Berthelier et al. BMC Genomics (2018) 19:378
https://doi.org/10.1186/s12864-018-4763-1

BMC Genomics

RESEARCH ARTICLE Open Access

# A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*

Jérémy Berthelier[1,2*], Nathalie Casse[2], Nicolas Daccord[3,5], Véronique Jamilloux[4], Bruno Saint-Jean[1] and Grégory Carrier[1]

## Abstract

**Background:** Transposable elements (TEs) are mobile DNA sequences known as drivers of genome evolution. Their impacts have been widely studied in animals, plants and insects, but little is known about them in microalgae. In a previous study, we compared the genetic polymorphisms between strains of the haptophyte microalga *Tisochrysis lutea* and suggested the involvement of active autonomous TEs in their genome evolution.

**Results:** To identify potentially autonomous TEs, we designed a pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements, download: https://doi.org/10.17882/51795), and conducted an accurate TE annotation on a new genome assembly of *T. lutea*. PiRATE is composed of detection, classification and annotation steps. Its detection step combines multiple, existing analysis packages representing all major approaches for TE detection and its classification step was optimized for microalgal genomes. The efficiency of the detection and classification steps was evaluated with data on the model species *Arabidopsis thaliana*. PiRATE detected 81% of the TE families of *A. thaliana* and correctly classified 75% of them. We applied PiRATE to *T. lutea* genomic data and established that its genome contains 15.89% Class I and 4.95% Class II TEs. In these, 3.79 and 17.05% correspond to potentially autonomous and non-autonomous TEs, respectively. Annotation data was combined with transcriptomic and proteomic data to identify potentially active autonomous TEs. We identified 17 expressed TE families and, among these, a TIR/Mariner and a TIR/hAT family were able to synthesize their transposase. Both these TE families were among the three highest expressed genes in a previous transcriptomic study and are composed of highly similar copies throughout the genome of *T. lutea*. This sum of evidence reveals that both these TE families could be capable of transposing or triggering the transposition of potential related MITE elements.

**Conclusion:** This manuscript provides an example of a de novo transposable element annotation of a non-model organism characterized by a fragmented genome assembly and belonging to a poorly studied phylum at genomic level. Integration of multi-omics data enabled the discovery of potential mobile TEs and opens the way for new discoveries on the role of these repeated elements in genomic evolution of microalgae.

**Keywords:** Transposable elements, Genome assembly, Pipeline, Tool, Annotation, Algae, Haptophyte, *Tisochrysis lutea*

* Correspondence: berthelier.j@laposte.net
[1]IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Ile d'Yeu, 44311 Nantes, France
[2]Mer Molécules Santé, EA 2160 IUML - FR 3473 CNRS, Le Mans University, Le Mans, France
Full list of author information is available at the end of the article

## Background

Transposable Elements (TEs) are defined as DNA sequences able to move and spread within eukaryotic and prokaryotic genomes. These repeated elements constitute a variable fraction of eukaryotic genomes, ranging from 3% in the yeast *Saccharomyces cerevisiae*, 45% in human, to 80% in maize [1–3]. TEs were discovered by Barbara McClintock in the late 1940s, refuting the idea that genomes are stable but are, on the contrary, dynamic entities [4]. TEs are highly diverse and an unified classification system for eukaryotic TEs has been proposed, establishing two TE classes according to their transposition mechanisms, structures and similarities [5]. Class I (Retrotransposons) groups elements moving by a copy-paste mechanism through an RNA that is reversed transcribed. Class I is composed of several TE orders, named LTR, DIRS, PLE, LINE and SINE. Class II (DNA transposons) is composed of TEs using different cut-paste mechanisms to transpose. These elements are grouped into the orders TIR, Crypton, Helitron and Maverick. Although intact retrotransposons and DNA transposons are autonomous elements that can move by themselves, SINE elements are non-autonomous TEs and rely on LINE for their mobility, even though their origin is distinct. Other non-autonomous elements can also be distinguished. LTR elements can degenerate into non-coding structures known as LARD (> 4 kbp) or TRIM (< 4 kbp), and TIR elements can also degenerate into non-coding structures known as MITE. LARD, TRIM and MITE elements have intact termini and can thus move by exploiting the molecular machinery of related autonomous TEs [6]. Genomes also contain highly diverged TE fossils, accumulated over time and having no mobility capacity. Due to their mobility, TEs generate mutations in their host genome through new insertions/deletions and participate in genome evolution by impacting the DNA sequence, genome size [7, 8] and chromosome structure [9]. TE activity is known to be triggered during stressful events and, while the majority of transpositions are neutral or harmful to the organisms, transposition events are recognized to promote beneficial mutations [10]. New TE insertions can impact gene function and gene regulation [11]. They can also create new genes and participate in the rise of new phenotypes. The role of TEs has been widely studied in animals [12], land plants [13] and insects [14, 15], but work on their impact on microalgal genomes is only just beginning [16–19]. Microalgae form a diverse polyphyletic group composed of eukaryotic, unicellular and multicellular, photosynthetic organisms [20]. They live in all aquatic habitats whether these have fresh, brackish or salt water and have colonized different extreme habitats, ranging from hot springs, high altitude streams, ice sheets and desert sand crusts, highlighting their evolutionary ability to adapt to broad range of ecosystems [21–25]. Currently around 150,000 species of algae have been described (http://www.algaebase.org), but the number of non-described species is likely to number hundreds of thousands or millions of species [26]. They are divided among different eukaryotic phyla, in Archaeplastidia (green and red lineage), Rhizaria, Alveolates, Stramenopiles (brown lineage), Cryptophytes, Haptophytes and Excavates [27]. Despite their high number and diversity, few genome-wide TE annotations have been performed for microalgae. For the green lineage, this task was realized for ten Chlorophyte species [28–37]. For the red lineage, TE annotation was only done for the Rhodophyte *Cyanidioschyzon* sp. [38]. TEs were annotated in three diatom genomes (brown lineage) [18, 39–41] and also in five dinoflagellate species [42–46]. In Haptophytes, TE annotation has only been performed for one species [47]. These studies reveal that the TE content of microalgae genomes is diverse and includes both retrotransposons and DNA transposons.

Concerning TE activity in microalgae, a few studies have reported evidence of expression or transposition events. Expression of two LTR/Copia families was identified under nitrate starvation or exposure to diatom-derived reactive aldehydes in the diatom species *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* [18]. Moreover, expression of LTR/Copia or TIR/Mariner elements was also reported under thermal stress in *P. tricornutum*, *Amphora acutiuscula*, *Amphora coffeaeformis* and *Symbiodinium microadriaticum* [16, 48–50]. Evidence of transposition events was only identified for a MITE element in a clone of *Chlamydomonas reinhardtii* in the presence of vitamin $B_{12}$, resulting in a new phenotype [17].

Concerning TE activity in Haptophytes, we previously compared genetic polymorphisms between genomes of several strains of *Tisochrysis lutea* [51]. We identified new insertions/deletions and suggested the implication of autonomous TEs in the genome evolution of this species. In this context, the goal of the present study was to inventory TEs in the *T. lutea* genome and to identify potentially autonomous TEs. This marine microalga is commonly used as a feed in aquaculture [52] and is particularly studied for biotechnological applications such as food and biofuel production [53, 54]. In addition, several domesticated strains of *T. lutea* have been obtained with different processes [55] and a large amount of omics data has been collected [51, 56–60].

In this study, we present a detailed TE annotation of the *T. lutea* genome. To achieve this, we designed a new pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements). The efficiency of the detection and classification steps of PiRATE was evaluated with data of the model species *Arabidopsis thaliana*.

Moreover, to be as exhaustive as possible about the repeated content of *T. lutea*, a new genome assembly was performed by combining Pacific Bioscience and Illumina data. Finally, available transcriptomic and proteomic data were used to reveal potential active TE families.

## Results

### PiRATE: Pipeline to Retrieve and Annotate Transposable Elements of non-model organisms

The goal of the present study was to inventory the TE content of the *T. lutea* genome and study the activity of potentially autonomous TEs. Annotation of TEs is a challenging task because of their diversity, their repetitive nature and the complexity of their structures (i.e. GC-rich regions, homopolymers and repeated motifs). Numerous tools have been designed to identify TEs (Additional file 1: Table S1), which can be grouped into four approaches according to their TE detection method: (1) similarity-based detection such as RepeatMasker [61], (2) structure-based detection such as MITE-Hunter [62], (3) repetitiveness-based detection such as RepeatScout [63], and (4) tools building repeated elements from unassembled data such as dnaPipeTE [64].

