De novo annotation of transposable elements : tackling the fat genome issue

Véronique Jamilloux, Josquin Daron, Frédéric Choulet, Hadi Quesneville

INVITED PAPER

De Novo Annotation of Transposable Elements: Tackling the Fat Genome Issue

By Véronique Jamilloux, Josquin Daron, Frédéric Choulet, and Hadi Quesneville

ABSTRACT | Transposable elements (TEs) constitute the most dynamic and the largest component of large plant genomes: for example, 80% to 90% of the maize genome and the wheat genome may be TEs. *De novo* TE annotation is therefore a computational challenge, and we investigated, using current tools in the REPET package, new strategies to overcome the difficulties. We tested our methodological developments on the sequence of the chromosome 3B of the hexaploid wheat; this chromosome is ~1 Gb, one of the "fattest" genomes ever sequenced. We successfully established various strategies for annotating TEs in such a complex dataset. Our analyses show that all of our strategies can overcome the current limitations for *de novo* TE discovery in large plant genomes. Relative to annotation based on a library of known TEs, our *de novo* approaches improved genome coverage (from 84% to 90%), and the number of full length annotated copies from 14 830 to 15 905. We also developed two new metrics for qualifying TE annotation: NTE50 involves measuring the number, and LTE50 the smallest sizes of annotations that cover 50% of the genome. NTE50 decreased the number of annotations from 124 868 to 93 633 and LTE50 increased it from 1839 to 2659. This work shows how to obtain comprehensive and high-quality automatic TE annotation for a number of economically and agronomically important species.

KEYWORDS | Bioinformatics; genomics

Manuscript received November 25, 2015; revised April 20, 2016; accepted July 6, 2016.
V. Jamilloux and **H. Quesneville** are with the UR1164 URGI Research Unit in Genomics-Info, INRA, INRA de Versailles, 78026 Versailles, France (e-mail: hadi.quesneville@versailles.inra.fr).
J. Daron and **F. Choulet** are with INRA, UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 63039 Clermont-Ferrand, France, and also with the Université Blaise Pascal, UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 63039 Clermont-Ferrand, France.

Digital Object Identifier: 10.1109/JPROC.2016.2590833

I. INTRODUCTION

Transposable elements (TEs) constitute by far the most dynamic and the largest component of large plant genomes. They make up an estimated 85% of the maize genome [1] and 88% of the wheat genome [2]–[4]. They are important in genome evolution, with consequences for gene function and regulation. Elucidating the role of TEs in a genome is now clearly essential for any detailed understanding of the biology of a species. Precise annotation of TEs in genomic sequences is therefore a prerequisite to many analyses.

TE annotation software and automatic TE annotation pipelines have been designed (see [5] and [6] for review). Most software for *de novo* identification of TEs works well with genomes up to 400 Mb. However, larger genomes are computationally challenging, requiring a huge amount of memory and computing time: the "fat genome" issue. This study presents strategies to overcome this problem by exploiting a feature specific to TEs in large genomes: only a small number TE families populate most of the TE space.

We developed new strategies based on this observation to overcome the genome size limitation and built them on the existing tools of the REPET package. This package includes two pipelines: TEdenovo that builds a TE consensus library, and TEannot that annotates TEs in the genome. It has been extensively used in a number of genome projects [7]–[25].

We chose chromosome 3B from bread wheat [4] as an experimental model to test our strategies. We describe the principles and the results from three different approaches building TE libraries with TEdenovo using: 1) a library of known TEs concatenated with a *de novo* library built from a genome spliced from their known TEs; 2) a *de novo* LTR-retrotransposon library obtained with the LTRHarvest software then concatenated with the *de novo* library built from a genome spliced for the already identified LTRs;

and 3) a *de novo* TE consensus library obtained from a genome subset (300 or 150 Mbp of the longest contigs). In each case, the consensuses are classified and used in an iterative process. Our strategies allowed us to overcome the current memory and time limitations for *de novo* TE discovery using REPET on large plant genomes. This study paves the way to comprehensive and high-quality automatic TE annotation in various economically and agronomically important species.

II. MATERIAL AND METHODS

A. Sequences

We used the sequence of wheat chromosome [4], which is the first fully sequenced wheat chromosome. We used the V2.1 assembly available at URGI repository (https://urgi.versailles.inra.fr/download/wheat/3B/previous_versions/allLargeContigsV2.1.fna.gz). This draft assembly corresponds to an automated scaffolding of 294691 contigs representing 986.1 Mbp. It does not correspond to the final high-quality published sequence, but an intermediate state of quality more comparable to what is generally available for TE annotation. Indeed, the published sequence has been manually improved to reach a high-quality assembly not generally available in standard genome projects. Thus, we consider an assembly that is more fragmented than the final output, making TE identification and annotation even more challenging; this is more difficult than necessary, but is a realistic and the most often encountered situation.

