



HAL
open science

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

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

► **To cite this version:**

Anaïs Gouin, Anthony Bretaudeau, Emmanuelle d'Alençon, Claire Lemaitre, Fabrice Legeai. Improvement of the assembly of heterozygous genomes of non-model organisms. *Genome Informatics*, Oct 2015, Cold Spring Harbor Laboratory, United States. 2015. hal-01231793

HAL Id: hal-01231793

<https://inria.hal.science/hal-01231793v1>

Submitted on 20 Nov 2015

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.

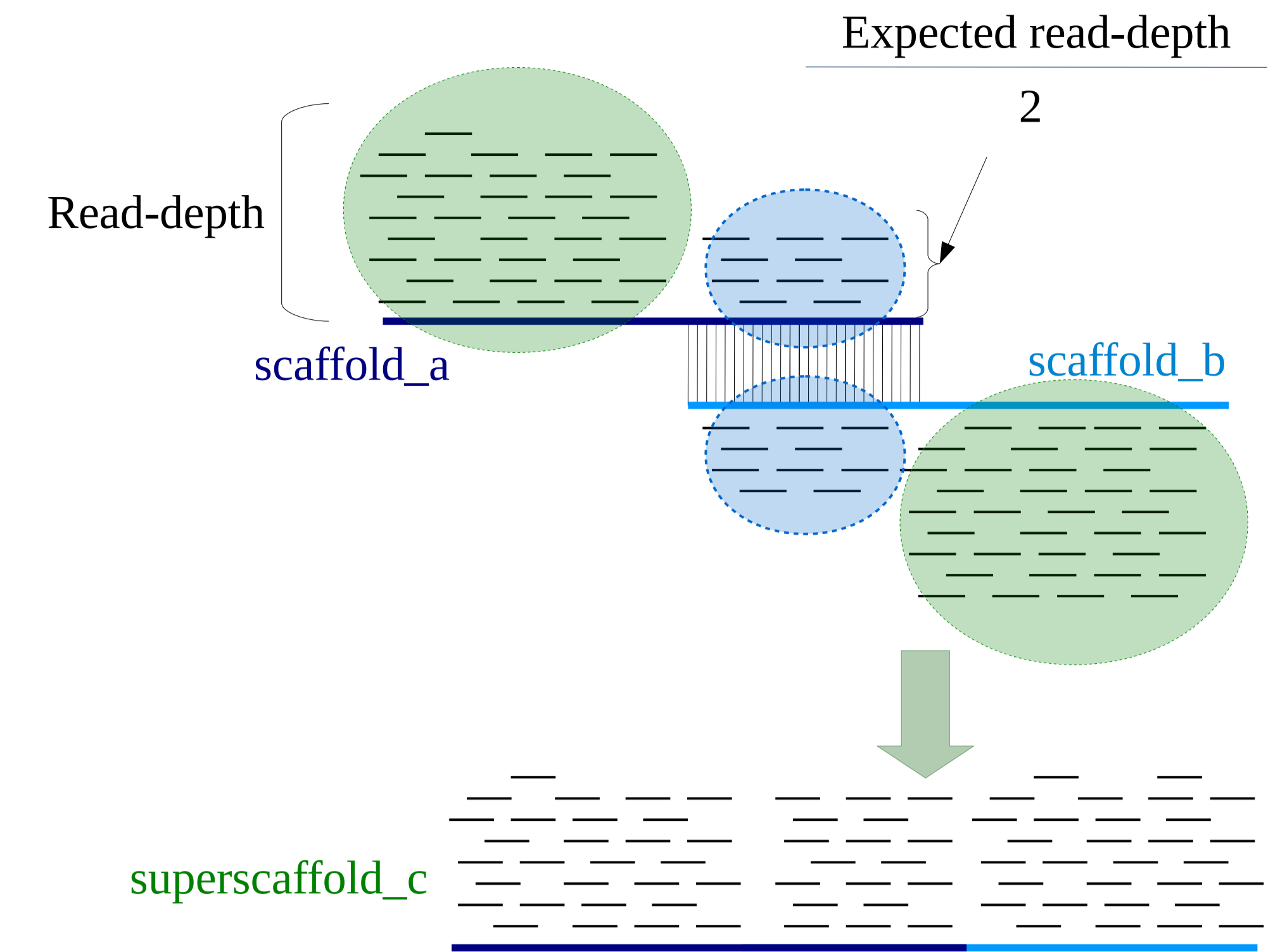
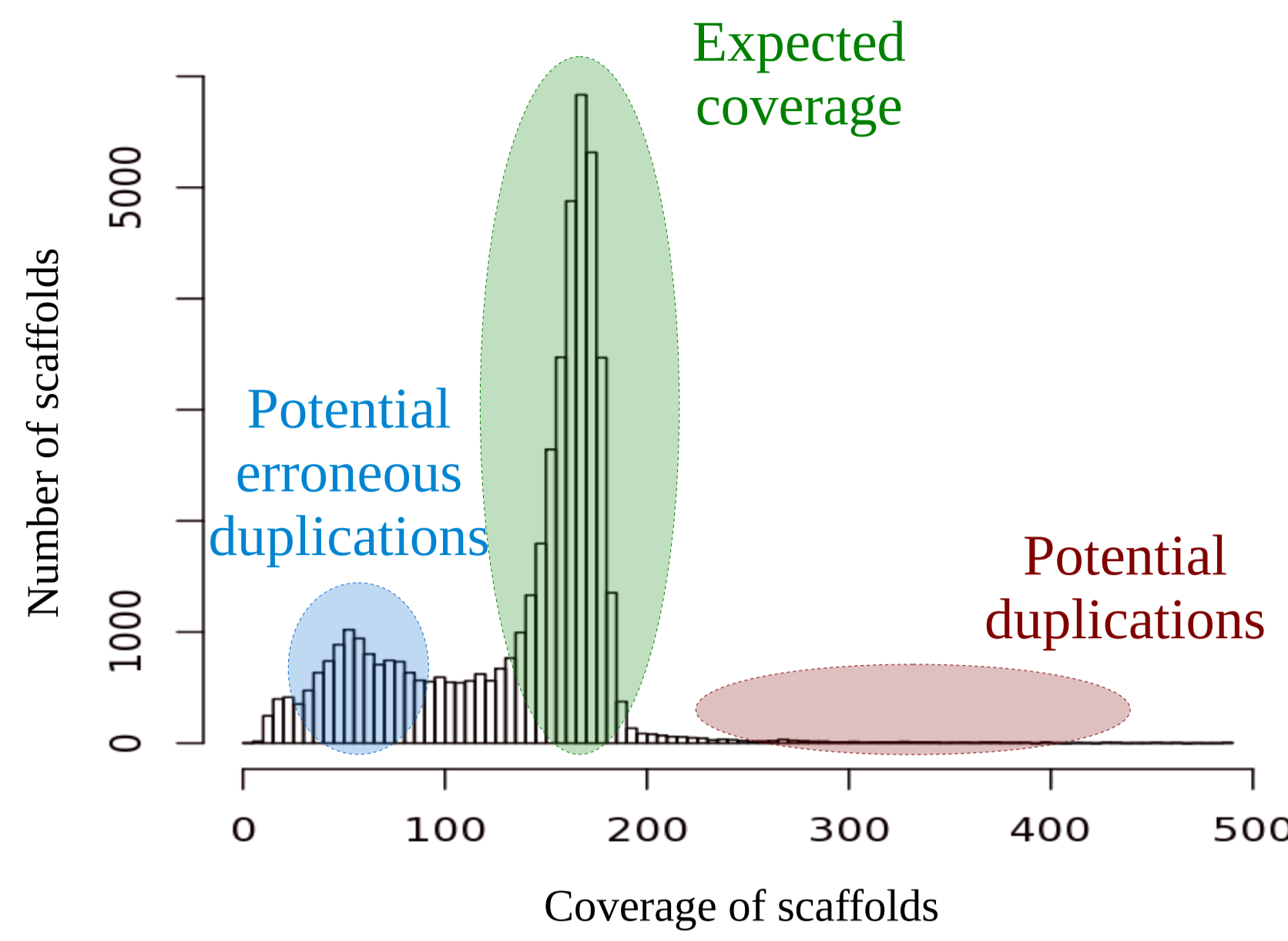
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

