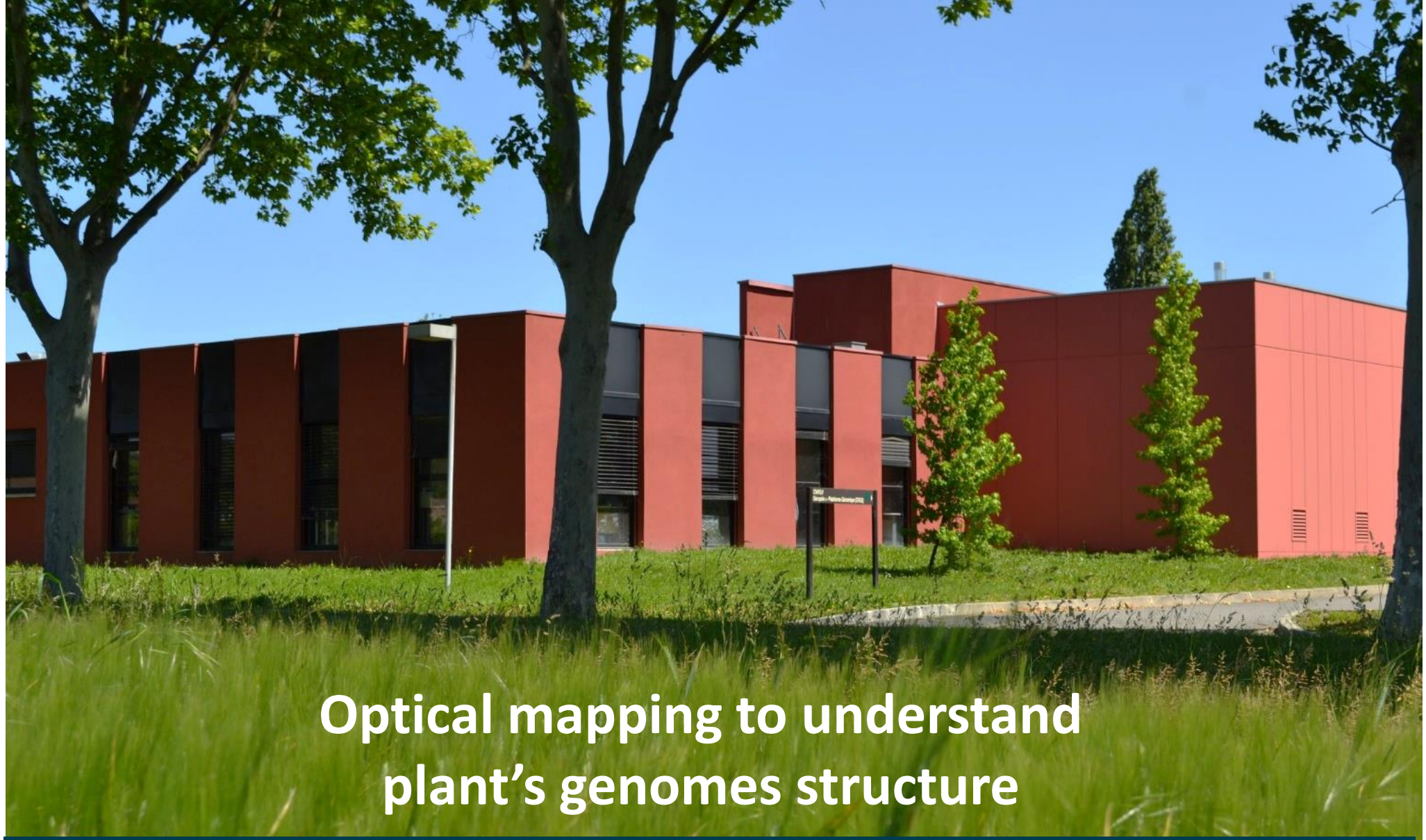
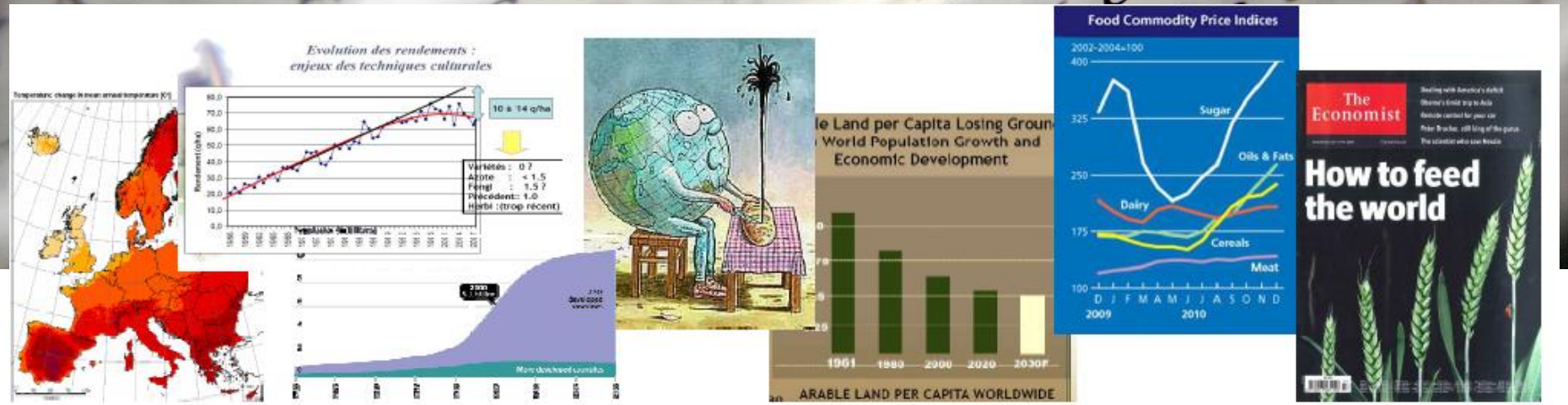


Optical mapping to understand plant's genomes structure

CNRGV-INRA
Chemin de Borde Rouge/ B.P.
52627 / 31326 Castanet-Tolosan

William Marande

- In a context of climate change, population growth and limited energy resources, increasing **plant genomes knowledge** is essential for a better understanding of mechanisms driving plant adaptation and evolution
- Exploration of plant genomes remains challenging : high level of **genome complexity**
- Single reference genome **is not enough** : high intra-species variability
- **Reliable sequence** information linked to a **trait** of interest in specific genotypes is essential to understand the role of a genomic region in a phenotype



Plant genome sequencing success: Use complementary technologies

To obtain a high quality genome sequence, it is necessary to **combine** several **sequencing technologies**

- Long read sequencing: Pacific Bioscience or Oxford Nanopore technology
  ➔ Sequence information and skeleton assembly

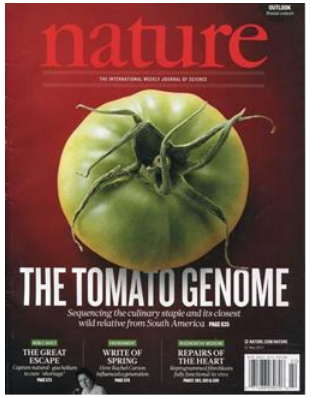
- Illumina sequencing
  ➔ Sequence accuracy

- **Optical maps**, Hi-C sequencing and 10X Genomics
  ➔ Scaffolding

Collaboration projects at the CNRGV

Whole genome sequencing project

BAC library / HMW DNA / Optical maps



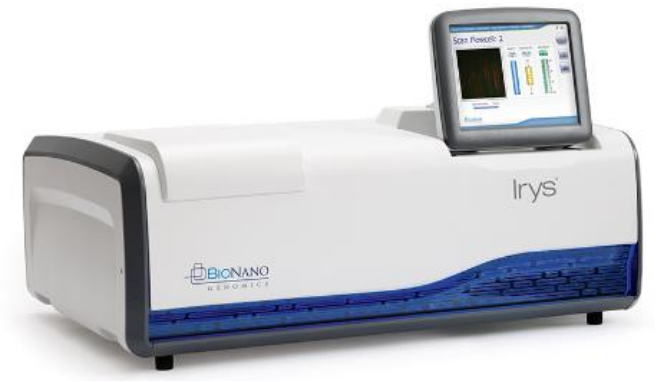
Characterization of specific genomic region of interest

QTL cloning/ Sequence capture/ Optical maps

TARGETED BAC LIBRARY CONSTRUCTION

BAC CLONE CHARACTERIZATION

Optical mapping



The BioNano Irys / Saphyr system

Direct visualization of long DNA molecules (> 100 kb)

Real physical distance information

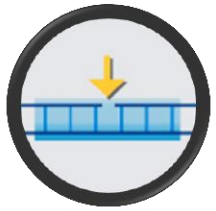
Applications:

Whole Genome scaffolding

Visualization of Structural Variations

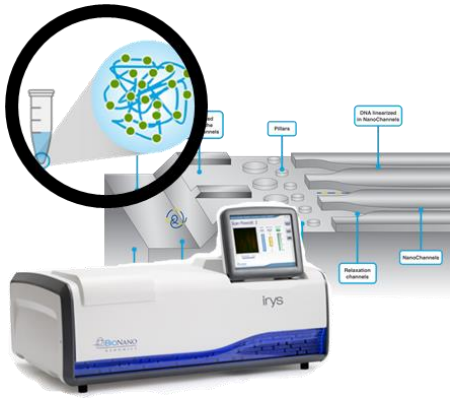
Optical map production

Isolation of HMW DNA
>150kb – 2Mb



Nicking of HMW DNA at spectific sites

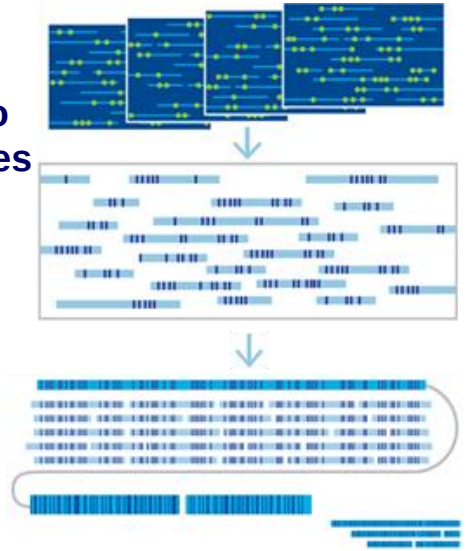
Fluorescent labelling of enzyme recognition sites