Currently, the tool used most frequently to perform a TE annotation is RepeatMasker, which provides a rough estimation of the TE content in a genome assembly [61, 65]. However, this tool compares the genomic sequences with a databank of known TEs to realize the annotation and is therefore not suitable for realizing a de novo TE annotation [65–67]. To perform a de novo TE annotation, pipelines employing repetitiveness-based methods of detection, such as RepeatModeler and REPET, are commonly recommended [66–69]. Here we built PiRATE (Fig. 1) to conduct a de novo TE annotation in the genome of non-model species *T. lutea*. PiRATE is composed of detection, classification and annotation steps.

#### *Detection of TEs*

To date, genome assembly of non-model organisms has usually not been performed at the level of complete chromosomes but is instead highly fragmented. This fragmentation is recognized to be partly the result of a bad assembly of the TE copies due to their high

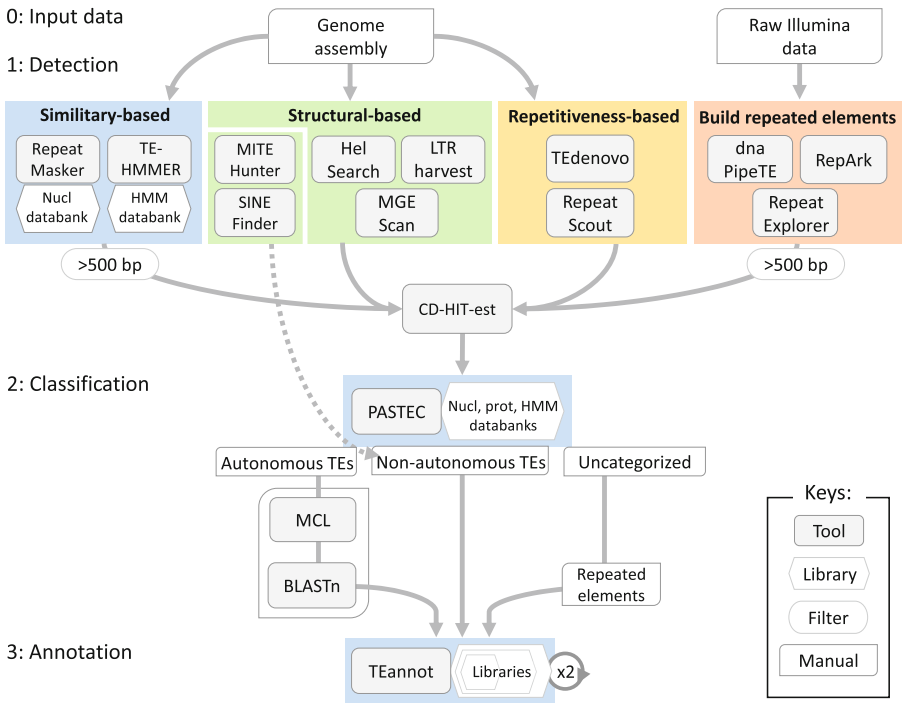


**Fig. 1** Overview of the PiRATE pipeline. Step 0: genome assembly and raw Illumina data are used as input data. Step 1: The detection of putative TEs and repeated sequences is performed using 12 tools, combining four detection approaches. Detected sequences from approaches 1 and 4 are filtered according to their length (minimum 500 bp). Detected sequences from the tools MITE-Hunter and SINE-Finder are directly saved as non-autonomous TEs. Other detected sequences are clustered with CD-HIT-est to reduce redundancy. Step 2: Putative TE sequences are automatically classified with PASTEC as potentially autonomous TEs, non-autonomous TEs or uncategorized sequences. The potentially autonomous TEs are manually checked and grouped into TE families. Step 3: Three libraries are manually constructed with a "Russian doll" strategy: 1) a "potentially autonomous TEs library", a "total TEs library" and a "repeated elements library". A double-run of TEannot is carried out for each library to select sequences that align with a full-length (FLC) on the genome assembly and finally obtain three independent annotations

repetitive content, which increases the difficulty of their detection [70]. The optimization of the detection step of PiRATE was therefore a priority. We made an overview of tools related to TE detection (Additional file 1: Table S1) and 12 tools were selected according to the specificity and efficiency of their algorithms. These tools represent the four major TE detection approaches (presented above), so as to be as exhaustive as possible. Combining tools is recognised to improve TE detection efficiency [66, 67, 71]. We then applied a clustering method to decrease the redundancy of the detected sequences, by selecting the larger detected sequences of each cluster. The goal of this step was to promote the detection of full-length TE sequences. The detection of complete TE sequences bearing recognizable conserved domains or specific structures and motifs makes the classification step easier. Moreover, a complete TE sequence indicates a potentially autonomous element.

#### *Classification of TEs*

The classification step of PiRATE is performed by PASTEC [72], which partly uses databanks of known TEs to establish an automated classification of the detected sequences. To improve the classification step of PiRATE, its default databanks were upgraded, by adding 1240 TE sequences from other public databanks, non-inventoried algal TEs and by building 78 new profile HMMs (Hidden Markov Model). Adding non-inventoried data is important for improving the TE classification of species belonging to poorly studied phyla, which often have few described TEs in the databanks. This is common for numerous microalgal phyla (i.e. Haptophyta, Euglenophyta and Dynophyta). In our case, only 17 TE families belonging to the Haptophyte phylum are present in the most frequently used and complete TE databank Repbase [73, 74]. We also estimated that only 2609 TE families are described for microalgal taxa in Repbase. Compared with other taxa, this number is very low, for examples 29,503 TE families are described for Metazoa and 12,620 for Viridiplantae (Repbase, 10/29/2017). The putative TE sequences are classified following the Wicker et al. classification [5] and can be grouped as 1) potentially autonomous TEs, 2) non-autonomous TEs or 3) uncategorized sequences. Because we were interested in potentially autonomous TEs, these sequences were manually checked and grouped into families.

#### *Annotation of TEs*

For the annotation step, we built three libraries in order to then apply a method that we named "Russian doll", due to its nesting strategy (Additional file 1: Figure S1). We built a "potentially autonomous TEs library" containing checked potentially autonomous TEs, a "total TEs library" also containing the non-autonomous TEs, and a "repeated elements library" also containing the uncategorized repeated sequences. These nested libraries made it possible to perform several independent annotations in order to avoid a competition effect among sequences aligning on the same genomic regions.

### Evaluation of PiRATE with *A. thaliana* genomic data

#### *Evaluation of the detection step*

The detection and classification steps of PiRATE were evaluated to highlight their strengths and weaknesses. This evaluation made it possible to define suitable rules for the manual check step. As a control, we used 359 consensus sequences of the described TE families of *A. thaliana,* available in Repbase. Genomic data of the model plant *A. thaliana* provided a suitable control because of its high quality genome assembly and high TE diversity. Class I and Class II *A. thaliana* TE families are well described for both autonomous and non-autonomous TEs. Detected sequences covering less than 40% of the full-length of a consensus sequence were considered too short to be efficiently classified and were not taken into account. The proportion of TE families detected with a complete length (coverage score ≥ 70%) or detected with at least a partial length (coverage score ≥ 40%) is given in Fig. 2a. PiRATE detected ~ 81% (292/359) of the TE families described in *A. thaliana* genome (Fig. 2a). PiRATE was especially effective for detecting sequences belonging to LTR (96%), LINE (79%), non-autonomous TIR (81%) and non-autonomous Helitron (94%) (Fig. 2a). It had a good efficiency for detecting TIR (62%) and Helitron (60%). However, it had difficulty detecting SINE elements (27%) (Fig. 2a). In addition, we compared the detection step of PiRATE to TEdenovo [68], LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77] (Fig. 2b). Overall, the detection step of PiRATE detected the highest percentage of TE families of *A. thaliana*. Compared to TEdenovo, which displayed the second highest percentage of detected TE families, PiRATE detected 21 additional TE families (+ 6%) (Fig. 2b and Additional file 1: Figure S2). PiRATE was particularly more effective for detecting LINE (+ 32%) and TIR (+ 10%) (Additional file 1: Figure S2).

