SUNRISE
UNE CULTURE POUR LE FUTUR

Result of the *de novo* Sequencing of the Complex Sunflower Genome Using PacBio Technology (100X)

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

INRA Toulouse

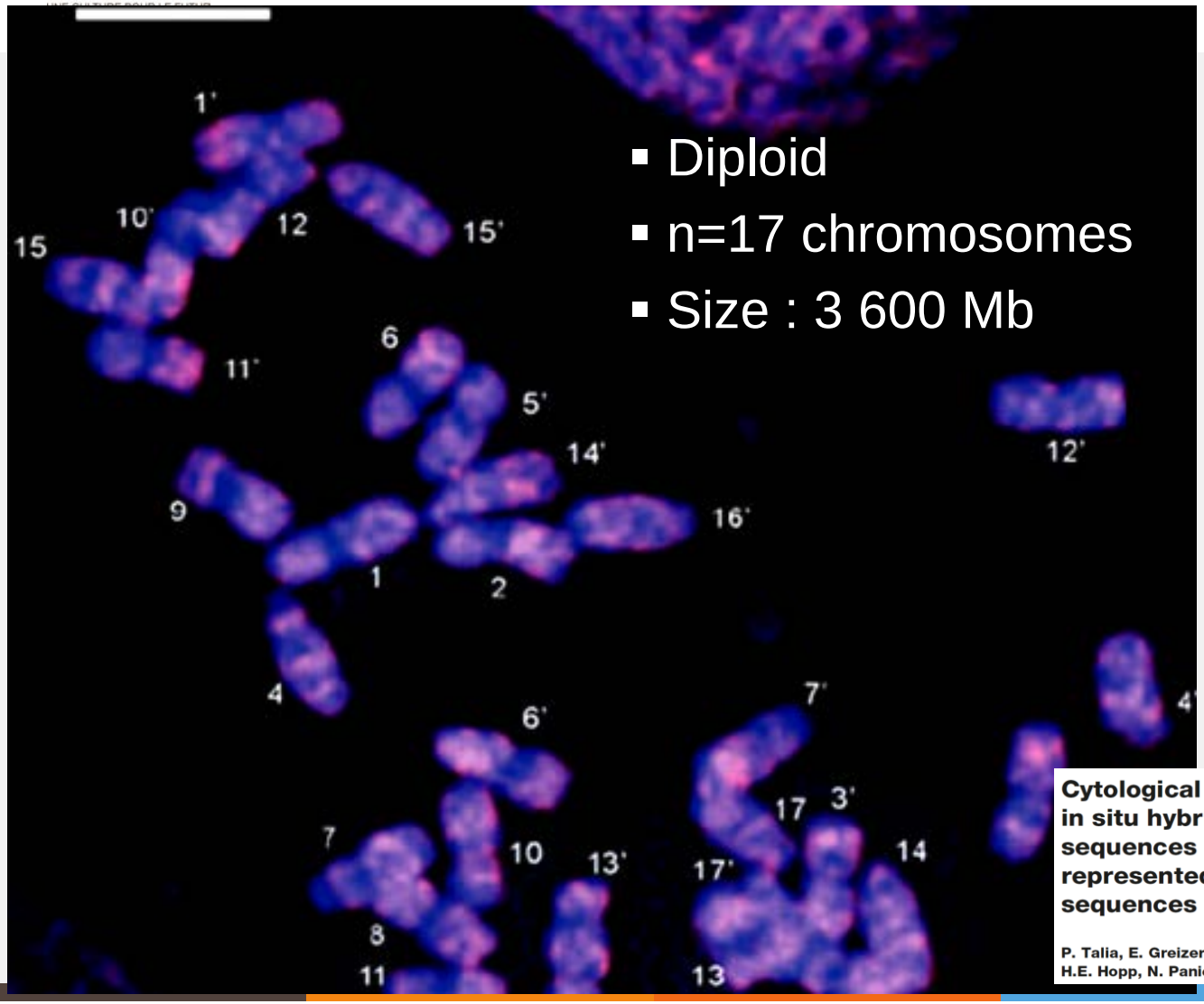
Navdeep Gill, Thuy Nguyen, Nolan Kane, Loren H. Rieseberg

UBC Vancouver

John E. Bowers, John M. Burke

UGA Athens

Sunflower genome background



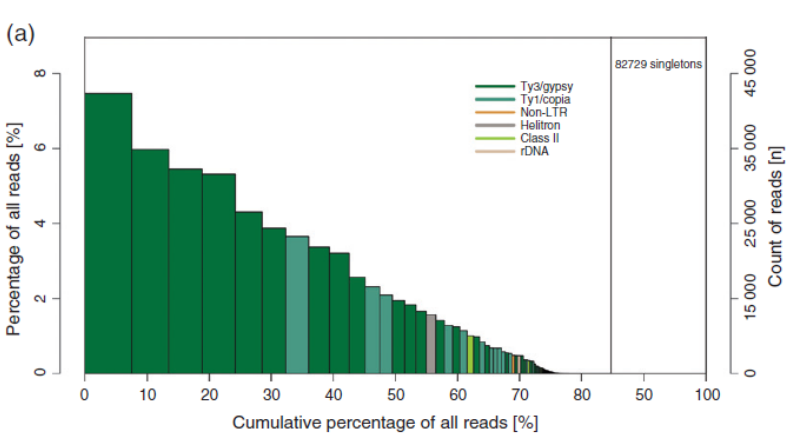
- Diploid
- n=17 chromosomes
- Size : 3 600 Mb