DNA molecules are linearized in NanoChannels

Imaging of linearized DNA molecules

Images converted into digital molecules



***De Novo assembly of* a consensus genome map**

The Optical Mapping Service at CNRGV

GENOME
OPTICAL MAPPING

Since 2016 : 29 optical maps for 14 species

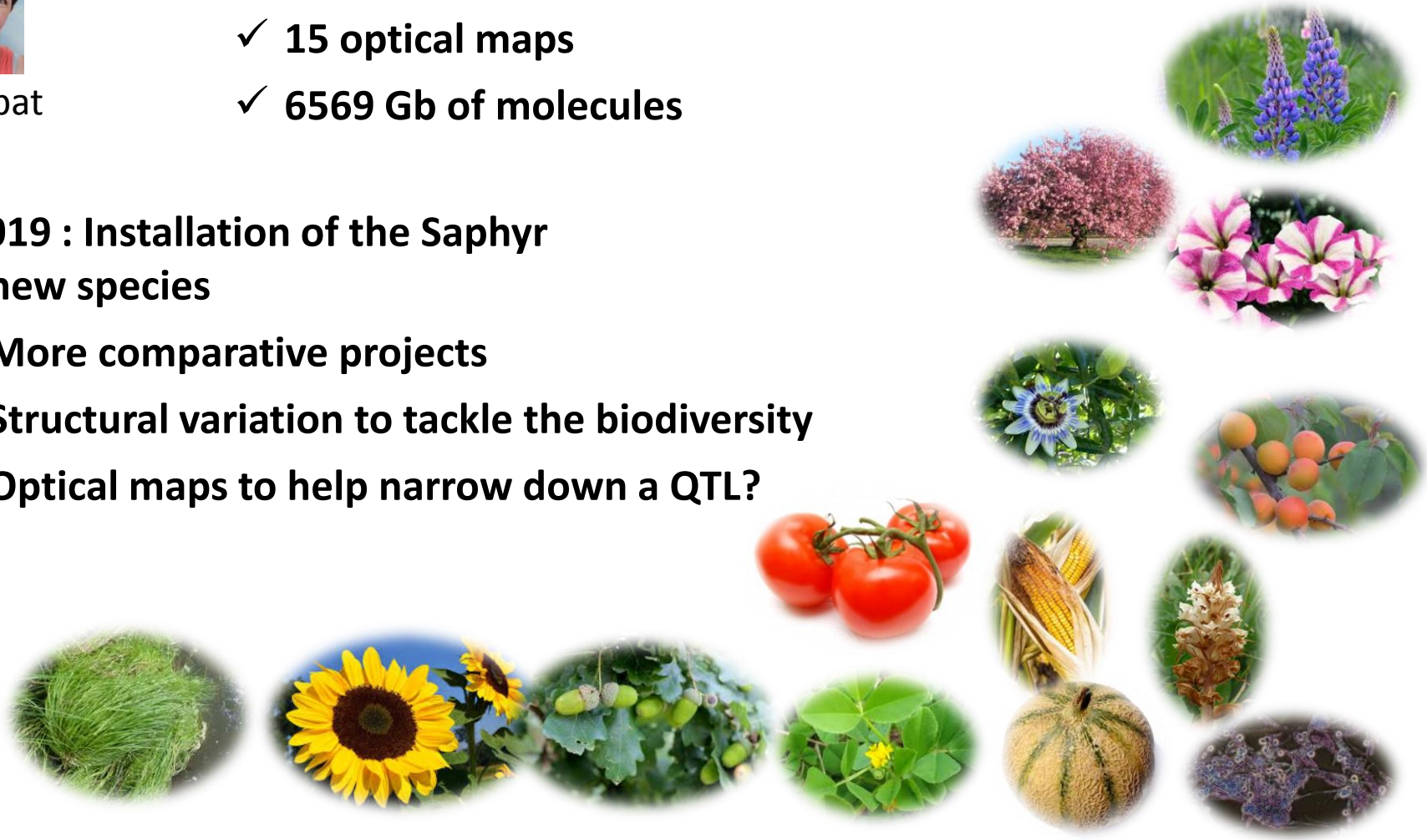
S. Arribat

2017 :

- ✓ **8 species**
- ✓ **15 optical maps**
- ✓ **6569 Gb of molecules**

2018-2019 : Installation of the Saphyr

- ✓ **new species**
- ✓ **More comparative projects**
- ✓ **Structural variation to tackle the biodiversity**
- ✓ **Optical maps to help narrow down a QTL?**



Tomato genome: PacBio with optical map

BspQI Raw data: 120X, N50 = 250 kb
>150kb: 85X, N50= 350kb
Nb label /100kb= 7,3

BssS1 Raw data: 180X, N50 = 205 kb
>150kb: 115X, N50= 260 kb
Nb label /100kb= 9,9

With PacBio data (RS2 system)

	PacBio Assembly	Optical map BspQ1	Hybrid scaffold BspQ1	Optical map BssS1	Hybrid scaffold 2 Step
Count	743	301	74	444	40
N50 length (Mb)	3,4	3,7	17,7	2.7	34
Total length (Mb)	0.79	0,77	0,79	0,88	0,82

x 10

Tomato: Mohamed ZOUINE; Sequencing at Get- PlaGe Genotoul

PacBio compared with 10x genomics: the Tomato genome

10X data (Illumina HighSeq)

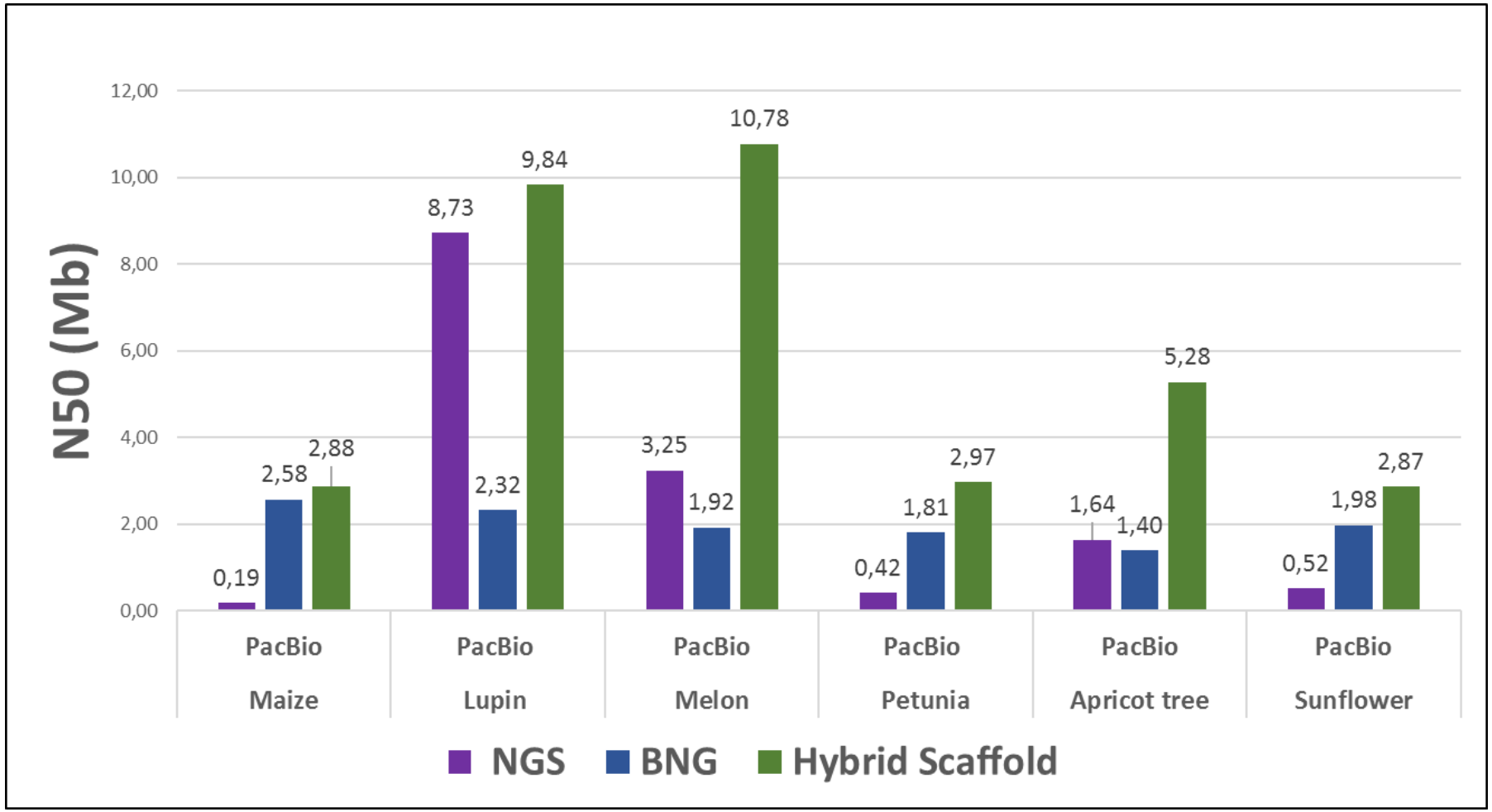
	PacBio Assembly	Optical map BspQ1	Hybrid scaffold BspQ1	Optical map BssS1	Hybrid scaffold 2 Step
Count	24500	301	116	444	80
N50 length (Mb)	1,8	3,7	9,8	2.7	17
Total length (Mb)	0.79	0,77	0,77	0,88	0,84

x 10

➔ 10x genomic N50 (1.8 Mb) is smaller than PacBio (3.4 Mb) but the result is still very good for 10 times less money
Same improvement than with PacBio data (X10)
Several advantages: few DNA (10-20ng) and same DNA for optical maps and 10X genomic

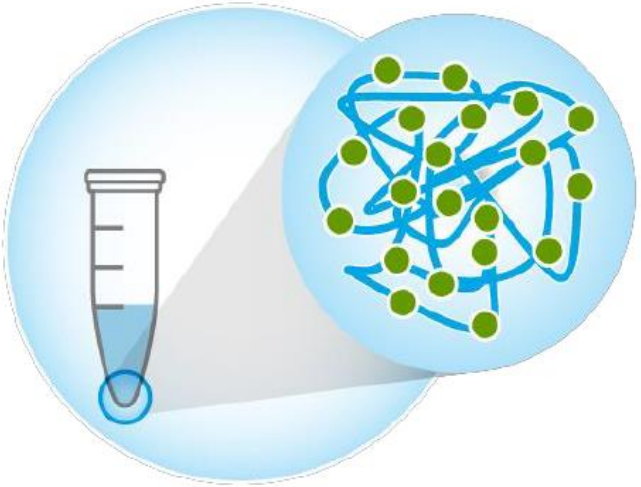
Tomato: Mohamed ZOUINE; Sequencing at Get- PlaGe Genotoul

The optical map to improve the Genome sequence assembly



Maize within Amaizing project: Clémentine VITTE ; Lupin: Benjamin PERET for ERC LUPIN ROOTS ; Melon: Abdelafid Bendhamane
Petunia: Michel Moser; Apricat tree: Véronique DECROOCQ; Sunflower within SUNRISE project (Nicolas Langlade, Stéphane Munos et Jérôme Gouzy)

Bionano Technology Revolution: DLE

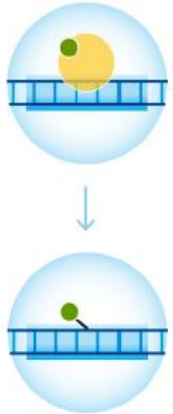


Nick
Label
Repair
Stain
(NLRS)

2

Label DNA at specific sequence motifs

Direct
Label
and
Stain
(DLS)



bionano GENOMICS

The optical map to improve the Genome sequence assembly

- Species: *Helianthus annuus* Sunflower
- 3.6 Gb
- 2n=34 chromosomes
- Genome sequence >100X PacBio (XRQ genotype)

N. Langlade

# contigs	LEN Max	N50 BP	#>N50	MEDIAN	BP
12 318	3,35 Mb	524 kb	1 684	120 kb	2,93

=> 80% of the genome inside contigs

Gouzy et al., 2016

Two major repeats in the sunflower genome: 8 kb and 11.5 kb

The 2-steps hybrid scaffolding strategy improves significantly the resulting N50

	PacBio Assembly	BioNano BspQ1 Assembly	Hybrid scaffold BspQ1	BioNano BssS1 Assembly	Hybrid scaffold 2 Step
Count	12318	2228	1430	4287	1069
Median length (Mb)	0.120	0.999	1.442	0.551	1.914
N50 length (Mb)	0.524	1.979	2.87	0.968	4.166
Max length (Mb)	3.35				24.670
Total length (Mb)	2930	3191	2922	3112	2960
% genome	81%	88%	81%	86%	82%