We also used the maize genome sequence with ten chromosomes, ZmB73_RefGen_v3 (Maize RefGen database V3 (release 5b.+) from Gramene.org) totalizing 2059701728 bp. We extracted 125147 contigs (2046696042 bp) from the original genome sequence.

B. The Known TE Library

A library of known wheat TEs was manually compiled from the TREP database (http://wheat.pw.usda.gov/ITMI/Repeats/) and previous annotation of BAC sequences [2], [3]. The previously identified TE copy sequences cluster in 586 groups using MCL after manual curation. We performed multiple alignments of each cluster, resulting in 586 TE consensuses.

C. TEdenovo

TEdenovo from the REPET package (release 2.1, see documentation and availability at https://urgi.versailles.inra.fr/Projects/URGI-softwares/REPET) was used to identify TE families in genomic sequences. It runs five functional steps: 1) detection of repeats combining a similarity approach with Blaster [13], [26], [27], and a structural approach with LTRharvest [28]; 2) hit clustering with Grouper [26], [27], Recon [29], and Piler [30] for the results of the similarity detection, and with BlastClust from the NCBI-BLAST suite (NCBI Software Development Toolkit) for the results of the structural detection; 3) multiple alignment of each cluster with MAP [31] and building a consensus by cluster; 4) consensus classification with PASTEClassifier [32] based on Wicker classification [33]; and 5) consensus filtering (see TEdenovo documentation). The TEdenovo outputs are a *denovo* classified consensus library, each consensus representing several TEs from a TE family, and a classification information file.

D. TEannot

TEannot is the annotation pipeline of the REPET package (release 2.1, see documentation and availability at https://urgi.versailles.inra.fr/Projects/URGI-softwares/REPET). It runs six functional steps: 1) aligning DNA sequence libraries on the genome by similarity searching using Blaster [13], [26], [27] CENSOR [34], and RepeatMasker (http://www.repeatmasker.org/); 2) combining previous High Score Pairs (HSP) and filtering them; 3) searching for Short Simple Repeats (SSR) with TRF [35], mreps [36], and RepeatMasker (http://www.repeatmasker.org/); 4) removing spurious HSP; 5) connecting distant HSP to build copies of TE elements; and 6) export the results as annotation. The TEannot output is a GFF3 annotation file containing all copies of the input TE library, directly usable in genome browser.

E. RepeatScout

RepeatScout [37] was run separately from the REPET package (release 2.1). The obtained consensuses were then formatted so that the TEdenovo pipeline can be run on it starting from step 5. Hence RepeatScout results were classified and filtered as the standard TEdenovo pipeline did. Note that today, RepeatScout has been integrated into the REPET package (available in the release 2.2) as a new step of the TEdenovo pipeline. Integration allows an input sequence to be formatted as done for other TEdenovo tools, and output sequences are in a format that can be used by subsequent steps.

F. Metrics for Qualifying TE Annotations

In this work, we developed two new metrics. TE annotations were sorted in decreasing order of length, and their cumulative length was computed. NTE50 is the number of TE annotations needed to reach a cumulative length corresponding to 50% of the genome. LTE50 is the size of this TE annotation that reaches this limit.

In addition, we also considered Full Length Copies, hereafter called FLC, as a hallmark of a good TE annotation. We defined an FLC as a TE annotation that is aligned (potentially with large gaps if there are insertions) over more than 95% of their cognate consensus. An annotation with a high FLC number indicates a good recovery of complete copies fragmented because of insertions in their sequence and a capacity to build the long consensuses able to find them.