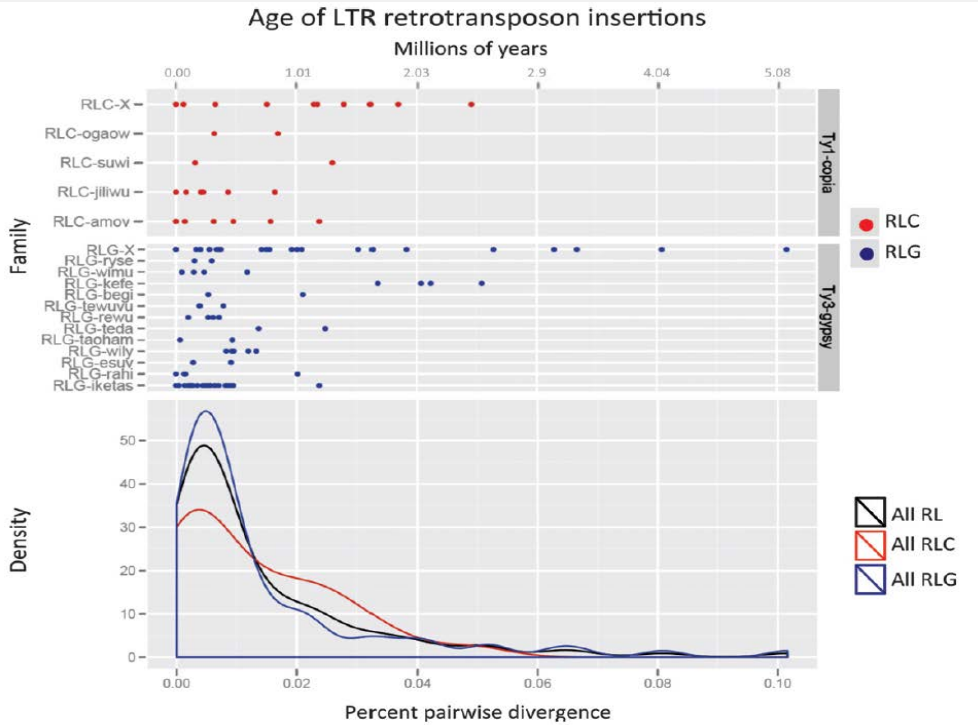
Species	Size
Rice	430 Mb
Rapeseed	1 100 Mb
Maize	2 300 Mb
H. sapiens	3 200 Mb
Sunflower	**3 600 Mb**
Wheat	17 000 Mb

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 is highly repeated





~ 80% known repeated sequences

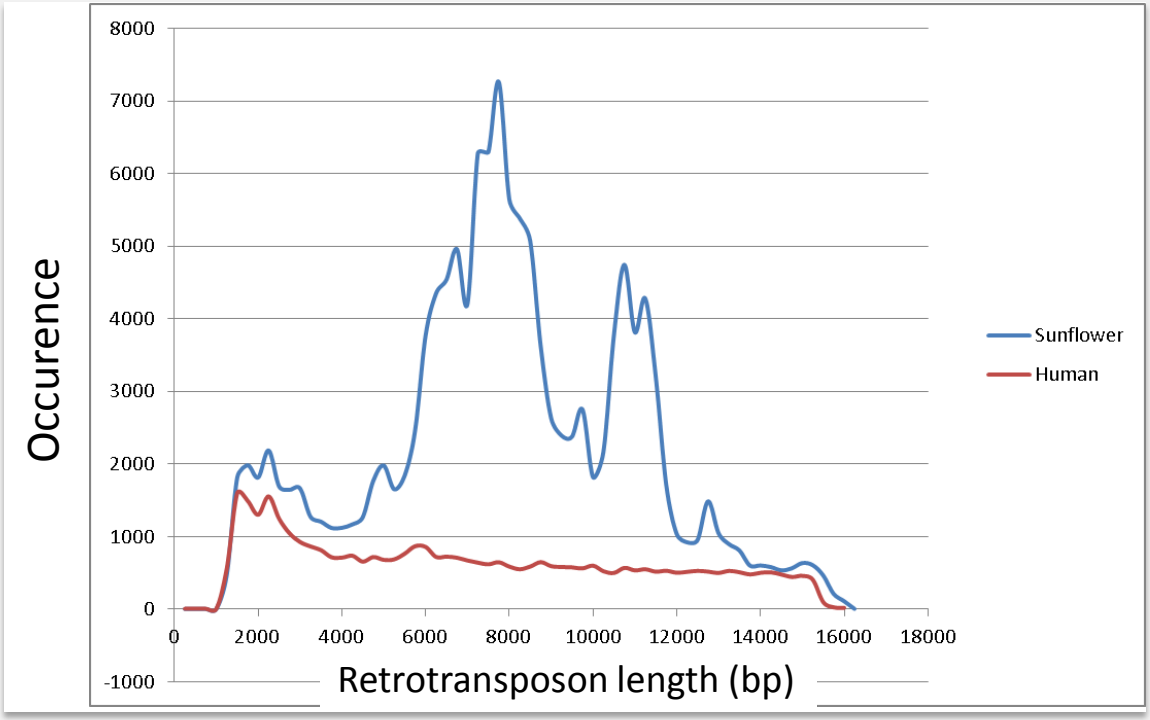
Majority of LTR < 0.5My

The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements

S. Evan Staton[1], Bradley H. Bakken[2], Benjamin K. Blackman[3,†], Mark A. Chapman[4,‡], Nolan C. Kane[5], Shunxue Tang[6,§], Mark C. Ungerer[2], Steven J. Knapp[6,¶], Loren H. Rieseberg[5] and John M. Burke[4,*]