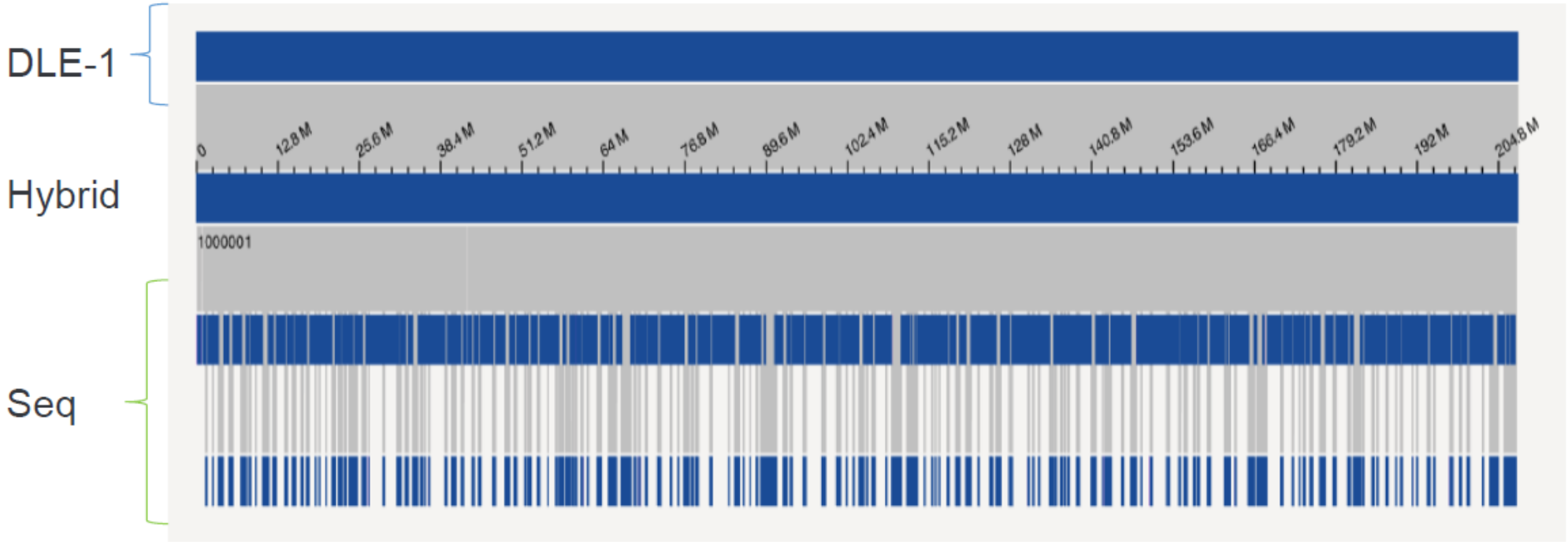
More than 7 fold increase

Gouzy et al., 2016

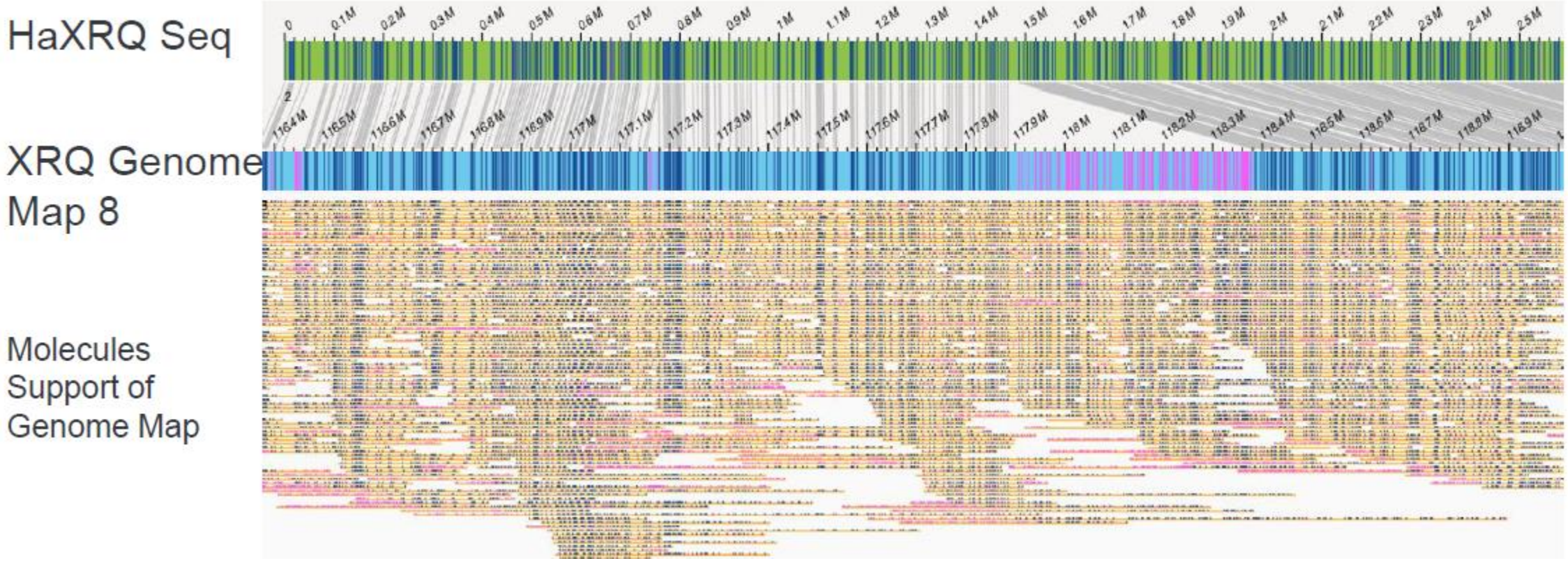
A new revolution: optical map with Direct Labelling Enzyme (DLE)

	Statistic	Original BNG	Original sequence	Sequence used in hybrid scaffold	Hybrid scaffold	Hybrid A + leftover unscaffolded sequence
DLE-1	Number of maps	69	11676	8738	25	5317
	N50 (Mb)	175.21	0.52	0.46	176.33	175.95
	Total length (Mb)	3057.67	2926.51	2792.45 (95.42%)	3000.44	3134.36

2359 cuts were made on 1167 sequences during chimera detection.
Cuts can be due to chimera or allelic difference.



Genome assembly improvment: sunflower XRQ



➔Validation of the optical map assembly versus the NGS assembly

➔XRQ genome: 17 scaffolds representing the 17 chromosomes

Optical map with the Saphyr and the DLE

➔ Essential tool to scaffold and validate the plant genome sequences assembly

➔ Genotype comparison at the genome level is feasible

Optical maps to study structural variations between genotypes: from the megabases (chromosomes) to the kilobases level (genes and repeated sequences)

Optical maps to study structural variations: Sunflower genotypes comparison

Optical map assembly with DLE1

Sample	XRQ	LSS	LSR	Arikara	wild sunflower
Data collected (Gb)	310	360	360	380	313
Assembly size (Gb)	3,06	3,05	3,07	4,92	5,43
Genome map N50 (Mb)	**175**	**175**	**9,9**	**77,9**	**55,3**
Coverage*	101	118	117	77	58

*estimation based on the assembly size

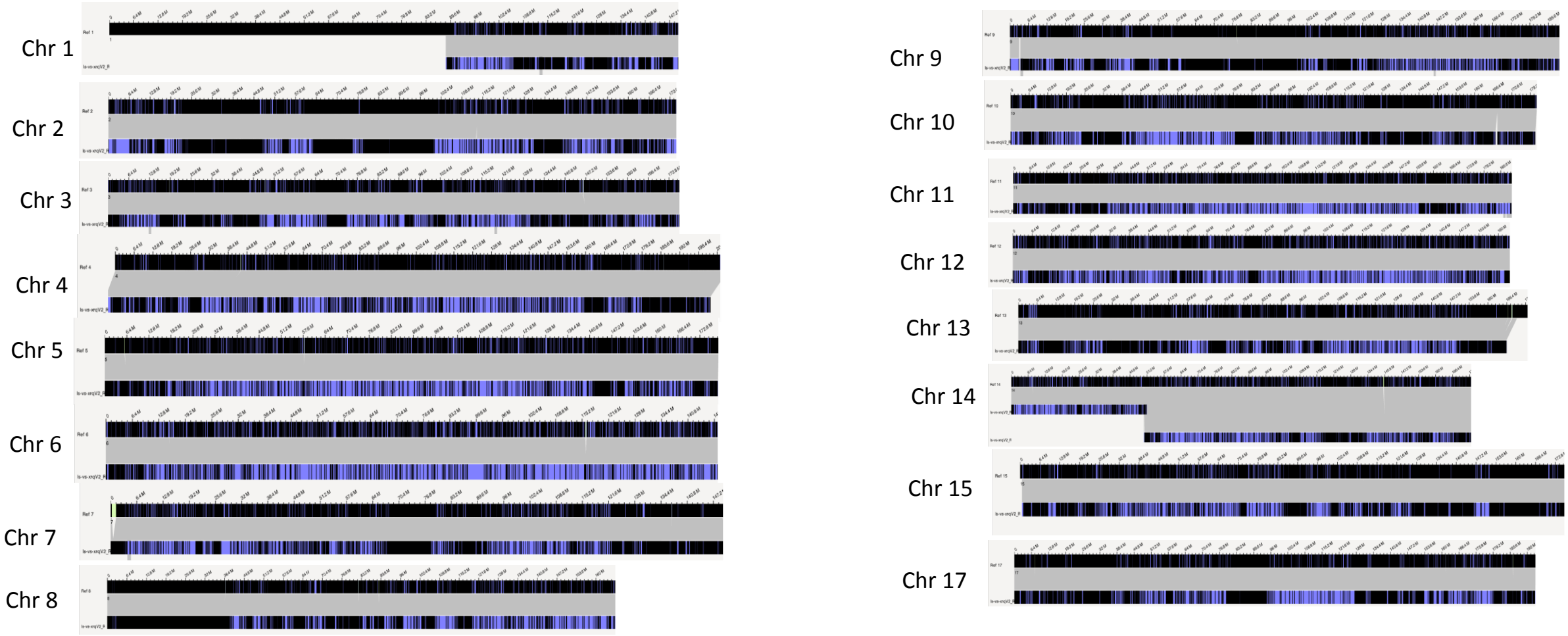
5 sunflower genomes with optical maps assembly at the chromosome level

➔Definition of variable and conserved region amongst the chromosomes

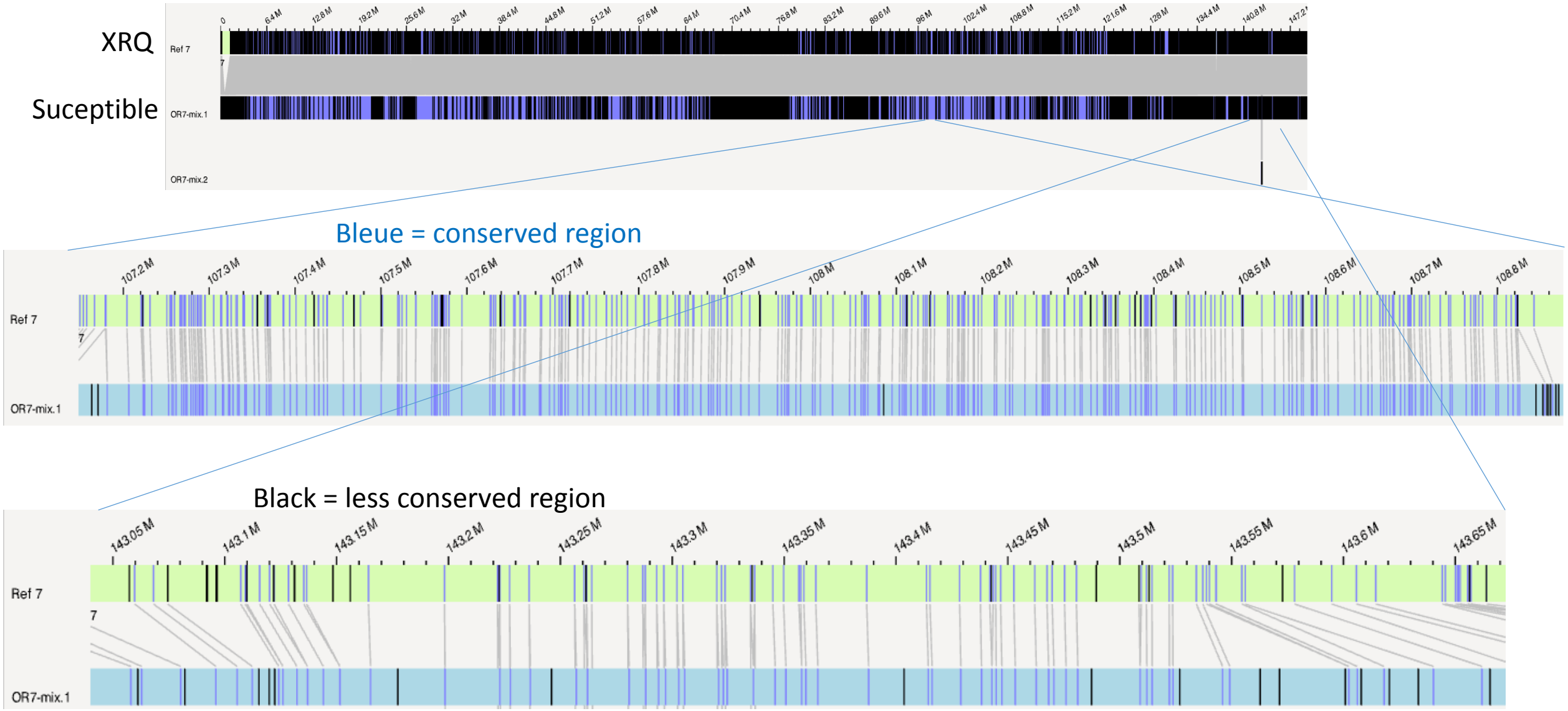
➔Suceptible (LSS) vs resistant genotype (LSR) structural variation analysis to understand resistance mechanism

Optical map alignment of 2 sunflower genotypes

Preliminary analysis: alignment of XRQ genome (upside) and LSS optical map (downside) using Refaligner software