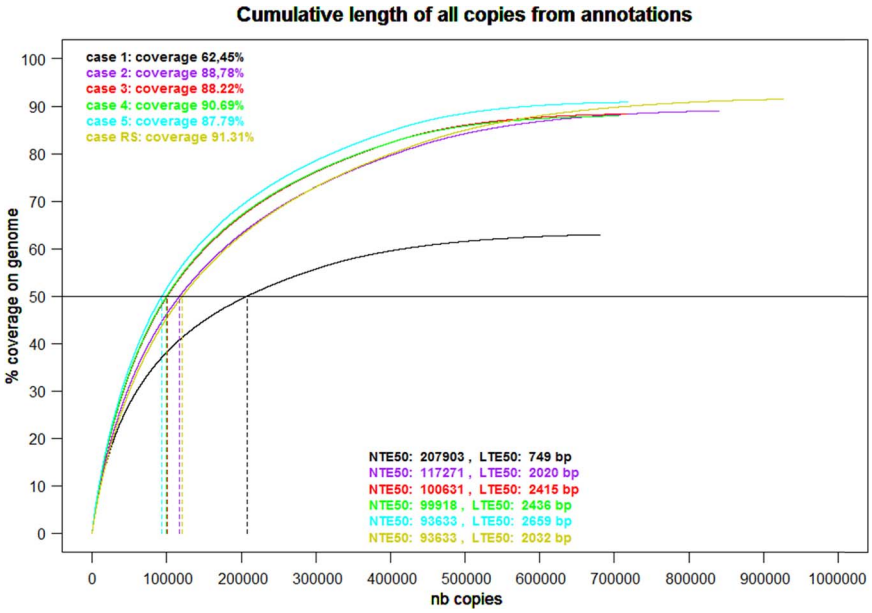


Fig. 1. ***Cumulative length coverage for TE annotations.***

III. RESULTS

A. Case 1: Obtaining a Reference Automated TE Annotation

To obtain a reference TE annotation, we annotated the wheat chromosome 3B with a library containing 586 consensus sequences of previously identified TEs (see Material and Methods); the TEannot pipeline with default parameters was used. We call this reference annotation case 1. We obtained an annotation that covers 84.19% of the genome, and included 14 830 FLC (fragmented and unfragmented annotation aligned over more than 95% of the consensus TE sequence; see Section II).

The plot of cumulative length of all annotated TE copies is a graphical way to compare several annotations (Fig. 1). Derived from this graphical view, we designed a new measure of TE annotation quality, by introducing a new statistic that we called NTE50. Analogous to N50 used to assess assembly quality, NTE50 is the number of the largest copies needed to annotate 50% of the genome. Lower NTE50 values indicate that annotated copies are longer, both a good indication that the annotation has successfully captured large unfragmented TE insertions (a major issue of TE annotation), annotated with long TE consensuses, and a good proxy for well-identified TE. We also define LTE50 as the length of the shortest TE annotation that annotates 50% of the genome; high-value LTE50 indicate better quality results.

The reference annotation exhibits a NTE50 of 124 868 elements and an LTE50 of 1839 bp. These two metrics, both of which are quality measures, the percentage of genome covered by the annotation and the number of FLC TE copies found, will be used as criteria to estimate the accuracy of approaches tested.

B. Automated TE Curation Strategy

Standard TE annotation involves three main steps (see [6]c for details). The first is to detect TE families to build a consensus sequence for each family. In the second step, these consensus sequences are manually curated to remove false positives (i.e., repeated sequences that are not a TE, such as segmental duplication or multigene families), and discard fragments of poorly identified TEs. The third step is to search for similarity to the curated consensus sequences in the genome sequence.

We tested a new strategy, which we call the *second TEannot process*, that is able to reduce the amount of work needed for manual curation (in the second step). By running TEdenovo and then TEannot directly, only consensuses that detect at least one FLC are selected (i.e., copies aligned over more than 95% of the consensus built during the previous step). These consensuses are kept for a second TEannot pass. Indeed, consensuses are built from clusters of repeated sequences. Sometimes some sequences of one cluster overlap sequences found in another cluster (being included for example). When annotating a genome, one consensus only is chosen at each position. In case of cluster overlap, the consensus built from sequences that are included in larger repeated sequences (i.e., found in another cluster), may not have a corresponding FLC. In some case, some consensus may annotate even no copy at all. We can then consider that only one consensus is useful and well built.

To test this consensus number reduction approach, we build a *de novo* TE consensus library using the

RepeatScout software [37]. This software is a less resource-intensive approach than the TEdenovo pipeline that cannot be used directly on the whole chromosome 3B sequence because it is too long and complex. RepeatScout is also known to provide consensuses that are very fragmented and redundant. We built our TE consensus library for the wheat chromosome 3B with RepeatScout (called RS library), and annotated it by the *second TEannot process*. Initially, the RS library contained 15 132 sequences used for the first TEannot pass. In the second TEannot pass, we used only the 9827 consensuses with at least one FLC. Then some consensus appears to have no counterpart in the genome, but this indicate that a better consensus built from another cluster is.

TE genome coverage decreased from 91.57% to 91.31% between the two passes, and the number of FLC annotations increased from 70 577 to 74 379. Indeed, with fewer consensuses, some copies that were previously partially annotated by two consensuses, are annotated in the second step with only one, as the second was removed, and then could appear as a FLC. Consequently NTE50 was 117 822 for pass 1 and 120 976 for pass 2, and LTE50 was 2104 and 2032.

These results show that the number of consensus used for the annotation can be substantially reduced with very little loss of annotation quality measured as the TE annotated fraction and NTE50 or LTE50. This validates the proof of concept for this second TEannot pass as a way to reduce TE consensus complexity, and thereby make annotation simpler. Note that the FLC number is higher with the RS library than with the reference annotation and then might seem better. However, this result is due to the small consensus sizes of the RS library, a consequence of the RepeatScout tool. Consequently, FLC can only be used to compare annotations obtained with the same TE library.

C. General Outline to Reduce Computational Load on Large Genomes

The "classical" TE annotation approach described above is computationally demanding. We tested strategies to reduce the computational load for large genomes, such as plant genomes. Large genomes are mostly comprised of a large number of repeats of a small number of TE families [2], [38], [39]. We exploited this feature by implementing a two-step iterative approach. The first step aims to detect "easy-to-find" highly repeated TEs and to splice them out of the genome; this reduces the length of sequence to be searched for *de novo* detection of "hard-to-find" TEs. Consequently, more sensitive detection and annotation parameters can be applied to a reduced dataset.