Sunflower genome contains long repeated elements

Length distribution of LTR retrotransposons



33% of sunflower genome

8% human genome

LTR up to 12kb

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

2014: Sunflower genome Ha412.v1.1

- International Consortium

  UBC Vancouver, INRA Toulouse, UGA Athens

- Produced from 454 and Illumina sequencing

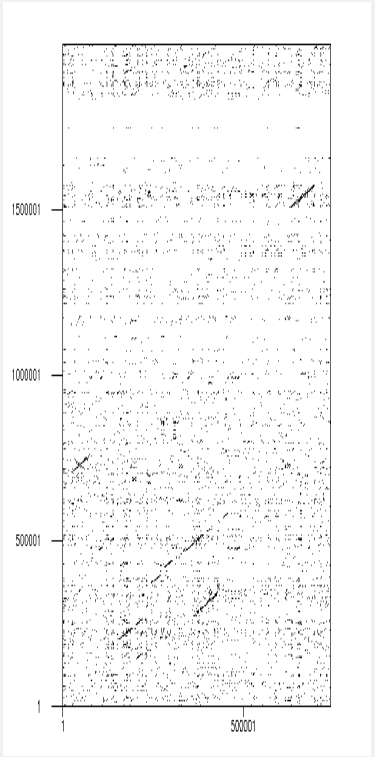
- Genome browser and annotation on

  www.heliagene.org

- 1 989 Mb (55% of 3 600Gb)

- Good at macro scale but local assembly problems

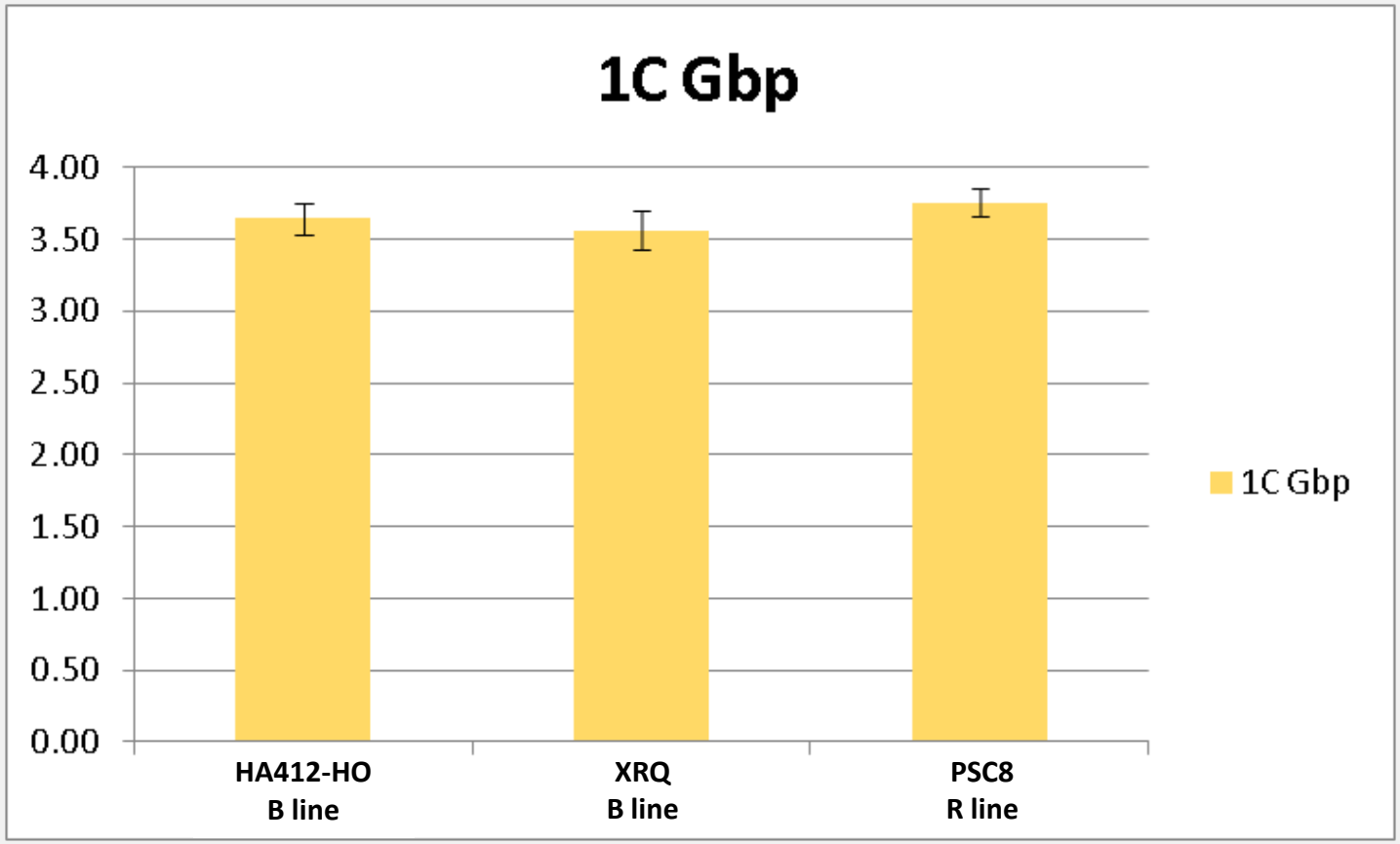
Ha412v1.1
(80% N)
1 900kb



PacBio from BAC contig
(No N)
700kb

Genome size estimation

Flux cytometry: Olivier Catrice (LIPM, INRA Toulouse)



2015: PacBio sunflower genome

SUNRISE Project (2012-2019)

INRA Toulouse (LIPM, CNRGV, Genomics Platform)

