Improvement of the assembly of heterozygous genomes of non-model organisms

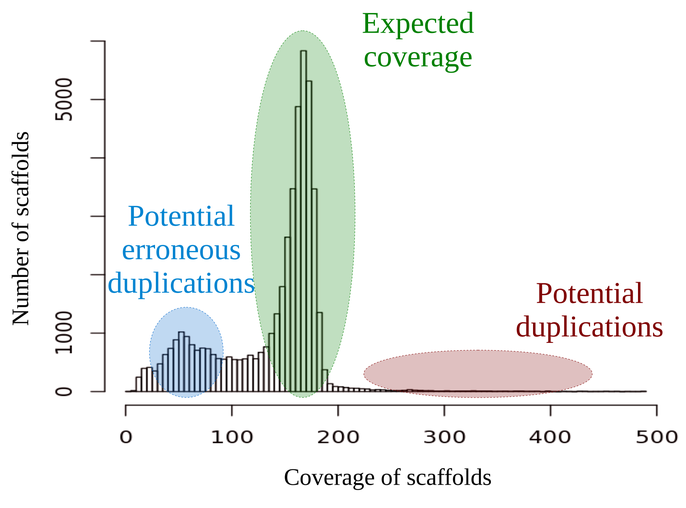
Anaïs Gouin, Anthony Bretaudeau, Emmanuelle d'Alençon, Claire Lemaitre, Fabrice Legeai

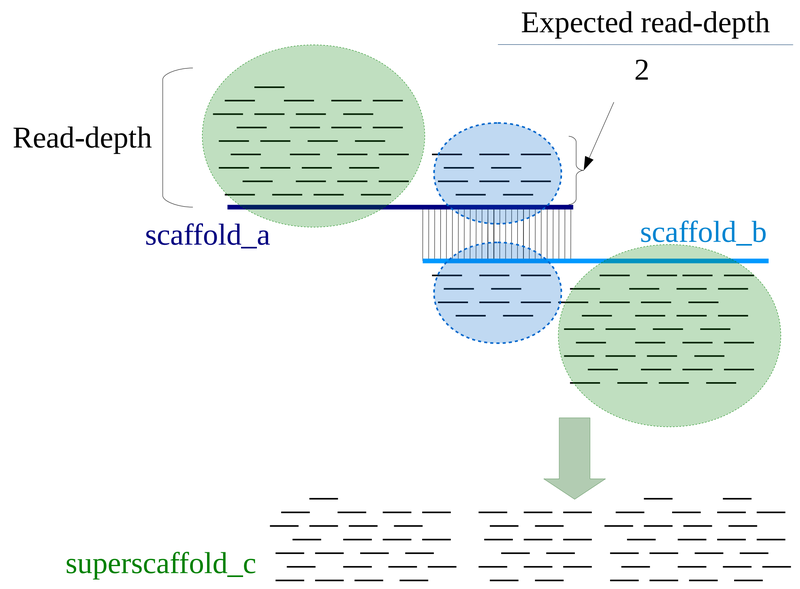
Improvement of the assembly of heterozygous genomes of non-model organisms

Anaïs GOUIN[1], Anthony BRETAUDEAU[2], Emmanuelle d'Alençon[3], Claire LEMAITRE[1] and Fabrice LEGEAI[2]
[1]INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes cedex, France
[2]INRA, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Domaine de la Motte – 35653 Le Rheu
[3]INRA DGIMI, université de Montpellier 1, 34000 Montpellier

