GenomeHarvest
diversity, organization and dynamics

FLAGSHIP PROJECT

Genome Harvest

January 2016-October 2019

→ Mobilizing biomathematics/bioinformatics and genomics/genetics

→ to decipher genome organization and dynamics

→ as pathways to crop improvement

Laboratoire Informatique Robotique Microélectronique Montpellier

CBGP

isem
Institut des Sciences de l'Evolution-Montpellier

agap

diade

IPME

GAFL
INRA AVIGNON

Bioversity International

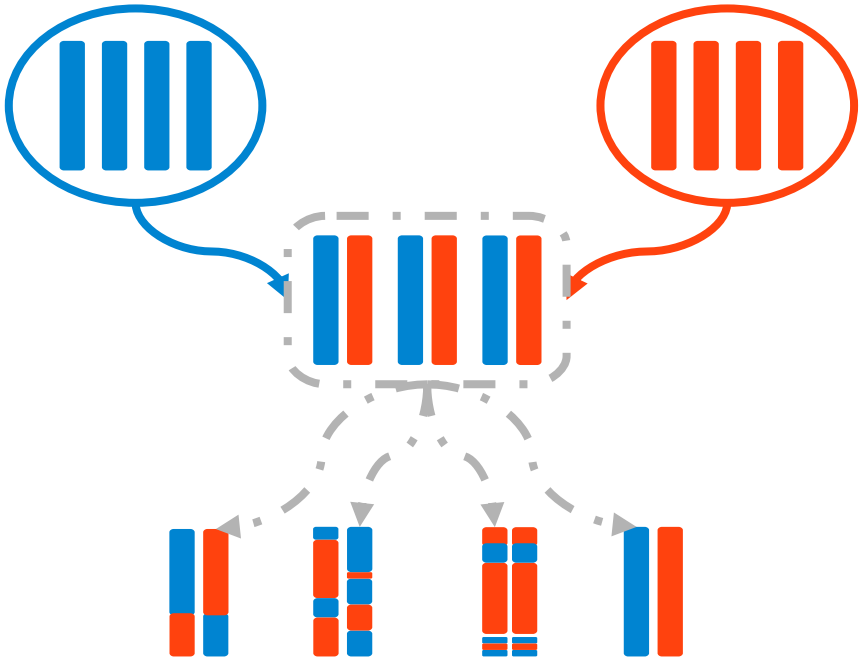
Introduction

- Les évènements d'hybridation entre espèces et sous-espèces sont largement répandus chez les plantes cultivées.

- Brassage génétique → **génomes mosaïques** ayant des origines ancestrales différentes

- **Intérêts ?**

  - Histoire de la domestication des plantes cultivées
  - Origines ancestrales de certains traits phénotypiques.



Introduction

- Trois modèles biologiques:
  - Nombre d'ancêtres différent (3 à 6)
  - 3 niveaux de structure
  - Génomes ancestraux plus ou moins bien connus

Eureka chr2

PTBA00008 chr4

IRIS 313 11030 chr8

Left to right:
Curk thesis, chapter 4 (2014)
Martin et al, in prep
Santos et al, in prep

Methods to characterize plant mosaic genomes

Mosaic complexity

Different biologic systems	Sexual & vegetative propagation = few meiosis Heterozygous A few known ancestors	Sexual & vegetative propagation = several meiosis ? Heterozygous Several ancestors, some unknown?	Sexual propagation = many meiosis Homozygous (autogamy) Several ancestors, some unknown

ARCAD RNAseq 26 accessions

Different types of data

Projet France Genomique DynaMo

→ *2-4 new de novo references*

→ *130-160 resequenced accessions*

→ *GBS 100-2000 progeny accessions*

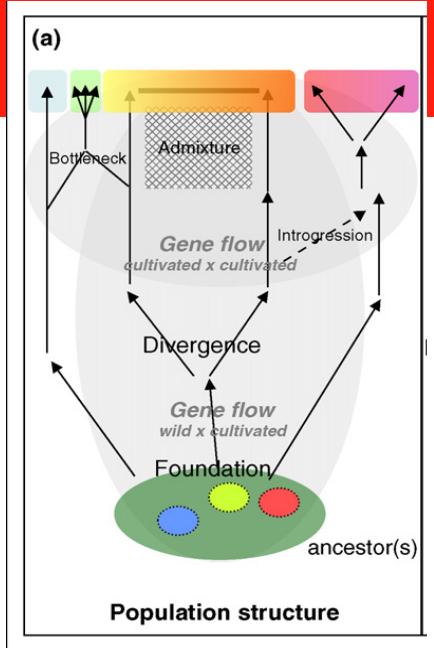
IRIGIN Project
International RIce Genomics INitiative

DATA NOTE
The 3,000 rice genomes project
Global Rice Science Partnership (GRiSP)

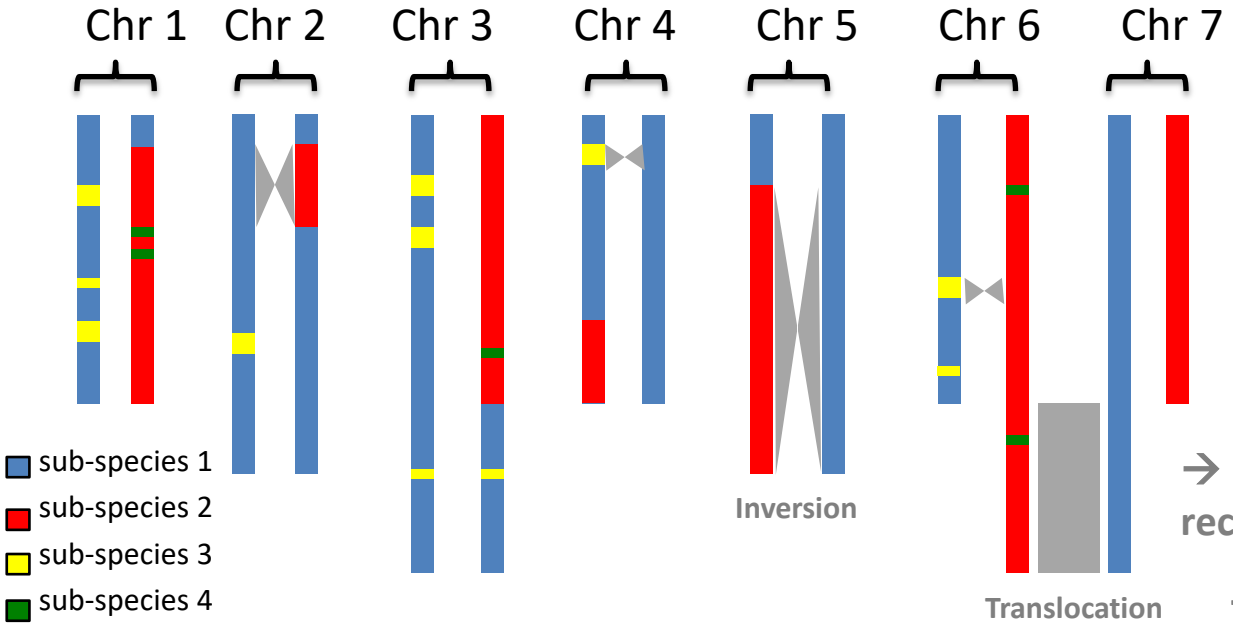
Focus

→ Focus: questions liées aux **fréquents événements d'hybridations inter(sub)spécifiques** au cours de l'histoire des plantes cultivées

frequent during the history of the cultivated plants (also animals, human)



