Deciphering species-level phylogenetic relationships in the evolutionary complex genus *Rosa* using an amplicon-sequencing approach

Kevin Debray[1], Marie-Christine Le Paslier[2], Aurélie Bérard[2], Tatiana Thouroude[1], Gilles Michel[1], Fabrice Foucher[1] and Valéry Malécot[1]

[1]IRHS, Agrocampus-Ouest, INRA, Université d’Angers, SFR 4207 QuaSaV, Beaucouzé, France
[2]INRA, US 1279 EPGV, Université Paris-Saclay, Evry, France

PHY-ROSE
2016-2019

Colloque EPGV
October, 3rd 2018

Background – *The Tree and Classification of Life*

Domain: Eukaryota

Kingdom: Plantae

Subkingdom: Tracheobionta

Superdivision: Spermatophyta

Division: Magnoliophyta

Class: Magnoliopsida

Subclass: Rosidae

Order: Rosales

Family: Rosaceae

Subfamily: Rosoideae

Tribe: Roseae

Genus: *Rosa*

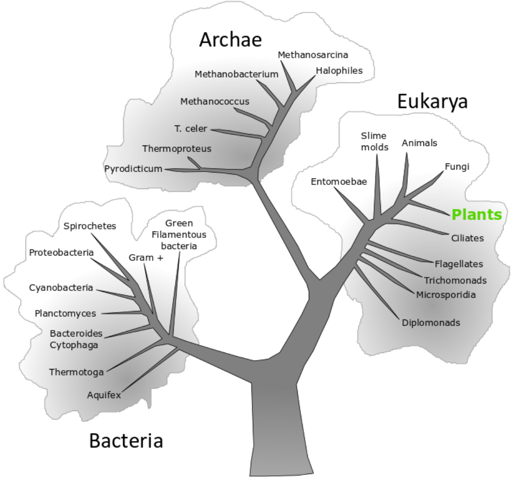
Subgenus: *Rosa*

Section: *Caninae*

Subsection: *Caninae*

Species: *Rosa canina* L.

Levels of interest

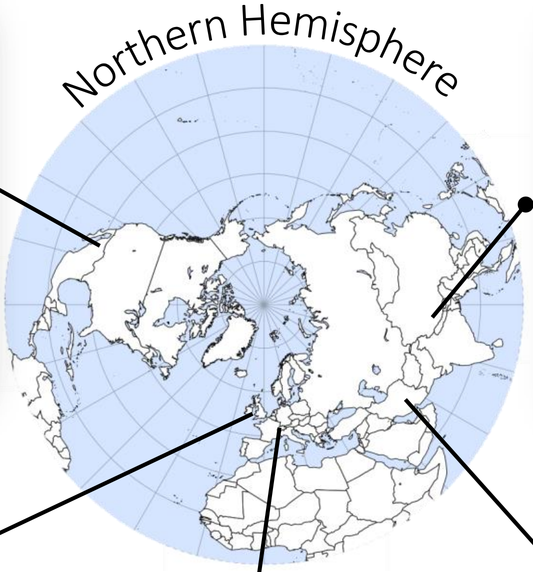


Background – *The genus* Rosa

150-200 species

Background – *Wild roses distribution*

Mexican valleys



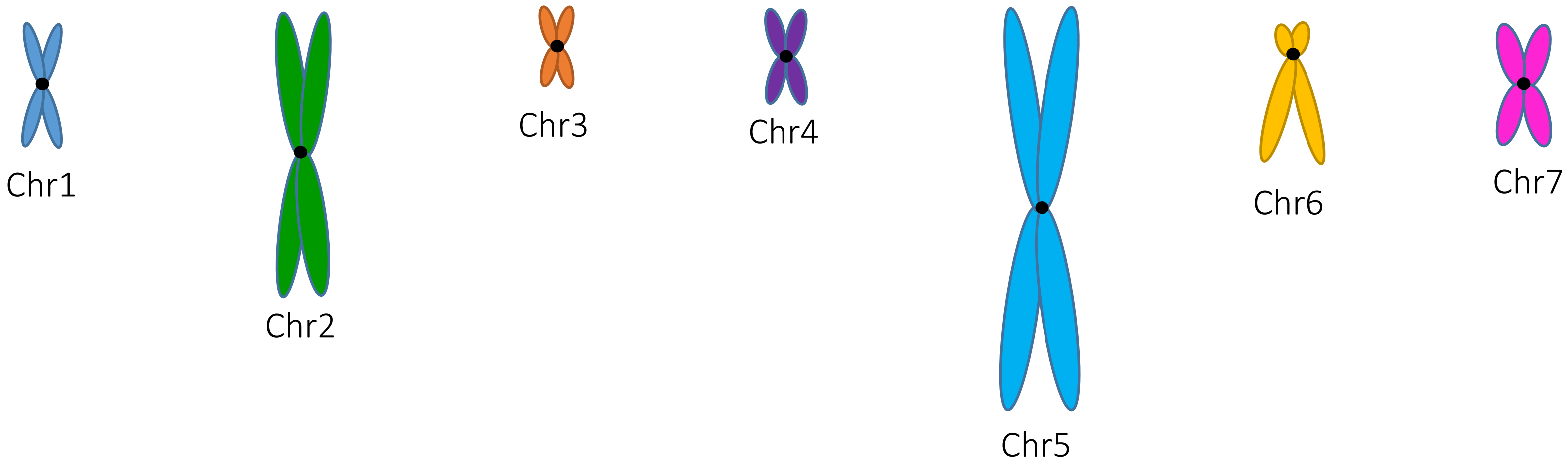
Himalayas

British coast

Alpine mountains

Iranian desert

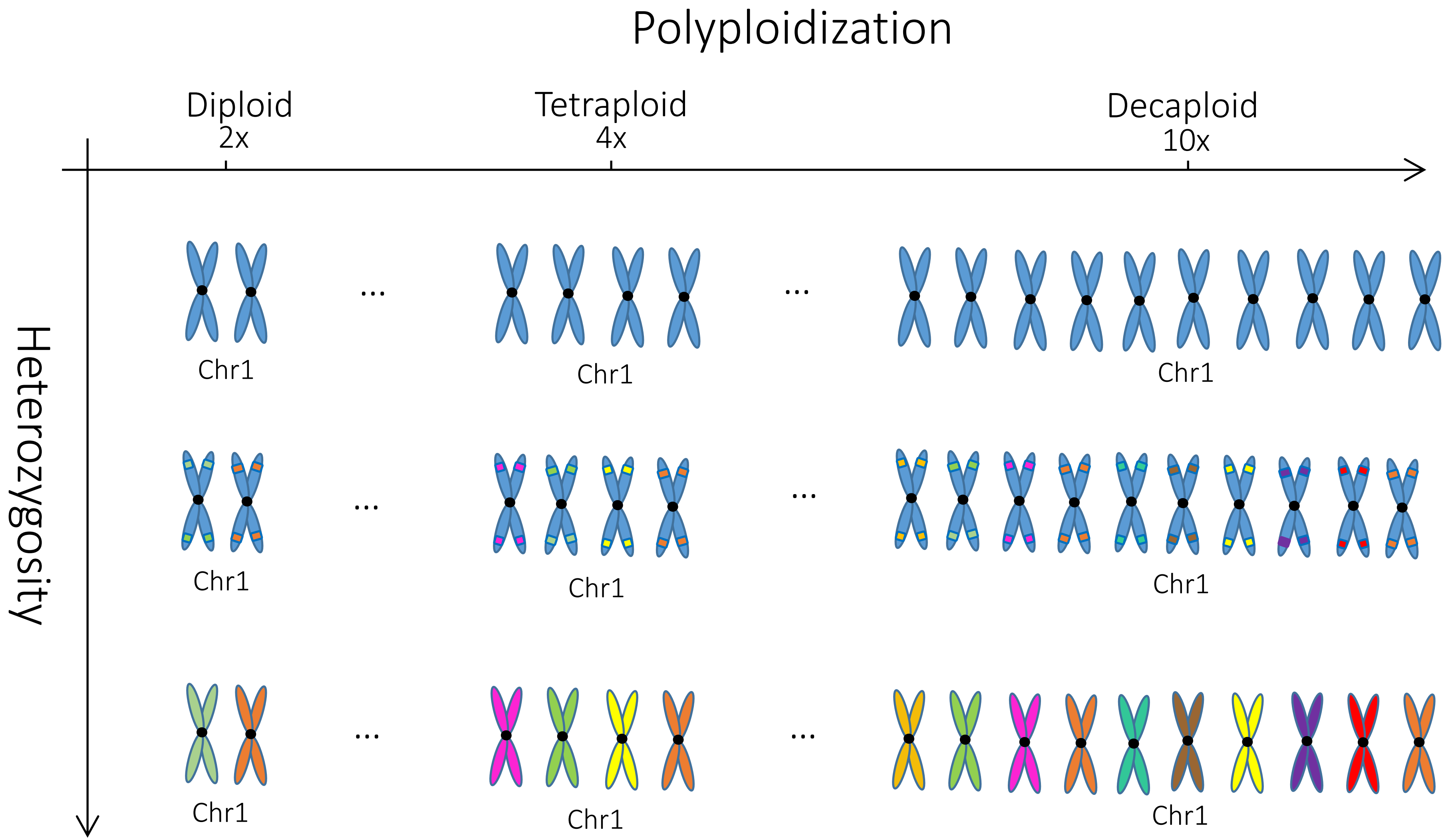
Background – *The Rose genome*



Haploid chromosome number: n=**7** (*Voanioala gerardii* n=298)

Monoploid genome size: 1Cx = **0,4-0,6 Gb** (*Fritallaria platyptera* 1Cx = 84 Gb)

Background – *Adding complexity*



Polyploidization

Diploid
2x

Tetraploid
4x

Decaploid
10x

Heterozygosity

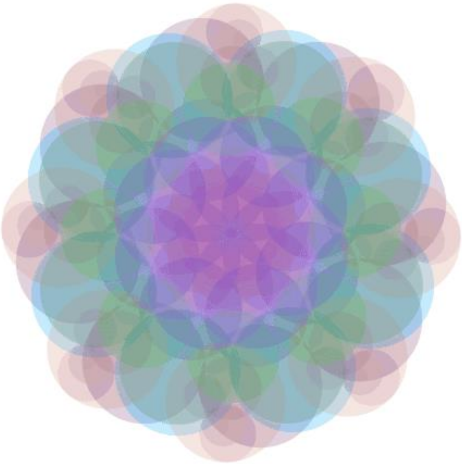
Chr1 … Chr1 … Chr1

Chr1 … Chr1 … Chr1

Chr1 … Chr1 … Chr1

Background – *Complex genomes*

Roses genomes are very **flexible**…
…as modeling clay!



… And this questions the notion of 'species'…

Background – *Current phylogenetic issues*

Morphological issues

Rosa canina L.

R. nitidula Besser

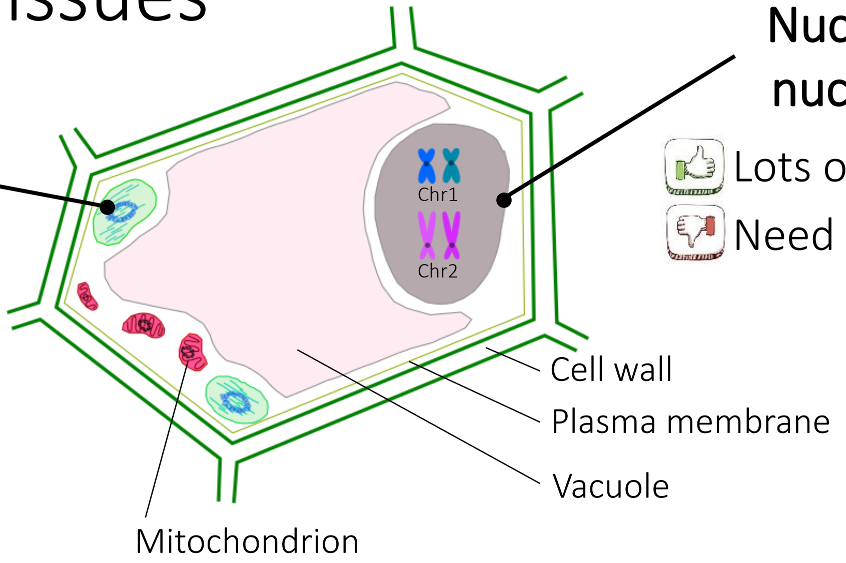
R. caesia Smith

R. corymbifera Borkh.

Plastid sequences issues

Chloroplast with plastid DNA

- 👍 Easy to target and analyze
- 👎 No individual polymorphism
- 👎 Slow rate of evolution
- 👎 Putative maternal heredity



Nucleus with nuclear DNA

- 👍 Lots of allelic variations
- 👎 Need prior knowledge

Background – *Current phylogenetic issues*

Morphological traits and plastid sequences have **significant flaws** to study roses relationships…

…One solution would be to analyze **many nuclear loci** across rose genomes…

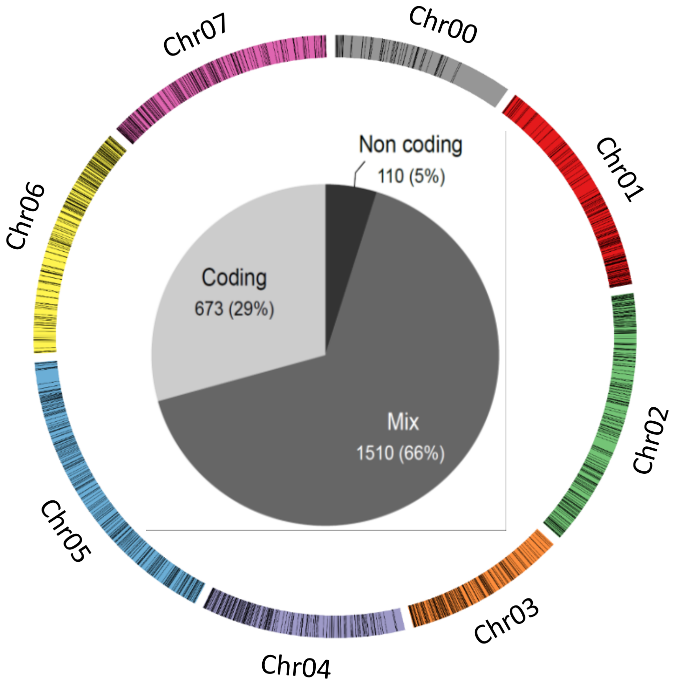
…Welcome to **Phylogenomics**!

