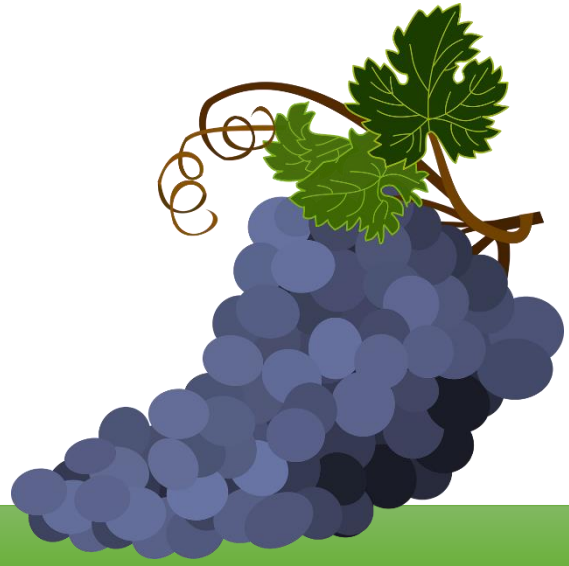
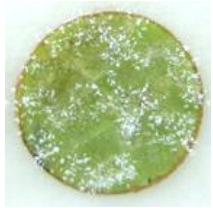
Assemblage diploïde du génome de Muscadinia rotundifolia cv Regale

Guillaume Barnabé, Amandine Velt, Camille Rustenholz, Didier Merdinoglu

INRA
SCIENCE & IMPACT



Contexte scientifique

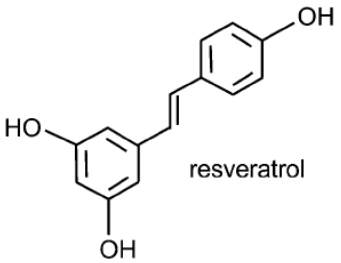


V. vinifera

X

Vitis sp.

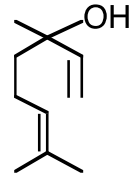
Résistance

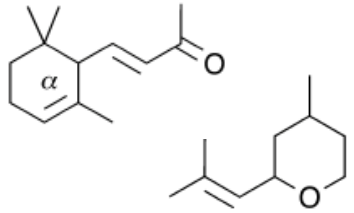


Défenses

Arômes

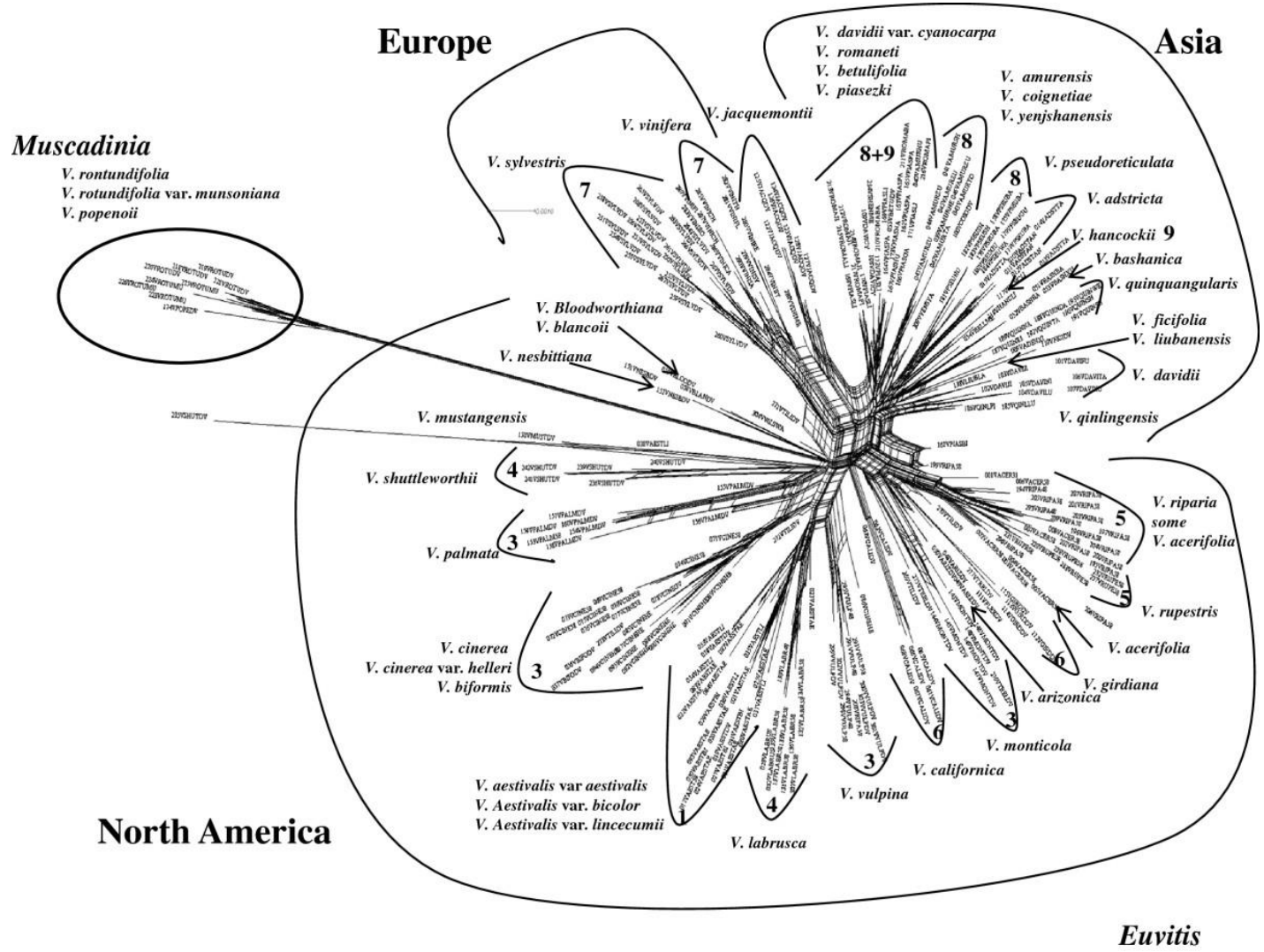
Qualité





INRA
SCIENCE & IMPACT

Contexte scientifique



Wan et al., BMC Evol Biol, 2013

01

Séquençage du génome Regale

Technologie PacBio P6C4

Données de séquençage PacBio du génome Regale

Génotype	Nombre bases séquencées (b)	Profondeur	Longueur maximale (b)	Longueur minimale (b)	Longueur moyenne (b)
Regale	33,810,444,526	67.62	53,524	35	9,741