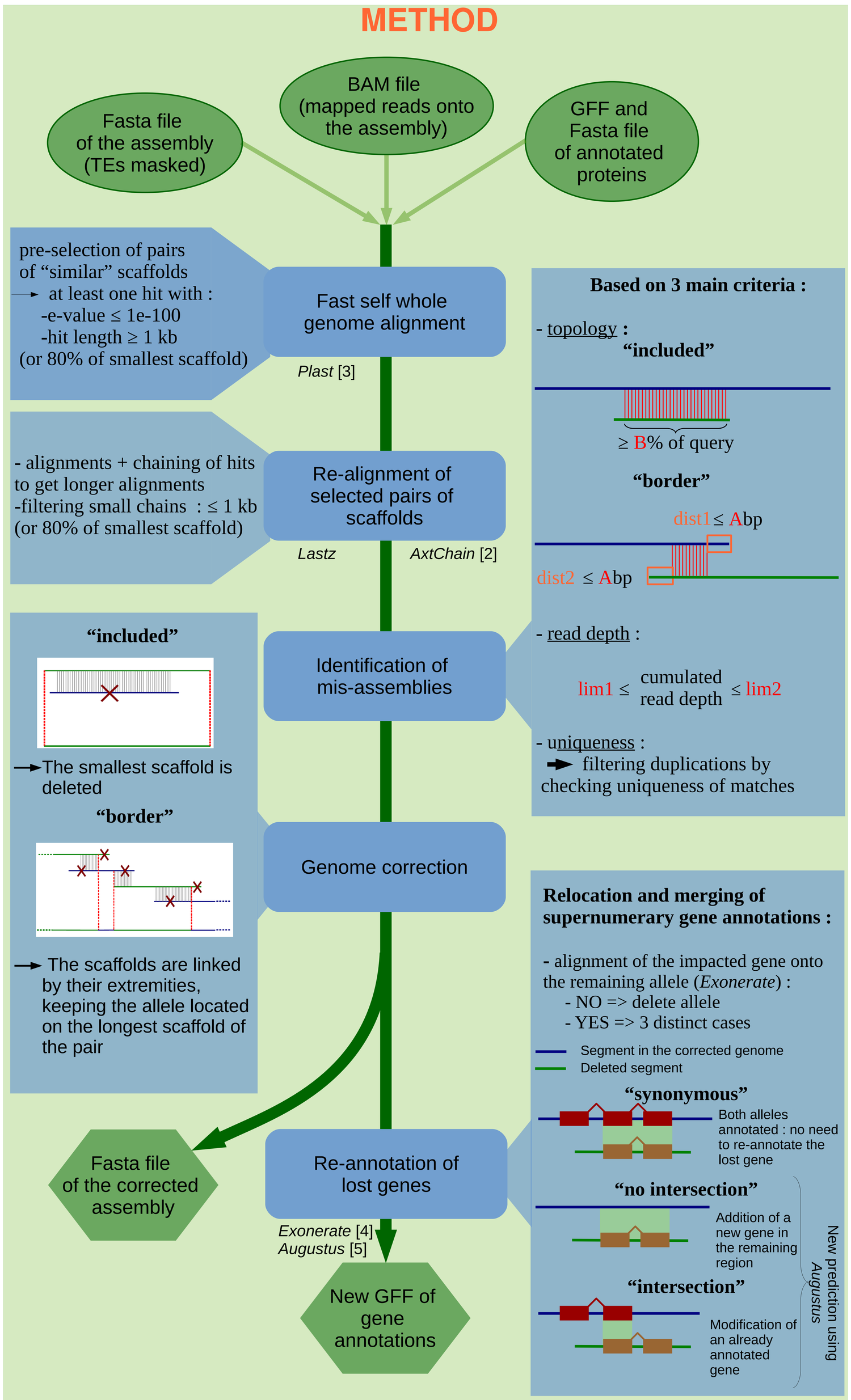
Motivation : Some heterozygous regions have a significant divergence between the two haplotypes and the assembly process can lead to the construction of two different contigs, instead of one consensus sequence.

Objective : Set up a strategy to detect and correct false duplications in already-built assemblies.





METHOD



APPLICATION

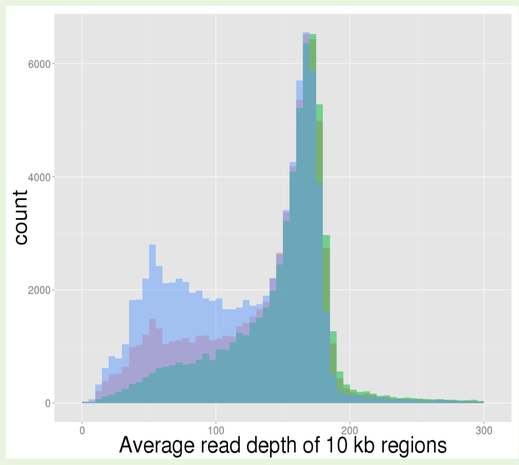
Spodoptera frugiperda genome

Genome correction

★ Comparison with another method : *Haplomerger* [1]
Expected size : ~ 400 Mb

	Initial assembly *Allpaths*	Corrected assembly	*Haplomerger*
Total size (Mb)	526.0	434.9	369.5
Nb. scaffolds	48,272	41,577	37,797
N50 (kb)	39.6	52.8	58.4

★ Read depth analysis : before/after correction



➡ Improvement of the initial assembly for both methods

➡ *Haplomerger* merged more regions, leading to a smaller final assembly

★ BUSCO statistics : Benchmarking sets of Universal Single-Copy Orthologs (2,675 for Arthropoda species) [6]

	Initial assembly	Corrected assembly	*Haplomerger*
Missing	363	**336** *	562
Single copy	1,246	**1,586** *	1,242
Fragmented	476	**457** *	771
Duplicated	590	296	**100** *

* best result by category

- **Reduction of the genome size (17%), increase of the N50 and more single copies for important genes**
- **Reduces less than *Haplomerger* ➡ gain of numerous BUSCO genes**

➡ **Our method: more conservative, preserves genome consistency and allows easier re-annotation of impacted genes**

Annotation stats

Previous release : 25,041 genes
==> 3,746 genes to re-annotate

	# genes	% success
"no alignment"	34	0
"synonymous"	747	100
"no intersection"	643	45.4
"intersection"	2,322	86.3

==> Overall success of 80% / New release : 21,578 genes

[1] Huang S. Et al, Genome research, 2012
[2] Kent WJ. Et al, Proceedings of the National Academy of Sciences, 2003.
[3] Nguyen V.H et al, *BMC Bioinformatics*, 2009
[4] Slater G. S. et al, BMC bioinformatics, 2005
[5] Stanke M. Et al, Genome Biol, 2006
[6] Waterhouse et al, *Nucleic Acids Research*, 2013