- 407 SMRT Cells with P6/C4

-102X ➔ 4.7Tb raw data

-Acquisition of PacBio RSII at INRA Toulouse

-April – July 2015

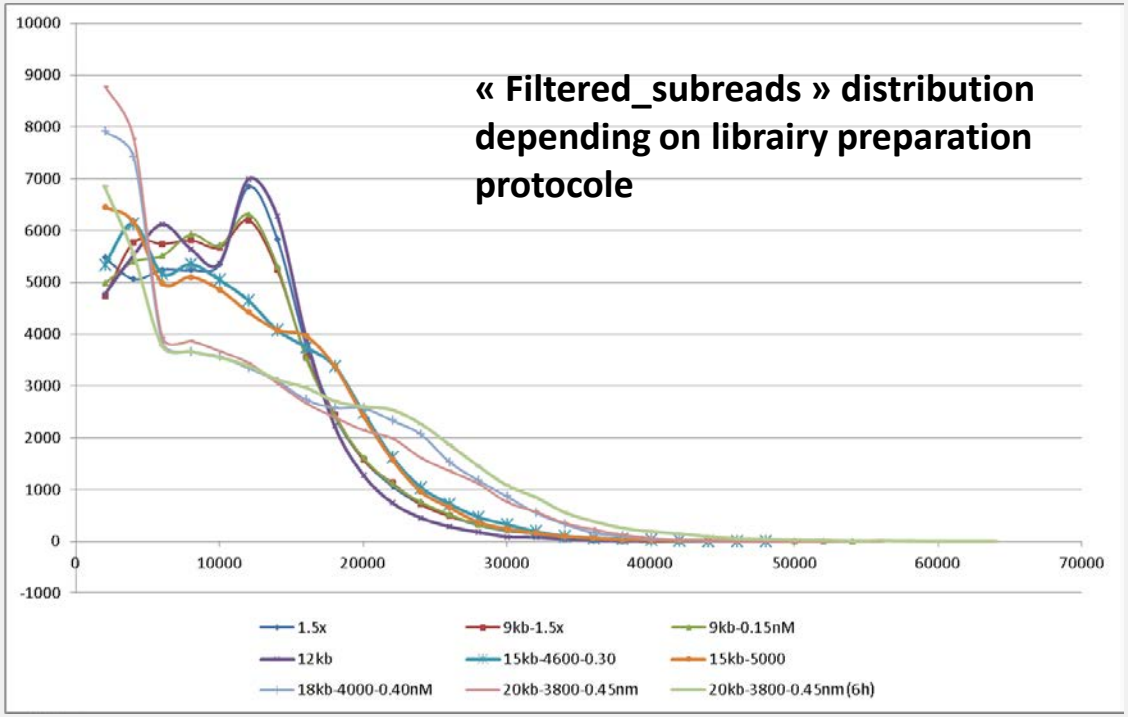
- Sequencing paralleled on 3 sites:
  - IGM (UC San Diego): 202 SMRT Cells
  - **INRA Toulouse GET-Plage: 159 SMRT Cells**
  - Lausanne University: 59 SMRT Cells

] New library prep

Development of long-fragment librairies

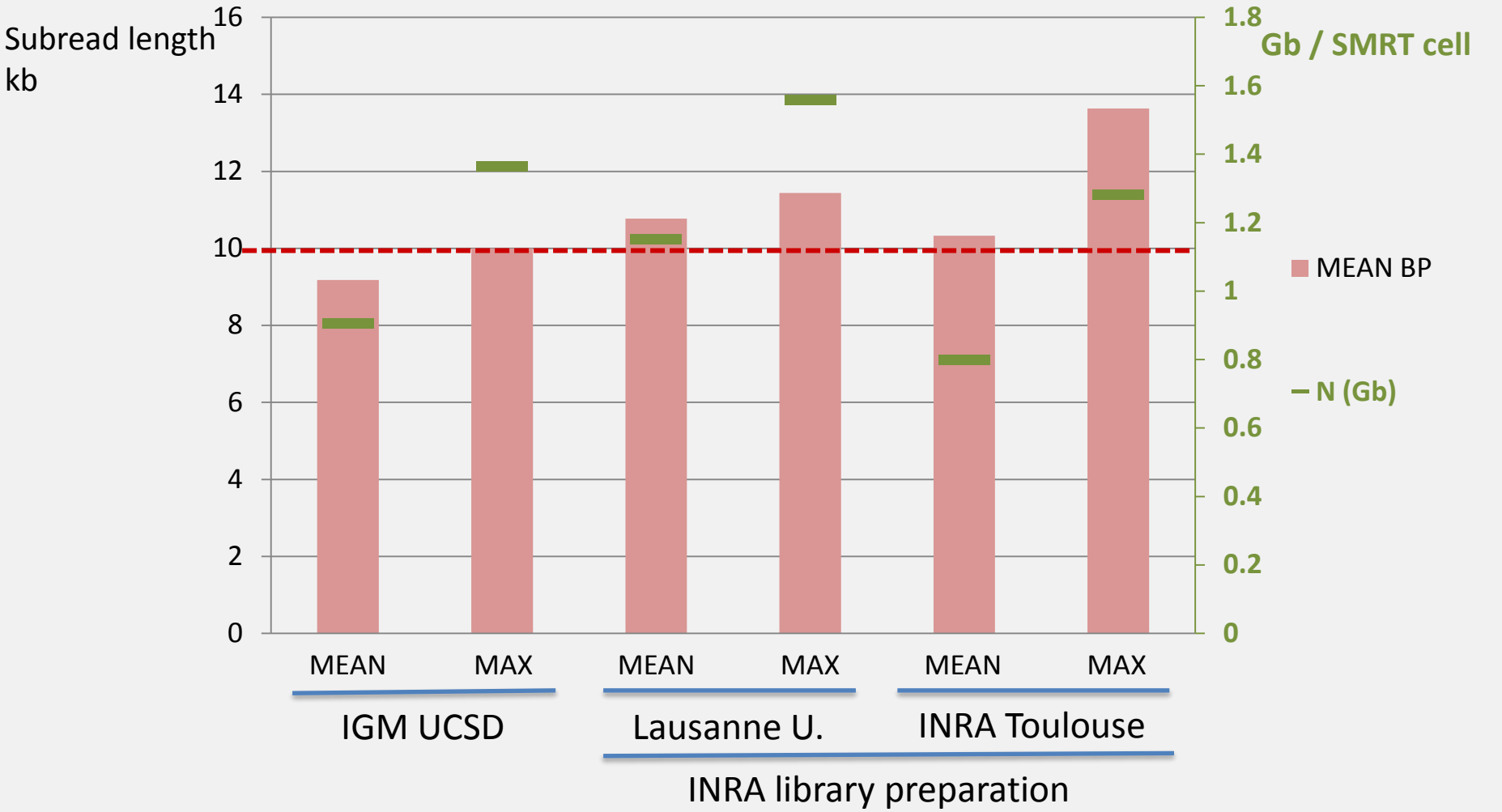
Accute need for long and pure DNA fragments

