashima, K., Miura, S., Harukawa, Y., Ito, T., and Kakutani, T. (2013). Mechanism for full-length rna processing of arabidopsis genes containing intragenic heterochromatin. Nature communications, 4:2301.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009). The b73 maize genome: complexity, diversity, and dynamics. science, 326(5956):1112–1115.

Sigman, M. J. and Slotkin, R. K. (2016). The first rule of plant transposable element silencing: location, location, location. The Plant Cell, 28(2):304–313.

Smit, A. F., Hubley, R., and Green, P. (1996). Repeatmasker.

Staton, E. (2018). Tephra a tool for discovering transposable elements and describing patterns of genome evolution.

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., Song, W., Zhang, M., Cui, Y., Dong, X., et al. (2018). Extensive intraspecific gene order and gene structural variations between mo17 and other maize genomes. Nature genetics, 50(9):1289.

Vicient, C. M. and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. Annals of Botany, 120(2):195–207.

Wang, X., Weigel, D., and Smith, L. M. (2013). Transposon variants and their effects on gene expression in arabidopsis. PLoS genetics, 9(2):e1003255.

Wicker, T., Matthews, D. E., and Keller, B. (2002). Trep: a database for triticeae repetitive elements. Trends in Plant Science, 7(12):561–562.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12):973.

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

Xiong, W., He, L., Lai, J., Dooner, H. K., and Du, C. (2014). Helitronscanner uncovers a large overlooked cache of helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences, page 201410068.

Yan, L., Helguera, M., Kato, K., Fukuyama, S., Sherman, J., and Dubcovsky, J. (2004). Allelic variation at the vrn-1 promoter region in polyploid wheat. Theoretical and applied genetics, 109(8):1677–1686.

Yang, L. and Bennetzen, J. L. (2009). Distribution, diversity, evolution, and survival of helitrons in the maize genome. Proceedings of the National Academy of Sciences, 106(47):19922–19927.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24(8):1586–1591.

Table 1: TEs filtering parameters. Parameters established to filter copies element for "Autonomous TEs" and "All TEs" which includes autonomous and non-autonomous elements. *min-len*: minimum element length; *max-len*: maximum element length; *min-pid*: minimum query identity percentage; *min-threshold*: minimum threshold length between query and subject; *max-threshold*: maximum threshold length between query and subject; *min-cov*: minimum query coverage percentage; *overlap*: margin of plus or minus nucleotides overlap between query and subject to be considered as duplicates.

Family	min-len (nt)	max-len (nt)	min-pid (%)	min-threshold (ratio)	max-threshold (ratio)	min-cov (%)	overlap (nt)
Autonomous TEs							
LTR	650	∞	95	0.9	1.1	95	5
LINE	1500	∞	95	0.9	1.1	95	5
TIR	700	∞	95	0.9	1.1	95	5
Helitron	2000	∞	95	0.9	1.1	95	5
All TEs							
LTR	650	∞	80	0.5	1.5	50	5
LINE	1500	∞	80	0.5	1.5	50	5
SINE	150	800	80	0.8	1.2	80	5
TRIM	600	∞	90	0.9	1.1	90	5
LARD	4000	∞	80	0.5	1.5	50	5
TIR	700	∞	80	0.5	1.5	50	5
MITE	50	800	80	0.85	1.15	90	5
Helitron	2000	∞	80	0.5	1.5	50	5

Table 2: Quantity and diversity of all identified TEs. Number of TE families, number of TE copies and TEs genome coverage (%) of all types of TEs.

Class	Order/Family	TEs identified	TEs copies	Genome coverage (%)
I	LTR/Copia	2,541	13,044	2.46
I	LTR/Gypsy	5,736	170,380	10.68
I	LTR/Unclassified	29	604	0.06
Total LTRs		8,306	184,028	13.2
I	LINE	65	248	0.1
I	SINE	24	2,477	0.07
I	TRIM	6,908	13,491	0.67
I	LARD	2	18	0.002
Total Class I		15.305	200.262	14.04
II	TIR/hAT	112	382	0.05
II	TIR/Mariner	1,466	1,591	0.29
II	TIR/Harbiner	16	219	0.02
II	TIR/Mutator	575	941	0.13
II	TIR/CACTA	131	117	0.02
II	TIR/Unclassified	3	110	0.01
Total TIRs		2,303	3,380	0.51
II	MITEs	3,515	38,205	0.72
II	Helitrons	1,322	1,163	0.72
Total Class II		7,140	42,748	1.99
Total TEs		22,445	243,010	16.03

