A targeted sequencing approach to capture the resistance gene clusters diversity in melon

Javier Belinchon-Moreno

A targeted sequencing approach to capture the resistance gene clusters diversity in melon

Javier Belinchon-Moreno1,2, Aurelie Berard1, Aurelie Canaguier1, Isabelle Le-Clainche1, Véronique Chovelon2, Jacques Lagnel2, Vincent Rittener-Ruff2, Massyly Lebbat1, Damien Hinsinger1, Nathalie Boissot2, Patricia Faivre-Rampant1

Plant and Animal Genomics Conference, San Diego, CA, USA

Cucurbi-genomics session

12th January 2023

NLR resistance genes

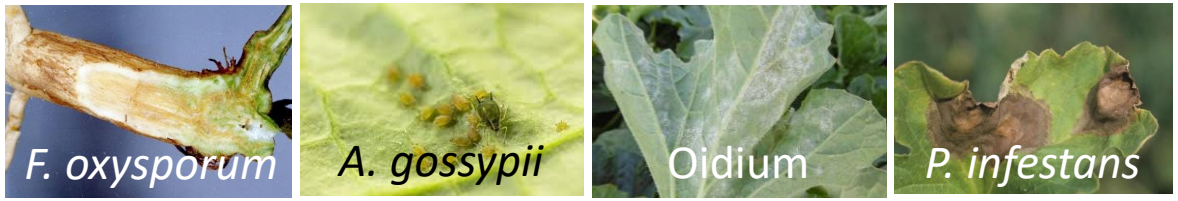
Nucleotide binding site (NB) -leucine rich repeat (LRR)

Adapted from (Gottin et al., 2021)

Intracellular immune receptor proteins

Major family of plant-resistance (R) genes

Broad range of **resistance**



Usually grouped into **clusters**

High level of presence/absence polymorphisms (PAV)

Low frequency in the Cucurbitaceae family

81 NLR genes identified in DHL95

45% of them grouped into 9 clusters

Vat cluster in melon

Most studied NLR cluster in melon

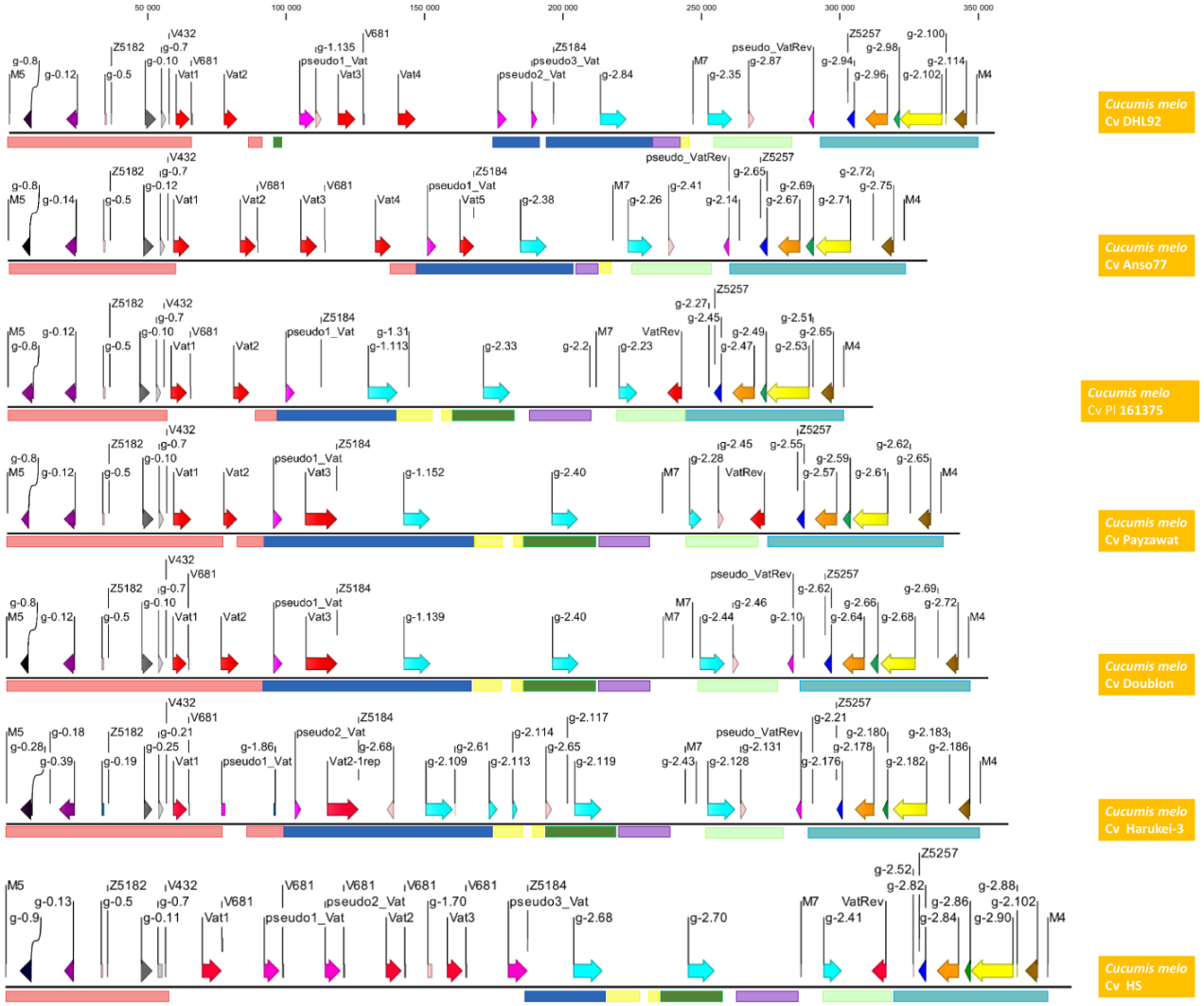
Region of **1Mb** on **chromosome 5**

(focus on 350Kb – M4M5)

Highest NLR gene density (23 genes by Gonzalez et al., 2014)

Highest PAV polymorphism density

Presence of TE inside the genes (pseudogenes)



(Chovelon *et al.*, 2021)

NLR clusters: a challenging characterization

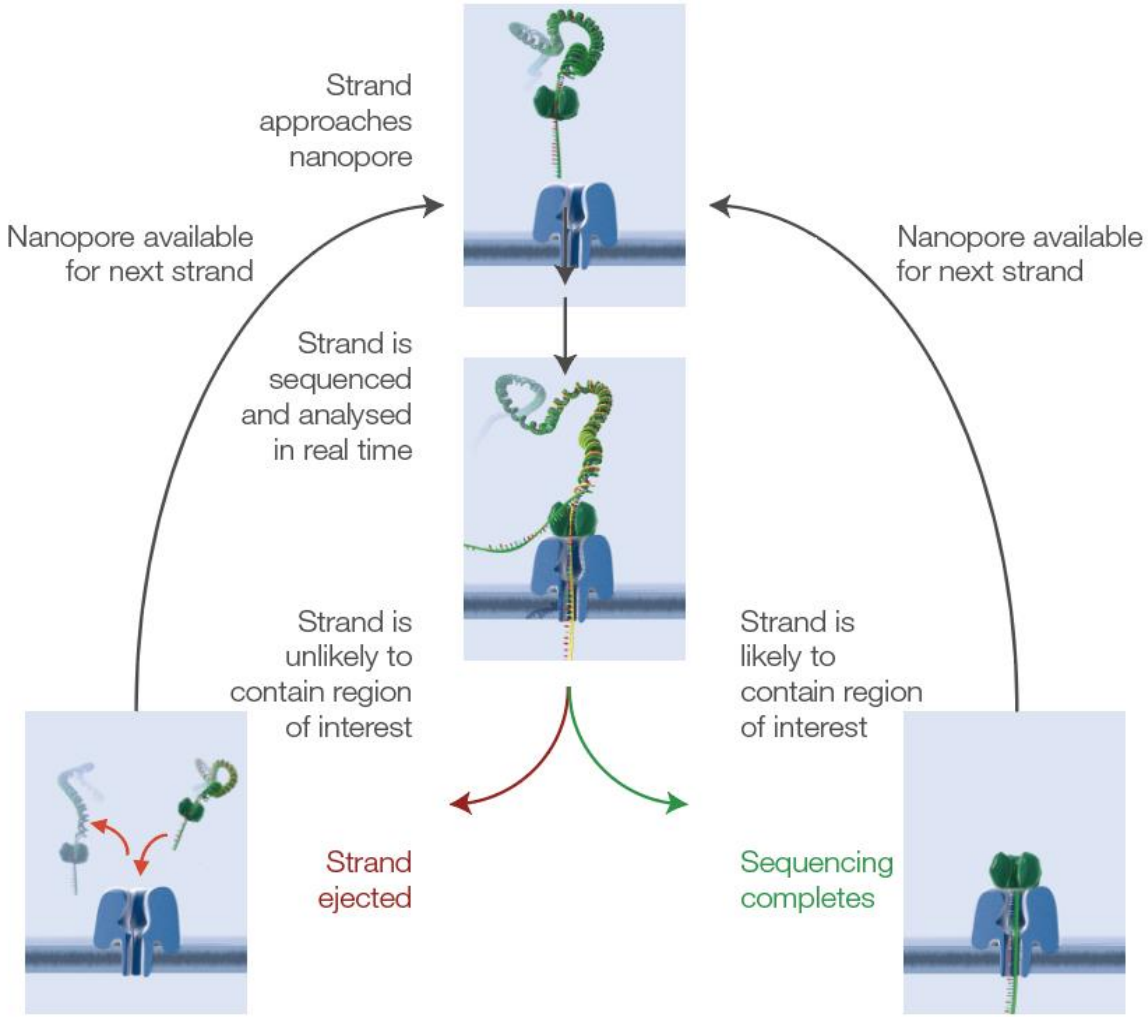
- Specific role of each NLR remains largely unknown
- Characterize the NLR-type resistance in melon ⇢ Contribution for sustainable agriculture
- Single reference genome cannot represent the full diversity of a species ⇢ Construction of **NLRome**
- Big challenge ⇢ **Complex gene structure** (repetitions)
  Grouped into **gene clusters,** with plenty of **repetitive elements** (TE)
  Prone to **duplication** and **transposition**