- New DNA extraction protocole / fragmentation / purification
- Optimisation of loading
- Increase run time to 4 → 6h

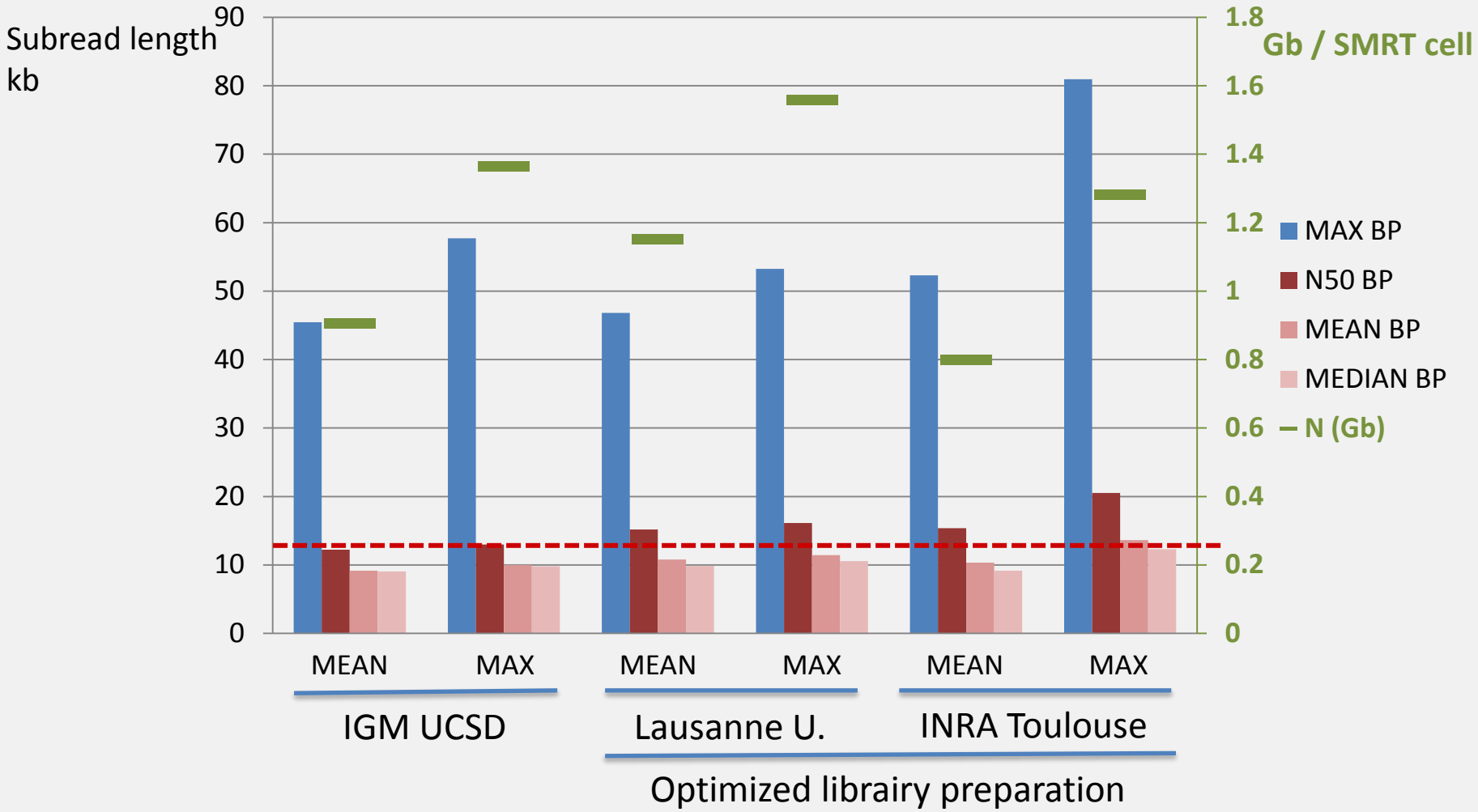


B. Mayjonade

PacBio raw data production



PacBio raw data production



Production of contigs

J. Gouzy

100% PacBio data for assembly

PacBio Corrected Reads (PBcR) pipeline

Koren *et al.* 2012 Nat Biotech

Read correction: MinHash Alignment Process (MHAP) / falcon_sense (PacBio)

Berlin *et al.* 2015 Nat Biotech

Contiguing: WGS/CABOG (overlap – *layout* – consensus)