#### *Evaluation of the classification step*

To evaluate the classification step of PiRATE, we used the 292 sequences detected by PiRATE during the evaluation of the detection step, which represent the largest detected sequences of the 292 TE families of *A. thaliana*. These 292 sequences were classified with PASTEC using the PiRATE databanks (excluding data from *Arabidopsis* species). To estimate the classification efficiency, we counted the number of detected TEs with correct classification at the order level and the number of sequences

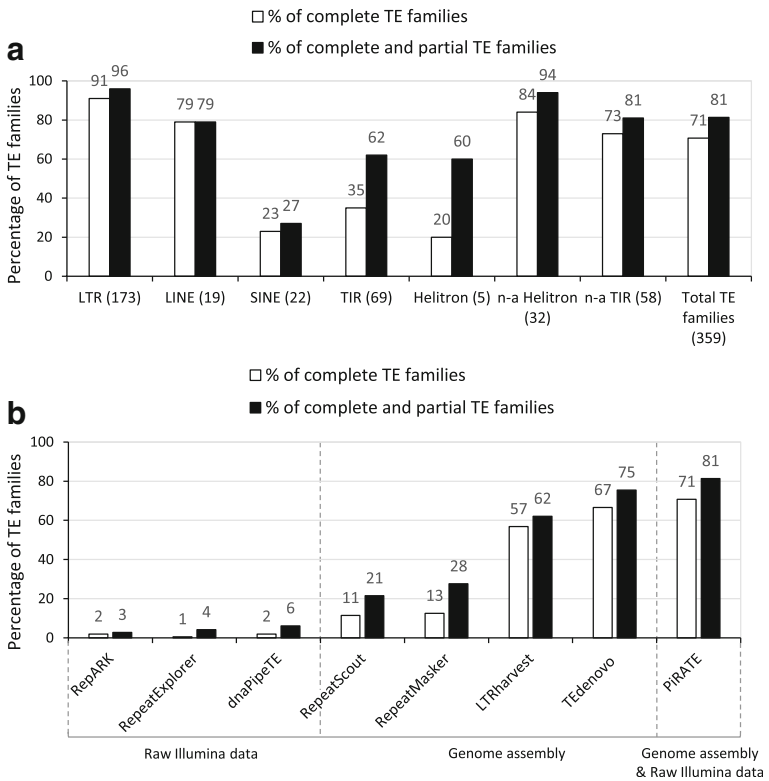


**Fig. 2** Evaluation of the detection step of PiRATE with genomic data of *Arabidopsis thaliana*. **a**) Percentage of TE families detected in *A. thaliana*. For each TE order (x-axis) is indicated the percentage of TE families detected with a complete length (coverage score ≥ 70%, white bars) or detected with a partial and a complete length (coverage score ≥ 40%, black bars). The x-axis indicates the number of TE families for each order; "n-a" means non-autonomous. **b**) Comparison of the percentage of TE families of *A. thaliana* detected by PiRATE (Step 1), RepARK, RepeatExplorer, dnaPipeTE, RepeatScout, RepeatMasker, LTRharvest and TEdenovo. For each tool is indicated the percentages of TE families of *A. thaliana* detected with a complete length (coverage score ≥ 70%, white bars) or detected with a partial and a complete length (coverage score ≥ 40%, black bars). The x-axis indicates the tools and nature of the input data

that had an incorrect classification or that were uncategorized. We observed that 75% (218/292) of the detected TEs were correctly classified, 7% (21/292) were incorrectly classified and 18% (53/292) were uncategorized. The classification step of PiRATE was therefore efficient at correctly classifying autonomous TEs belonging to LTR (98%), LINE (87%), TIR (91%) and Helitron (100%) but had difficulty correctly classifying SINE (50%), non-autonomous TIR (27%) and non-autonomous Helitron (7%) (Additional file 1: Figure S3). Taking into account all of the above results, PiRATE is efficient enough to detect and correctly classify the majority of the autonomous TE families.

## A new genome assembly of *T. lutea* to improve the TE annotation

We recently published a draft genome assembly of *T. lutea* obtained with Illumina short-read technology [51]. To obtain an improved genome assembly, the genome of *T. lutea* was re-sequenced with Pacific Bioscience long-read technology. A new genome assembly was performed from the long reads and was improved with the Illumina short-read data, used to build the draft genome assembly [51]. The new genome assembly of *T. lutea* is composed of 193 contigs and is 82 Mb in size. A gain of around 30 Mb was obtained (+ 34%), compared with the previous 54 Mb genome assembly, which was composed of 7659 contigs [51]. The size of the coding regions increased slightly between these genome versions. While the new genome assembly encodes for 15,972 genes, corresponding to a coding region length of 32 Mb, the gene proportion of the previous draft genome version was 25 Mb, suggesting that the new assembled regions are mostly repeated elements. This new larger version of the genome seems to incorporate more assembled TEs.

## Effect of genome quality on TE detection approaches

To estimate the contribution of each TE detection approach of PiRATE depending on the level of fragmentation of the genome assembly, the detection step (Fig. 1) of PiRATE was applied with raw Illumina data of *T.*

*lutea* and, either the draft genome version of *T. lutea* (7659 contigs) [51] or the new genome assembly of *T. lutea* (193 contigs). In both cases, the detected sequences were compared to the referent sequences of the TE families of *T. lutea* (described below). For each TE detection approach in PiRATE, we counted the number of *T. lutea* TE families detected, with the largest length (i.e. the most complete sequences, having the highest percentage of coverage compared to the reference TE sequences) and divide this number by the total of detected TE families. This provided an estimation ratio of the contribution of each TE detection approach depending on the input data (Fig. 3). With both types of dataset, the similarity-based approach had the weakest percentage and contributed to detecting only 2 or 3% of the *T. lutea* TE families. Using the draft genome assembly and the raw Illumina data, the structural-based approach contributed to detecting 1% of the TEs families of *T. lutea,* but 20% of the TE families of *T. lutea* with the new genome assembly and the raw Illumina data (Fig. 3). The repetitiveness-based approach contributed to detecting 7% of the TE families of *T. lutea* with the draft genome assembly and the raw Illumina data. However, it was the most efficient approach with the new genome and contributed to detecting 63% of the *T. lutea* TE families (Fig. 3). When a draft genome assembly is used as input, the fourth detection approach, using raw Illumina data to build repeated elements, was the most useful approach and contributed to detecting 67% of the TE families (Fig. 3).

### Annotation of the repeated elements content of the *T. lutea* genome

We applied PiRATE to the new genome assembly of *T. lutea* and raw Illumina data. After the classification step, we manually curated the sequences as potentially autonomous TEs, non-autonomous TEs or uncategorized repeated elements. Because we were interested in characterizing their activity, the potentially autonomous TEs were manually checked and grouped into families (see Methods). We identified six potentially autonomous families of LTR/Copia and four families of LTR/Gypsy (Table 1). We found 14 potentially autonomous families of LINE elements, similarly close to Tx1 elements, belonging to the L1 superfamily [78, 79]. We identified seven potentially autonomous families of TIR/Harbinger, six families of TIR/PiggyBac and eight families of TIR/Mariner. A high number of potentially autonomous hAT elements were detected. Due to their divergence, they were grouped into 129 putative families.