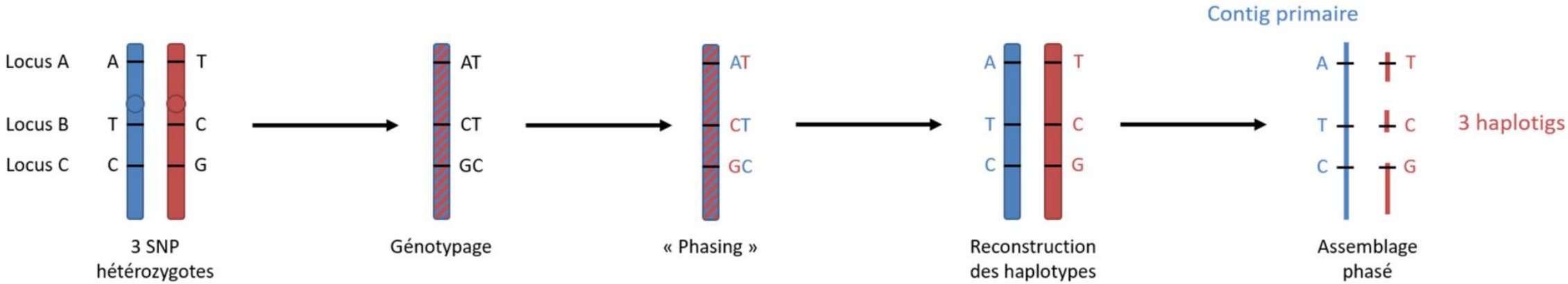
Objectif : générer un assemblage diploïde du génome Regale

INRA SCIENCE & IMPACT

01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

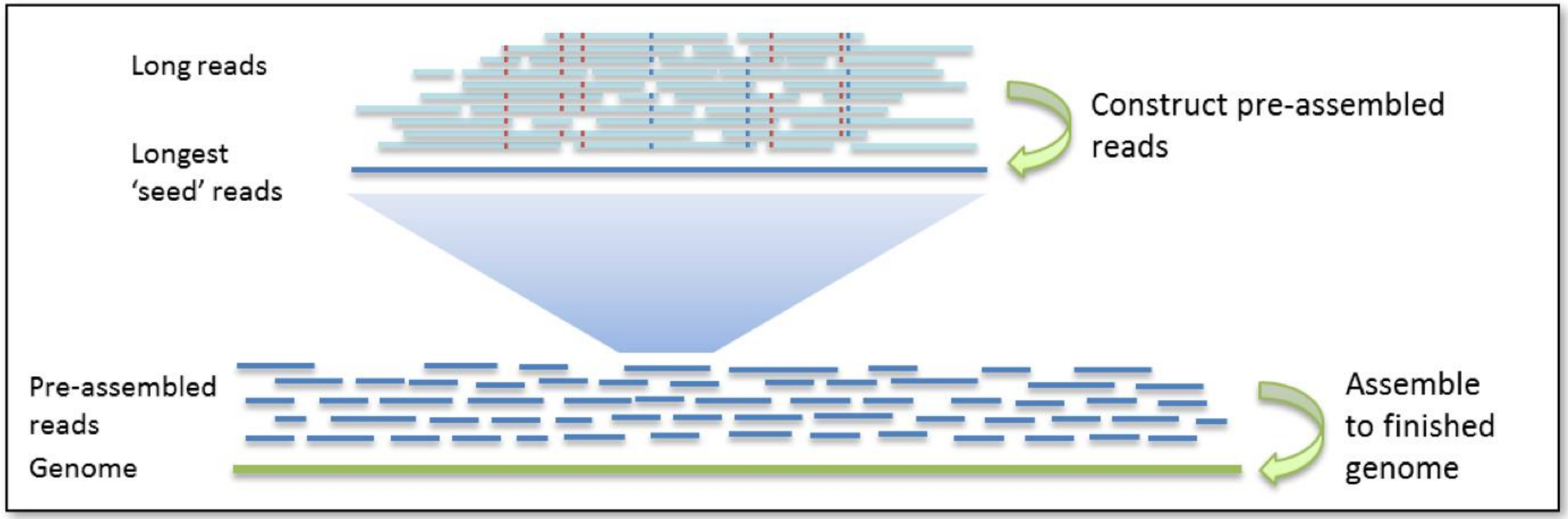
Qu’est ce qu’un assemblage diploïde ?



