Toward a better understanding of plant genomes structure: combining NGS and optical mapping technology to improve the sunflower assembly

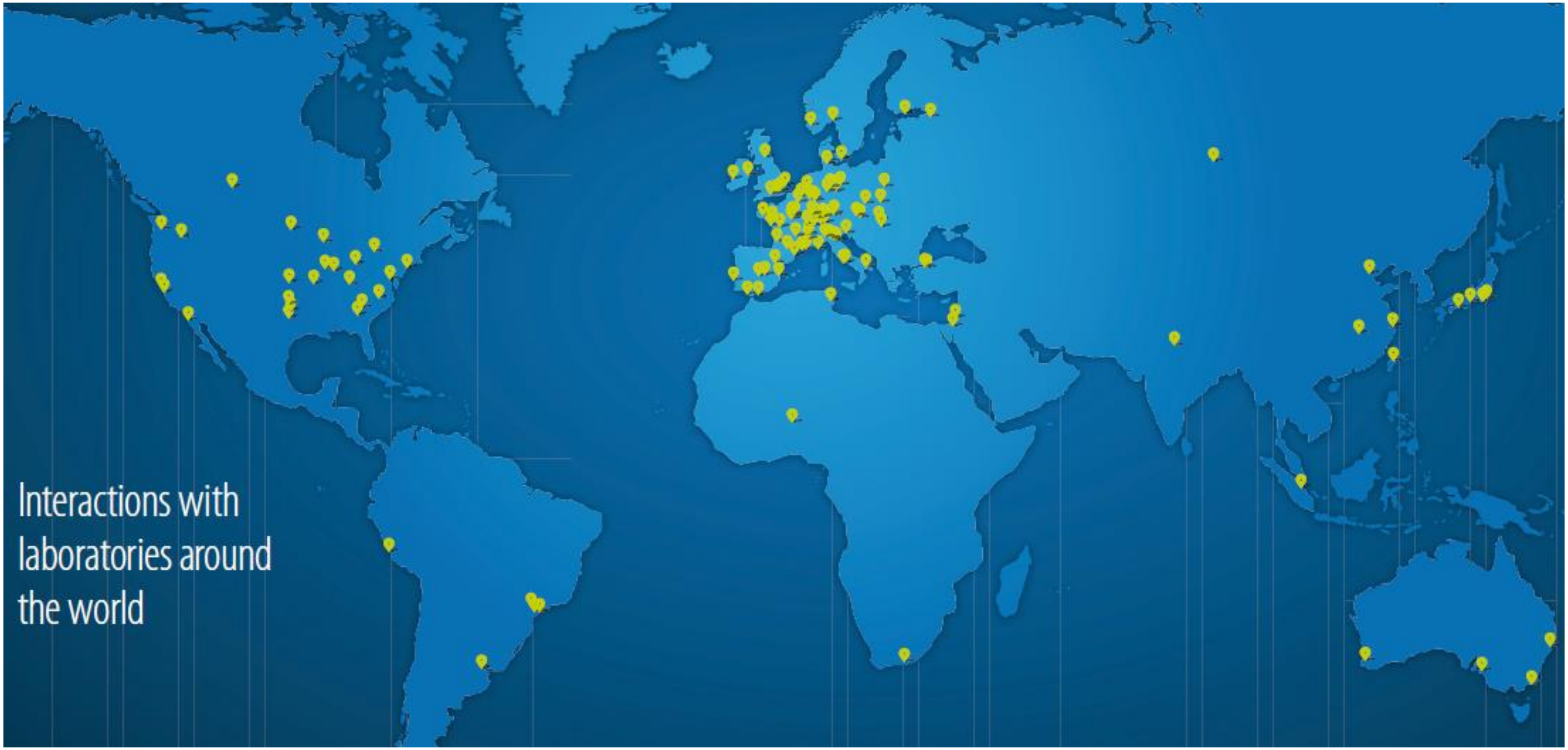
Céline CHANTRY-DARMON

CNRGV
The French Plant Genomic Center

- Created in 2004 by INRA
- A dedicated structure to assist plant genomic programs
  - ➢ Distribute the genomic resources at the international level
  - ➢ Provide high quality research material and efficient tools and services
  - ➢ Develop genomic projects in collaboration
  - ➢ Host scientists
  - ➢ Develop innovative solutions

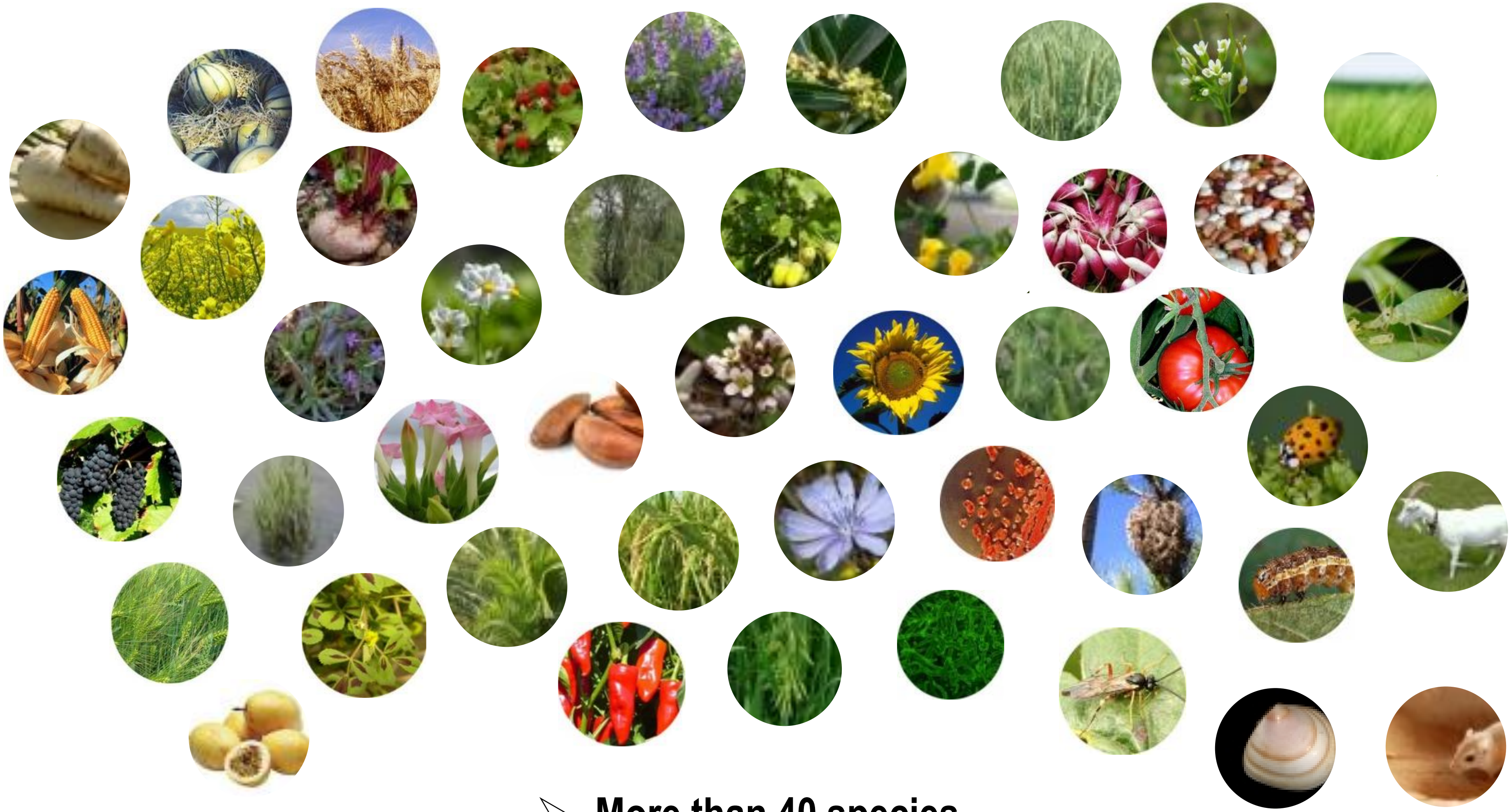
ISO 9001:2008
Octobre 2005

Interactions with laboratories around the world



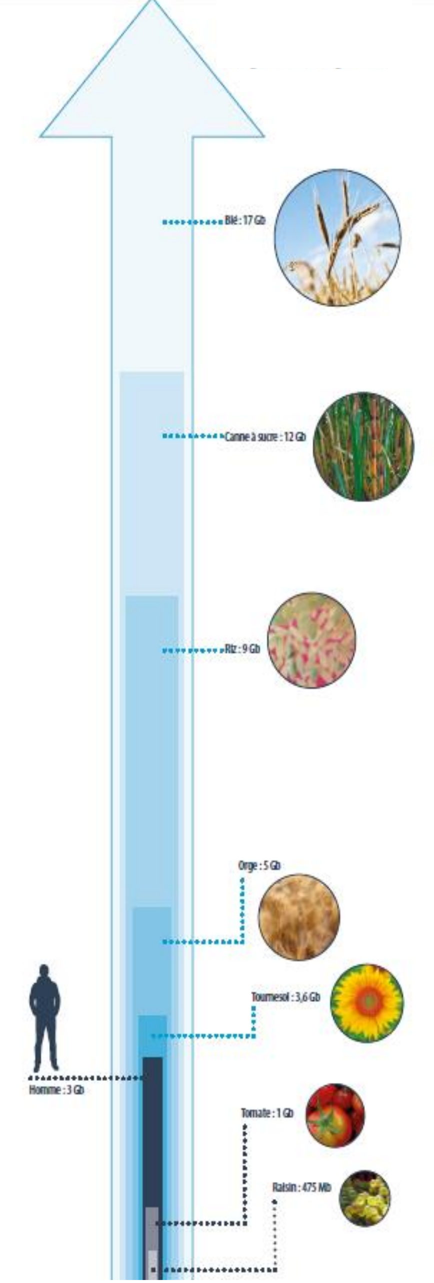
- More than 3 millions BAC clones distributed during the last 5 years