Figure legends

Figure 1. Overview of the TE Discovery pipeline. *1_Input data*: last genome assembly, available TEs sequences from Repbase Giri and sRNA-seq Illumina data are used as input data. *2_TEs family detection*: the detection of putative TEs is performed by using four tools combining two detection approaches. The resulted sequences are merged into multi-fasta files for each type of TEs. *3_TEs clustering*: Each TE type sequences are clustered with Vsearch to reduce redundancy. *4_TEs copy detection*: blastn is performed to the clustered sequences with specific parameters according to each type of TEs to detect copies across the genome. *5_TEs copy filter*: the detected copies are subjected to filtering steps to detect ”potential autonomous TEs” and/or ”All TEs” including non- autonomous TEs. *6_TEs annotation*: TEs annotation *gff3* files are generated for each type of TEs with detailed descriptions and merged into one single file.

Figure 2. Comprehensive circos ideograms of TEs of the potato genome. Left panel: Retrotransposon TEs (Class I). Right panel: DNA TEs (Class II). Each concentric circle represents a different type of TE with their own color pallet and range of coverage so the distribution across the chromosomes can be appreciated. Each line of within the circles represents coverage percentage per Mb. (1) Gene distribution. (a) LTR distribution. (b) SINE distribution. (c) LINE distribution. (d) TRIM distribution. (e) TIR distribution. (f) MITE distribution. (g) Helitron distribution.

Figure 3. Frequency histograms of TEs distance to the nearest gene. Left panel: Retrotransposon TEs (Class I). Right panel:DNA TEs (Class II). Left side of the red line represents the TEs distance within the first 5kb to the nearest gene.

Figure 4. Chromosomal ideograms of LTRs age per family. Upper panel: chromosomal distribution of insertion age of all superfamilies of full-length LTR retrotransposons. Lower panel: chromosomal distribution of insertion age of Gypsy, Copia and Unclassified LTR retrotransposons separately.

Figure 5. Number of each LTR family by insertion age according to genome chromatin state. Upper panel: frequency histrograms of Gypsy, Copia and Unclassified LTR families by their insertion age (in millons of years) separated by heterochromatin regions (blue) and euchromatic regions (red). Lower panel: Scatter plots of ten random independent families from Gypsy, Copia and Unclassified LTRs assessing the age

distribution of individual elements by their heterochromatin (blue) and euchromatin (red) state.