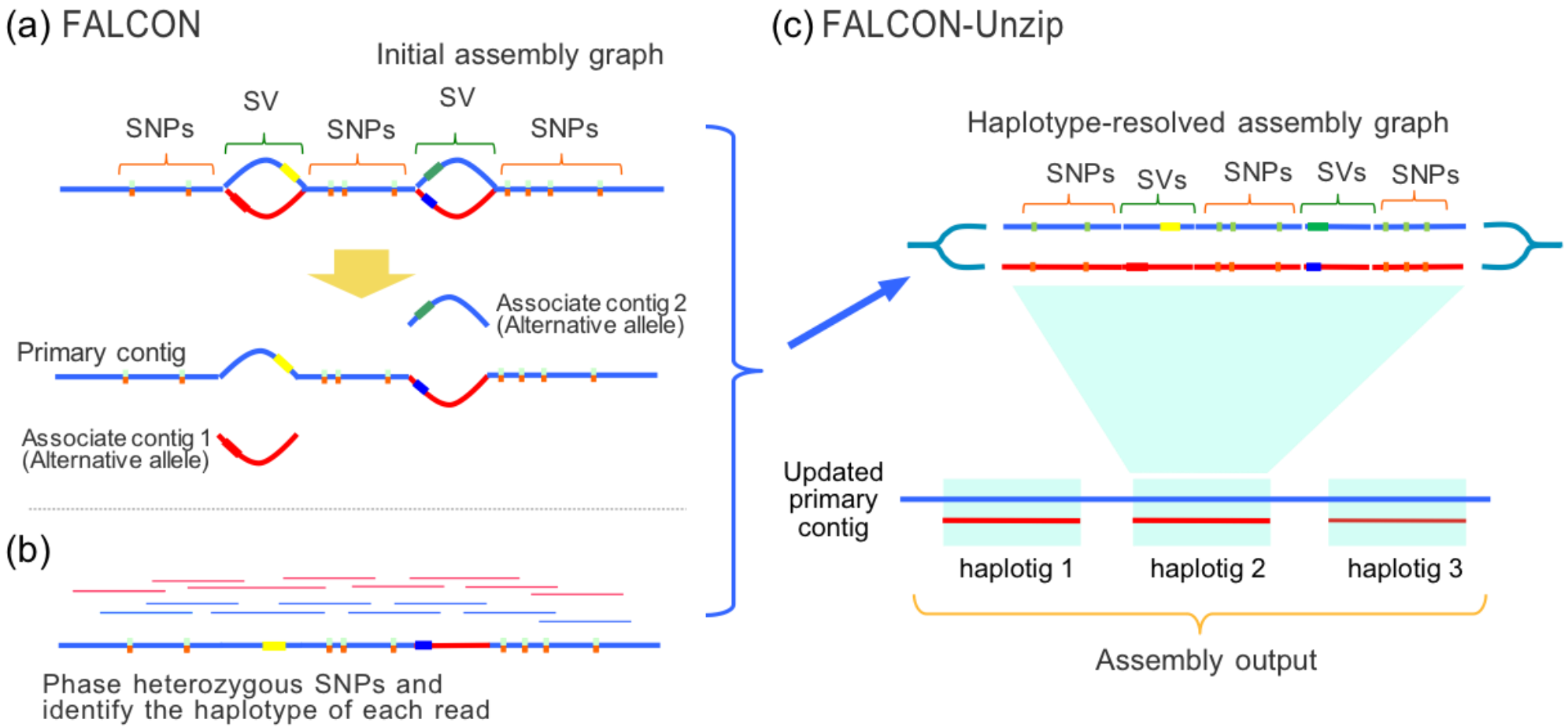
Assemblage diploïde du génome avec FALCON

Reads bruts → Reads pré-assemblés → Génération des contigs

Reads corrigés



Assemblage diploïde du génome avec FALCON



Assemblage diploïde du génome avec FALCON

Phased diploid genome assembly with single-molecule real-time sequencing

Chen-Shan Chin[1,10], Paul Peluso[1,10], Fritz J Sedlazeck[2], Maria Nattestad[3], Gregory T Concepcion[1], Alicia Clum[4], Christopher Dunn[1], Ronan O'Malley[5], Rosa Figueroa-Balderas[6], Abraham Morales-Cruz[6], Grant R Cramer[7], Massimo Delledonne[8], Chongyuan Luo[5], Joseph R Ecker[5], Dario Cantu[6], David R Rank[1] & Michael C Schatz[2,3,9]

While genome assembly projects have been successful in many haploid and inbred species, the assembly of noninbred or rearranged heterozygous genomes remains a major challenge. To address this challenge, we introduce the open-source FALCON and FALCON-Unzip algorithms (https://github.com/PacificBiosciences/FALCON/) to assemble long-read sequencing data into highly accurate, contiguous, and correctly phased diploid genomes. We generate new reference sequences for heterozygous samples including an F1 hybrid of *Arabidopsis thaliana*, the widely cultivated *Vitis vinifera* cv. Cabernet Sauvignon, and the coral fungus *Clavicorona pyxidata*, samples that have challenged short-read assembly approaches. The FALCON-based assemblies are substantially more contiguous and complete than alternate short- or long-read approaches. The phased diploid assembly enabled the study of haplotype structure and heterozygosities between homologous chromosomes, including the identification of widespread heterozygous structural variation within coding sequences.

De novo genome assembly is a fundamental pursuit in genome research[1–3] that has led to the creation of high-quality reference genomes for many haploid or highly inbred species and has promoted gene discovery, comparative genomics, and other studies[4–6]. However, currently available genome assemblies rarely capture the heterozygosity present within a diploid or polyploid species[7]. Most assemblers output a mosaic genome sequence that arbitrarily alternates between parental alleles[8]. projects shift toward more heterogeneous samples such as outbred, wild-type diploid, and polyploid nonmodel organisms, as well as to highly rearranged disease samples including samples from human cancers.

While the problem of assembling diploid and polymorphic genomes is not new[12,13], it lacks a universal and scalable solution. Computational methods for diploid assembly tend to generate short contigs averaging from just a few hundred bases to several kilobases[12,14,15]. Approaches such as sequencing both parents and offspring (i.e., trios)[16], haploid sex cells[17], clonal fosmid[18], and synthetic long reads[19,20] are labor intensive and costly, and they often produce assemblies with limited contiguity. Long-range scaffolding technologies such as optical mapping and chromatin assays are often inapplicable to heterozygous short-read assemblies, as they demand well-assembled contig sequences (minimal contig N50 size of 50 kbp to 100 kbp) and can leave unresolved regions (N characters) inside the scaffolds.

Single-molecule real-time (SMRT) Sequencing is commonly used to finish bacterial genomes and provide high-contiguity assemblies for mammalian-scale genomes[21,22]. The long reads (currently ~10 kbp, on average, with some approaching 100 kbp) can span many repetitive elements and help resolve complicated diploid genomes. Nonetheless, existing assemblers do not take advantage of the long reads to resolve haplotypes. In this paper, we present FALCON, a diploid-aware long-read assembler, and FALCON-Unzip, an associated haplotype-resolving tool, to assemble haplotype contigs or 'haplotigs' that represent the diploid genome with correctly phased homologous chromosomes (**Fig. 1**).