Plants project diversity

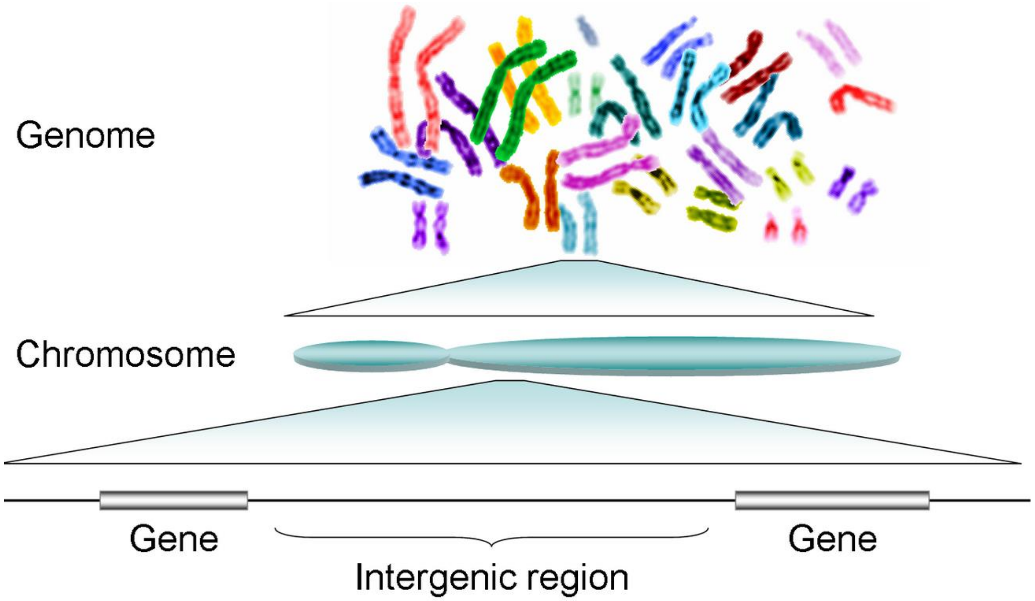


- More than 40 species

The goal for the Plant Genomic Center



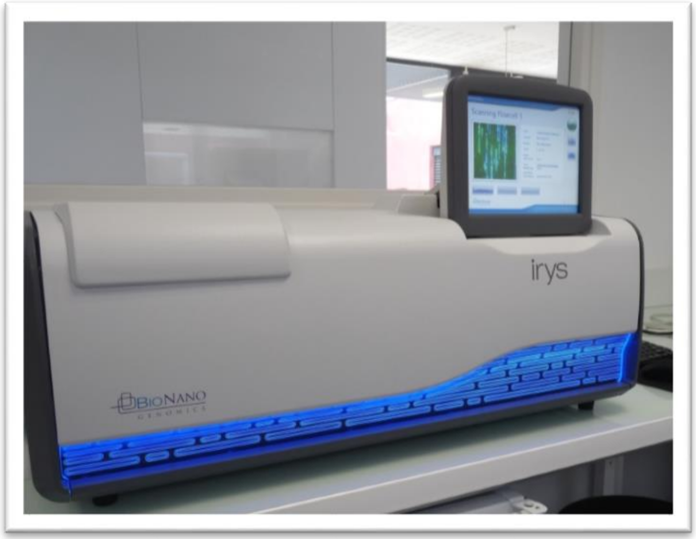
- Large genome size
- Repeats elements
- Polyploidy

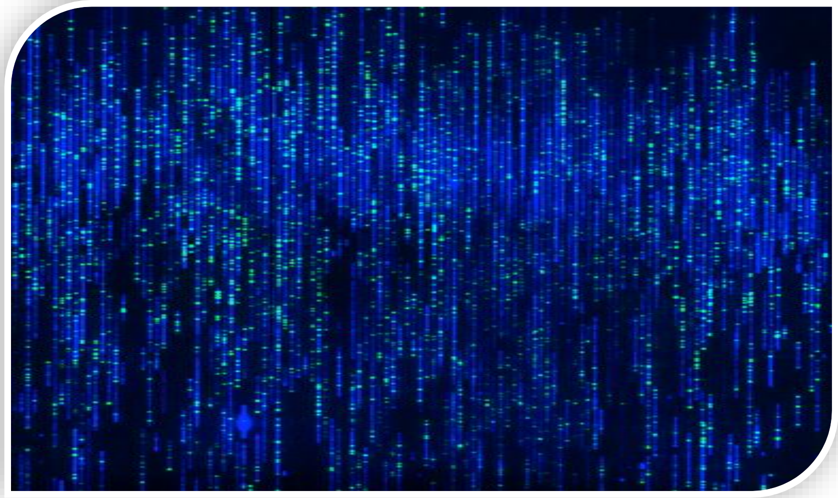


- ➢ Manage genome size and diversity
- ➢ Decrease genome complexity
- ➢ Target genomic region of interest

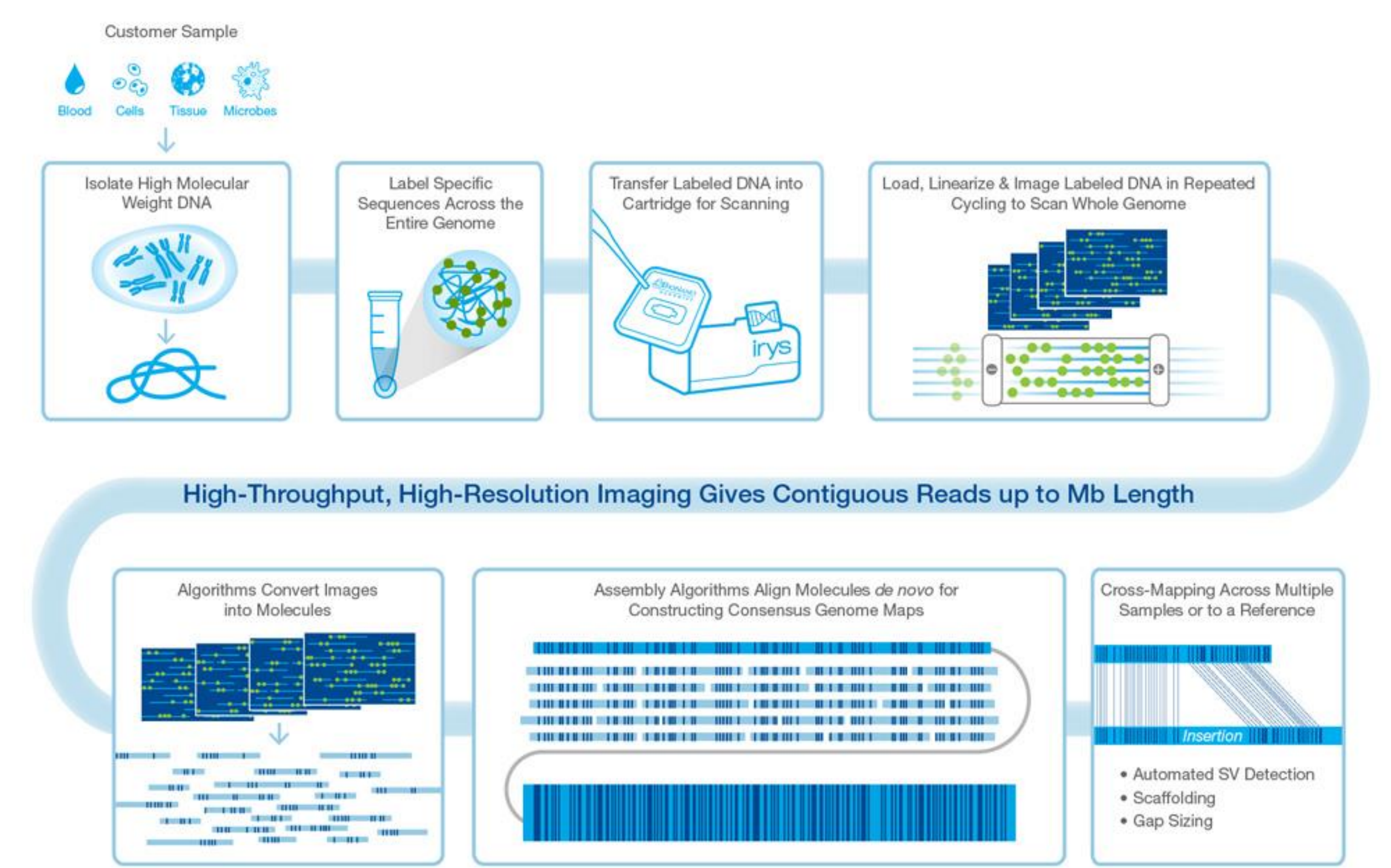
Focus on the optical mapping with BioNano