All the chromosomes can be aligned and conserved region (bleue) vs less conserved (black) can be observed



Important information For region of interest analysis,
For the syntheny analysis
To evaluate the genetic difference between genotypes

Understanding the resistance of Sunflower to a parasitic plant

Sunflower

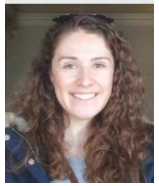
O. cumana

S. Munos

P. Duriez

Lipm Laboratoire Interactions Plantes Micro-organismes

syngenta

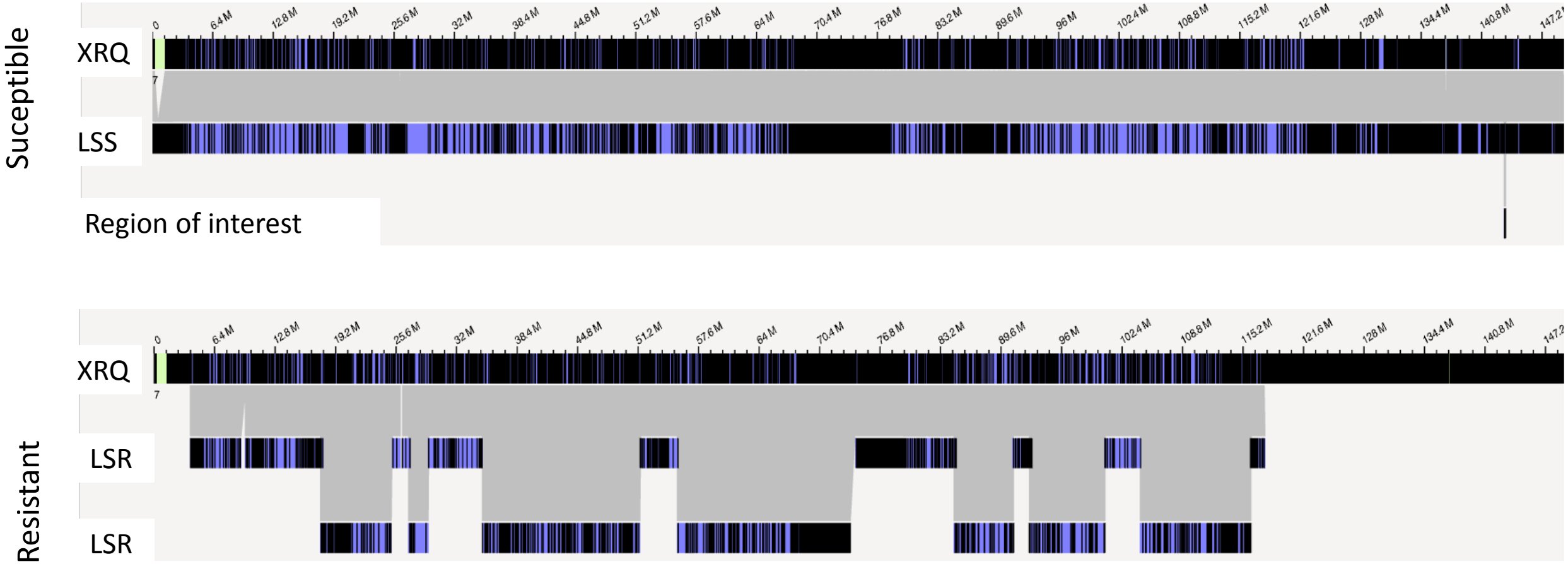


C. Satgé

S. Arribat

- *Orobanche cumana* : root-parasitic plant
- Important yield loss for sunflower crops in Europe
- Identification of QTL controlling the parasitic plant

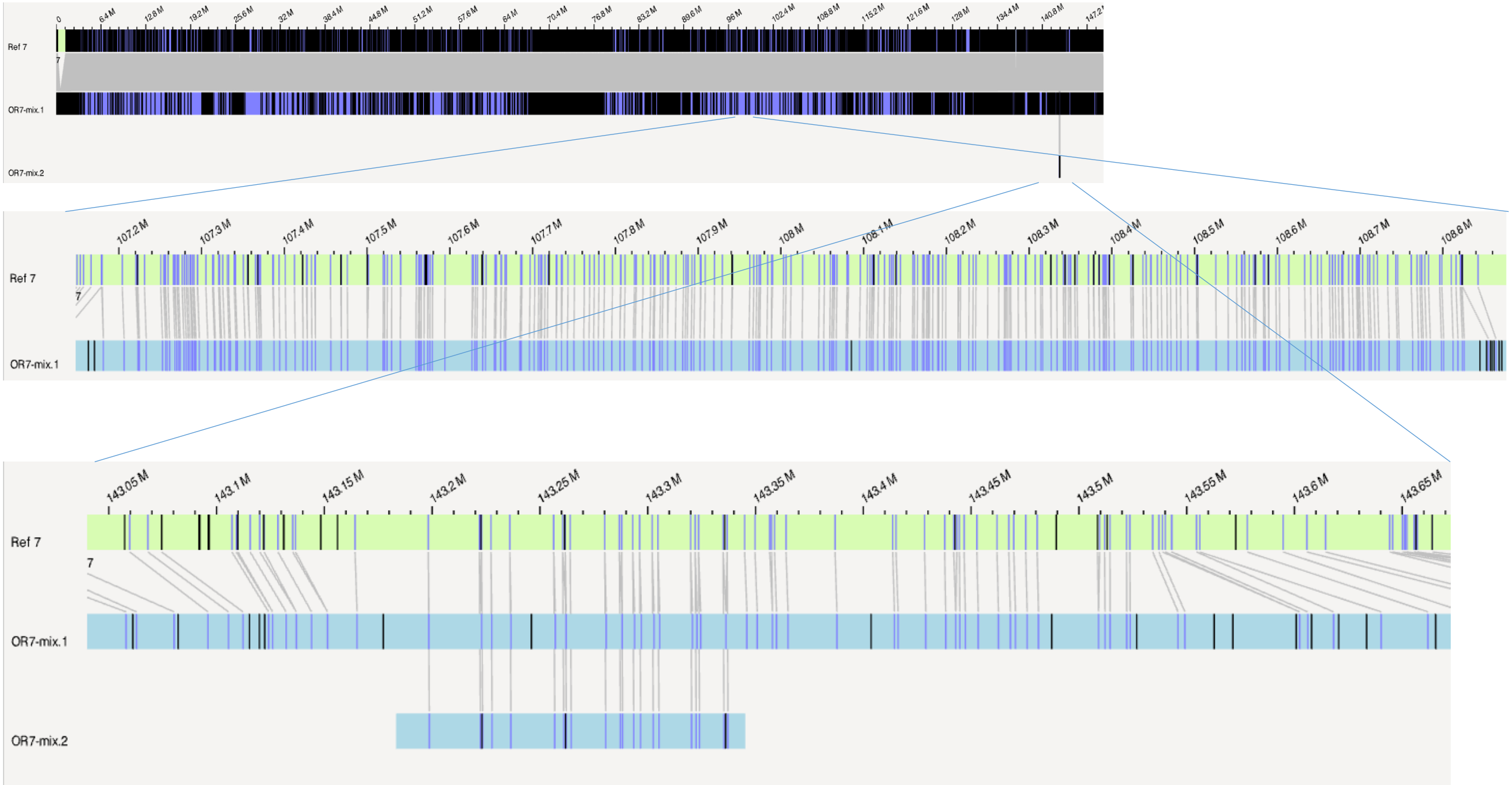
Where is the region of interest ?



The region of interest is in a variable region that do not align with the LSR genotype

➔Several structural variations (inversion, translocation) more than insertion or deletion

➔ This software do not analyse structural variation


Ref 7
OR7-mix.1
OR7-mix.2
0
6.4 M
12.8 M
19.2 M
25.6 M
32 M
38.4 M
44.8 M
51.2 M
57.6 M
64 M
70.4 M
76.8 M
83.2 M
89.6 M
96 M
102.4 M
108.8 M
115.2 M
121.6 M
128 M
134.4 M
140.8 M
107.2 M
107.3 M
107.4 M
107.5 M
107.6 M
107.7 M
107.8 M
107.9 M
108 M
108.1 M
108.2 M
108.3 M
108.4 M
108.5 M
108.6 M
108.7 M
108.8 M
143.05 M
143.1 M
143.15 M
143.2 M
143.25 M
143.3 M
143.35 M
143.4 M
143.45 M
143.5 M
143.55 M
143.6 M
143.65 M

Identification of true variants between LSR and LSS

Pipeline developped by bionano to identify true variants

Structural Variation Results between LSR (anchor) and LSS (query)

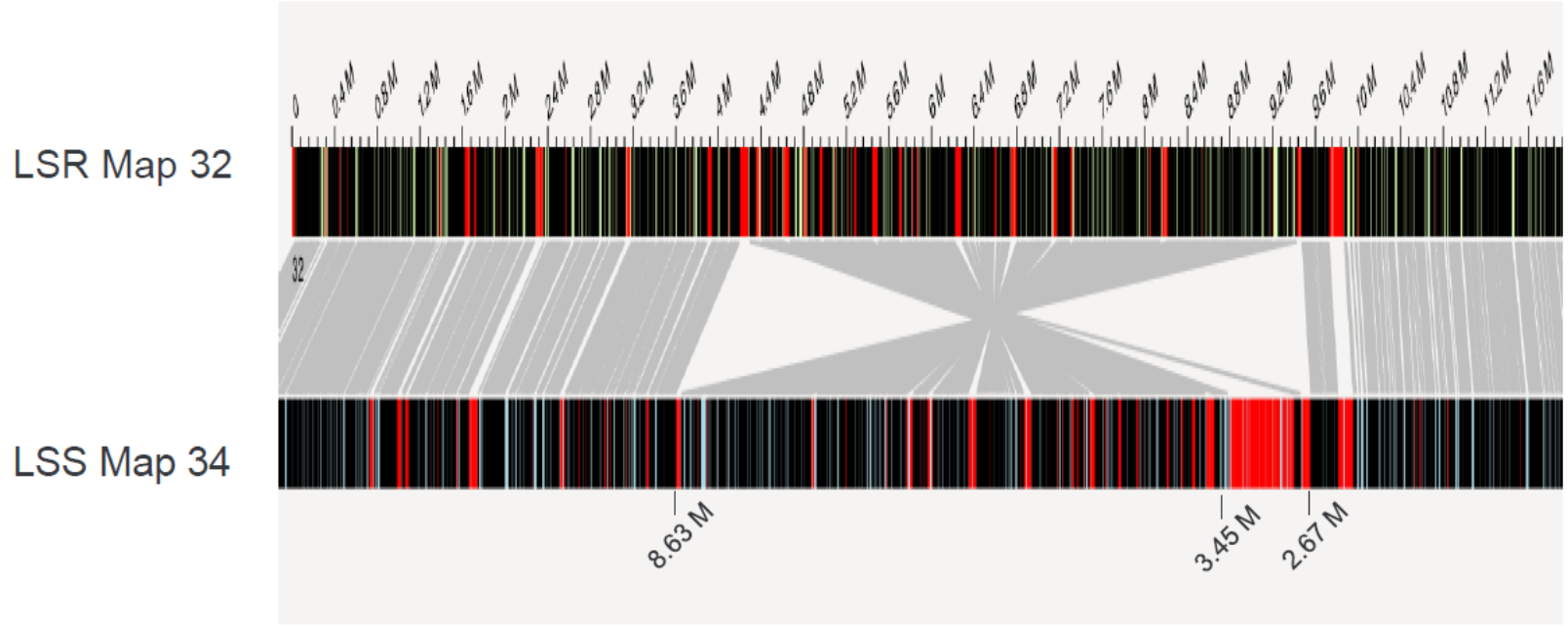
SVs	LSS to LSR (ref)
Insertion	2457
Deletion	2472
Inversion breakpoints	20
Inter-chr translocation	3 (further interpreted)
Intra-chr translocation	0