Three annotations were conducted with three nested libraries (Additional file 1: Figure S1). From the "potentially autonomous TEs library" composed of 240 referent sequences, we estimated that the proportion of the potentially autonomous TEs represent 3.79% of the *T. lutea* genome (Table 1). The annotation of the TE content was performed with the "total TEs library" containing 459 supplementary sequences corresponding to 14 sequences of potential SINE elements, 188 sequences of potential MITE, 240 sequences of potential TRIM and 17 sequences of potential LARD (Table 1). From this annotation, we estimated that the genome of *T. lutea* contains 20.84% of potentially autonomous and non-autonomous TEs (Table 1 and Additional file 1: Figure S4). Class I and Class II TEs represent 15.89 and 4.95%, respectively (Table 1). We found a large quantity of Gypsy (4.65%), LINE (3.87%) and hAT (2.12%) copies, suggesting ancient burst events for these elements (Table 1). We established that the proportion of non-autonomous TEs is 17.05% (Table 1). Then, we performed the annotation

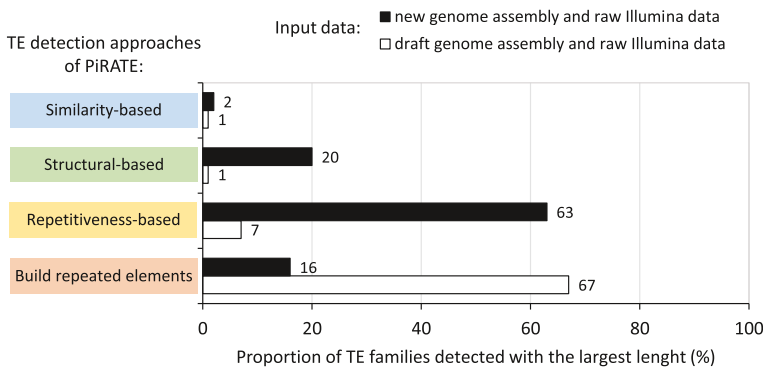


**Fig. 3** Comparison of the contribution of the four TE detection approaches of PiRATE on the detection of the TE families of *Tisochrysis lutea*, depending on the input data. For each TE detection approach, we calculated the number of TE families detected with the largest length and divide this number by the total of detected TE families of *T. lutea*. The input dataset was either the draft genome assembly of *T. lutea* and raw Illumina data of *T. lutea* (white bars) or the new genome assembly of *T. lutea* and raw Illumina data of *T. lutea* (black bars). The similarity-based detection, structural-based detection and the repetitiveness-based detection use a genome assembly as input data. The last approach builds repeated elements from raw Illumina data

**Table 1** Diversity and proportion of transposable element orders and classes in the genome assembly of *Tisochrysis lutea*. The abbreviations "a" and "n-a" indicate autonomous and non-autonomous transposable elements respectively

| | | Orders/ Superfamilies | Number of families (f) or detected sequences (s) | Number of potentially autonomous TEs | Proportion of the potentially autonomous TEs (%) | Proportion of total genome (%) |
|---|---|---|---|---|---|---|
| Class I | a | LTR/Copia | 6 f | 45 | 0.37 | 1.09 |
| | | LTR/Gypsy | 4 f | 242 | 2.56 | 4.65 |
| | | LINE/L1 | 14 f | 59 | 0.25 | 3.87 |
| | n-a | SINE | 14 s | | | 0.04 |
| | | LTR/LARD | 17 s | | | 0.76 |
| | | LTR/TRIM | 240 s | | | 5.48 |
| Total Class I | | | | | | 15.89 |
| Class II | a | TIR/hAT | 129 f | 145 | 0.41 | 2.12 |
| | | TIR/Mariner | 8 f | 41 | 0.11 | 0.19 |
| | | TIR/Harbinger | 7 f | 26 | 0.05 | 0.34 |
| | | TIR/PiggyBac | 7 f | 14 | 0.04 | 0.26 |
| | n-a | MITE | 188 s | | | 2.04 |
| Total Class II | | | | | | 4.95 |
| Total TEs | | | | 572 | 3.79 | 20.84 |

of every repeated element by using the "repeated elements library" containing an additional 2680 uncategorized repeated sequences. From this annotation, we estimated that 17.79% of the *T. lutea* genome is represented by uncategorized repeated elements (Additional file 1: Figure S4). To estimate the proportion of the simple tandem repeats, we used the tool RepeatMasker and found that they made up 5.97% of the genome assembly of *T. lutea* (Additional file 1: Figure S4). By adding together the proportions of all the annotated repeats, we estimated that the total proportion of repeated elements in the *T. lutea* genome was 44.6%. Knowing that the coding gene proportion is of 38.49%, we estimated that 16.91% of the genome is non-characterized (Additional file 1: Figure S4).

### Discovery of potentially active autonomous TEs in the *T. lutea* genome

In this study we chose to focus on the identification of potentially autonomous TEs to reveal potentially active elements. From the annotation obtained with the "potentially autonomous TEs library", we performed the cartography of the 572 annotated TEs that are potentially autonomous (Fig. 4).

To identify potentially active TEs and have an estimation of the genome dynamic of *T. lutea*, transcriptomic data were mapped on the new genome assembly and crossed with the annotation of the 572 potentially autonomous TEs. Expression was identified for 17 TE families: one LTR/Copia, four TIR/Mariner, four TIR/Harbinger and eight TIR/hAT. These families represent 73 potentially autonomous TEs and their genomic position is illustrated in Fig. 4 and is indicated in Additional file 2. Putative ancient transpositions were studied by looking for similarities between copies belonging to these 17 expressed TE families (Additional file 2). We identified that the Mariner-3 family is composed of 24 highly similar copies, which share a mean pairwise identity of 99.7%. Among them, 20 copies seem to be complete (Additional file 3). This high number of similar copies suggests that this family was/is active. The hAT-2 family is composed of three highly similar copies that share a mean pairwise identity of 99.8%. Moreover, eight similar copies were identified for the Harbinger-6 family and five similar copies for the Copia-3 family. Other details can be found in Additional file 2. TE copies belonging to these 17 expressed TE families were submitted to BLASTx on proteomic data of *T. lutea*, that we previously obtained under nitrogen limitation [58]. We identified that transposase proteins were synthesized for the Mariner-3 family and the hAT-2 family. The transposases of the Mariner-3 and hAT-2 families match with six and 36 peptides, respectively. The alignments with the matching peptides can be found for both families in Additional file 4. From transcriptomic data of a previous study, we highlight that these families were among the three higher expressed genes [58].

## Discussion

### PiRATE: Pipeline to Retrieve and Annotate Transposable Elements of non-model organisms

The goal of the present study was to inventory the TE content in the genome of *T. lutea* genome and study the activity of potentially autonomous TEs. We built PiRATE to counter the lack of knowledge about TEs in Haptophytes and the difficulty of identifying TEs in a

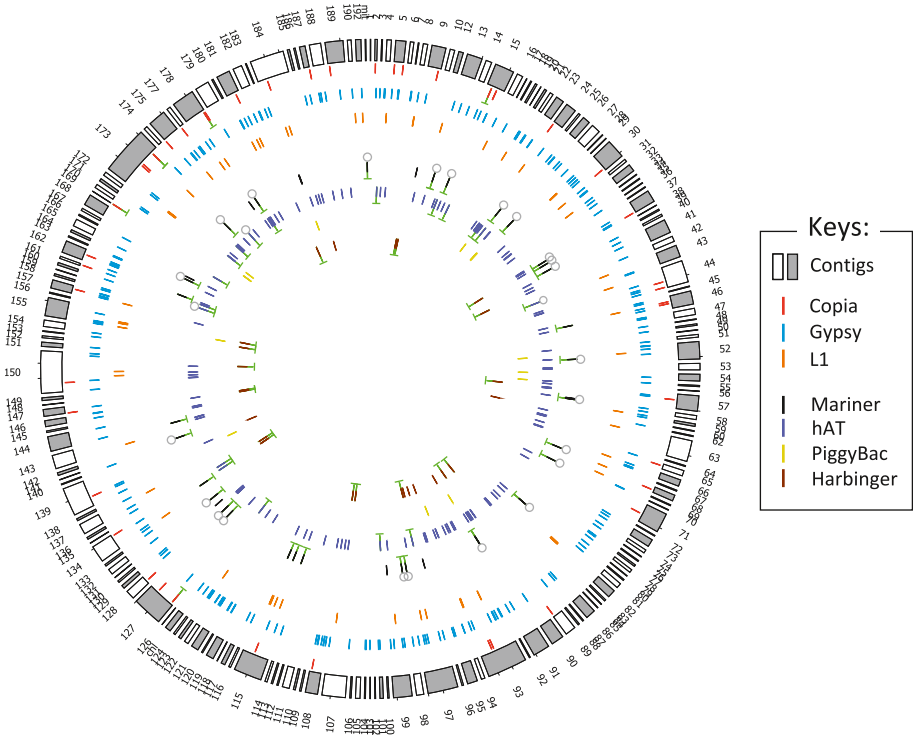


**Fig. 4** Cartography of the 572 potentially autonomous TEs in the genome assembly of *Tisochrysis lutea*. The contig position is random. TEs belonging to the same superfamily are represented with the same colour. The 73 potentially autonomous TEs belonging to the 17 expressed TE families are highlighted with a green "T". Elements belonging to the TIR/Mariner *Luffy* family and TIR/hAT *Ace* family are marked by a grey circle. Transposase proteins were synthesized for these both TE families

