de novo sequencing of the sunflower genome

Stéphane Muños

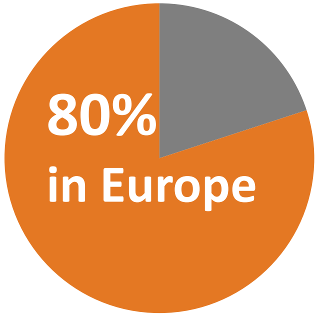
LIPM – INRA Toulouse

stephane.munos@toulouse.inra.fr

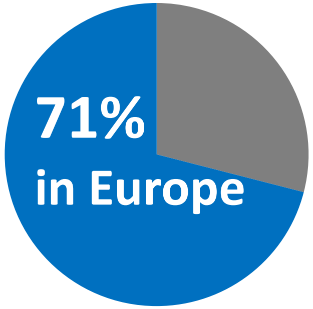
@stephane_munos
@SUNRISE_France

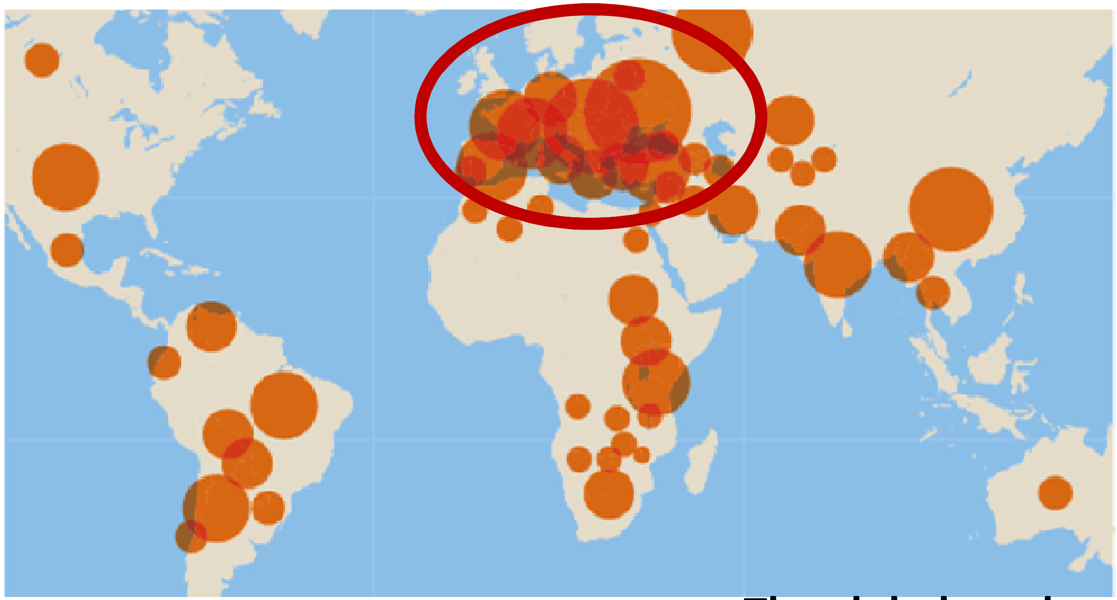
Sunflower, an important crop for Europe

39 Million tons of seed produced worldwide



30 Million hectares worldwilde



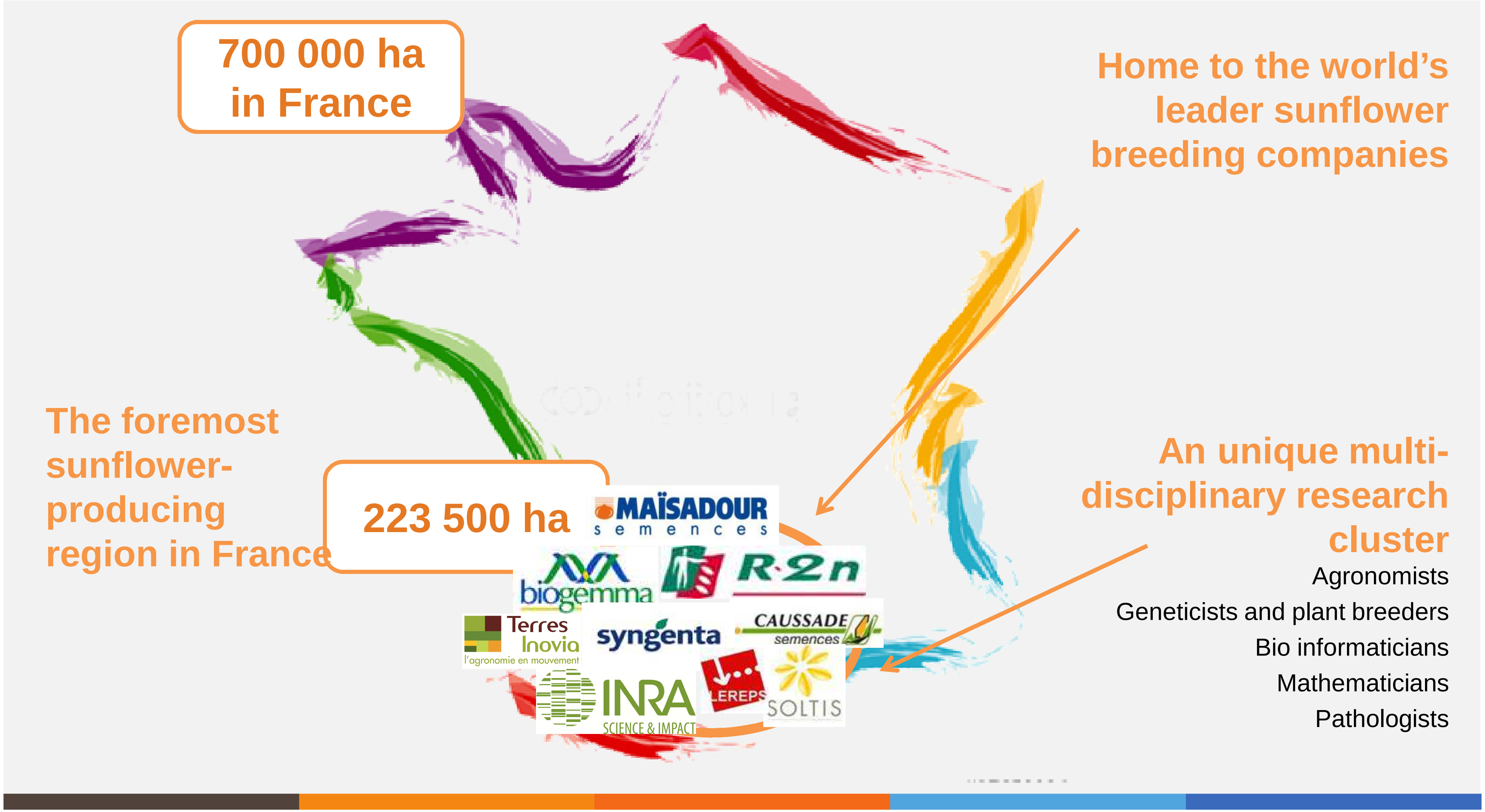


Societal challenge

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

SUNRISE
UNE CULTURE POUR LE FUTUR

The french region Midi-pyrénées, a key player in global sunflower production



700 000 ha in France

Home to the world’s leader sunflower breeding companies

The foremost sunflower-producing region in France

223 500 ha

An unique multi-disciplinary research cluster

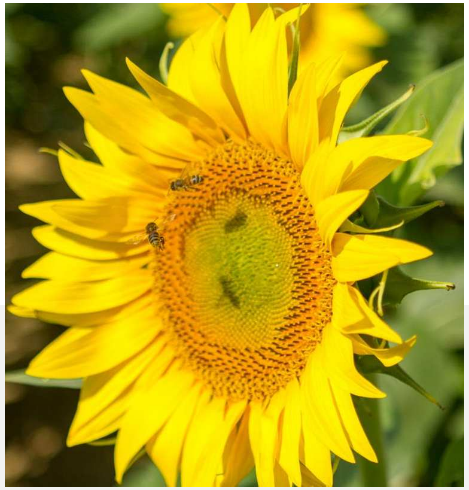
Agronomists

Geneticists and plant breeders

Bio informaticians

Mathematicians

Pathologists



Sequencing of the sunflower genome

Jérôme Gouzy, Baptiste Mayjonade, Christopher J. Grassa, *Sébastien Carrère, Erika Sallet, Ludovic Legrand, Hélène Badouin, Nicolas Pouilly, Marie-Claude Boniface, Nicolas Blanchet, Brigitte Mangin, Cécile Donnadieu, Hélène Bergès,* ***Stéphane Muños****,* ***Patrick Vincourt, Nicolas Langlade***

SUNRISE
UNE CULTURE POUR LE FUTUR

INRA Toulouse

P. Vincourt

N. Langlade

Christopher J. Grassa, *Navdeep Gill, Thuy Nguyen, Nolan Kane,* ***Loren H. Rieseberg***