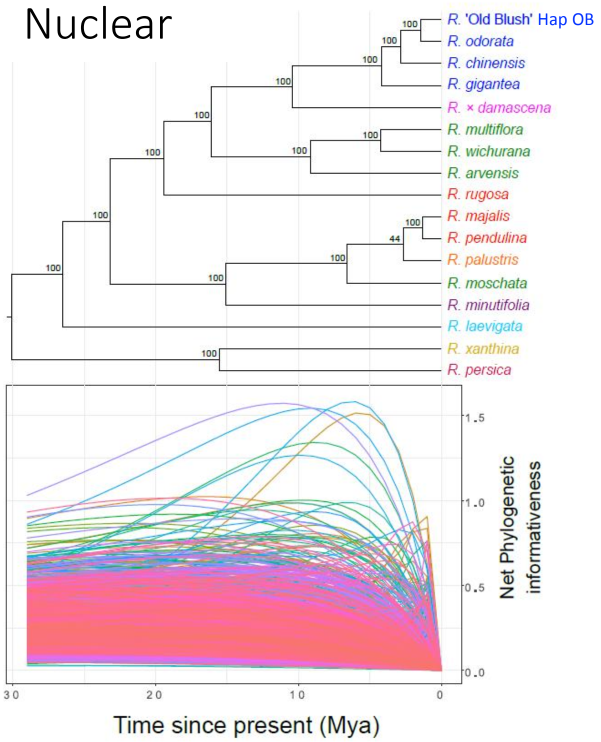
Objectives – *Whole thesis objectives*

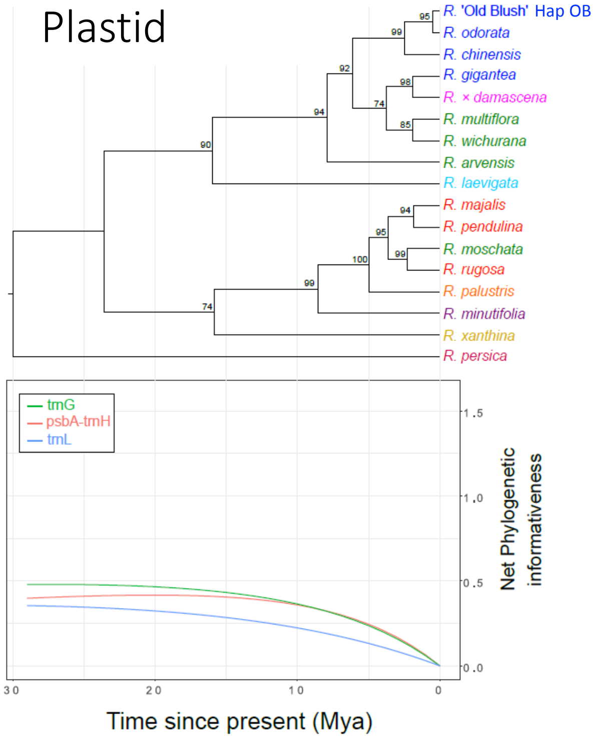
1. Identifying and assessing of **single-copy nuclear loci** for *Rosa* phylogenomics

2. Building a **phylogenomic network** of *Rosa* species highlighting **polyploidizations and hybridizations** as well as **traits evolution and biogeography**

Methods – *"Previously in PHY-ROSE…"*







A genome-mining strategy identifies **2,293 variable single-copy nuclear loci** for rose phylogenomics

Debray *et al.* In Prep.

Methods – *"Previously in PHY-ROSE…"*

Royal Botanic Garden Edinburgh

Centre sur la biodiversité de l'Université de Montréal

La Roseraie du Val-de-Marne

Missouri Botanical Garden

Universität Bonn

The Arnold Arboretum of Harvard University

Conservatoire Botanique National Franche-Comté

Ville de Nantes

arboretum du vallon de l'Aubonne

Quarryhill Botanical Garden

Conservatoire botanique national de Franche-Comté

Université de Strasbourg

arboretum Kalmthout

LUOMUS
Luonnontieteellinen keskusmuseo
Naturhistoriska centralmuseet
Finnish Museum of Natural History

Wageningen University & Research

Cambridge University Botanic Garden

Roses Loubert
www.pepiniere-rosesloubert.com

MUN Botanical Garden
Memorial University of Newfoundland

The Morton Arboretum

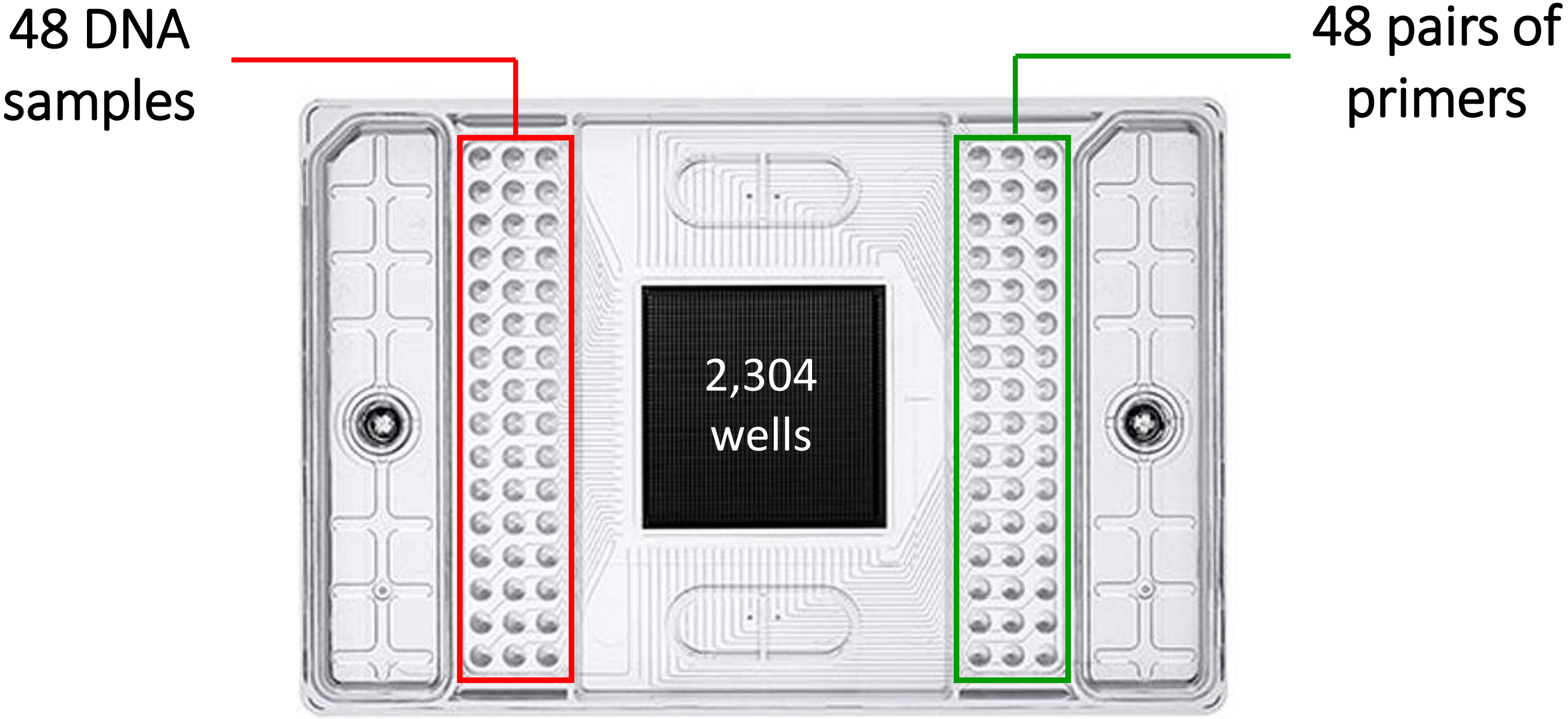
Plantentuin Meise

Rancho Santa Ana Botanic Garden

Silica-dried leaves bank of ca. 300 accessions (131 species)

Methods – *The Amplicon sequencing technique*

Step 1/2: Target PCR amplification of loci

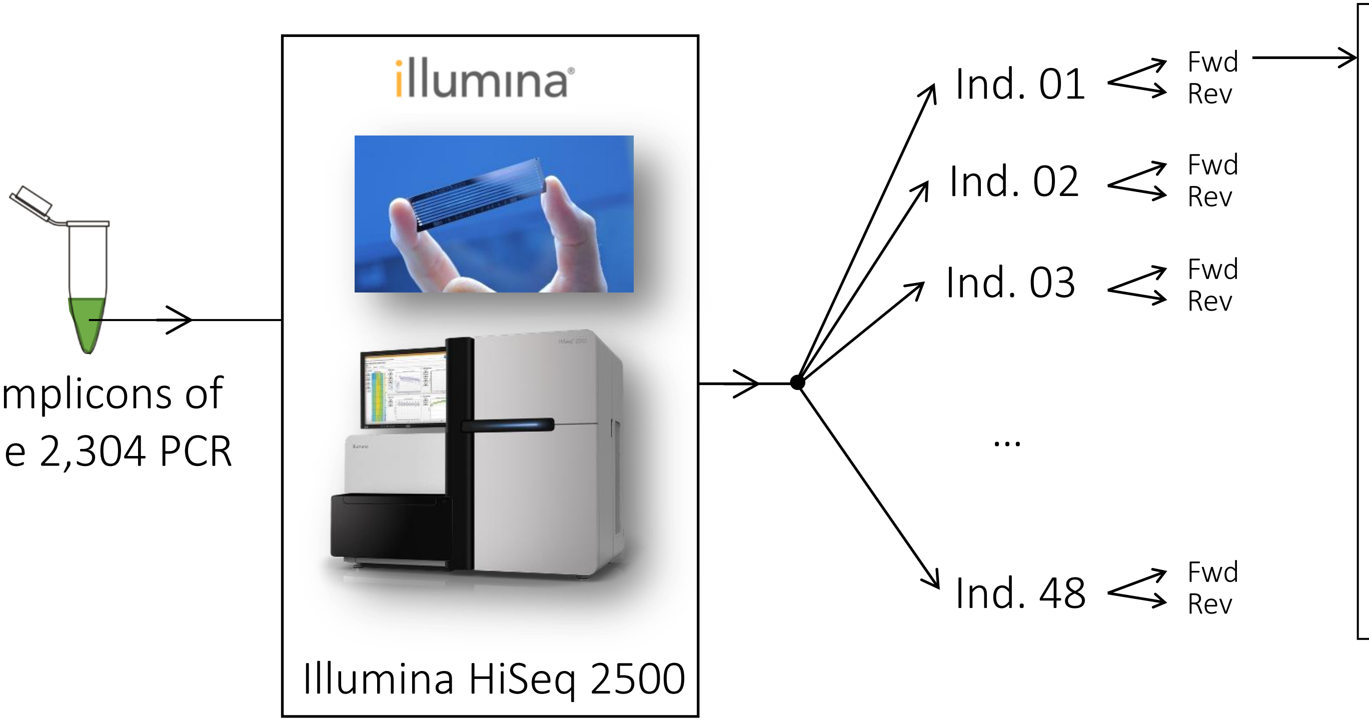


FLUIDIGM®

Access Array 48*48

Methods – *The Amplicon sequencing technique*

Step 2/2: Illumina sequencing 2*300 bp



IND01_FWD.fastq

```
@SEQ00000001\1
GATTTGGGGTTCAAAGCAG
+
!''*((((***+))%%%++)(%%%%)
@SEQ00000002\1
GATTTGGGGTTCAAAGCAG
+
=><=\*()**%$=)('%++)(%A'@
@SEQ00000003\1
GATTTGGGGTTCAAAGCAG
+
@()+??((*6*+>AC!!#@'&@@##
…
…
…
@SEQ25489732\1
GATTTGGGGTTCAAAGCAG
+
AC!!#@'&@@##*((((*!!?#';
```

96 files to demultiplex

Methods – *Post-sequencing workflow*

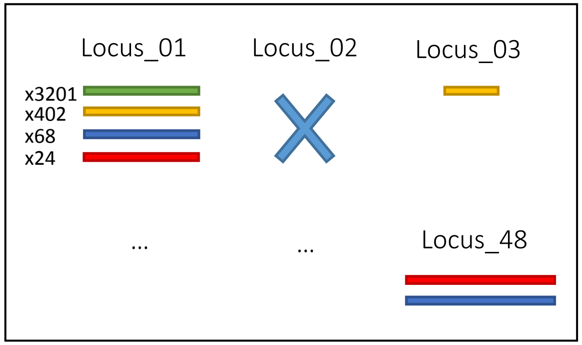
IND01_FWD.fastq

```
@SEQ00000001\1
GATTTGGGGTTCAAAGCAG
+
!''*((((***+))%%%++)(%%%%)
@SEQ00000002\1
GATTTGGGGTTCAAAGCAG
+
=><=\*()**%$=)(‘%++)(%A’@
@SEQ00000003\1
GATTTGGGGTTCAAAGCAG
+
@()+??((*6*+>AC!!#@’&@@##
…
…
…
@SEQ25489732\1
GATTTGGGGTTCAAAGCAG
+
AC!!#@’&@@##*((((*!!?#’;
```