fragmented genome assembly [70]. The detection step of PiRATE has been optimized to promote the detection of full-length TE sequences and its classification step has been improved for algal genomes. The detection step of PiRATE combines multiple, existing analysis packages representing all major approaches for TE detection. The detection step of PiRATE was evaluated with genomic data of *A. thaliana* and compared to TEdenovo [68], LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77] (Fig. 2b). Overall, the detection step of PiRATE detected the highest percentage of TE families (81%) with a partial and complete length compare to the other tools used alone (Fig. 2b). This confirms that the combining of multiple tools, using several approaches improves the detection of different TE families, with complete sequences, as previously indicated [66, 67, 71]. In this comparison, TEdenovo was efficient and displayed the second highest percentage of detected TE families (75%) (Fig. 2b). LTRharvest also showed a good capacity to detect TE families of *A. thaliana* (62%) (Fig. 2b). This is due to the high content of LTR elements in the *A. thaliana* genome and because this tool detected TE families belonging to other TE orders. In this comparison, the least effective tools were RepARK (3%), RepeatExplorer (4%) and dnaPipeTE (6%), which used raw illumina data as input (Fig. 2b). This is not surprising considering the challenge of building repeated elements from raw Illumina data, compared to the other tools using the complete genome assembly of *A. thaliana*.

### A new genome assembly of *T. lutea* to improve the TE annotation

We recently published a draft genome assembly of *T. lutea* obtained with Illumina short-read technology [51]. While this technology has a very low sequencing error rate, its use alone often leads to fragmented assemblies, especially in TE-rich genomes, due to the incapacity of short-reads to entirely span repetitive elements [80]. To obtain an improved genome assembly, the genome of *T. lutea* was re-sequenced with Pacific Bioscience long-read technology and the assembly was corrected with short-read Illumina data. Indeed, the use of long-reads leads to a more complete and accurate assembly of long repeated elements such as TEs [81–83]. However, to date, this technology has a high sequencing error rate and its combination with short-read Illumina data has become a common way of partially overcoming this problem [84–86]. Compare to the previous draft genome

assembly, this new genome assembly is larger, less fragmented and seems to incorporate more assembled TEs.

### Effect of genome quality on TE detection approaches

To estimate the contribution of the four TE detection approaches of PiRATE depending on the level of fragmentation of the genome assembly, the detection step (Fig. 1) of PiRATE was applied with raw Illumina data of *T. lutea* and, either the draft genome version of *T. lutea* (7659 contigs) [51] or the new genome assembly of *T. lutea* (193 contigs). The four TE detection approaches showed different contribution according to the level of fragmentation of the genome assembly (Fig. 3). By gathering these four detection approaches, PiRATE improves the TE detection of organisms having a genome assembly which is highly fragmented.

### Annotation of the repeated elements content of the *T. lutea* genome

With PiRATE, we established that the total proportion of repeated elements in the *T. lutea* genome is represented by 20.84% of TEs, 17.79% of uncategorized repeated elements and 5.97% of simple tandem repeats (Additional file 1: Figure S4). The high percentage of uncategorized repeated elements could indicate the presence of unknown TEs. A high number of uncategorized sequences (30.9%) was also reported in the *Emiliania huxleyi* genome [40]. Here, we choose to focus on the identification of potentially autonomous TEs to reveal potentially active elements. The proportion of the potentially autonomous TEs represents 3.79% of the *T. lutea* genome, corresponding to 572 annotated TEs (Fig. 4). Interestingly, we found a potentially autonomous TIR/Mariner in the predicted mitochondrial genome and a potentially autonomous LTR/Copia and TIR/hAT in the predicted chloroplast genome.

### Identification of potentially active TEs in *T. lutea*

Few studies have investigated TE activity in microalgal genomes and their role is poorly known. Regarding Class I TEs, some studies reported expression of LTR elements in dinoflagellate and diatom species under thermal stress or nitrogen limitation [16, 48–50]. Concerning Class II elements, a previous study reported a case of phenotypic evolution for the microalga *Chlamydomonas reinhardtii* caused by the transposition of a MITE in the presence of vitamin $B_{12}$ [17]. In the present study, we identified 17 expressed TE families and, among these, a TIR/Mariner *Luffy* and a TIR/hAT *Ace* family were able to synthesize their transposase under nitrogen starvation [58]. We highlight the presence of highly similar copies (Additional file 3) suggesting that these elements are able to transpose or could be able to trigger the transposition of potential derived MITE elements. Although we cannot draw conclusions about their mobility, the investigation of the TE expression is a good indicator of the potential activity of TEs. Nitrogen limitation has been previously described as a stress condition in the diatom *Phaeodactylum tricornutum*, triggering overexpression of the LTR/Copia family named *Blackbeard* [18]. Although we cannot draw conclusions about de novo insertions, the evidence presented here indicates that these both TEs families are suitable candidates for mobility and could participate in the genome evolution of *T. lutea*.

## Conclusion

Genome-wide TE annotation has rarely been performed in microalgae compared with animals, insects and land plants. This study opens the way to new searches about the role of TEs in the genome evolution of *Tisochrysis lutea* and their contribution to the microalgal adaptation process. In the present study, we built PiRATE to counter the lack of knowledge about TEs in Haptophytes and the difficulty of identifying TEs in a fragmented genome assembly. With PiRATE, we conducted a genome-wide detection and annotation of the repeated elements in a new genome assembly of *Tisochrysis lutea* and established that it is composed of 3.8 and 15.95% of potentially autonomous and non-autonomous TEs, respectively. The annotation of the potentially autonomous TEs was crossed with transcriptomic and proteomic data and evidence of expression was identified for 17 TE families. Among these, we discovered that transposase proteins were synthesized for both a Mariner (*Luffy*) and a hAT (*Ace*) family. Both these families have several highly similar copies throughout the genome and were among the three highest expressed genes in a previous transcriptomic study. All of this suggests that both these families could be able to transpose themselves or trigger the transposition of potential derived MITE elements.

## Methods

### Microalga strain and culture conditions

The *T. lutea* strain was provided by the Culture Collection of Algae and Protozoa (CCAP 927/14). This strain was isolated by Haines in the late 70s and stored in the algae bank. The strain was grown in two 1-L flasks, bubbled with 0.22 mm filtered-air. The culture was maintained at a constant temperature of 21 °C, under a constant irradiance of 50 μmol $m^{-2}$ $s^{-1}$.

### DNA extraction, sequencing, genome assembly and gene annotation

Total DNA was extracted from the *T. lutea* WT-strain using a phenol/chloroform protocol. DNA quality and concentration were assessed with gel electrophoresis and

Qubit Fluorometric Quantitation (ThermoFisher, Massachusetts, USA), respectively. *T. lutea* genome sequencing was performed with a PacBio RSII sequencer (Pacific Bioscience, California, USA) at the Plateforme GeT PlaGe (Toulouse, France); seven SMRT cells were performed. Filtered subreads were assembled using Canu1.3 [82]. The assembly was polished with Quiver (https://github.com/PacificBiosciences/GenomicConsensus) and its accuracy was improved with Pilon [87] using previous Illumina Hiseq mate-pair reads of *T. lutea* ([51]; SRA: SRR3156597). The annotation of the coding-gene region was performed with the pipeline MAKER2 [88–91].