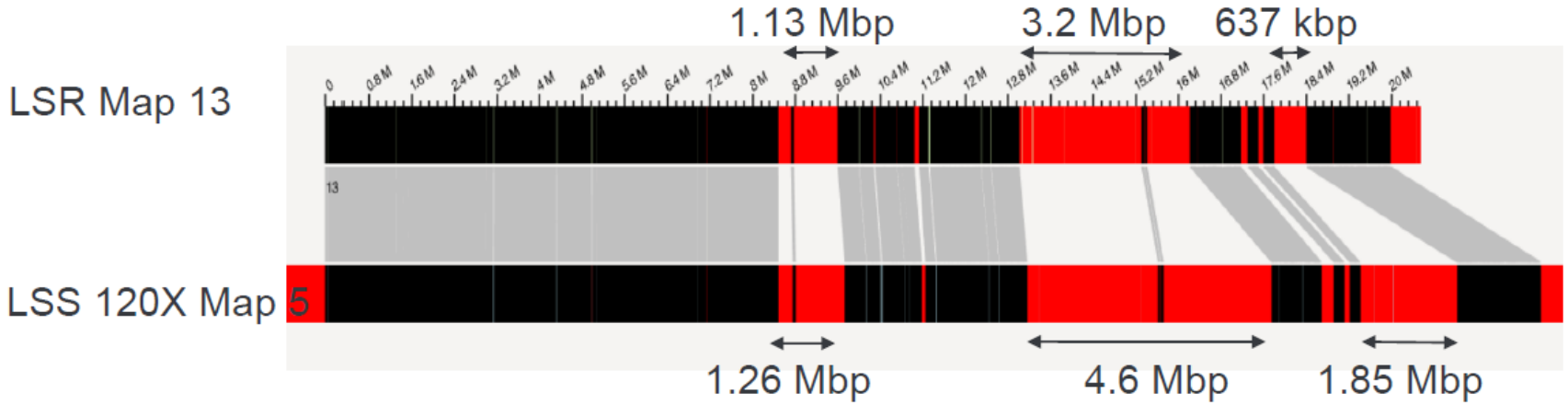
Indel with confidence > -1, Inversions with confidence >0, translocations with confidence >=0.1

bionano GENOMICS

Detection and Visualization of large genomic rearrangement

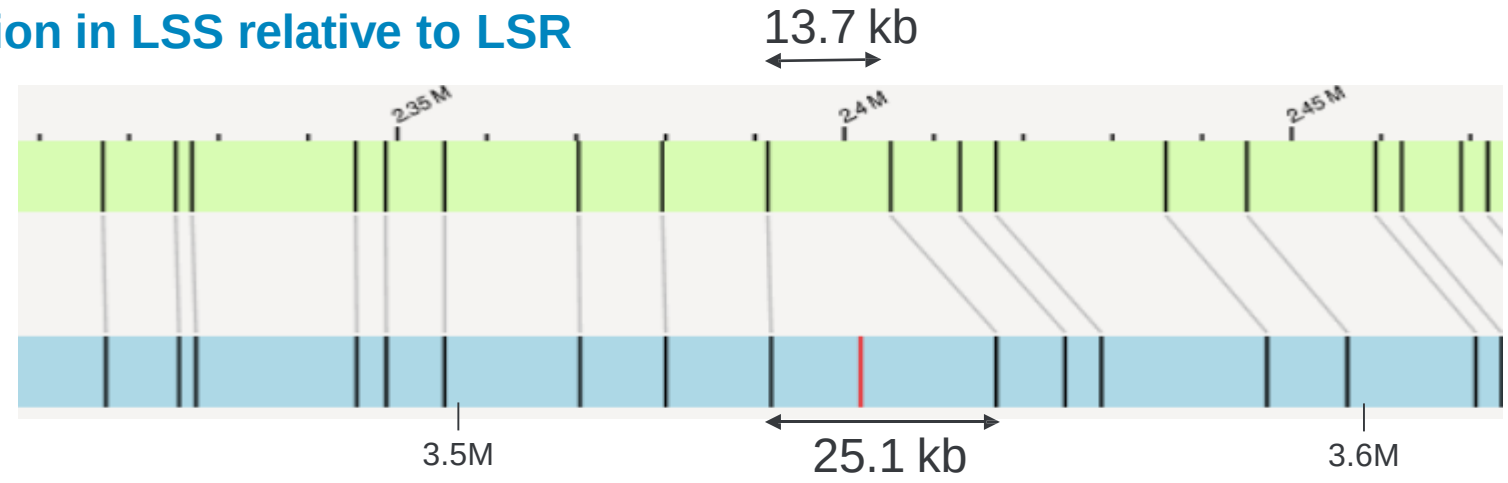
An 5.2 Mb inversion



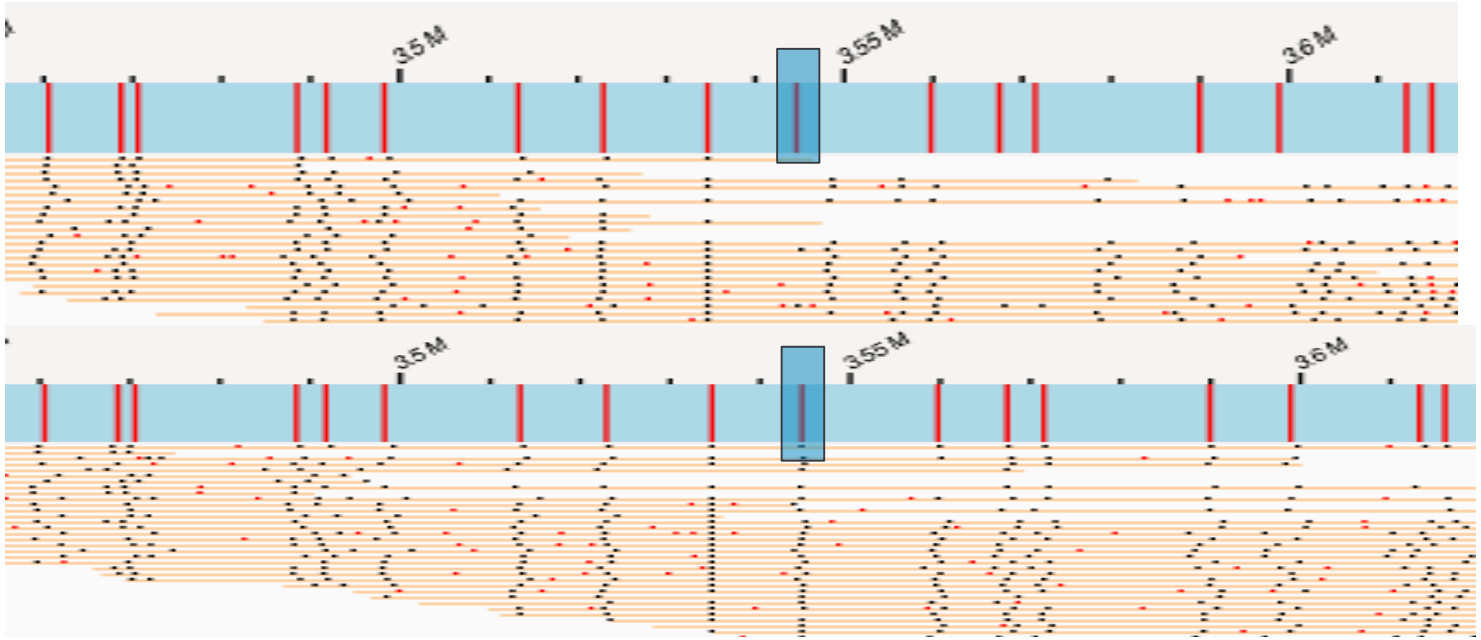


An example of insertion

A 11kbp insertion in LSS relative to LSR

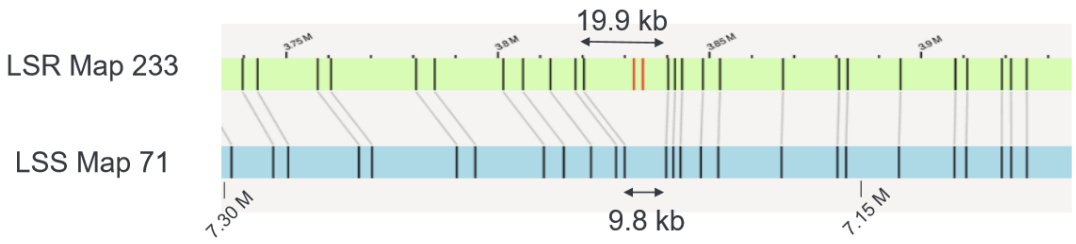


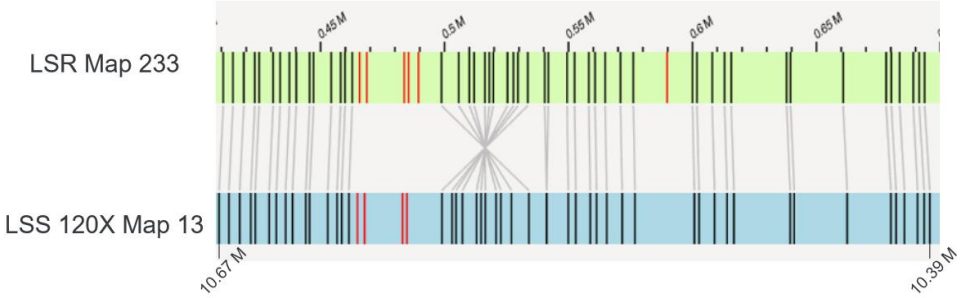
Cross Strain Molecule Check:



Structural variations examples

1Okb deletion in LS/LR





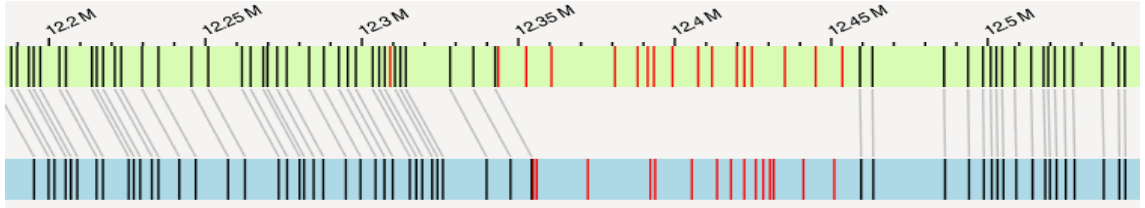
35kb inversion between LS/LR

Preliminary results

LSR Map 5

12kbp deletion in LSS / LSR

LSS Map 6



The way we use the optical maps

The BioNanoIrys system

Applications

Whole Genome scaffolding

Targeting a specific genomic region / Comparison to a reference genome, looking for changes in the patterns:

- **reveal insertion, deletion, inversion, translocation of genome segments**

Developments

- Dual labeling to target the region of interest with markers
- Analysis of the structural variations between genotypes
- Specific labeling such as epigenetic, telomere, gene cluster …
- Evaluate the possibility to construct optical map starting from a population of individuals and compare with another pop. : “core optical map” to highlight dedicated structural variations explaining a specific phenotype?