TEdenovo with a lightweight computing resource requirement configuration (see below) allows "easy-to-find" TEs to be found quickly; these can then be used to build a "first" consensus library. The library is then used to identify the copies in the genome with the TEannot pipeline. This first annotation is then used to splice the annotated sequence segments out of the initial genome sequence to reduce its size and complexity.

In the second iteration, other TEs (i.e., "hard-to-find" TEs) are detected in the reduced genome using TEdenovo with default parameters and used to build a "second" consensus library. To finish, TEannot annotates the initial unspliced genome sequence with a sequence library corresponding to the concatenation of the "first" and "second" libraries. A second TEannot pass allows automatic curation of the TE library concatenation.

Here, we present four implementations of this strategy (see Fig. 2) to identify the best one for the wheat chromosome 3B.

D. Case 2: Iterative Approach With Known TEs Library

We used the reference annotation obtained with the known TE library of 586 consensuses (see above). We selected consensuses with Full Length Fragment Copies (unfragmented copies aligned on 95% of the consensus), called hereafter *FLF consensuses*, and selected annotated copies having more than 80% identity with these *FLF consensuses*, as a proxy for a well-annotated copy. These copies were then spliced out of the genome, and only contig sequences longer than 500 bp were retained, reducing the genome sequence to be analyzed from 986.1 to 252.7 Mb. We then built a *de novo* library (7009 consensus) using the reduced genome. Finally, we annotated the initial chromosome 3B sequence with a concatenation of the 586 known TEs and the 7009 *de novo* consensuses using the *second TEannot process*. We reduced the first concatenated library containing 7595 consensuses to the 2159 consensus with at least one FLC. The resulting TE annotation shows 88.78% genome coverage with 20 311 FLC. The NTE50 was 117 271 and LTE50 2020.

Compared with the reference annotation (case 1), the coverage is greater, and NTE50 and LTE50 are better indicating better annotation with longer TE copies.

E. Case 3: Iterative Approach With a First LTR *de novo* Library

First, we only looked for LTR families, which are abundant in wheat. Methods that search for their LTR structure are computationally efficient with large genomes, such that the whole wheat chromosome 3B sequence can be searched. We used LTRharvest [28] that uses an efficient suffix array structure to speed up the LTR search phase. This tool was launched in the structural branch of the TEdenovo pipeline.

The first iteration thus builds a *de novo* library using only the structural branch of TEdenovo using sequence contigs whose length is over 5 kb. We obtained 2100 consensuses to build a "LTR" library to be used to annotate the whole chromosome sequence with TEannot. This improved the sensitivity of TE annotation over that

