Identification and correction of genome mis-assemblies due to heterozygosity

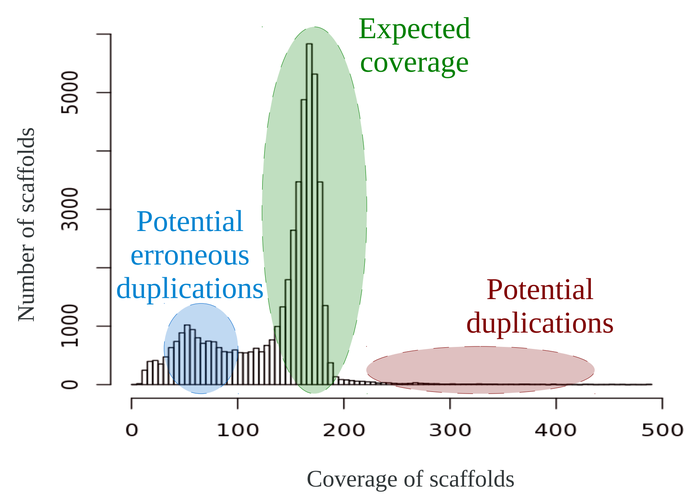
Anaïs Gouin, Anthony Bretaudeau, Claire Lemaitre, Fabrice Legeai

Identification and correction of genome mis-assemblies due to heterozygosity

Anaïs GOUIN[1], Anthony BRETAUDEAU[2], 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

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.**



PIPELINE

Pre-selection of "similar" scaffolds — *Plast* [3]

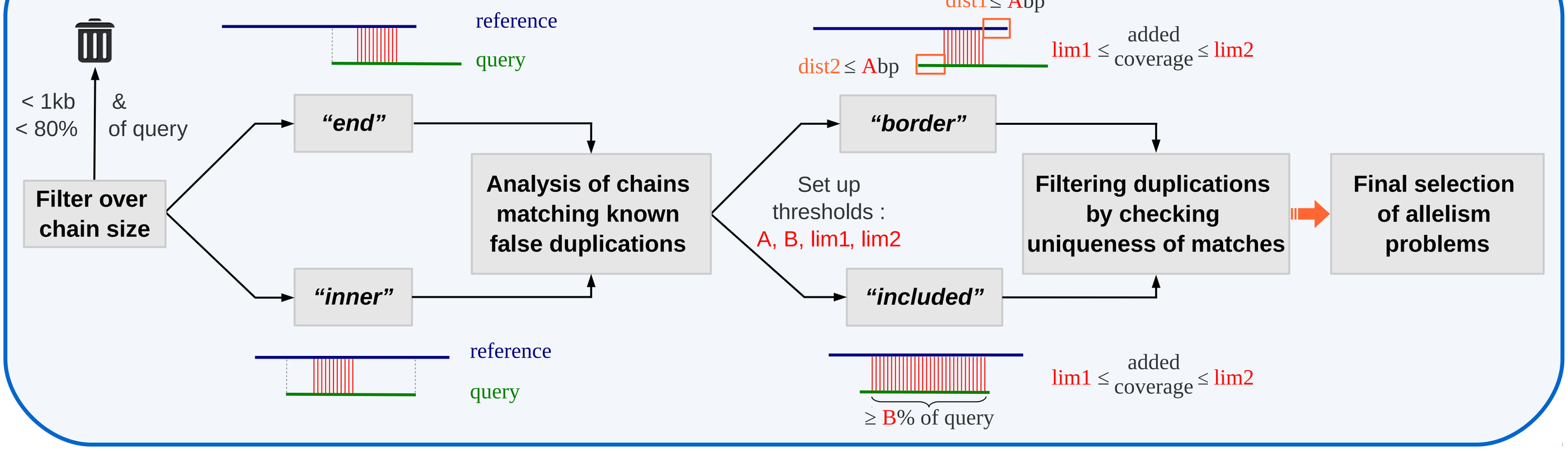
Re-alignment of selected pairs — *Lastz* [1]

Chaining

Identification of mis-assemblies

Genome correction — Remove scaffolds (entire or parts)

Identification of mis-assemblies

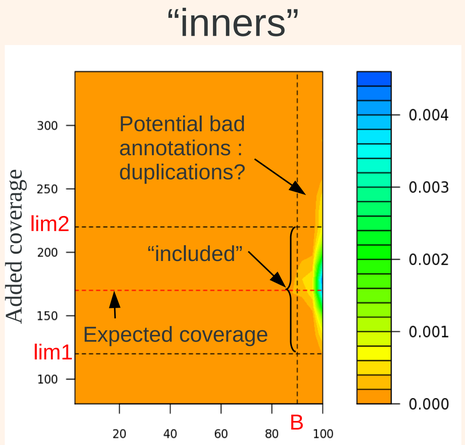


Application : *Spodoptera frugiperda* genome

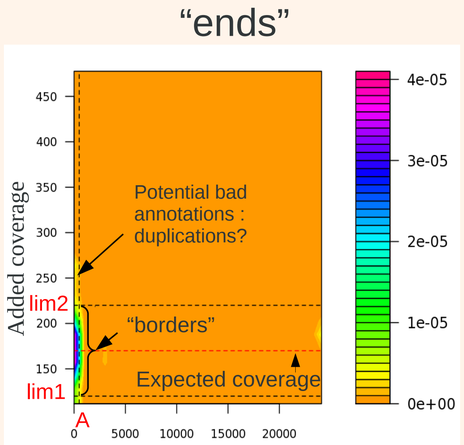
Set up thresholds

206 chains matching known false duplications (manually curated) :
153 "inners" / 53 "ends"

"inners"



"ends"



→ 114 "included" / 38 "borders"

➠ ~80% of known allelic regions with chosen thresholds

Genome correction

Expected size : ~ 375 Mb

	Initial assembly *Allpaths*	*Platanus* [2] assembly	Corrected assembly
Total size (Mb)	526,0	470,1	434,9
Nb. scaffolds	48 272	41 633	41 536
N50 (bp)	39 593	75 578	52 733
BUSCO* stats — No hit	14	9	15
BUSCO* stats — Single hit	1497	2047	1904
BUSCO* stats — Multi hit	782	237	374

*BUSCO : Benchmarking sets of Universal Single-Copy Orthologs (Arthropoda species) [4]

→ **Our strategy allows a good improvement of the initial assembly.**

→ **Performing a new assembly with a tool handling heterozygosity (such as *Platanus*) is still more efficient.**

Perspectives :

Non-studied here : potential problems of allelism within a scaffold, at contig level?

Other parameters to take into account : SNPs count (to select regions) / mate pairs (to validate corrected assembly)

[1] Harris RS: Improved pairwise alignment of genomic DNA. Ann Arbor: ProQuest; 2007:84

[2] Kajitani R. et al, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, *Genome Res.* 2014; 24(8):1384-95

[3] Nguyen V.H., Lavenier D., PLAST: parallel local alignment search tool for database comparison, *BMC Bioinformatics*, vol 10, no 329, 2009

[4] Waterhouse et al, OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs, *Nucleic Acids Research*, 2013, PMID:23180791