→ **Mosaic genomes ,** hybrids between (sub)species



Large structural variations

→ **Impact on gene expression and thus phenotypes**

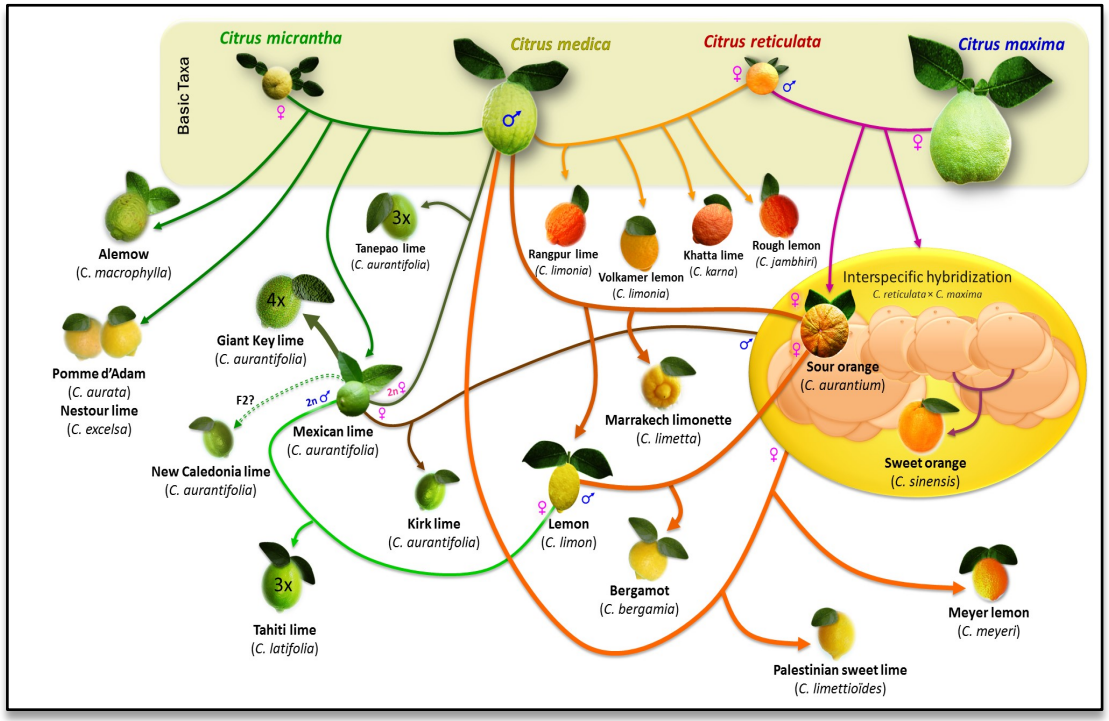
→ Impact on chromosome recombination and transmission

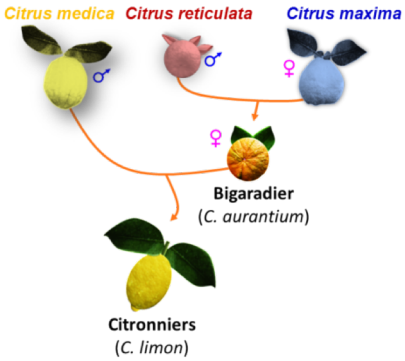
→ Impact on gene and thus character transmission

→ Impact on genetic analysis (QTL, GWAS, ...)

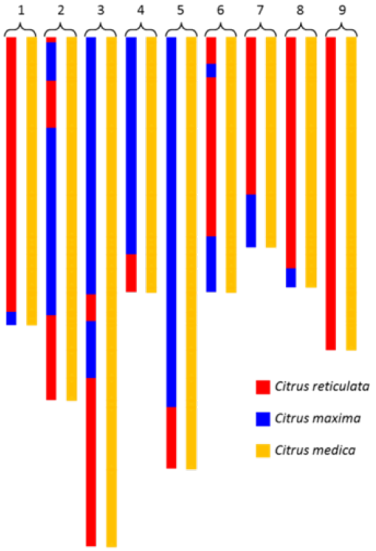
→ Impact on fertility

Context agrumes





Structure phylogénomique du citronnier Eureka



- Evolution par hybridation à partir 4 taxa ancestraux principaux

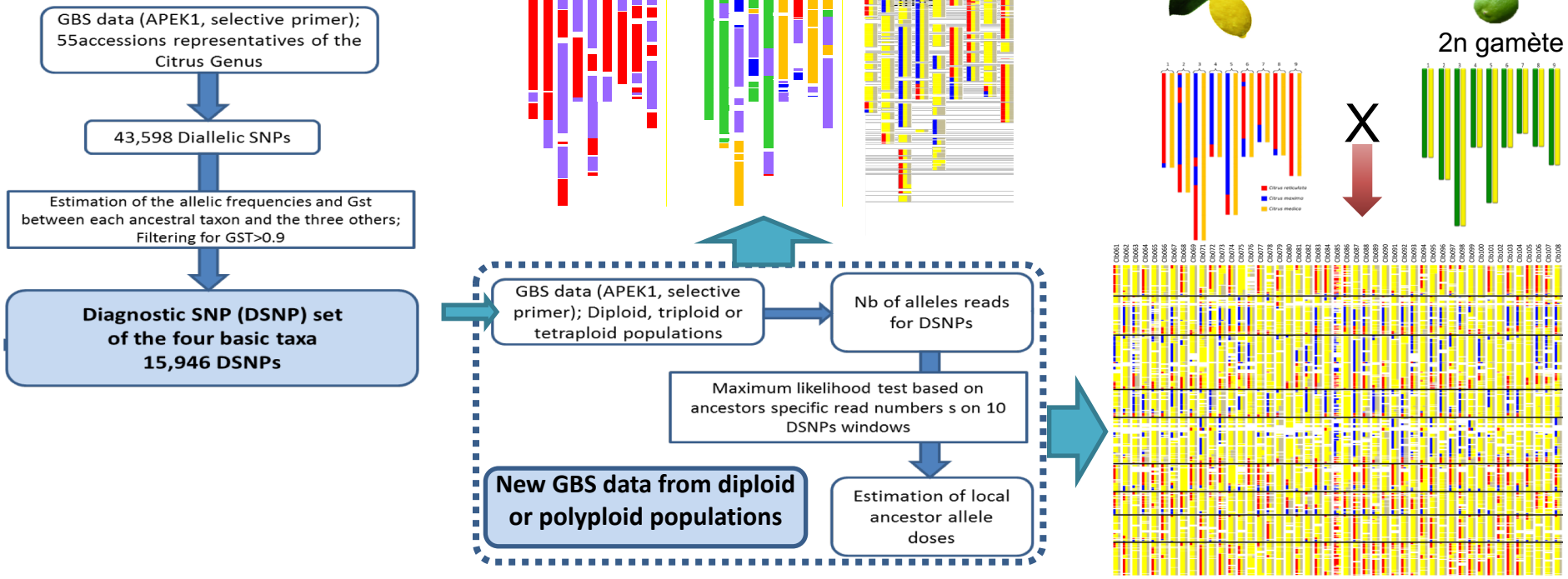
- Hybridation suivie de peu d'évènements de méioses interspécifiques → structure mosaïque assez simple

- Identification de SNP diagnostiques des taxons ancestraux connus et bien différenciés

Décryptage mosaïques agrumes

Franc Curk/P. Ollitrault/…

TraceAncestor :pipeline bioinformatique pour l'analyse des structures en mosaïque du génome de grandes populations d'hybrides diploïdes ou polyploïde à partir de données GBS

