

# Identification and correction of genome mis-assemblies due to heterozygosity

Anaïs Gouin, Anthony Bretaudeau, Claire Lemaitre, Fabrice Legeai

► **To cite this version:**

Anaïs Gouin, Anthony Bretaudeau, Claire Lemaitre, Fabrice Legeai. Identification and correction of genome mis-assemblies due to heterozygosity. European Conference on Computational Biology (ECCB), Sep 2014, Strasbourg, France. ECCB 2014, 2014. <hal-01092959>

**HAL Id: hal-01092959**

**<https://hal.inria.fr/hal-01092959>**

Submitted on 9 Dec 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Identification and correction of genome mis-assemblies due to heterozygosity

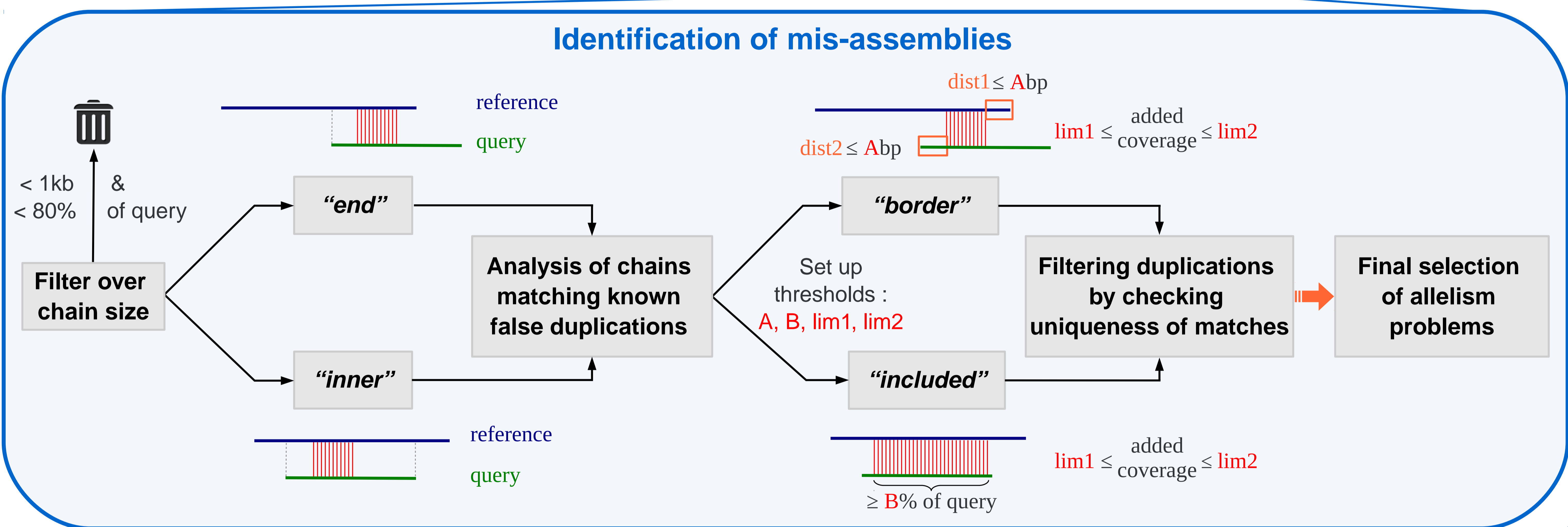
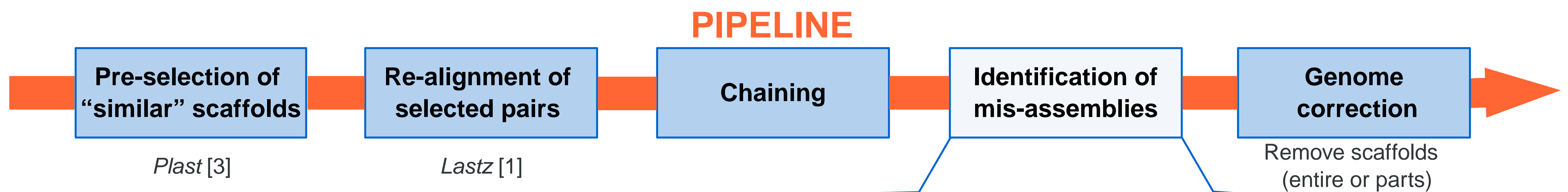
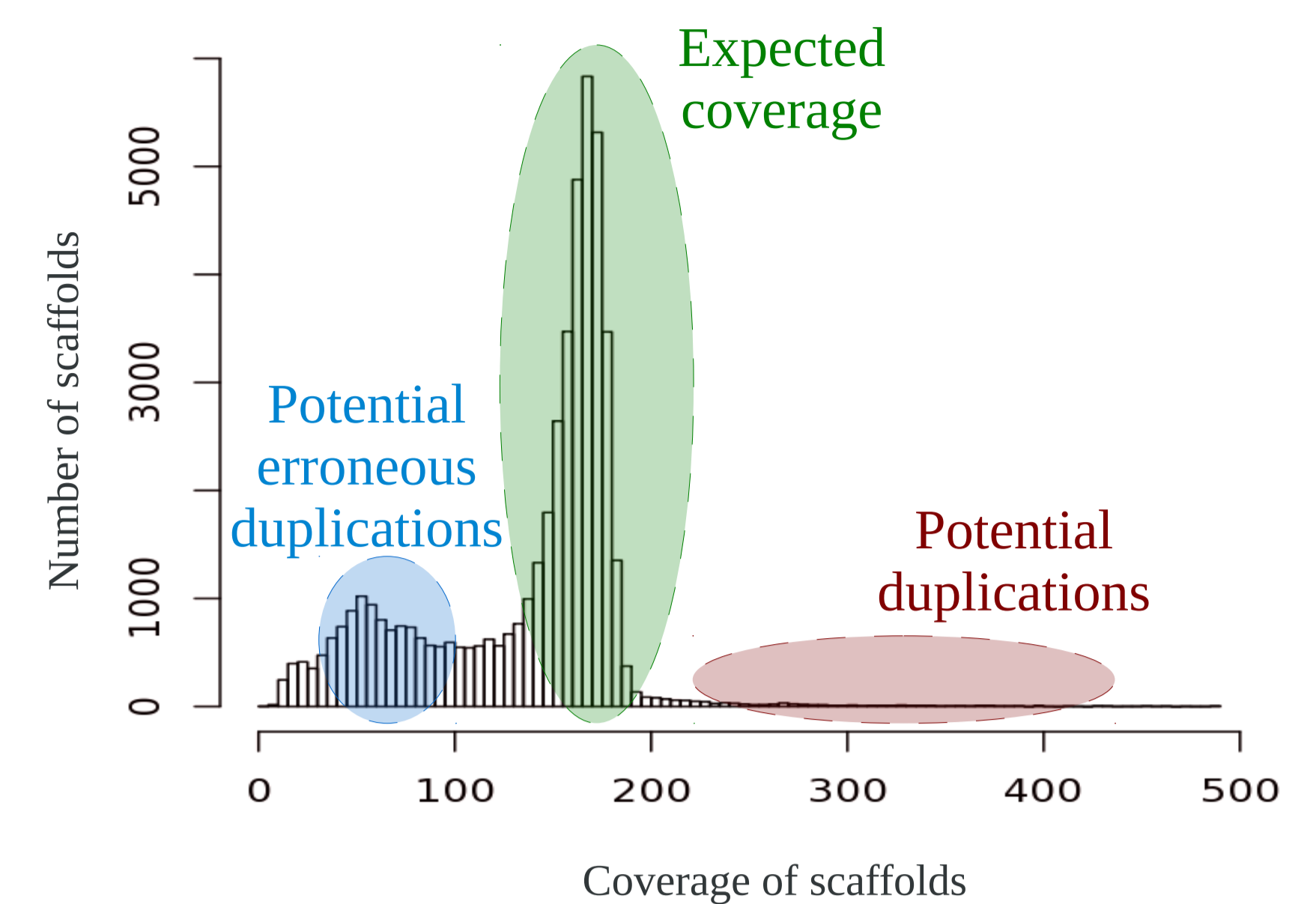
Anais GOUIN<sup>1</sup>, Anthony BRETAUDEAU<sup>2</sup>, Claire LEMAITRE<sup>1</sup> and Fabrice LEGEA<sup>1</sup>

<sup>1</sup>INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes cedex, France

<sup>2</sup>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.

