Toward a better understanding of plant genomes structure

INRA SCIENCE & IMPACT

CNRGV

Hélène BERGES

Director of the Plant Genomic Center

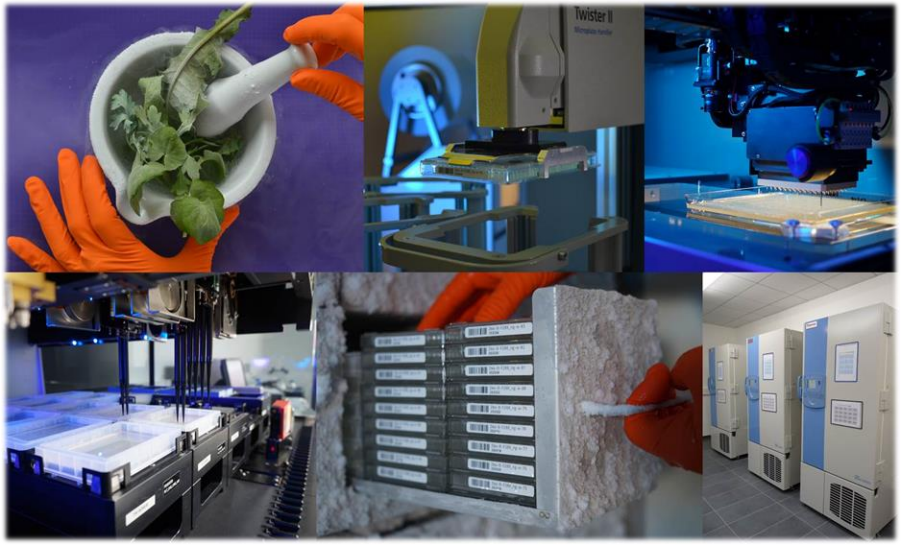
The French Plant Genomic Center

Created in 2004 by INRA (French National Institute for Agricultural Research)

- **Depository of genomic libraries for the scientific community**

⇨ BAC libraries

ISO 9001:2008
Octobre 2005



- **A dedicated structure to assist plant genomic programs**

⇨ Distribute the genomic resources at the international level

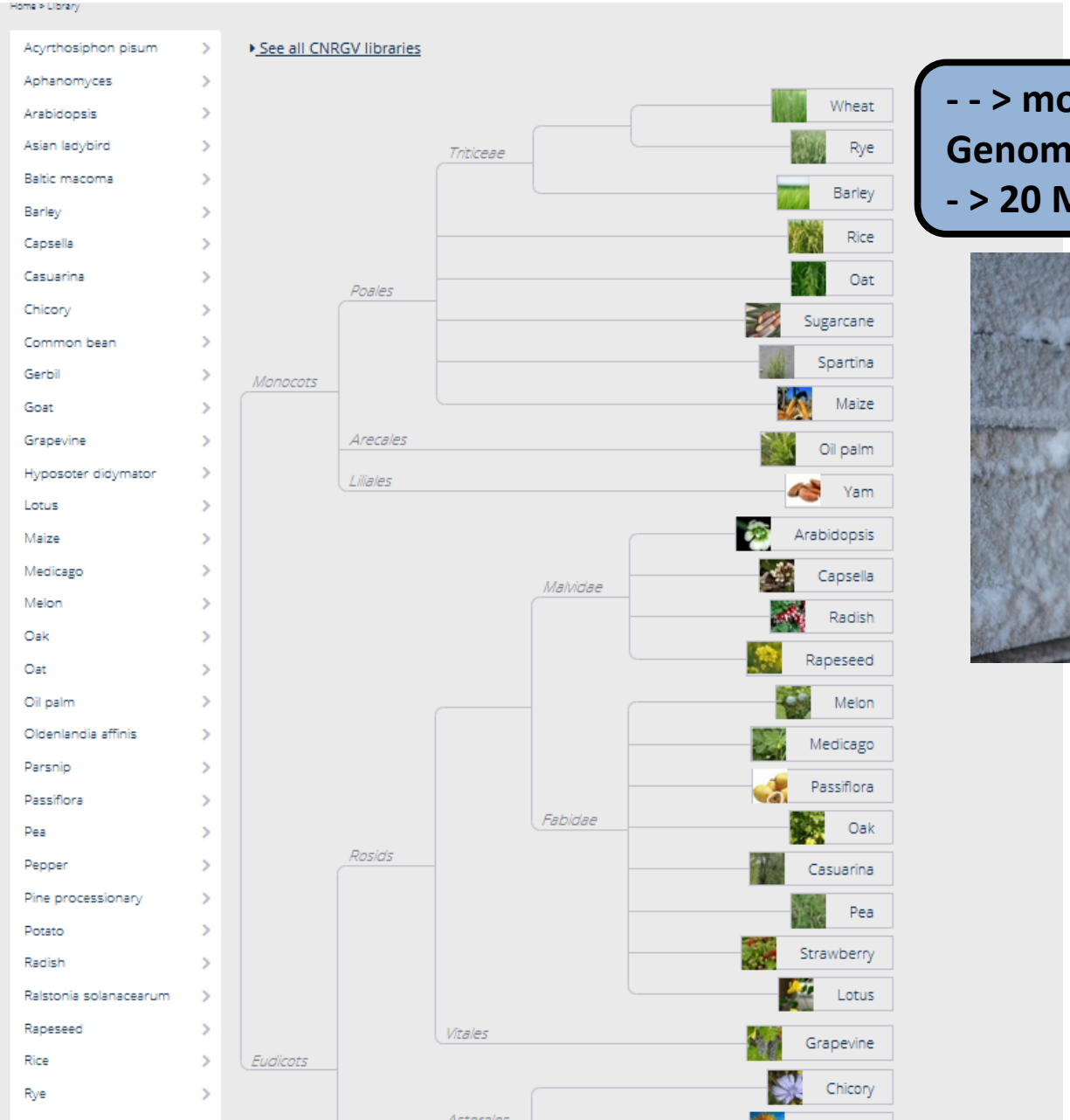
⇨ Provide high quality research material and efficient tools and services for studying plant genomes

⇨ Develop innovative solutions

⇨ Develop genomic projects in collaboration

⇨ Host scientists in the frame of collaborations

Genomic Libraries available at CNRGV



- - > more than 430 different Genomic Libraries
- > 20 M unique clones

Services at CNRGV

INRA SCIENCE & IMPACT CNRGV PLANT GENOMIC CENTER

YOUR ACCOUNT FR EN

Communication Library Services Projects Team Equipment Quality Contact

Home > Services

Services

BAC LIBRARY CONSTRUCTION | GENOMIC SAMPLES DISTRIBUTION | SCREENING SERVICES | TARGETED BAC LIBRARY CONSTRUCTION | BAC CLONE CHARACTERIZATION | ROBOTIC SC

BAC library construction

Plant BAC library construction

Ask Services

Last update: 04 August 2015

Plant BAC library construction

Libraries of large genomic DNA inserts are essential to genomics research. The cloning of large genomic DNA fragments into BAC vectors facilitates handling and multiplication as well as long-term conservation. BAC libraries help to identify and isolate genes of interest but also to carry out the physical mapping and sequencing of plant genomes.

We perform whole steps required to produce BAC libraries, from plant tissue to clones ordered in microtiter plates:

- Cell nuclei extraction from frozen material of your plant of interest
- HMW DNA partial digestion
- Sizing of DNA fragments
- Ligation in a vector
- Cloning in phage resistant bactery
- Colony picking using high throughput automated station
- Quality tests and characterization of the library

All steps are tracked via CNRGV's information management system: Genolims.