INRA SCIENCE & IMPACT

01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

03 **Contrôle qualité de l'assemblage Regale**
Alignements de reads Illumina, BUSCO, …

Statistiques sur l'assemblage Regale

	Nombre de contigs	Longueur moyenne (b)	N50	L50 (b)	Somme des longueur de tous les contigs (b)
Contigs primaires	1,572	274,596	160	688,163	431,666,254
Haplotigs	2,323	109,914	344	206,731	255,331,801

INRA SCIENCE & IMPACT

Contrôle qualité : alignements des reads bruts PacBio et de reads Illumina sur les contigs primaires

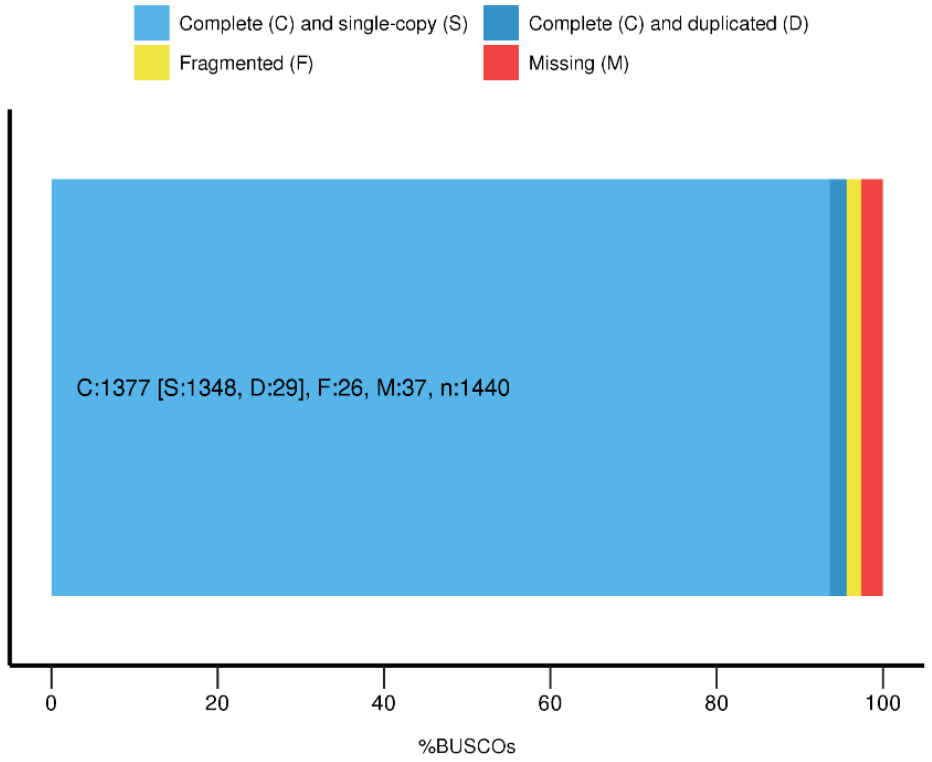
	Nombre total de reads	% de reads alignés
Reads PacBio corrigés	3,470,682	95,5%
Reads Illumina	113,548,726	97,7%

Alignements des reads bruts PacBio avec blasr
Alignements des reads Illumina avec bwa mem

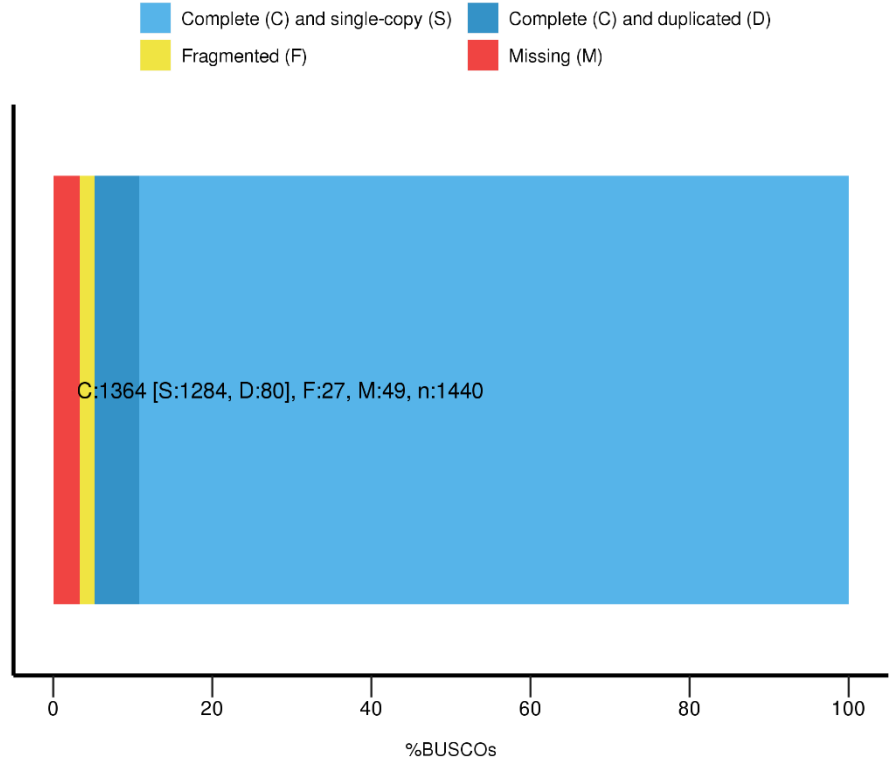
INRA SCIENCE & IMPACT

Contrôle qualité : BUSCO

PN40024



Regale



01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

03 **Contrôle qualité de l'assemblage Regale**
Alignements de reads Illumina, BUSCO, …

04 **Polishing de l'assemblage Regale**
PacBio-utilities et PILON

~ 139 millions de reads MiSeq
~ 92 millions de reads HiSeq (101 nts)
~ 113 millions de reads GAIIx (76/114 nts)

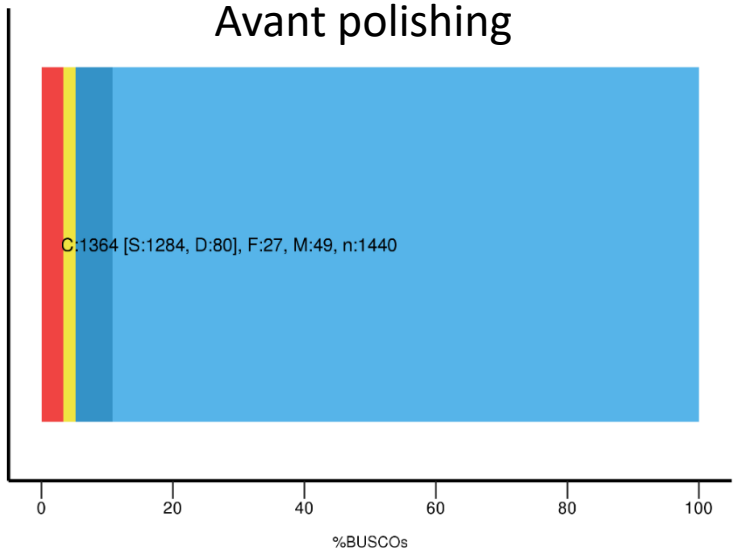
Polishing de l'assemblage avec Pacbio-utilities et PILON