Targeted sequencing ⇐

- Sequencing using short reads are **ineffective**
- Genotyping approaches do **not provide reliable information**
- WGS using long reads may be so **expensive (wetlab, info & bioinfo)**

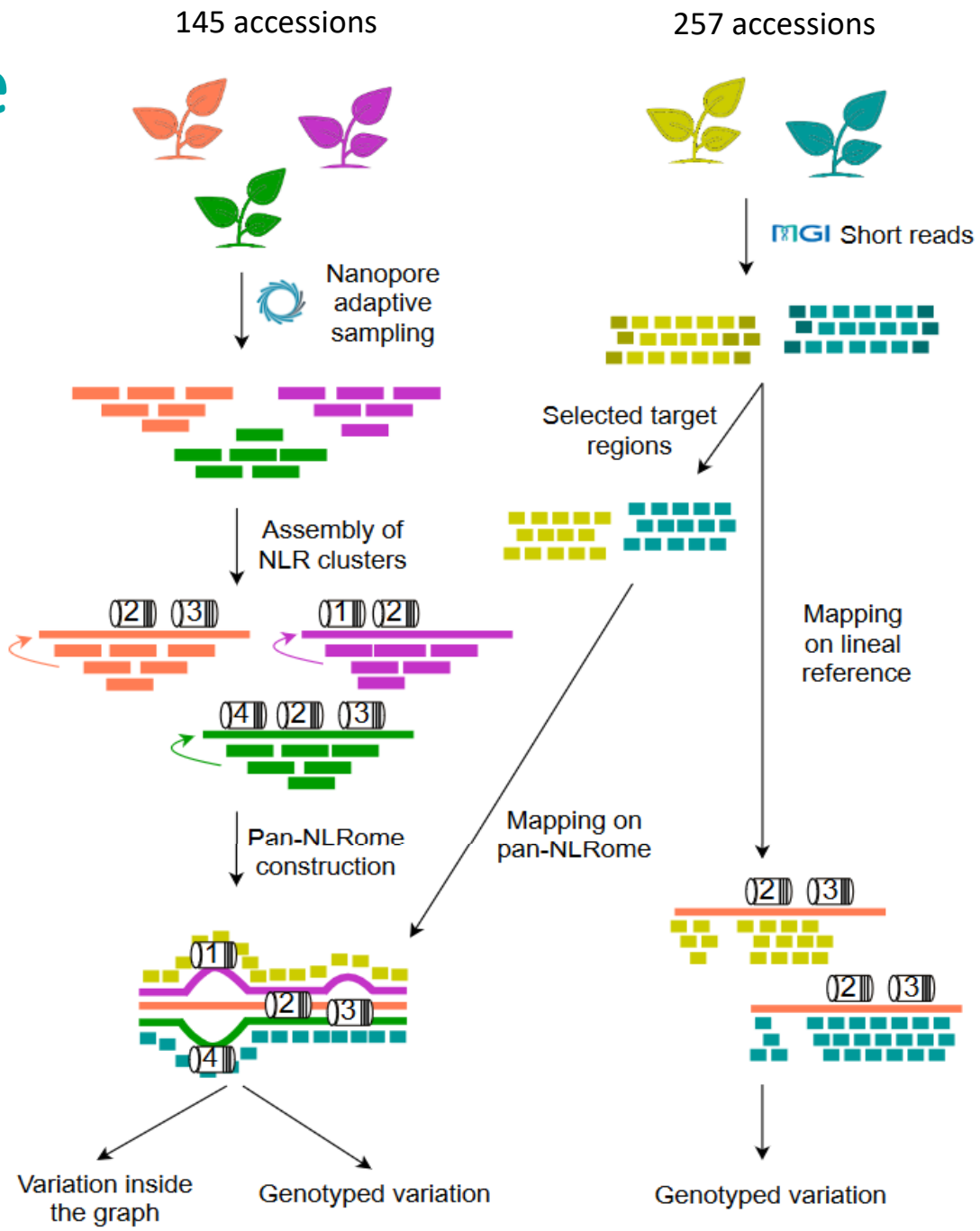
Nanopore adaptive sampling

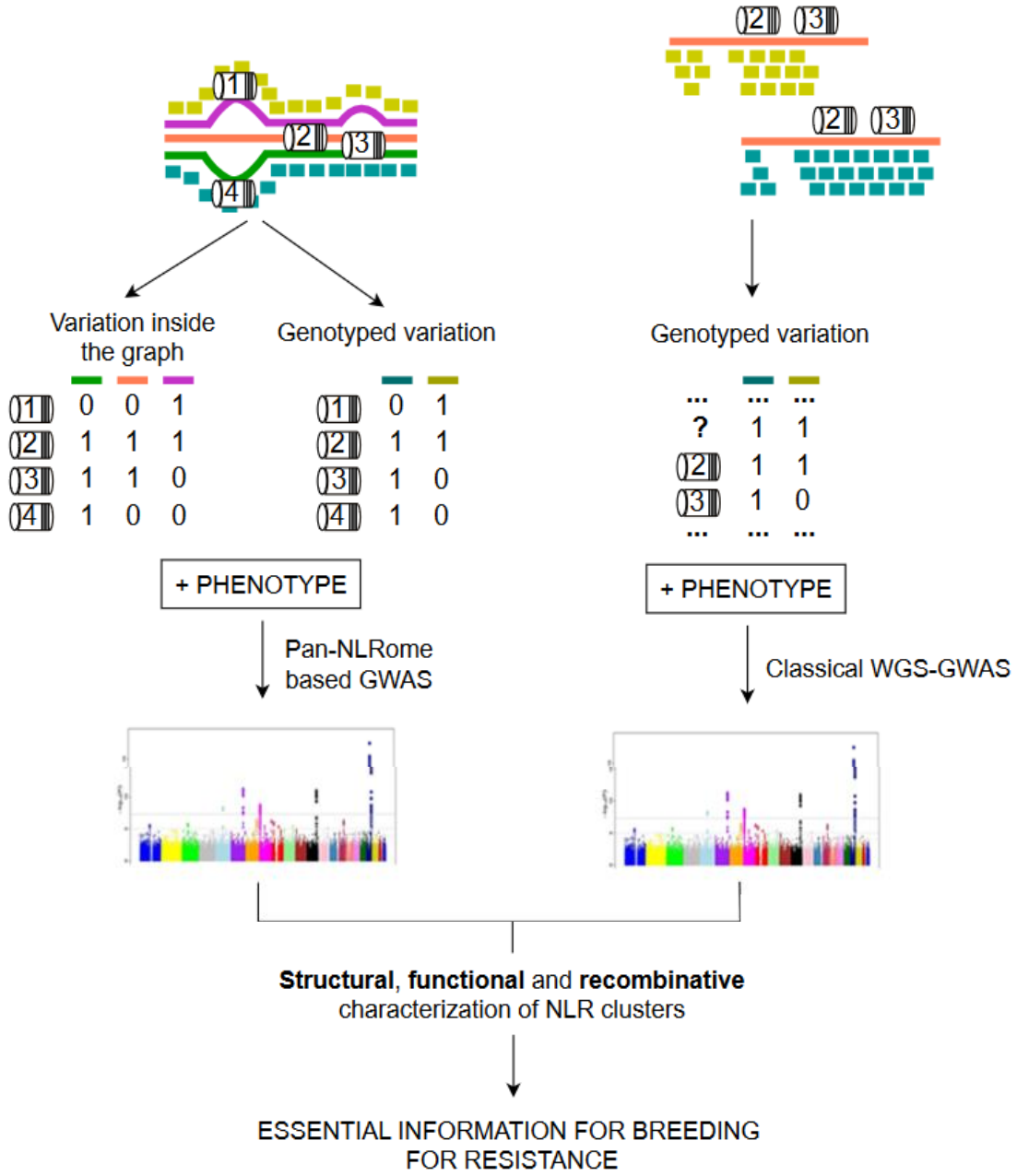
- **Real time** target selection: **easy** library preparation.
- Accept/reject molecules based on **small initial part** of sequence (~800 bp)
- **Enrich/deplete** specific regions of interest.
- Increase on-target data, reduce time-to-answer.
- No need of DNA amplification
- No need of laborious or expensive experimental design