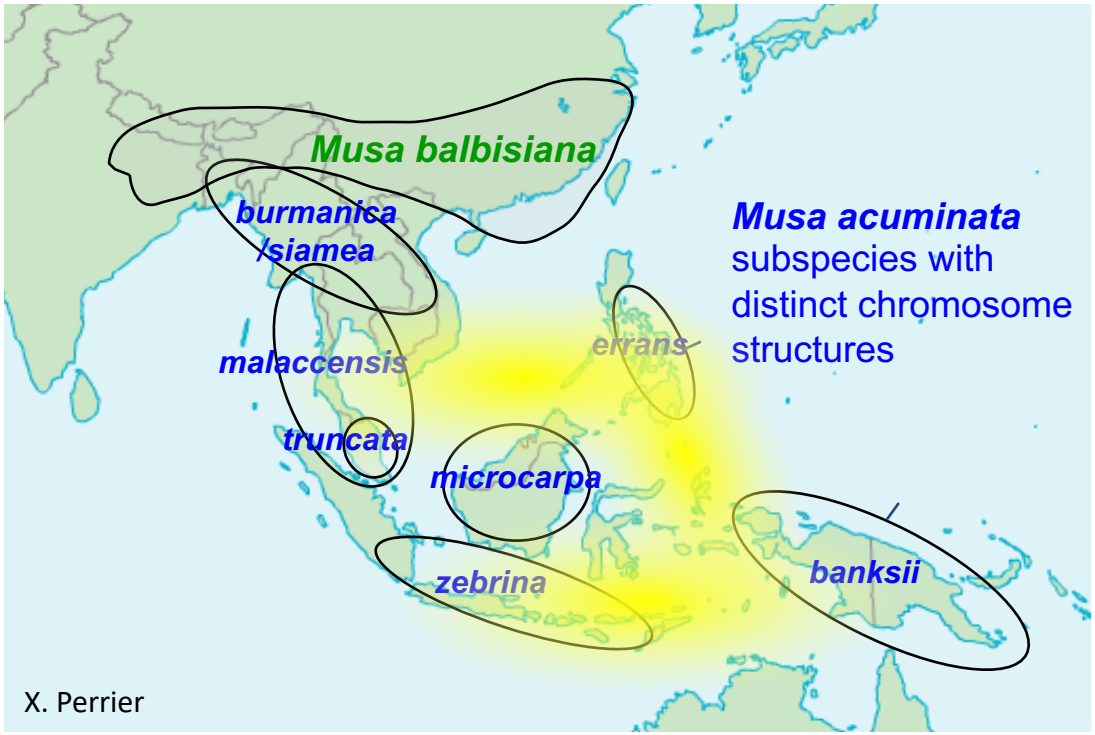


INRA SCIENCE & IMPACT

cirad LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT

IRD Institut de Recherche pour le Développement FRANCE

Context Bananier



Domestication involved:

- **hybridization** between species and subspecies made possible by **human migration**

- selection of **diploid and triploid, seedless, parthenocapic hybrids** by early farmers

- Hybridation suivie de peu d'évènements de méioses interspécifiques → structure mosaïque assez simple

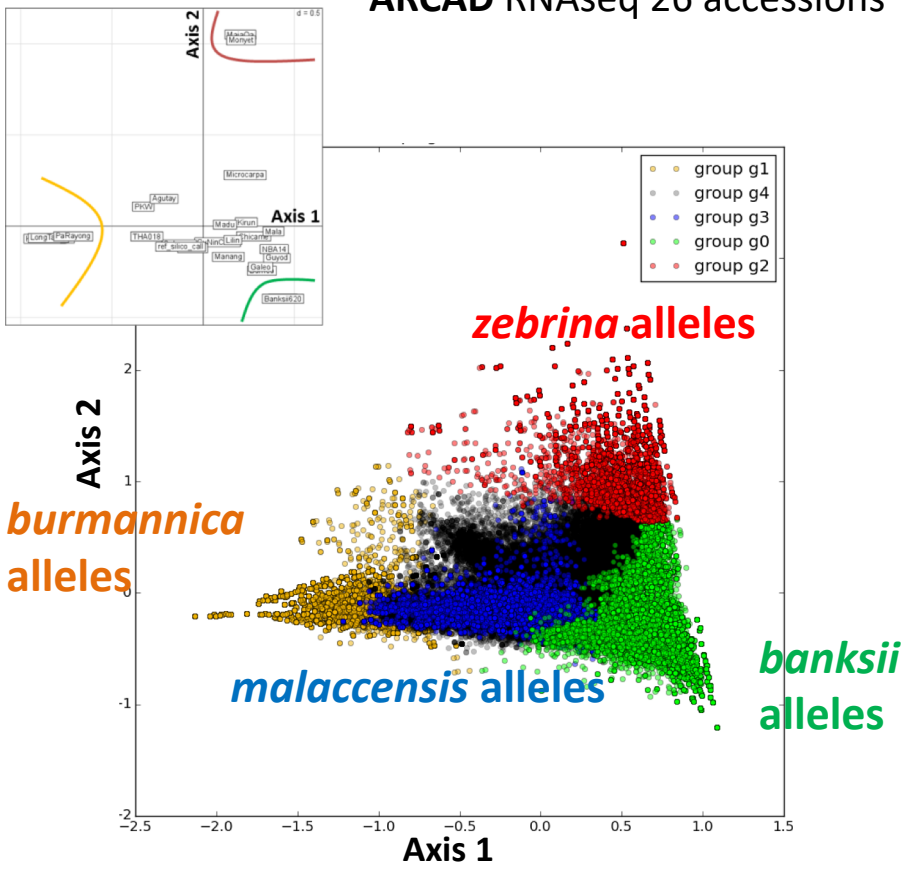
- Identification de SNP diagnostiques des taxons ancestraux, taxon ancestraux potentiellement déjà introgressés et pas tous connus