- The Bionano Irys system: a new tool to study complex genomes
- Advantages of BioNano optical mapping:
  - Direct visualization of long DNA molecules (>100 kb)
  - Provides real physical distance information





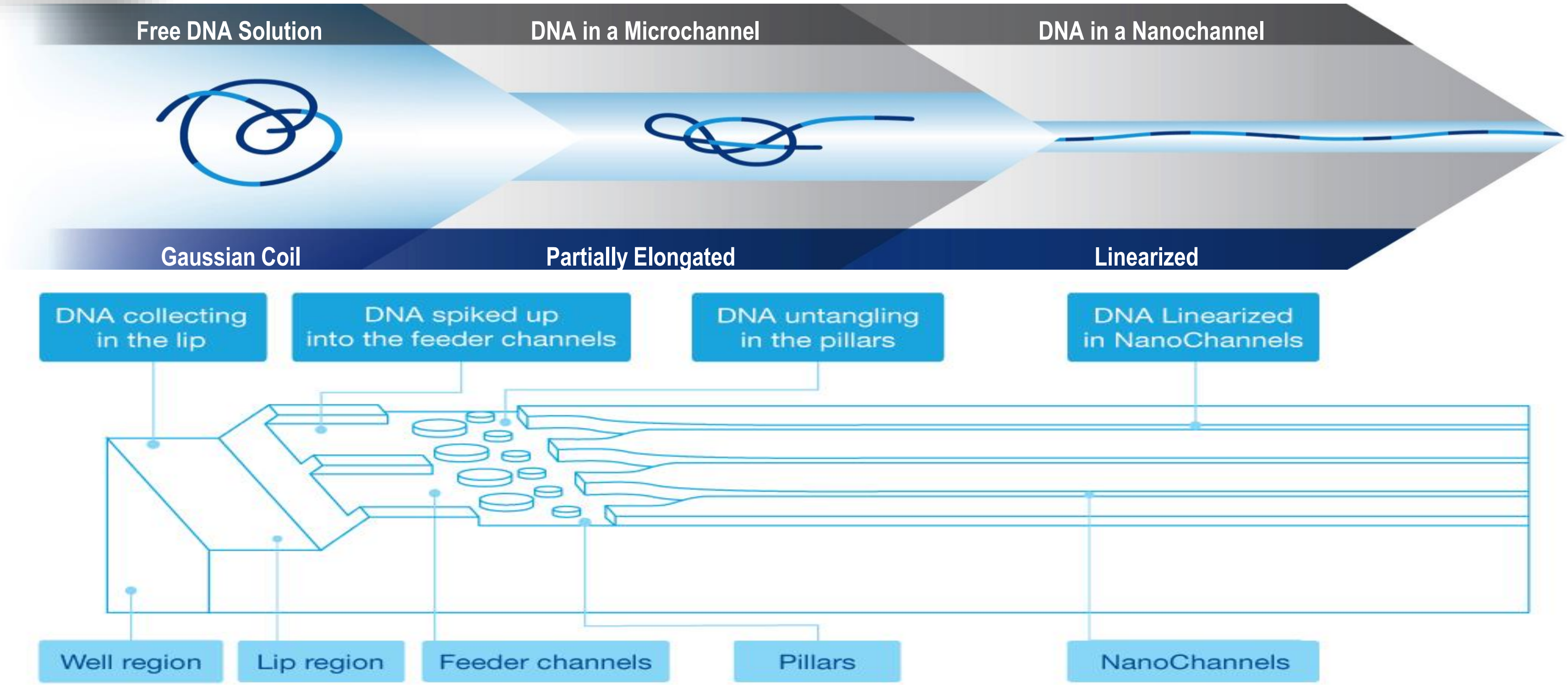
The Workflow

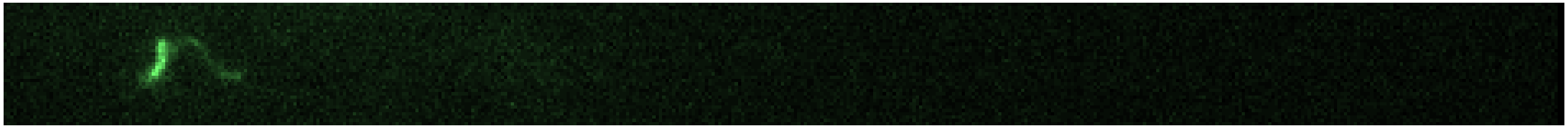

Customer Sample
Blood
Cells
Tissue
Microbes
Isolate High Molecular Weight DNA
Label Specific Sequences Across the Entire Genome
Transfer Labeled DNA into Cartridge for Scanning
irys
Load, Linearize & Image Labeled DNA in Repeated Cycling to Scan Whole Genome
High-Throughput, High-Resolution Imaging Gives Contiguous Reads up to Mb Length
Algorithms Convert Images into Molecules
Assembly Algorithms Align Molecules *de novo* for Constructing Consensus Genome Maps
Cross-Mapping Across Multiple Samples or to a Reference
Insertion
• Automated SV Detection
• Scaffolding
• Gap Sizing

- ➢ **50Gb data generated per flowcell (=> 100Gb / chip)**

BIONANO GENOMICS

HMW DNA molecules in nanochanels



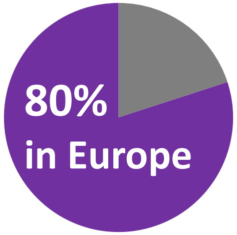


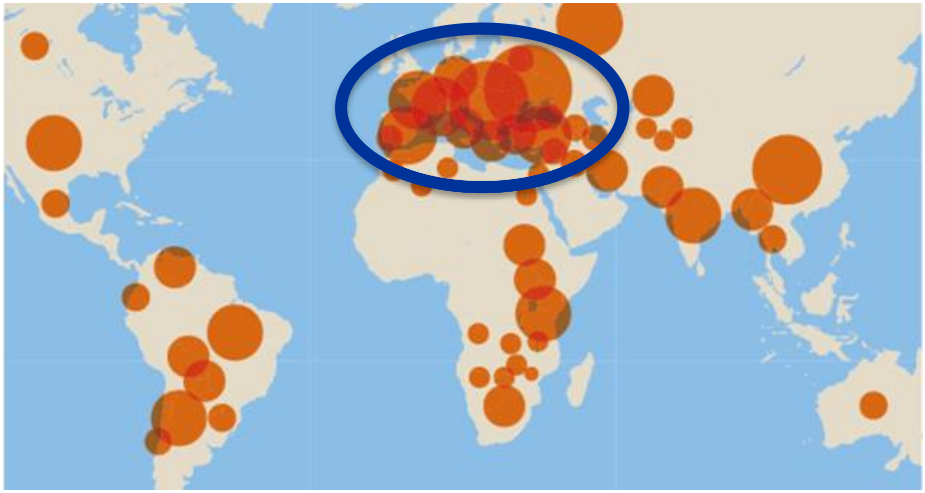
- HMW DNA molecules from 100kb to >2Mb

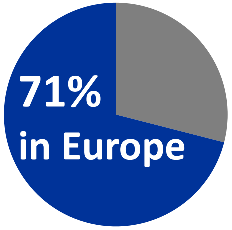
The Sunflower : an important crop for Europe