IND01_REV.fastq

```
@SEQ00000001\2
CTGCTTTGAACCCCAAATC
+
DJ!!#@’&@@##*((((*!!?#’;
@SEQ00000002\2
CTGCTTTGAACCCCAAATC
+
*+>AC!!#@’&%$=)(‘%++)(%
@SEQ00000003\2
CTGCTTTGAACCCCAAATC
+
@()+??((*6*+>AC!!#@’&@@
…
…
…
@SEQ25489732\2
CTGCTTTGAACCCCAAATC
+
DJ!!#@’&@@##*@()+??((*6
```

GOAL



1 QUALITY CHECKS

2 DEMULTIPLEXING

3 TRIMMING

4 ASSEMBLY

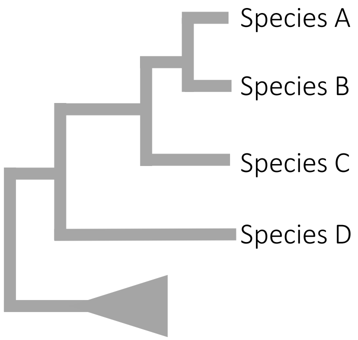
5 CLUSTERING

6 HAPLOTYPING

Methods – *Phylogenomics networks*

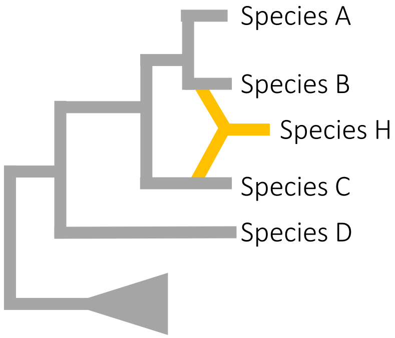
Method 1: **with** a priori

1. Diploid non-hybrid tree
   = Backbone diploid progenitors tree



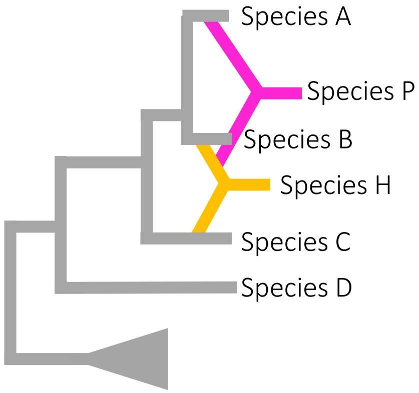
2. Graft alleles from diploid hybrids and merge

Diploid-Hybrids detection?



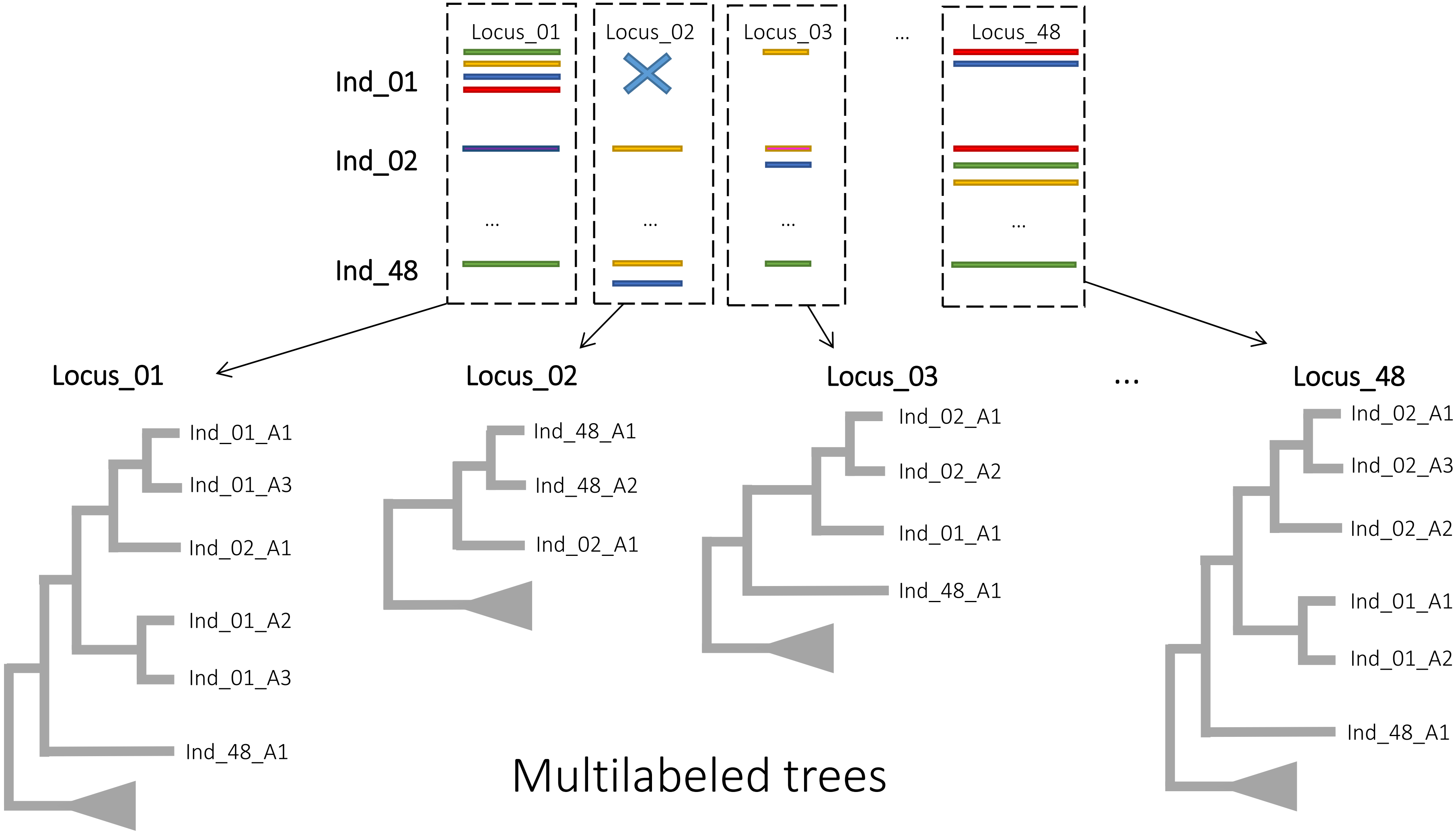
3. Graft alleles from polyploids and merge

Polyploids detection?