Décryptage mosaïques bananiers

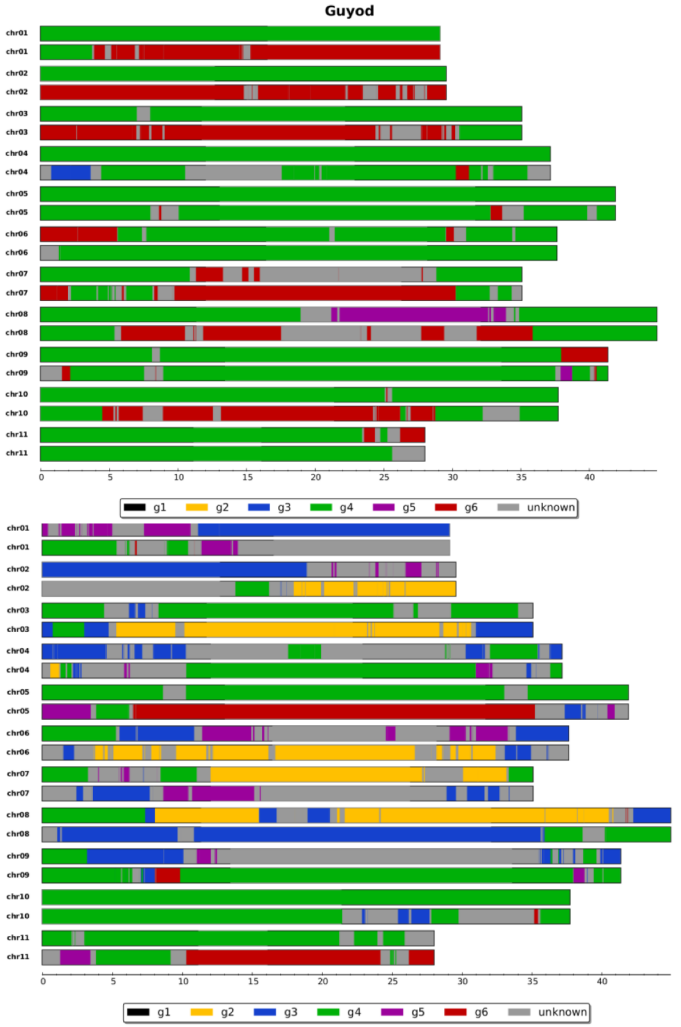
G. Martin/N. Yahiaoui/…

Factorial analysis and K-mean clustering

ARCAD RNAseq 26 accessions



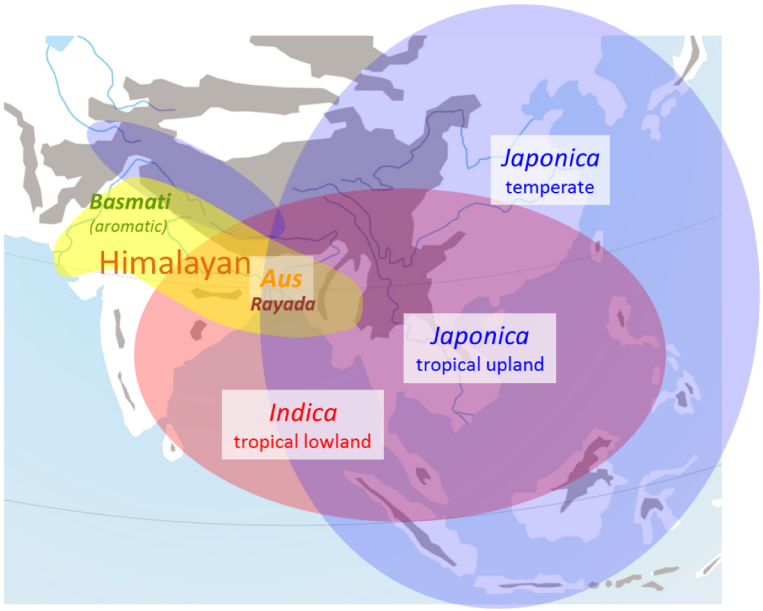
Chromosome painting

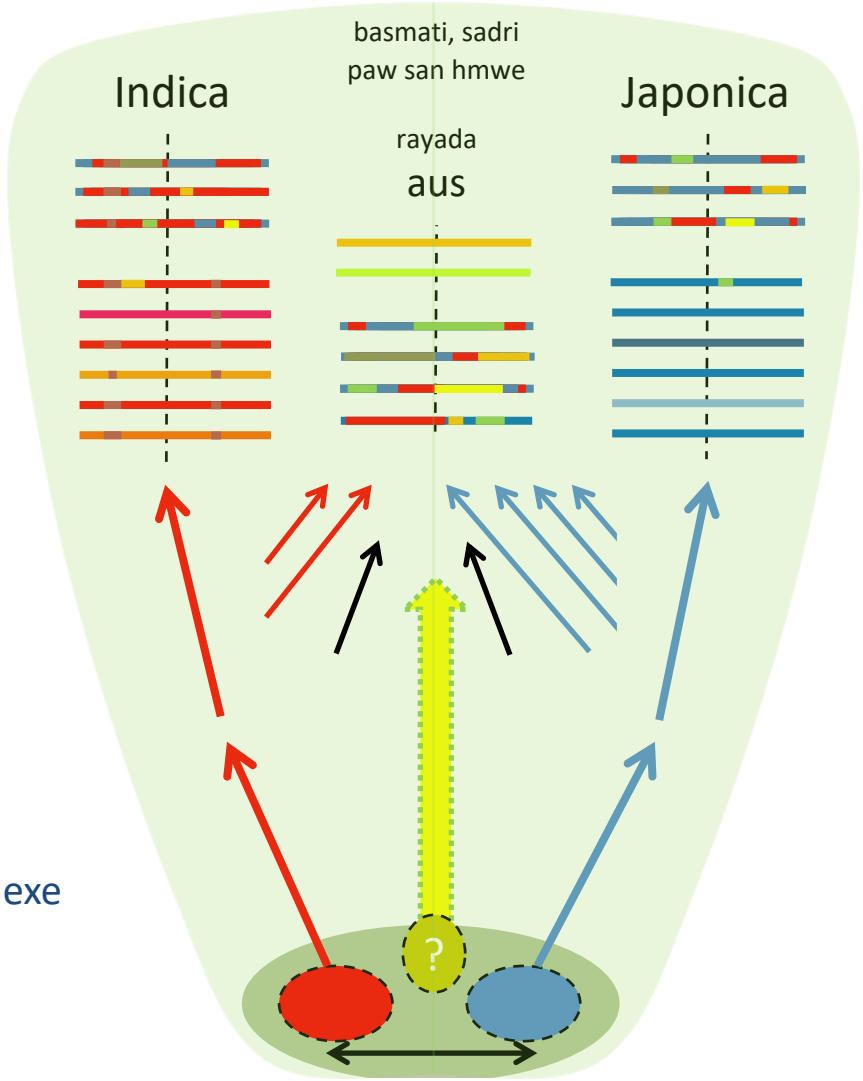


→ mosaïque plus complexe qu'anticipée= plus de générations de brassage méiotique

→ plus de contributeurs sauvages ancestraux qu'anticipés 2 à 5-6, dont un ou deux non identifiés

Context Riz



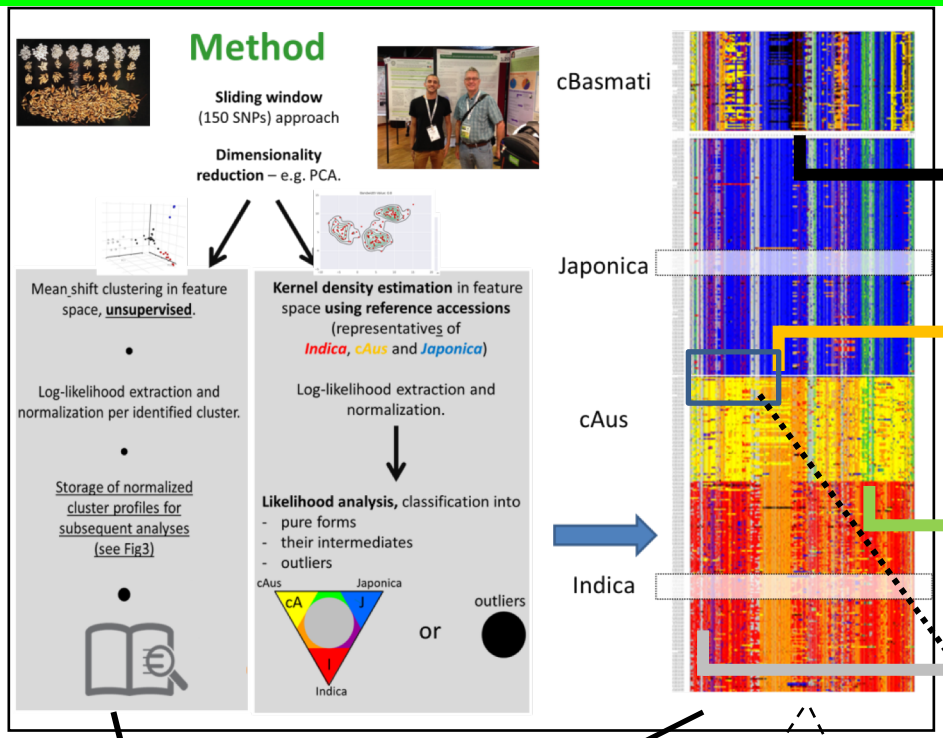


- Hybridation suivie de beaucoup d'évènements de méioses interspécifiques → structure mosaïque complexe

