Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

MS28

Software for the analysis of repeats

Due date of milestone: August 31, 2013

Start date of project: 1.9.2011 Duration: 48 months

Contributor: P5 (INRA)

Context

The recent successes of new sequencing technologies allow today to sequence increasingly large genomes at reduced costs. Transposable elements (TEs) constitute the most structurally dynamic components and the largest portion of nuclear sequences of these large genomes, e.g. 85% of the maize genome (Schnable et al. 2009), and 88% of the wheat genome (Choulet et al. 2010). Therefore, TEs annotation should be considered as a major task in these genome projects.

However, this still remains a major challenge, since a good TE annotation relies critically on an expertly assembled reference sequence set, data that currently cannot be obtained in an automatic fashion. This crucial step is now a bottleneck for many genome analyses.

Results

We scaled-up a repeat detection and an annotation pipeline, both part of the REPET package (Flutre, Duprat, Feuillet, & Quesneville, 2011), now at its v2.2 release (http://urgi.versailles.inra.fr/Tools/REPET). The two pipelines called TEdenovo and TEannot respectively build a TEs library and annotate TE copies in the genome. The TEdenovo pipeline strategy is to find as much as possible potential TEs, and then to classify putative TEs in order to filter out false positives. The pipeline starts by the detection of repeated sequences comparing by alignments the genome with itself. These alignments are independently clustered according to RECON (Bao & Eddy, 2002), GROUPER(Flutre et al., 2011; Quesneville, Nouaud, & Anxolabéhère, 2003), PILER (Edgar & Myers, 2003).Then, it builds multiple alignments from the clusters, from which a consensus sequence is derived. These consensus are classified according to TE features and redundancy is removed. Finally, there is the possibility to remove false-positives according to the classification (SSR, host genes, rDNA and under-represented unclassified consensus).

Two steps have been improved since REPET v1 (Flutre et al., 2011). A structural TE detection approach is now implemented. LTRharvest (Ellinghaus, Kurtz, & Willhoeft, 2008) is used to search for LTR retrotransposons, using structural features of this TE category. Potential TEs hence detected and all other derived consensus are put together before the classification and a redundancy removal step. Classification has been also improved with the development of PASTEC, a new classifier that we have developed. It tests all TE classifications, each result being weighted according to the evidences found. In addition to similarities to known TEs in Repbase Update and the search for repeated structures, it also uses HMM profiles, which are interesting to classify TEs and to detect host genes. PASTEC gives precisions about completeness and indicates if TEs are potentially chimerics.

We also propose now a new pipelines based on Tallymer (Kurtz, Narechania, Stein, & Ware, 2008), called TallymerPipe, as pre-processing tool for a fast repeated region detection. We also propose SegDup, a pipeline to detect segmental duplications, based on our previous work (Fiston-Lavier, Anxolabehere, & Quesneville, 2007).

Using these pipelines, we apply a new strategy (with tools), to cope with very large genomes such as the wheat. This strategy is an iterative approach and can be summarized as follows:

1) Detection of the most easy to found TEs, with stringent parameters, to build a first TE library. They often corresponds to young TEs and the less degenerate ones,

2) TE annotation and splicing of the corresponding sequences from the initial contigs. We then obtain a reduced genome sequence.

3) Detection of the other TEs with sensitive parameters on the reduced genome sequence to build a second TE library.

4) Annotation of the original contigs with the concatenation of the two TE libraries.

The rational here is that these large genomes are mostly made of few TE families easy to found because present in number of copies. They will be detected in the first step and this will allow reducing the genome size by an important factor. Using this approach we were able to reduce the wheat 3B chromosome from 986Mbp to ~230Mb, a reasonable size for a detection of TEs with sensitive parameters.

Literature cited

Bao, Z., & Eddy, S. R. (2002). Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes, 1269–1276. doi:10.1101/gr.88502.

Edgar, R. C., & Myers, E. W. (2003). BIOINFORMATICS PILER : identification and classification of genomic repeats, 1–7.

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, *9*, 18. doi:10.1186/1471-2105-9-18

Fiston-Lavier, A.-S., Anxolabehere, D., & Quesneville, H. (2007). A model of segmental duplication formation in Drosophila melanogaster. *Genome research*, *17*(10), 1458–70. doi:10.1101/gr.6208307

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS one*, *6*(1), e16526. doi:10.1371/journal.pone.0016526

Kurtz, S., Narechania, A., Stein, J. C., & Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics*, *9*(1), 517. doi:10.1186/1471-2164-9-517

Quesneville, H., Nouaud, D., & Anxolabéhère, D. (2003). Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *Journal of molecular evolution*, *57 Suppl 1*, S50–9. doi:10.1007/s00239-003-0007-2