Acknowledgements

INRA SCIENCE & IMPACT

CNRGV PLANT GENOMIC CENTER

William MARANDE
Sandrine ARRIBAT
Carine SATGE
Céline CHANTRY-DARMON
Stéphane CAUET
Arnaud BELLEC
Sonia VAUTRIN
Nathalie RODDE
Elisa PRAT
Caroline CALLOT
Joëlle FOURMENT
Nadine GAUTIER
Nadège ARNAL
Roseana RODRIGUES
Isabelle DUFAU
David PUJOL
Laetitia HOARAU

Lipm Laboratoire Interactions Plantes Micro-organismes

SUNRISE UNE CULTURE POUR LE FUTUR

Jérôme GOUZY
Nicolas LANGLADE
Stéphane MUNOS
Pauline DURIEZ

@CNRGV
@SUNRISE_France
http://cnrgv.toulouse.inra.fr/

Jean-Christophe ROUSSEAUX
Joël PIQUEMAL
Jan GIELEN

syngenta

bionano GENOMICS

LA RÉGION OCCITANIE Pyrénées-Méditerranée

UNION EUROPÉENNE

INSTITUT CARNOT Plant2Pro

Comparison of NRLS and DLS technologies on the building of LSR and LSS optical maps

	LSR007 Assembly NRLS labelling[1]		LSS007 Assembly NRLS labelling[1]	
Réalisation	CNRGV		CNRGV	
Median length (Mb)	0,76		0,59	
Mean length (Mb)	1,0		0,79	
N50 length (Mb)	1,4		1,1	
Total length (Gb)	3,2		3,1	
Effective coverage of assembly (x)	88,9		94,3	

[1] : **N**ick, **R**epair, **L**abel, **S**tain
[2] : **D**irect **L**abel and **S**tain (non compatible with Irys system)

→ Better metrics and contiguity with the new chemistry DLS

Optical maps to study structural variations at the genome level

De novo genome map assemblies

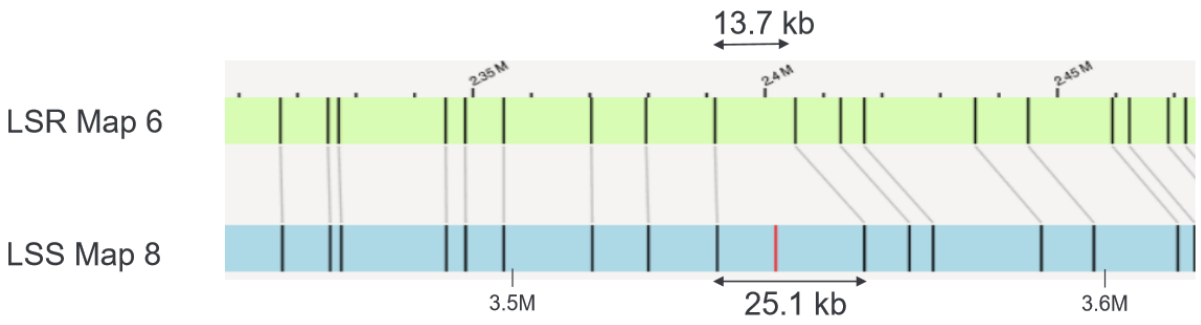
Sample	LSR (120X)	LSS (70X)	LSS (120X)
Labelled with	DLE-1	DLE-1	DLE-1
Data collected (molecules >150 kb)	362 Gbp	214 Gbp	345 Gbp
Molecules N50 (molecules > 150 kb)	299 kbp	314 kbp	317 kbp
Assembly size	3.07 Gbp	3.05 Gbp	3.05 Gbp
Genome map N50	9.88 Mbp	21.27 Mbp	174.93 Mbp

Structural variations analysis

SVs	LSS to LSR (ref)
Insertion	1946
Deletion	1955
Inversion breakpoints	6

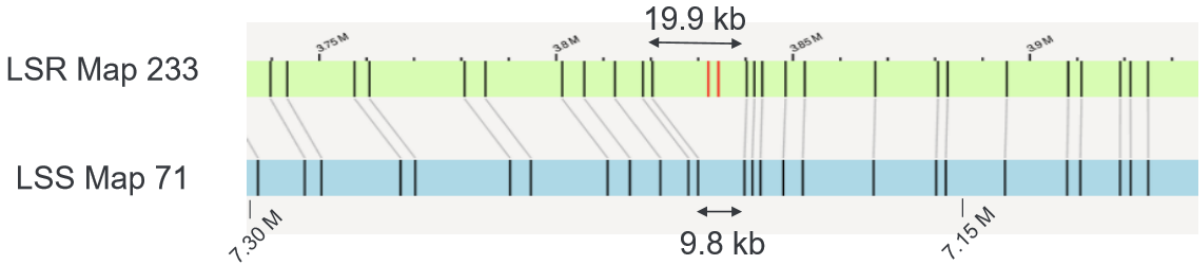
bionano GENOMICS

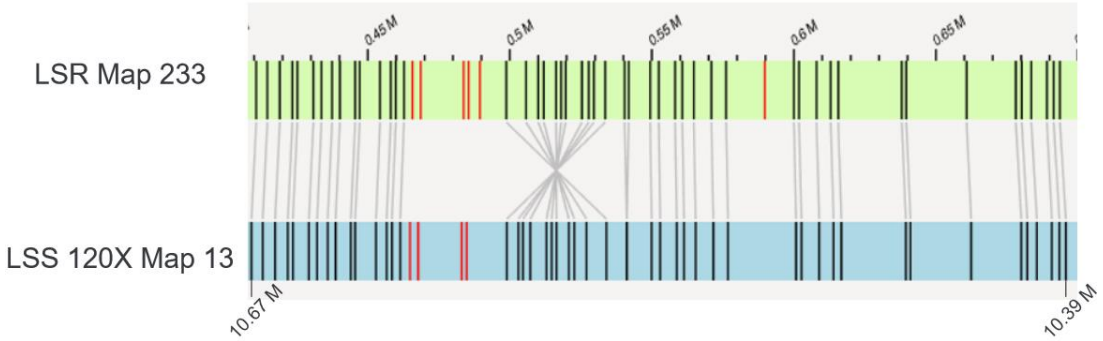
Structural variations examples



11kb insertion in LS/LR

1Okb deletion in LS/LR





35kb inversion between LS/LR

Preliminary results

=> How to be more efficient in focusing on genomic regions?

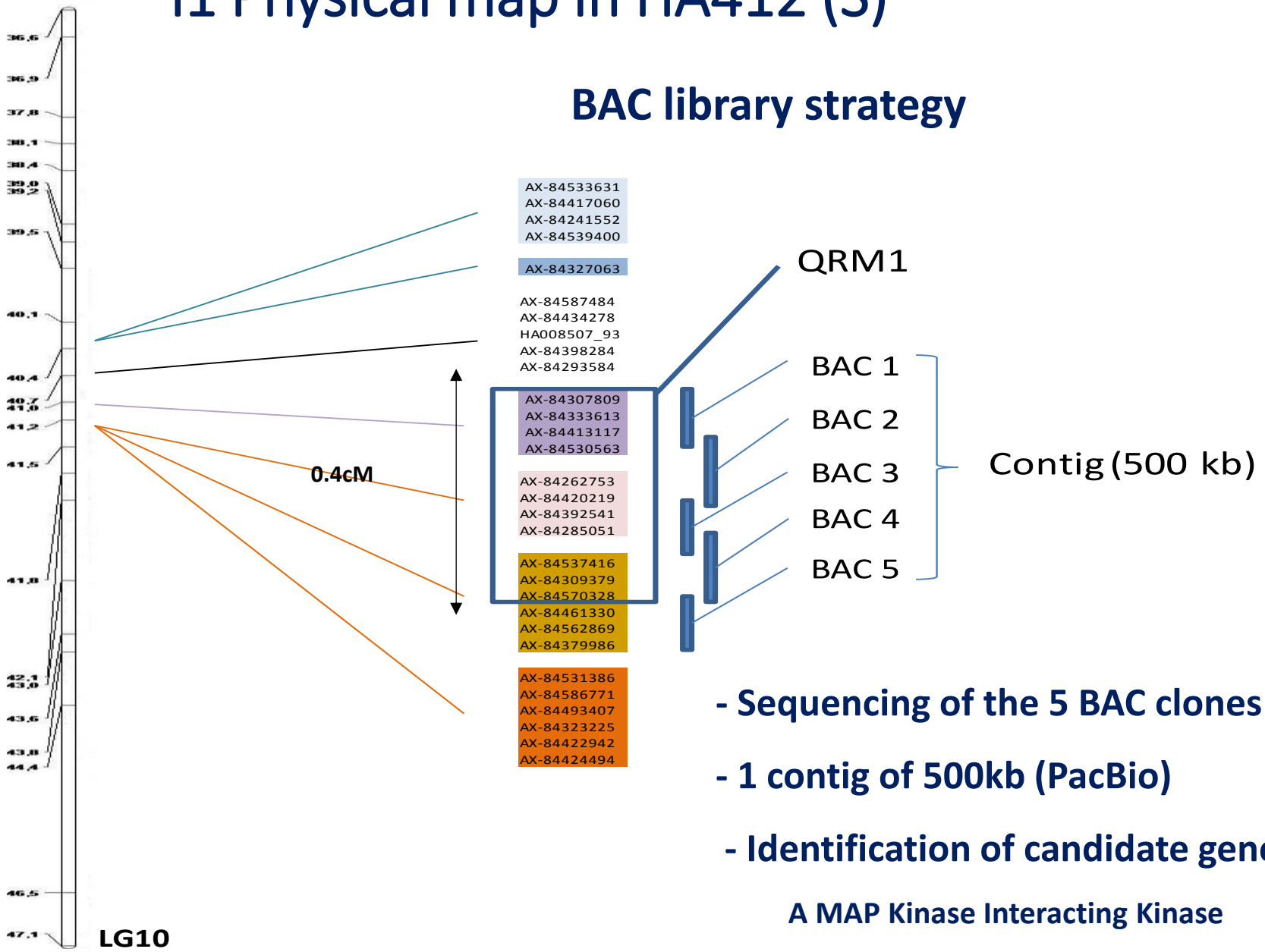
Focusing on a genomic region of interest in Sunflower


S. Munos S. Vautrin

- QRM1 controls quantitative resistance to downy mildew
  Susceptible (HA412) /Resistant (XRQ)
- Establishment of a genetic map (0.4 cM window on LG10)
- Markers definition on the QMR1 locus
- XRQ : *in silico* analysis of the 2Mb sequence on chromosome 10 (based on 20 markers alignment) composed of 14 scaffolded Pacbio contigs separating by N gaps (10k missing nucleotides)

QRM1 Physical map in HA412 (S)

BAC library strategy

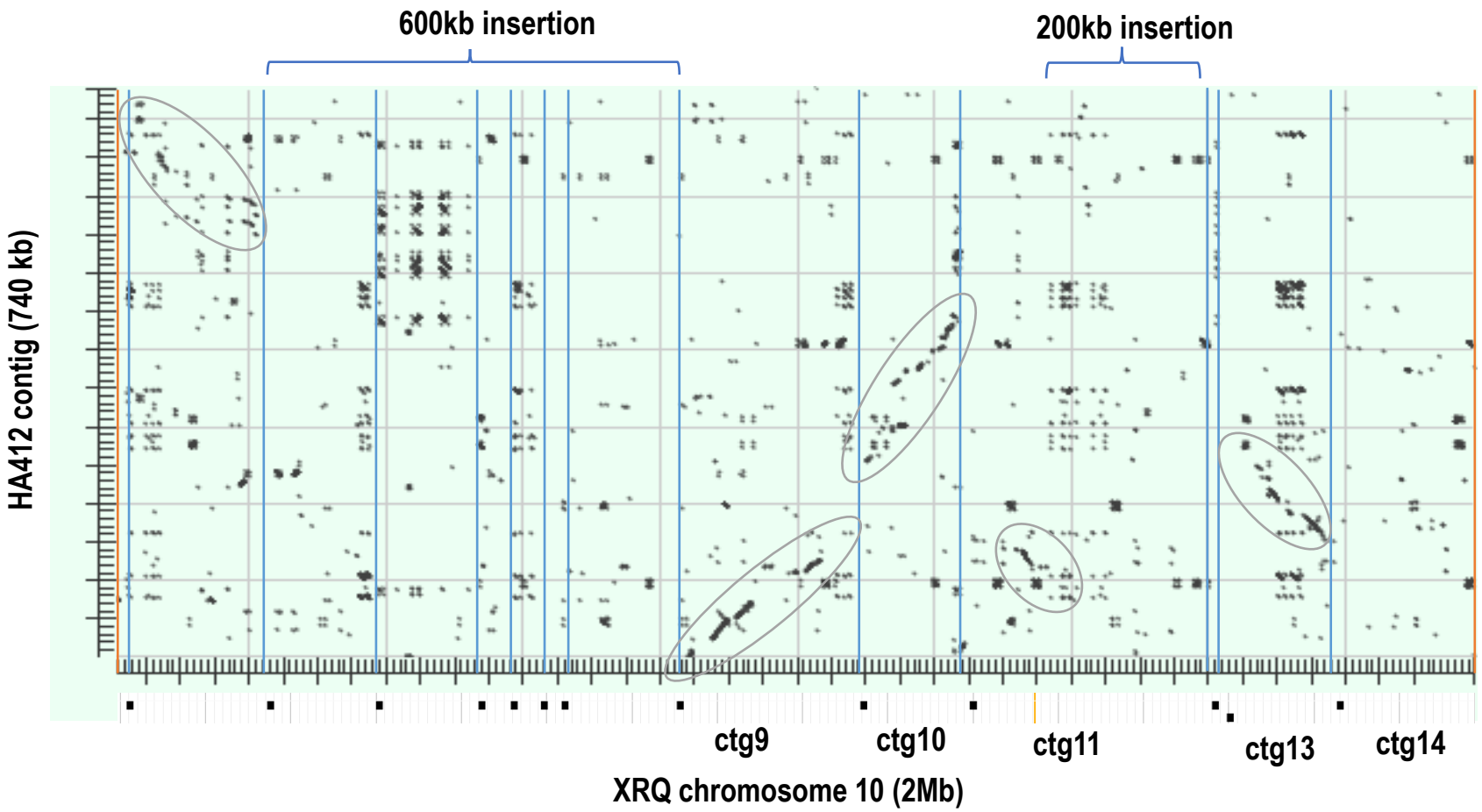


- **Sequencing of the 5 BAC clones (HA412)**
- **1 contig of 500kb (PacBio)**
- **Identification of candidate genes:**