We propose different strategies according to the level of genomic data available. Thus if there are no studies on your plant of interest, the BAC library will constitute the gateway to the genes and sequence of interest.

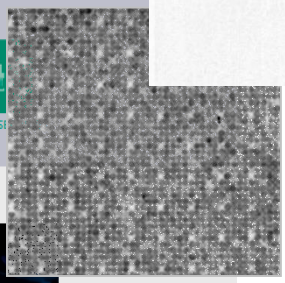
Linked to a genetic map, a BAC library will constitute a reliable basis to establish a physical map and eventually sequence the genome.

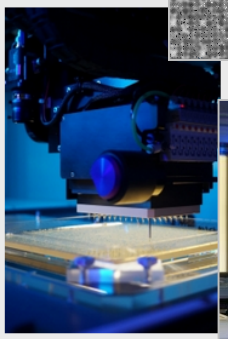
If the sequence of the plant genome of your interest or close relative exists, by constructing a BAC library it will be very easy to reveal the genetic diversity of ecotypes or related species of interest.

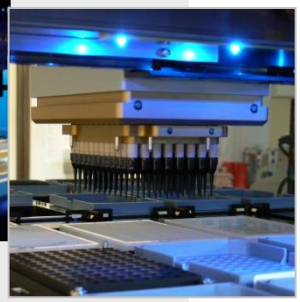
The number of clones of the BAC library and consequently its coverage will depend on the sequence availability and the goal of your project.

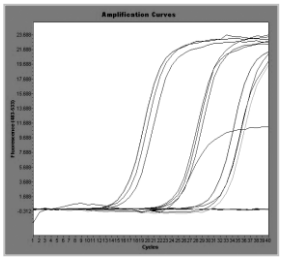
BAC library construction can be associated with the production of 3D-pools or macroarray, their screening and almost any services provided at CNRGV.

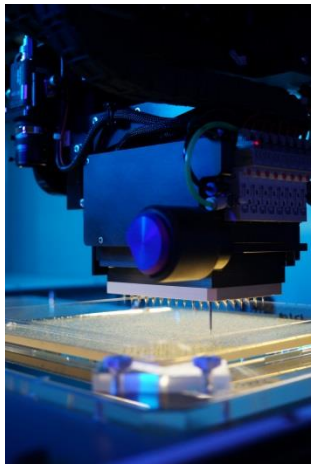
Various projects on model or crop plants are already under process.