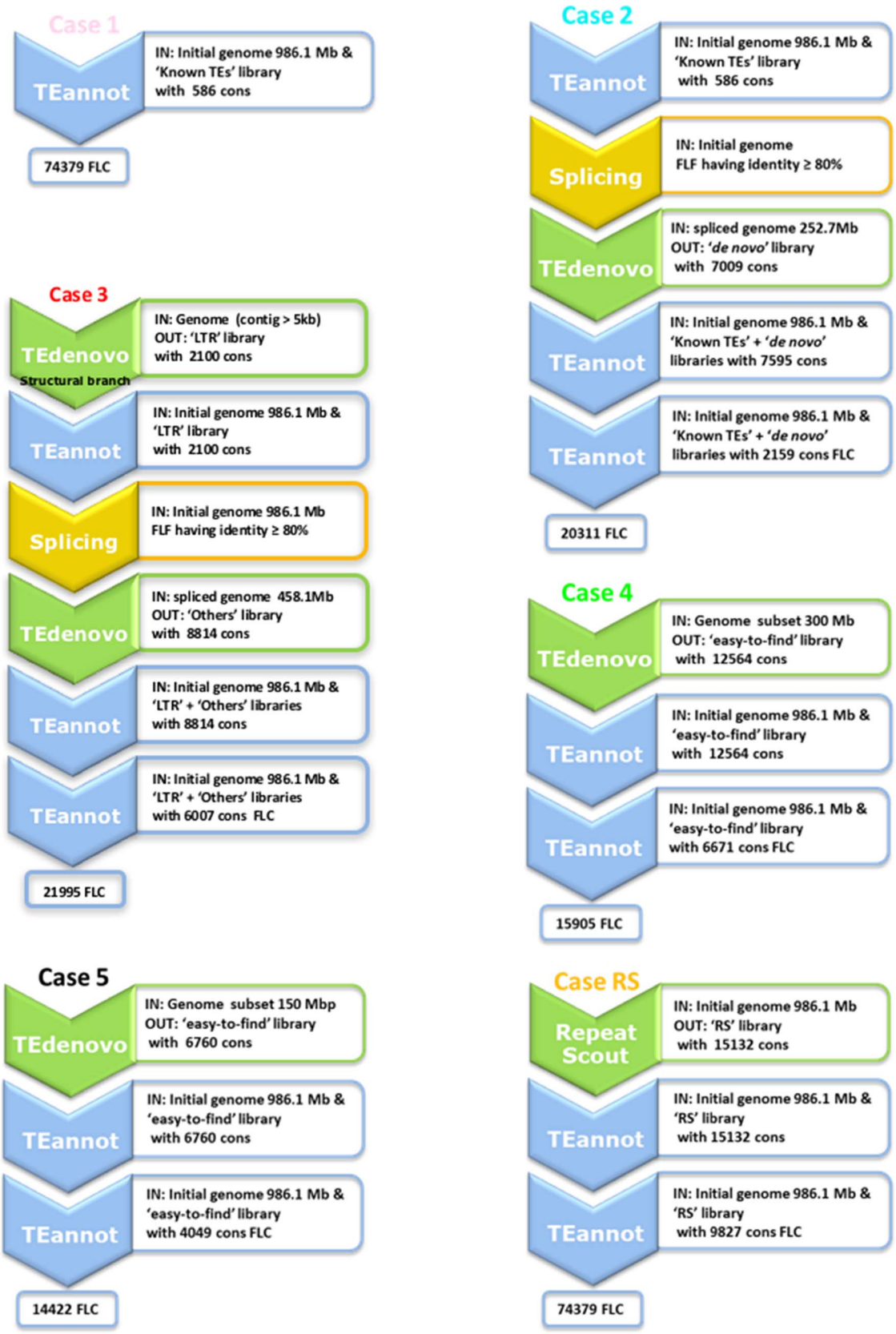


Fig. 2. ***Workflows of the tested strategies. "IN:" refers to input data for the step. "OUT:" indicates the results. "cons" stand for consensus. FLC are full length copies defined as aligning with more than 95% of the consensus.***

of the LTRharvest outputs. The TE annotation obtained was used to reduce the working length of the chromosome sequence by removing TE copies with more than 80% identity to *FLF consensuses*, as in case 2. The second iteration built an "others" library from the working sequence using the complete TEdenovo pipeline with default parameters. We annotated the initial sequence with the concatenation of the "LTR" and the "others" libraries using the *second TEannot process*. We reduced the number of consensuses from 8814 to 6007 having at least one FLC. The resulting TE annotation showed 88.22% genome coverage with 21995 FLC. NTE50 and LTE50 were 100631 and 2415 bp, respectively. Surprisingly, this approach gave a better annotation quality than cases 1 and 2, where some TE are already known and curated.

F. Cases 4 and 5: Iterative Approaches With a *de novo* Library Computed From Sequence Subsets

The rational here is that easy-to-find TEs are so abundant that they can be found on a sequence subset. Consequently, using the TEdenovo pipeline, we built a *de novo* library from a part of the initial sequence (the largest contigs).

In case 4, the longest contigs were concatenated until the cumulative length was 300 Mbp. We used the TEdenovo pipeline(similarity and structural branches) with default parameters, and then annotate the initial sequence using the *second TEannot process*; this reduced the TE library from 12564 consensuses to 6671 consensuses having at least one FLC. The resulting TE annotation showed 90.69% genome coverage with 15905 FLC. The NTE50 and LTE50 were 93633 and 2659 bp, respectively. These results are better than those for cases 1, 2, and 3.

In case 5, we applied the same process with the largest contigs, but with a cumulative length of only 150 Mbp. We used the TEdenovo pipeline with default parameters and annotated the initial sequence using the *second TEannot process*. The *second TEannot process* reduced the TE library from 6760 consensuses to 4049 consensuses having at least one FLC. The resulting TE annotation showed 87.79% genome coverage with 14422 FLC. The NTE50 and LTE50 were 99918 and 2436 bp, respectively. Here, the results are not better than those for case 4, but still better than those for cases 1, 2, and 3.

G. Case 4 Iterative Approach on the Maize Genome

We tested the case 4 approach on the 2 Gb maize genome sequence. We concatenated the longest contigs from all 10 chromosomes until the cumulative length added up to 300 Mbp (2153 contigs represente 299999029 bp). We use the TEdenovo pipeline (similarity and structural branches) with default parameters, and then annotate the initial genome using the *second TEannot process*, reducing the TE library from 15145 consensus to 7319 consensus having at least one FLC. Resulting TE annotation shows 85.87% genome coverage with 41666 FLC. NTE50 and LTE50 are respectively 141391 and 3896 bp. Genome coverage is similar to the 85% that has been described for this genome [1].