- Identification de SNP diagnostiques des taxons ancestraux, taxon ancestraux déjà introgressés et potentiellement pas tous connus

Décryptage mosaïques Riz

J. Santos/JC Glaszmann/…



« alien »

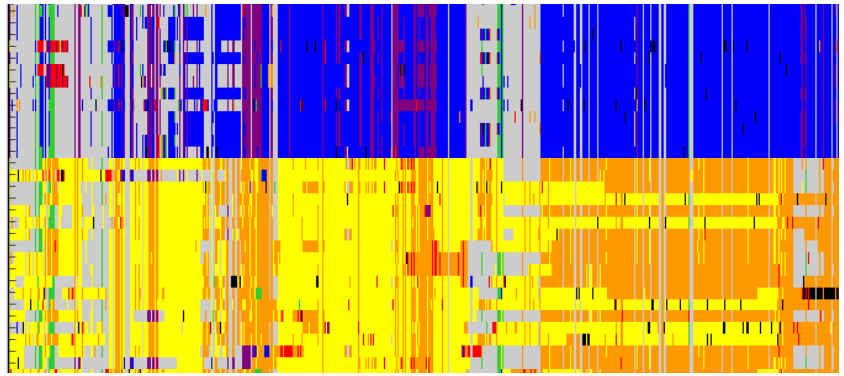
Introgression

among

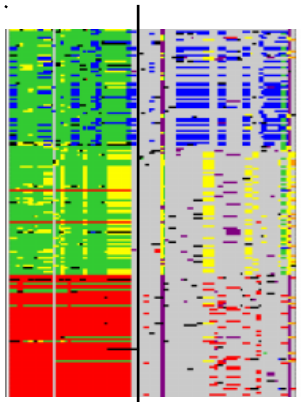
varietal

groups

patterns



SH4, shattering



A repertoire for further in-depth analysis without *a priori*

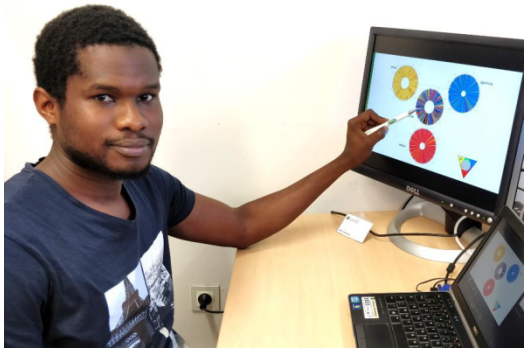
CultiVar : où en est-on ?

(épisode n°13 – 6 juin 2018)

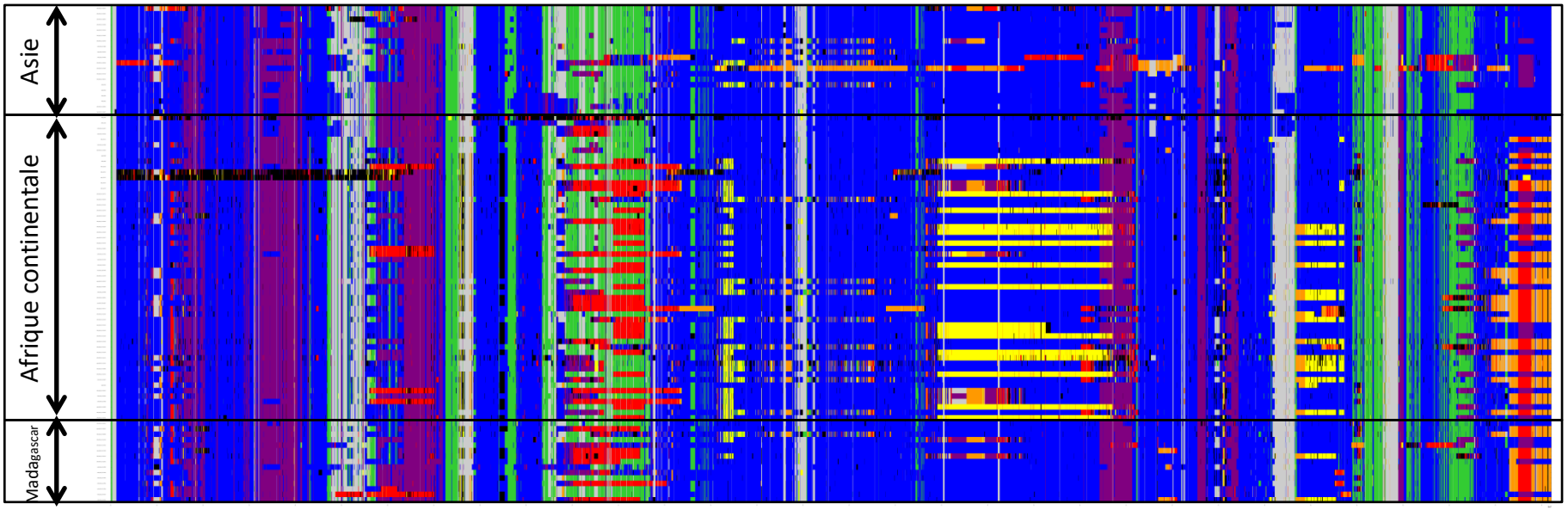
CultiVar
Training in plant breeding

www.cultivar-flagship.net

Juin, la saison des stages



Abdoulaye Beye étudie les introgressions entre groupes variétaux lors de la migration et la dispersion des riz asiatiques en Afrique

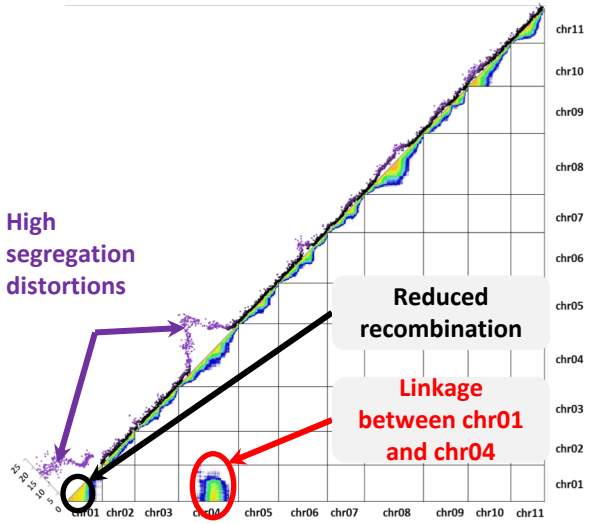


Les Japonicas d'Afrique (les fameux riz pluviaux) portent des introgressions spécifiques issues vraisemblablement de **cAus** et de **Indica**

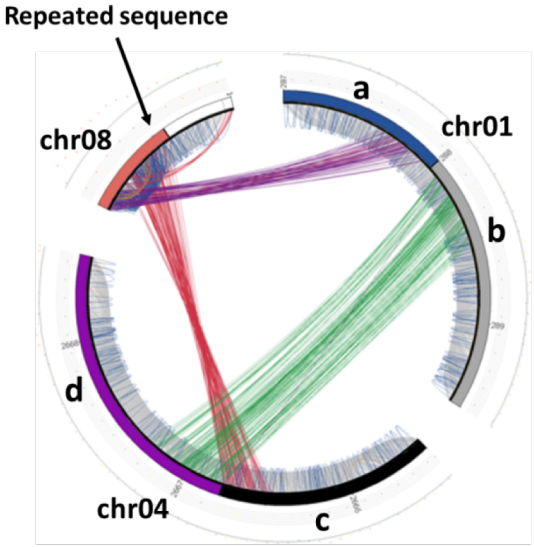
Characterization of large structural variations