Methods – *Phylogenomics networks*

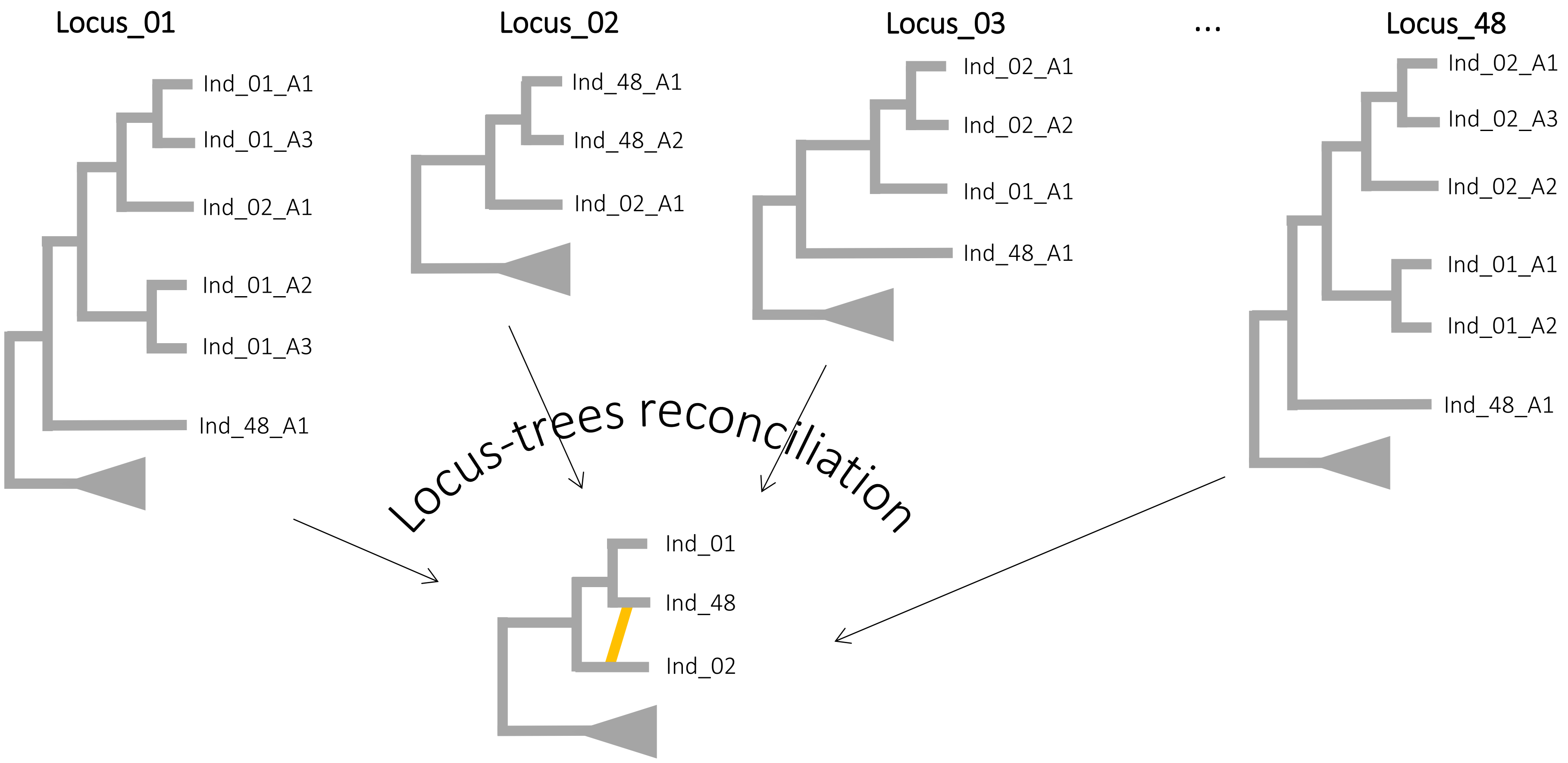
Method 2: **without** a priori



Multilabeled trees

Methods – *Phylogenomics networks*

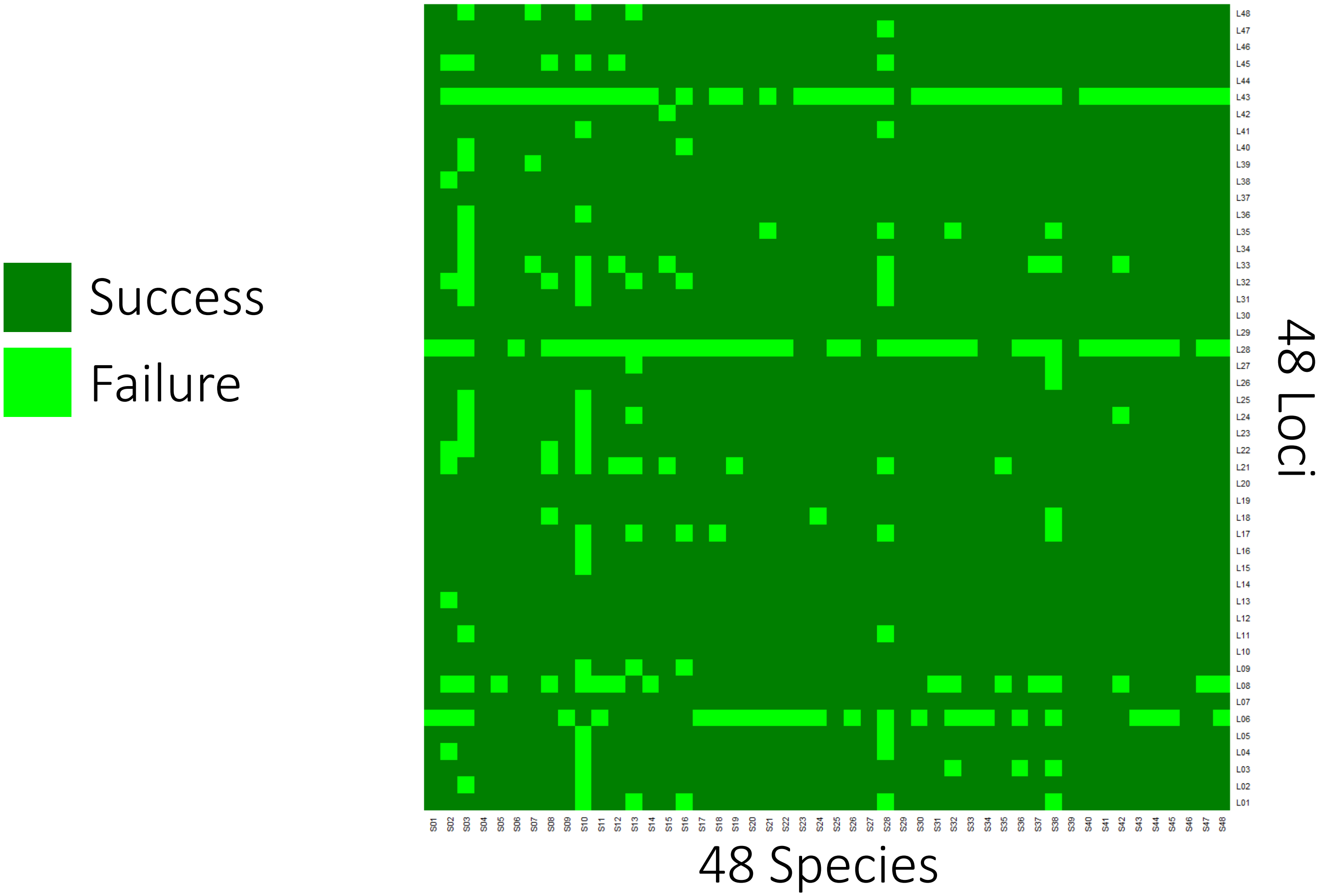
Method 2: **without** a priori



Species networks

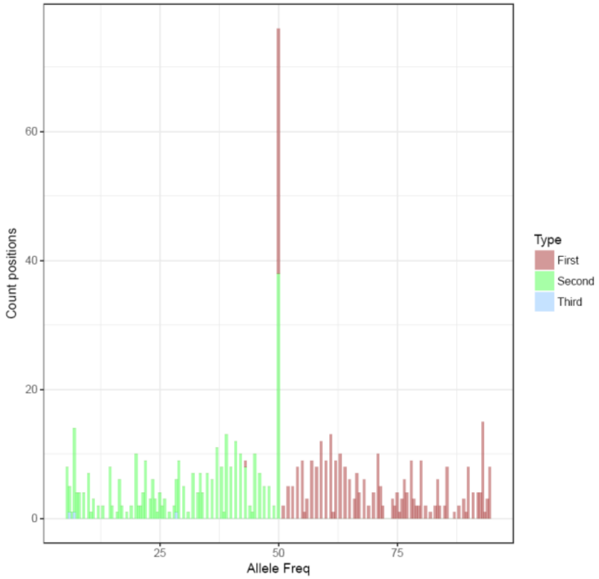
Results – *Amplicon Sequencing*

1 pilot run already analyzed: **90% of success**

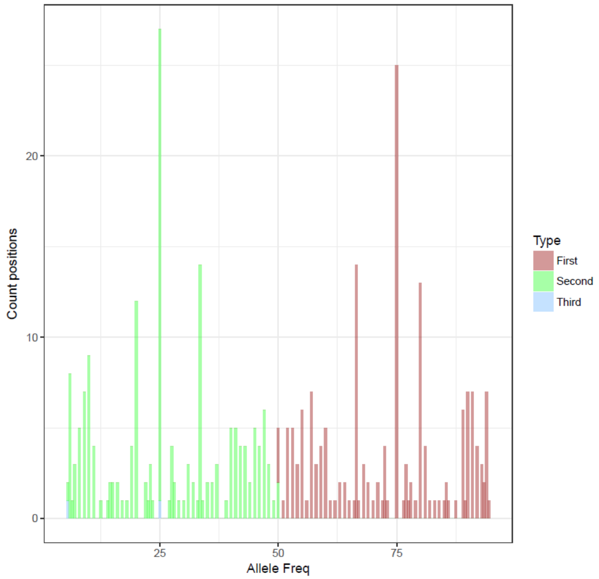


Results – *Method 1: with a priori*

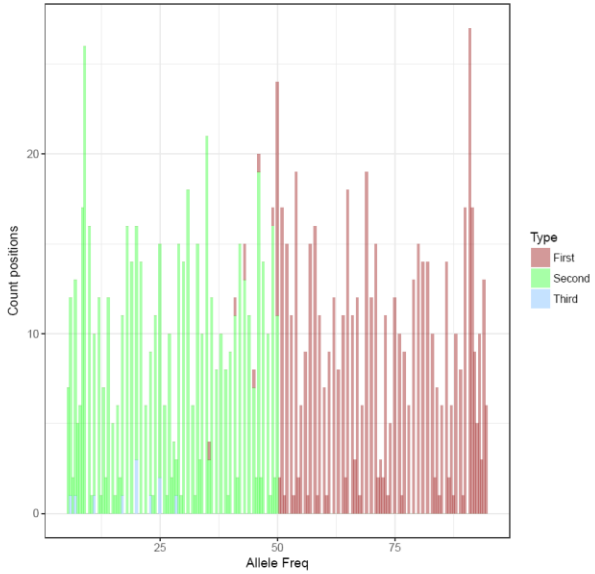
Polyploid detection?



50:50
2x



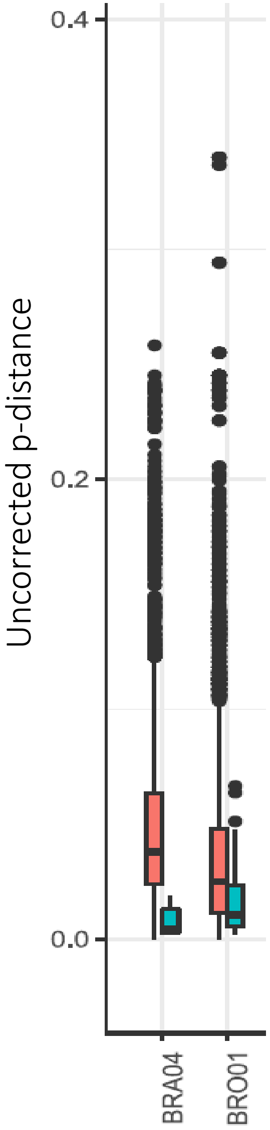
25:75
4x



?
> 2x

Alleles frequencies in combination with literature give an idea of the **ploidy level**

Results – *Method 1: with a priori*



Diploid-hybrids detection?

Comparison of **intra/inter** specific sequence **variations**

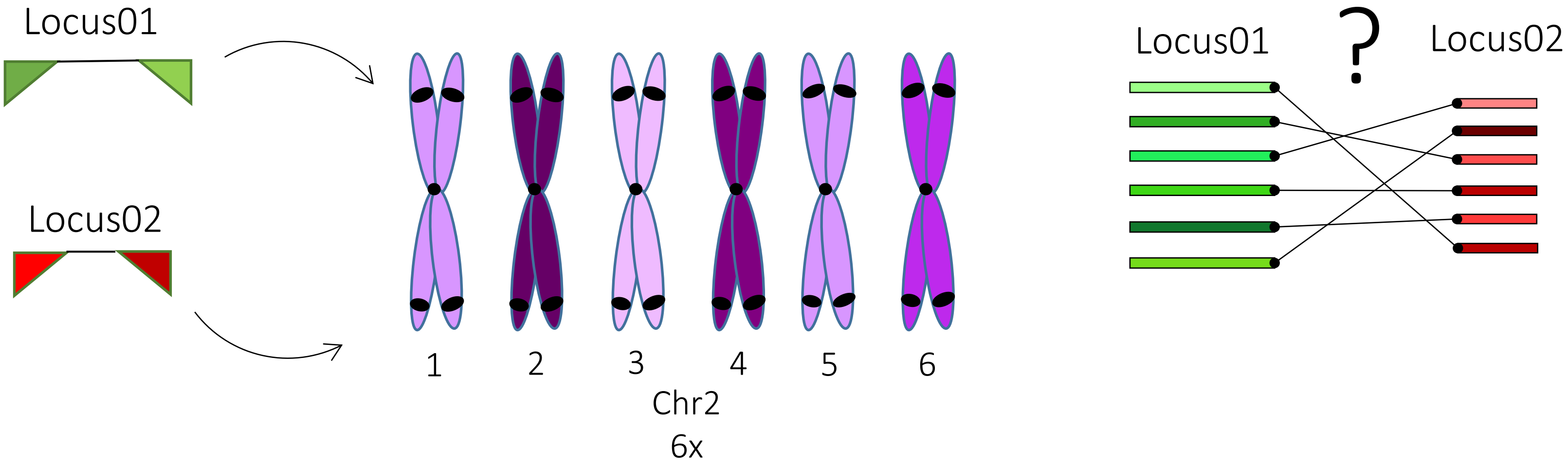
*“**Intra**-individual variations << **Inter**-individual variations in diploids non-hybrids”*

Discussion – *Amplicon sequencing*

- Prior (genomic) knowledge on the taxa studied

  Barstia (Orobanchaceae) Uribe-Convers *et al*. 2016 PLoS ONE 11(2):28pp
  Cucurbita (Cucurbitaceae) Kates *et al*. 2017 MPE 111:98-109

- Targeting alleles



👍 Allele variation
👎 Allele phasing

Discussion – *Amplicon sequencing*

- Robustness toward DNA quality

  Carefree sampling and storage
  Herbarium samples can be used
  Some accessions were sampled before 2000

- High-quality sequencing for low price

  Deep sequencing coverage (up to 11,591 X – mean was **3,308 X**)

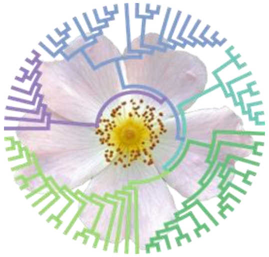
  0,83€/locus/species *vs*. ca. 3 €/locus/species using 1X Sanger sequencing

Perspectives – *On the origin of* Rosa persica

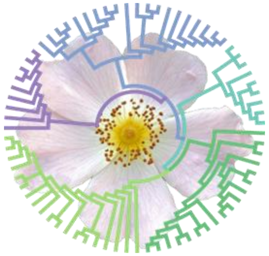
Rosa persica: Living fossil or super evolved rose?

Dr Zahra Karimian
Research Center for Plant Sciences
Ferdowsi University of Mashhad - Iran

Acknowledgments



Thanks for your attention!



Curators and collectors

Royal Botanic Garden Edinburgh

LUOMUS
LUONNONTIETEELLINEN KESKUSMUSEO
NATURHISTORISKA CENTRALMUSEET
FINNISH MUSEUM OF NATURAL HISTORY

Conservatoire Botanique National
FRANCHE-COMTÉ

UNIVERSITÄT BONN

Centre sur la biodiversité
de l'Université de Montréal

MISSOURI BOTANICAL GARDEN

Plantentuin Meise

The Morton Arboretum

Roses Loubert
www.pepinière-rosesloubert.com

arboretum KALMTHOUT

La Roseraie du Val-de-Marne

MUN Botanical Garden
Memorial University of Newfoundland

Ville de Nantes

The ARNOLD ARBORETUM of HARVARD UNIVERSITY

Conservatoire botanique national de Franche-Comté

WAGENINGEN UNIVERSITY & RESEARCH

Quarryhill BOTANICAL GARDEN

UNIVERSITÉ DE STRASBOURG

CAMBRIDGE UNIVERSITY Botanic Garden

arboretum du vallon de l'Aubonne

Sequencing

EPGV
ETUDE DU POLYMORPHISME
DES GENOMES VÉGETAUX

Bioinformatics

GenoToul Bioinfo

Funding

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'ENSEIGNEMENT SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

objectif végétal
Recherche, Formation & Innovation
en PAYS de la LOIRE

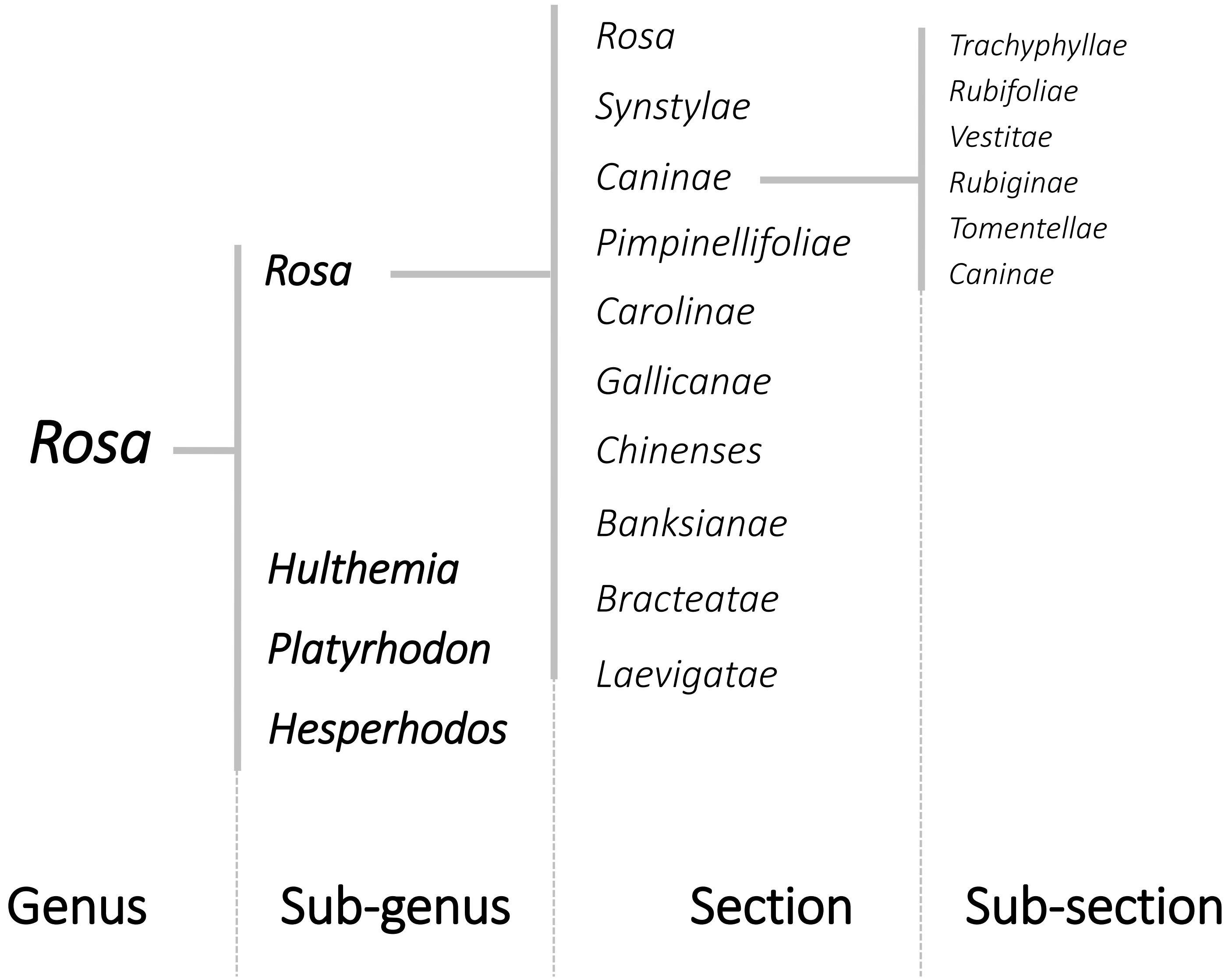
IRHS
Institut de Recherche en Horticulture et Semences