UBC Vancouver

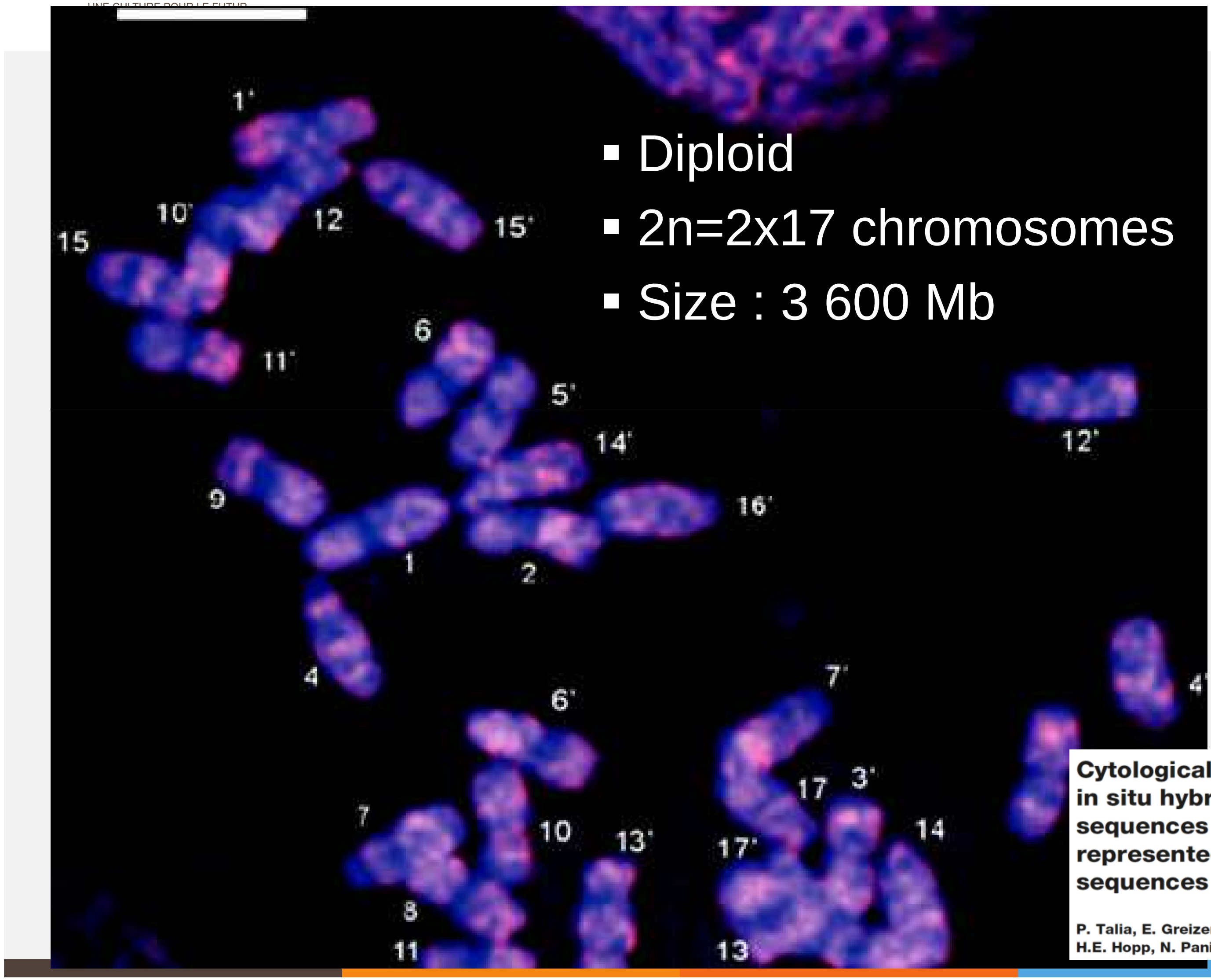
L. Rieseberg

John E. Bowers, ***John M. Burke***

UGA Athens

J. Burke

Sunflower genome background



- Diploid
- 2n=2x17 chromosomes
- Size : 3 600 Mb

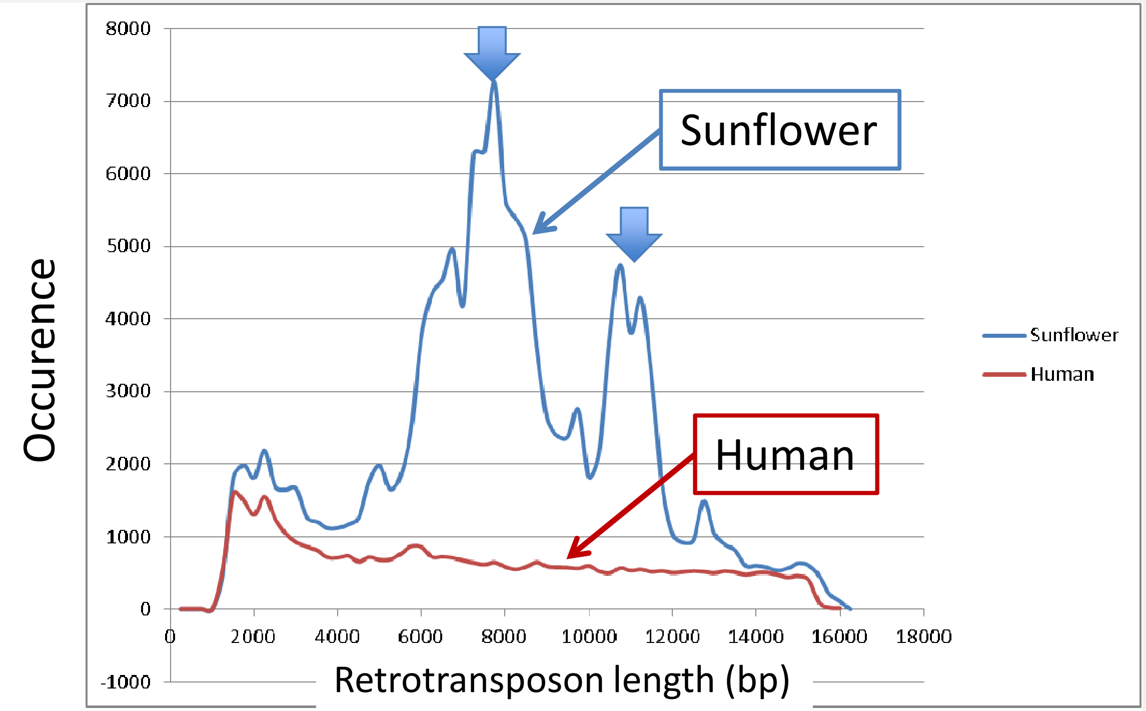
Species	Size
Rice	430 Mb
Rapeseed	1 100 Mb
Maize	2 300 Mb
H. sapiens	3 200 Mb
Sunflower	**3 600 Mb**
Wheat	17 000 Mb

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: 8kb and 11.5kb

The repeats make the assembling very difficult

2014: Sunflower genome Ha412.v1.1

- Sunflower line : HA412
- International Consortium

  UBC Vancouver, INRA Toulouse, UGA Athens

- Produced from 454 and Illumina sequencing

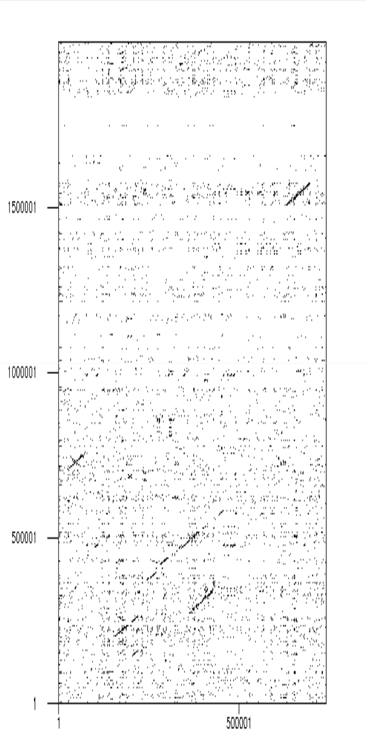
- 1 989 Mb (55% of 3.6Gb)

- Genome browser and annotation on

  www.heliagene.org

- Good at macro scale but local assembly problems

Ha412v1.1
(80% N)
1 900kb



HA412 BAC sequences
(No N)
700kb

2014: Sunflower genome Ha412.v1.1

•Sunflower line : HA412

•International Consortium 🇨🇦 🇫🇷 🇺🇸

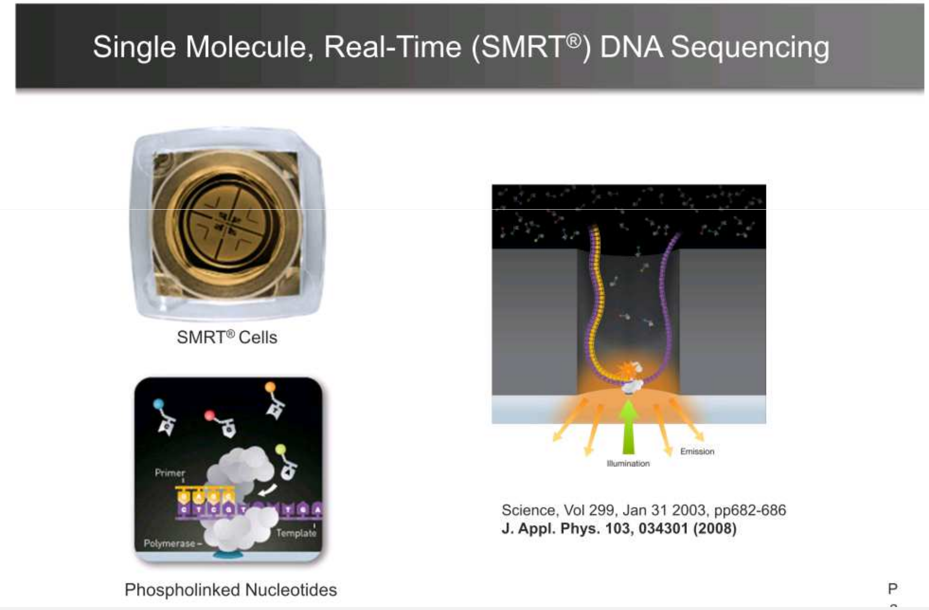
UBC Vancouver, INRA Toulouse, UGA Athens