(Nanopore London Callings, 2022)

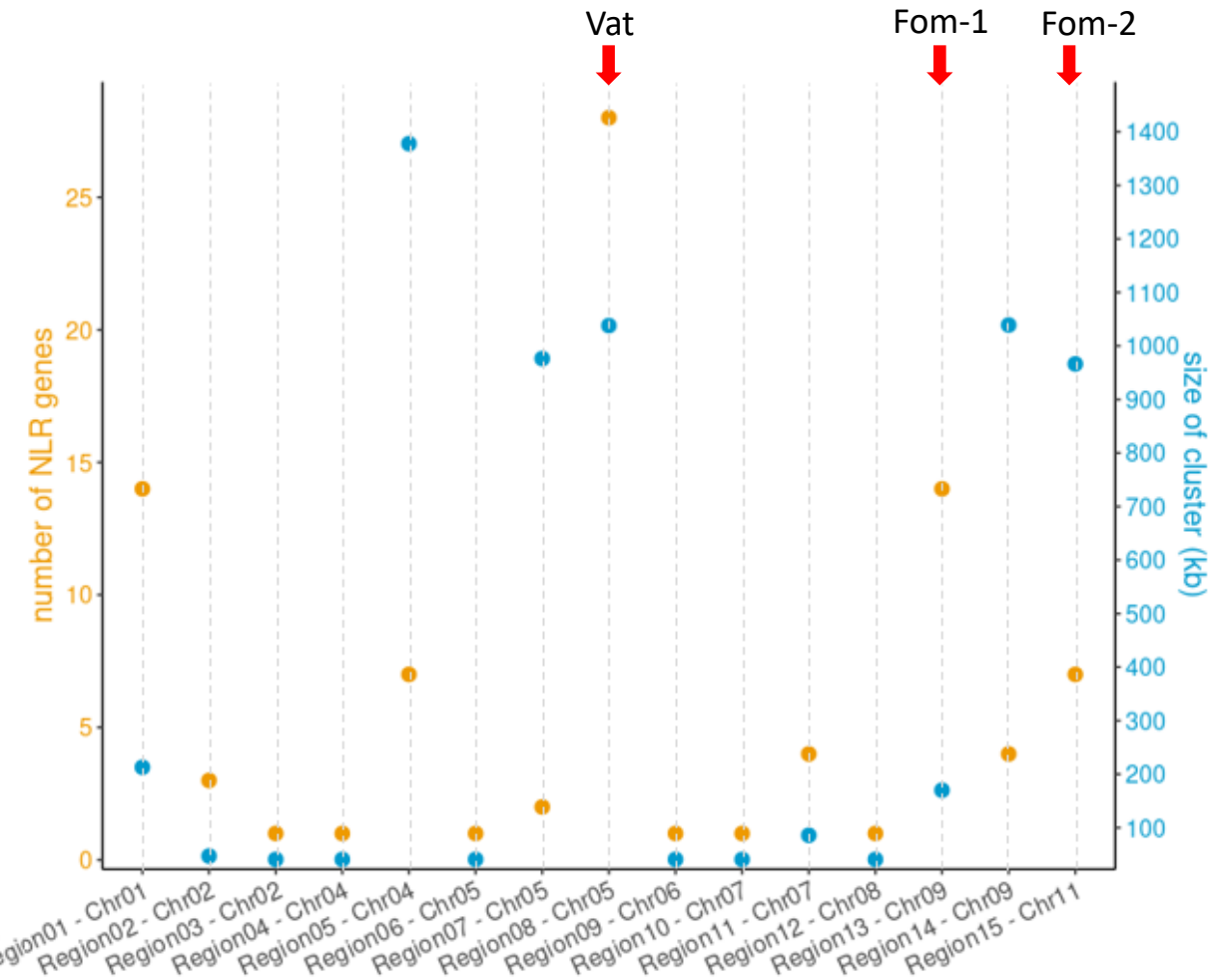
Our objective




1
2
3
4
Variation inside the graph
Genotyped variation
0 0 1
1 1 1
1 1 0
1 0 0
0 1
1 1
1 0
1 0
+ PHENOTYPE
Pan-NLRome based GWAS
Genotyped variation
... ...
? 1 1
1 1
1 0
... ...
+ PHENOTYPE
Classical WGS-GWAS
Structural, functional and recombinative characterization of NLR clusters
ESSENTIAL INFORMATION FOR BREEDING FOR RESISTANCE

Nanopore adaptive sampling: our target regions

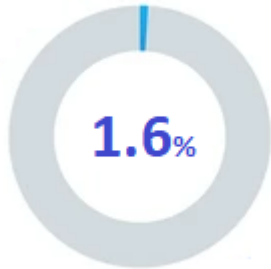
Reference genome → Variety **Anso77** (draft genome produced at INRAE-GAFL/EPGV)



Clusters:
NLR genes + 20kb-extra flanking
Variable in size
Variable in number of genes

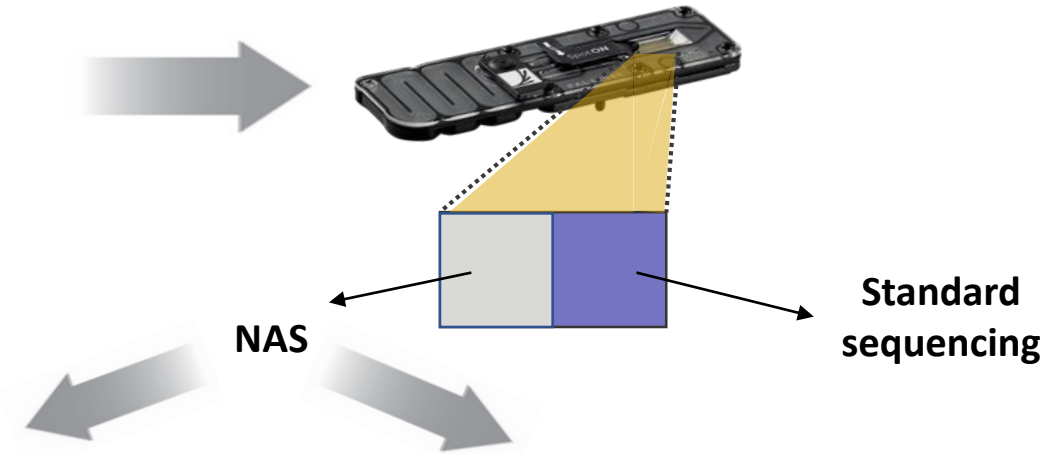
NLRGenomeSweeper (Toda *et al.*, 2020)