Données Illumina	Nombre total de reads
GAIIx 2*76 nts + 2*114 nts	113,000,000
Hiseq 2*100	92,000,000
Miseq (300-300/400-100/450-100)	139,000,000

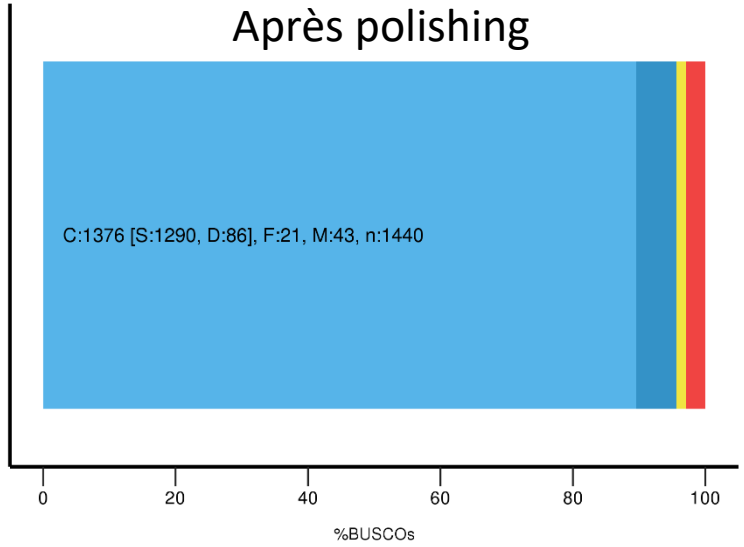
↓

62% des indels corrigés grâce à PacBio-utilities + Pilon

Avant polishing



Après polishing



01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

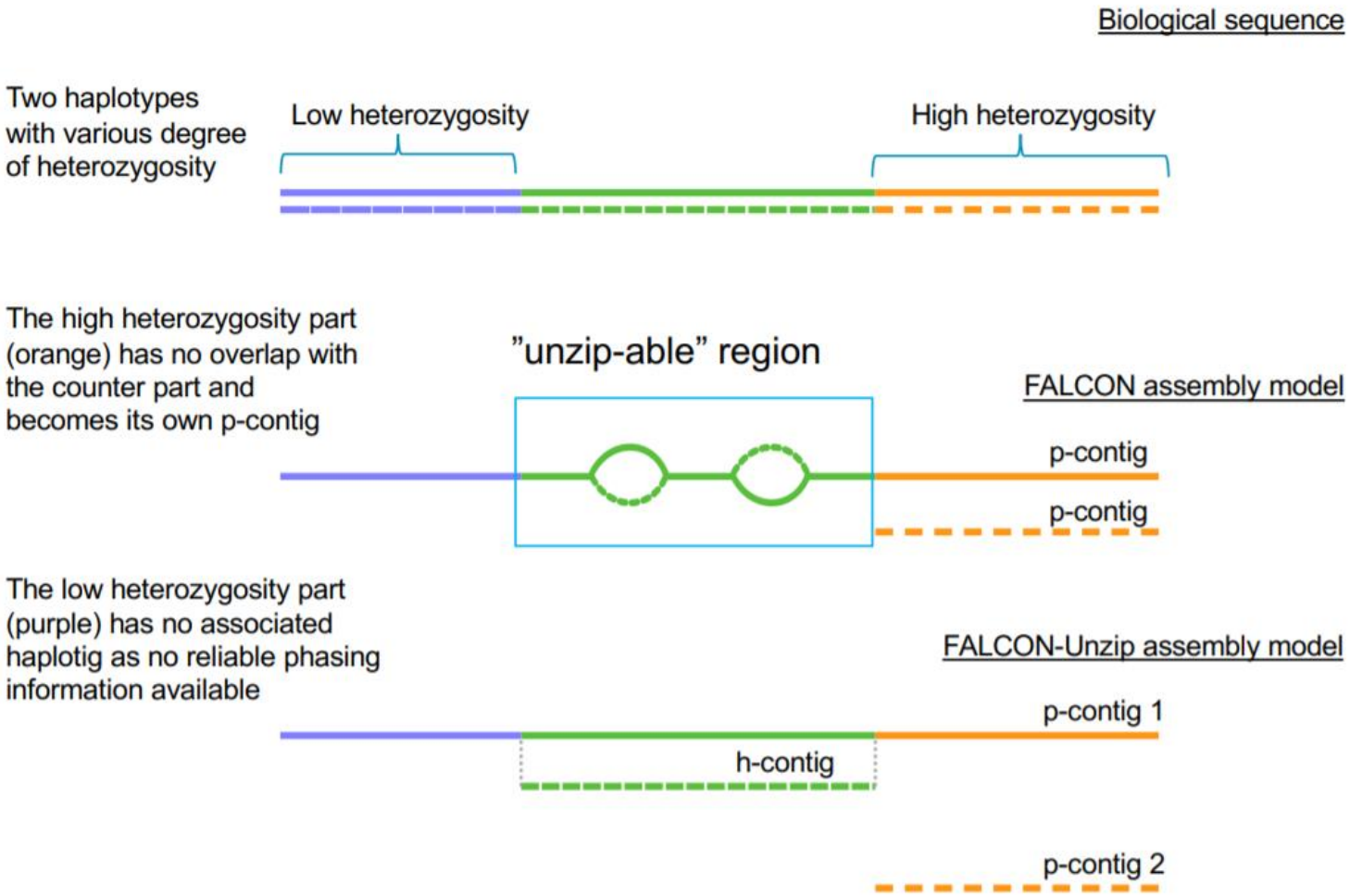
03 **Contrôle qualité de l'assemblage Regale**
Alignements de reads Illumina, BUSCO, …

04 **Polishing de l'assemblage Regale**
PacBio-utilities et PILON

05 **Suppression des contigs homologues**
Scripts maison

INRA
SCIENCE & IMPACT

Suppression des contigs primaires homologues



Suppression des contigs primaires homologues

Annotations des contigs primaires avec Eugene

Blast des CDS contre eux-mêmes, suppression des self-hits et détection des homologues