Figure 1

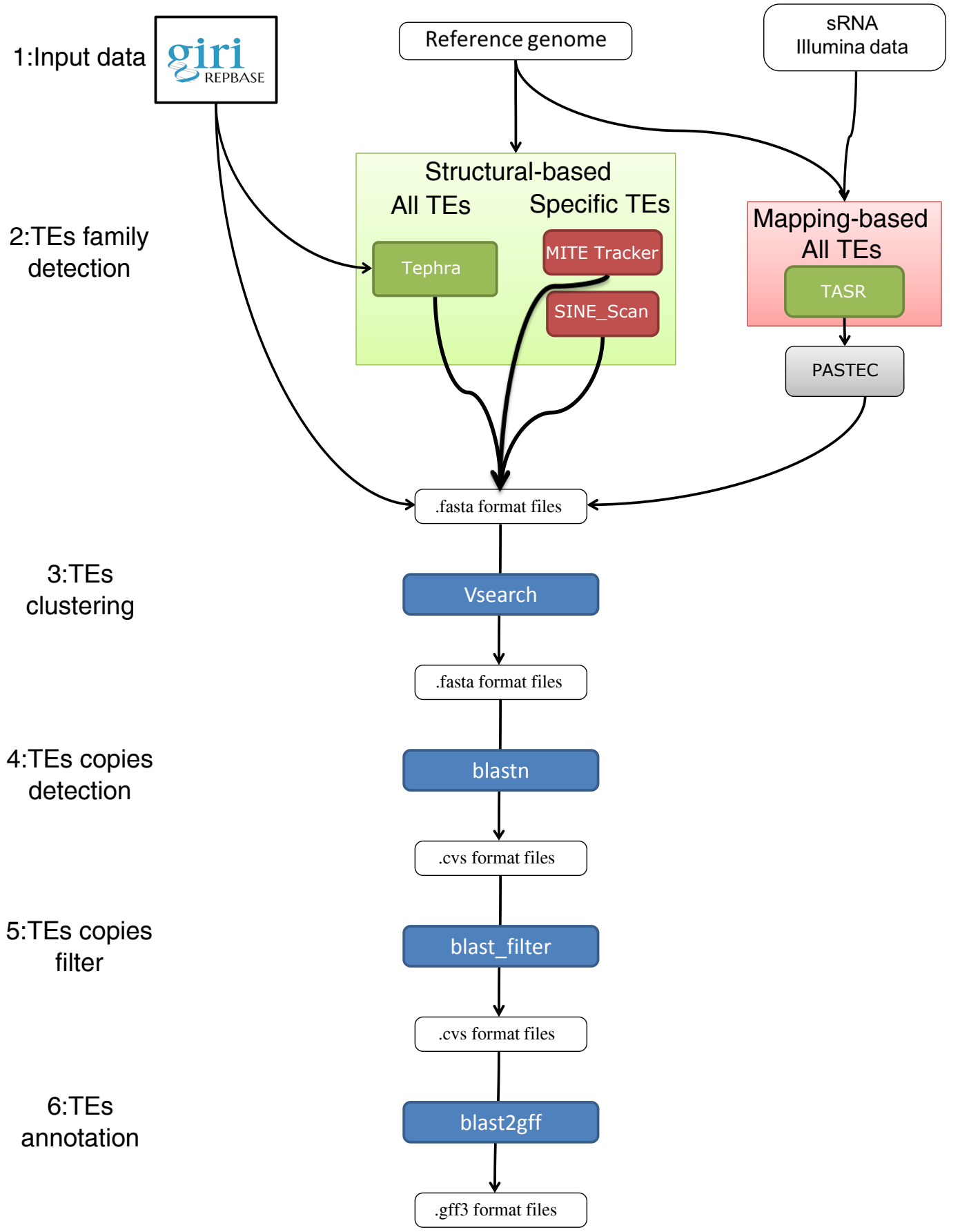
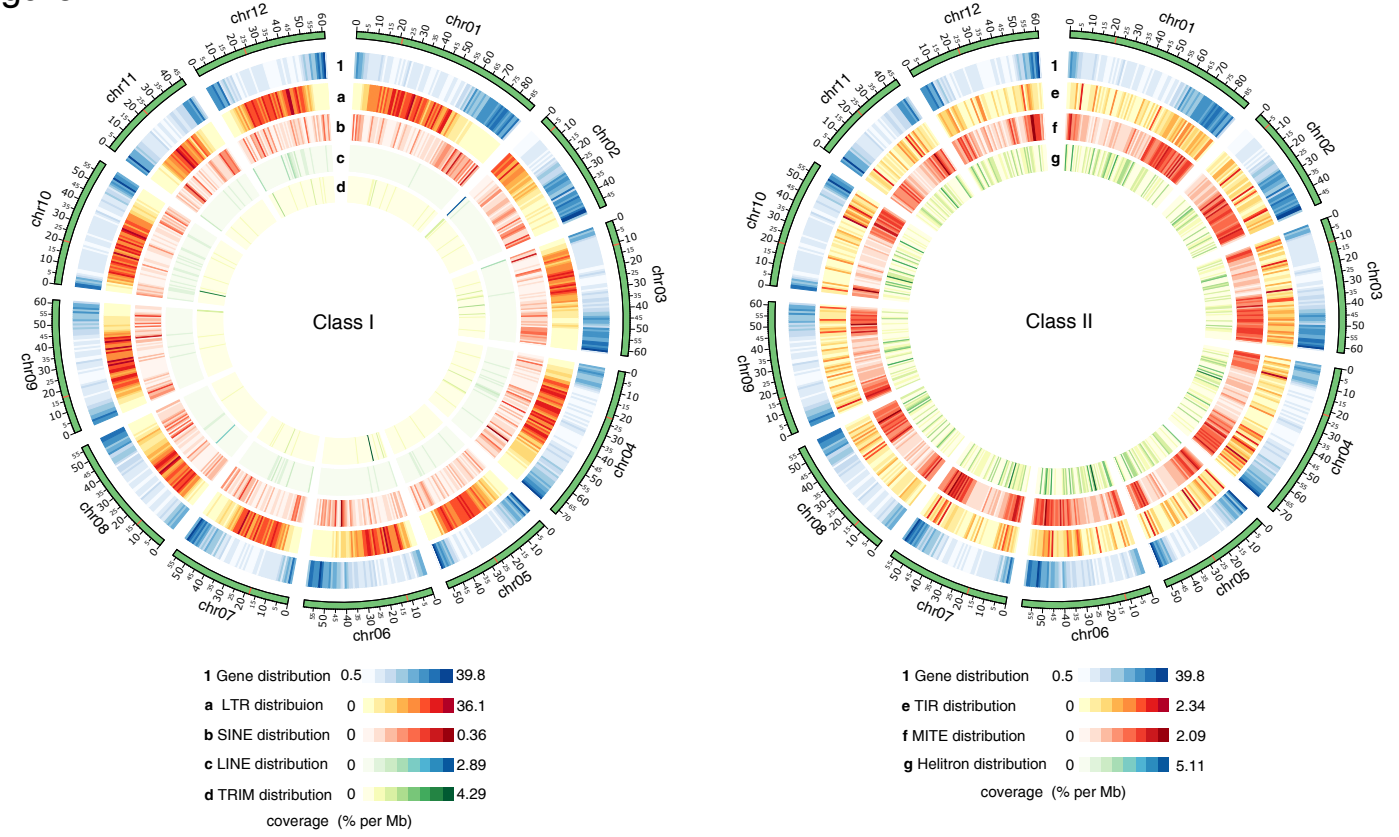