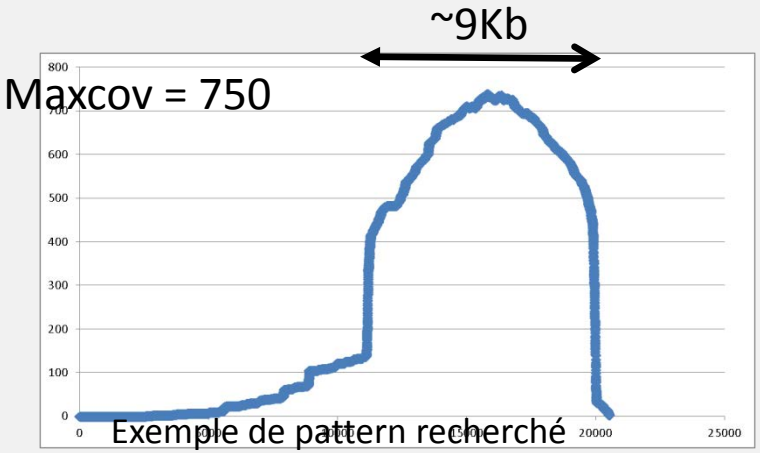
Polishing: quiver (PacBio)

Need to optimize PBcR and layout step due to proportion of repeated and highly conserved sequences

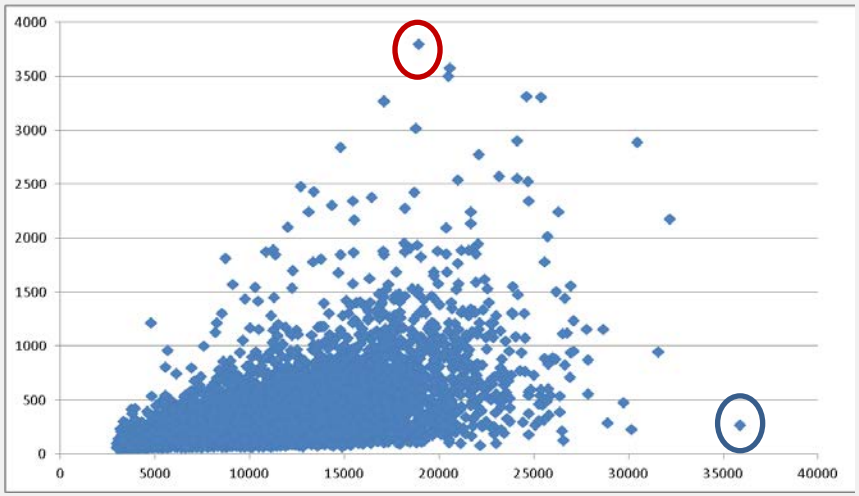
Step 0: suppression of non-informative data to avoid spending time and storage space on repetitive elements

- Mapping de 1x de données sur 2x de lectures longues (>= 20Kb)
- Analyse de la couverture des lectures « longues » en ne considérant que les hits de plus de 3kb
- Recherche d'un pattern de type « unité de répétition » (MHAP/MinHash)

➔Construction d'une banque d'unité de répétition « draft »



Exemple de pattern recherché



➔ Suppression au fil de l'eau des séquences brutes non informatives = complètement incluses dans une unité de répétition

#	MAX	N50 BP	NUM >= N50	MEAN	BP
32,8M	80,9kb	13,7kb	9,1M	10,3kb	339 Gb (94x)

General principle of 100% PacBio assembly

- 2 pipelines très proches conceptuellement et techniquement
  - Les lectures sont corrigées avant d'être assemblées par WGS(CABOG). Les consensus sont corrigés avec « quiver » de PacBio
  - Les différences se situent dans les paramétrages par défaut et les versions utilisées

	HGAP 3 (PacBio=PB)	PBcR (Koren *et al.*)
Read correction		
Alignement	**PB/BLASR**	**MHAP (Berlin et al.) ou PB/BLASR**
Correction	**PB/dagcon**	**PBcR (PB/falconcns|PB/dagcon)**
Contiguing		
Overlap	**CA/overlap**	**CA/overlap**
Layout	**CA/unitigger**	**CA/unitigger (bogart)**
Consensus	**CA/utgcns**	**CA/utgcns (pbutgcns)**
Assembly correction		
Polishing	**PB/Quiver**	**PB/Quiver**

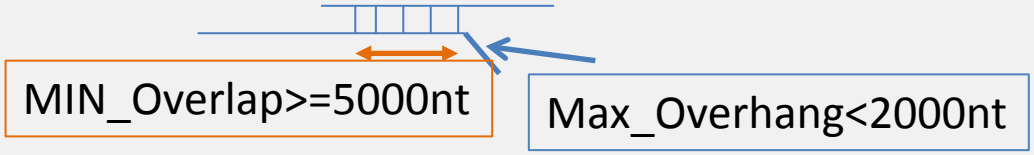
- + Falcon: pipeline développé par PacBio utilisant sur DALIGNER développé par Gene Meyers. Le principe reste le même: correction puis contigage

XRQ (3,6Gb): read correction

Basé sur une version de PBcR modifiée (perl)

- pour accélérer une étape qui sature les IO du filesystem clusterisé (utilisation d'un cache local ➔ 13x plus rapide)
- rajouter des filtres sur les overlaps pour ne pas saturer les 20Tb de quota