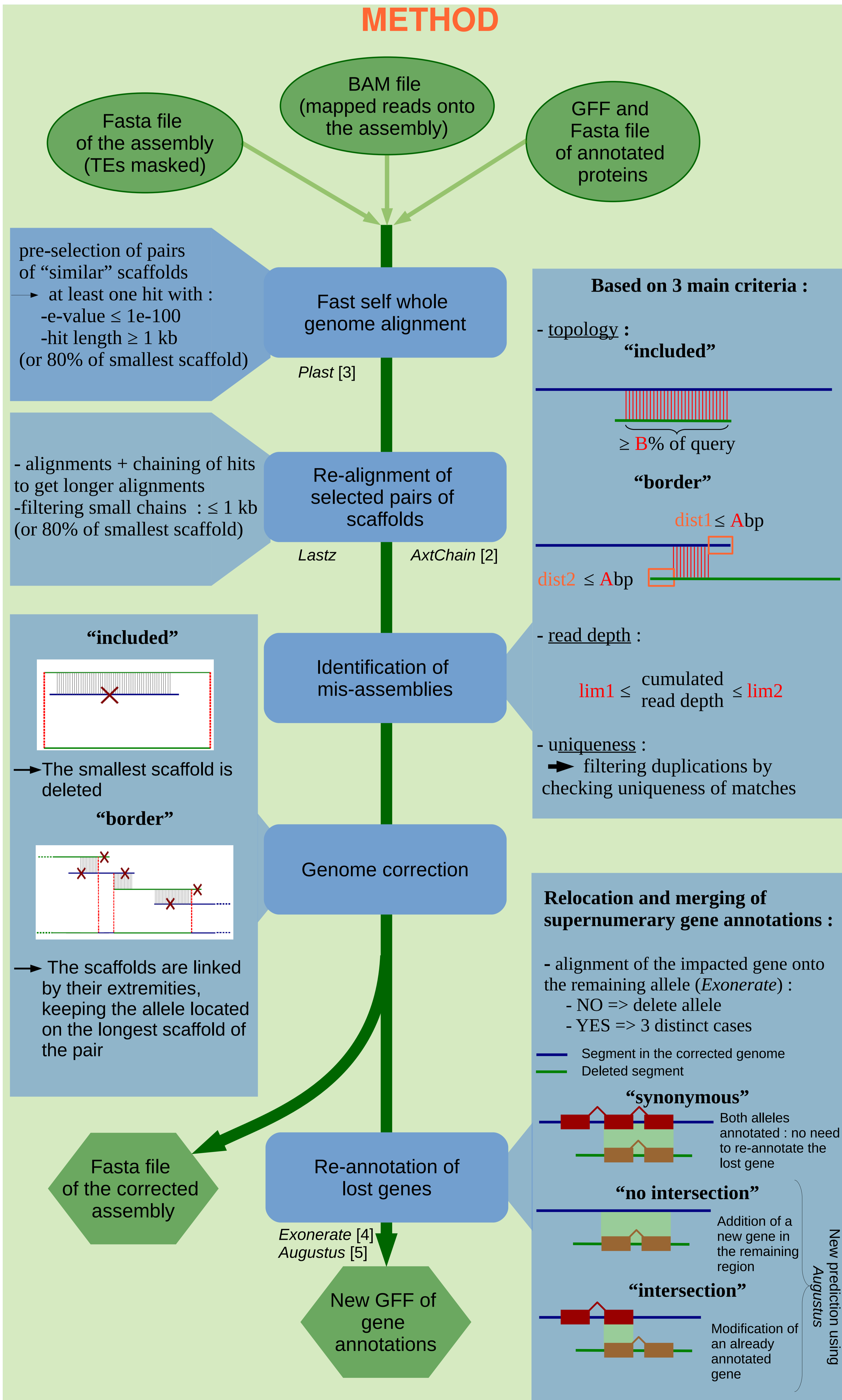
Anaïs GOUIN¹, Anthony BRETAUDEAU², Emmanuelle d'Alençon³, Claire LEMAITRE¹ and Fabrice LEGEAI²
¹INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes cedex, France
²INRA, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Domaine de la Motte – 35653 Le Rheu
³INRA DGIMI, université 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