A lot of genetic and genomic ressources produced: BAC libraries, physical map, high-density genetic maps, SNPs from re-sequenced genomes…

Difficulties were due to the short length of the sequences used for assembly that cannot span the long repeats

At the end of 2014 : a technological breakthrough

PacBio RSII (Pacific Biosciences)

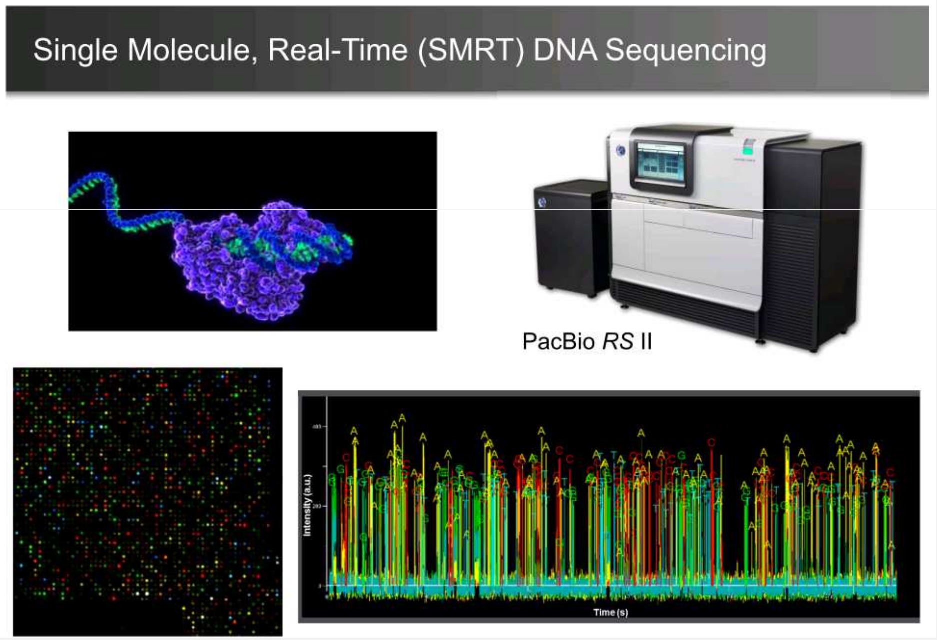


Single Molecule, Real-Time (SMRT®) DNA Sequencing

SMRT® Cells

Phospholinked Nucleotides

Science, Vol 299, Jan 31 2003, pp682-686
J. Appl. Phys. 103, 034301 (2008)



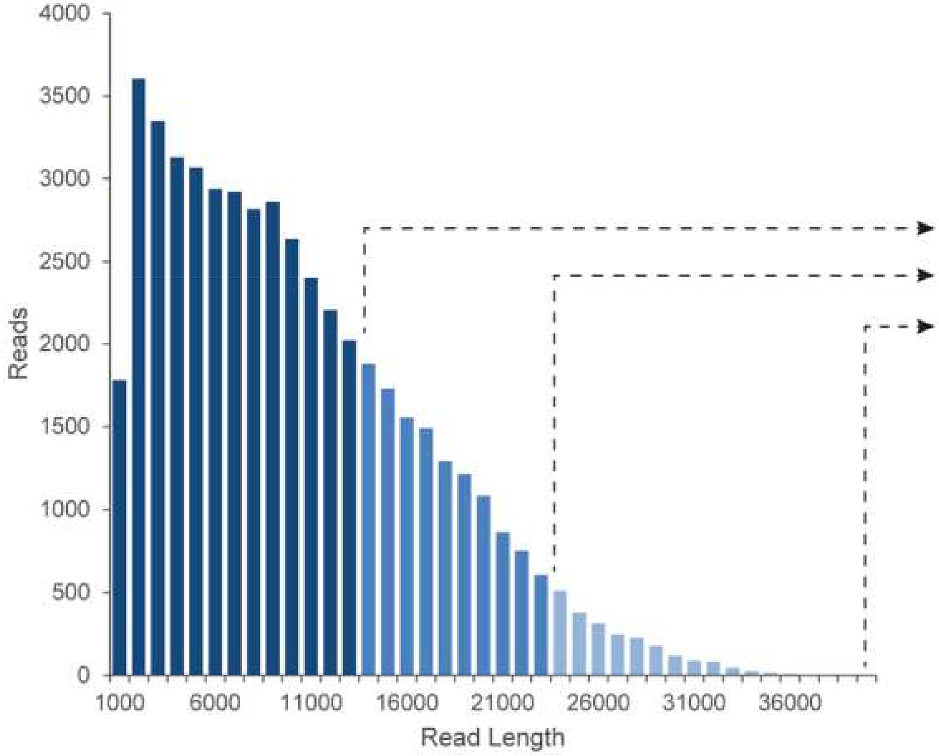
Single Molecule, Real-Time (SMRT) DNA Sequencing

PacBio *RS* II

A movie (fluorescence incorporation during the synthesis) is recorded and then converted to DNA sequence

At the end of 2014 : a technological breakthrough

P6-C4: Read Length Performance



Half of data in reads:	> 14 kb
Top 5% of reads:	> 24 kb
Maximum read length:	> 40 kb
Data per SMRT® Cell:	500 Mb – 1 Gb (in 4 hours)

PacBio produces sequences longer than the known repeats

P6-C4, 4-hr movie, 20-kb BluePippin™ size-selected *E. coli* library (1 SMRT Cell)

SUNRISE
UNE CULTURE POUR LE FUTUR

2015: Acquisition of the PacBio RSII at INRA Toulouse in march 2015

SUNRISE Project (2012-2019)

INRA Toulouse (LIPM, CNRGV, Genomic Platform)

First PacBio in France (GeT-PlaGe)

Sunflower line: XRQ

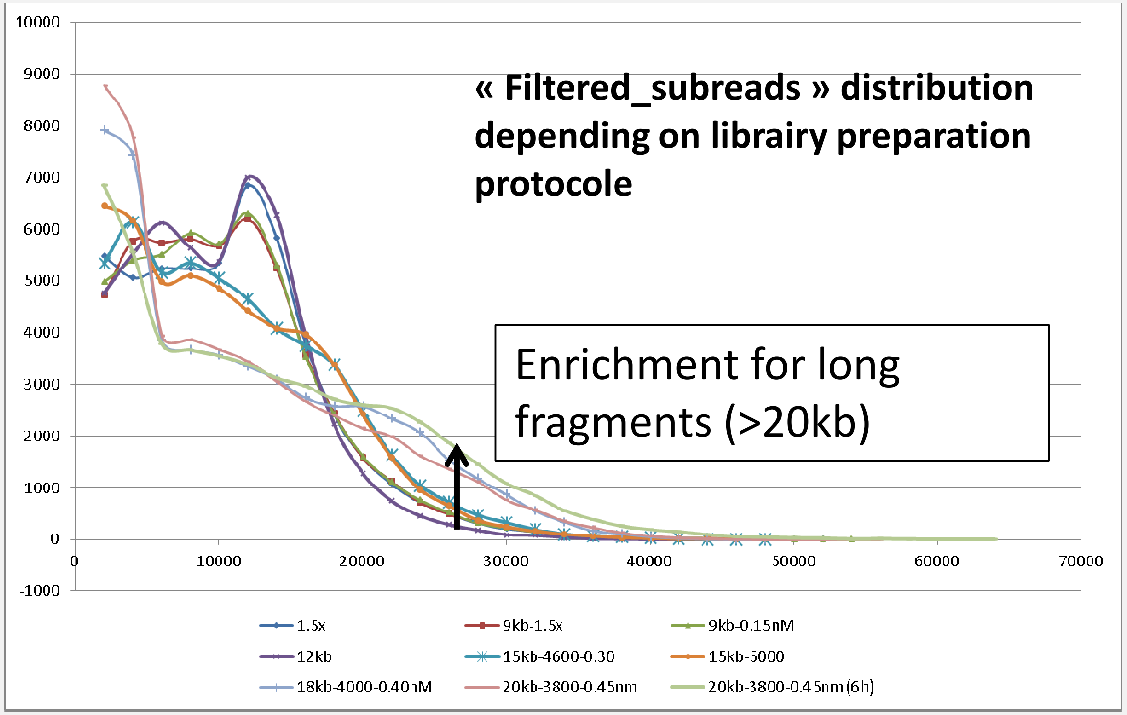
100% PacBio data used for the assembling.