Genome length ~380Mb
Target regions length ~6.16Mb

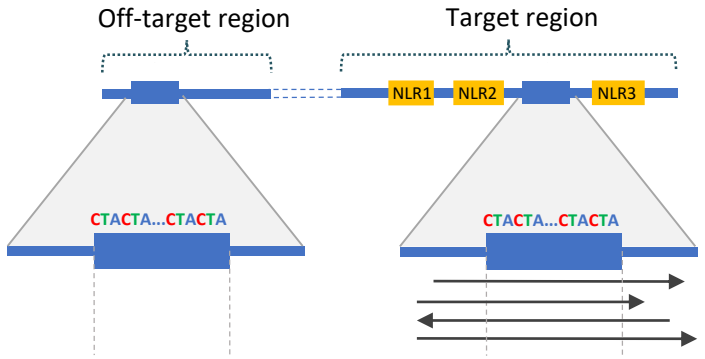


Our experiences

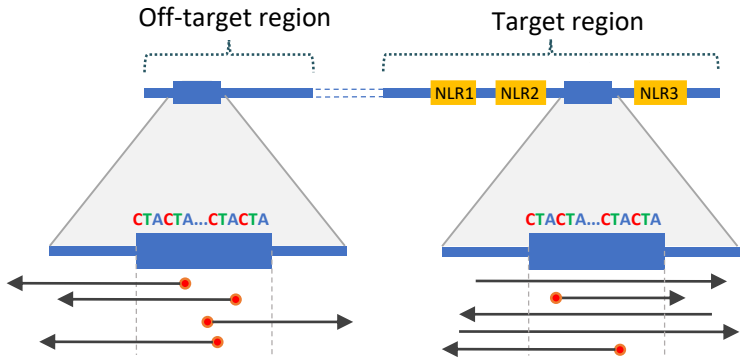
Testing on ANSO-77 (same as provided reference)



<u>Masking</u> repetitive elements >200kb



<u>Not</u> masking repetitive elements

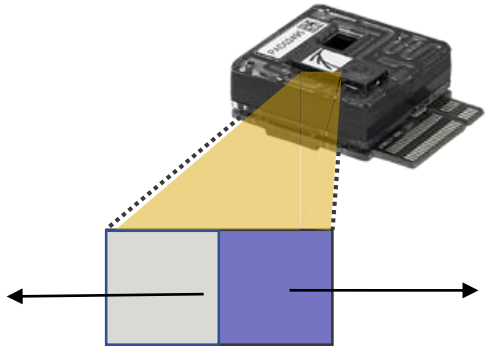


Our experiences

From ANSO-77 (same as provided reference) to more genetically distant accessions

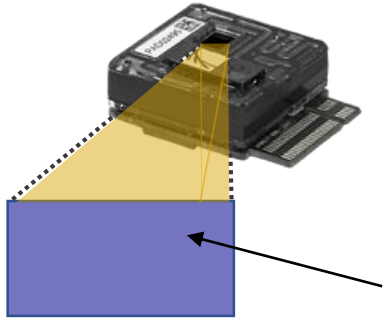
Anso77 + Doublon

Standard sequencing



NAS

Multiplexing



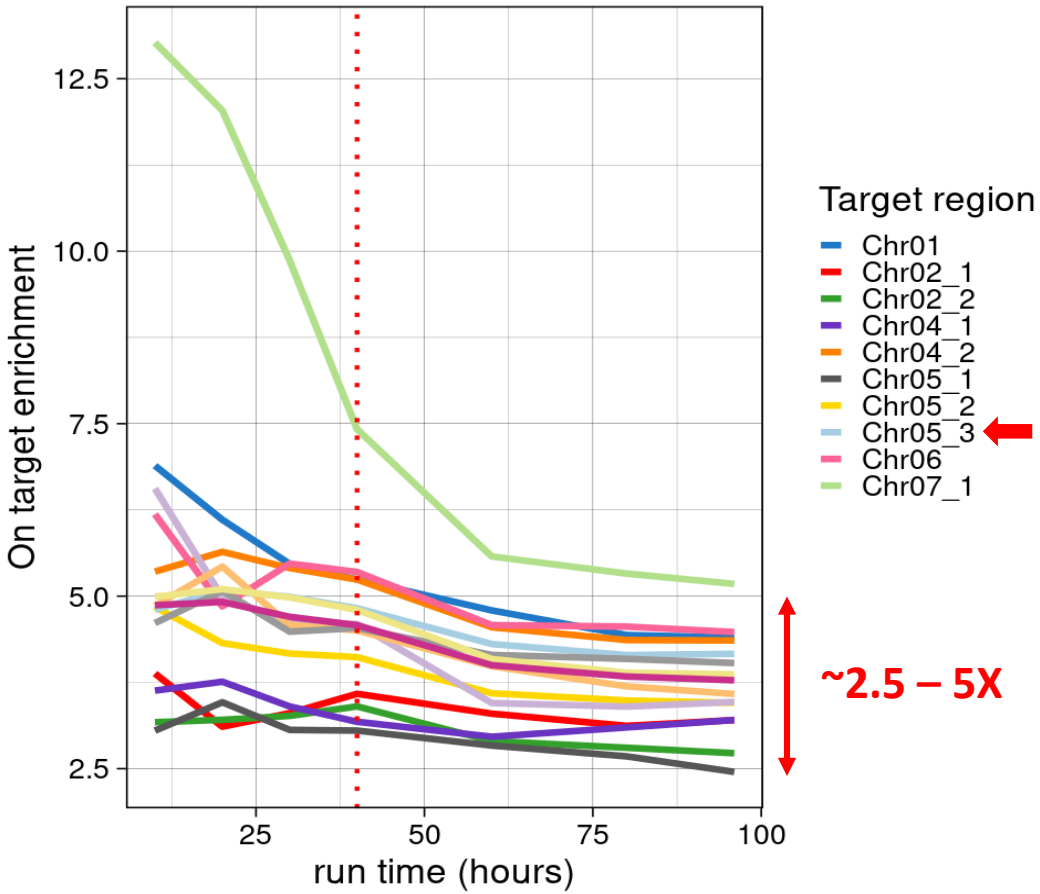
NAS

Up to 10 accessions/flowcell → 145 accessions already sequenced and assembled

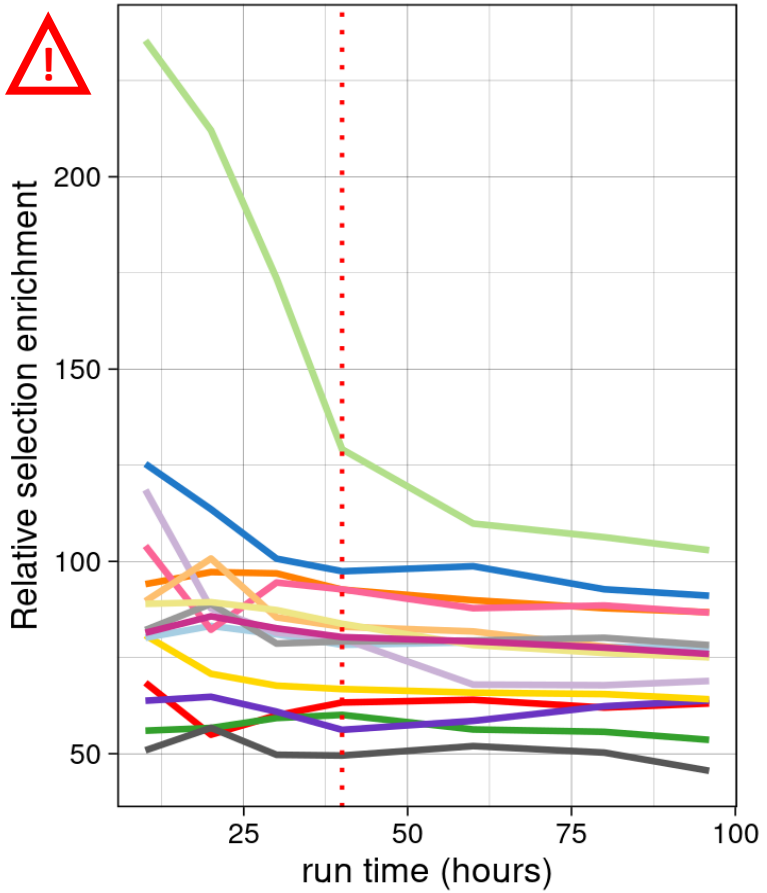
Good enrichment of our target regions

Effectively reducing the off-target volume of data

ANSO-77
Standard sequencing



ANSO-77
Adaptive sampling



$$OTE = \frac{region\_cov\_NAS}{region\_cov\_standard}$$

$$RSE = \frac{\frac{region\_cov\_NAS}{chrom\_cov\_NAS}}{\frac{region\_cov\_standard}{chrom\_cov\_standard}}$$

