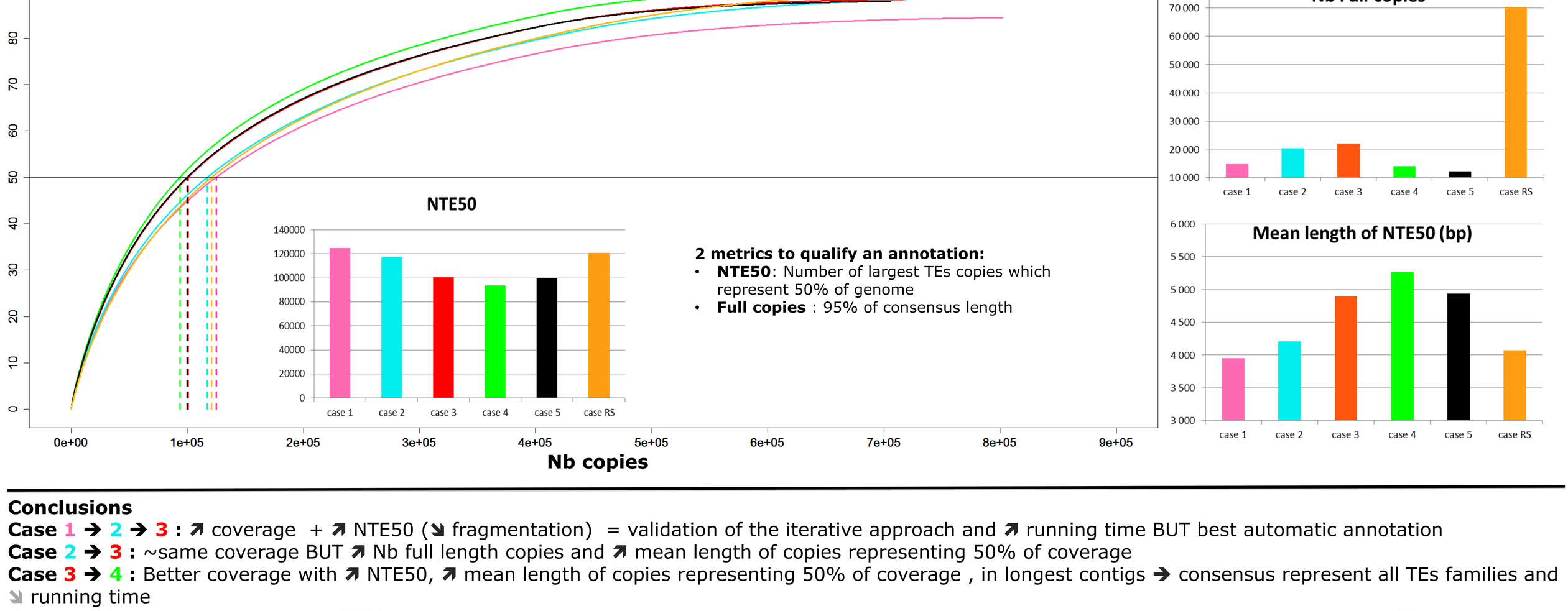
De novo TE annotation : tackling the fat genomes issue

Véronique Jamilloux[1], Olivier Inizan[1], Timothée Chaumier [1], Mark Moissette [1], Sandie Arnoux[2], Hadi Quesneville[1]

[1] INRA-URGI (Unité de Recherche en Génomique-Info), UR1164, INRA de Versailles-Grignon, Route de Saint-Cyr, Versailles, 78026 France
[2] INRA-BioEpAR (Biologie, Epidémiologie et Analyse du Risque), UMR1300, ONIRIS, route de Gachet la Chantrerie, Nantes, 44307 France
Contact: veronique.jamilloux@versailles.inra.fr

Abstract Transposable elements (TEs) constitute the most dynamic and the largest component of large plant genomes, e.g. 85% of the maize genome [1], and 88% of the wheat genome [2]. *De novo* TEs annotation is therefore a computational challenge. We designed new strategies to this end based on the functionalities of the REPET package [3]. This package includes 2 pipelines: TEdenovo that builds a repeated sequences consensus library, and TEannot that annotates copies of library in the genome.
For our methodological developments, we chose the wheat 3B chromosome sequence (~1Gbp) as an experimental model. We describe the principles and the results of several strategies by different use cases:

Case 1: A library of known TEs [4]
Case RS: RepeatScout [6] consensus library. In each cases the consensus are classified and used in an iterative annotation process. **Case 2**: A library of known TEs concatenated to a *de novo* library built from a genome spliced from their known TEs,
Case 3: A *de novo* LTR-retrotransposon library obtained with the LTRHarvest software [5] then concatenated with the *de novo* library built from a genome spliced from an already identified LTR-Rns library,
Case 4 & **5**: A *de novo* TEs consensus library obtained from a genome part (300 Mbp and 150 Mbp longest contigs),



Legend

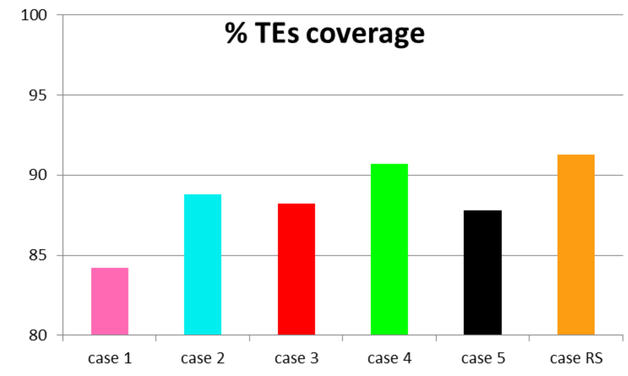
★ 3B chromosome V2.1, first to be fully sequenced, 294 691 contigs, 986.1 Mbp [1]

Cons FLC: consensus having at least 1 Full Length Copy

REPET package V2.0 [7]

2 metrics to qualify an annotation:

- **NTE50**: Number of largest TEs copies which represent 50% of genome
- **Full copies** : 95% of consensus length



Conclusions

Case 1 ➔ 2 ➔ 3 : ↗ coverage + ↗ NTE50 (↘ fragmentation) = validation of the iterative approach and ↗ running time BUT best automatic annotation
Case 2 ➔ 3 : ~same coverage BUT ↗ Nb full length copies and ↗ mean length of copies representing 50% of coverage
Case 3 ➔ 4 : Better coverage with ↗ NTE50, ↗ mean length of copies representing 50% of coverage , in longest contigs ➔ consensus represent all TEs families and ↘ running time
Case 4 ➔ 5 : ↘ coverage with ↗ NTE50 (↗ fragmentation) with ↘ running time = consensus from too small genome part don't represent enought TEs
Case 4 ➔ 6 : ↗ coverage with ↗ NTE50 (↗ fragmentation) with same running time, ↗ nb full length copies but they are too small.
Our analyses show that all our strategies enable us to overcome the current memory and time limitations for *de novo* TEs discovery and annotation using REPET on large plant genomes, effectively. This study paves the route towards comprehensive and high quality automatic TEs annotation in a number of economically and agronomically important species.

Perspectives : We are testing these strategies on the maize genome (2,3 Gbp)

[1] Schnable, PSD WareRS Fulton, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115
[2] Choulet, F, T Wicker, C Rustenholz, et al. 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686-1701
[3] Flutre, T, E Duprat, C Feuillet, H Quesneville. 2011. Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526.(urgi.versailles.inra.fr/download/repet/REPET_linux-x64-2.2.tar.gz)
[4] Daron J, 3b manual curated TEs library based on Trep, 2012
[5] Ellinghaus D, Kurtz S, Willhoeft U 2008. LTRHarvest, an efficient and flexible software for denovo detection of LTR retrotransposons. BMC Bioinformatics doi:10.1186/1471-2105-9-18
[6] Price A. L., Jones N. C. and Pevzner P. A. 2005. Denovo identification of repeat families in large genomes. 13 Annual Conference on Intelligent System for Molecular Biology (ISMB 05) Daron J, 3b manual curated TEs library based on Trep, 2012
[7] REPET package V2.2 is downloadable https://urgi.versailles.inra.fr/download/repet/REPET_linux-x64-2.2.tar.gz