INRA
SCIENCE & IMPACT

AGRO CAMPUS OUEST

ua université angers

Background – *The genus* Rosa



Rosa

Rosa

Hulthemia

Platyrhodon

Hesperhodos

Rosa

Synstylae

Caninae

Pimpinellifoliae

Carolinae

Gallicanae

Chinenses

Banksianae

Bracteatae

Laevigatae

Trachyphyllae

Rubifoliae

Vestitae

Rubiginae

Tomentellae

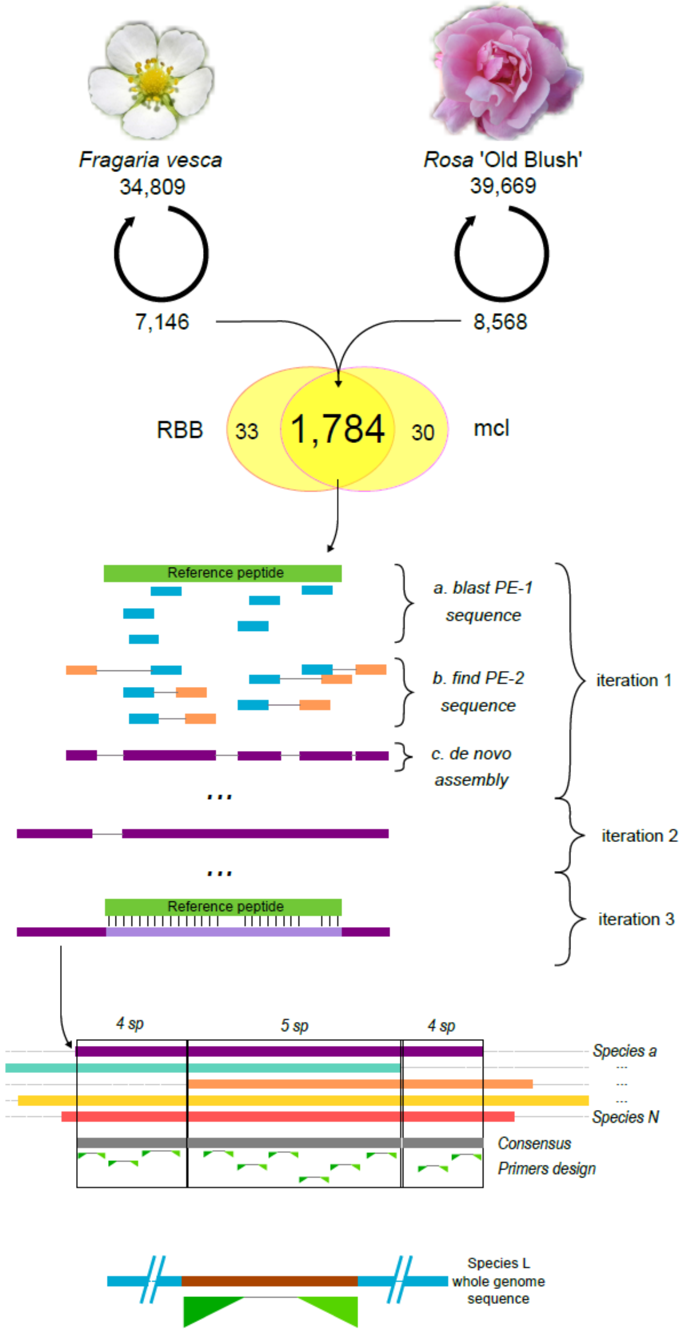
Caninae

Genus

Sub-genus

Section

Sub-section



STEP 1:

Find Single Copy Genes (SCGs) in both species

STEP 2:

Find shared Single Copy Orthologs (SCOs)

STEP 3:

Target assembly of the 1784 SCOs in **unassembled** genomes

STEP 4:

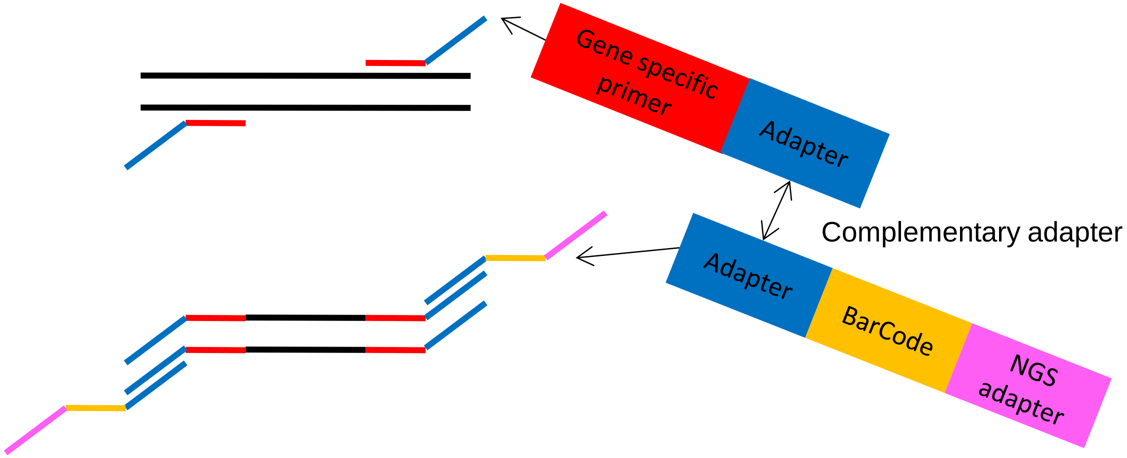
Align contigs, find blocs with ≥4 species (including *R.* 'Old Blush' and *R. persica*) and design primers on consensus sequences

STEP 5:

Check primers specificity on *R.* 'Old Blush' genome and find theoretical amplicons on **assembled** genomes

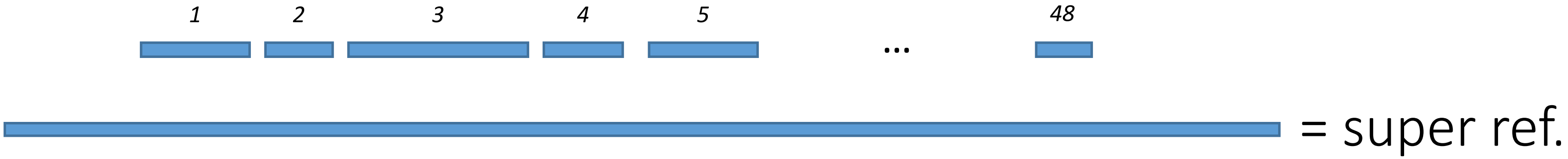
Debray *et al.* In Prep.

Methods – *The 4-primer amplification*

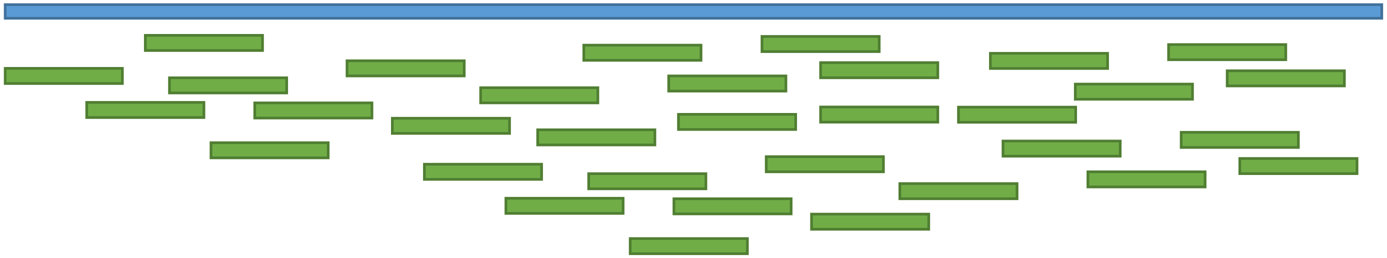


Methods – *Allelic frequencies and ploidy*

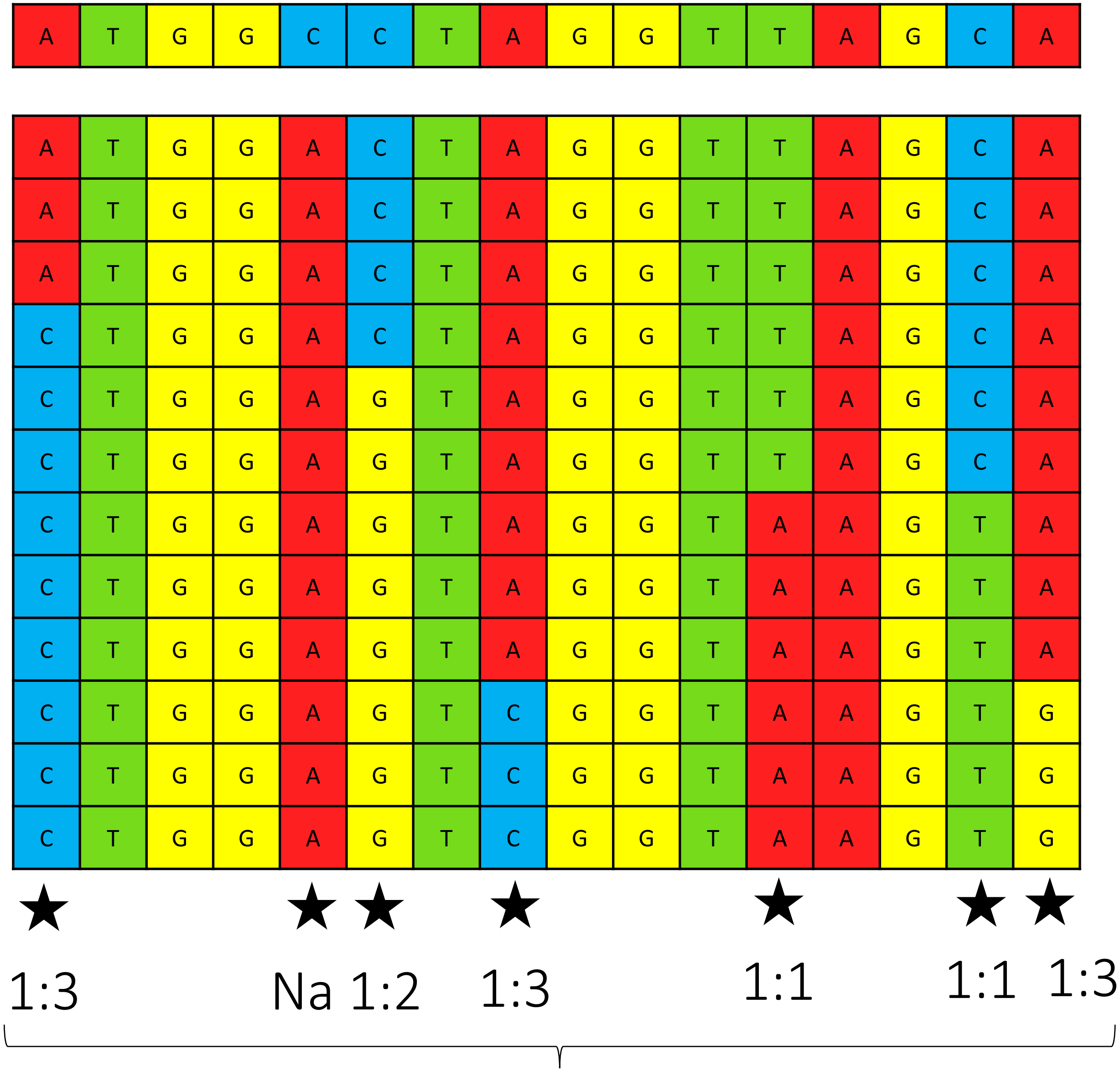
1 Concatenation of nuclear loci in the ref. hap OB



2 Mapping all reads of 1 individual



3 Count Allelic variations



A	T	G	G	C	C	T	A	G	G	T	T	A	G	C	A

= super ref.