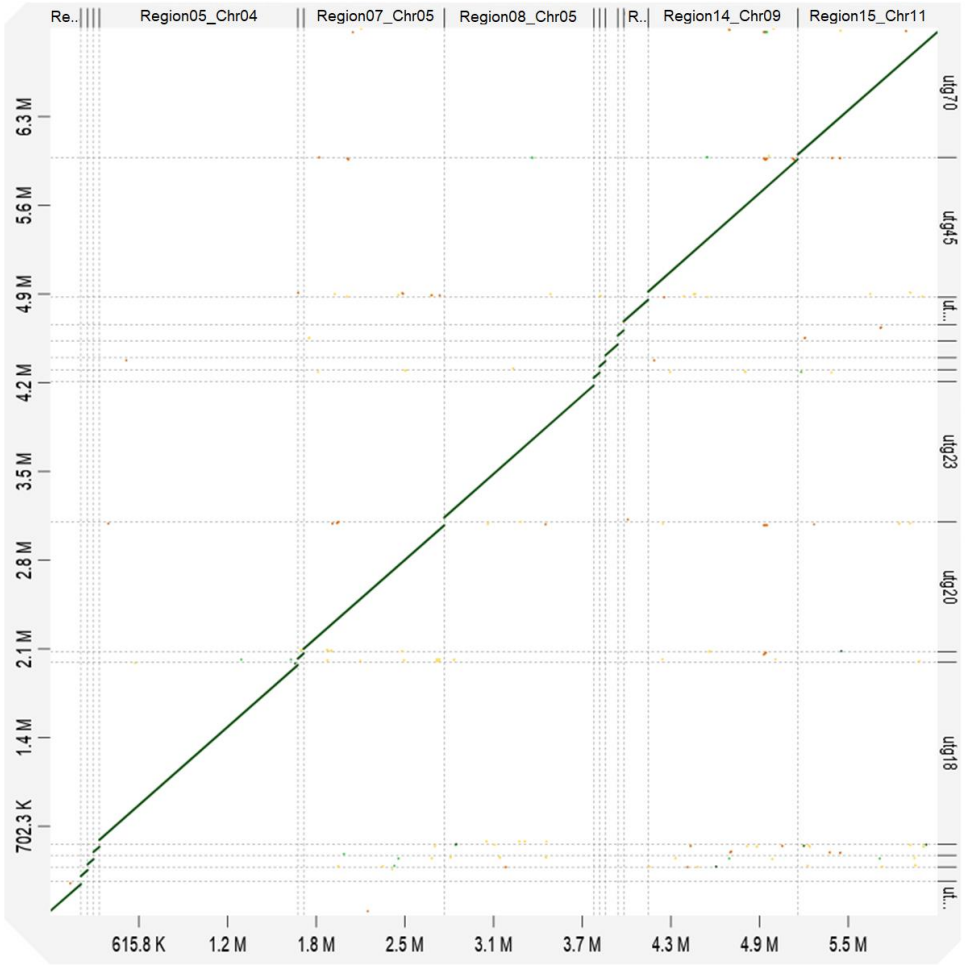
Variable between clusters

Very high in average

Assembly of 15 target regions of Anso77

Complete and continuous assembly

Anso-77 reference: 15 gene clusters



Anso-77 assembly of regions of interest

ANSO-77 – inodorus (ssp. melo)

Target regions fully assembled

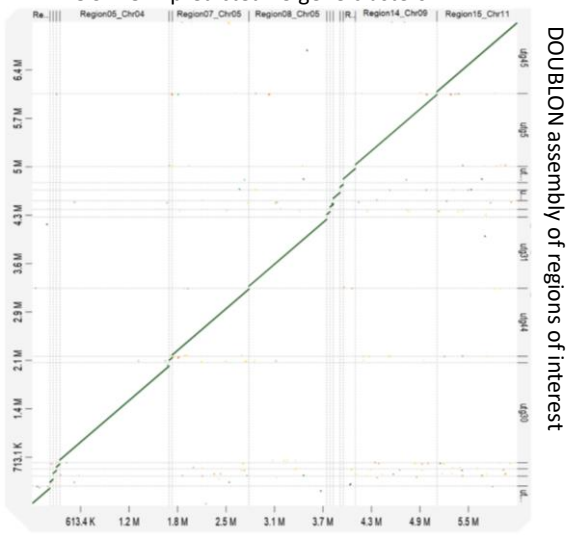
All predicted NLR genes (84) found in the assembly

number_contigs	15
total_length	7.023 Mb
largest_contig	1.443 Mb
GC(%)	33.23
N50	1.107 Mb
N75	1.026 Mb
L50	3
L75	5

Results using different varieties

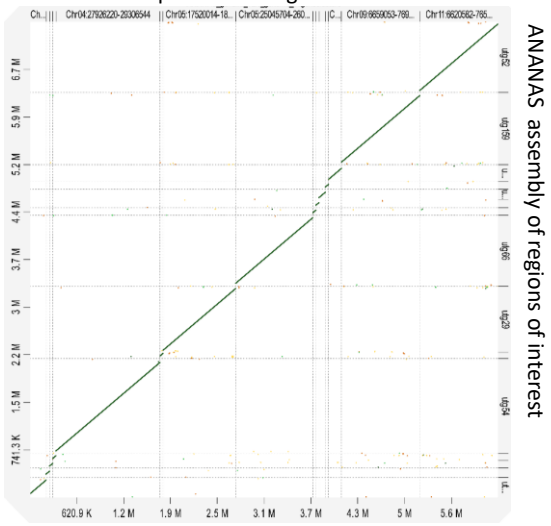
Also complete and continuous assembly of all the NLR-clusters

DOUBLON: predicted 15 gene clusters



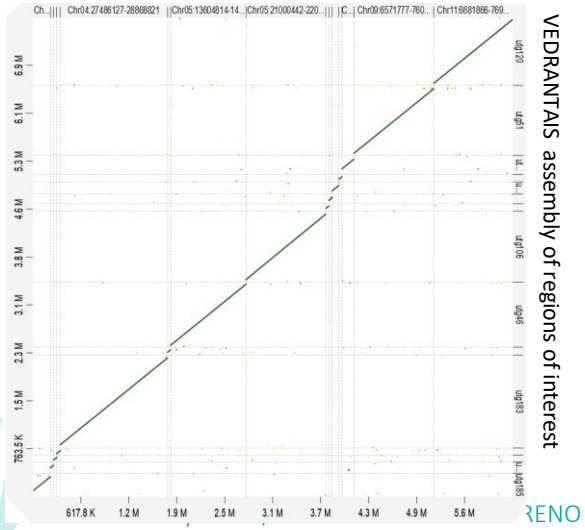
DOUBLON assembly of regions of interest

ANANAS: predicted 15 gene clusters



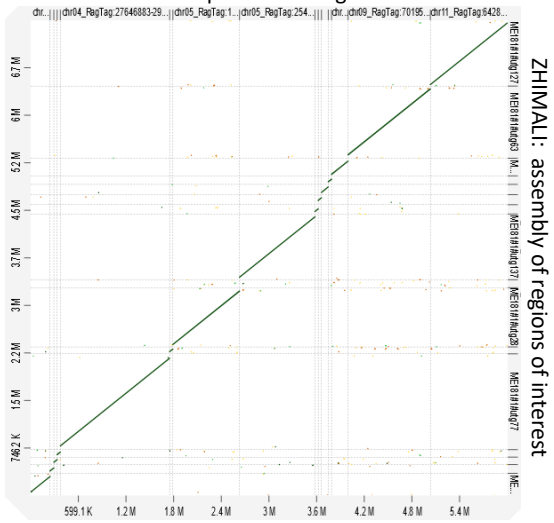
ANANAS assembly of regions of interest

VEDRANTAIS: predicted 15 gene clusters



VEDRANTAIS assembly of regions of interest

ZHIMALI: predicted 15 gene clusters



ZHIMALI: assembly of regions of interest

DOUBLON – cantalupensis (ssp. melo)

Target regions fully assembled

All predicted NLR genes (76) found in the assembly

ANANAS – ameri (ssp. melo)

Target regions fully assembled

85/85 NLR genes found in the assembly

VEDRANTAIS – cantalupensis (ssp. melo)