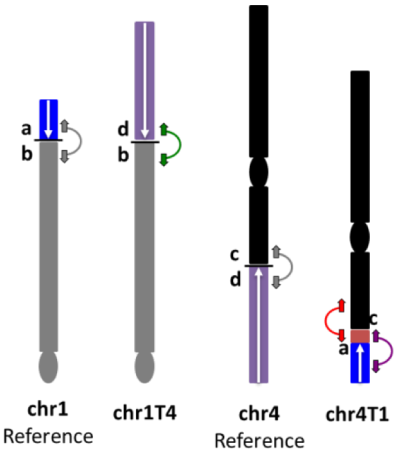
G. Martin/F. Baurens/
A. D'Hont/…

Linkage analysis



Mate Pair sequencing

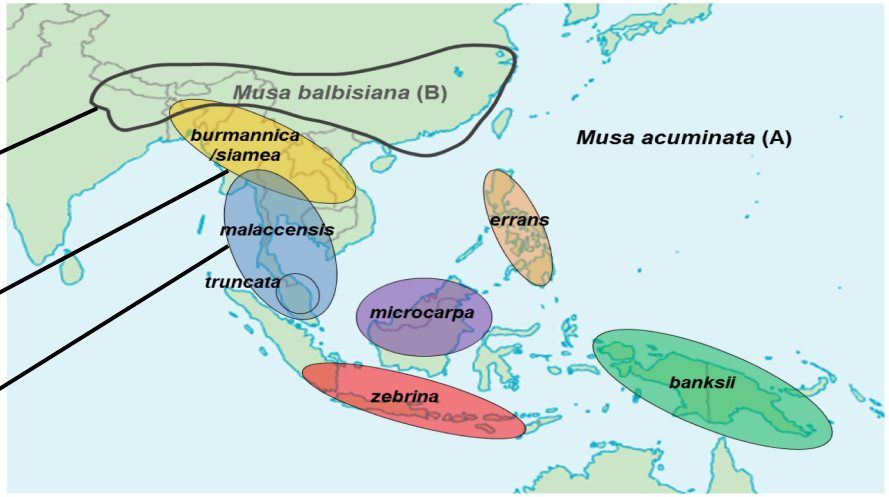




Translocation 1/3
Inversion 5
Baurens et al., sub.

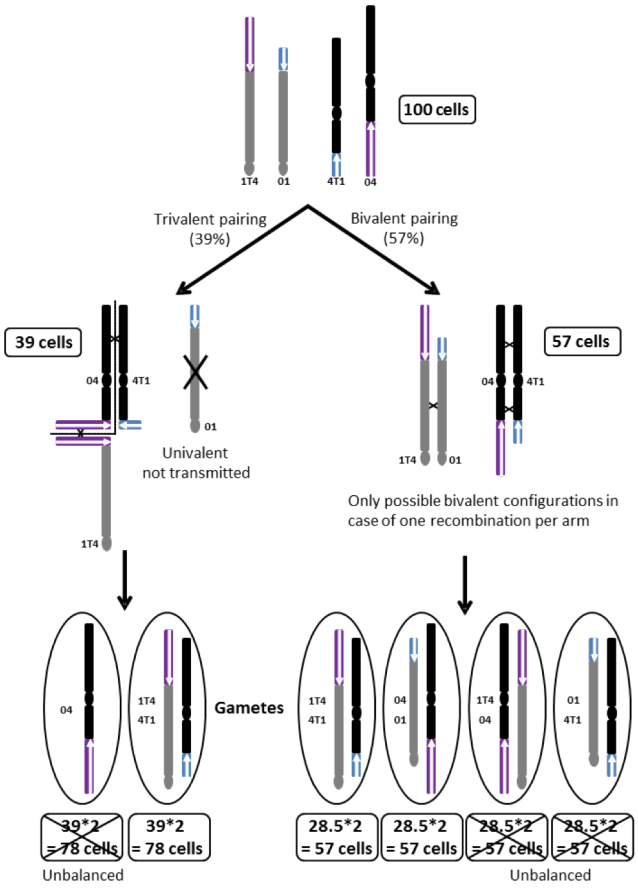
Translocation 1/9
Translocation 2/8
Dupouy et al., in prep

Tanslocation 1/4
Martin el al., 2017

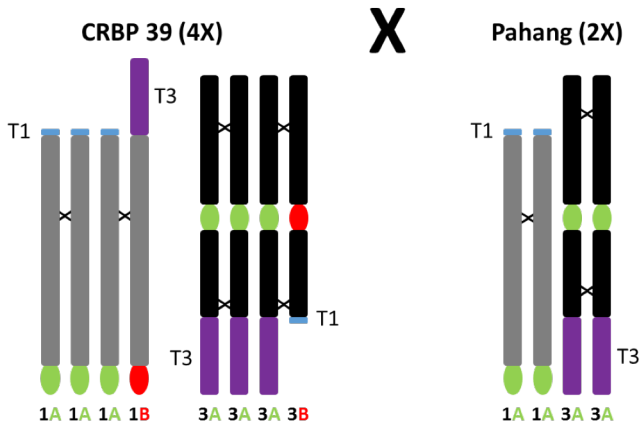


Impact des variations structurale sur les ségrégations chromosomiques

Contexte diploïde



Contexte triploïde



→ formation aneuploides =
perte ou gain d'un chromosome
ou d'un segment de chromosome

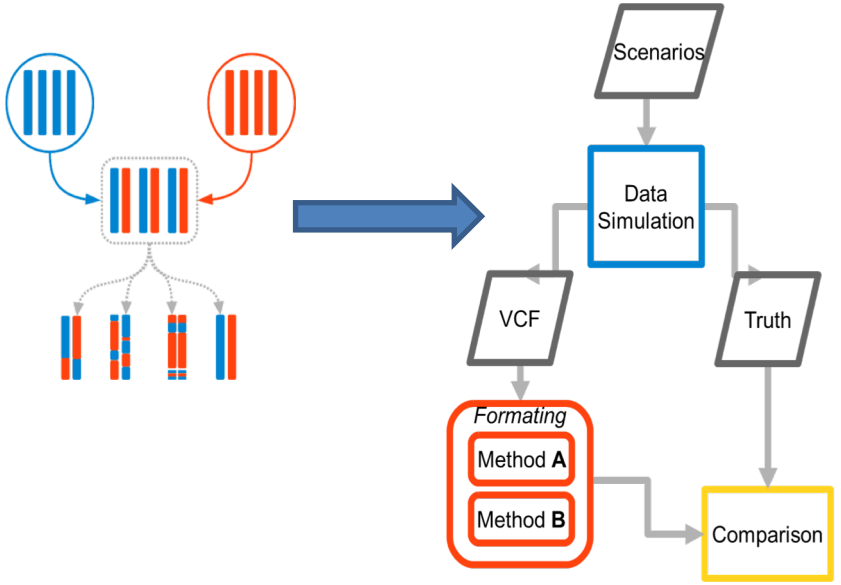
→ Réduction de fertilité
→ Biais de ségrégations =certaines combinaisons d'allèles moins ou non représentées

En cours/perpectives

V. Berry/JC Glaszmann/M. Gautier/Nabila Yahiaoui/G. Martin/F. Baurens/A. D'Hont/…

Décryptage des mosaïques:

- autres méthodes en cours de développement:
  - méthodes basés sur la phylogénétique : CE Rabier Post Doc LIRMM
  - basée sur la génétique des populations: Thèse A Cottin, AGAP/ GBGP