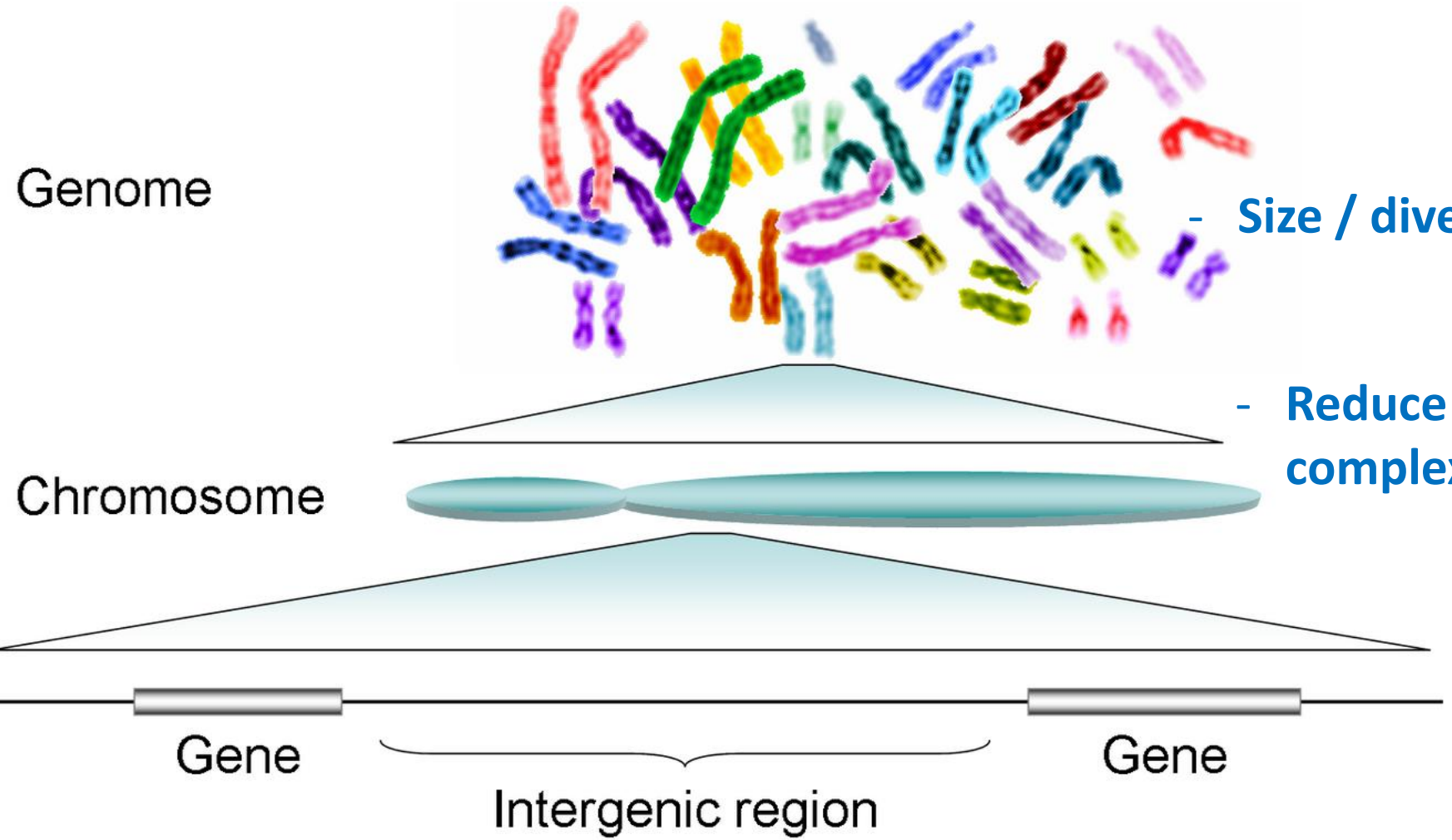




Interactions with laboratories around the world

- **3 233 242 Clones distributed during the last 5 years (2011-2015)**

The challenges - The expectations

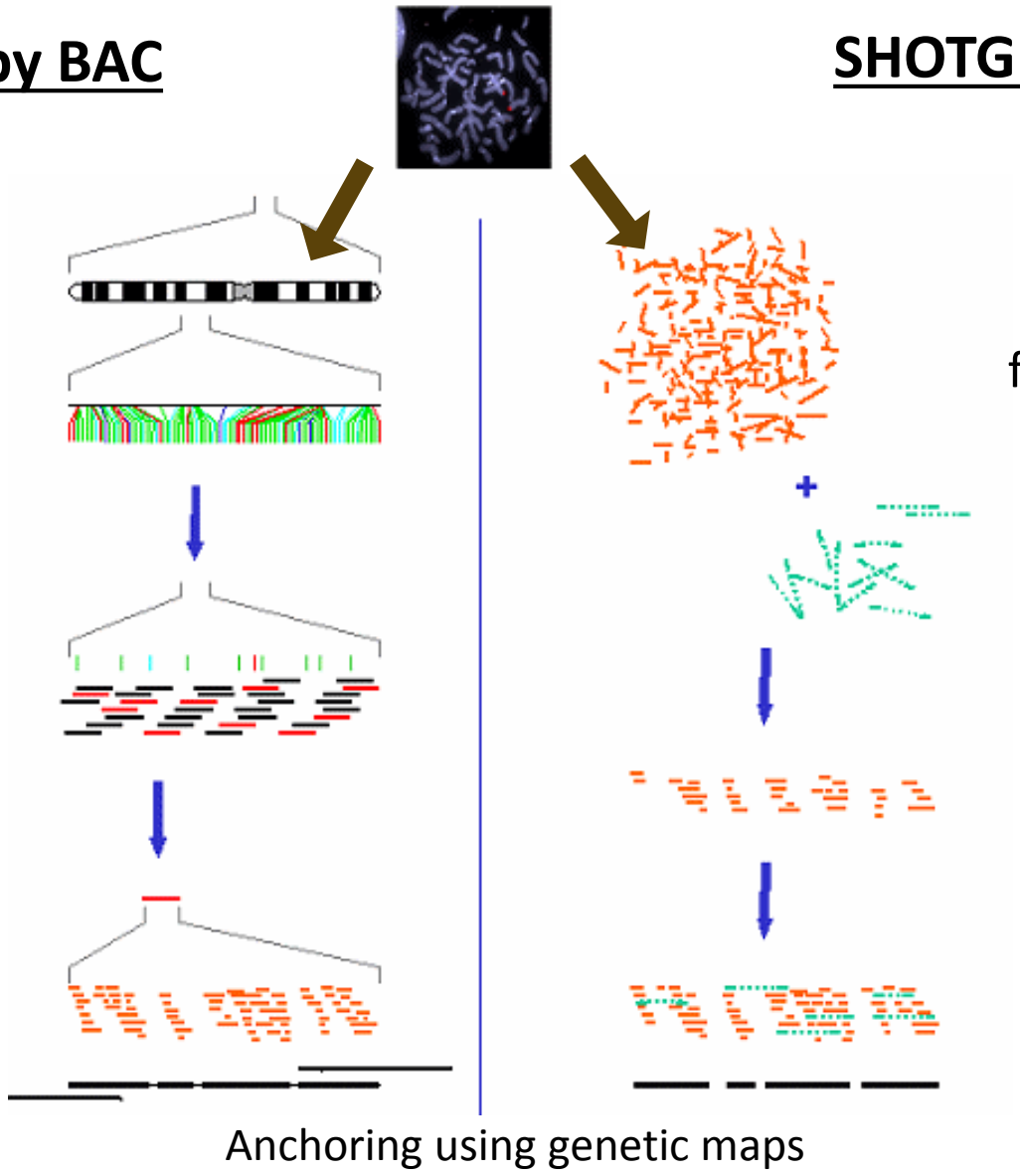


- Size / diversity
- Reduce the complexity

- Target specific Genomic regions
- Manage diversity

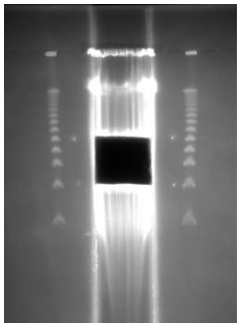
- Focus / Target / Markers

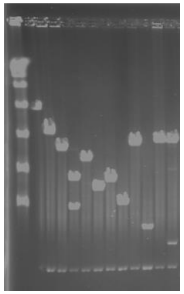
Genome sequencing strategies



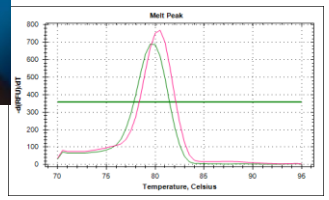
The BAC library strategy

- Specific Markers of genomic regions





BAC library from various genotypes



Screening
Identification of BAC clones

```
TATTTACCATATCAGATTCACATTCAGTCCTCAGCAAAATGAAGGGCTCCATTTTCACTCTGTTTTTATT
CTCTGTCCTATTTGCCATCTCAGAAGTGCGGAGCAAGGAGTCTGTGAGACTCTGTGGGCTAGAATACATA
CGGACAGTCATCTATATCTGTGCTAGCTCCAGGTGGAGAAGGCATCAGGAGGGGATCCCTCAAGCTCAGC
AAGCTGAGACAGGAAACTCCTTCCAGCTCCCACATAAACGTGAGTTTTCTGAGGAAAATCCAGCGCAAAA
CCTTC                                CTTTGGGGTGGACAGATGCCCACTGAAGAGCTTTGG
AAGTC                                ATTTACAAACTTTGTGTTGCACTGATGGCTGTTCCA
TGACT                                CAAATACCCAATGGGTGGCAGAGCTTTATCACATGT
TTAAT                                AATATTGTGTTATTAAAATGATGGCTTTTGGGTAGG
CAAAA                                TTGAAACCACAGTGATCTCTATTTTCTCCCTTTGCC
AAGGT                                ATGCTTTGAAATTTCAAATGCTGCGCAAAATTGCAA
TAAAA
```

- BAC-Pool Sequencing (PacBio)
- 35 to 100 x coverage
- PacBio Technology - 1BAC : 1 contig
- MTP of BACs : 1 contig

➢ **Essential and efficient tools for understanding the organization and function of specific genomic regions**

Targeting a genomic region of interest

Partners :
INTA Argentina, Maria Fernanda Pergolesi and Maria Jose Dieguez

Gamma 6 (Sensitive) | Sinvalocho (Résistant)

Objectives:

Physical map of the locus confering resistance to *Puccinia triticina in the* resistant wheat cultivar Sinvalocho (The pathogen Puccinia is a rust fungus)

Data : LrSV2 gene for adult plant wheat leaf rust resistance was mapped on chromosome 3BS

- *LrSV2* target interval : 262 Kb in CS
- CS sequence available
- PCR markers genome specific available

1. **Construction of the NG-BAC library from Sinvalocho**
   440 samples representing 637 440 clones **(3.39 X)**
   Instead of 1650 plates

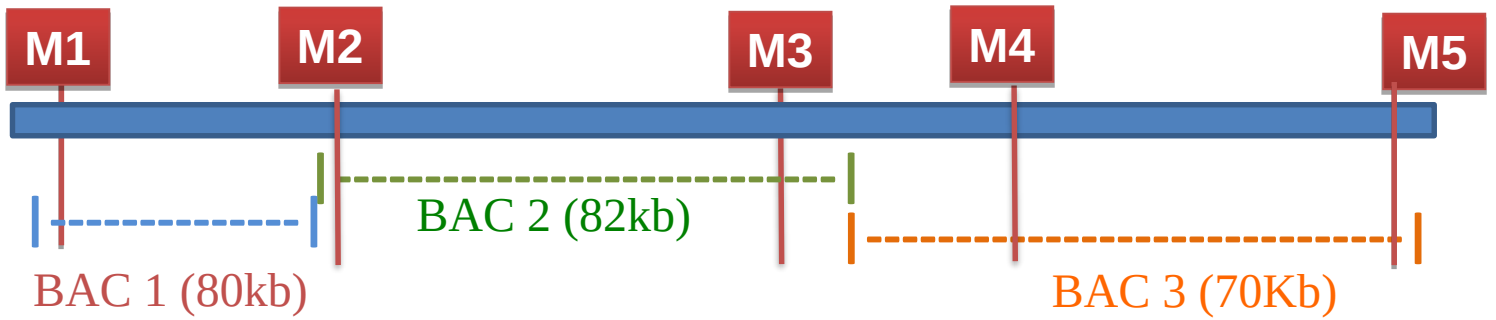
2. **Screening with 5 markers**

3. **Sequencing of positive BACs clones**

Map based cloning of LrSV2 in wheat

3 positive BAC clones identified spanning the *LrSV2* interval

-> characterization, validation



-> BAC sequencing using 454 Technology – Definition of news markers

-> reduction of the interval to 71kb

-> Size of *Sinvalocho* locus smaller than in CS

Identification of candidate genes (annotation)