Target regions fully assembled

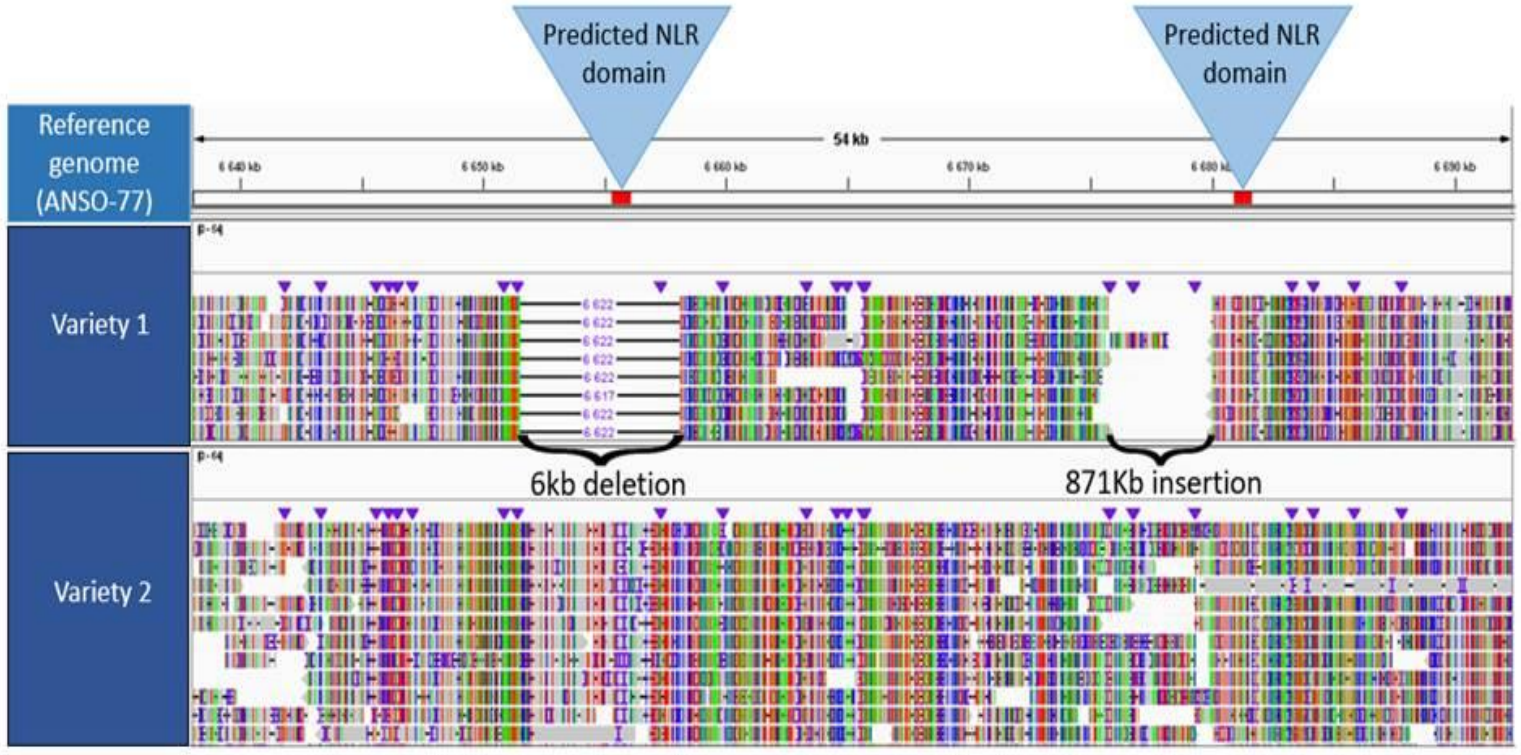
82/81 NLR genes found in the assembly

ZHIMALI – chinensis (ssp. agrestis)

Target regions fully assembled

79/79 NLR genes found in the assembly

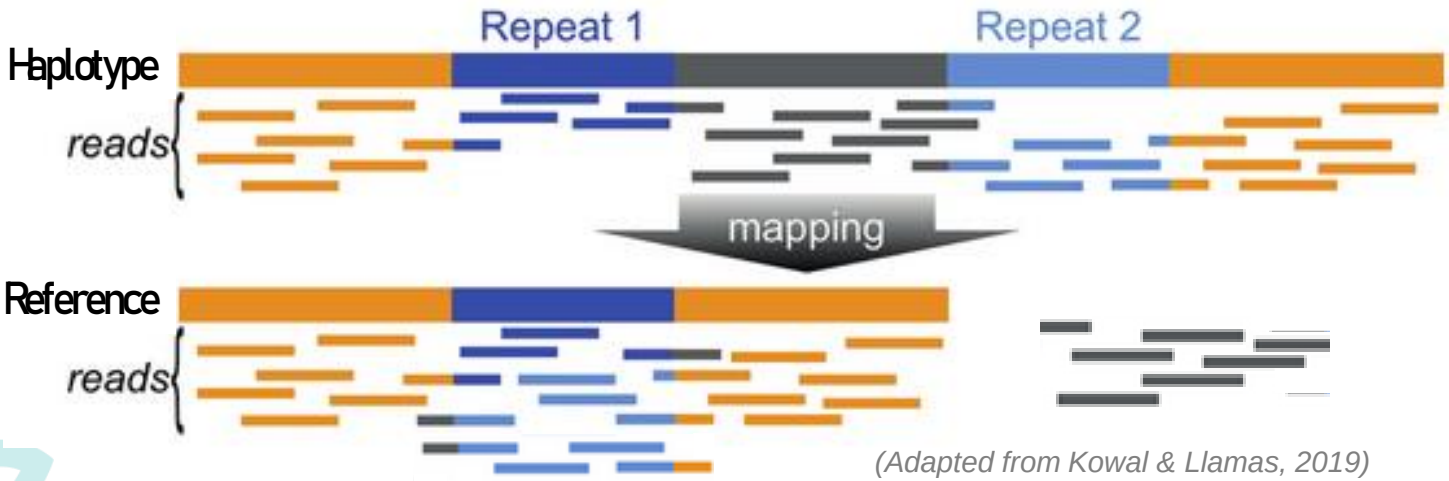
Adaptive sampling limitations



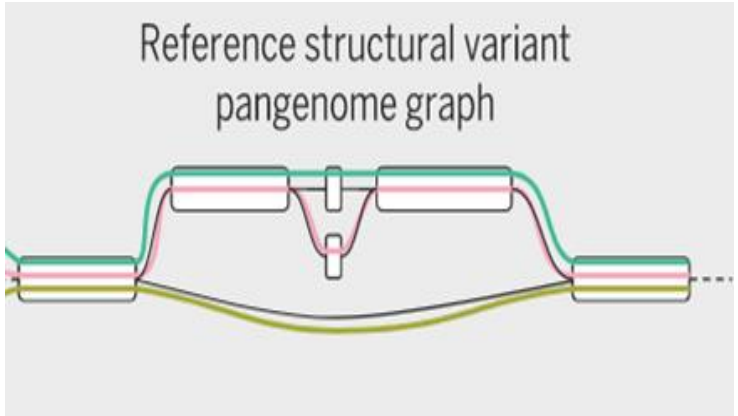
Very long insertions not present in the reference can not be captured ⇢ Minor impact in our study

A single reference is not enough: Necesity of a pan-NLRome

- Accession differing from the reference by SV → Reference may contain no location to correctly map the reads → **Mapping bias**
- **True** even with the newer **long-read** sequencing approaches
- For this reason, SVs are much more poorly characterized than SNPs and short InDels (Sirén *et al.,* 2021).
- The problem grows in highly complex regions with a large number of presence/absence polymorphisms and many **repetitive elements** → NLR clusters in melon



(Adapted from Kowal & Llamas, 2019)



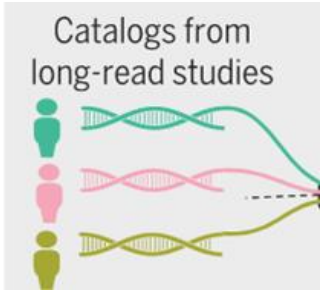
Pangenome construction tools

Using 7 varieties sequenced with NAS

	Minigraph *(Li et al., 2020)*	**Minigraph-Cactus** *(Hickey et al., 2022)*	**PGGB** *(Garrison et al., 2023)*
Sequence comparison	reference-based, progressive	reference-based, progressive	symmetric, all-vs-all
Resolution	SV only (variations >50 bp)	base-level (SNPs)	base-level (SNPs)
Scope	full assemblies	Non-centromeric	full assemblies
Full reconstruction possible	yes	yes	no
Short read mapping	untested	yes (fast)	untested
Long read mapping	yes (fastest)	yes	yes (slowest)

(Adapted from https://github.com/human-pangenomics/hpp_pangenome_resources)

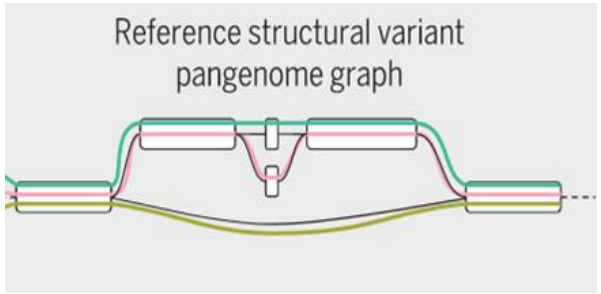
Variant genotyping on the pangenome graph



Assembly panel

Pangenome graph inference

- minigraph
- minigraph-cactus
- pggb
- ...