## TE annotation in the *T. lutea* genome using PiRATE

### *Step 1: TE detection*

The new genome assembly of *T. lutea* and previous raw Illumina data ([51]; SRA: SRR3156597) were used as input. Putative TE sequences were detected using four approaches (Fig. 1). The first approach was represented by two tools using similarity-based detection: RepeatMasker (setting: -s, –no_low, –lib; with the PiRATE nucleotide databank; [61]) and TE-HMMER (with a homemade profile HMMs databank). TE-HMMER is a homemade tool using HMMER (default setting, [92]) and tBLASTn (setting: -evalue 10E-300, [93]). The second approach consisted of five tools using structural-based detection: LTRharvest (default setting, [75]), Helsearch (default setting, [94]), MGEScan-nonLTR (default setting, [95]), MITE-Hunter (default setting, [62]) and SINE-Finder (default setting, [96]). The third approach combines tools using repetitiveness-based detection: TEdenovo (steps 1 to 4, default setting, [68] and Repeat Scout (default setting, [63]). These tools cluster repeated sequences from a genome assembly to build consensus sequences. The last approach was composed of tools performing the assembly of repeated sequences from raw Illumina data (fasta or fastq). We used RepARK (default setting, [77]), dnaPipeTE (setting: %coverage: 0.6, [64]) and RepeatExplorer (setting: -paired, [76]). The sequences detected by the first and the last approaches that were below 500 bp in length were removed with a perl script. The sequences detected with SINE-Finder and MITE-Hunter were directly saved for the second step. Other detected sequences were concatenated into a single FASTA file and clustered with CD-HIT-est (settings: -aS 1 -c 1 -r 1 -g 1 -p 0, [97]) to reduce the redundancy. This made it possible to remove shorter sequences that aligned with 100% of identity on a part of the larger sequences.

### *Step 2: TE classification*

In the second step, sequences were automatically classified with PASTEC [72], following the Wicker et al. classification system [5] This tool was improved with custom databanks (described below). Three libraries were manually constructed with a "Russian doll" strategy in order to perform separate annotations (Additional file 1: Figure S1): a "potentially autonomous TEs library", a "total TEs library" containing the potentially autonomous TEs and the non-autonomous TEs and a "repeated elements library" also containing the uncategorized repeated sequences. Sequences classified as LTR, LINE and TIR were manually sorted by superfamily (according to the evidence section produced by PASTEC). To facilitate their manual check, sequences belonging to the same putative superfamily were grouped into families with MCL (MCL_inflation: 1.5; MCL_coverage: 0). The percentage of identity between sequences belonging to the same family were checked with Blastn (–identity: 80%). We followed the 80–80-80 Wicker rules to form families [5]. Finally, larger sequences from each TE family were checked and selected for the "potentially autonomous TEs library" according to the presence of TE domains or similarities with Pfam (http://pfam.xfam.org), NCBI-BLASTx and Censor (http://www.girinst.org/censor). We defined as potentially autonomous LTR, sequences bearing at least a reverse transcriptase and an integrase domain and having similarity to known LTR elements. We defined as potentially autonomous LINE, sequences bearing at least a reverse transcriptase domain and sharing similarity to known LINE elements. We defined as potentially autonomous TIR, sequences with evidence of a transposase domain or similarity with known TIR elements.

No manual checks were performed for sequences classified as non-autonomous TEs. Sequences classified as SINE, MITE and TRIM were directly selected for the "total TEs library". Only sequences classified as LARD, which were obtained with the repetitiveness-based approach of TE detections (TEdenovo or Repeat Scout), were selected. Sequences detected by SINE-Finder and MITE-Hunter were also directly selected for the "total TEs library". Finally, the sequences classified as noCat (uncategorized) and obtained with the repetitiveness-based approach at the TE detection step were selected for the "repeated elements library".

### *Step 3: TE annotation*

Three libraries were built (Additional file 1: Figure S1): 1) a "potentially autonomous TEs library" 2) a "total TEs library" and 3) a "repeated elements library". A first run of TEannot ([68], default setting, steps 1, 2, 3, 7 and 8) was performed for each library to known sequences matching with a full-length size on the genome (FLC sequences) and to remove potential chimeric data. A second run of TEannot was performed with these FLC sequences for each of the final libraries (default setting,

steps 1, 2, 3, 4, 5, 7 and 8) and three annotations were obtained.

### Proportion of TEs and repeated elements in *T. lutea*

From the annotation file obtained with the "potentially autonomous TEs library", we manually selected 572 sequences and calculated their proportion in the genome of *T. lutea*. TEs. The different criteria used are detailed in Additional file 1: Method S1 and Table S2. An illustration of the position of these sequences on the *T. lutea* genome assembly was built with the tool Circos [98]. The annotations obtained with the "total TEs library" and the "repeated elements library" were used to estimate the total proportion of TEs and to calculate the proportion of uncategorized repeated elements in the genome of *T. lutea*. Details on the method are available in Additional file 1: Method S2 and Table S3. The proportion of simple repeats was calculated with the tool RepeatMasker (setting: -s -noint -no_is, [61]).

### PiRATE databanks

#### *Nucleotide and protein databanks*

A nucleotide and a protein databank of TEs were built with sequences from Repbase (REPET version 20.05, http://www.girinst.org/repbase), the P-MITE database (http://pmite.hzau.edu.cn) and SINE base (http://sines.eimb.ru). Because algae originally arose from the predation of a cyanobacterial organism by a eukaryotic heterotrophic organism, cyanobacterial TE sequences were also added from the IS-finder database (http://www-is.biotoul.fr) (Additional file 5). Moreover, we added non-inventoried TEs of microalgae and macroalgae, retrieved from the NCBI database (Additional file 5).

#### *Profile HMMs databank*

A homemade databank of profile HMMs was built with sequences of the protein databank. Multiple protein alignments were performed with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). When possible, TE protein sequences from algae were favoured. 78 profile HMMs were performed with the HMMbuild tool of HMMER [92] for 62 TE categories displayed on the Browse Repbase tool (http://www.girinst.org/repbase/update/browse.php). This databank was used with TE-HMMER at the detection step. At the classification step, we combine this databank with the default databank of PASTEC (ProfilesBankForREPET_Pfam27.0_GypsyDB.hmm, https://urgi.versailles.inra.fr/download/repet).

### Evaluation of PiRATE

The efficiency of the detection and classification steps of PiRATE were evaluated with genomic data of the model plant *A. thaliana*. We used the genome assembly TAIR10 available on the TAIR project (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_chromosome_files) and the raw Illumina data available at the 1001 genome project http://1001genomes.org/data/SLU/SLUHenning2014/releases/current/strains/Seattle-0). These data of *A. thaliana* were submitted to the step 1 of PiRATE (RepeatMasker and TE-HMMER were used without data from *Arabidopsis* species in the databanks). The detected sequences were submitted to PASTEC [72] and compared to the 359 TE families described in *A. thaliana* and available on Repbase (http://www.girinst.org/repbase). We didn't include the terminal repeated sequences of the LTR TE families and the heterologous TE named DRL1. From the classification file, we selected each of the sequences matching to a TE consensus sequences of *A. thaliana*. Those covering less than 40% of the full-length of a consensus sequences were considered as too short to be efficiently classified and were not taken in account. We considered as a partial or complete detection the detected sequences covering at least 40% or 70% of the full-length of a consensus TE family of *A. thaliana*, respectively. For the comparison of the detection step of PiRATE with TEdenovo [68] (steps 1 to 4, with LTRharvest [75]), LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77], the number of detected TE families was calculated with the same method previously described for the evaluation of the PiRATE detection step. For the evaluation of the classification step of PiRATE, we used the longest detected sequences of the 292 TE families detected by PiRATE during the evaluation of the detection step as a control. These 292 sequences were classified with PASTEC using modified versions of the three PiRATE databanks (Nucleotide, protein and profile HMMs), without data from *Arabidopsis* sp. We calculated the percentages of correct classification, incorrect classification or uncategorized classification. Details on the impact of the genome assembly quality on the efficiency of the TE detection step of PiRATE are available in Additional file 1: Method S3.