Détection des contigs primaires contenu entièrement dans un autre

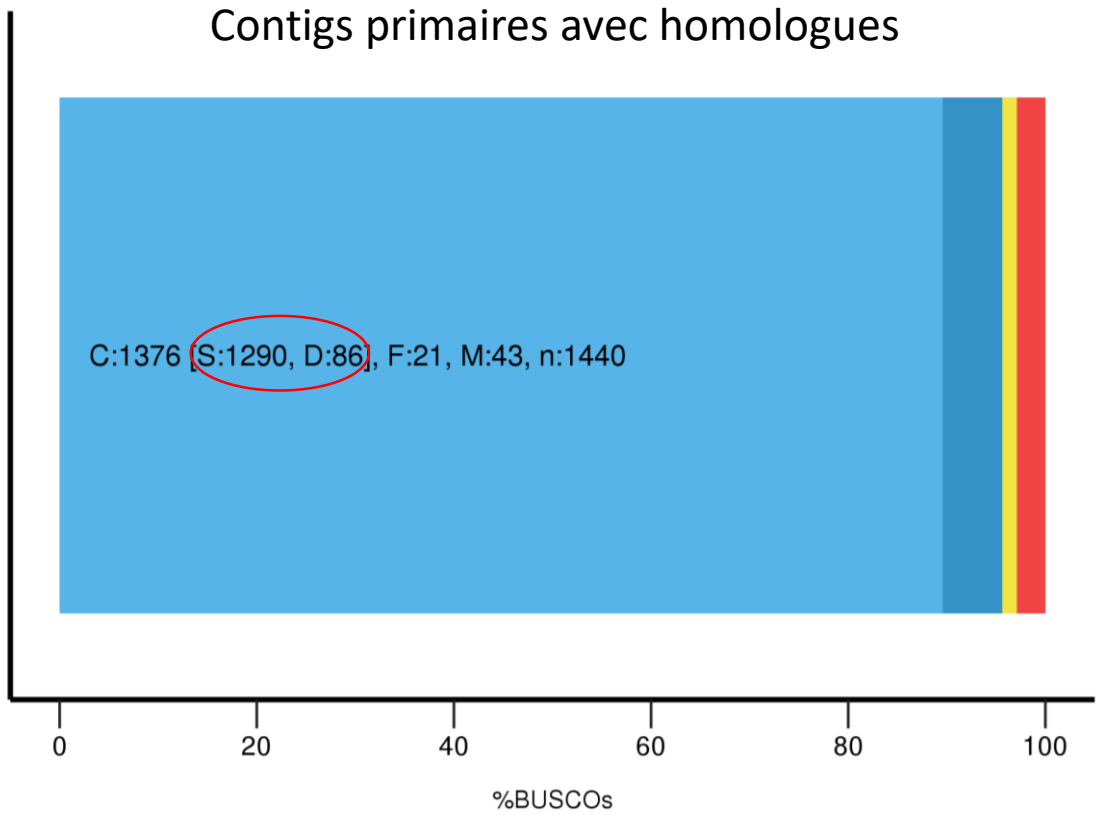
Alignement de reads Illumina sur les contigs primaires et détection de contigs primaires avec une couverture réduite = homologues

INRA SCIENCE & IMPACT

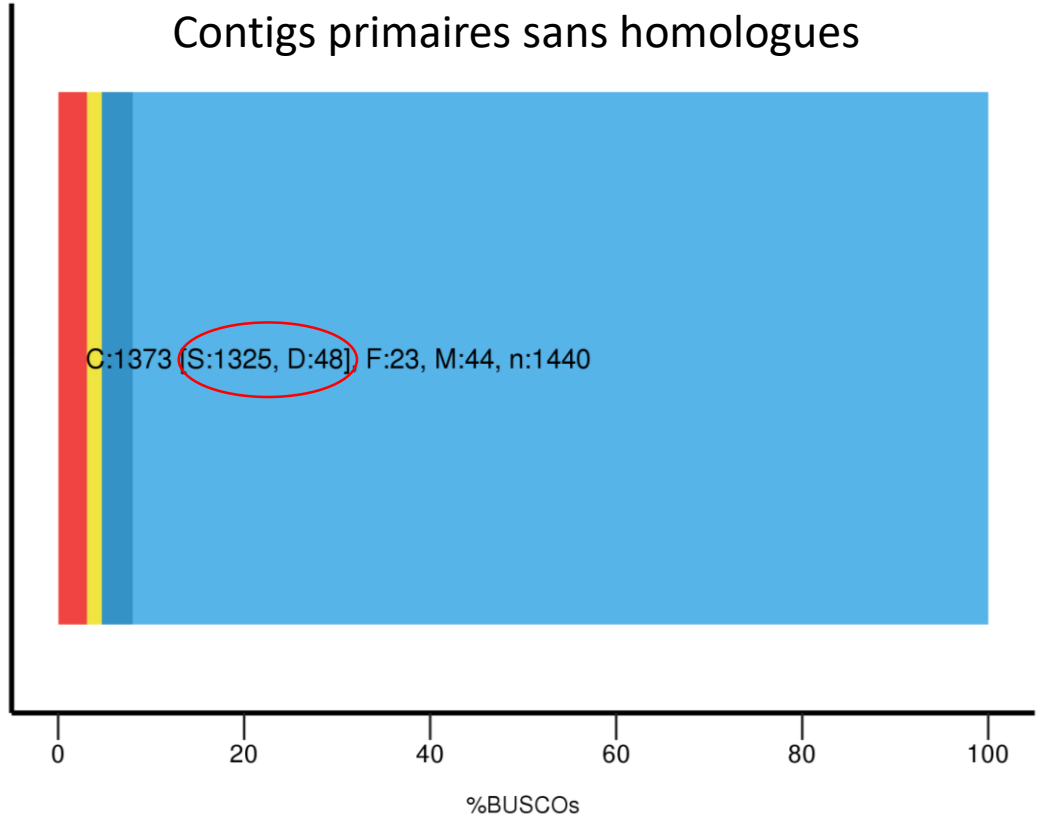
Suppression des contigs primaires homologues

276 contigs primaires homologues trouvés

Contigs primaires avec homologues



Contigs primaires sans homologues



Complete (C) and single-copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)