IV. DISCUSSION

A. Metrics for TEs Annotation Assessment

We used two novel metrics to assess annotation quality. The first, called NTE50, is the number of the largest copies needed to annotate 50% of the genome. We also defined LTE50 as the length of the shortest TE annotation that annotates 50% of the genome. We propose that, for genomes with a small TE coverage (less than 50%), the metrics can be redefined as largest TE copies (or shortest for LTE50) needed to annotate 50% of the TE genomic space (instead of the full genome space).

These measures favor methods able to annotate the larger copies, considered here as a hallmark of good quality TE annotation. Indeed, poor TE annotations are generally characterized by TE fragmentation in the reconstructed consensus or the annotated genomic copies. The rational behind the quality measure of an annotation with these two metrics is based on this assumption.

Because these metrics depend of the length of the consensus, they can only be used to compare annotations of the same genomic sequence. When comparing two species, we expect different TE family composition and different length distribution of the TE consensus. If one species has many large TEs such as LTRs, both NTE50, and LTE50 will be large, but not because of successful TE consensus reconstruction and copy annotation. Hence, comparison to a species with few large TEs will be uninformative.

The FLC number depends on the consensus sizes of the TE library and is then influenced by the TE detection methods used to build this library. Consequently, this metrics can only be used, contrarily to NTE50 and LTE50, to compare annotation obtained with the same TE detection tools.

B. Conditions for Strategy Efficiency

The REPET pipelines were run on a computer cluster with 75 nodes—totalizing 800+ cores—having between 16 and 512 Go of RAM. On a such cluster it took between three weeks and a few hours to fully annotate a genome according to genome size, repeat density, and cluster usage by other users. For the ~1 Gb wheat 3B chromosomes, we reduced computing time from 3 weeks to 1 with the case 4 strategy, making possible now to run REPET on large genomes of few gigabases. The annotation of the 2 Gb maize genome took two weeks. Disk usage is not a big issue as only 800 Gb is needed temporarily for the maize analysis.

We tested different strategies that may be useful for tackling the fat genome issue. The strategies we used here are efficient because they rely on a feature of large genomes that can be found, i.e., the wheat is populated by few TE families with large copy numbers. Interestingly, we

confirmed here that *de novo* approaches do not need to recover all the TE copies to build a good consensus. They only need to recover few well-conserved TE copies (minimum of three for clustering algorithms, and of one for LTRharvest when used with REPET). If the number of copies in a genome is high, there is a good chance of finding some of them; indeed, even with a subset of the whole sequence, it is still possible to find some of the copies. The probability of finding them also depends on the age of the TE families. Old TE families have divergent copies degraded by numerous mutations and deletions; well-conserved copies of these families may be rare. A limitation of this approach is consequently expected for genomes dominated by a few old TE families. Nevertheless, observations of the TE content of large genomes suggest that the strategy should work well in most cases.

Surprisingly, we show that full *de novo* approaches work better than those using a known TE library. Possibly, because full *de novo* strategies build more consensuses, they are each built from fewer copies. Consequently, these copies are closer to the consensus they built, making alignment better and consequently improving the length of the annotations. Known TEs may be: 1) identified in different genetic background; and/or 2) established from a single sequence as a reference representative of the sequence diversity; or 3) built manually with some inherent subjectivity. In all such cases, they may fail to represent well the diversity of the TE sequence to be annotated. Each TE copy may consequently have a closer reference sequence allowing proper annotation in full *de novo* approaches, than is the case with a library of known TEs built with subjectivity, in particular when restricting each family representative to one or a very small number of sequences.

Surprisingly again, the best strategy was that using a subset of the genome for TE identification. This approach appears to work better than when the full genome was used after its splicing. Possibly, a better consensus is built from a smaller set of sequences than from a larger number. This may be a limitation of the multiple alignment software used. Because of the number of alignments that had to be built, we favored speed over sensitivity. This may be a consequence of this, but obviously, using more sensitive approaches will not solve the problem of computational power required for a large genome.

Interestingly, these iterations reduced the quantity of data to analyze without losing important information, and even improving results in some cases. Consequently, computing time was shortened proportionally.

C. Mixed Strategies

From this study, we propose mixed strategies for the most difficult cases. For genomes longer than 1 Gb, we suggest analysis chromosome by chromosome, or by set of chromosomes, each set being shorter than 1 Gb. Each set should be analyzed independently as a new genome, applying the genome subset strategy if needed. We would expect to obtain as many TE consensus libraries as chromosome sets used, and they should be concatenated into one library for the annotation phase. The *second annotation process*, described above, should remove the inherent redundancy introduced by working independently on each set.

Sometimes the scaffold assignment to chromosomes is not available. This is often the case for draft sequences. In this case, the same approach can be used, but by using sets of scaffolds.

Availability

The REPET package can be downloaded at https://urgi.versailles.inra.fr/Projects/URGI-softwares/REPET. TE consensus and annotation can be downloaded at https://urgi.versailles.inra.fr/Data/Transposable-elements/wheat for the wheat and https://urgi.versailles.inra.fr/Data/Transposable-elements/maize for the maize. ■

Acknowledgment

The auhors would like to thank the REPET team for their help with solving technical issues.