Functional validation (transformation of a susceptible variety to confirm the involvement of the gene in the resistance) in progress

Time for the project : 3 months

Comparison of short and long reads Sequencing

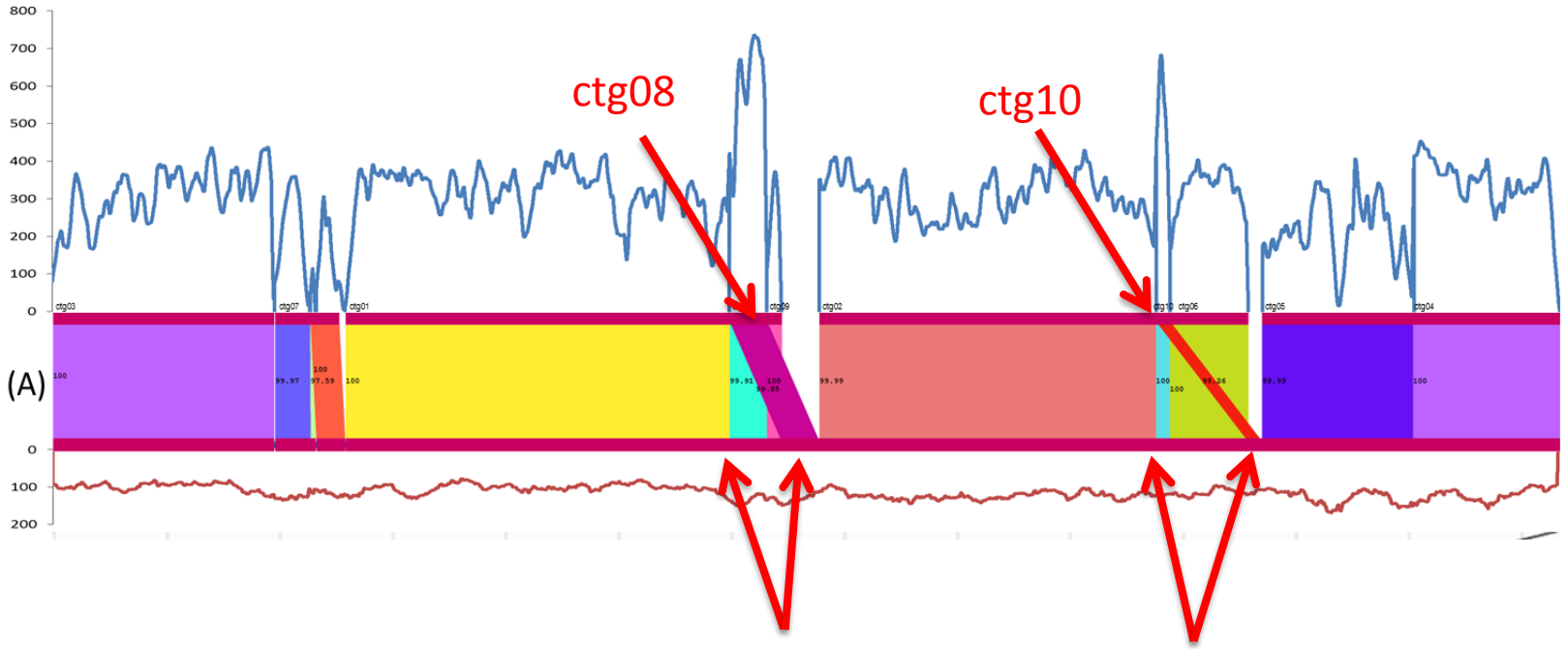
Clone name	Estimated insert size (kb)	Roche-454 contigs [a]	PacBio RS II contigs	Roche-454 size (bp)	PacBio RS II size (bp)	Roche-454/PacBio RS II size ratio
Frag-55O19	90	2	1	90248	90557	0.99659
Heli-337E08	170	6	1	173471	175563	0.98808
Hord-155N13	155	10	1	155553	155889	0.99784
Sacc-241H10	150	17	1	142570	152547	0.93460
Sacc-276O20	100	1	1	105854	105851	1.00003
Trit-136P19	120	2	1	124789	125714	0.99264
Trit-131J6	130	17	1	125436	133334	0.94077
Zeam-34K24	135	10	1	128036	133221	0,96108
Zeam-100L1	85	1	1	86647	86651	0.99995

PacBio RS II reads assembly performed following HGAP workflow.
Newbler assembly for Roche-454 reads (filtering low quality, *E.Coli*, and vectors reads).

- **Assembly of PacBio RS II sequences of pool of untagged BAC clones led to one contig per BAC assigned with BAC-end sequences**

Interest of the long reads for complex genomic region

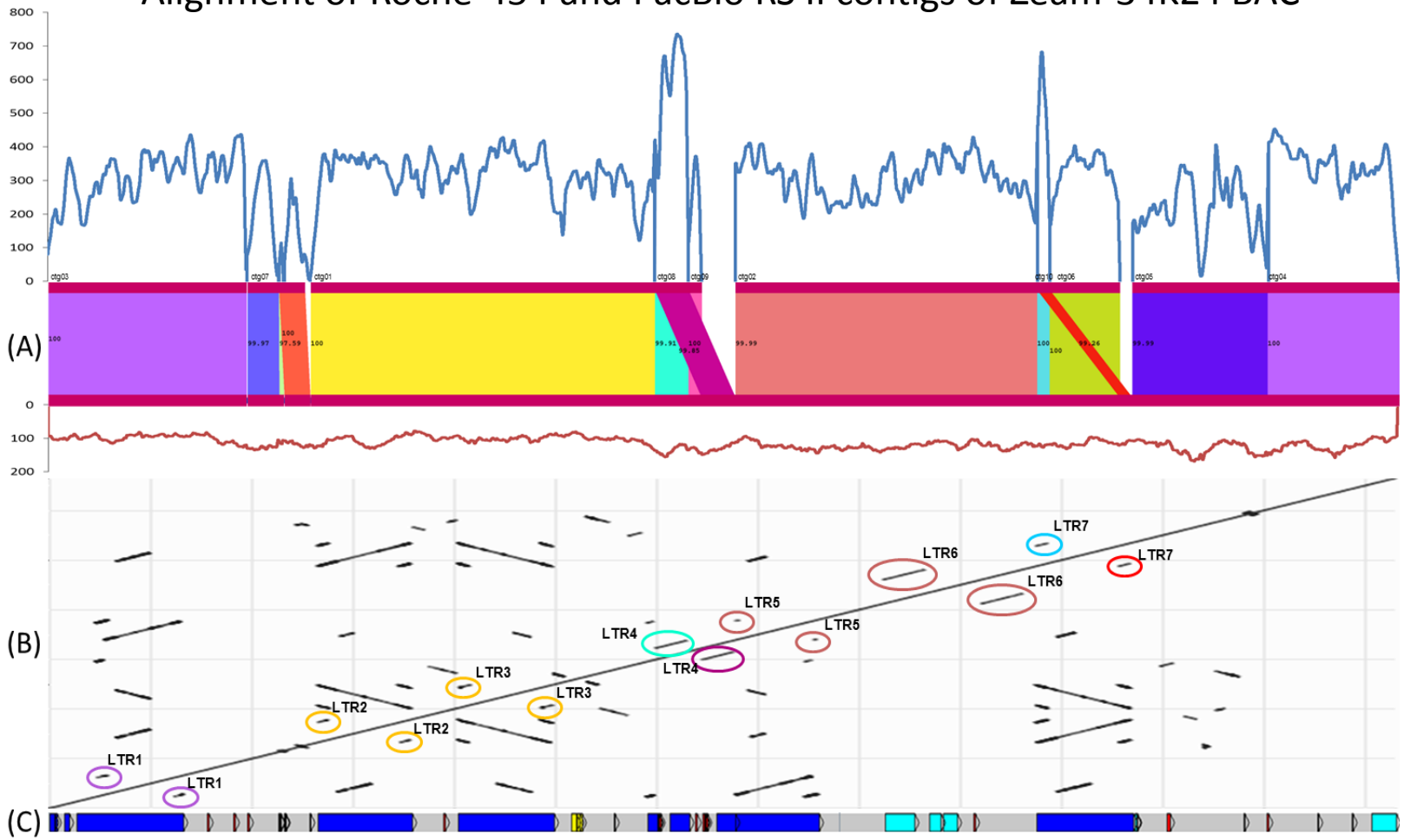
Alignment of Roche-454 and PacBio RS II contigs of Zeam-34K24 BAC