CR1 Reads >= 12kb +



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

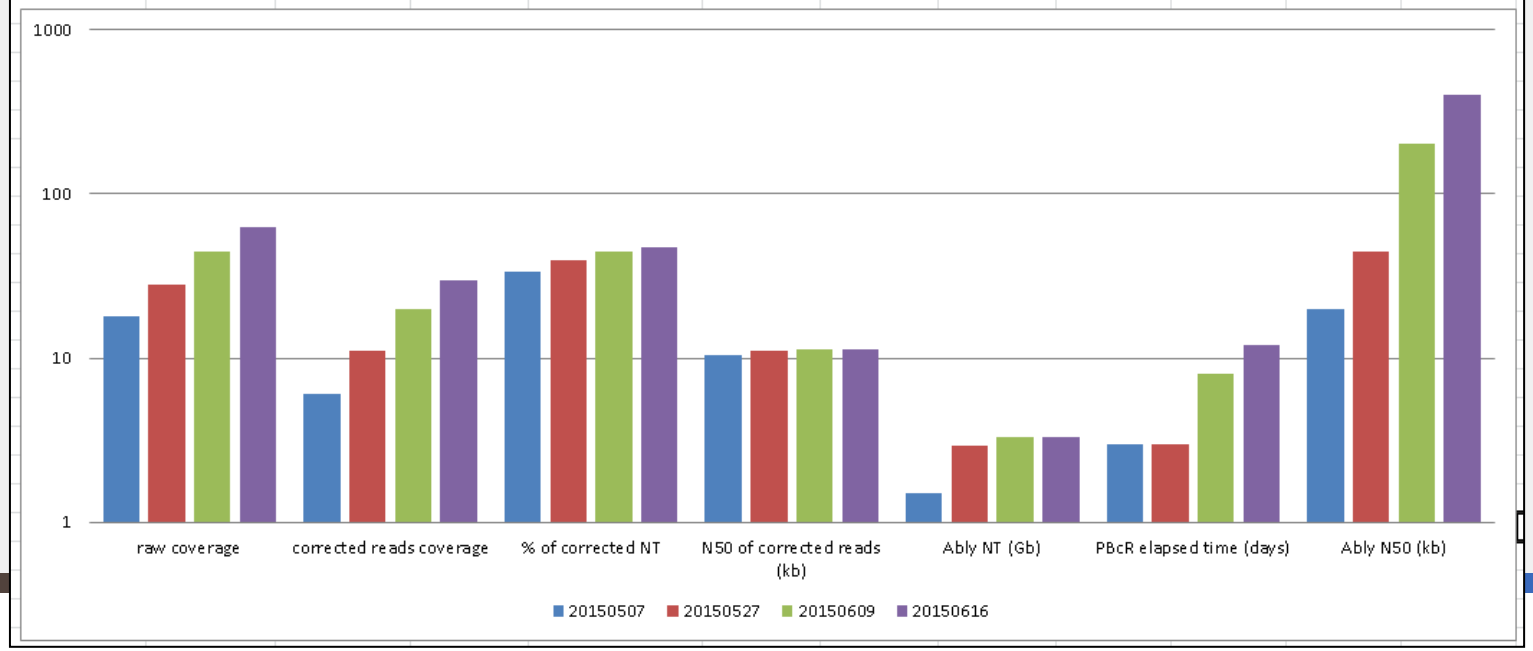
CR2 Reads >= 3kb, MIN_Overlap >= 3000nt, LEN_Overlap >= 50% de la lecture la plus courte

Post -process	#	MAX	N50 BP	MEAN	BP
	19,7M	58kb	11,5kb	9kb	180 Gb (50x)
Trimming + Len>= 9kb	7,2M	58kb	13,2kb	13,3kb	95.6 Gb (26x)

XRQ: Assembly using corrected reads

Preliminary assembly very encouraging and better than Illumina-based genome

	raw coverage	corrected reads coverage	% of corrected NT	N50 of corrected reads (kb)	Ably NT (Gb)	PBcR elapsed time (days)	Ably N50 (kb)			
20150507	18	6	33	10.5	1.5	3	19.9	ParamSet1		
20150527	28	11	39	11.1	2.9	3	44.4	ParamSet1		
20150609	45	20	44	11.4	3.3	8	201	ParamSet2	max(tmp): 2.5Tb	3j cln, 5 ably
20150616	63	29.6	47	11.2	3.3	12	402	ParamSet2	max(tmp): 4.5Tb	7j cln, 5j ably



XRQ: Assembly using corrected reads

Using dataset of corrected reads CR1 (34x, N50=13,6kb)

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
97 028	1.65M	121 kb	8 737	20 kb	4.3 !!