39 Million tons of seed produced worldwide

30 Million hectares worldwide



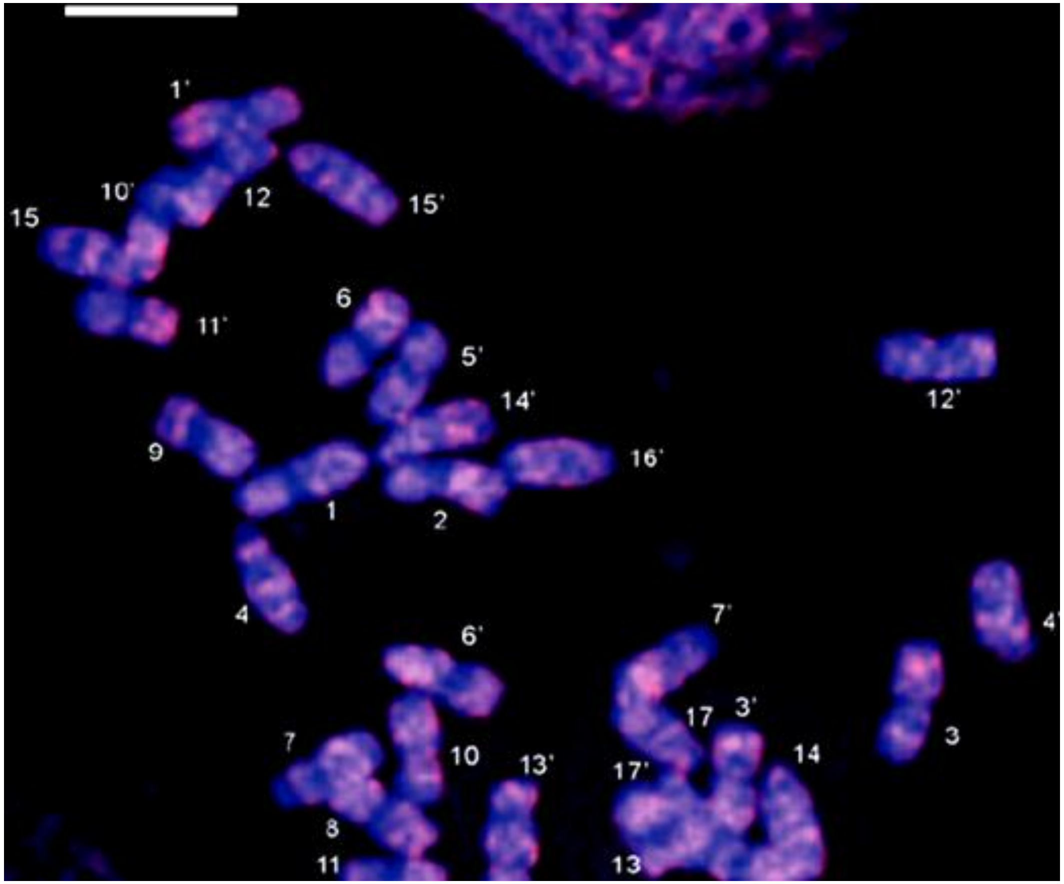




Societal challenge

The global production of sunflower seeds has to increase to meet growing demand *(human food, animal feed, green chemistry...)*

The Sunflower genome



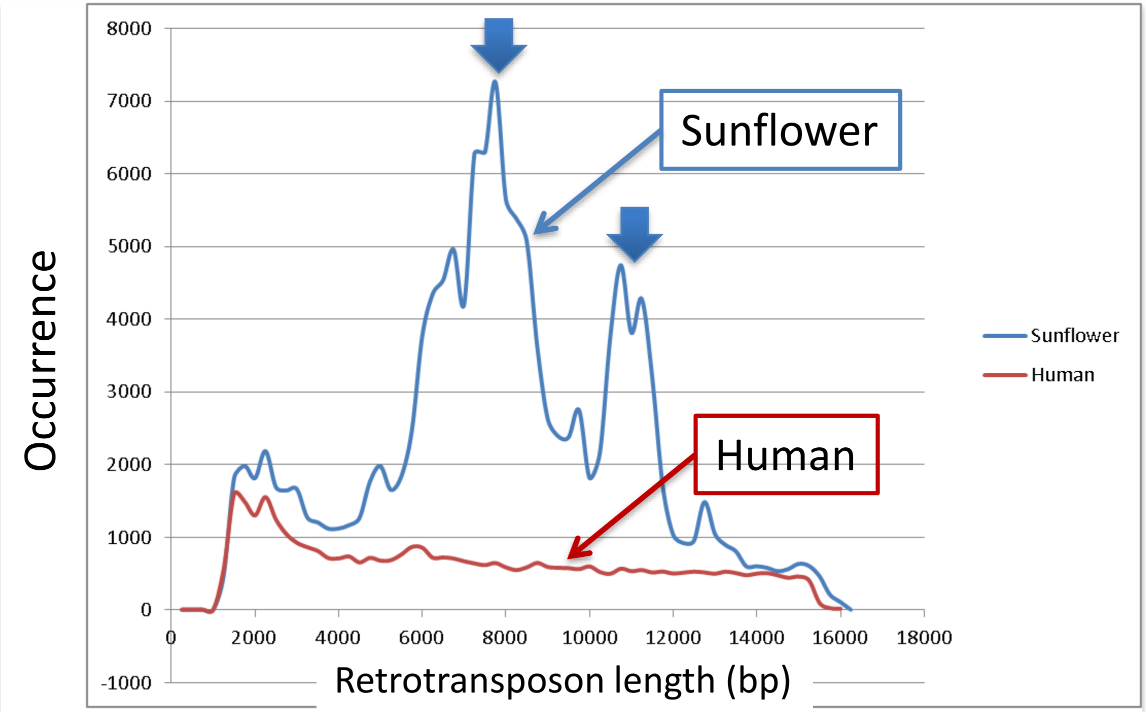
- *Helianthus annuus*
- 3.6 Gb
- 2n=34 chromosomes

Cytological characterization of sunflower by in situ hybridization using homologous rDNA sequences and a BAC clone containing highly represented repetitive retrotransposon-like sequences

P. Talia, E. Greizerstein, C. Díaz Quijano, L. Peluffo, L. Fernández, P. Fernández, H.E. Hopp, N. Paniego, R.A. Heinz, and L. Poggio

Sunflower genome contains long repeated sequences

Length distribution of LTR retrotransposons



LTRharvest (Ellinghaus *et al.* 2008, default parameters)

J. Gouzy

Repeats = 33% of the sunflower genome

Repeats = 8% of the Human genome

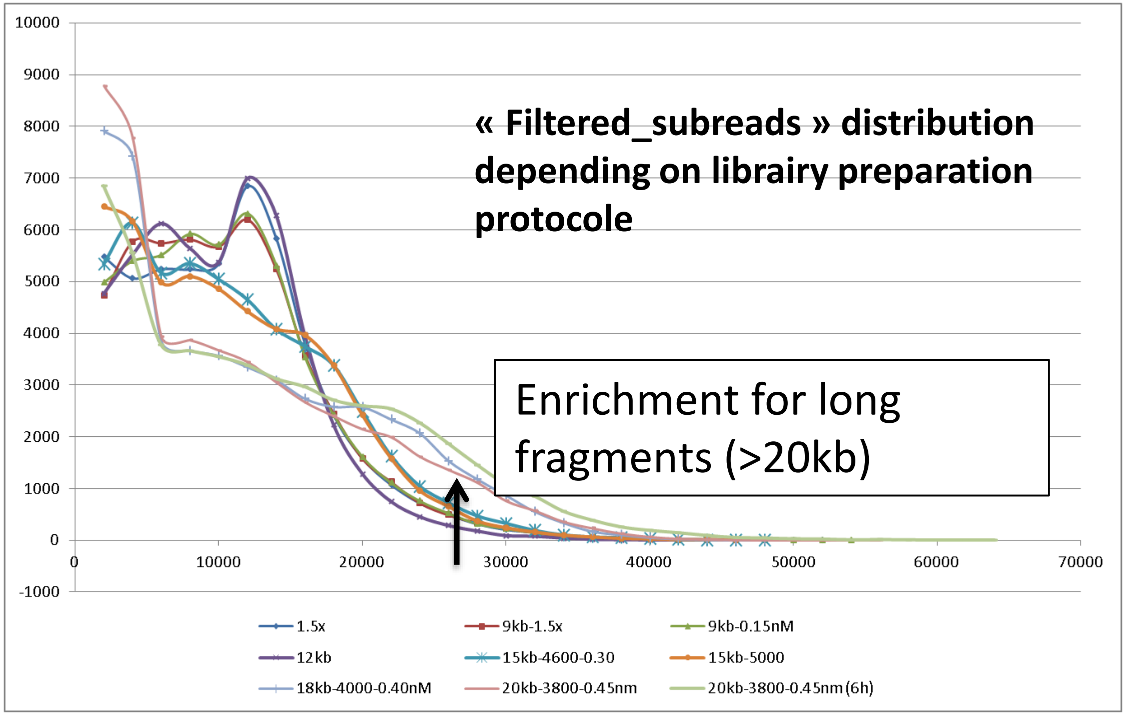
Two major repeats in the sunflower genome:
8 kb and 11.5 kb

The repeats make the assembling very difficult

Development of long-fragment libraries

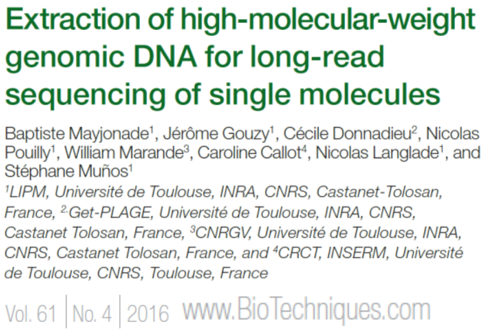
The longer the PacBio sequences are, the better it is to cross the LTR :