- Two contigs (454) displaying strong homologies with two distinct regions (PacBio)
- Roche-454 reads coverage exhibited two spikes / PacBio RS II reads coverage is stable corresponding to missambled data

Interest of the long reads for complex genomic region

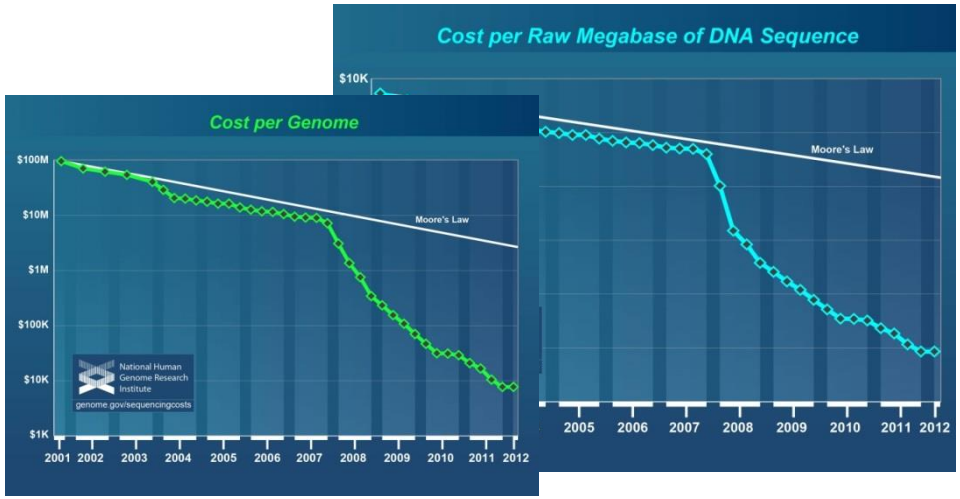
Alignment of Roche-454 and PacBio RS II contigs of Zeam-34K24 BAC

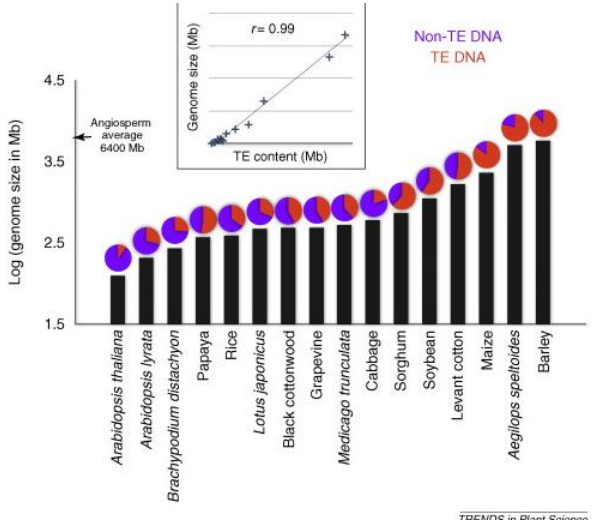


- Collapsing region corresponding to strong similar repeated element (LTR – copia superfamily)

The Next-generation DNA sequencing technologies

Complex plant genomes sequencing projects became possible (many billions of bases per day for hundreds or thousands of dollars per gigabase instead of millions or billions of dollars per gigabase)





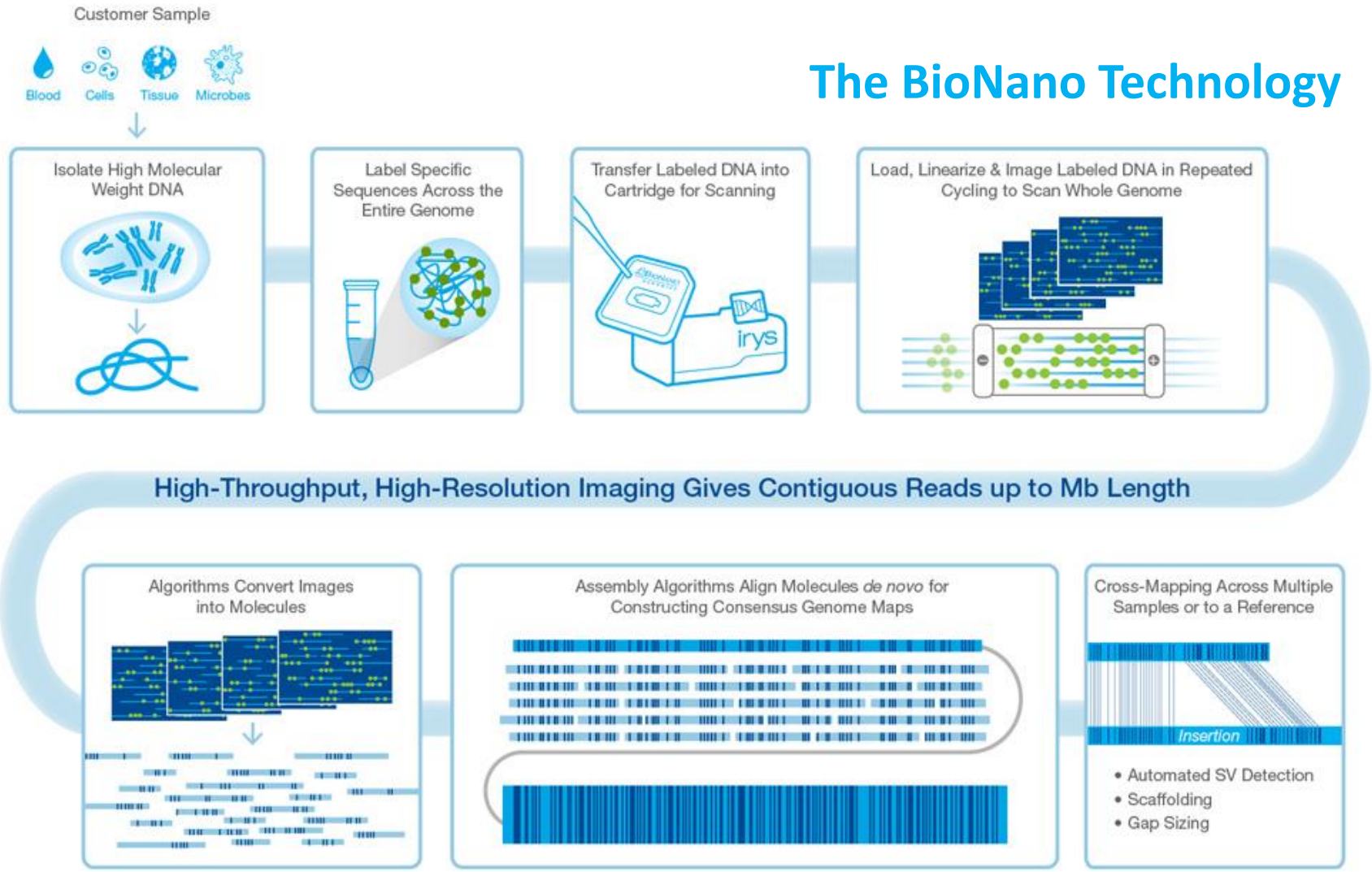
but de novo assembling of plant genomes remains challenging

- **Size of the reads**
- **Gene families are difficult to assemble and may collapse into a mosaic**
- **Repeat elements**
- **How to assemble these genomes accurately?**

-> Despite the progress made with the NGS technologies we still don't have enough reference plant genomes with high confident data (false conclusions ?)