1:Input data
giri
REPBASE
Reference genome
sRNA
Illumina data
2:TEs family detection
Structural-based
All TEs
Specific TEs
Tephra
MITE Tracker
SINE_Scan
Mapping-based
All TEs
TASR
PASTEC
.fasta format files
3:TEs clustering
Vsearch
.fasta format files
4:TEs copies detection
blastn
.cvs format files
5:TEs copies filter
blast_filter
.cvs format files
6:TEs annotation
blast2gff
.gff3 format files

Figure 2


Class I
Class II
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
1 Gene distribution 0.5 39.8
a LTR distribuion 0 36.1
b SINE distribution 0 0.36
c LINE distribution 0 2.89
d TRIM distribution 0 4.29
coverage (% per Mb)
1 Gene distribution 0.5 39.8
e TIR distribution 0 2.34
f MITE distribution 0 2.09
g Helitron distribution 0 5.11
coverage (% per Mb)

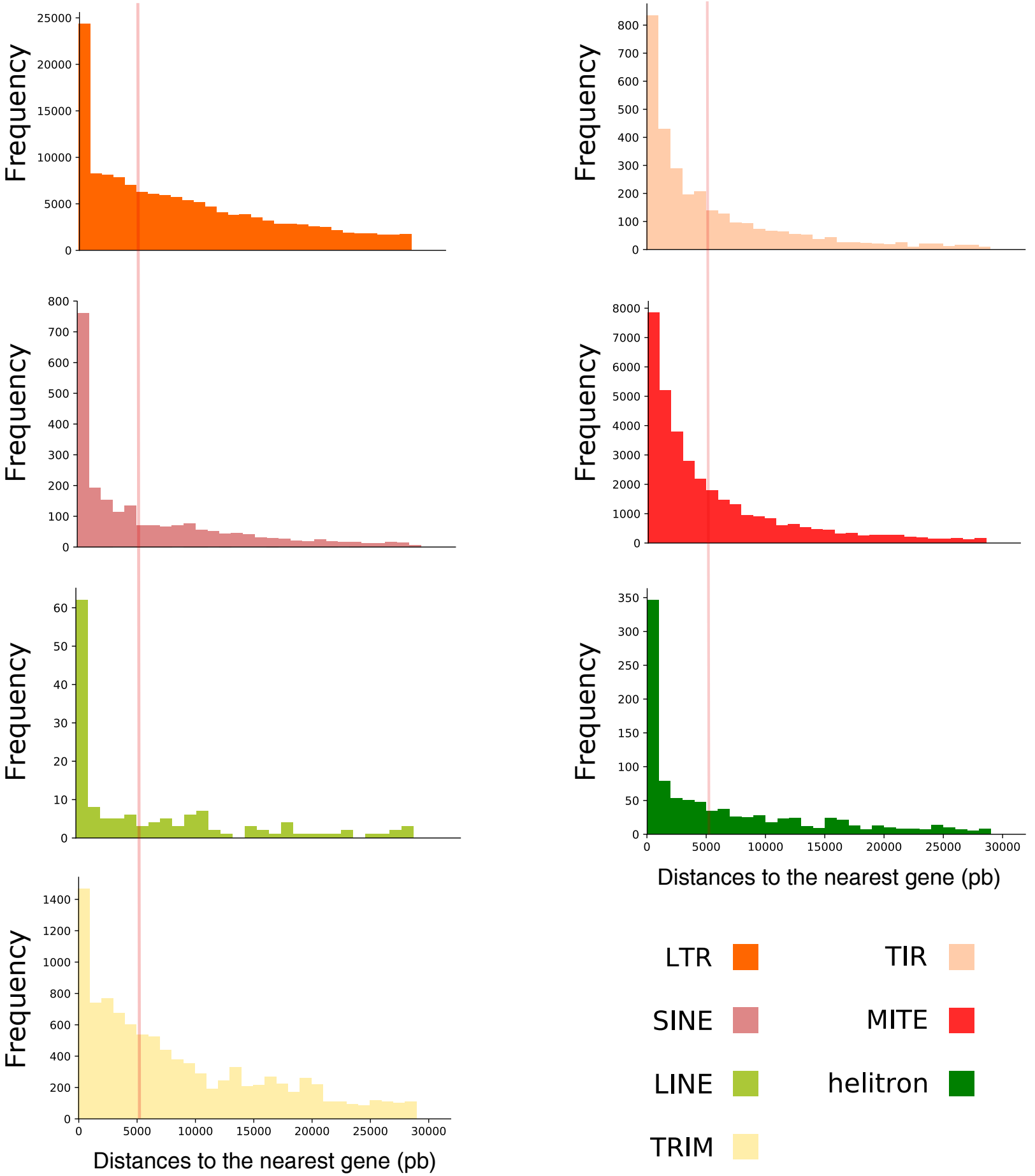

Frequency
Distances to the nearest gene (pb)
LTR
TIR
SINE
MITE
LINE
helitron
TRIM