- New DNA extraction protocol
- Optimization of fragmentation, purification, loading
- Increase running time from 4 to 6h



« Filtered_subreads » distribution depending on librairy preparation protocole

Enrichment for long fragments (>20kb)



Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules

Baptiste Mayjonade[1], Jérôme Gouzy[1], Cécile Donnadieu[2], Nicolas Pouilly[1], William Marande[3], Caroline Callot[4], Nicolas Langlade[1], and Stéphane Muños[1]

[1]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France, [2]Get-PLAGE, Université de Toulouse, INRA, CNRS, Castanet Tolosan, France, [3]CNRGV, Université de Toulouse, INRA, CNRS, Castanet Tolosan, France, and [4]CRCT, INSERM, Université de Toulouse, CNRS, Toulouse, France

Vol. 61 | No. 4 | 2016 www.BioTechniques.com

B. Mayjonade

PacBio Genome Assembly

N. Langlade

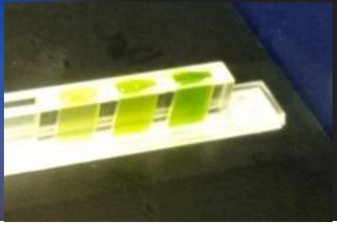
- XRQ sunflower line
- Genome sequence >100X PacBio

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

➢ 80% of the genome inside contigs

SUNRISE
UNE CULTURE POUR LE FUTUR

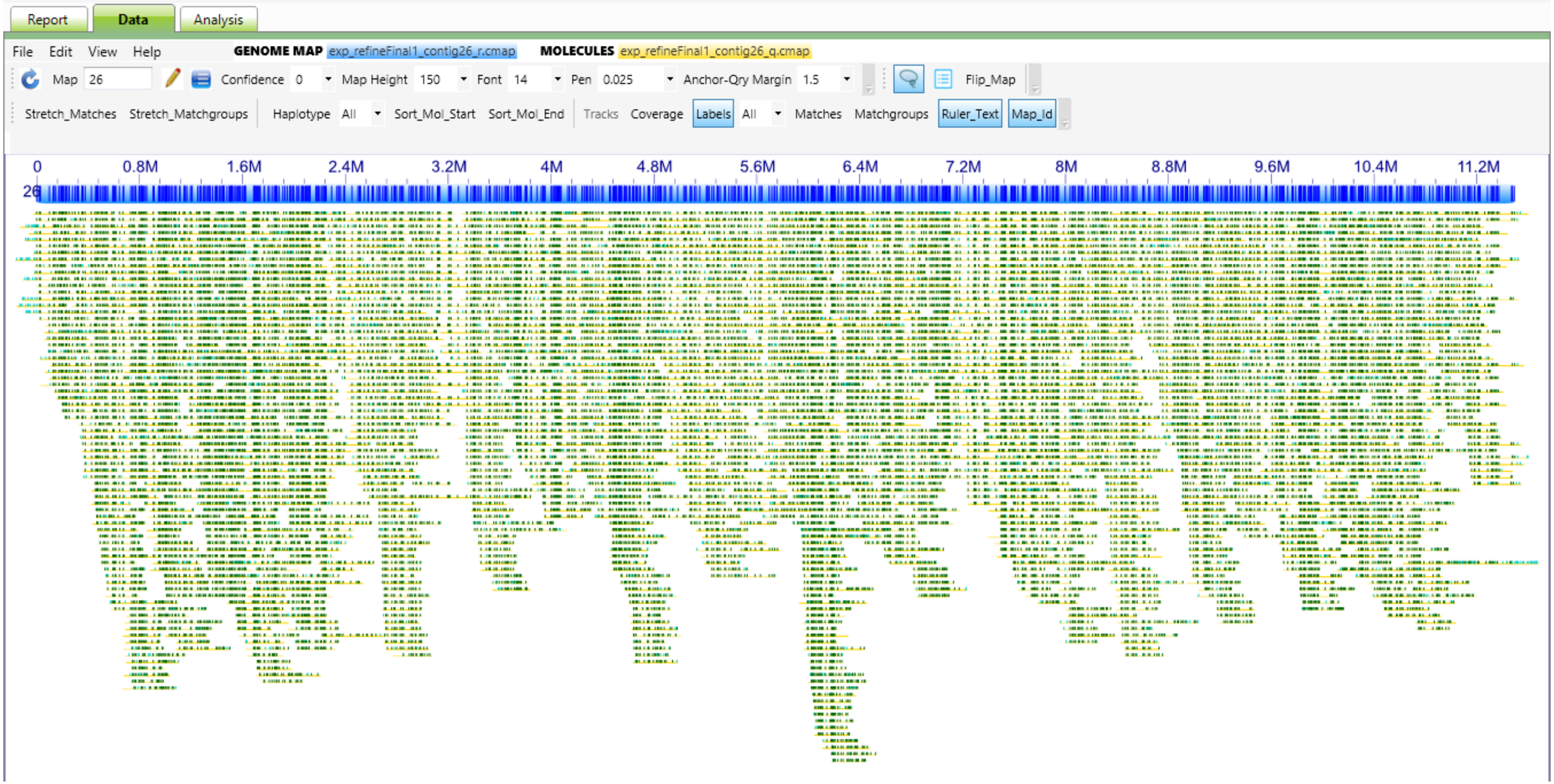
BioNano analyses



- HMW DNA Extraction of fresh young dark treated leaves
- 2 nicking enzymes (BspQ1 & BssS1)

	BspQ1	Bsss1
	5′...GCTCTTCN▼...3′ 3′...CGAGAAGN...5′	5′...CACGAG...3′ 3′...GTGCT▲C...5′
Theoretical nb labels / 100kb	7,2	17,2
Real nb labels / 100kb	6,4	12,8
Raw data (Gb)	846 (235X)	845 (235X)
Filtered data >100kb (Gb)	635 (176X)	600 (167X)
Molecules N50 (kb)	206	187

Example of a BioNano map



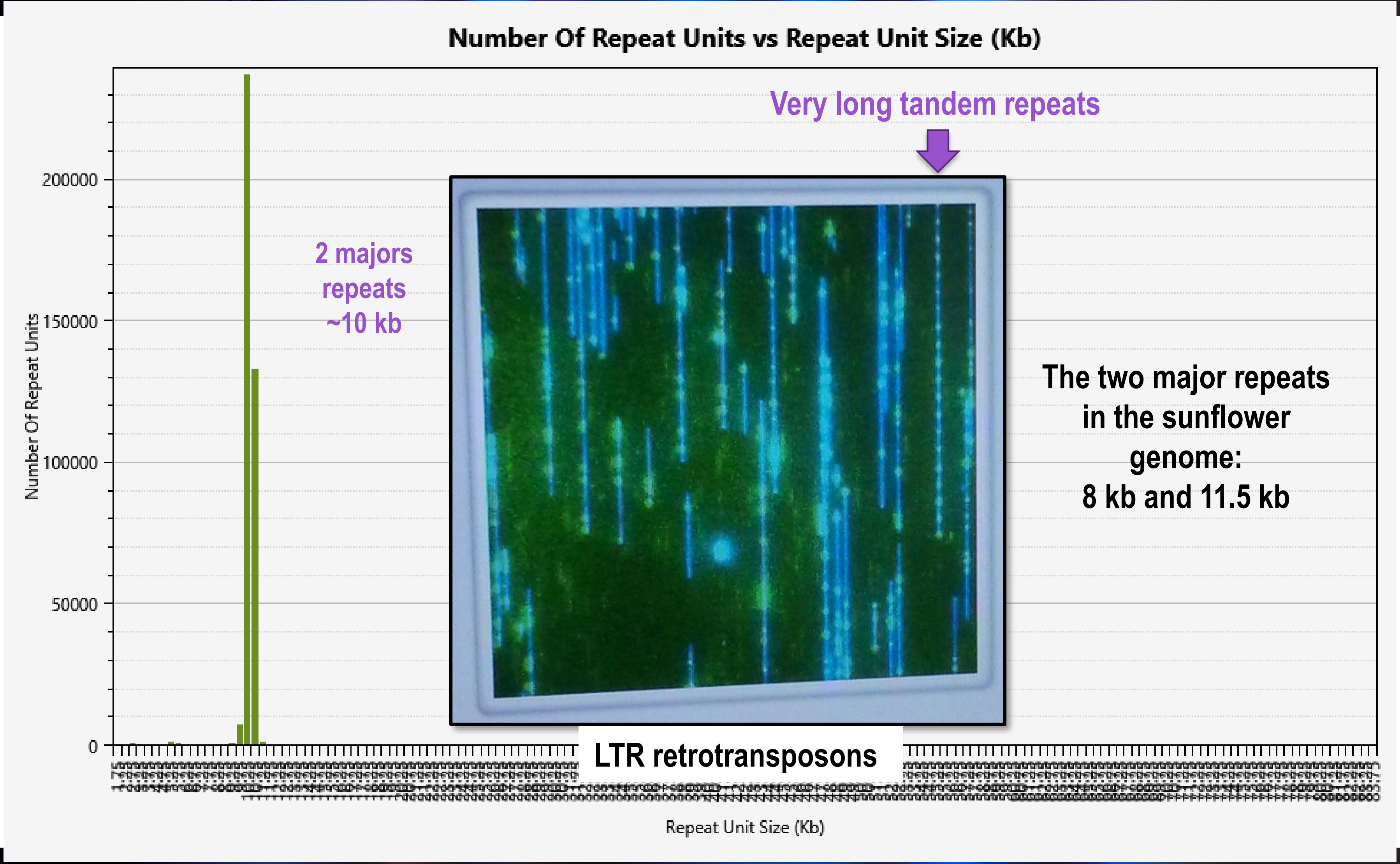
GENOME MAP exp_refineFinal1_contig26_r.cmap MOLECULES exp_refineFinal1_contig26_q.cmap