A	T	G	G	A	C	T	A	G	G	T	T	A	G	C	A
A	T	G	G	A	C	T	A	G	G	T	T	A	G	C	A
A	T	G	G	A	C	T	A	G	G	T	T	A	G	C	A
C	T	G	G	A	C	T	A	G	G	T	T	A	G	C	A
C	T	G	G	A	G	T	A	G	G	T	T	A	G	C	A
C	T	G	G	A	G	T	A	G	G	T	T	A	G	C	A
C	T	G	G	A	G	T	A	G	G	T	A	A	G	T	A
C	T	G	G	A	G	T	A	G	G	T	A	A	G	T	A
C	T	G	G	A	G	T	A	G	G	T	A	A	G	T	A
C	T	G	G	A	G	T	C	G	G	T	A	A	G	T	G
C	T	G	G	A	G	T	C	G	G	T	A	A	G	T	G
C	T	G	G	A	G	T	C	G	G	T	A	A	G	T	G

= mapped reads

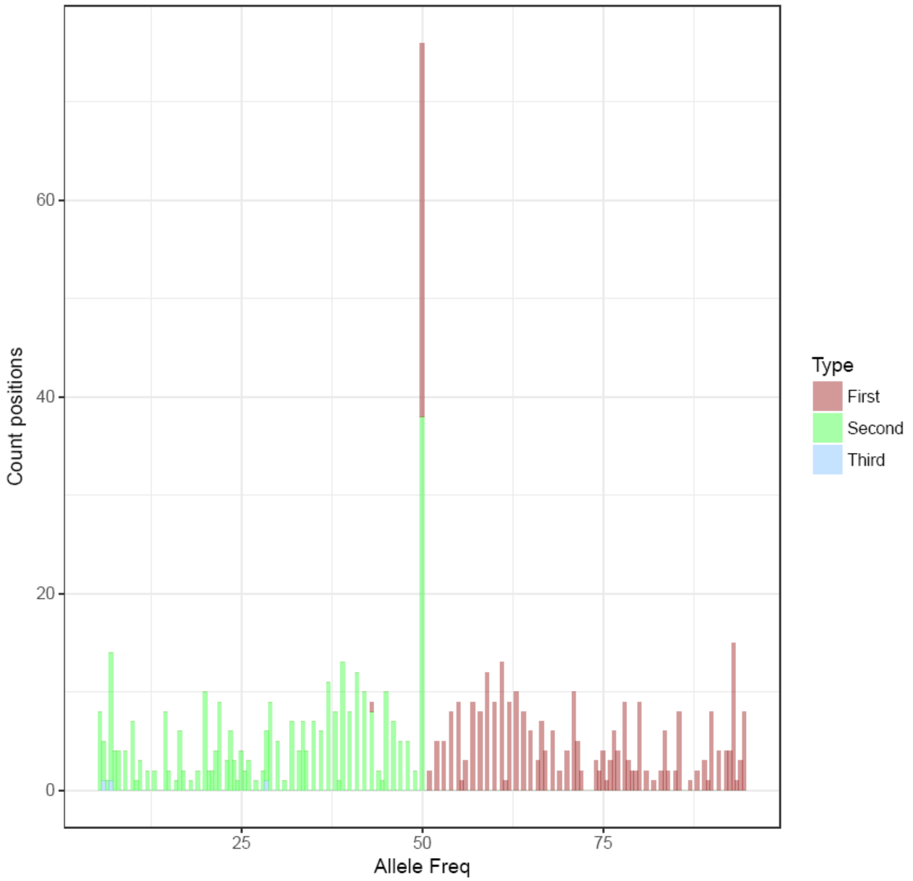
freq	Count A1
0	0
…	…
0,25	15
…	…
0,5	30
…	…
0,75	45
…	…
1	60

freq	Count A2
1	60
…	…
0,75	45
…	…
0,5	30
…	…
0,25	15
…	…
0	0

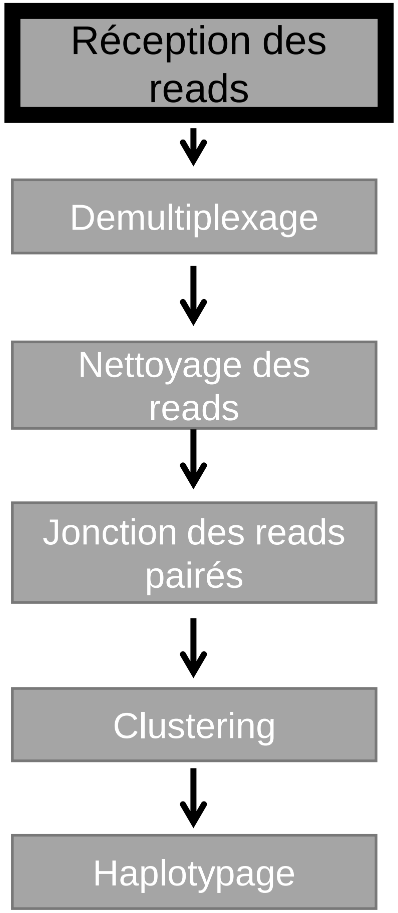
3 Count Allelic variations

freq	Count A1
0	0
…	…
0,25	15
…	…
0,5	30
…	…
0,75	45
…	…
1	60

freq	Count A2
1	60
…	…
0,75	45
…	…
0,5	30
…	…
0,25	15
…	…
0	0



Methods – *Recovering alleles from reads*



Réception des reads

Demultiplexage

Nettoyage des reads

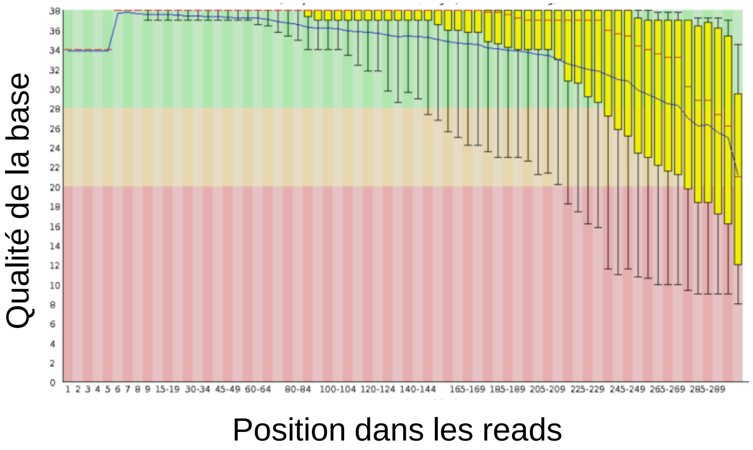
Jonction des reads pairés

Clustering

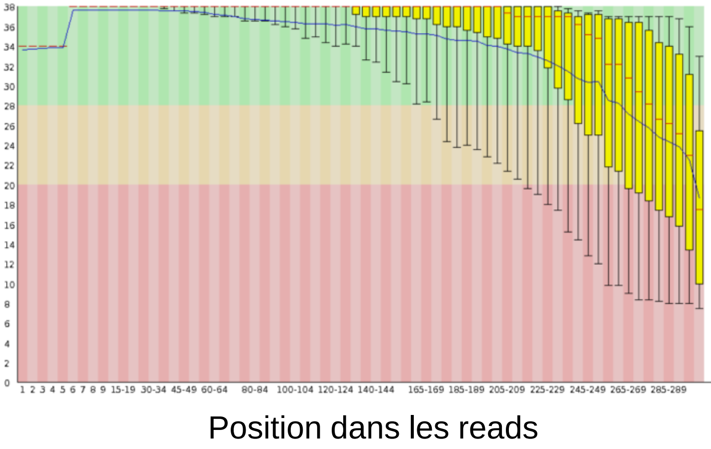
Haplotypage

Séquençage Illumina paired-end
2 x 300bp

Total : 18,876,898 reads ~ 2,7Go

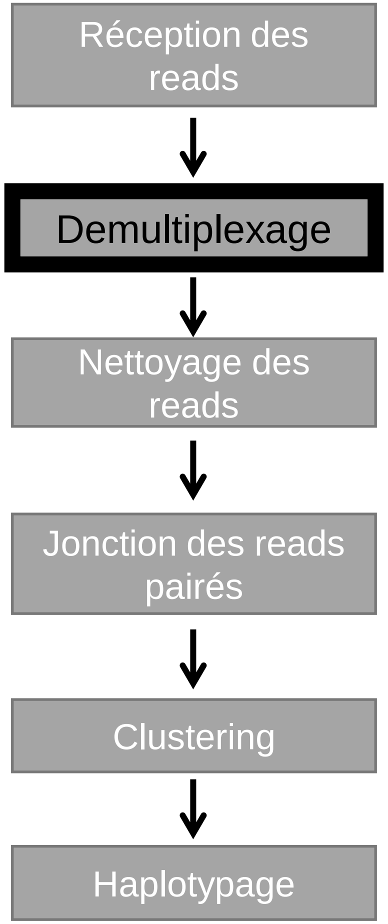


Forward



Reverse

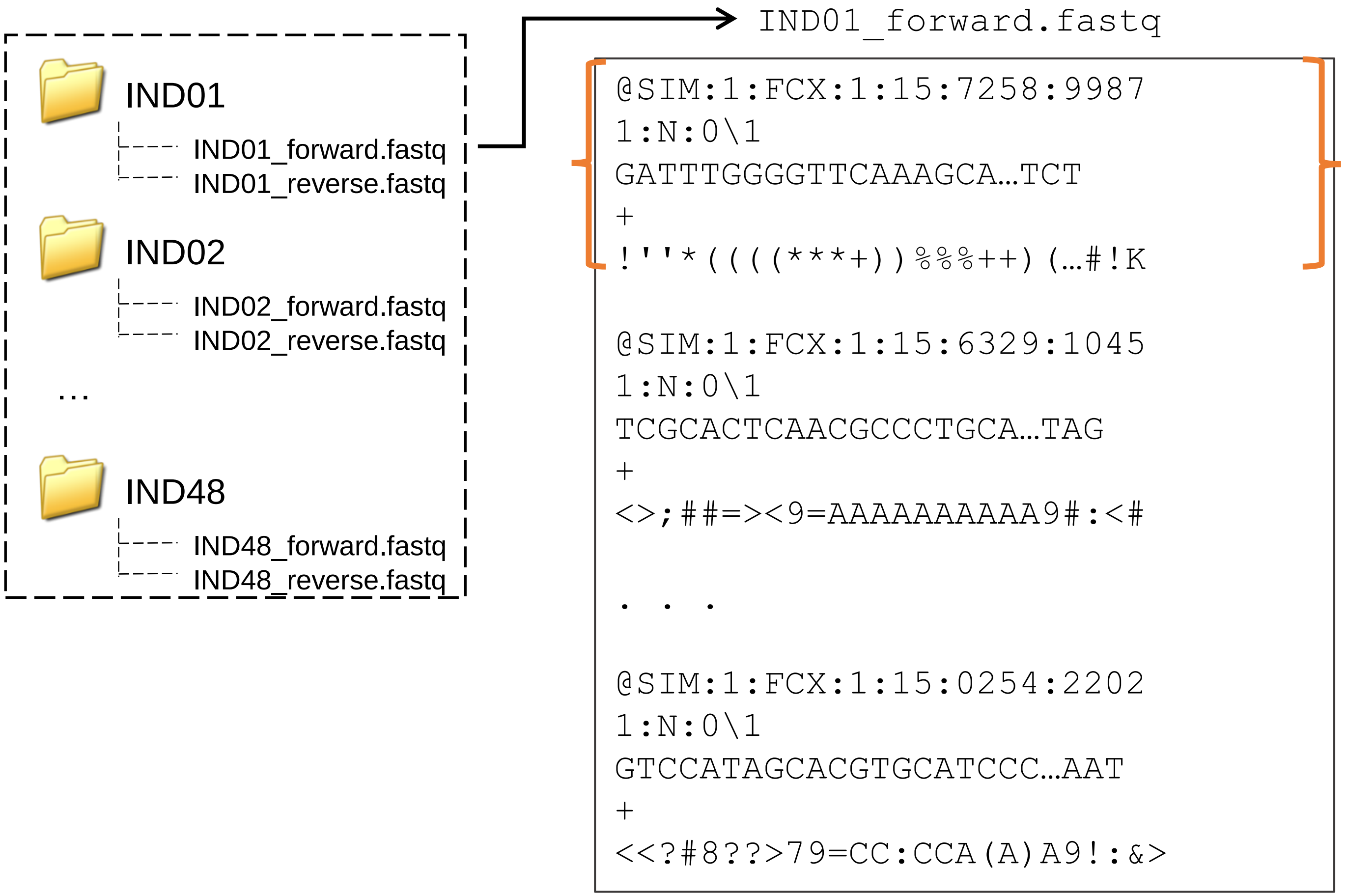
Methods – *Recovering alleles from reads*



```
IND01
├--- IND01_forward.fastq
└--- IND01_reverse.fastq
IND02
├--- IND02_forward.fastq
└--- IND02_reverse.fastq
…
IND48
├--- IND48_forward.fastq
└--- IND48_reverse.fastq
```

