The role of transposable elements on gene expression regulation in maize and its response to drought

Maud Fagny, Véronique Jamilloux, Johann Joets, Clémentine Vitte

Optimizing the *de novo* annotation of transposable elements in large, TE-rich genomes

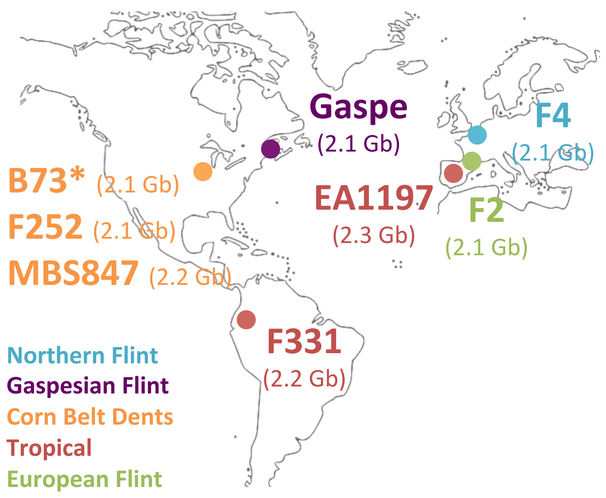
Maud Fagny[1], Véronique Jamilloux[2], Johann Joets[1], Clémentine Vitte[1]
[1]Génomique Quantitative et Evolution – Le Moulon, INRA, Université Paris-Saclay, CNRS, AgroParisTech, Université Paris-Sud, Gif-sur-Yvette, France. [2]Unité de recherche Génomique Info, INRA, Versailles, France

Background

Transposable elements (TEs) are major constituents of plant genomes and contribute to their dynamics and evolution. Existing softwares packaging tools using TE structural features detection and repetitiveness approaches have proven useful for TE annotation of small genomes. Attempts have been made to adapt them for large genomes analysis (Jamilloux *et al*., 2016 - *Proceedings of the IEEE*). However, this remains challenging, mainly due to the amount of genomic data to analyze and to the structural complexity of these genomes. In particular, tools based on repetitiveness can be run only on a subset of the genome because of the required amount of memory and computation times. Tools based on structural features have been successfully applied on wheat (Wicker *et al*., 2018 - *Genome Biology*) and maize (Anderson *et al*., 2019 – *bioRχiv).* However*,* these studies were based on a succession of in-house scripts, and no pipeline has emerged so far to annotate all TE classes in large genomes. Consequently, while half of angiosperm species have a genome size above 1.6 Gb (Pellicer *et al*., 2018 - *Genes*), only a handful of large plant genomes have been annotated so far. To better characterize the potential role of TEs in maize local adaptation, we aim at characterizing TE polymorphisms in seven maize inbred lines of contrasted origins. With its 2.1 Gb genome, we use maize as a model for TE annotation of an average angiosperm genome.

Material & Methods

Genetic material



*Maize reference genome – not sequenced by us

Genome assembly

***De novo* genome sequence assembly**
(nrgene)
Illumina short reads
(coverage >180X)

De novo TE annotation**

Structural features + TE-specific genes
(TIRs, LTRs, *gag*, *pol*, Tase, RPA, …)
+
TE Clustering and redundancy removal
(by sequence similarity 80-80-80)

Repetitiveness
(all-against-all alignment of the genome)
+
TE Classification
(by annotation of TE-specific features)

Database-dependent TE annotation

Similarity
(annotation of old, degraded and incomplete TE copies)

**The approaches shown here are built to *de novo* annotate TEs on assembled genome sequences. Other approaches exist to detect TEs from unmapped reads.

Test of *de novo* TE annotation tools on the B73 genome

		Structural features	Repetitiveness	Similarity
	Examples *tested* (*not tested*)	*LTRharvest, TEA, detectMITE, SINE-finder, HelitronScanner* (*SINE_scan, MGEScan, LTR_retriever, …*)	*REPET-TEdenovo* (*RepeatScout, RepeatModeler*)	*RepeatMasker, REPET-TEannot* (*TE-HMMER*)
All genomes	**Independency from TE database**	+	+	-
	Detection of old, incomplete TE copies	--	+/-	+
	Accuracy of TE boundaries	+	-	+/-
	Detection of highly variable TE superfamilies (Helitrons)	+	--	+
Large genomes	**Detection of nested TE**	++	+/-	+/-
	Sensitivity to low copy-number TE families	+	-	+
	Specificity (low false positive rate)	--	+/-	+/-
	Memory usage	+	--	+
	Computation time (32 cores, 96GB RAM) for maize B73 genome (2.1Gb)	Weeks	Months	Days

Table 1: Comparison of different *de novo* TE annotation approaches

- Structural-based tools are relatively fast to run on large genomes, but have high false-positive rates and necessitate additional filtering steps.
- Repeats-based tools have a lower false-positive rate, but are slow and must be run on a subset of the genome (<300Mb), hampering the detection of low copy-number TEs, which may play an important role in adaptation.

Conclusions

Finding the right accuracy/computing time balance to *de novo* annotate TEs in large plant genomes is challenging. A few softwares exist that integrate both structure-based and repetitiveness-based tools, such as *REPET-Tedenovo* (Quesneville *et al.*, 2015 - *PLoS Comp Bio*) or *PiRATE* (Berthelier *et al.*, 2018 – *BMC Genomics*). But their application to large genomes is hampered by the long computation time (up to months), the high amount of memory necessary to run the repetitiveness-based tools on the whole genome, and the absence of filtering steps to filter out the false positives generated by structural features-based tools.

Future directions

New pipeline based on structural features

- Package several existing structure-based tools
- Provide automatized, tunable filtering steps
- Include automatized TE clustering steps

Modified integrated pipelines

- Modify data storage structures
- Improve memory usage for repetitiveness-based tools
- Add tunable filters to clean up structure-based results