### Transcriptomic and proteomic analyses

The expression analysis was performed using eight sets of previously published transcriptomics data [56, 58]. These data were concatenated and normalized using the tool insilico_read_normalization.pl of Trinity [99]. Reads were then mapped on the new genome assembly of *T. lutea* with TopHat [100] and crossed with the annotation of the potentially autonomous TEs. HTseqCount [101] was used to count the number of mapped reads for each potentially autonomous TEs. With a homemade script we retrieved the TE families with transcripts covering at least 90% of the annotated sequences. Sequences

of the TE copies of these expressed TE families were then compared with BLASTx to published proteomic data [58]. Sequence alignments of the peptides of the Mariner (*Luffy*) and hAT (*Ace*) elements on the predicted transposases were performed with ClustalOmega and visualized with Geneious (Additional file 4). With the global-alignment tool of Geneious [102] (setting: free end gaps), a mean pairwise identity was calculated for each expressed TE family having at least three annotated copies (Additional file 3).

### PiRATE is automated through a stand alone Galaxy

All tools used in PiRATE are automated in a standalone Galaxy [103]. The PiRATE-Galaxy is available through a virtual machine at https://doi.org/10.17882/51795. A tutorial file can be download.

## Additional files

**Additional file 1:** Additional supporting information. This file contains the additional supporting figures, tables, results and materials and methods. (PDF 594 kb)

**Additional file 2:** Percentage of identity between copies of the expressed TE families. This file lists the percentage of identity between the TE copies of the 17 expressed TE families identified in the genome of *Tisochrysis lutea*. (XLSX 19 kb)

**Additional file 3:** Sequences alignment of TE copies of the TIR/Mariner *Luffy* family. This file contains the sequence alignment of the copies belonging to the TIR/Mariner *Luffy* described in the genome of *Tisochrysis lutea*. (PDF 17427 kb)

**Additional file 4:** Sequences alignment of the peptides matching on the predicted TE proteins. This file contains the alignment of the peptides matching on the predicted proteins of the TIR/Mariner *Luffy* and the TIR/hAT *Ace*. (PDF 996 kb)

**Additional file 5:** List of non-inventoried sequences added to the databanks used by the pipeline PiRATE. This file lists the non-inventoried sequences added to the databanks used by the pipeline PiRATE, they belong to algae and cyanobacteria. (XLSX 23 kb)

### Abbreviations

HMM: Hidden Markov Model; LARD: Large Retrotransposon Derivative; LINE: Long Interspersed Element; LTR: Long Terminal Repeat; MITE: Miniature Inverted-repeat Transposable Element; PiRATE: Pipeline to Retrieve and Annotate Transposable Element; PLE: Penelope; SINE: Short Interspersed Nuclear Element; TE: Transposable element; TIR: Terminal Inverted Repeat; TRIM: Terminal-repeat Retrotransposons In Miniature

### Acknowledgements

This work was supported by the French Region of Pays de la Loire with the Atlantic Microalgue program and the French Research Institute for Exploitation of the Sea (IFREMER). We thank the platform Genotoul GeT-PlaGe for the genome sequencing of *T. lutea*. We also thank the URGI Team for their advice about REPET as well as Jonathan Filée and Etienne Bucher for their advice on this study. We thank Helen McCombie for the proofreading. The authors are grateful to the anonymous reviewers for their critical comments, which have greatly improved the manuscript.

### Funding

This work was supported by the French region of Pays de la Loire and the French Research Institute for Exploitation of the Sea (IFREMER).

### Availability of data and materials

Datasets relating to the identification of TEs, as well as the improved genome assembly of *T. lutea*, are available at https://doi.org/10.17882/52231. The virtual machine of the PiRATE-Galaxy and the tutorial are available at https://doi.org/10.17882/51795.

### Authors' contributions

JB developed PiRATE, carried out the TE annotation and the expression analyses, participated in the genome assembly and drafted the manuscript. NC participated in the coordination and contributed to draft the manuscript. ND designed the pipeline for the genome assembly, contributed for the analysis of the genome assembly and contributed to draft the manuscript. VJ contributed to parameter the classification and the annotation steps of PiRATE (PASTEC and TEannot), contributed to interpret the classification and annotation results and contributed to draft the manuscript. BSJ participated in the coordination of the study, the expression analyses and contributed to draft the manuscript. GC coordinated the design of PiRATE, the control, the TE annotation and the expression analyses, participated in the genome assembly, realized the gene annotation and helped to draft the manuscript. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note



### Author details

[1]IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Ile d'Yeu, 44311 Nantes, France. [2]Mer Molécules Santé, EA 2160 IUML - FR 3473 CNRS, Le Mans University, Le Mans, France. [3]Institut de Recherche en Horticulture et Semences, INRA of Angers, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université d'Angers, Angers, France. [4]Research Unit in Genomics-Info, INRA of Versailles, Versailles, France. [5]Université Bretagne Loire, Angers, France.

Received: 8 December 2017 Accepted: 7 May 2018
Published online: 22 May 2018

### References

1. Adams M, Kerlavage A, Fleischmann R, Fuldner R, Bult C, Lee N, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature. 1995;377:3–174.
2. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112.
3. Carr M, Bensasson D, Bergman CM. Evolutionary Genomics of Transposable Elements in *Saccharomyces cerevisiae*. Stajich JE, editor. PLoS ONE. 2012;7: e50978.
4. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci. 1950;36:334–55.
5. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
6. Bureau TE, Wessler SR. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell. 1994;6:907–16.
7. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115:49–63.
8. Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012;509:7–15.
9. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen F. Transposons in place of telomeric repeats at a Drosophila telomere. Cell. 1993;75:1083–93.
10. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol Ecol. 2013;22:1503–17.
11. Kazazian HH. Mobile elements: drivers of genome evolution. Science. 2004; 303:1626–32.