Map 26 Confidence 0 Map Height 150 Font 14 Pen 0.025 Anchor-Qry Margin 1.5

Haplotype All Labels All

- ➢ 176X coverage, molecules from 150kb to 2,3Mb

Visualization of the Sunflower repeats



Very long tandem repeats

2 majors repeats ~10 kb

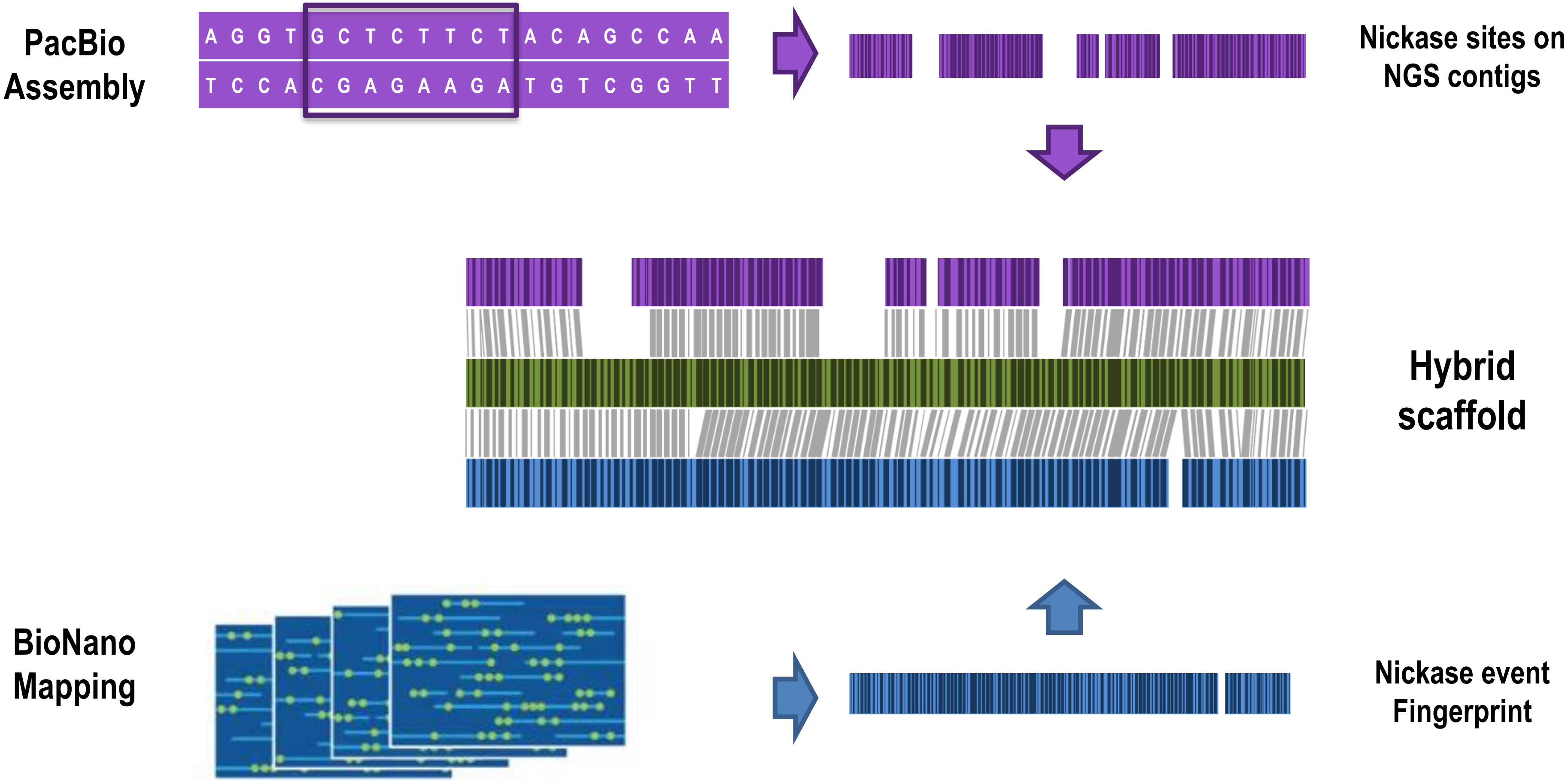
LTR retrotransposons

The two major repeats in the sunflower genome:
8 kb and 11.5 kb

PacBio assembly and BioNano Maps

	PacBio Assembly	BioNano BspQ1 Assembly	BioNano BssS1 Assembly
Count	12318	2228	4287
Median length (Mb)	0.120	0.999	0.551
N50 length (Mb)	0.524	1.979	0.968
Max length (Mb)	3.35	11.49	5.322
Total length (Mb)	2930	3191	3112
% genome coverage	81%	88%	86%

Hybrid Assembly


PacBio Assembly
A G G T G C T C T T C T A C A G C C A A
T C C A C G A G A A G A T G T C G G T T
Nickase sites on NGS contigs
Hybrid scaffold
BioNano Mapping
Nickase event Fingerprint

Sunflower Hybrid Assembly

	PacBio Assembly	BioNano BspQ1 Assembly	Hybrid scaffold
Count	12318	2228	1430
Median length (Mb)	0.120	0.999	1.442
N50 length (Mb)	0.524	1.979	2.87
Max length (Mb)	3.35	11.49	17.45
Total length (Mb)	2930	3191	2922
% genome	81%	88%	81%

Sunflower Hybrid Assembly

	PacBio Assembly	BioNano BspQ1 Assembly	Hybrid scaffold
Count	12318	2228	1430
Median length (Mb)	0.120	0.999	1.442
N50 length (Mb)	0.524	1.979	2.87
Max length (Mb)	3.35	11.49	17.45
Total length (Mb)	2930		2922
% genome	81%	88%	81%

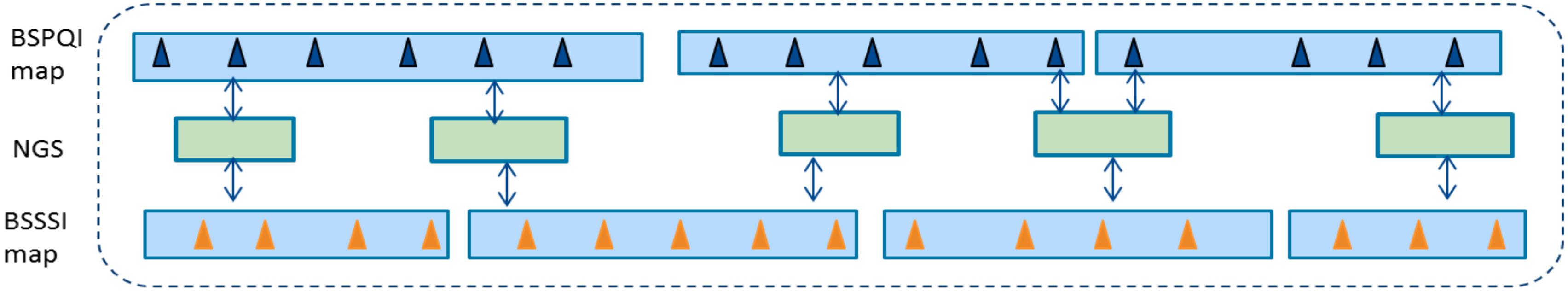
More than 5 fold increase

Sunflower Hybrid Assembly

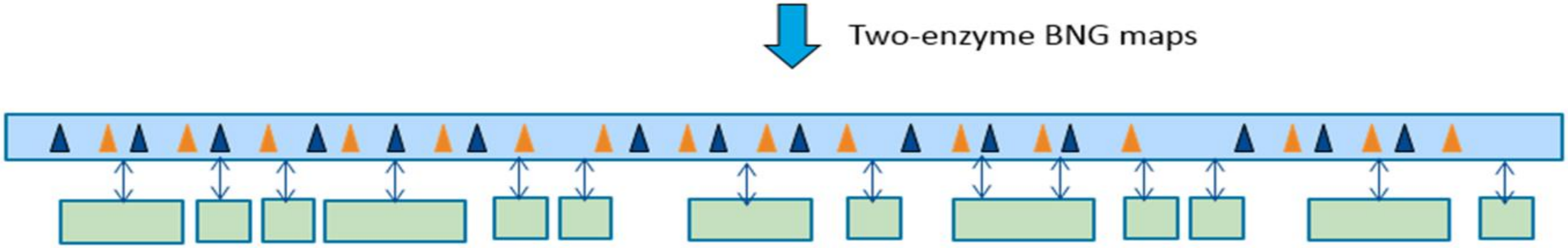
	PacBio Assembly	BioNano BspQ1 Assembly	BspQ1 Hybrid scaffold
Count	12318	2228	1430
Median length (Mb)	0.120	0.999	1.442
N50 length (Mb)	0.524	1.979	2.87
Max length (Mb)	3.35	11.49	17.45
Total length (Mb)	2930	3191	2922
% genome	81%	88%	81%