-> Third-generation sequencing technologies ?

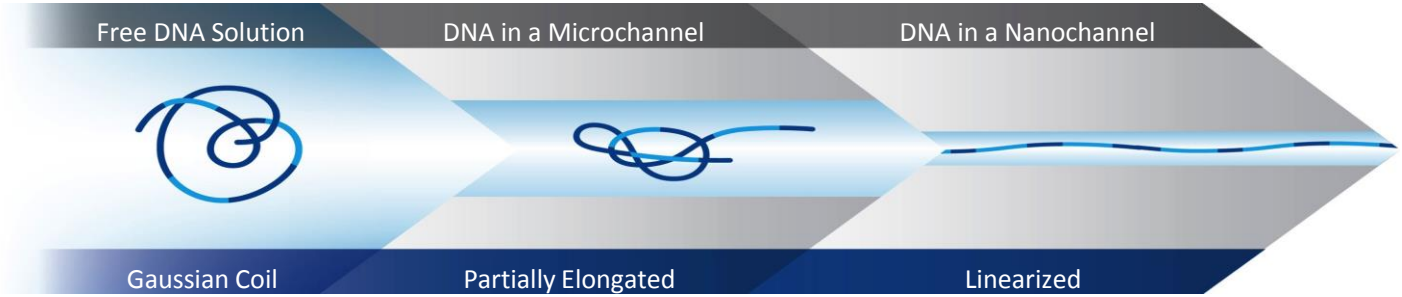
Looking for structural variations by physical mapping



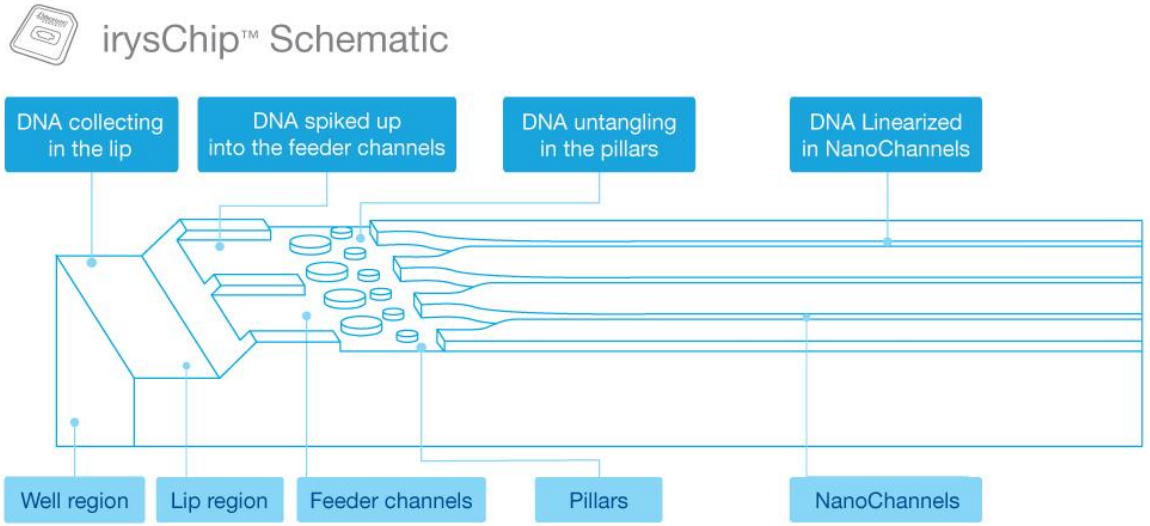
50Gb data generated per flowcell (=> 100Gb per chip)

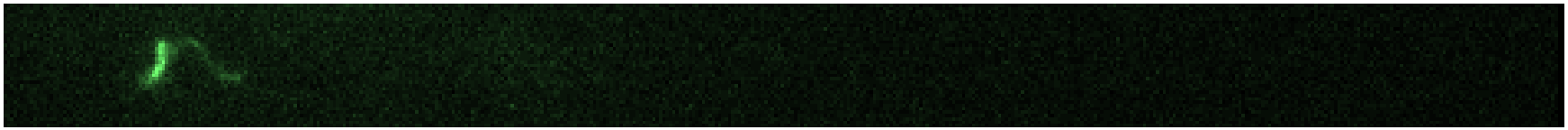
Looking for structural variations by physical mapping

The BioNano Technology



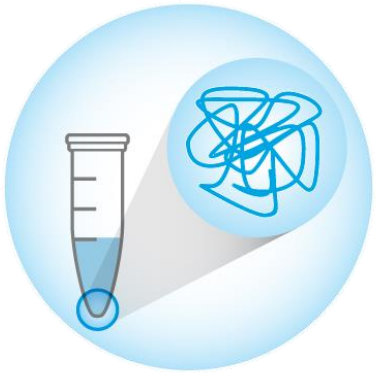
DNA molecules in the nanochanels

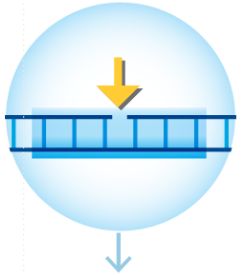




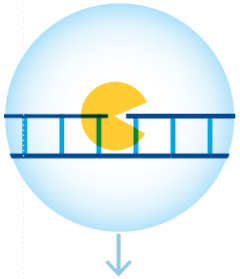
Looking for structural variations by physical mapping

Purified Long DNA → Nick & Label → Labeled DNA



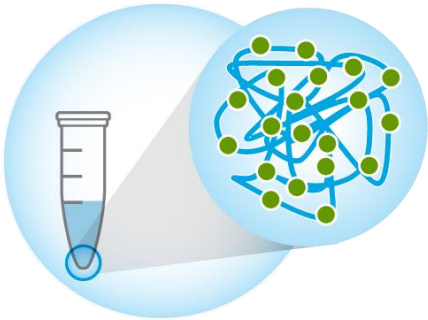


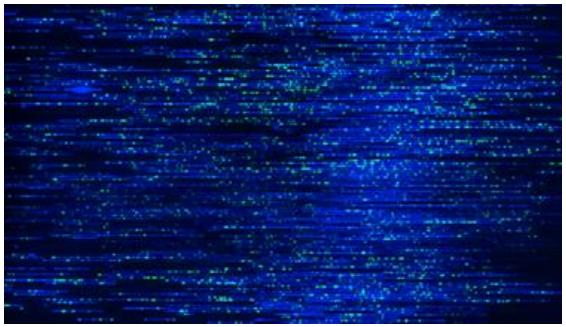
A nicking endonuclease creates single-strand nicks at recognition sites



Polymerase initiates strand displacement and polymerization

Fluorescent nucleotides are incorporated t<o label the enzyme recognition sites





Looking for structural variations by physical mapping

BioNano system

Workflow

Start with non-amplified native genomic DNA

Label seq. specific sites (e.g. nickase motifs)