SUNRISE
UNE CULTURE POUR LE FUTUR

Development of long-fragment librairies

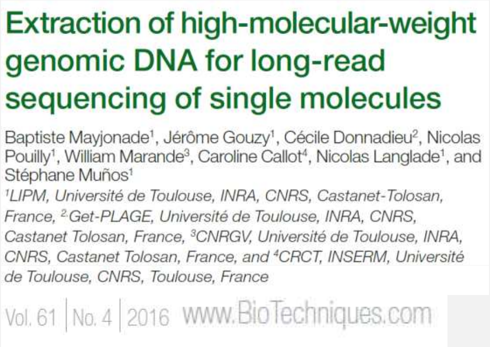
The longer the PacBio sequences are, the better it is to span LTR :

- New DNA extraction protocole (submitted to BioTechniques)
- Optimisation of fragmentation, purification, loading
- Increase run time to 4 → 6h (movie-length)



« 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

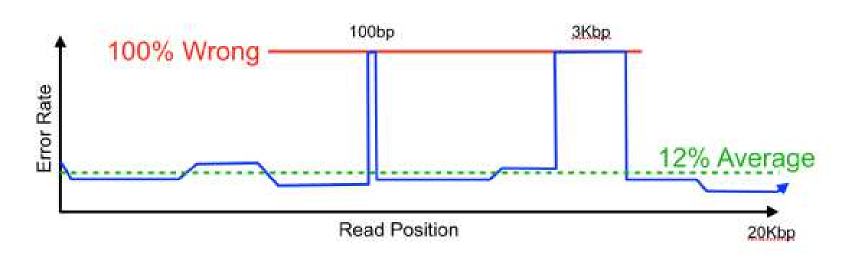
Vol. 61 | No. 4 | 2016 www.BioTechniques.com

B. Mayjonade

Production of contigs

J. Gouzy

1) Correction of the raw sequences



https://dazzlerblog.wordpress.com/2015/11/06/intrinsic-quality-values/

1.1: Pairwise comparison of raw sequences (**`PBcR/MHAP, minimap, falcon`**)
1.2: The sequences are corrected

2) Assembling of the corrected sequences

2.1: Pairwise comparison of the corrected sequences (**`WGS/Falcon`**)
2.2: Sequences are aligned (based on the overlap between sequences), the contig is breaked on the point where the repeat is not spanned.
2.3: consensus sequences of the contigs

3) « Polishing » of the consensus contigs sequences

3.1: mapping of the raw data on the consensus sequences (**`Blasr`**)
3.2: correction of the consensus sequence based on the error rate model of the polymerase (**`quiver`**)

PacBio raw data produced (April – July 2015)

B. Mayjonade

Raw data (407 SMRT cells)

#	MAX	N50 BP	MEAN	Gb
37,5M	80,9kb	*13,7kb*	*10,3kb*	367 (102x)

/3

Corrected reads (PBcR)

	#	MAX	N50 BP	MEAN	Gb
CR1	11,2M	59kb	13,6kb	11,2kb	125 Gb (34x)

Third of the raw data are conserved after correction and used for assembling

WORLD WINNER

80 974 nucleotides

New challenge

VS

Assembly result of the contigs

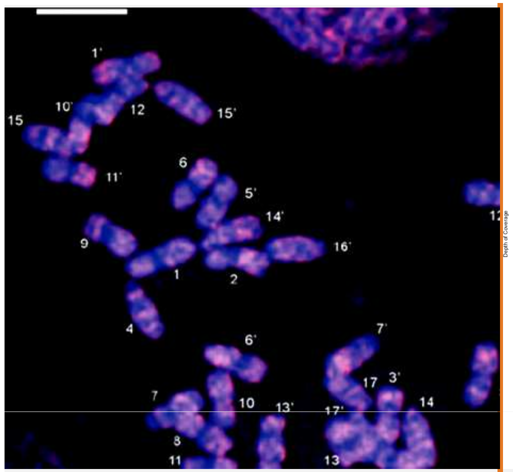
#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
13 124	4.4M	498 kb	1700	118 kb	3.03

➔80% of genome in the contigs (No Ns)

➔20% not assembled (likely concatemeres of rDNA, TE, telomeres, centromeres)

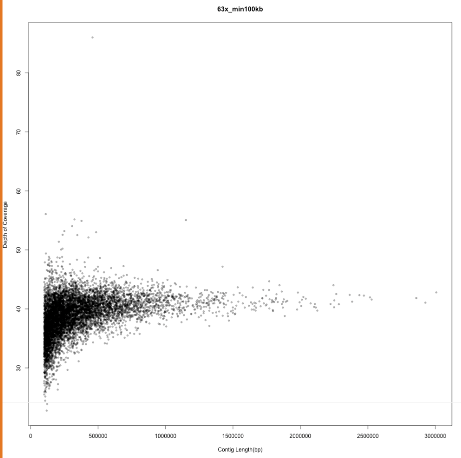
SUNRISE
UNE CULTURE POUR LE FUTUR

From contigs to chromosome sequences



Sunflower Genome

- 17 chromosomes
- 3.6G base pairs

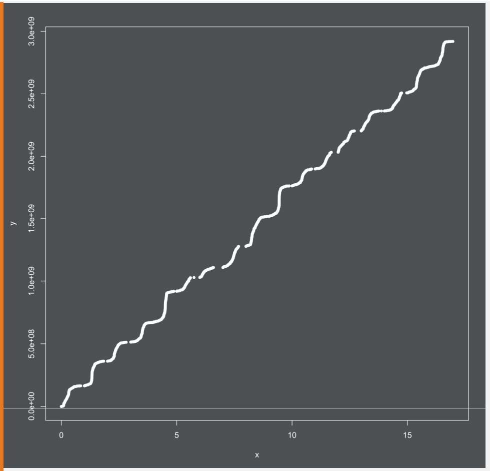


Genome assembly

- 12,318 Contigs (putative chimeric contigs discarded)
- 2.93 Gb

Chris Grassa

FROM CONTIGS TO PSEUDO-MOLECULES

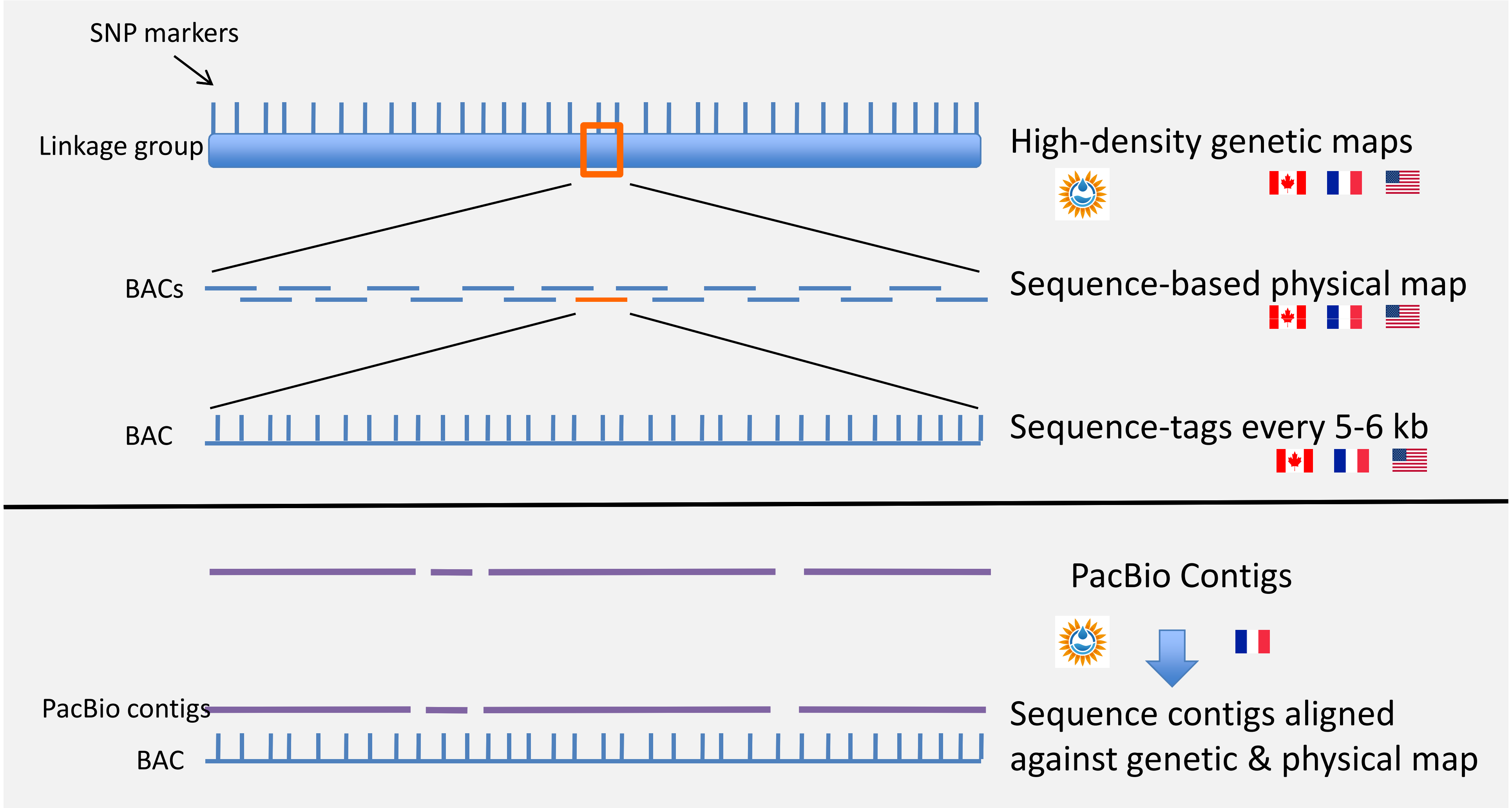


Reference Genome

- 17 pseudomolecules
- +chloroplast and mitochondrion

SUNRISE
UNE CULTURE POUR LE FUTUR

Chromosome production strategy



SNP markers

Linkage group

High-density genetic maps

BACs

Sequence-based physical map

BAC

Sequence-tags every 5-6 kb

PacBio Contigs

PacBio contigs

BAC

Sequence contigs aligned
against genetic & physical map

SUNRISE
UNE CULTURE POUR LE FUTUR

Integration of High-Density Genetic Maps

INRA
(Muños)

- 86,223 markers
- 3 Populations:
  - HA89 x LR1
  - XRQ x PSC8 - 2014
  - XRQ x PSC8 - 2015

UGA
(Bowers)

- 10,080 markers
- 4 Populations:
  - HA412 x RHA415
  - HA412 x ANN1238
  - NMS373 × Hopi
  - RHA280 x RHA801

USDA
(Talduker)

- 5,019 RAD-tag markers
- 3 F2 Populations:
  - HA89 x RHA464
  - B-line x RHA464
  - CR29 x RHA468