- un simulateur en construction pour tester les approches en fonctions des spécificités biologiques des plantes étudiées: Thèse A Cottin, AGAP/ GBGP

- applications aux nouveaux jeux de données citrus, bananier, riz, sorgho (projet Muse),…

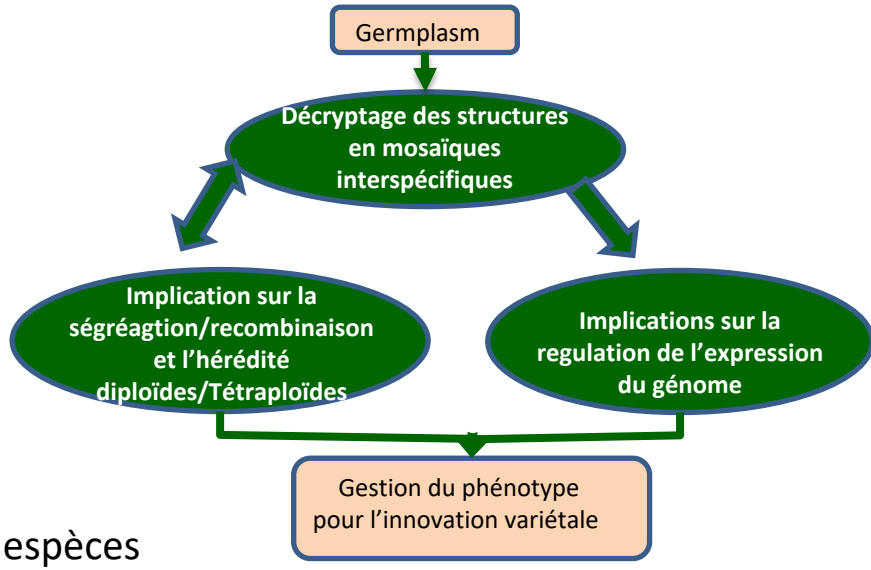
- impact sur l'expression des allèles

Décryptage des variations structurales:

- autres méthodes en cours de développement

Thèse M. Dupouy

- applications aux nouveaux jeux de données bananier, citrus
- impact sur ségrégation des chromosomes, la détection de QTL,…



→ implications pour programmes d'amélioration de ces espèces

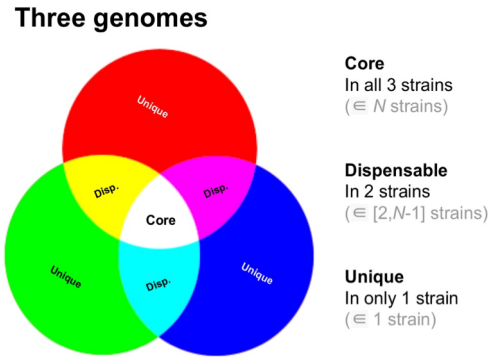
En cours/perpectives

M. Ruiz/A. Mancheron/…

On dispose maintenant souvent de **plusieurs séquences de références** pour une plante/espèce

→ il existe de nombreuses variations aux niveau de ces séquences, SNP, INDEL mais aussi dans le **contenu en gènes** = notion de pan et core génomes

Sur les quelque 24000 familles de gènes connues chez le riz, seulement 60% sont communes à toutes les variétés; les autres sont présentes, ou absentes, selon les variétés



→ développement de méthodes de stockage, d'interrogation et visualisation de ces données

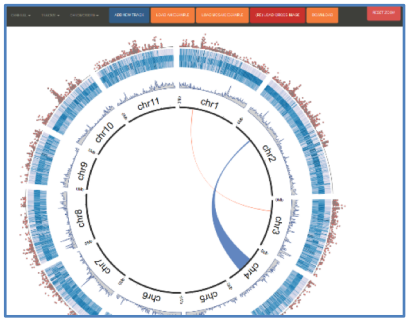
→ Thése C. Agret, collaboration LIRMM/AGAP

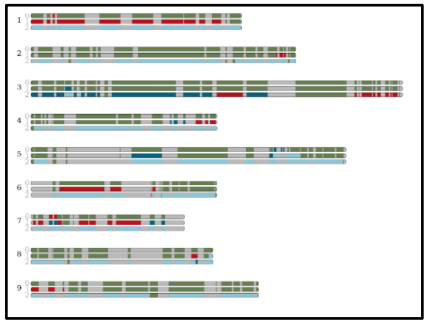
RedOak : Structure d'indexation génomique

Intégration/mise à disposition des outils

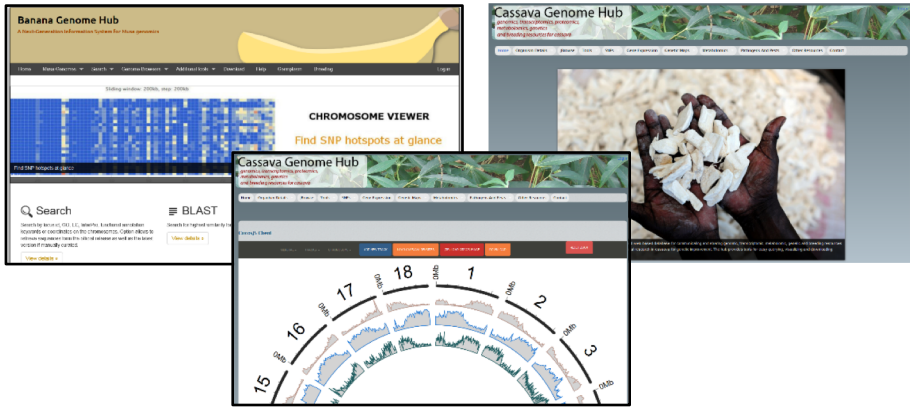
A. Dereeper/G. Droc/
A. Comte/M. Rouad/…

1) Développement d'outils de visualisation





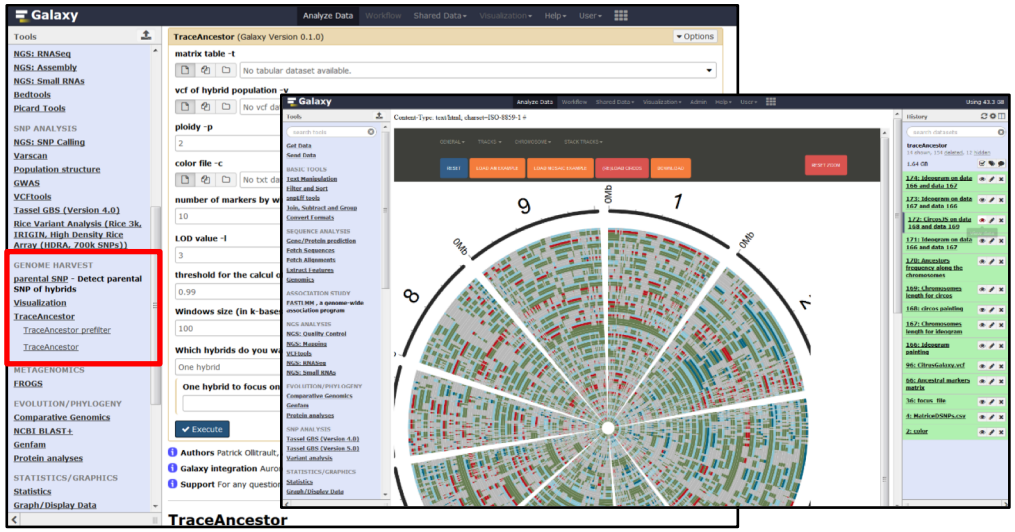
3) Connexion des outils aux Genome Hub



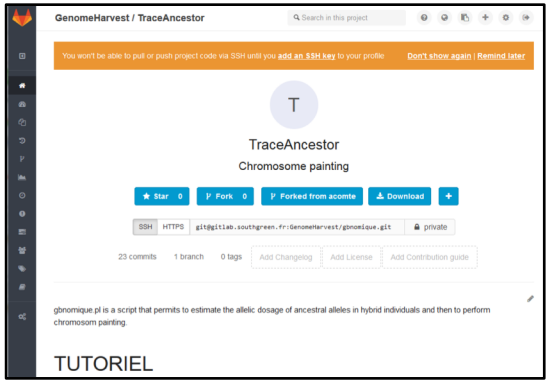
2) Accès aux outils par Galaxy.
Développement de wrappers/workflows