INRA SCIENCE & IMPACT

01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

03 **Contrôle qualité de l'assemblage Regale**
Alignements de reads Illumina, BUSCO, …

04 **Polishing de l'assemblage Regale**
PacBio-utilities et PILON

05 **Suppression des contigs homologues**
Scripts maison

05bis **Scaffolding des contigs**
Carte optique (~250/300 scaffolds)

01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

03 **Contrôle qualité de l'assemblage Regale**
Alignements de reads Illumina, BUSCO, …

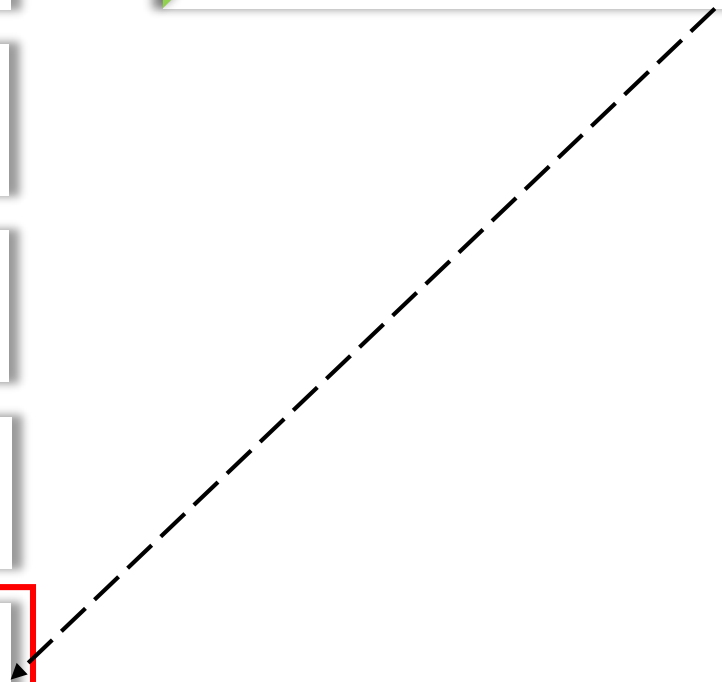
04 **Polishing de l'assemblage Regale**
PacBio-utilities et PILON

05 **Suppression des contigs homologues**
Scripts maison

06 **Annotation automatique des contigs et haplotigs**
EuGene

01bis **Séquençage du transcriptome Regale**
RNA-seq Illumina 2*100b

02bis **Assemblage du transcriptome Regale**
DRAP (De novo RNA-seq Assembly Pipeline)



INRA SCIENCE & IMPACT

Assemblage du transcriptome Regale

Séquençage Illumina 2*100pb
6 conditions expérimentales : fleurs, racines, baies jeunes, baies mures, jeunes feuilles, jeunes feuilles infectées par le mildiou
704,890,885 paires de reads séquencées
Assemblage du transcriptome avec DRAP (De novo RNA-seq Assembly Pipeline)

1) Assemblage condition par condition
2) Création d'un méta-assemblage

Nb transcrits	Longueur moyenne (b)	Longueur max (b)	Longueur min (b)
46 572	1 575	15 602	205

Cabau C, Escudié F, Djari A, Guiguen Y, Bobe J, Klopp C. (2017) **Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ 5:e2988.**

Annotation automatique des contigs primaires et haplotigs avec EuGene-EP

- Données
  - Assemblage corrigé avec PILON
  - Transcriptome DRAP
  - Annotations PN40024
  - Banques protéiques :
    - Swissprot,
    - Uniprot plant,
    - TAIR
  - RepBase
- Annotation avec Eugene-EP (paramètres par défaut)
  - Contigs primaires et haplotigs annotés séparément

INRA SCIENCE & IMPACT

Annotation automatique des contigs primaires et haplotigs avec EuGene-EP

	total number of genes	% genome length	%GC	number of protein coding gene	gene length (b)	intergenic length (b)
primary contigs	34,741	29%	33,7%	33,119	3,799	8,791
haplotigs	22,986	33%	33,8%	21,978	3,787	7,474

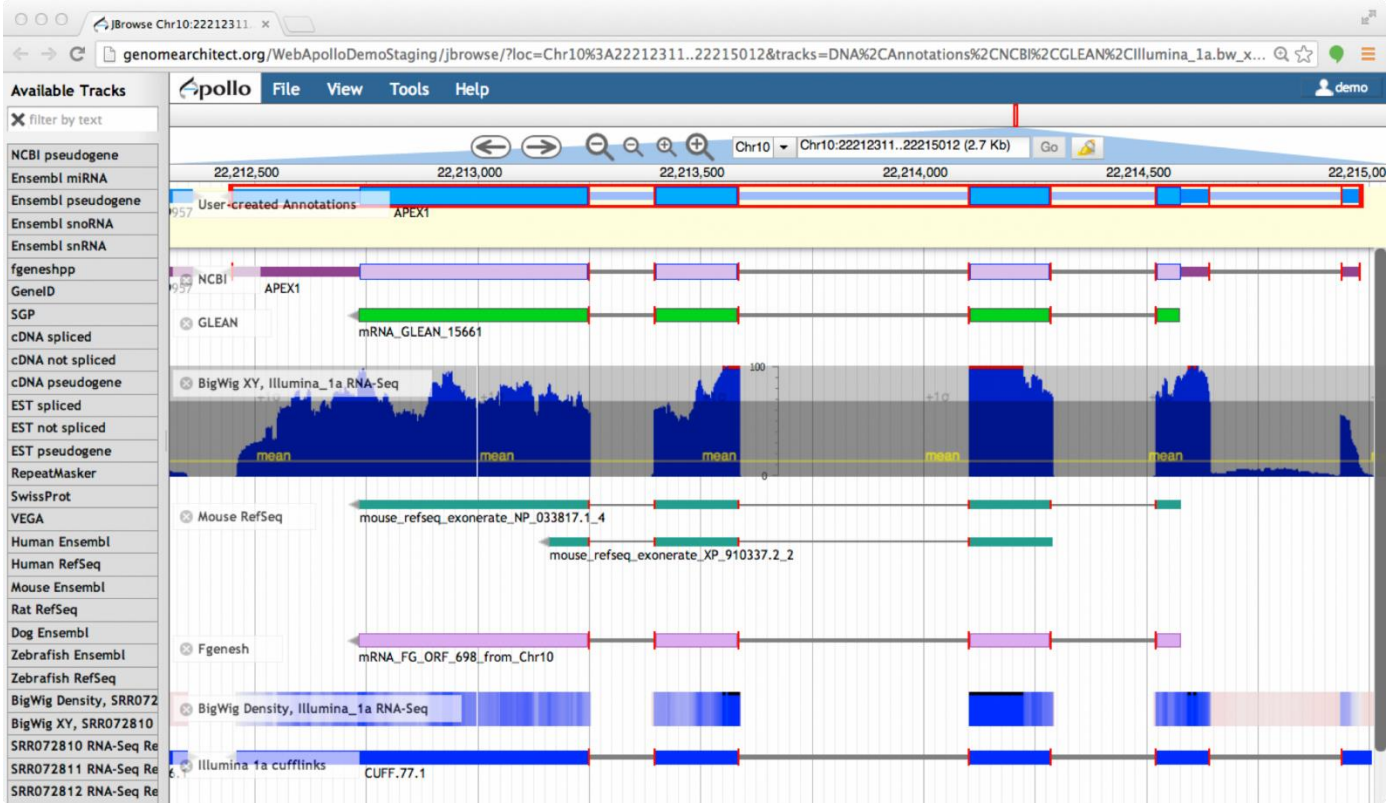
	number of non protein coding genes	ncRNA gene length (b)
primary contigs	1,622	334
haplotigs	1,008	360