Linearize & Image

Convert images to digitized molecules:

- **Convert label locations to distances between labels**
- **Create molecular barcodes (100kb to >1 Mb)**

Assemble the molecular barcodes into consensus maps/contigs:

- **Map lengths can be as long as 30 Mb**

Applications

For Genome Finishing, the maps serve as a scaffold:

- **Sequencing contigs are converted in silico into molecular barcodes by highlighting the same sequence motifs**
- **These sequencing based barcodes are then aligned to the BioNano maps**

For SV discovery/detection, compare to a reference or gold standard, looking for changes in the patterns:

- **Shifts in barcode patterns reveal insertion (addition), deletion (subtraction), inversion (re-orientation, translocation of genome segments**

Sunflower Genome Finishing

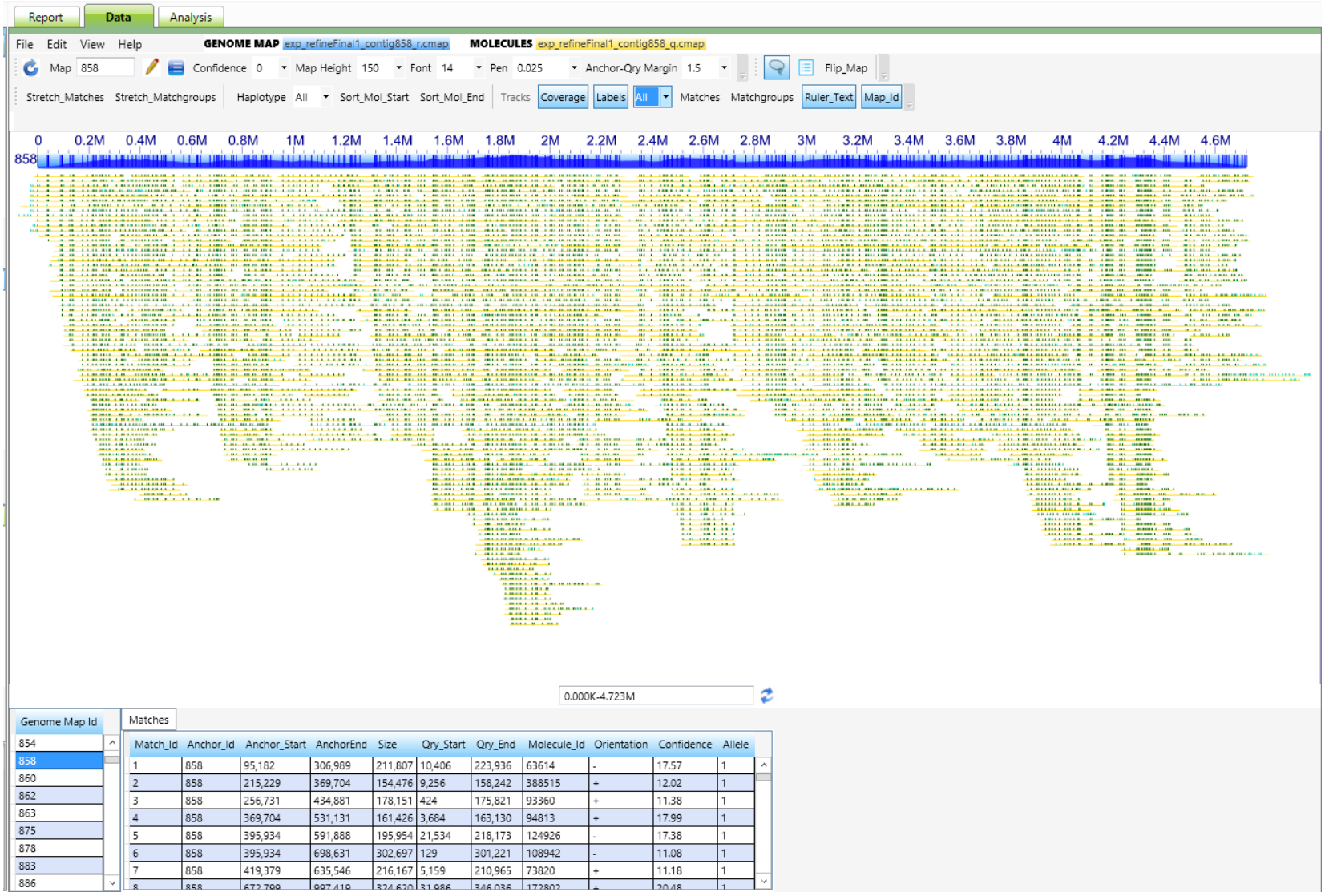
- Species: *Helianthus annuus*
- 3,6 Gb
- 2n=34 chromosomes
- Genome sequence + 100X PacBio

# 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

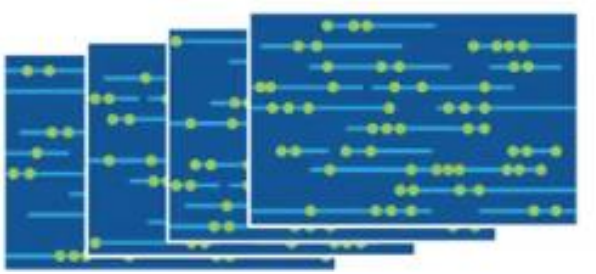
Langlade et al., PAG Jan 2016

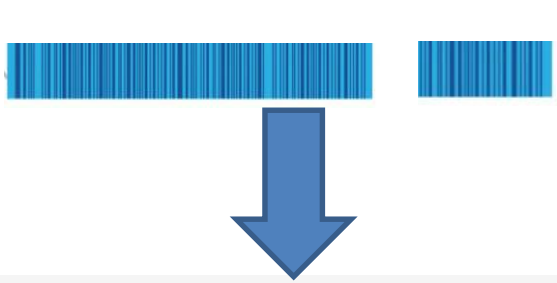
Optical map of the sunflower genome



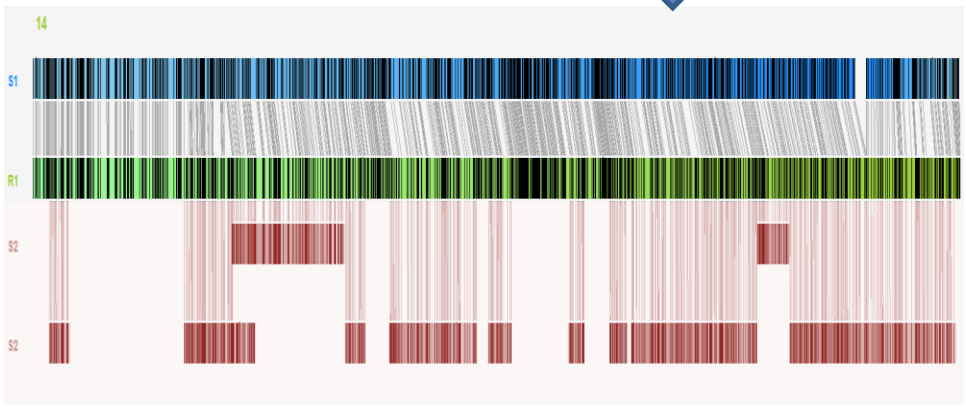
Match_Id	Anchor_Id	Anchor_Start	AnchorEnd	Size	Qry_Start	Qry_End	Molecule_Id	Orientation	Confidence	Allele
1	858	95,182	306,989	211,807	10,406	223,936	63614	-	17.57	1
2	858	215,229	369,704	154,476	9,256	158,242	388515	+	12.02	1
3	858	256,731	434,881	178,151	424	175,821	93360	+	11.38	1
4	858	369,704	531,131	161,426	3,684	163,130	94813	+	17.99	1
5	858	395,934	591,888	195,954	21,534	218,173	124926	-	17.38	1
6	858	395,934	698,631	302,697	129	301,221	108942	-	11.08	1
7	858	419,379	635,546	216,167	5,159	210,965	73820	+	11.18	1
8	858	672,799	997,419	324,620	31,986	346,036	172802	+	20.48	1