Galaxy

http://galaxy.southgreen.fr/galax



4) Mise à disposition du code et documentation via GitHub/GitLab



github SOCIAL CODING

GitLab

Activité 2
(characterize inter(sub)specific mosaic genome structures)

Activité 3
(impact of genome structure on gene expression)

KDE classifier

python

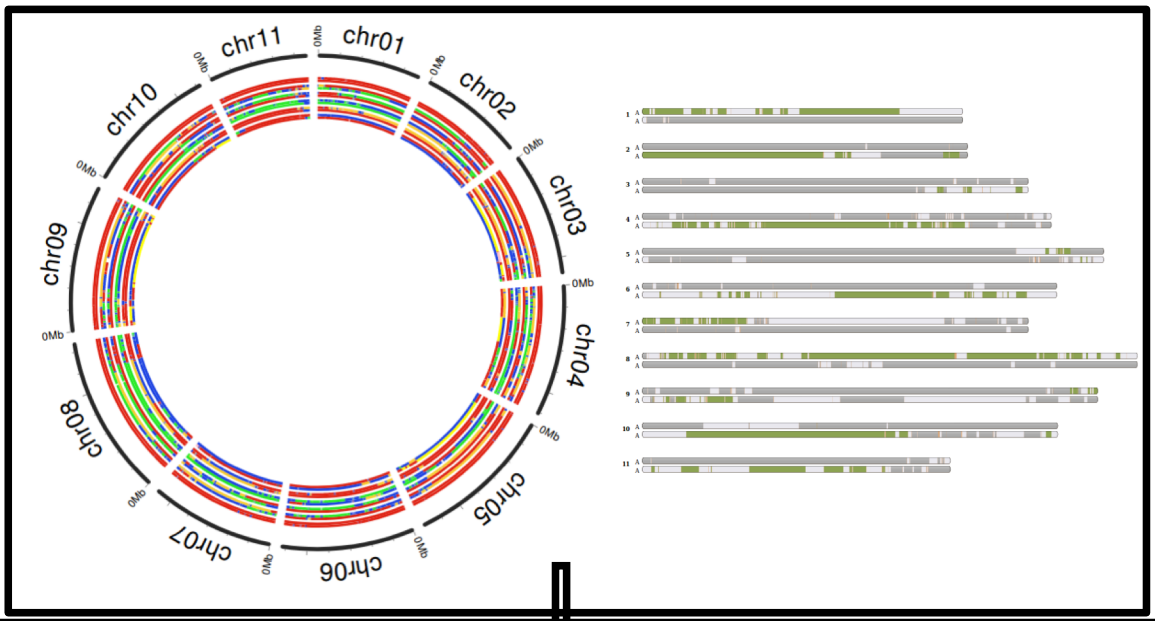
VCF Hunter

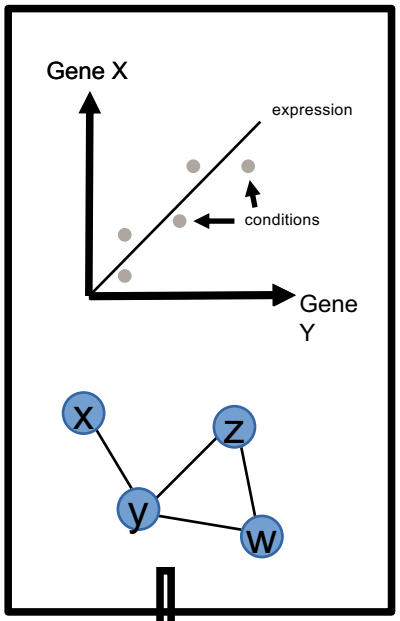
python

TraceAncestor

perl

ASE co-expression





Galaxy

Banana Genome Hub
A Next-Generation Information System for Musa genomics

GitLab

github
SOCIAL CODING

Introduction

Outil	Organisme sur lequel il a été développé	Nombre d'haplotypes pris en charge	Permet le phasing ou prend en charge données phasées	Méthode	inputs
TraceAncestor	agrume	2 à 4	non	Estimation de la fréquence des allèles ancestraux identifiés par indice GST	VCF + Liste de SNPs diagnostiques
KDE_Classifier	riz	1	non	Kernel Density Estimation	VCF / géno + fichier structure
VCFHunter	banane	N	non	ACP + clusterisation	VCF

GenomeHarvest
diversity, organization and dynamics

Formation sur les logiciels de reconstruction de génomes mosaïques sous galaxy

http://cc2-web1.cirad.fr/galaxydev et http://galaxy.southgreen.fr/galaxy/

Le 2 juillet 2018 - Aurore Comte

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools

NGS: Quality Control
NGS : Mapping
NGS: GATK Tools
NGS: GATK2 Tools
NGS: SAM/BAM Manipulations
NGS: RNASeq
NGS: Assembly
NGS: Small RNAs
Bedtools
Picard Tools

SNP ANALYSIS

NGS: SNP Calling
Varscan
Population structure
GWAS
VCFtools
Tassel GBS (Version 4.0)
Rice Variant Analysis (Rice 3k, IRIGIN, High Density Rice Array (HDRA, 700k SNPs))

GENOME HARVEST

parental SNP – Detect parental SNP of hybrids
Visualization
TraceAncestor
KDE_classifier

METAGENOMICS

FROGS

EVOLUTION/PHYLOGENY

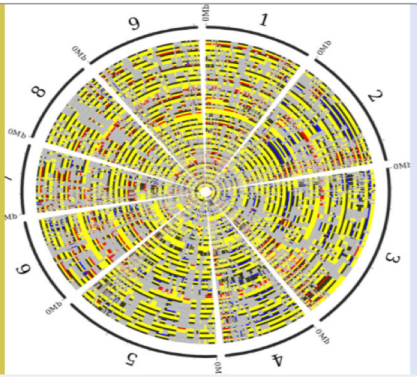
Comparative Genomics
NCBI BLAST+
Genfam

SouthGreen bioinformatics platform

Welcome to GALAXY

Our pre-configured and validated workflows

NGS analyses | SNP calling | SNP analysis | GWAS | Structural variations | Chrom. reconstruction | Metagenomics | Gene families | Mosaic Genomes



Mosaic genome reconstruction

TraceAncestor / KDE_Classifier : Two approaches to analyze the mosaic structure of plant genomes

Input: VCF file + structure file

Access workflow

These workflows as part of the services provided by South Green

How to load big datasets?

Core values

- **Accessibility**
  - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
  - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
  - Users can share or publish their analyses (histories, workflows, visualizations)
  - Pages: online Methods for your paper

Pages: interactive, web-based documents that describe a complete analysis.

=> Diffusion des wrappers via le Galaxy

Toc 

Les structures en mosaïques interspécifiques et leurs implications au coeur des questions de recherche en amont des projets d'innovation variétale