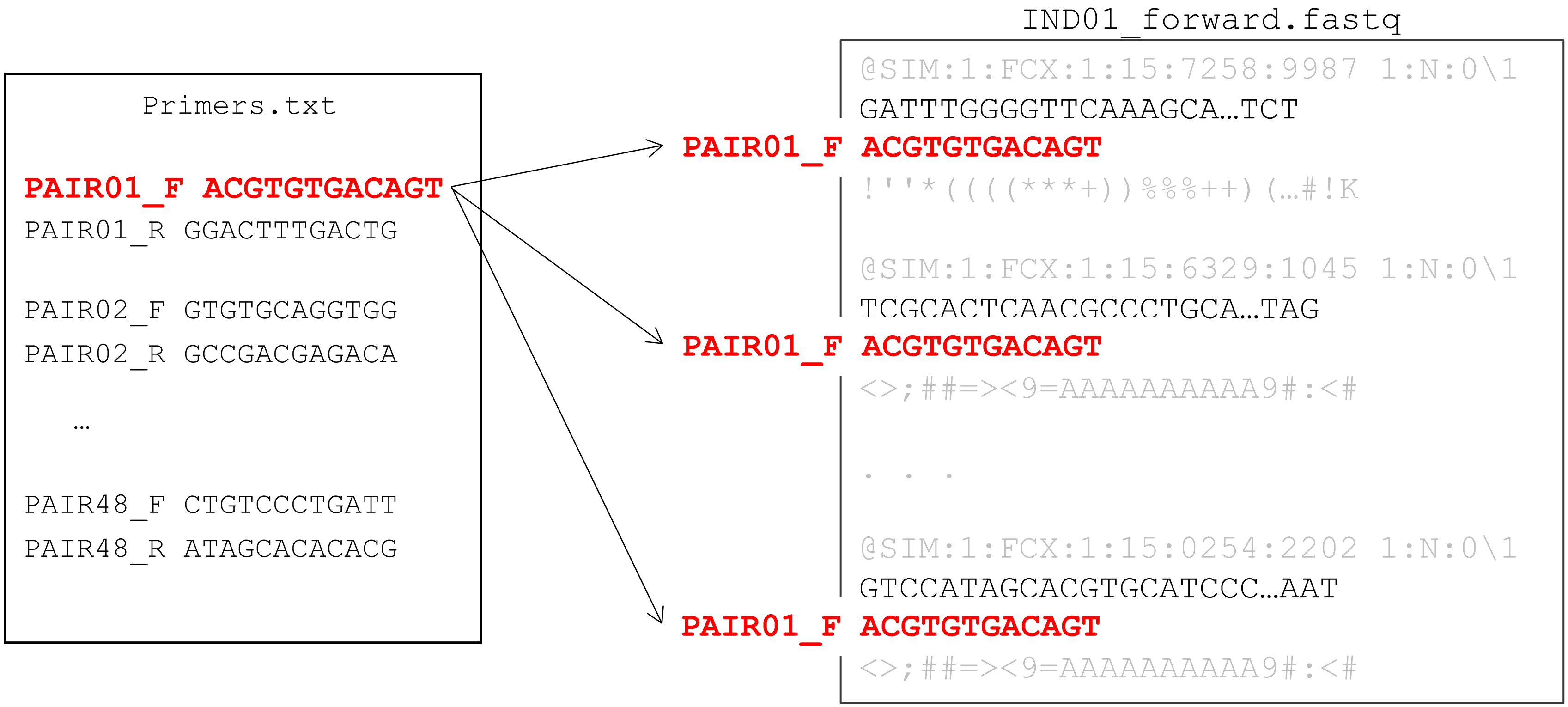
```
            Primers.txt

PAIR01_F ACGTGTGACAGT
PAIR01_R GGACTTTGACTG

PAIR02_F GTGTGCAGGTGG
PAIR02_R GCCGACGAGACA

    …

PAIR48_F CTGTCCCTGATT
PAIR48_R ATAGCACACACG
```

Methods – *Recovering alleles from reads*



```
@SIM:1:FCX:1:15:7258:9987
1:N:0\1
GATTTGGGGTTCAAAGCA…TCT
+
!''*((((***+))%%%++)(…#!K

@SIM:1:FCX:1:15:6329:1045
1:N:0\1
TCGCACTCAACGCCCTGCA…TAG
+
<>;##=><9=AAAAAAAAAA9#:<#

. . .

@SIM:1:FCX:1:15:0254:2202
1:N:0\1
GTCCATAGCACGTGCATCCC…AAT
+
<<?#8??>79=CC:CCA(A)A9!:&>
```

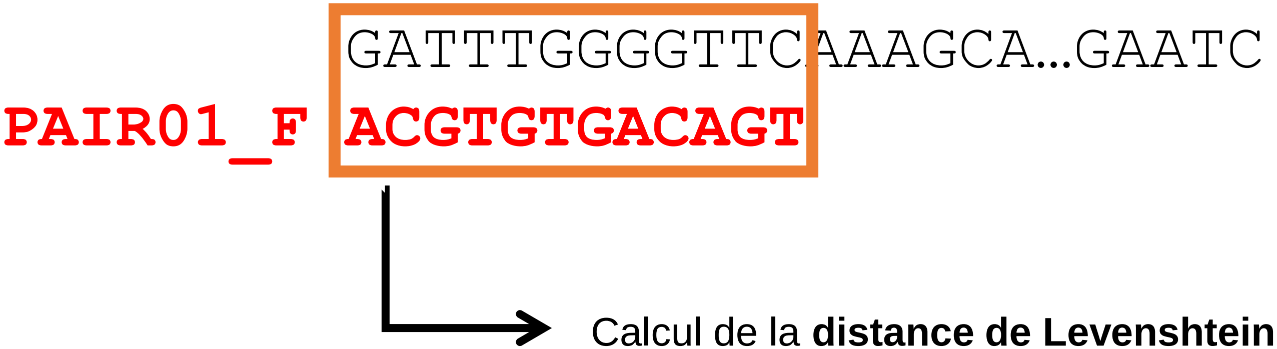
Methods – *Recovering alleles from reads*



```
Primers.txt

PAIR01_F ACGTGTGACAGT
PAIR01_R GGACTTTGACTG

PAIR02_F GTGTGCAGGTGG
PAIR02_R GCCGACGAGACA

   …

PAIR48_F CTGTCCCTGATT
PAIR48_R ATAGCACACACG
```

```
IND01_forward.fastq
@SIM:1:FCX:1:15:7258:9987 1:N:0\1
GATTTGGGGTTCAAAGCA…TCT
PAIR01_F ACGTGTGACAGT
!''*((((***+))%%%++)(…#!K

@SIM:1:FCX:1:15:6329:1045 1:N:0\1
TCGCACTCAACGCCCTGCA…TAG
PAIR01_F ACGTGTGACAGT
<>;##=><9=AAAAAAAAAA9#:<#

. . .

@SIM:1:FCX:1:15:0254:2202 1:N:0\1
GTCCATAGCACGTGCATCCC…AAT
PAIR01_F ACGTGTGACAGT
<>;##=><9=AAAAAAAAAA9#:<#
```

Methods – *Recovering alleles from reads*



GATTTGGGGTTCAAAGCA…GAATC

PAIR01_F ACGTGTGACAGT

Calcul de la **distance de Levenshtein (LD)**
entre les 2 chaînes de caractères
(ie. nombre minimal de caractères qu'il faut supprimer,
insérer ou remplacer pour passer d’une chaîne à l’autre)

Si LD ≤ 4:

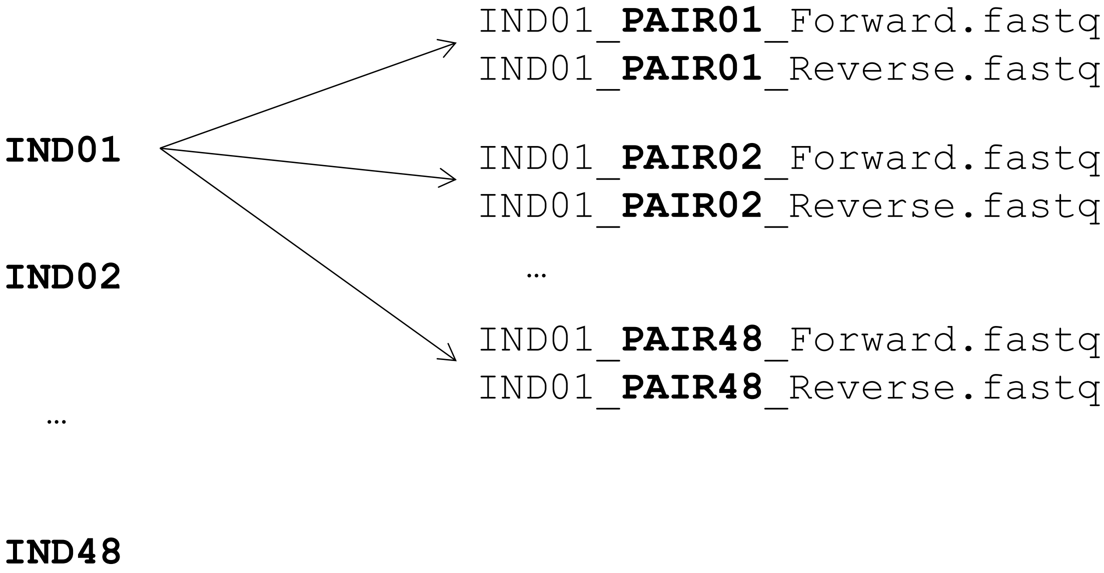
Le read est associé à la PAIR01 en partie forward

Sinon:

Le read est abandonné *(et sa paire reverse aussi…)*

Methods – *Recovering alleles from reads*

Au final j’obtiens une arborescence de fichiers de reads :



IND01

IND01_**PAIR01**_Forward.fastq
IND01_**PAIR01**_Reverse.fastq

IND01_**PAIR02**_Forward.fastq
IND01_**PAIR02**_Reverse.fastq

…

IND01_**PAIR48**_Forward.fastq
IND01_**PAIR48**_Reverse.fastq

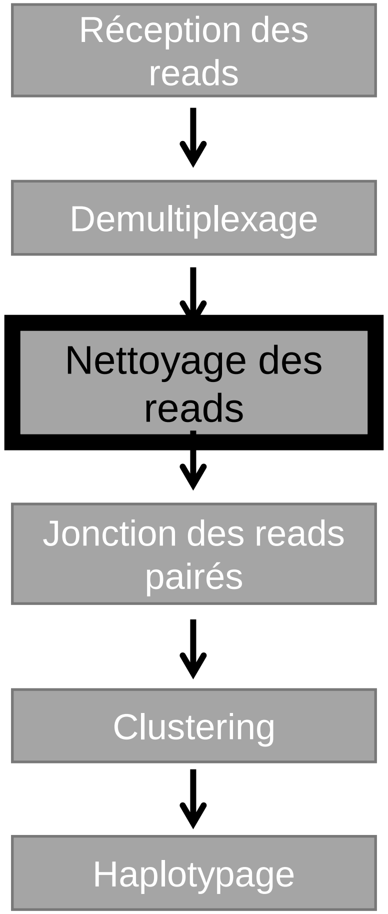
IND02

…

IND48

48 individus x 48 paires x 2 sens = 4 608 fichiers

Methods – *Recovering alleles from reads*



Réception des reads

Demultiplexage

Nettoyage des reads

Jonction des reads pairés

Clustering

Haplotypage

ENTRÉE

```
IND01_PAIR01_Forward.fastq
IND01_PAIR01_Reverse.fastq
```

- Retirer les morceaux d'adaptateurs Illumina
- Retirer les fenêtres de 4bp dont la moyenne de qualité est < 20 (99% de confiance dans l'identification des bases)
- Retirer les reads nettoyés < 30 bp

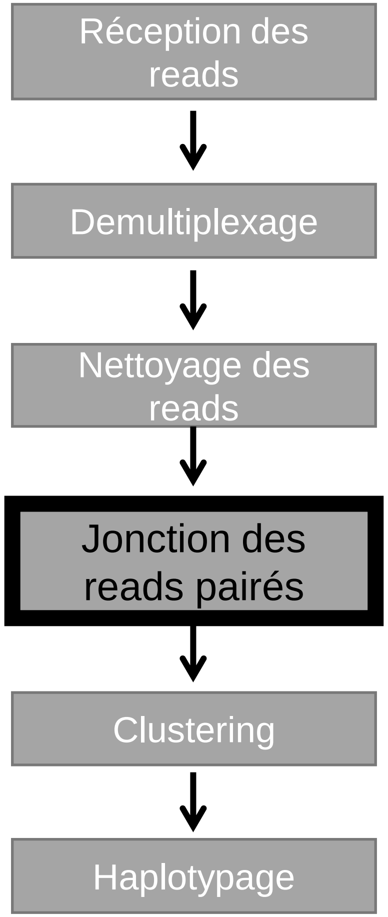
SORTIE

```
IND01_PAIR01_1P.fastq     IND01_PAIR01_1U.fastq
IND01_PAIR01_2P.fastq     IND01_PAIR01_2U.fastq
```

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Bolger *et al*. 2014, *Bioinformatics* 30(15):2114-2120

Methods – *Recovering alleles from reads*



Réception des reads

Demultiplexage

Nettoyage des reads

Jonction des reads pairés

Clustering

Haplotypage

```
IND01_PAIR01_1P.fastq

@read25486\1
CACCACATATGCTGTCTCTGGCAC
+
<>;##=><9=AAAAAAAAAA9#:<

@read12579\1
…
```

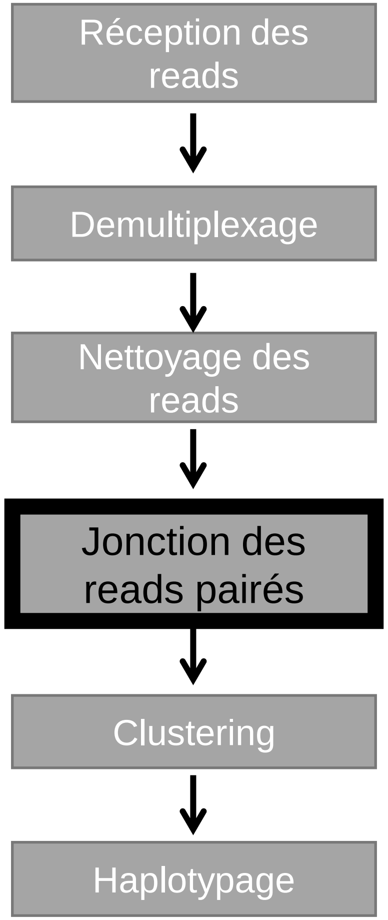
```
IND01_PAIR01_2P.fastq

@read25486\2
GGTTTAGAGGAATCAGATTCAAGT
+
;??#A=>C9!<<>()/,*-9#;;A

@read12579\2
…
```