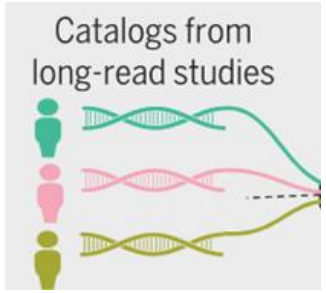


Graph reconstruction

Variant detection inside the graph

- VG deconstruct

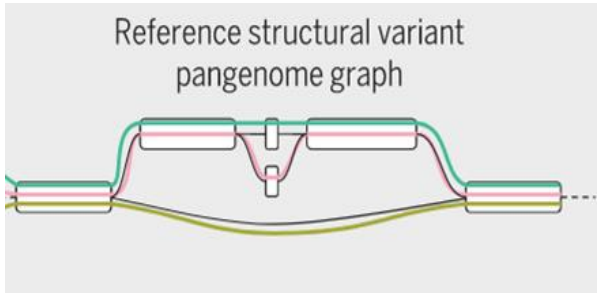
Variant genotyping on the pangenome graph



Assembly panel

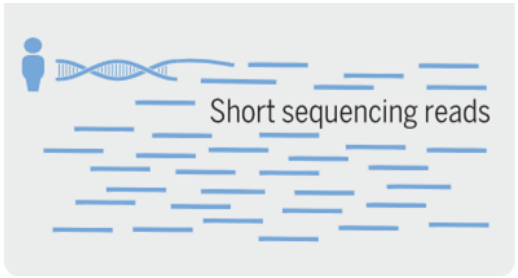
Pangenome graph inference

- minigraph
- minigraph-cactus
- pggb
- …



Graph reconstruction

Variant detection inside the graph



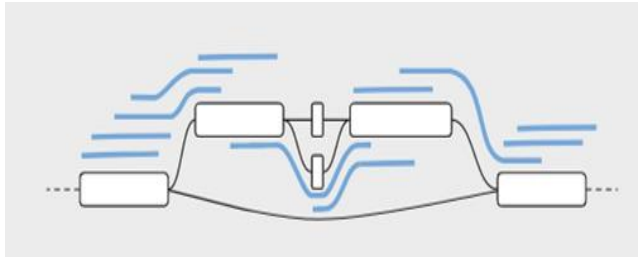
Genomes to genotype

mapping & genotyping

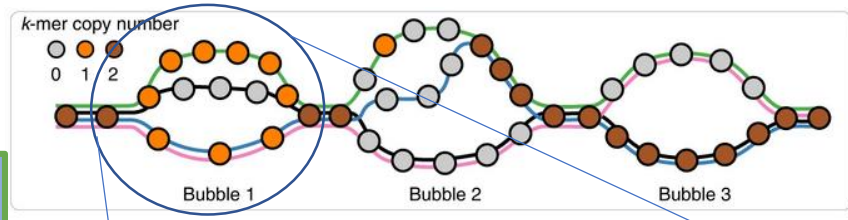
- VG giraffe/VG call
- GraphAligner/VG Call
- Pangenie
- …

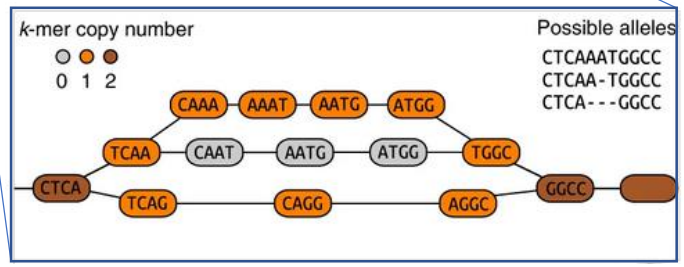
Genotypes **ALREADY** present in the graph