REFERENCES

[1] P. S. Schnable *et al.*, "The B73 maize genome: Complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, Nov. 2009.

[2] J. Daron *et al.*, "Organization and evolution of transposable elements along the bread wheat chromosome 3B," *Genome Biol.*, vol. 15, no. 12, p. 546, Jan. 2014.

[3] F. Choulet *et al.*, "Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces," *Plant Cell*, vol. 22, no. 6, pp. 1686–1701, Jun. 2010.

[4] F. Choulet *et al.*, "Structural and functional partitioning of bread wheat chromosome 3B," *Science*, vol. 345, no. 6194, Jul. 2014, Art. no. 1249721.

[5] E. Lerat, "Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs," *Heredity (Edinb)*, vol. 104, no. 6, pp. 520–533, Jun. 2010.

[6] E. Permal, T. Flutre, and H. Quesneville, "Roadmap for annotating transposable elements in eukaryote genomes," *Methods Mol. Biol.*, vol. 859, pp. 53–68, Jan. 2012.

[7] E.-M. Willing *et al.*, "Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation," *Nature Plants*, vol. 1, no. 2, 2015, Art. no. 14023.

[8] A. Bolger *et al.*, "The genome of the stress-tolerant wild tomato species Solanum pennellii," *Nature Genetics*, vol. 46, no. 9, pp. 1034–1038, Sep. 2014.

[9] T. Slotte *et al.*, "The Capsella rubella genome and the genomic consequences of rapid mating system evolution," *Nature Genetics*, vol. 45, no. 7, pp. 831–835, Jul. 2013.

[10] J. L. Olsen *et al.*, "The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea," *Nature*, vol. 530, no. 7590, pp. 331–335, Feb. 2016.

[11] B. A. Read *et al.*, "Pan genome of the phytoplankton emiliania underpins its global distribution," *Nature*, vol. 499, no. 7457, pp. 209–213, Jul. 2013.

[12] J. M. Cock *et al.*, "The Ectocarpus genome and the independent evolution of multicellularity in brown algae," *Nature*, vol. 465, no. 7298, pp. 617–621, Jun. 2010.

[13] H. Quesneville *et al.*, "Combined evidence annotation of transposable elements in genome sequences," *PLoS Comput. Biol.*, vol. 1, no. 2, pp. 166–175, Jul. 2005.

[14] T. Wicker *et al.*, "The wheat powdery mildew genome shows the unique evolution of an obligate biotroph," *Nature Genetics*, vol. 45, no. 9, pp. 1092–1096, Sep. 2013.

[15] F. Martin *et al.*, "Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis," *Nature*, vol. 464, no. 7291, pp. 1033–1038, Apr. 2010.

[16] T. Rouxel *et al.*, "Effector diversification within compartments of the Leptosphaeria maculans genome affected by repeat-induced point mutations," *Nature Commun.*, vol. 2, p. 202, Jan. 2011.

[17] C. A. Cuomo *et al.*, "The fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization," *Science*, vol. 317, no. 5843, pp. 1400–1402, Sep. 2007.

[18] P. Abad *et al.*, "Genome sequence of the metazoan plant-parasitic nematode meloidogyne incognita," *Nature Biotechnol.*, vol. 26, no. 8, pp. 909–915, Aug. 2008.

[19] J. Amselem *et al.*, "Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea," *PLoS Genetics*, vol. 7, no. 8, Aug. 2011, Art. no. e1002230.

[20] S. Duplessis *et al.*, "Obligate biotrophy features unraveled by the genomic analysis of rust fungi," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 22, pp. 9166–9171, May 2011.

[21] P. D. Spanu *et al.*, "Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism," *Science*, vol. 330, no. 6010, pp. 1543–1546, Dec. 2010.

[22] V. Nene *et al.*, "Genome sequence of aedes aegypti, a major arbovirus vector," *Science*, vol. 316, no. 5832, pp. 1718–1723, Jun. 2007.

[23] A. G. Clark *et al.*, "Evolution of genes and genomes on the drosophila phylogeny," *Nature*, vol. 450, no. 7167, pp. 203–218, Nov. 2007.

[24] J. A. Eisen, "Genome sequence of the pea aphid acyrthosiphon pisum," *PLoS Biol.*, vol. 8, no. 2, Feb. 2010, Art. no. e1000313.

[25] N. Buisine, H. Quesneville, and V. Colot, "Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets," *Genomics*, vol. 91, no. 5, pp. 467–475, May 2008.

[26] H. Quesneville, D. Nouaud, and D. Anxolabéhère, "Detection of new transposable element families in drosophila melanogaster and anopheles gambiae genomes," *J. Mol. Evol.*, vol. 57 Suppl 1, pp. S50–S59, Jan. 2003.