⇨ Hybrid scaffold + not scaffolded PacBio contigs : 3611Mb

2 enzymes Hybrid scaffolding



infer linkage of BSPQI maps from BSSSI maps (and vice versa) and further merge maps to generate two-enzyme map



Higher label information density on two-enzyme maps allow us to fill in gaps in the final scaffolds by anchoring shorter NGS contigs

2 Step Hybrid Assembly

	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	11.49	17.45	5.322	24.670
Total length (Mb)	2930	3191	2922	3112	2960
% genome	81%	88%	81%	86%	82%

2 Step Hybrid Assembly

	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	11.49	17.45	[illegible]	24.670
Total length (Mb)	2930	3191	2922	3112	2960
% genome	81%	88%	81%	86%	82%

Improvement

2 Step Hybrid Assembly

	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	[illegible]	[illegible]	[illegible]	24.670
Total length (Mb)	2930	3191	2922	3112	2960
% genome	81%	88%	81%	86%	82%

More than 7 fold increase

Improvement of the sunflower assembly

- More than 7 fold improvement of the N50 length
- The 2 step hybrid scaffolding strategy improves significantly the resulting N50

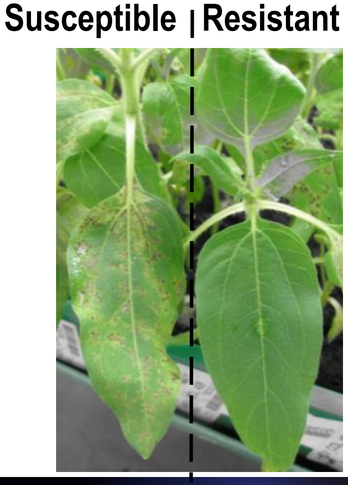
➡ **Characterized a region of interest**

Fine mapping of the QRM1 QTL in Sunflower

S. Munos S. Vautrin

Spring 2013:
7455 F2 segregating for QRM1 only.
901 recombinant plants identified.

Phenotypic analysis
(232 recombinant plants)

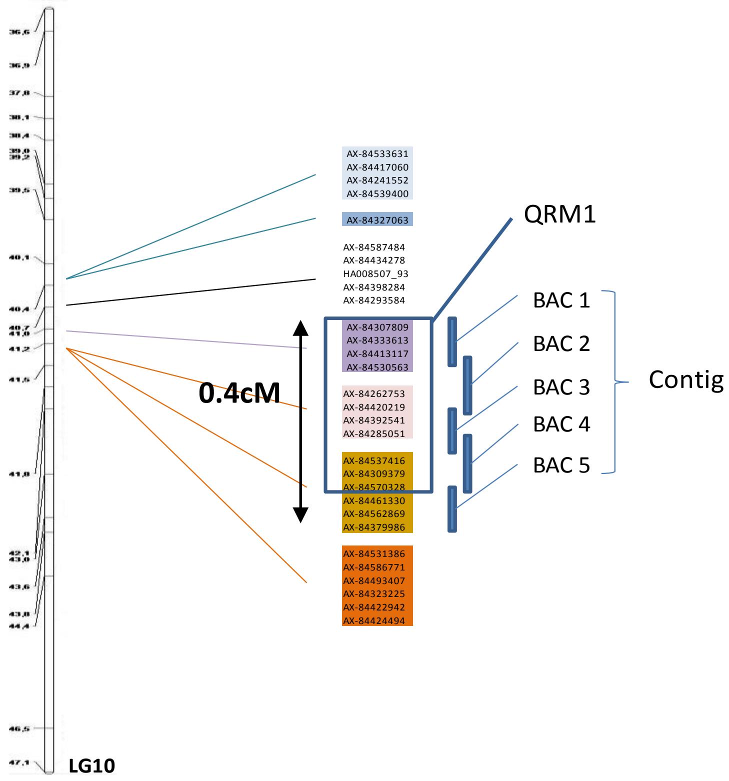


- QRM1 controls quantitative resistance to downy mildew
- Strong effect on LG10
- Explain 65% of the phenotypic variability
- 2 Near Isogenic Lines:
  - Susceptible (PSC8)
  - Resistant (XRQ)
- *In silico* physical mapping

⇒ reduction of the genetic mapping to a 0.4 cM window

Map-based cloning of QRM1

HA412 BAC sequence



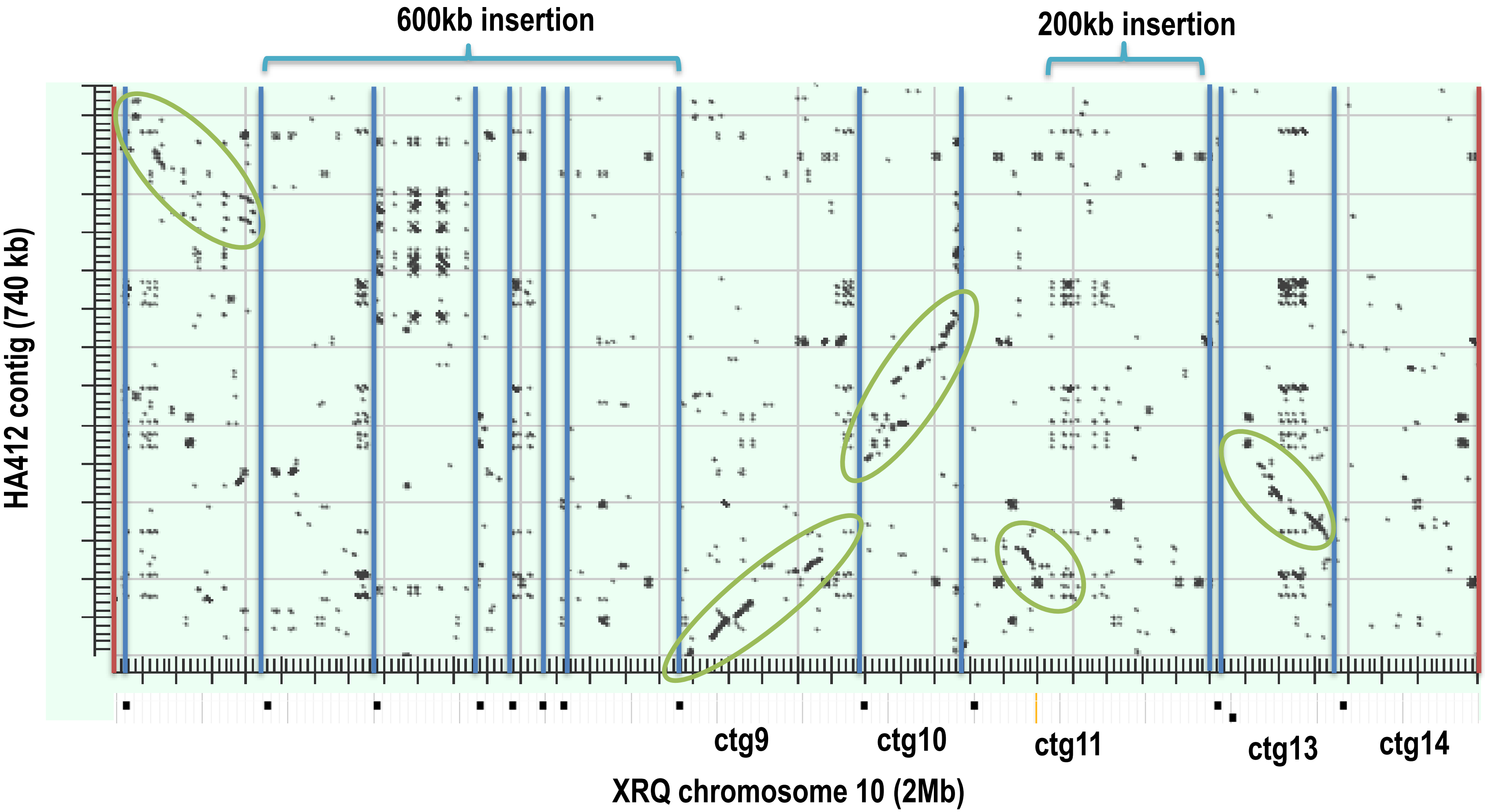
- Sequencing of the 6 HA412 BAC clones by Pacbio
- A contig of 740kb assembled in one unique sequence
- Highly accurate sequence