Figure 3

Figure 4

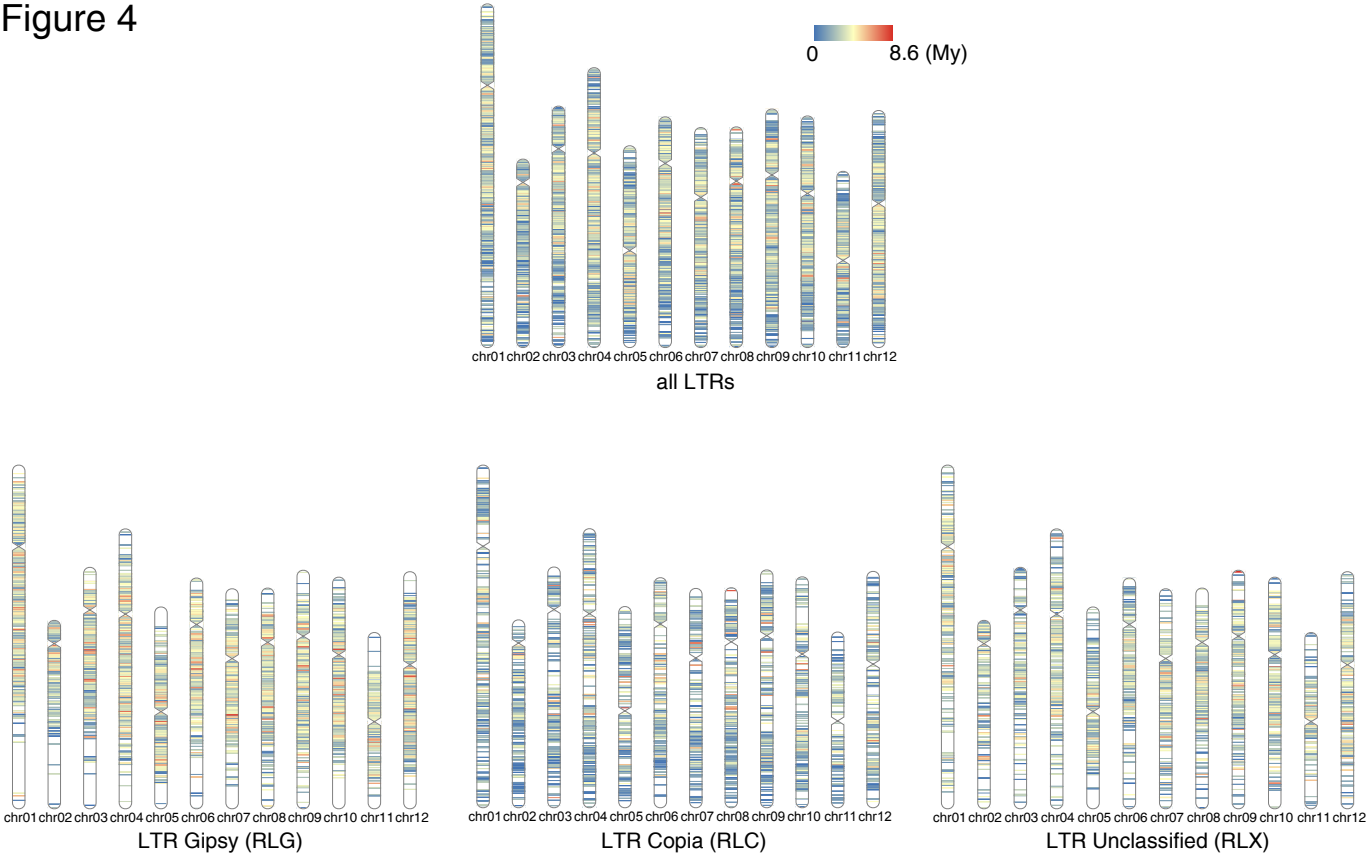