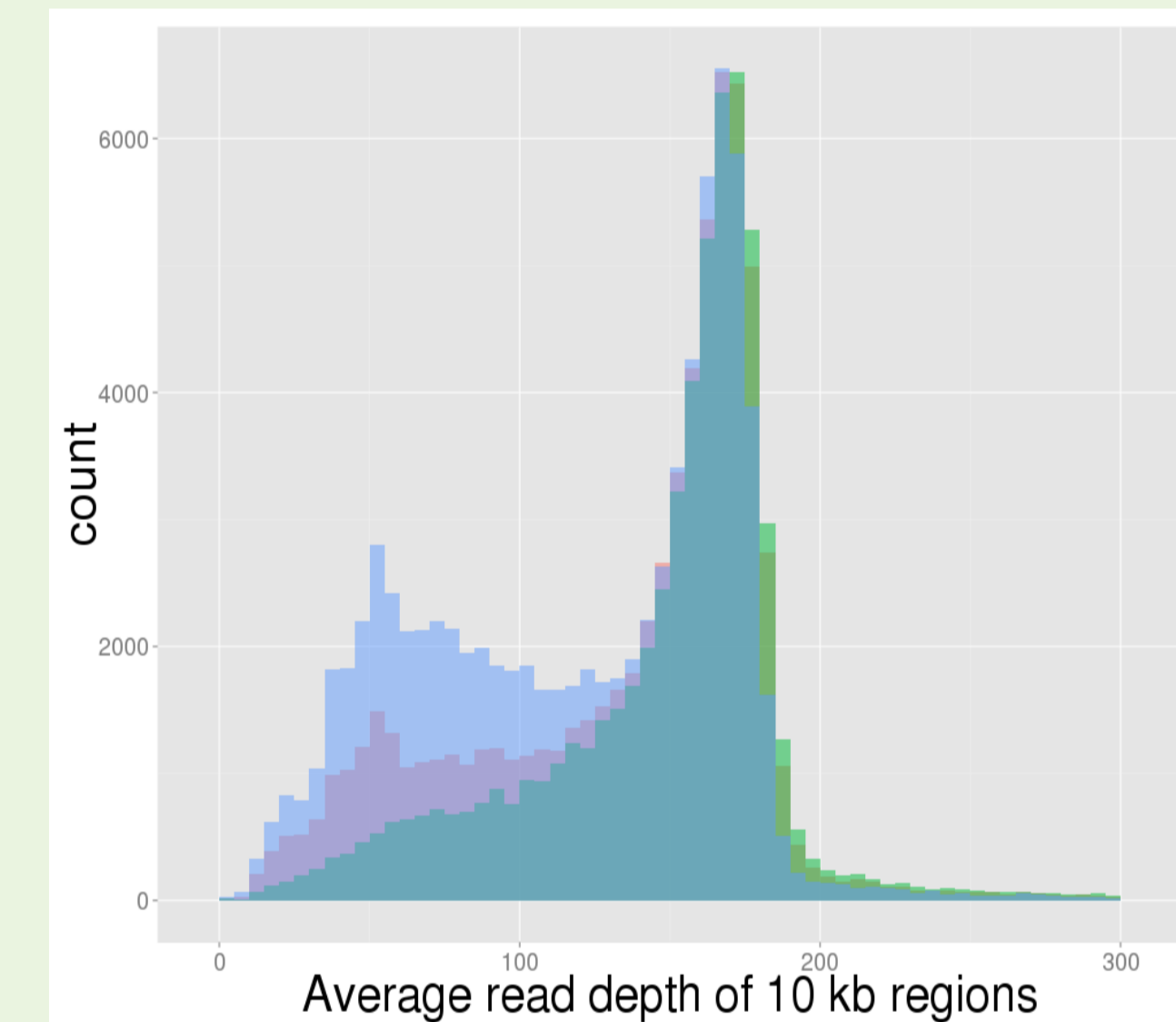
Spodoptera frugiperda genome

Genome correction

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

	Initial assembly <i>Allpaths</i>	Corrected assembly	<i>Haplomerger</i>
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	<i>Haplomerger</i>
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