[27] T. Flutre, E. Duprat, C. Feuillet, and H. Quesneville, "Considering transposable element diversification in de novo annotation approaches," *PLoS ONE*, vol. 6, no. 1, Jan. 2011, Art. no. e16526.

[28] D. Ellinghaus, S. Kurtz, and U. Willhoeft, "LTR harvest, an efficient and flexible software for de novo detection of LTR retrotransposons," *BMC Bioinf.*, vol. 9, p. 18, Jan. 2008.

[29] Z. Bao and S. Eddy, "Automated de novo identification of repeat sequence families in sequenced genomes," *Genome Res.*, vol. 12, no. 8, pp. 1269–1276, 2002.

[30] R. C. Edgar and E. W. Myers, "PILER: Identification and classification of genomic repeats," *Bioinformatics*, vol. 21 Suppl 1, pp. i152–i158, Jun. 2005.

[31] X. Huang, "On global sequence alignment," *Comput. Appl. Biosci.*, vol. 10, no. 3, pp. 227–235, Jun. 1994.

[32] C. Hoede *et al.*, "PASTEC: An automatic transposable element classification tool," *PLoS ONE*, vol. 9, no. 5, Jan. 2014, Art. no. e91929.

[33] T. Wicker *et al.*, "A unified classification system for eukaryotic transposable elements," *Nature Rev. Genetics*, vol. 8, no. 12, pp. 973–982, 2007.

[34] J. Jurka, P. Klonowski, V. Dagman, and P. Pelton, "CENSOR–a program for identification and elimination of repetitive elements from DNA sequences," *Comput. Chem.*, vol. 20, no. 1, pp. 119–121, Mar. 1996.

[35] G. Benson, "Tandem repeats finder: A program to analyze DNA sequences," *Nucleic Acids Res.*, vol. 27, no. 2, pp. 573–580, Jan. 1999.

[36] R. Kolpakov, G. Bana, and G. Kucherov, "Mreps: Efficient and flexible detection of tandem repeats in DNA," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3672–3678, Jul. 2003.

[37] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21 Suppl 1, pp. i351–i358, Jun. 2005.

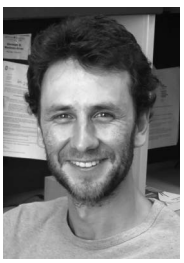
[38] R. Baucom *et al.*, "Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome," *PLoS Genetics*, vol. 5, no. 11, 2009, Art. no. e1000732.

[39] J. Wegrzyn *et al.*, "Unique features of the loblolly pine (pinus taeda L.) megagenome revealed through sequence annotation," *Genetics*, vol. 196, no. 3, pp. 891–909, 2014.

ABOUT THE AUTHORS

Véronique Jamilloux received the M.S. degree in applied physics in 1987.

She was recruited as software engineer in several informatics services and developments companies. In 2008, she joined, as bioinformatics engineer, the URGI, an INRA bioinformatics research unit dedicated to plants and crop parasites (Versailles, France). She is responsible of REPET software and methodological development, a pipeline package for repeats detection and classification.



Josquin Daron received the Ph.D. degree in 2015.

He worked in Feuillet's Lab at GDEC, an INRA research unit on Genetics Diversity and Ecophysiology of Cereals (Clermont-Ferrand, France). He contributed to the characterization of nonsyntenic genes in wheat compared to related species, and studied the evolutionary dynamics of transposable elements. He is currently working as a Postdoctoral Fellow in Slotkin's lab (Department of Molecular Genetics, Ohio State University, Columbus, OH, USA). His current research interests include evolution, transposable element dynamics, comparative genomics, and epigenetics.

Frédéric Choulet received the Ph.D. degree in bacterial genomics from the University of Nancy, France, in 2006.

He then joined the GDEC, an INRA research unit on Genetics Diversity and Ecophysiology of Cereals (Clermont-Ferrand, France) as a Postdoctoral Fellow to work on plant genomics and, more precisely, on the complex bread wheat genome. Since 2009, he has been a bioinformatics group leader at GDEC and an active member of the International Wheat Genome Sequencing Consortium, which aims at obtaining a reference sequence for this amazingly large and polyploid genome. He leads projects in wheat genome sequencing and is particularly interested in understanding the relationship between genome structure, expression, and evolution, mainly focusing on transposable element dynamics and gene duplications.

Hadi Quesneville received the Ph.D. degree in 1996.

Head of URGI, an INRA bioinformatics research unit dedicated to plants and crop parasites (Versailles, France), he has a long-standing track record in the field of bioinformatics and genomics. His research focuses on the analysis of genomes, developing methods and tools to annotate, store and explore repeated sequences through "omics" data analysis and data integration. He is an expert in transposable elements and repeats annotation for which he leads the development of the REPET package. His work contributes to understanding the impact of repeats on genome and epigenome structure and evolution. He is chair of the wheat information system expert working group aiming at providing a federation of information systems for the wheat genomics and genetics.