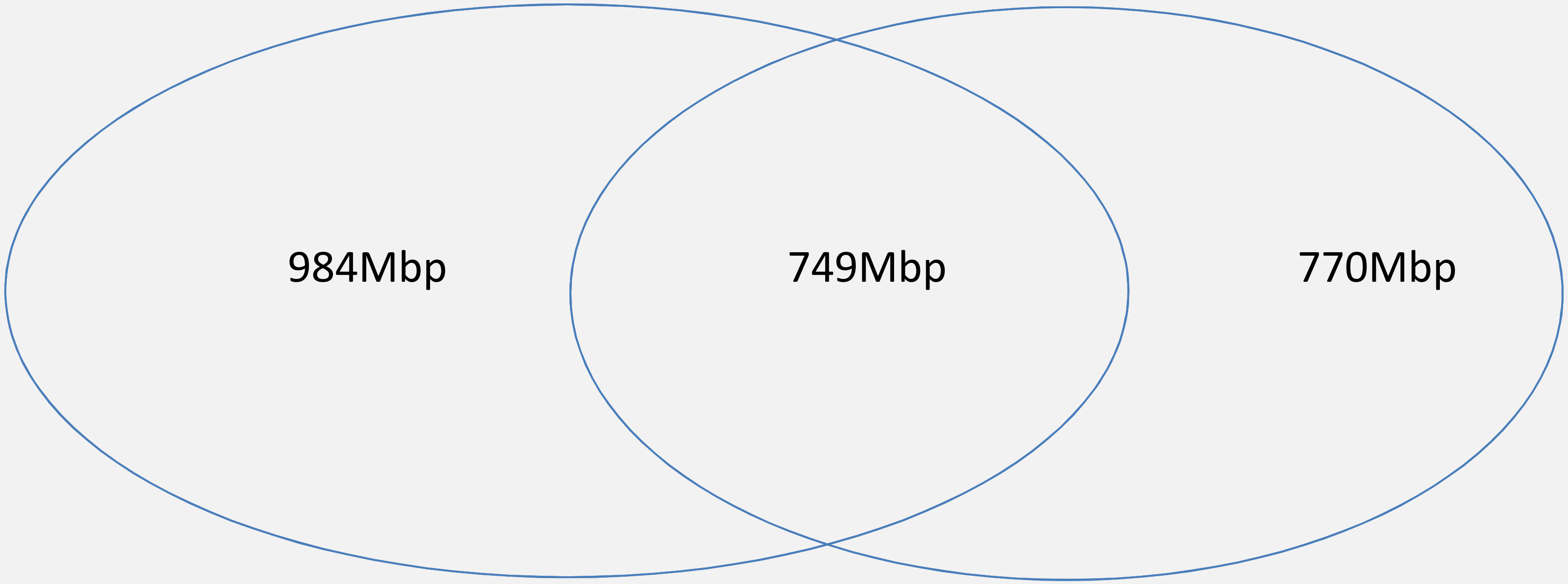
UBC
(Grassa)

- Sequenced-based (~2.5M SNPs)
- 1 Population:
  - RHA280 x RHA801

Complementarity of genetic maps

UBC map
#contigs: 12 209
bp placed: 1 733 Mb

INRA 2015 map
# contigs: 3 703
bp placed: 1 518 Mb



984Mbp

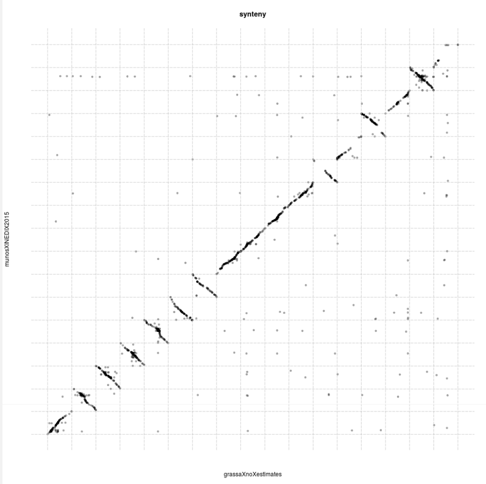
749Mbp

770Mbp

Other maps (INRA 2014, UGA, USDA): 415Mbp

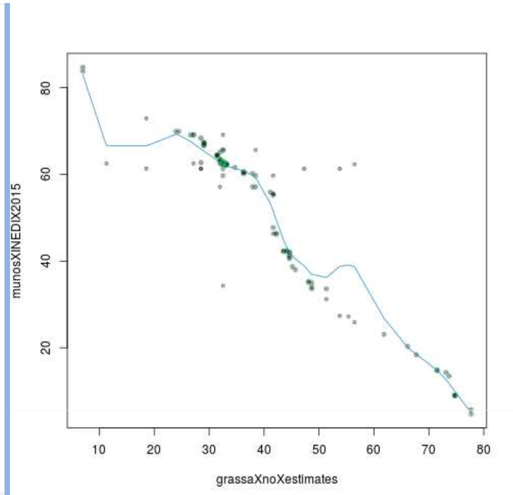
Build Consensus Map Units

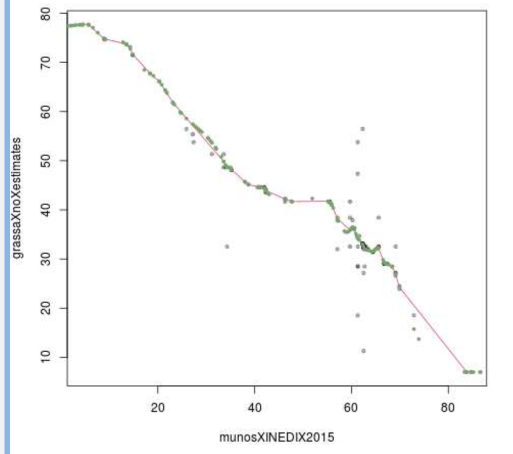
C. Grassa



Raw Maps

- Mostly agree
- Minor ordering differences
- Some LGs inverted
- Recombination rate varies





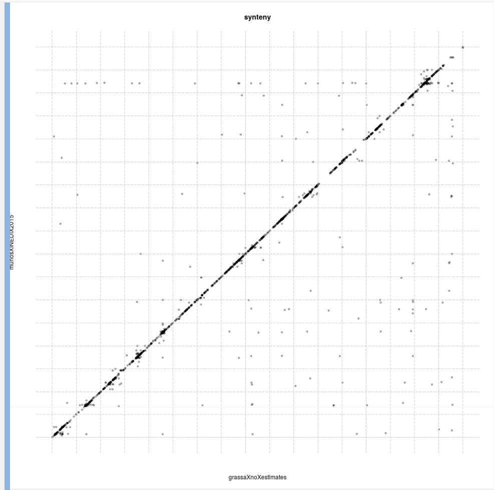
Machine-learn consensus units:

1) Loose fit curve using contigs with markers in both maps

2) Drop outliers

3) Train model

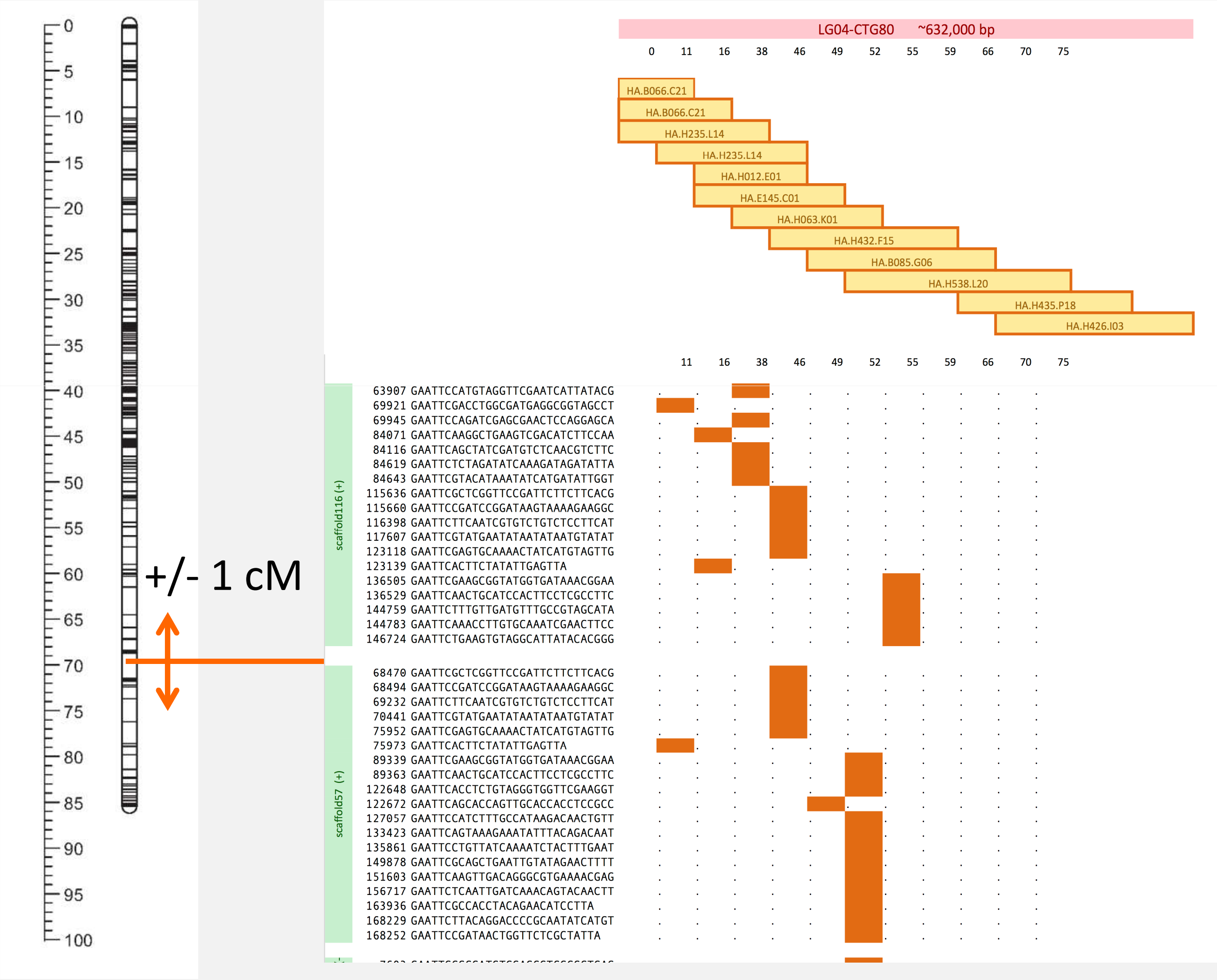
4) Predict positions in consensus units



Consensus Map

- Near colinearity
- **Common map units**

Physical Map Scaffolding



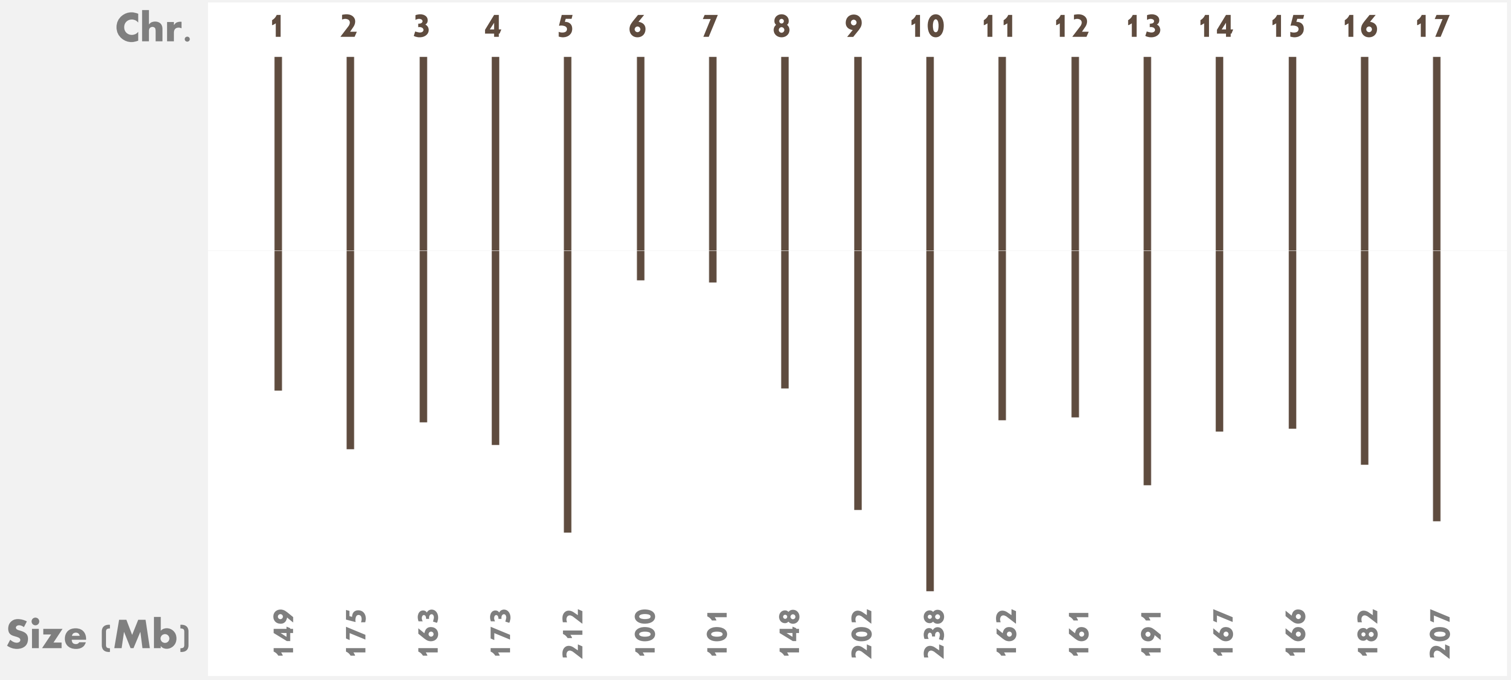
```
 63907 GAATTCCATGTAGGTTCGAATCATTATACG
 69921 GAATTCGACCTGGCGATGAGGCGGTAGCCT
 69945 GAATTCCAGATCGAGCGAACTCCAGGAGCA
 84071 GAATTCAAGGCTGAAGTCGACATCTTCCAA
 84116 GAATTCAGCTATCGATGTCTCAACGTCTTC
 84619 GAATTCTCTAGATATCAAAGATAGATATTA
 84643 GAATTCGTACATAAATATCATGATATTGGT
115636 GAATTCGCTCGGTTCCGATTCTTCTTCACG
115660 GAATTCCGATCCGGATAAGTAAAAGAAGGC
116398 GAATTCTTCAATCGTGTCTGTCTCCTTCAT
117607 GAATTCGTATGAATATAATATAATGTATAT
123118 GAATTCGAGTGCAAAACTATCATGTAGTTG
123139 GAATTCACTTCTATATTGAGTTA
136505 GAATTCGAAGCGGTATGGTGATAAACGGAA
136529 GAATTCAACTGCATCCACTTCCTCGCCTTC
144759 GAATTCTTTGTTGATGTTTGCCGTAGCATA
144783 GAATTCAAACCTTGTGCAAATCGAACTTCC
146724 GAATTCTGAAGTGTAGGCATTATACACGGG

 68470 GAATTCGCTCGGTTCCGATTCTTCTTCACG
 68494 GAATTCCGATCCGGATAAGTAAAAGAAGGC
 69232 GAATTCTTCAATCGTGTCTGTCTCCTTCAT
 70441 GAATTCGTATGAATATAATATAATGTATAT
 75952 GAATTCGAGTGCAAAACTATCATGTAGTTG
 75973 GAATTCACTTCTATATTGAGTTA
 89339 GAATTCGAAGCGGTATGGTGATAAACGGAA
 89363 GAATTCAACTGCATCCACTTCCTCGCCTTC
122648 GAATTCACCTCTGTAGGGTGGTTCGAAGGT
122672 GAATTCAGCACCAGTTGCACCACCTCCGCC
127057 GAATTCCATCTTTGCCATAAGACAACTGTT
133423 GAATTCAGTAAAGAAATATTTACAGACAAT
135861 GAATTCCTGTTATCAAAATCTACTTTGAAT
149878 GAATTCGCAGCTGAATTGTATAGAACTTTT
151603 GAATTCAAGTTGACAGGGCGTGAAAACGAG
156717 GAATTCTCAATTGATCAAACAGTACAACTT
163936 GAATTCGCCACCTACAGAACATCCTTA
168229 GAATTCTTACAGGACCCCGCAATATCATGT
168252 GAATTCCGATAACTGGTTCTCGCTATTA
```

- Tags aligned to physical map with simple scoring scheme
- Reciprocal best hits seed the scaffold position
- Successive matches searched +/- 1cM from seed

Sunflower reference genome (XRQ line)

3 027 Mb (3.38 % de N)



Mitochondrion

301 kb

Chloroplast

151 kb

98.5% of contigs in the pseudomolecules

Gene content and genome annotation

61 RNA-Seq librairies on the sequenced genotype (XRQ)

- **Organ-specific** expression (12 organs)
- **Abiotic stress** response: drought, osmotic stress, salt stress (in roots and leaves)
- **Hormone regulation**: (9 hormones in roots and leaves)

Gene annotation

- **52 243** protein coding genes (mRNA)
- **4 945** lncRNA genes
- **88** pre-miRNA genes (351 mature miRNA)
- **862** tRNA and rRNA genes