  A MAP Kinase Interacting Kinase

  A Proteinase inhibitor

Comparison of the XRQ genome vs HA412 BAC clones



Low collinearity

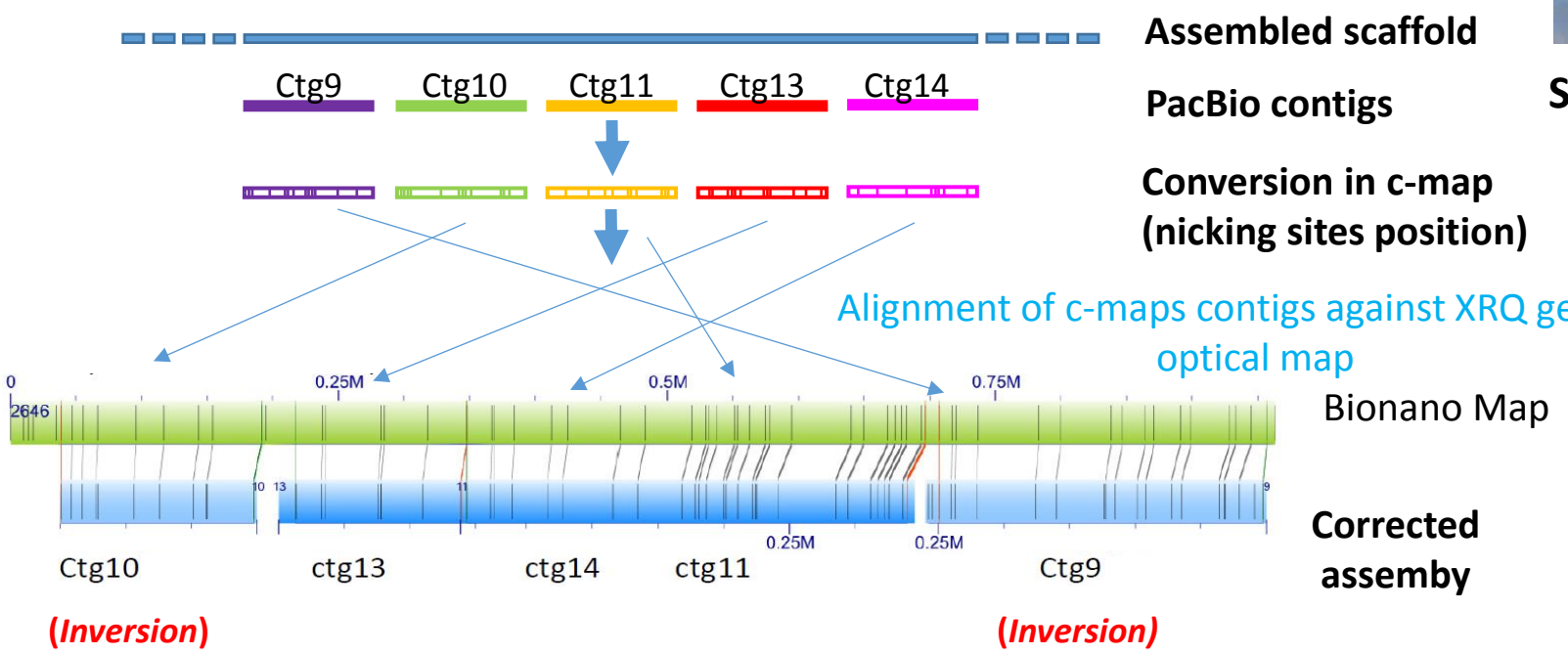
Fragmented alignment / orientation inconstancies :

Scaffolding errors OR true variability?

Optical maps to solve conflicts in the assembly

S. Cauet

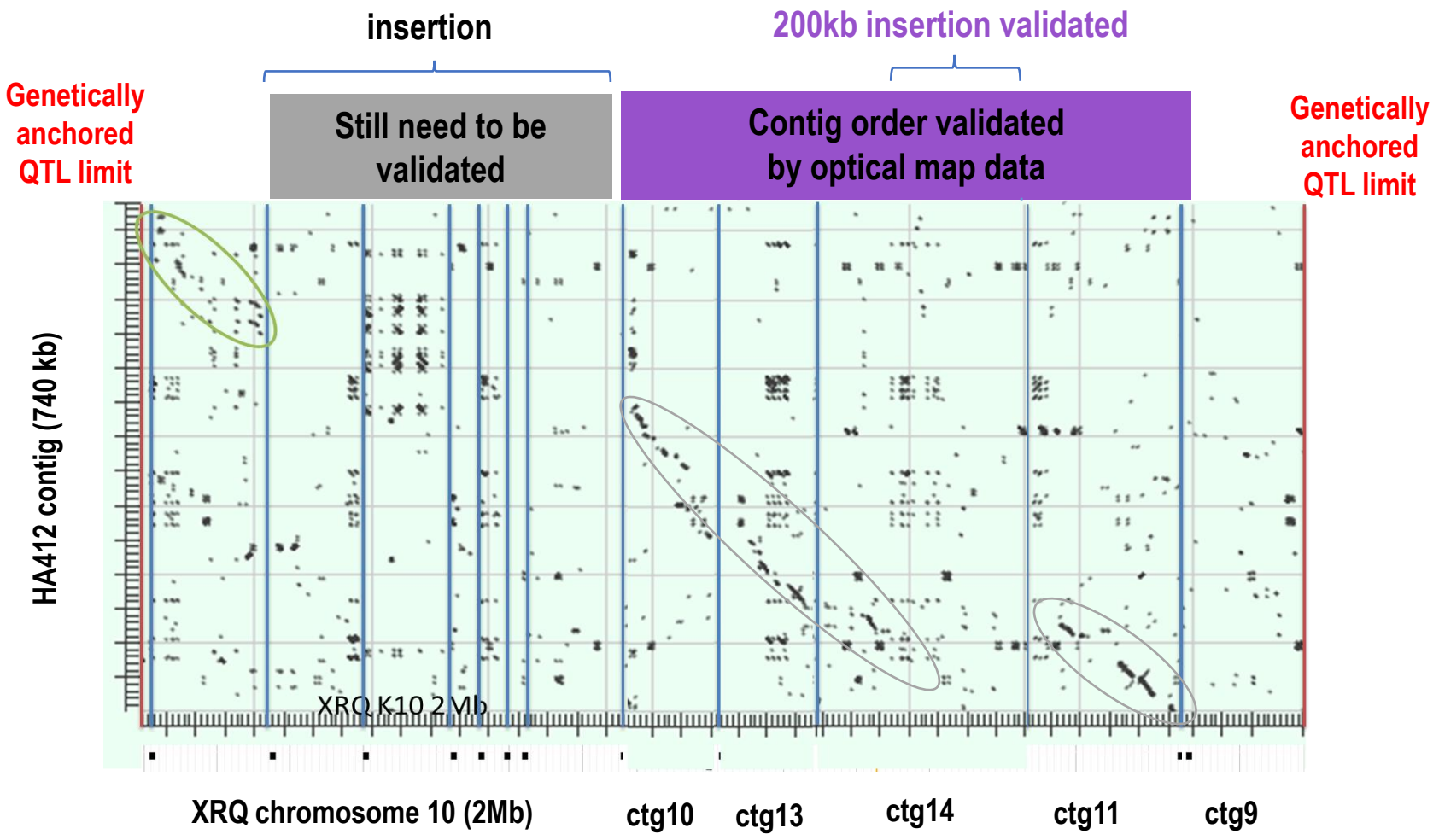
Alignment of the contig against the BioNano assembly of XRQ genome



On this targeted region, Optical Bionano map allowed:

- **to orientate some contigs**
- **to correct scaffolding of the PacBio contigs**

Comparison of the XRQ genome vs HA412 BAC clones

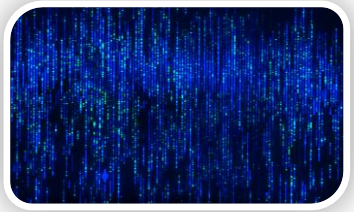


- Validation of the collinearity between XRQ and Ha412 sequences on QRM1 locus
- High variability observed: 2 major insertions of several hundreds kb in XRQ
- Annotation of the 2 sequences and comparative analysis are under progress (9 candidates genes have been identified)

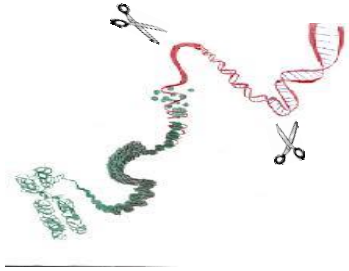
Genomics to help agriculture facing challenges

Toward a better understanding of plant genome's structure by combining complementary approaches

NGS
Ref sequence genomes

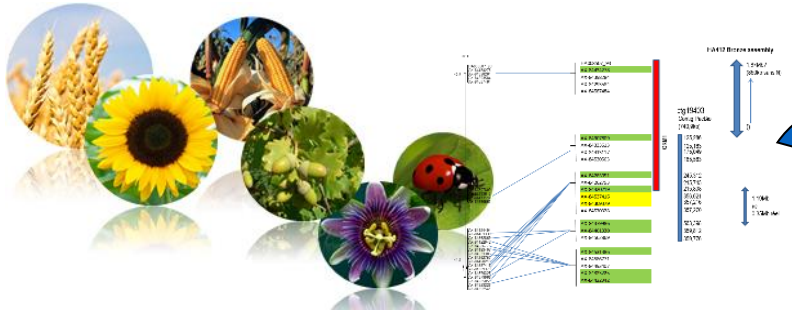


Optical maps



BAC library
Sequence Capture

Dedicated tools to better understand the role of regions of interest

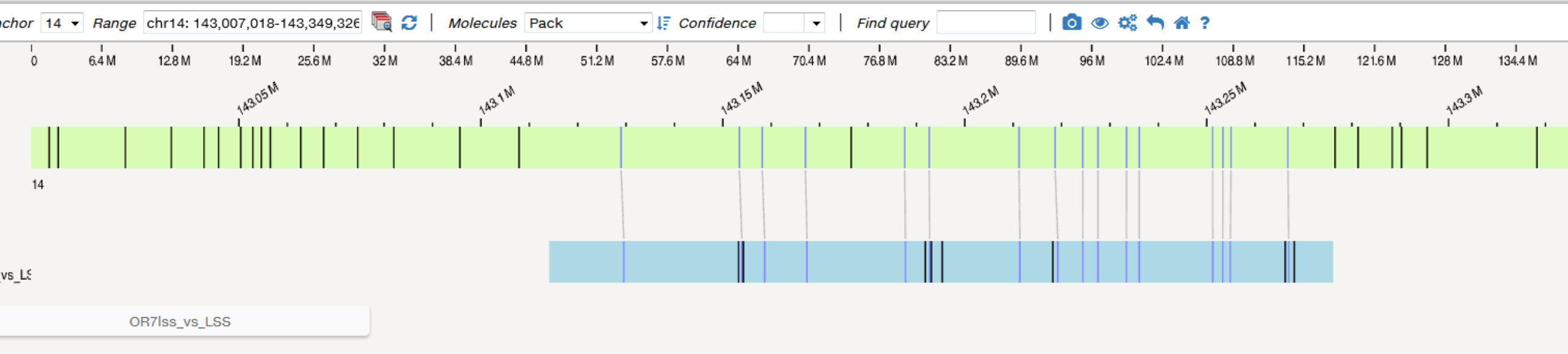


- Genetic map
- Specific markers available in the region of interest
- Physical map established on other genotypes

1. Optical maps
2. BAC library from various genotypes
3. Sequence Capture

- Physical caracterisation of regions of interest
- Isolation of the region of interest
- Identification of the region
- Comparison with reference map

Sequencing (NGS)
Comparison


chor
14
Range
chr14: 143,007,018-143,349,326
Molecules
Pack
Confidence
Find query
0
6.4 M
12.8 M
19.2 M
25.6 M
32 M
38.4 M
44.8 M
51.2 M
57.6 M
64 M
70.4 M
76.8 M
83.2 M
89.6 M
96 M
102.4 M
108.8 M
115.2 M
121.6 M
128 M
134.4 M
143.05 M
143.1 M
143.15 M
143.2 M
143.25 M
143.3 M
14
vs_LS
OR7lss_vs_LSS