INRA SCIENCE & IMPACT

Perspectives : finition de l'annotation du génome par l'URGI

Annotations des éléments répétés

Mise en place d'un WebApollo pour la curation manuelle des gènes et des éléments répétés



01 **Séquençage du génome Regale**
Technologie PacBio

02 **Assemblage diploïde du génome Regale**
FALCON

03 **Contrôle qualité de l'assemblage Regale**
Alignements de reads Illumina, BUSCO, …

04 **Polishing de l'assemblage Regale**
PacBio-utilities et PILON

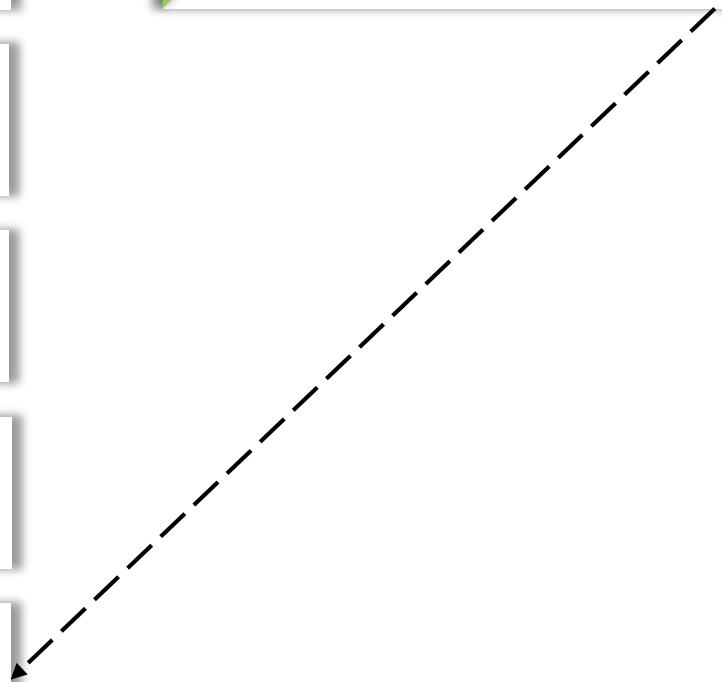
05 **Suppression des contigs homologues**
Scripts maison

06 **Annotation automatique des contigs et haplotigs**
EuGene

07 **Annotation manuelle des gènes R**
Etape qui sera réalisée par l'EPGV

01bis **Séquençage du transcriptome Regale**
RNA-seq Illumina 2*100b

02bis **Assemblage du transcriptome Regale**
DRAP (De novo RNA-seq Assembly Pipeline)



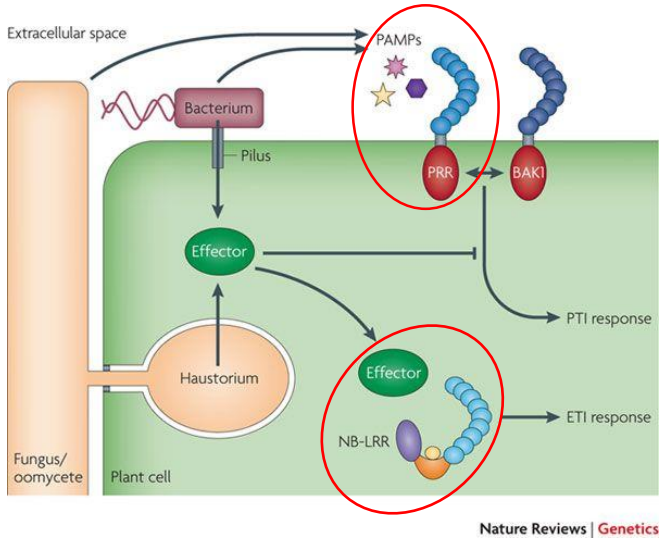
INRA SCIENCE & IMPACT

Annotation manuelle des gènes de résistance (gènes R)

PN40024

	Sub-families	Structure	No. of seq.
Complete genes	CN	CC NBS	5
	CNL	CC NBS LRR	211
	NBS	NBS	14
	NL	NBS LRR	102
	RNL	RPW8 NBS LRR	13
	TIR	TIR	9
	TNL	TIR NBS LRR	94
	TNLI	TIR NBS LIM	2
Pseudogenes	-	NA	379

NB-LRR et PRR (Pattern Recognition Receptors)



Dodds & Rathjen, Nature Reviews Genetics, 2010

Remerciements

L'INRA Grand Est-Colmar
- Guillaume Barnabé
- Gisèle Butterlin
- Vincent Dumas
- Didier Merdinoglu
- Pere Mestre
- Camille Rustenholz

Le CNRGV
- Sandrine Arribat
- Hélène Bergès
- William Marande

L'EPGV
- Patricia Faivre-Rampant
- Marie-Christine Le Paslier

Le LIPM
- Jérôme Gouzy
- Erika Sallet

L'URGI
- Joëlle Anselem

La Fondation Jean Poupelain