Hybrid scaffolding of the sunflower genome

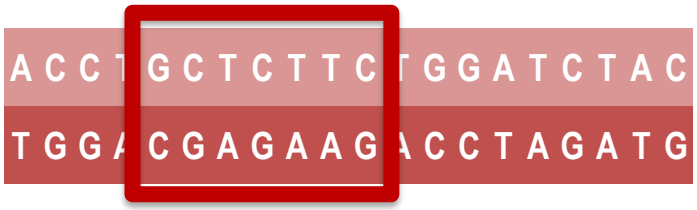


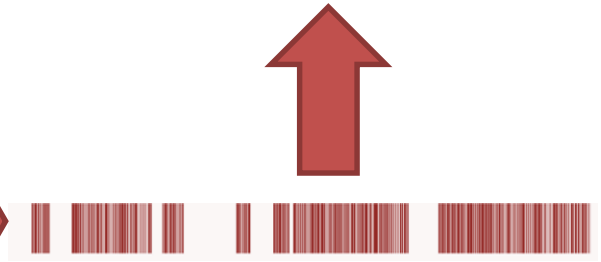


Nickase event
Fingerprint



Hybrid Scaffold





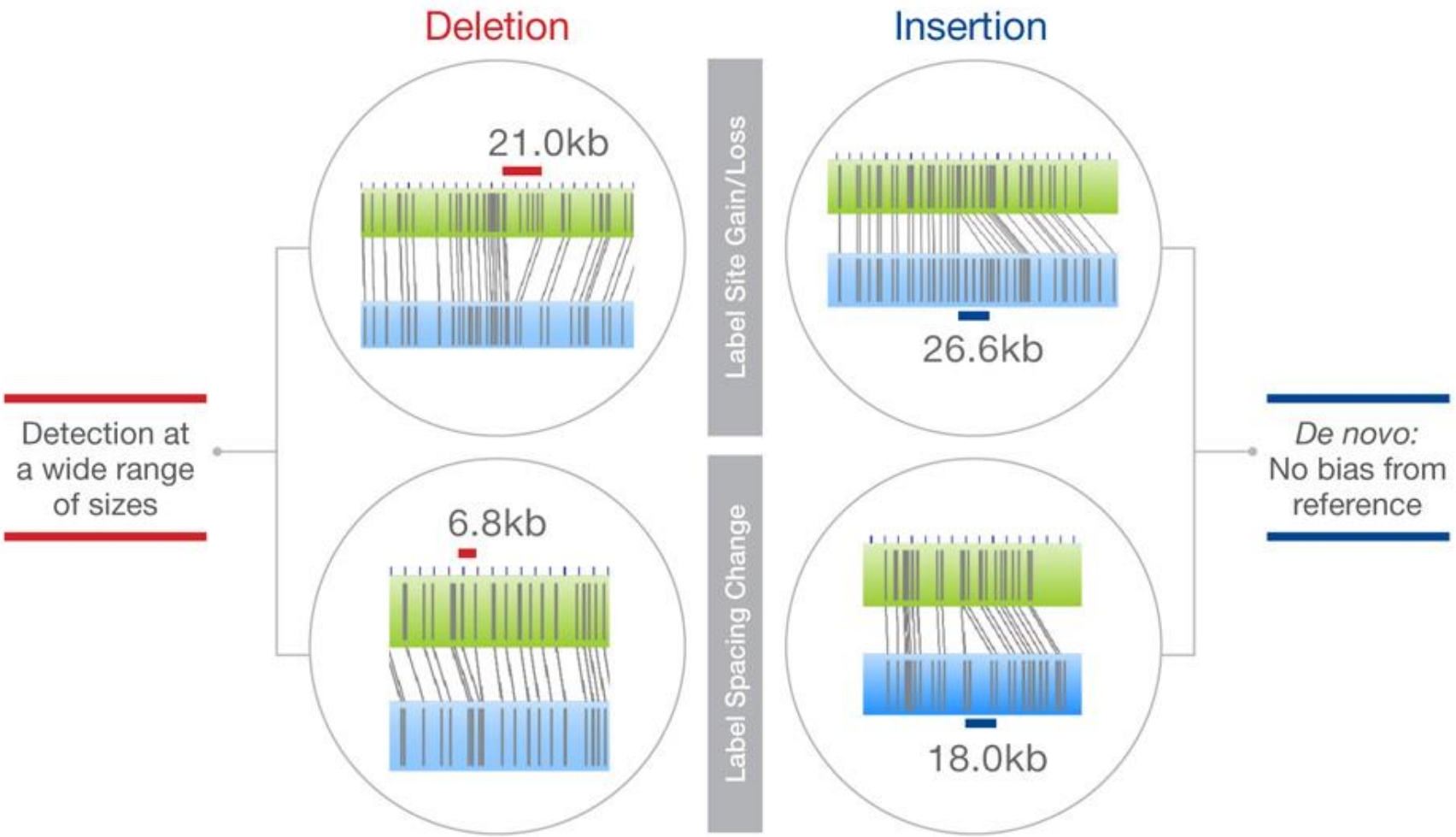
Nickase recognized sites on NGS scaffolds sequences

Delledonne et al., 2016

Sunflower Genome enhancement

	PacBio Map	BioNano Map	Hybrid scaffold
Count	12318	2959	1844
Median length (Mb)	0,120	0,621	1,49
N50 length (Mb)	0,524	1,444	2,227
Max length (Mb)	3,35	4,72	10,09
Total length (Mb)	2930	3202	3303
% genome	81%	89%	91%

Detection of structural variations



Conclusion

- Single molecule long reads technology resolves gaps and collapsing issues of shorter reads sequences assembly

Missing regions /Collapsing of duplicated regions

- Interest of long reads technology in the assembly of genome sequences and consequently in the accuracy of the data generated (*especially with complex genome such as plant genomes with TEs*)

- Combination of BACs and long read technology to solve some issues due to duplicated genes, high level of repetitive elements or polyploidie

- Physical maps to investigate structural variations

Aknowledgements

INRA SCIENCE & IMPACT

CNRGV PLANT GENOMIC CENTER

@CNRGV

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

Arnaud BELLEC
Sonia VAUTRIN
Gensèric BEYDON
Nathalie RODDE
William MARANDE
Joëlle FOURMENT
Elisa PRAT
Nadine GAUTIER
Nadège ARNAL
Audrey COURTIAL
Céline CHANTRY-DARMON
Céline JEZIORSKI
Stéphane CAUET
David PUJOL
Laetitia HOARAU
Hélène BERGES

LE MOULON Génétique végétale UMR

Clémentine Vitte

Lipm Laboratoire Interactions Plantes Micro-organismes

Nicolas Langlade
Stéphane Munos
Jérôme Gouzy

CSIRO

Wolfgang Spielemeyer

International Wheat Genome Sequencing Consortium

Kellye Eversole

INRA SCIENCE & IMPACT

Béatrice Denoyes

INTERNATIONAL BARLEY SEQUENCING CONSORTIUM

Nils Stein

Anete Pereira de Souza /
Danilo Augusto

UNICAMP