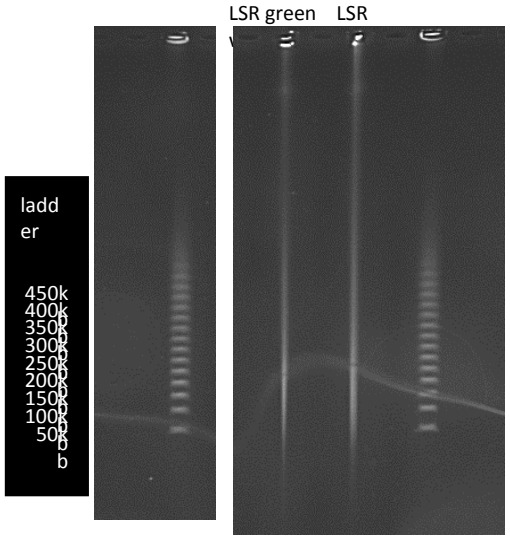
Alignment comparison

Sunflower DLE1 optical maps

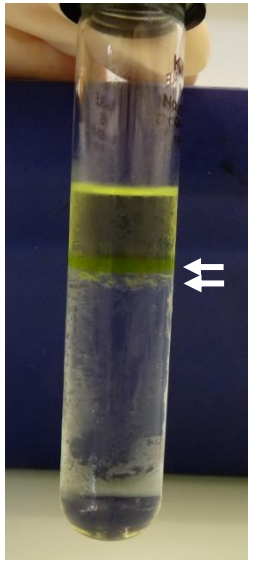
DNA extraction : key step

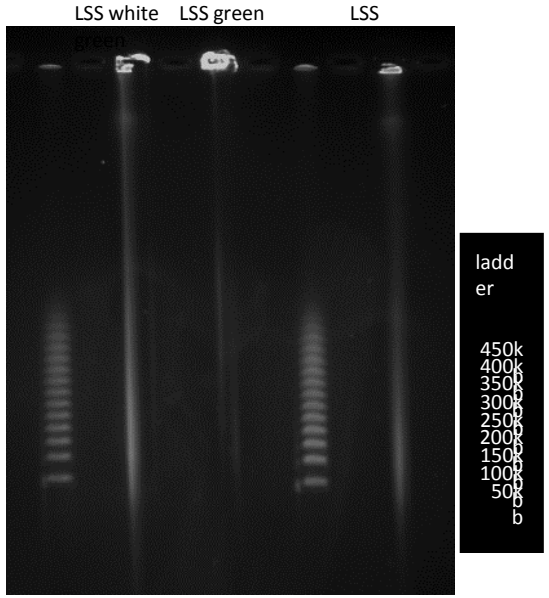
Sample	LSR (120X)	LSS (70X)	LSS (120X)
Labelled with	DLE-1	DLE-1	DLE-1
Data collected (molecules >150 kb)	362 Gbp	214 Gbp	345 Gbp
Molecules N50 (molecules > 150 kb)	299 kbp	314 kbp	317 kbp
Assembly size	3.07 Gbp	3.05 Gbp	3.05 Gbp
Genome map N50	9.88 Mbp	21.27 Mbp	174.93 Mbp

LSR007



LSS007





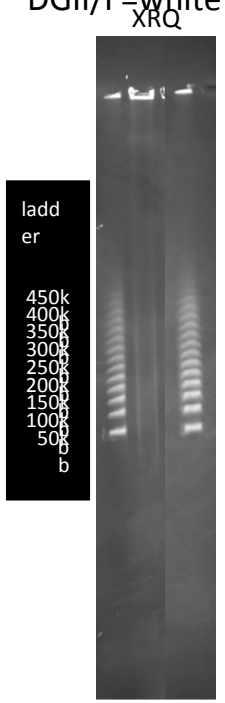
How to check the best Q of the HMW gDNA?

DNA extraction : key step

Sample	XRQ	Arikara	NB1	Broomrape
Data collected (molecules >150 kb)	310 Gbp	319 Gbp	300 Gbp	229 Gbp
Assembly size (non-haplotype-aware)	3.06 Gbp	4.64 Gbp	6.89 Gbp	3.31 Gbp
Genome map N50	175 Mbp	2.24 Mbp	4.33 Mbp	9.96 Mbp

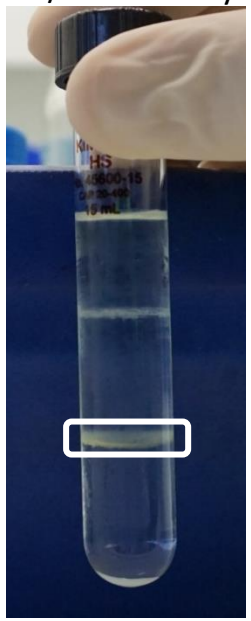
XRQ

1 layers
DGII/I =white layer



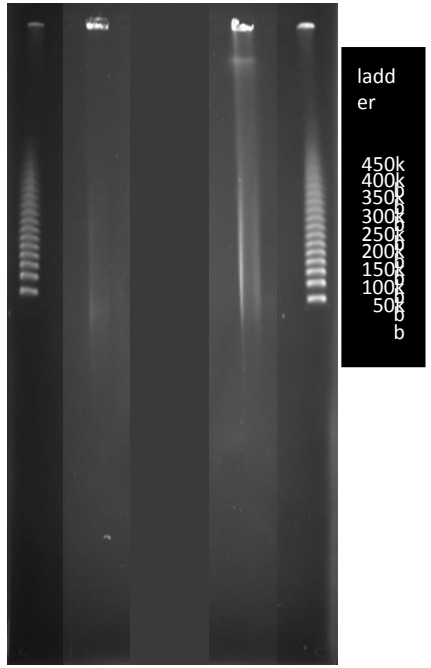
Broomrape

2 layers
DGI/Ech = white layer
DGII/I = white layer



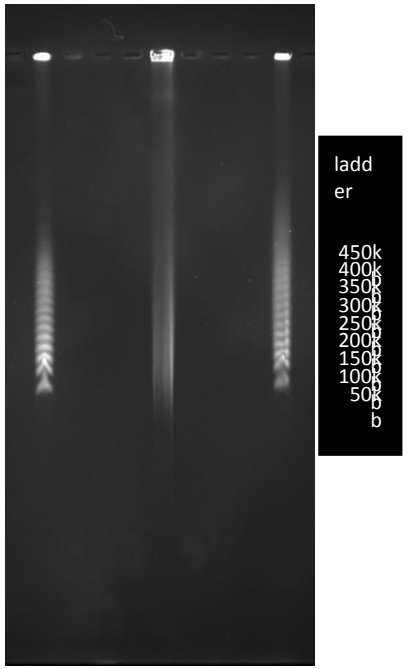
NB1

2 layers
DGI/Ech = green layer
DGII/I = light green layer



arikara

2 layers
DGI/Ech = dark green
DGII/I = green



Criteria to check the best Q of the HMW gDNA? Size? Range of sizes? Purity?

There is no standard protocole but we need standard procedure to check Q

Comparison of NRLS and DLS technologies on the building of LSR and LSS optical maps

	LSR007 Assembly NRLS labelling[1]		LSS007 Assembly NRLS labelling[1]	
Réalisation	CNRGV		CNRGV	
Median length (Mb)	0,76		0,59	
Mean length (Mb)	1,0		0,79	
N50 length (Mb)	1,4		1,1	
Total length (Gb)	3,2		3,1	
Effective coverage of assembly (x)	88,9		94,3	

[1] : **N**ick, **R**epair, **L**abel, **S**tain
[2] : **D**irect **L**abel and **S**tain (non compatible with Irys system)

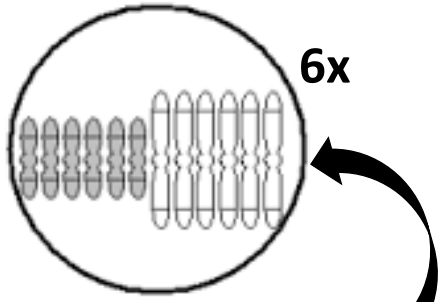
→ Better metrics and contiguity with the new chemistry DLS

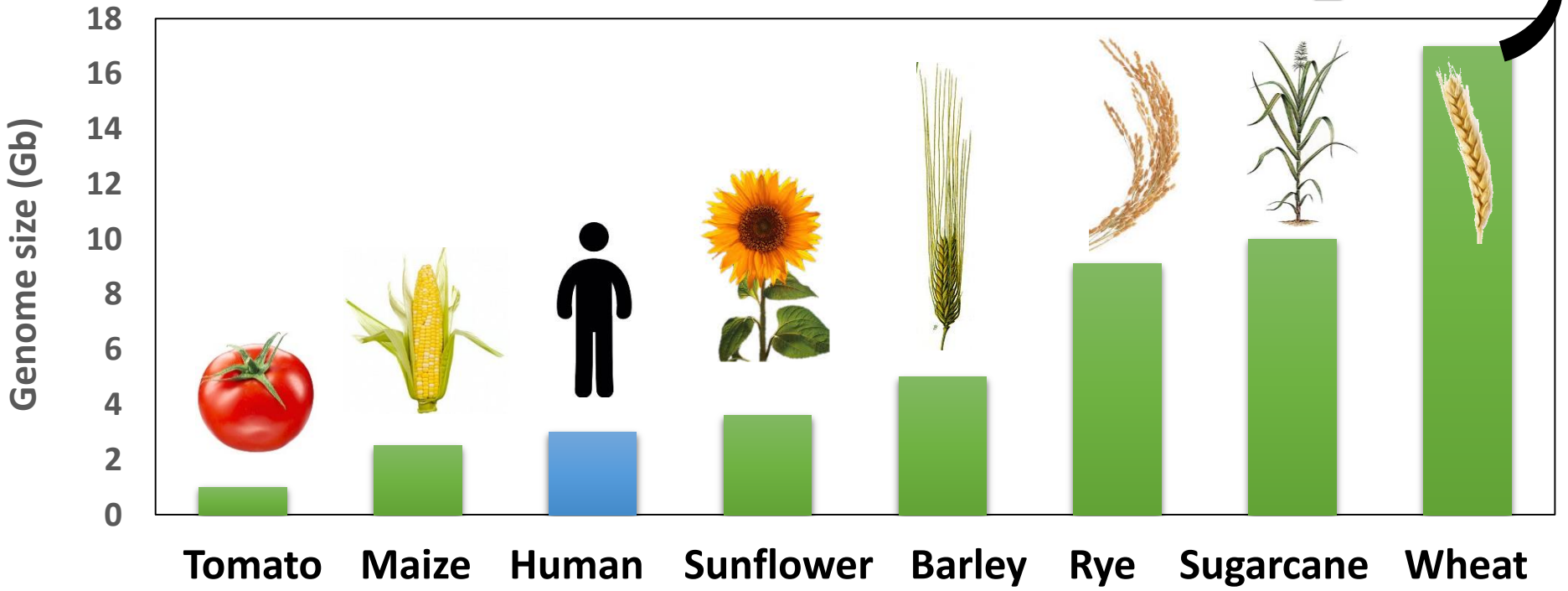
Plant's genome exhibits high levels of complexity

→ Large genome size

→ High level of transposable elements

→ Polyploïdy





% of repeat elements

Various targets for crop improvement

- **Yield potential and yield stability**
  - ✓ Reduce inputs…

- **Adaptation to climate change**
  - ✓ abiotic tolerance…

- **Durable resistance to biotic stress**
  - ✓ Virus, fungi, new desases…

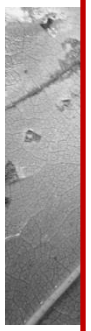
- **Quality**
  - ✓ Grain protein content, nutritional needs …

Various targets for crop improvement

➢ Yield potential and yield stability

Most research projects aim at linking genotype to specific phenotype :

- **Exhaustive sequence information on whole genome not required**
- **Reliable and quality information of the specific region necessary**
- **Structural variations related to a phenotype are essential to understand biological process (important genetic diversity)**



...

Genome assembly improvement to help linking genotype / phenotype

- Sunflower proves again to be a highly complex genome, showing very high diversity between genotypes

One reference genome is not enough!

- Despite long reads sequencing, assembly (scaffolding) has to be checked when working on reference genomes

- Optical map allowed to validate major rearrangements between the 2 genotypes

- Proven interest of complementary approaches (NGS – optical map – BAC)