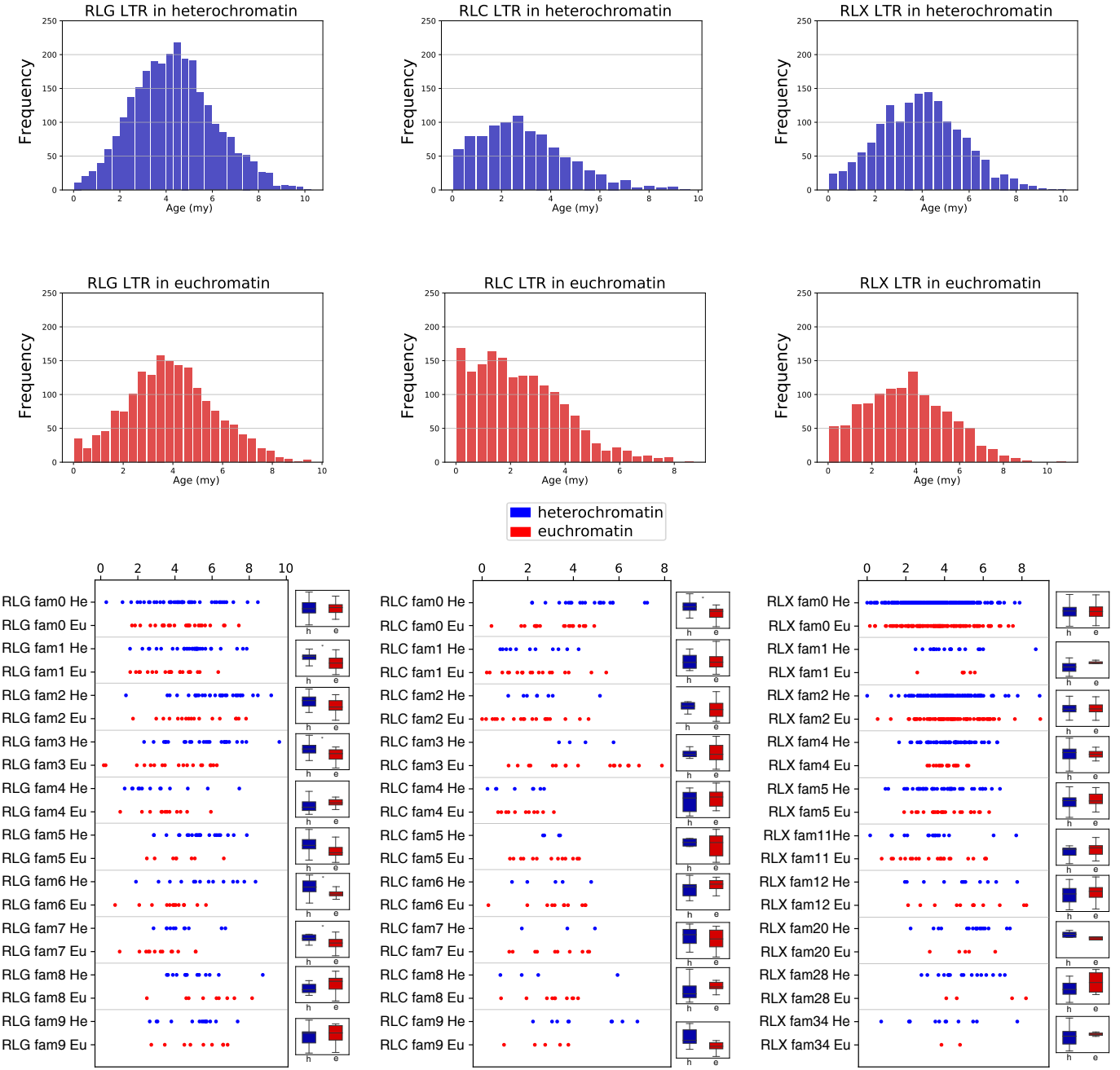