98% of transcripts mapped on pseudo-molecules

User interface and data visualization

Expression data

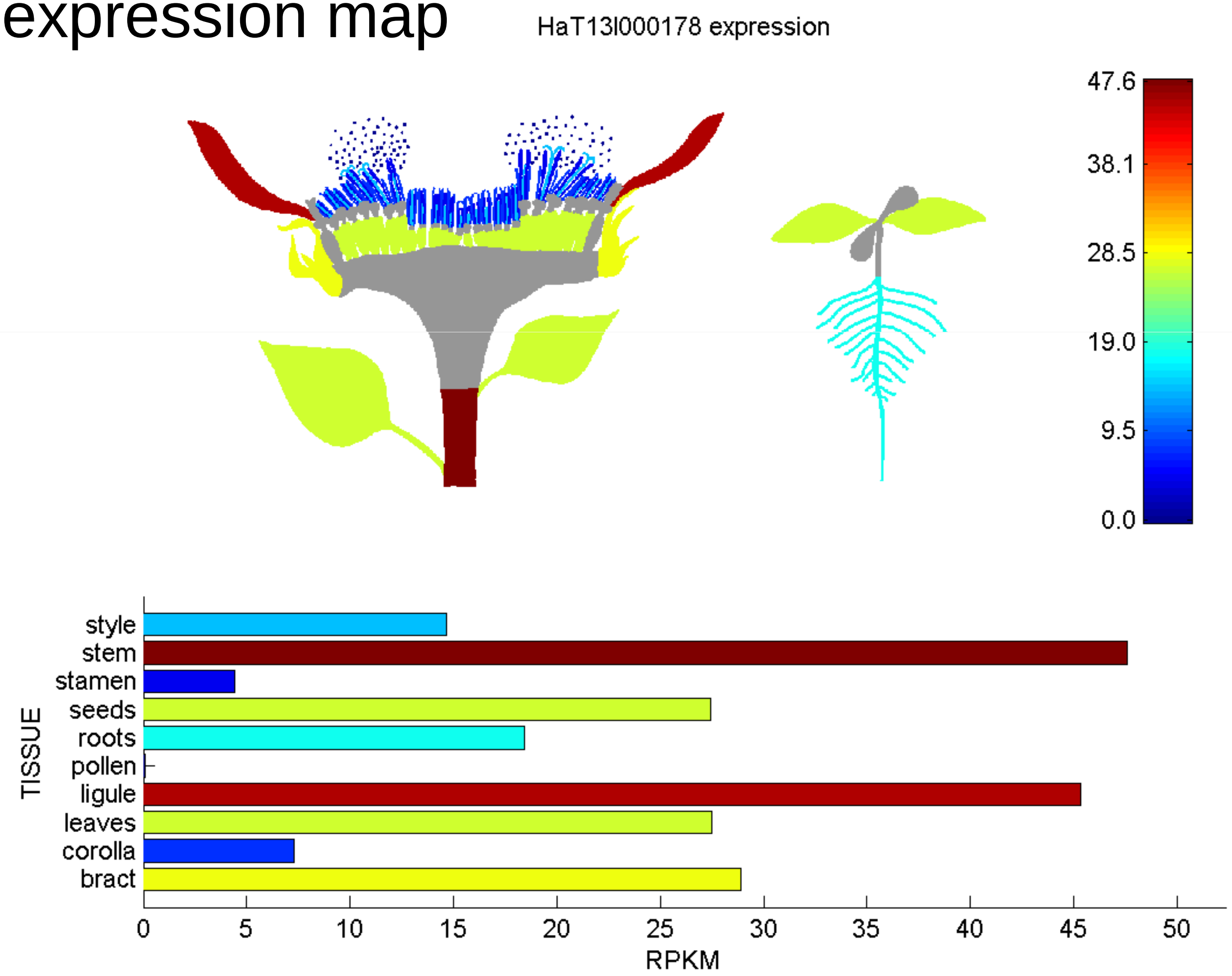
Maintain access to raw count data

Integrate visualization on www.heliagene.org

SUNRISE
UNE CULTURE POUR LE FUTUR

User interface and data visualization

Gene expression map



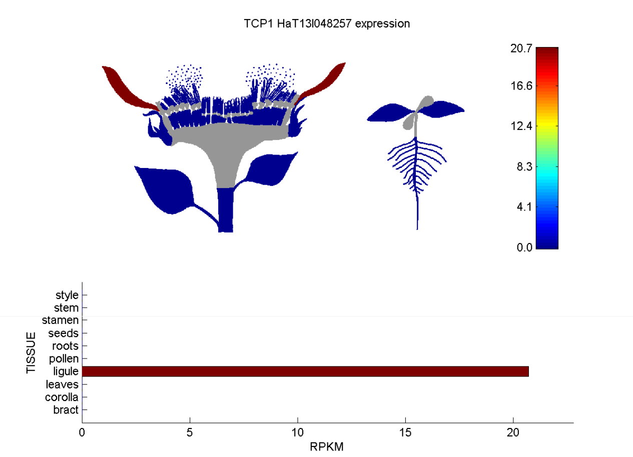
N. Langlade

SUNRISE
UNE CULTURE POUR LE FUTUR

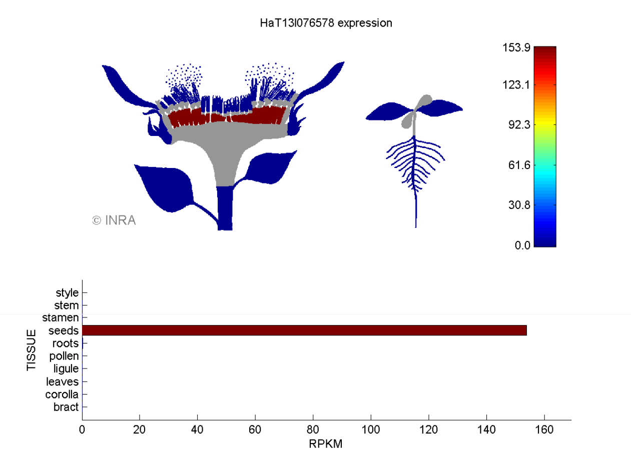
User interface and data visualization

Examples of organ-specific genes

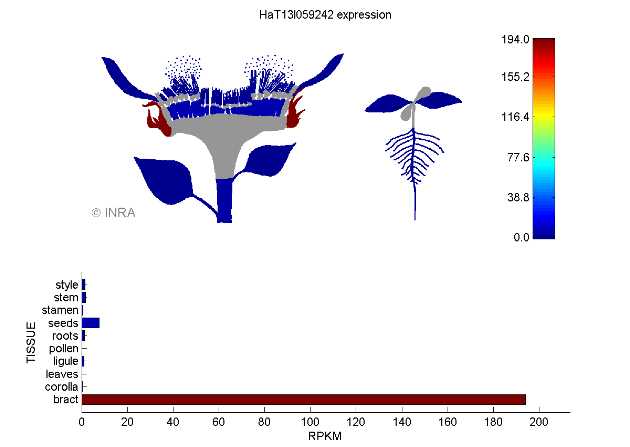
CYC1



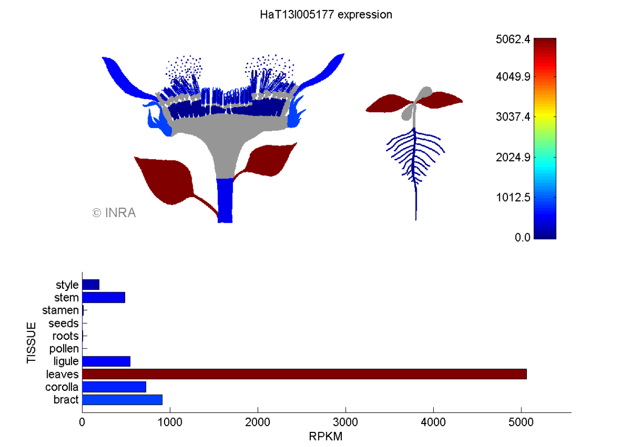
Glutenin



pre-miRNA



GAPDH



What next for sunflower?

Improving the sunflower assembly.

Optical mapping
CNRGV-INRA Toulouse (N. Rodde, C. Chantry, H. Bergès)
Irys system (Bionano) acquired in March 2016

NRGene on HA412 line

SUNRISE
UNE CULTURE POUR LE FUTUR

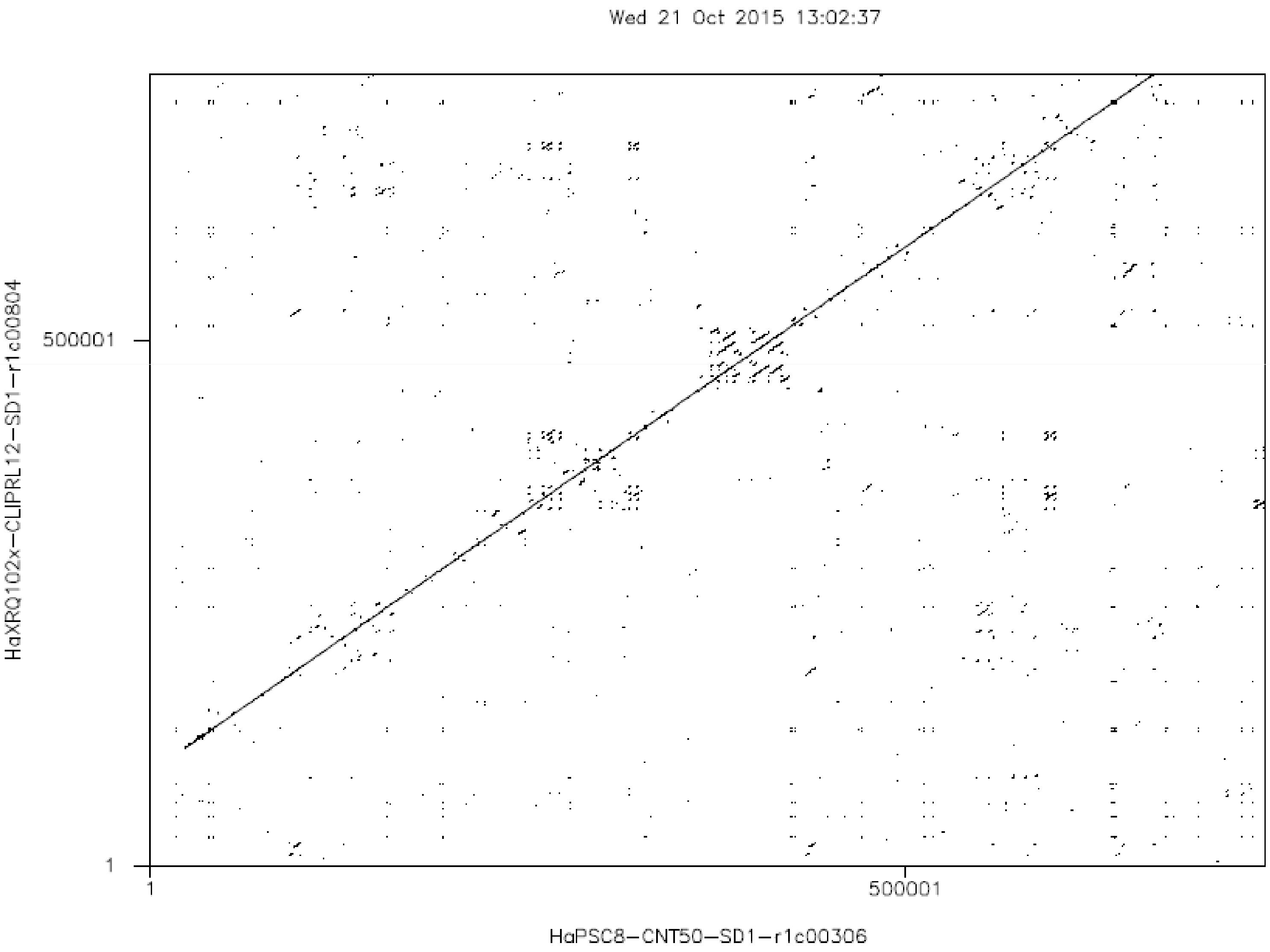
What Next for sunflower?