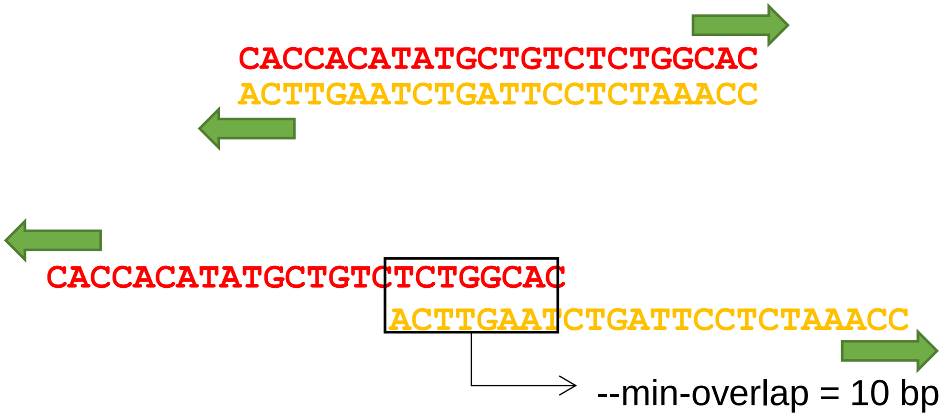
```
CACCACATATGCTGTCTCTGGCAC
```

```
GGTTTAGAGGAATCAGATTCAAGT
  ↓ Reverse complement
ACTTGAATCTGATTCCTCTAAACC
```

```
CACCACATATGCTGTCTCTGGCAC
                ACTTGAATCTGATTCCTCTAAACC
```

Methods – *Recovering alleles from reads*



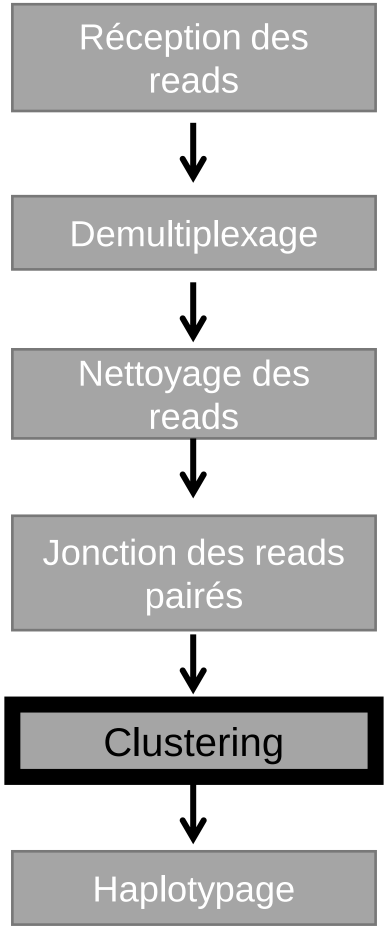


--min-overlap = 10 bp
--max-mismatch-density = 25%
Aucun gap

IND01_PAIR01_**combined**.fastq
IND01_PAIR01_**Notcombined**.fastq *(‘NNNNN’ insertion)*

FLASH (Fast Length Adjustment of SHort reads)
Magoc and Salzberg 2011, *Bioinformatics* 27(21):2957-2963

Methods – *Recovering alleles from reads*



`IND01_PAIR01_`**`combined`**`.fastq`
`IND01_PAIR01_`**`Notcombined`**`.fastq`

Processus itératif (3 répétitions) :

- Détection des chimères de PCR par alignement
- Regroupement des reads similaires avec un seuil croissant d'identité au cours des itérations

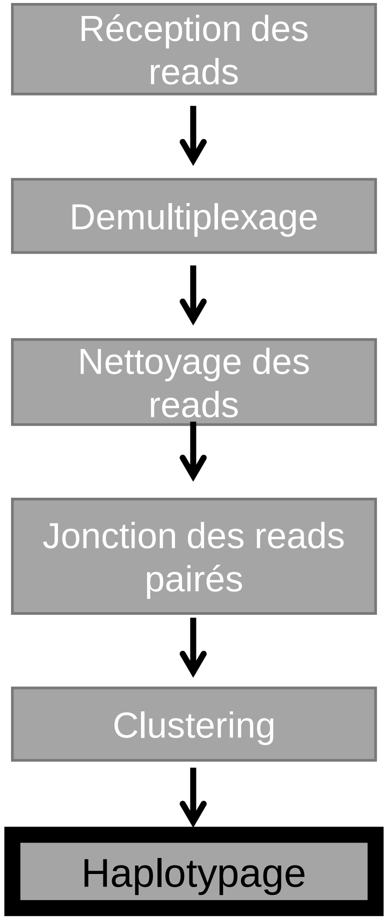
Pour chaque locus, pour chaque taxon :

Cluster 1 : 925 séquences
Cluster 2 : 854 séquences
Cluster 3 : 75 séquences
Cluster N : 2 séquences

`IND01_PAIR01_`**`PURC_clusters`**`.fastq`

Pipeline for Untangling Reticulate Complexes (PURC)
Rothfels et al. 2016, *New Phytologist* 213(1):413-429

Methods – *Recovering alleles from reads*



IND01_PAIR01_**PURC_clusters**.fastq

Contient tous les haplotypes possibles. Leur nombre est surestimé en raison des insertions de NNNN.

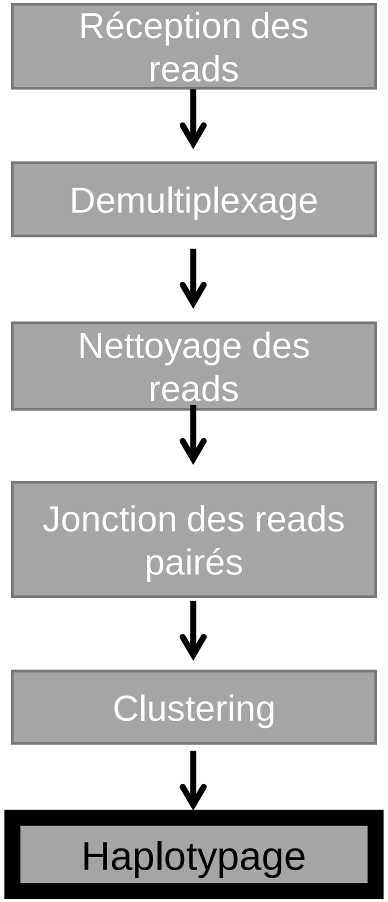
- Regrouper les cluster qui sont identiques (en ignorant les gaps)
- Déduire des haplotypes dans un contexte où la ploidïe n’est pas connue

IND01_PAIR01_**PURC_clusters_reduced**.fastq

Fluidigm2PURC

Blischak *et al.* 2018, *BioRXiv*

Methods – *Recovering alleles from reads*



Réception des reads

Demultiplexage

Nettoyage des reads

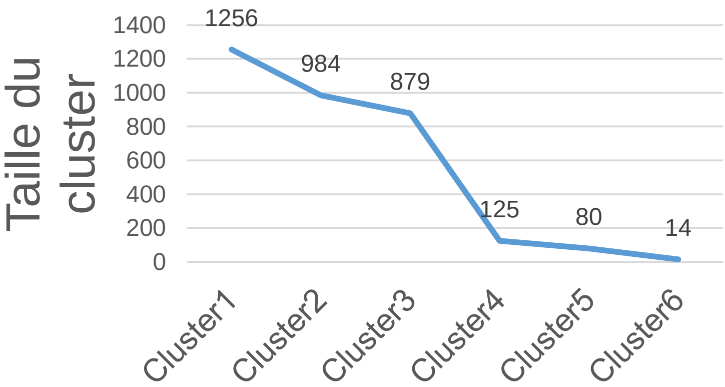
Jonction des reads pairés

Clustering

Haplotypage

Exemple : L’étape de Clustering identifie 6 clusters chez un individu 4x

On utilise comme variable la taille des clusters et comme paramètre le taux d’erreur de séquençage au locus.

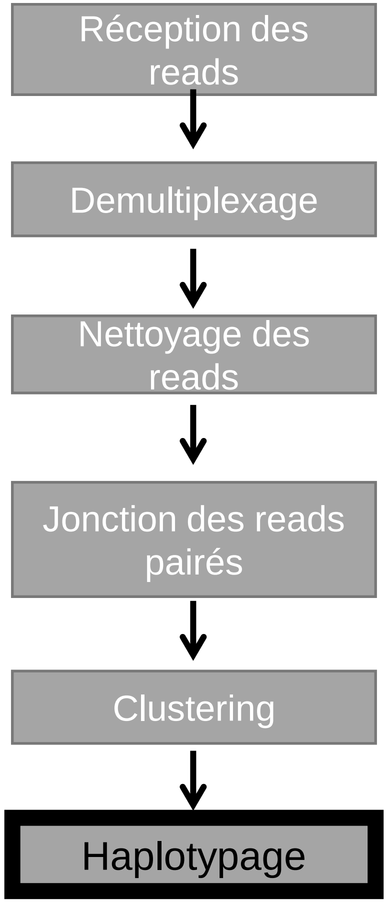


On modélise différents cas de figure :
Model1 : 0 0 0 0 0 0
Model2 : 1 0 0 0 0 0
Model3 : 1 1 0 0 0 0
Model4 : 1 1 1 0 0 0
Model5 : 1 1 1 1 0 0
Model6 : 1 1 1 1 1 0
Model7 : 1 1 1 1 1 1

Pour chaque modèle, on calcule son **maximum de vraisemblance** et on regarde si le gain obtenu est intéressant

Fluidigm2PURC
Blischak *et al.* 2018, *BioRXiv*

Methods – *Recovering alleles from reads*



Maximum de vraisemblance pour un modèle à H clusters :

Nb de clusters envisagés

Taille du cluster i

variable

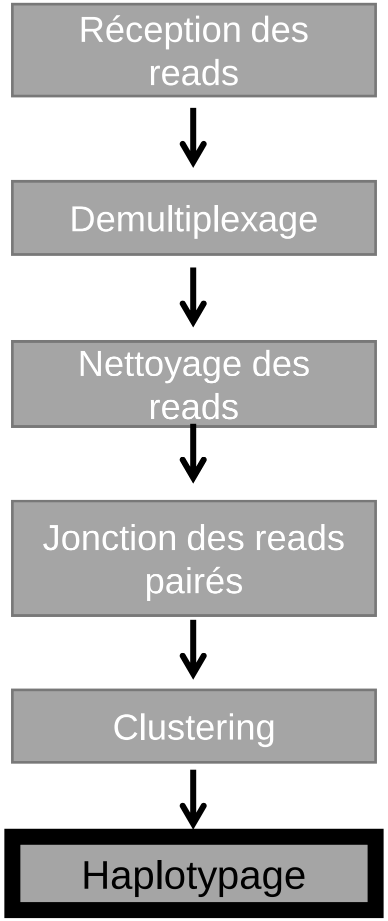
Niveau moyen d'erreur de tous les reads à ce locus

paramètre

$$l_H = \underbrace{\sum_{i=0}^{H} C_i \times \log(1-\epsilon)}_{\text{Les clusters de 0 à H sont de vrais clusters}} + \underbrace{\sum_{j>H}^{N} C_j \times \log(\epsilon)}_{\text{Les clusters suivants (>H) sont des erreurs}}$$

Fluidigm2PURC
Blischak *et al.* 2018, *BioRXiv*

Methods – *Recovering alleles from reads*



Réception des reads

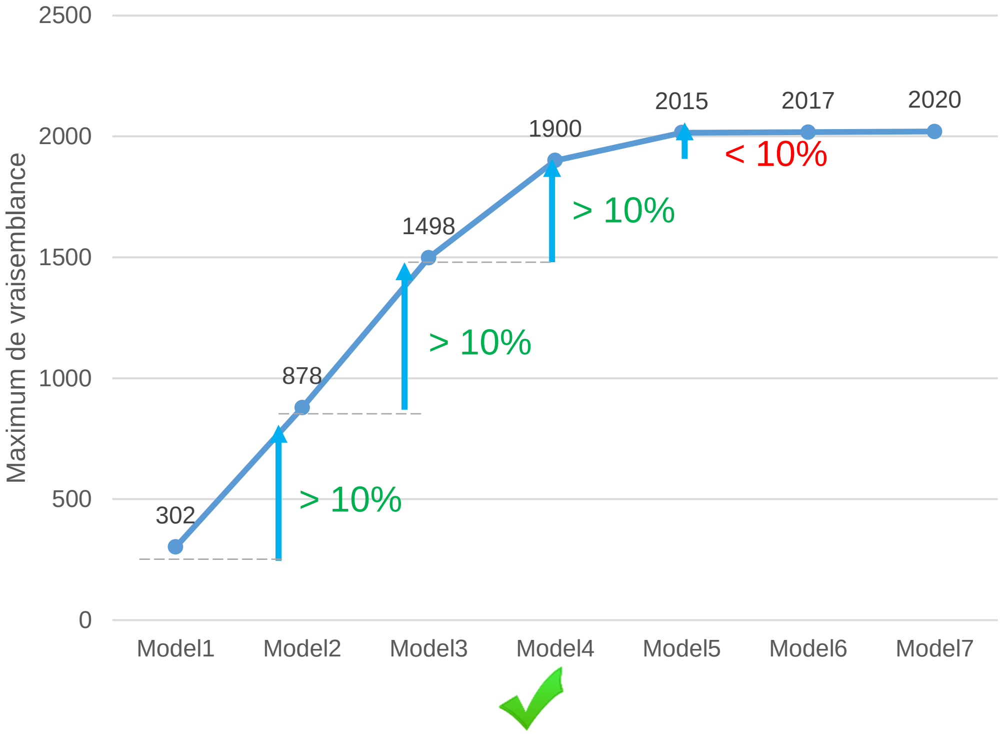
Demultiplexage

Nettoyage des reads

Jonction des reads pairés

Clustering

Haplotypage



On retient le Model4 (à 3 vrais clusters)

Fluidigm2PURC
Blischak *et al.* 2018, *BioRXiv*