0
8.6 (My)
chr01 chr02 chr03 chr04 chr05 chr06 chr07 chr08 chr09 chr10 chr11 chr12
all LTRs
chr01 chr02 chr03 chr04 chr05 chr06 chr07 chr08 chr09 chr10 chr11 chr12
LTR Gipsy (RLG)
chr01 chr02 chr03 chr04 chr05 chr06 chr07 chr08 chr09 chr10 chr11 chr12
LTR Copia (RLC)
chr01 chr02 chr03 chr04 chr05 chr06 chr07 chr08 chr09 chr10 chr11 chr12
LTR Unclassified (RLX)

Figure 5


RLG LTR in heterochromatin
RLC LTR in heterochromatin
RLX LTR in heterochromatin
RLG LTR in euchromatin
RLC LTR in euchromatin
RLX LTR in euchromatin
Frequency
Age (my)
heterochromatin
euchromatin
RLG fam0 He
RLG fam0 Eu
RLG fam1 He
RLG fam1 Eu
RLG fam2 He
RLG fam2 Eu
RLG fam3 He
RLG fam3 Eu
RLG fam4 He
RLG fam4 Eu
RLG fam5 He
RLG fam5 Eu
RLG fam6 He
RLG fam6 Eu
RLG fam7 He
RLG fam7 Eu
RLG fam8 He
RLG fam8 Eu
RLG fam9 He
RLG fam9 Eu
RLC fam0 He
RLC fam0 Eu
RLC fam1 He
RLC fam1 Eu
RLC fam2 He
RLC fam2 Eu
RLC fam3 He
RLC fam3 Eu
RLC fam4 He
RLC fam4 Eu
RLC fam5 He
RLC fam5 Eu
RLC fam6 He
RLC fam6 Eu
RLC fam7 He
RLC fam7 Eu
RLC fam8 He
RLC fam8 Eu
RLC fam9 He
RLC fam9 Eu
RLX fam0 He
RLX fam0 Eu
RLX fam1 He
RLX fam1 Eu
RLX fam2 He
RLX fam2 Eu
RLX fam4 He
RLX fam4 Eu
RLX fam5 He
RLX fam5 Eu
RLX fam11He
RLX fam11 Eu
RLX fam12 He
RLX fam12 Eu
RLX fam20 He
RLX fam20 Eu
RLX fam28 He
RLX fam28 Eu
RLX fam34 He
RLX fam34 Eu
h
e