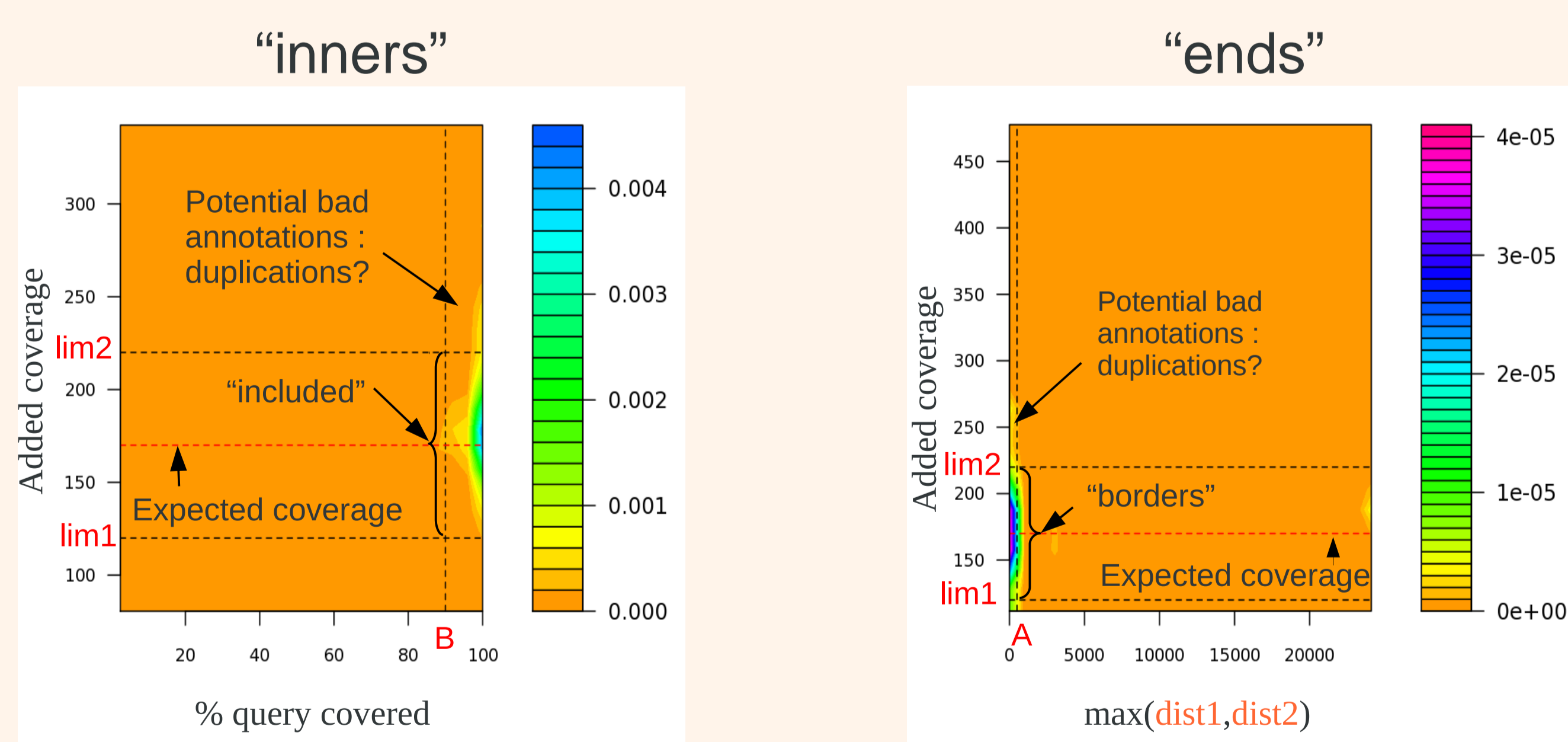


**Application : *Spodoptera frugiperda* genome**



## Set up thresholds

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



→ 114 "included" / 38 "borders"

→ ~80% of known allelic regions with chosen thresholds

## Genome correction

Expected size : ~ 375 Mb

	Initial assembly <i>Allpaths</i>	<i>Platanus</i> [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	15
	Single hit	1497	2047
	Multi hit	782	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