Map-based cloning of the QRM1 QTL

XRQ WHOLE GENOME PACBIO SEQUENCE

- Resistant genotype XRQ
- Retrieval of a 2Mb sequence on chromosome 10 (based on 20 markers alignment)
- This 2 Mb sequences is composed of 14 scaffolded Pacbio contigs separating by N gaps (10k missing nucleotides)
- Comparison of the HA412 sequence (BAC clones) and the XRQ resistant line (full genome sequence)

Comparison of the XRQ genome vs HA412 BAC clones



Low collinearity

Fragmented alignment / orientation inconstancies :

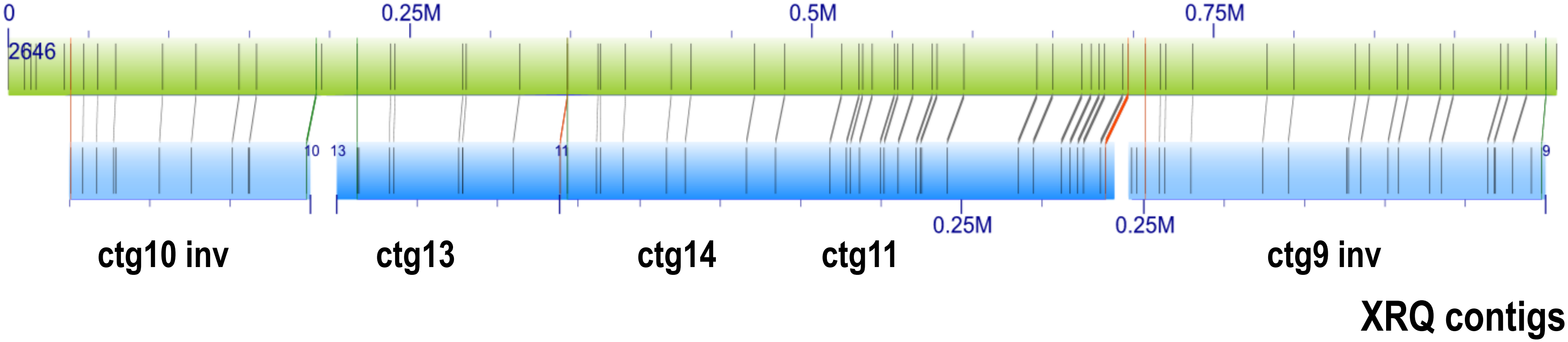
Scaffolding errors OR true variability?

XRQ Optical map data allowed to correct XRQ scaffolding

S. Cauet

Alignment of the contig cmaps against the BioNano assembly of XRQ genome

C-map obtained on a unique DNA map



On this targeted region, Optical Bionano map allowed:

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

Manually curated genome XRQ sequence *vs* BAC clones contig HA412

Genetically anchored QTL limit

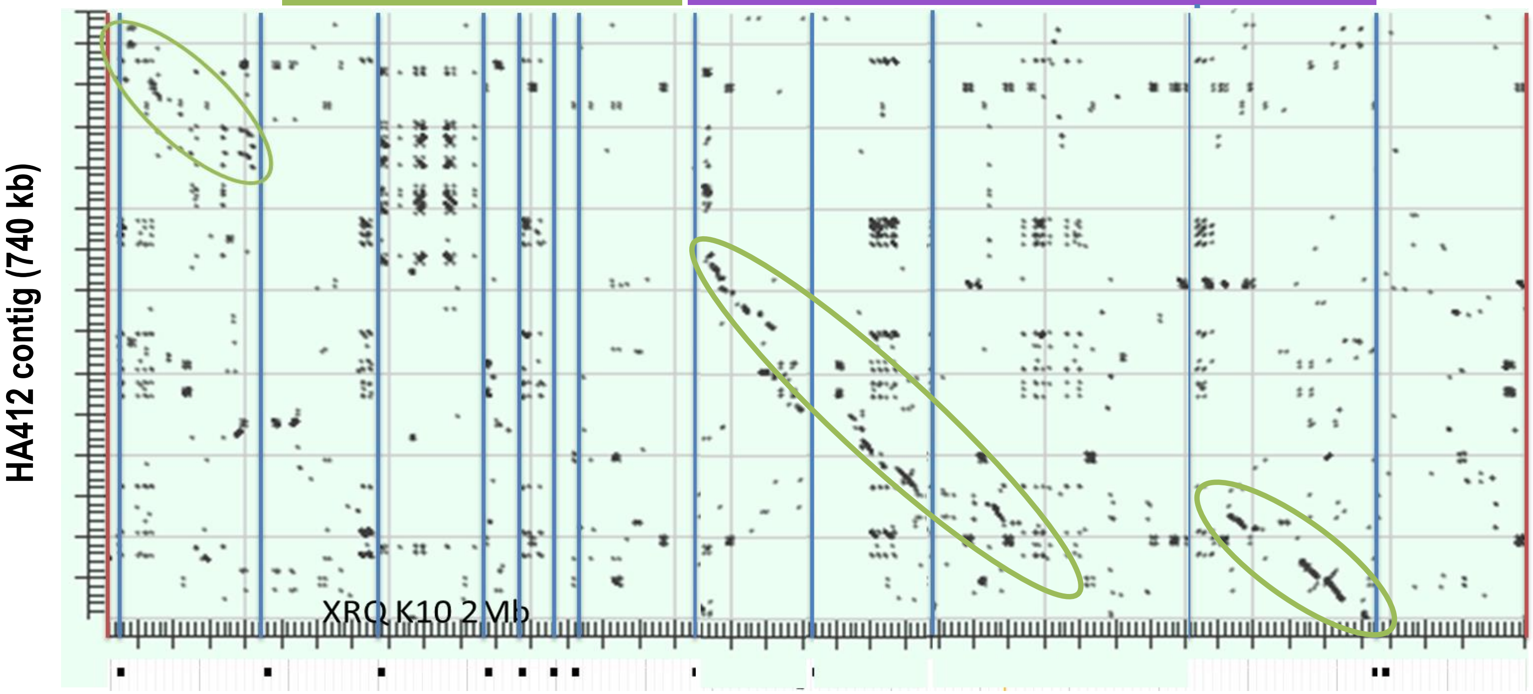
600kb insertion?

Still need to be validated

200kb insertion validated

Contig order validated by optical map data

Genetically anchored QTL limit



The scaffoling of PacBio contigs is more accurate (improved collinearity between the two sequences on QRM1 QTL)

But still high variability the two sunflower lines : 2 major insertions of several hundreds of kb in XRQ

Summary for the QRM1 QTL

- The mapping of the QRM1 QTL controlling downy mildew resistance in sunflower has been restricted to a 2Mb sequence
- The optical map allowed to validate major rearrangements between the 2 sunflower genotypes
- Annotation of the 2 sequences and comparative analysis are under progress but 9 candidates genes have been identified

At the full genome scale

- The optical map (2 enzymes, >150 X) will improve the sunflower genome sequence (orientation of PacBio contigs and scaffolding)
- More than 7 fold improvement of the N50 length
- The 2 step hybrid scaffolding strategy improves significantly the resulting N50
- We hope to obtain more improvement with the use of the new 1 step 2 enzymes hybrid scaffolding tool from BioNano
- Now we must look to the conflicts between the NGS assembly and the BioNano assembly more in details

Acknowledgements

@CNRGV
@SUNRISE_France

http://cnrgv.toulouse.inra.fr/

INRA SCIENCE & IMPACT

CNRGV PLANT GENOMIC CENTER

Hélène BERGES
Nadège ARNAL
Arnaud BELLEC
Genséric BEYDON
Caroline CALLOT
Stéphane CAUET
Céline CHANTRY-DARMON
Joëlle FOURMENT
Nadine GAUTIER
Laetitia HOARAU
Céline JEZIORSKI
William MARANDE
Elisa PRAT
David PUJOL
Nathalie RODDE
Roseana RODRIGUES
Sonia VAUTRIN

LIPM Laboratoire Interactions Plantes Micro-organismes

Marie-Claude BONIFACE
Félicity VEAR
Nicolas POUILLY
Patrcik VINCOURT
Jérôme GOUZY
Nicolas LANGLADE
Stéphane MUÑOS

SUNRISE UNE CULTURE POUR LE FUTUR

BIONANO GENOMICS

John BAETEN
Kees-Jan FRANCOIJS

UNION EUROPÉENNE

LA RÉGION OCCITANIE Pyrénées-Méditerranée