12. Nekrutenko A, Li W-H. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 2001;17:619–21.
13. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2012;14:49–61.
14. Darboux I, Charles J-F, Pauchet Y, Warot S, Pauron D. Transposon-mediated resistance to Bacillus sphaericus in a field-evolved population of Culex pipiens (Diptera: Culicidae). Cell Microbiol. 2007;9:2022–9.
15. Maumus F, Fiston-Lavier A-S, Quesneville H. Impact of transposable elements on insect genomes and biology. Current Opinion in Insect Science. 2015;7:30–6.
16. Egue F, Chenais B, Tastard E, Marchand J, Hiard S, Gateau H, et al. Expression of the retrotransposons *Surcouf* and *Blackbeard* in the marine diatom *Phaeodactylum tricornutum* under thermal stress. Phycologia. 2015;54:617–27.
17. Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG. Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution. The ISME journal. 2015;9:1446–55.
18. Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics. 2009;10:624.
19. Philippsen GS, Avaca-Crusca JS, Araujo APU, DeMarco R. Distribution patterns and impact of transposable elements in genes of green algae. Gene. 2016;594:151–9.
20. De Clerck O, Guiry MD, Leliaert F, Samyn Y, Verbruggen H. Algal taxonomy: a road to nowhere? J Phycol. 2013;49:215–25.
21. Sakai N, Sakamoto Y, Kishimoto N, Chihara M, Karube I. Chlorella strains from hot springs tolerant to high temperature and high CO 2. Energy Convers Manag. 1995;36:693–6.
22. Rott E, Cantonati M, Füreder L, Pfister P. Benthic algae in high altitude streams of the alps – a neglected component of the aquatic biota. Hydrobiologia. 2006;562:195–216.
23. Anesio AM, Laybourn-Parry J. Glaciers and ice sheets as a biome. Trends Ecol Evol. 2012;27:219–25.
24. Treves H, Raanan H, Finkel OM, Berkowicz SM, Keren N, Shotland Y, et al. A newly isolated *Chlorella* sp. from desert sand crusts exhibits a unique resistance to excess light intensity. FEMS Microbiol Ecol. 2013;86:373–80.
25. Berthelier J, Schnitzler CE, Wood-Charlson EM, Poole AZ, Weis VM, Detournay O. Implication of the host TGFβ pathway in the onset of symbiosis between larvae of the coral Fungia scutaria and the dinoflagellate Symbiodinium sp. (clade C1f). Coral Reefs. 2017;36:1263–8.
26. Guiry MD. How many species of algae are there? J Phycol. 2012;48:1057–63.
27. Not F, Siano R, Kooistra WHCF, Simon N, Vaulot D, Probert I. Diversity and Ecology of Eukaryotic Marine Phytoplankton. Advances in Botanical Research [Internet]. Elsevier; 2012 [cited 2015 Oct 29]. p. 1–53. Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780123914996000013
28. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 2007;318:245–50.
29. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science. 2010;329:223–6.
30. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The *Chlorella variabilis* NC64A genome reveals adaptation to Photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell. 2010;22:2943–55.
31. Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, et al. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc Natl Acad Sci. 2006;103:11647–52.
32. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci. 2007;104:7705–10.
33. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine Picoeukaryotes micromonas. Science. 2009;324:268.
34. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, et al. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. Genome Biol. 2012;13:R39.
35. Vieler A, Wu G, Tsai C-H, Bullard B, Cornish AJ, Harvey C, et al. Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga Nannochloropsis oceanica CCMP1779. Bhattacharya D, editor. PLoS Genetics. 2012;8:e1003064.
36. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, et al. Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. Genome Biol. 2012;13:R74.
37. Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, et al. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. Proc Natl Acad Sci. 2017;114:E4296–305.
38. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, et al. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. BMC Biol. 2007;5:28.
39. Armbrust EV. The genome of the diatom Thalassiosira Pseudonana: ecology, evolution, and metabolism. Science. 2004;306:79–86.
40. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008;456:239–44.
41. Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Maréchal E, et al. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. The Plant Cell Online. 2015;27:162–76.
42. McEWAN M, Humayun R, Slamovits CH, Keeling PJ. Nuclear genome sequence survey of the dinoflagellate *Heterocapsa triquetra*. J Eukaryot Microbiol. 2008;55:530–5.
43. Jaeckisch N, Yang I, Wohlrab S, Glöckner G, Kroymann J, Vogel H, et al. Comparative Genomic and Transcriptomic Characterization of the Toxigenic Marine Dinoflagellate *Alexandrium ostenfeldii*. Moustafa A. PLoS ONE. 2011;6:e28012.
44. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. Curr Biol. 2013;23:1399–408.
45. Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, et al. The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. Science. 2015;350:691–4.
46. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, et al. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Scientific Reports [Internet]. 2016 [cited 2018 Feb 16];6. Available from: http://www.nature.com/articles/srep39734
47. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature. 2013;499:209–13.
48. Hermann D. Caractérisation d'éléments transposables de type mariner chez les microalgues marines [Internet]. Université du Maine; 2011 [cited 2015 Nov 23]. Available from: https://tel.archives-ouvertes.fr/tel-00732952/
49. Nguyen DH. Caractérisation et expression de nouveaux éléments génétiques transposables de la superfamille Tcl-Mariner chez la microalgue marine *Amphora acutiuscula* (Bacillariophyta). 2014;
50. Chen JE, Cui G, Wang X, Liew YJ, Aranda M. Recent expansion of heat-activated retrotransposons in the coral symbiont *Symbiodinium microadriaticum*. The ISME Journal. 2017;
51. Carrier G, Baroukh C, Rouxel C, Duboscq-Bidot L, Schreiber N, Bougaran G. Draft genomes and phenotypic characterization of Tisochrysis lutea strains. Toward the production of domesticated strains with high added value. Algal Res. 2018;29:1–11.
52. Liu W, Pearce CM, McKinley RS, Forster IP. Nutritional value of selected species of microalgae for larvae and early post-set juveniles of the Pacific geoduck clam, *Panopea generosa*. Aquaculture. 2016;452:326–41.
53. Marchetti J, Bougaran G, Le Dean L, Mégrier C, Lukomska E, Kaas R, et al. Optimizing conditions for the continuous culture of Isochrysis affinis galbana relevant to commercial hatcheries. Aquaculture. 2012;326–329:106–15.
54. Sánchez Á, Maceiras R, Cancela Á, Pérez A. Culture aspects of Isochrysis galbana for biodiesel production. Appl Energy. 2013;101:192–7.
55. Bougaran G, Rouxel C, Dubois N, Kaas R, Grouas S, Lukomska E, et al. Enhancement of neutral lipid productivity in the microalga Isochrysis affinis Galbana (T-Iso) by a mutation-selection procedure. Biotechnol Bioeng. 2012;109:2737–45.
56. Carrier G, Garnier M, Le Cunff L, Bougaran G, Probert I, De Vargas C, et al. Comparative transcriptome of wild type and selected strains of the microalgae Tisochrysis lutea provides insights into the genetic basis, lipid metabolism and the life cycle. Abad-Grau MM. PLoS One. 2014;9:e86889.
57. Charrier A, Bérard J-B, Bougaran G, Carrier G, Lukomska E, Schreiber N, et al. High-affinity nitrate/nitrite transporter genes ( *Nrt2* ) in *Tisochrysis lutea* : identification and expression analyses reveal some interesting specificities of Haptophyta microalgae. Physiol Plant. 2015;154:572–90.

58. Garnier M, Bougaran G, Pavlovic M, Berard J-B, Carrier G, Charrier A, et al. Use of a lipid rich strain reveals mechanisms of nitrogen limitation and carbon partitioning in the haptophyte Tisochrysis lutea. Algal Res. 2016;20:229–48.
59. Thiriet-Rupert S, Carrier G, Chénais B, Trottier C, Bougaran G, Cadoret J-P, et al. Transcription factors in microalgae: genome-wide prediction and comparative analysis. BMC Genomics. 2016;17:282.
60. Thiriet-Rupert S, Carrier G, Trottier C, Eveillard D, Schoefs B, Bougaran G, et al. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of Tisochrysis lutea. Algal Res. 2018;30:59–72.
61. Smit, A. F., Hubley, R., & Green, P. (1996). RepeatMasker. [Internet]. Available from: http://www.repeatmasker.org.
62. Han Y, Wessler SR. MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38:e199–e199.
63. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21:i351–8.
64. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian Tiger mosquito (Aedes albopictus) Repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biology and Evolution. 2015;7:1192–205.
65. Ragupathy R, You FM, Cloutier S. Arguments for standardizing transposable element annotation in plant genomes. Trends Plant Sci. 2013;18:367–76.
66. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. Mobile DNA [Internet]. 2015 [cited 2017 Jun 28];6. Available from: http://www.mobilednajournal.com/content/6/1/13
67. Arensburger P, Piégu B, Bigot Y. 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. 2016;6:e1256852.
68. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in De novo annotation approaches. Xu Y. PLoS One. 2011;6:e16526.
69. Smit, AF, Hubley, R. RepeatModeler Open-1.0 [Internet]. 2010. Available from: http://www.repeatmasker.org.
70. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. Mob DNA. 2015;6:13.
71. Kamoun C, Payen T, Hua-Van A, Filée J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. BMC Genomics. 2013;14:700.
72. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An Automatic Transposable Element Classification Tool. Cordaux R. PLoS One. 2014;9:e91929.
73. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research. 2005;110:462–7.
74. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
75. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.
76. 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;29:792–3.
77. Koch P, Platzer M, Downie BR. RepARK–de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 2014;42:e80–e80.
78. Garrett JE, Carroll D. Tx1: a transposable element from Xenopus laevis with some unusual properties. Mol Cell Biol. 1986;6:933–41.
79. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008;9:411–2.
80. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2011;13:nrg3117.
81. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, et al. Illumina TruSeq synthetic long-reads empower De novo assembly and resolve complex, highly-repetitive transposable elements. Singh N. PLoS One. 2014;9:e106689.
82. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv. 2017:071282.
83. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res. 2017;27:709–21.
84. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:341.
85. Phillippy AM. New advances in sequence assembly. Genome Res. 2017;27:xi–xiii.
86. Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 2017;27:787–92.
87. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. Wang J. PLoS One. 2014;9:e112963.
88. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17:847–8.
89. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2007;18:188–96.
90. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 2011;12:491.
91. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale. bioRxiv. 2017;193490.
92. Eddy SR. Others. Multiple alignment using hidden Markov models. Ismb. 1995;3:114–20.
93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
94. Yang L, Bennetzen JL. Structure-based discovery and description of plant and animal Helitrons. Proc Natl Acad Sci. 2009;106:12832–7.
95. Rho M, Tang H. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Res. 2009;37:e143–e143.
96. Wenke T, Dobel T, Sorensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. THE PLANT CELL ONLINE. 2011;23:3117–28.
97. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
98. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
99. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
100. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
101. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.
102. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–9.
103. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451–5.