VG-Giraffe & GraphAligner



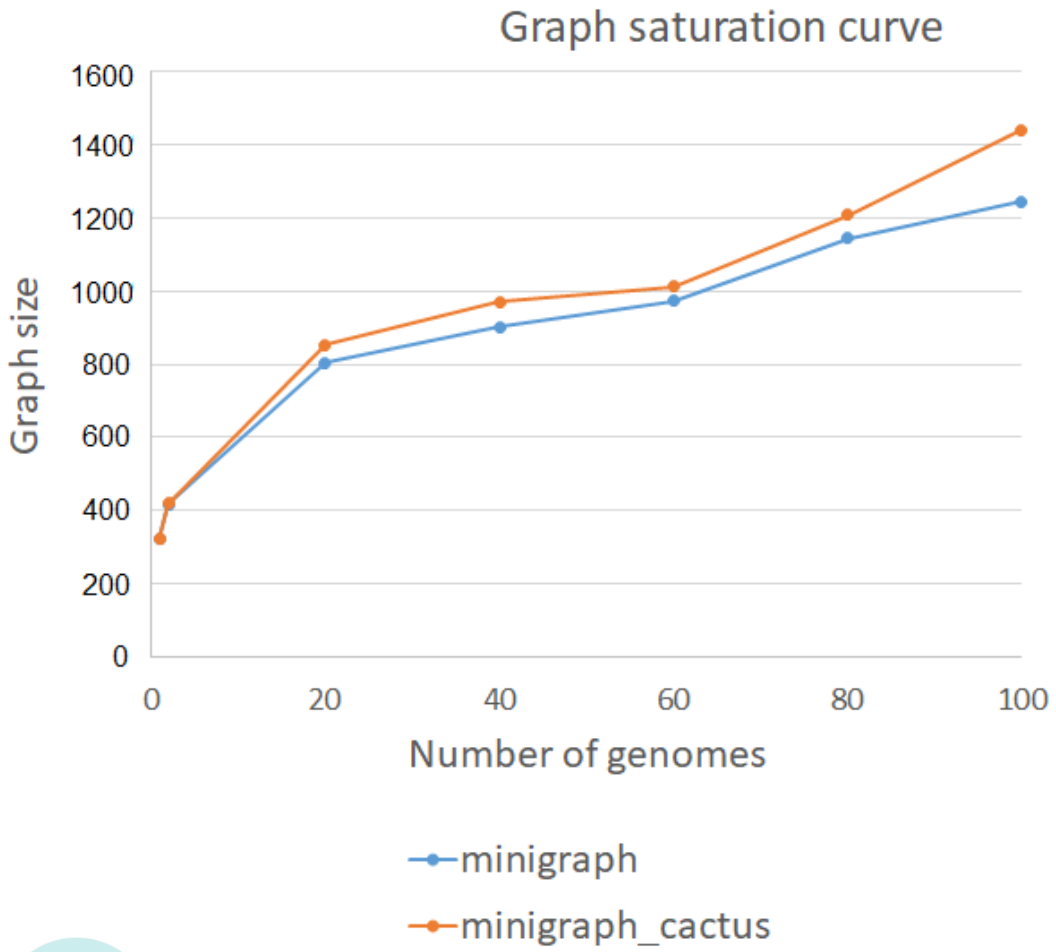
Pangenie



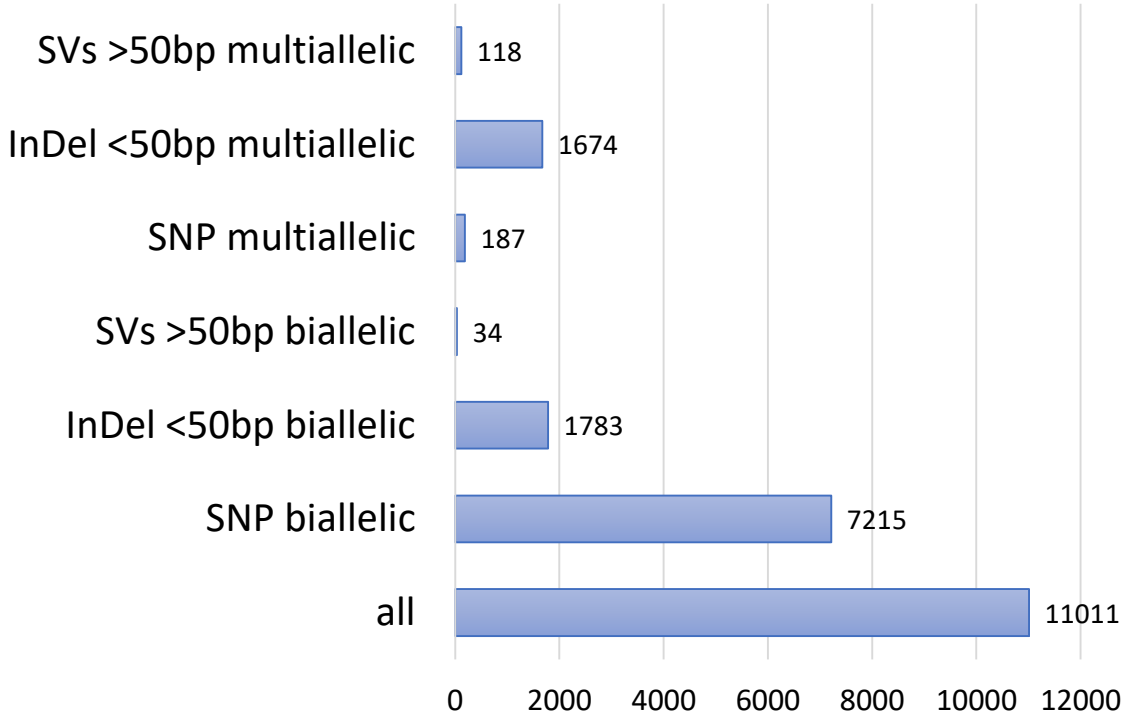


First attempts to build a pan-NLRome graph

Focus on the well studied M4-M5 region of Chr5 (350 Kb)



100 accessions graph (minigraph-cactus)



First attempts to build a pan-NLRome graph

Focus on the well studied M4-M5 region of Chr5 (350 Kb)

2 melon acccessions
Anso77 and Doublon

419 557 bp
6 033 segments
8 189 edges

Variation present in graph
2157 variants found
1665 SNPs
462 InDels 1-50 bp
31 SVs > 50 bp

Mapping of Anso77 short reads
1 variants PASS found
0 SNPs
1 InDels 1-50 bp
0 SVs > 50 bp

Mapping of Doublon short reads
2149 variants PASS found
1657 SNPs (99.5%)
462 InDels 1-50 bp (100%)
30 SVs > 50 bp (98%)

7 melon acccessions
Anso77 and Doublon

584 518 bp
24 820 segments
34 084 edges

Variation present in graph (only counting ANSO/DOUBLON)
1531 variants found
1199 SNPs
302 InDels 1-50 bp
30 SVs > 50 bp

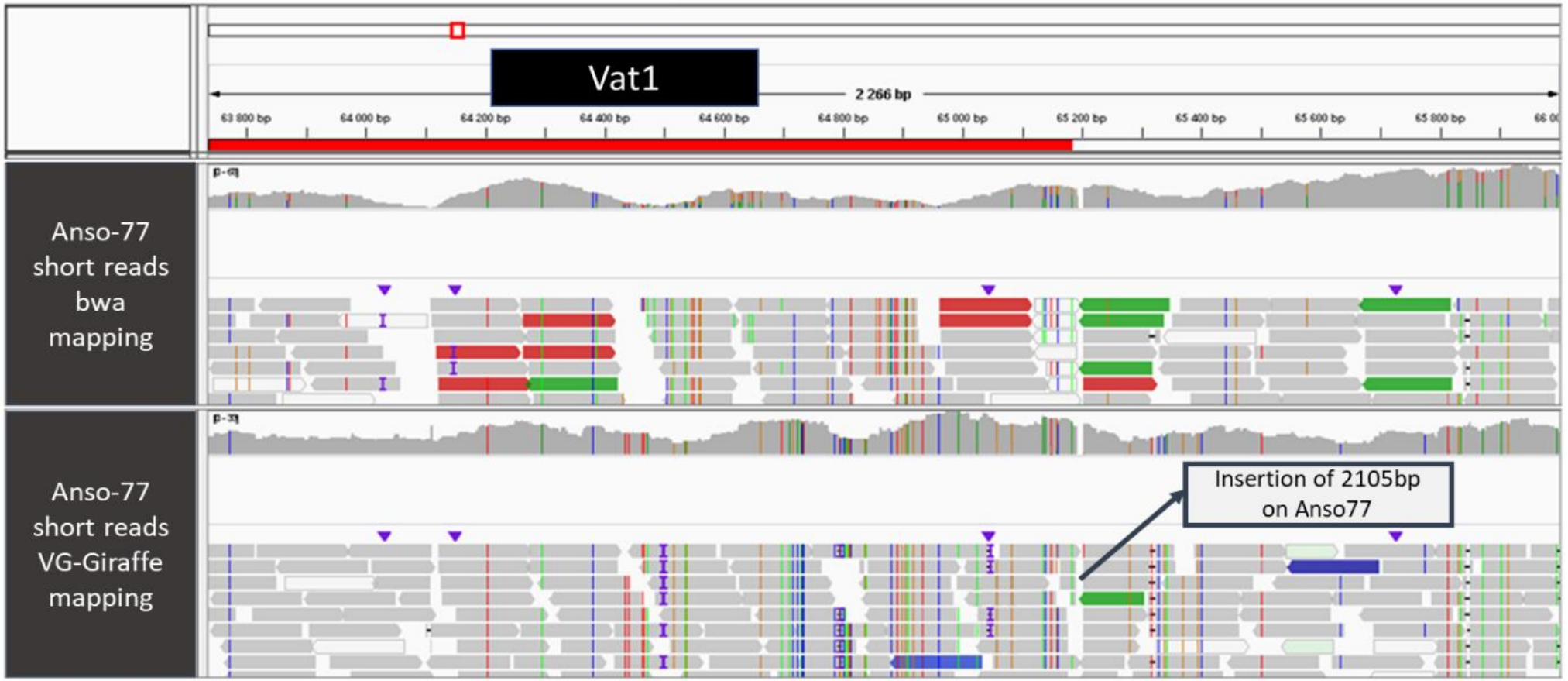
Mapping of Anso77 short reads
7 variants PASS found
0 SNPs
6 InDels 1-50 bp
1 SVs > 50 bp

Mapping of Doublon short reads
1703 variants PASS found
1305 SNPs (108.8%)
372 InDels 1-50 bp (123%)
26 SVs > 50 bp (86.7%)

First attempts to build a pan-NLRome graph

Focus on the well studied M4-M5 region of Chr5 (350 Kb)

2 melon acccessions
Anso77 and Doublon



In conclusion

- Nanopore adaptive sampling is a **simple**, **reliable**, **efficient** and **cost-saving** target sequencing approach
- Adaptive sampling allows to **efficiently retrieve** (**mapping and assembly**) our **15 ROI in melon**.
- A **reference genome is not enough** to characterise the diversity of the NLRome in melon
- Graph pangenomics still unstable and under developement
- The construction of a **pangenome graph** will allow the **characterisation of the complex NLRome clusters** using a **reference-free** approach that improves the common errors showing up in classical variant calling with a single reference

Acknowledgements

INRAE-EPGV

Patricia Faivre-Rampant
Aurelie Berard
Aurelie Canaguier
Isabelle Le-Clainche
Damien Hinsinger

INRAE-GAFL

Nathalie Boissot
Veronique Chovelon
Jacques Lagnel
Vincent Rittener-Ruff

Genoscope

Corinne Cruaud
Stefan Engelen

PRIVATE PARTNERS

@EPGV_INRAE

support-EPGV@inrae.fr
javier.belinchon-moreno@inrae.fr