Several reference genomes are needed because, nor XRQ nor HA412 represent the averall genetic variations in sunflower

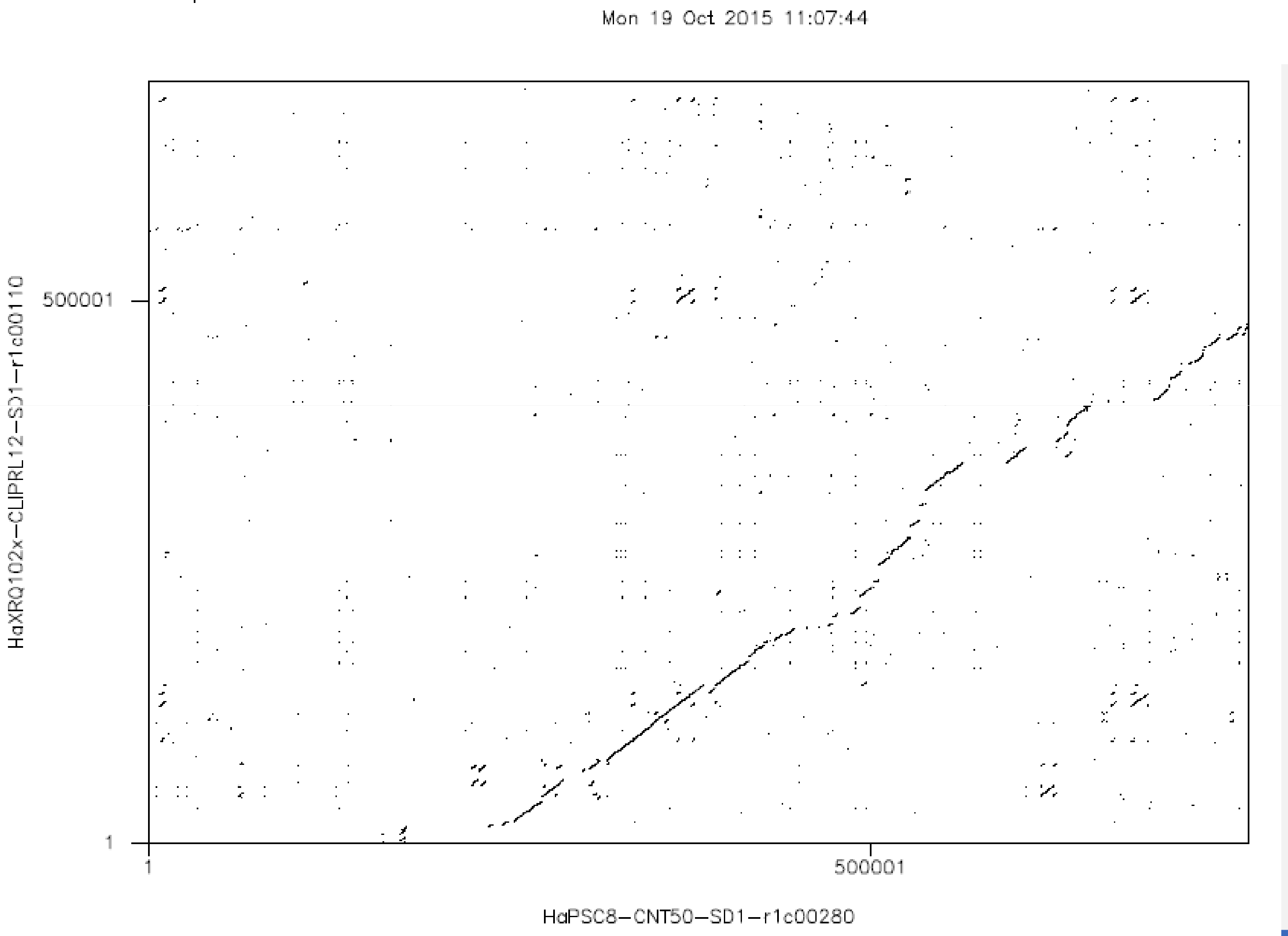
A first step : PSC8 sunflower line *de novo* sequenced (50X PacBio data, HeliOr project)

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
26 273	2.5M	223kb	3799	66 kb	3.15

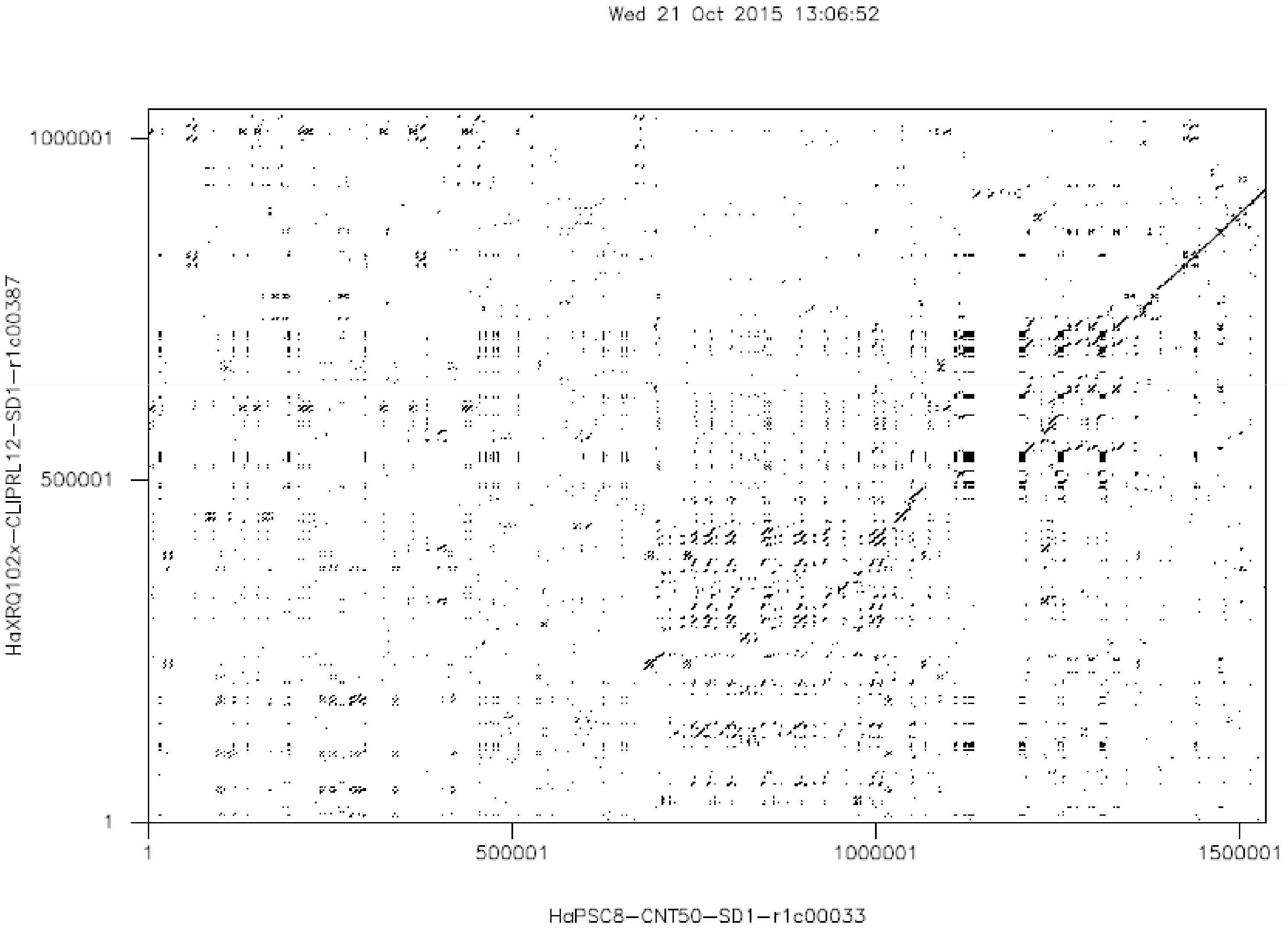
XRQ vs PSC8: some conserved regions



XRQ vs PSC8: regions with structural variations



XRQ vs PSC8: highly divergent regions



SUNRISE
UNE CULTURE POUR LE FUTUR

Summary

A high quality genome sequence produced (XRQ line)

www.heliagene.org

Sunflower could be a model plant like tomato became for fleshy fruits!

But breeding and genetic research need more genomes to be sequenced and more tools and data.

Many thanks to our colleagues now retired

Felicity Vear
(INRA Clermont-Ferrand)

Patrick Vincourt
(INRA Toulouse)

Thank you for your attention

Fundings

SOFIPROTEOL
L'engagement durable

P.R.O.M.O.S.O.L.
ASSOCIATION
POUR LA PROMOTION
DE LA SÉLECTION
DES PLANTES OLÉAGINEUSES

Partners