La taille de l'assemblage bien trop grande

L'assemblage devient fragmentaire

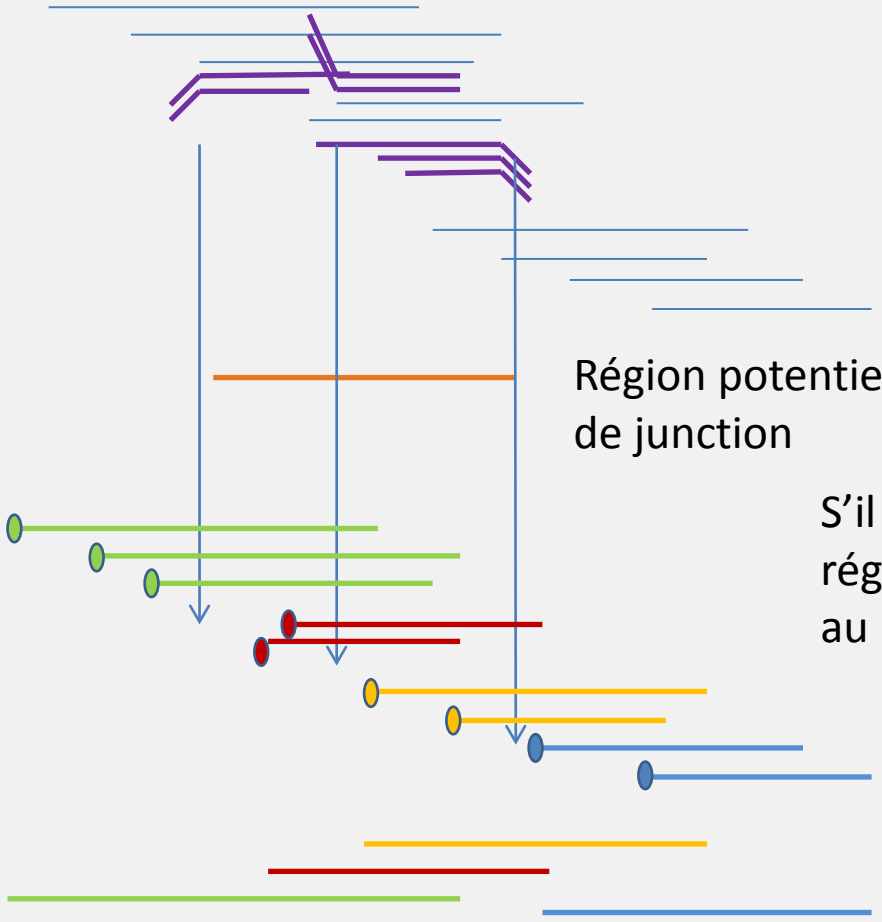
➔ Il y a un problème dans le processus

➔que l'on avait lors des analyses préliminaires mais que l'on ne voyait pas dans les statistiques des assemblages

➔que l'on ne voit pas du tout sur les autres espèces

Focus on the consistance analysis phase of unitigger « Bogart » - default functionning

Les unitigs sont vérifiés en analysant les lectures apparentées qui ne sont pas dans l'unitig et qui sont partiellement alignées



Région potentiellement problématique avec trois points de junction

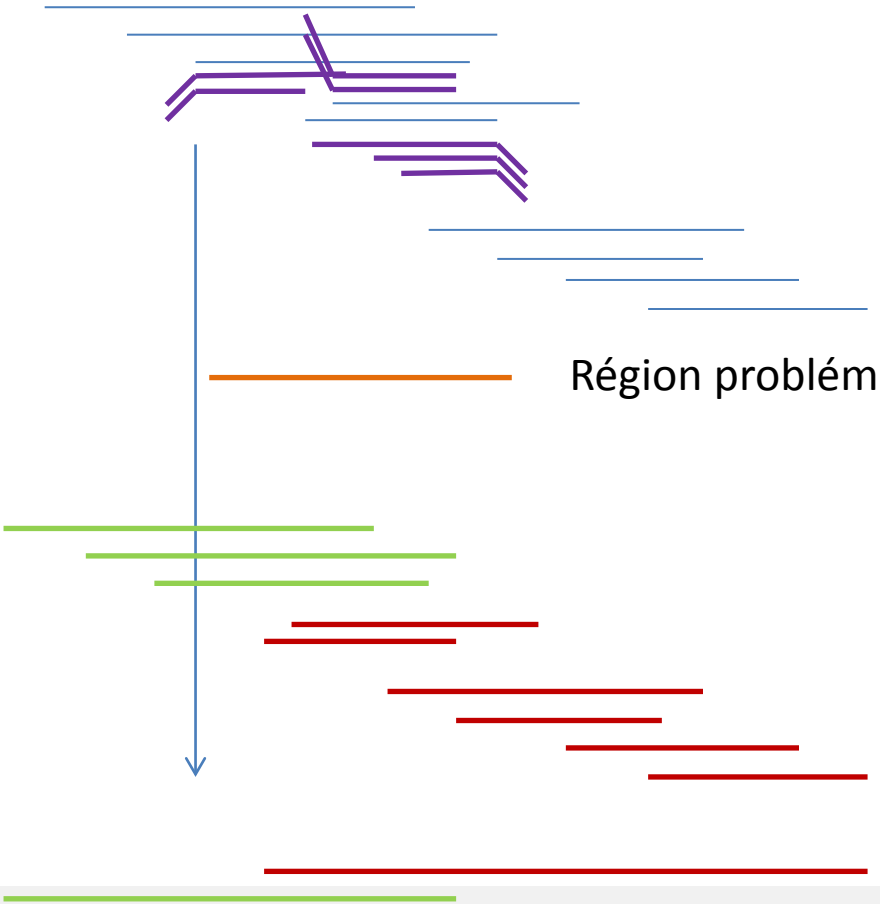
S'il n'y a pas de lecture qui chevauche la région (+/- ½ minoverlap!), l'unitig est coupé au niveau de chaque junction

➔ Génération de 4 unitigs largement chevauchant ➔ expansion artefactuelle de l'assemblage

Focus on WGS/Bogart – Modification of C code

➔ On a un « bruit » du à la quantité de répétitions très conservées dans le génome du tournesol: la « même » unité de répétition est associée a de très nombreux contextes et cela crée un signal d'incohérence ➔ être plus stringent lors de la prise en compte des reads apparentés

➔ on ne coupe qu'une fois pour diminuer l'effet du chevauchement des très longues lectures



Région problématique putative avec un seul point de jonction

➔ Génération de deux unitigs chevauchant mais le chevauchement est bien plus limité que précédemment.

XRQ : Assembly using corrected reads and modified code

Back to expected stats

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

Despite the important read length, « too » repeated regions are collapsed (rDNA, centromeres, telomeres)

➔ ~20% of genome

PSC: Comparison PBcR/WGS/ FALCON

Test on a second génotype de tournesol (52x) - HELIOR

Corrected reads by PBcR

#	MAX	N50 BP	MEAN	BP
7,5M	59kb	13,6kb	9kb	70,1 Gb (19,6x)

PBcR/WGS « Sunflower modified version »

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
26 273	2.5M	223 kb	3 799	66 kb	3.1

FALCON-default

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
35 066	1.0M	101 kb	6 212	38 kb	2.05

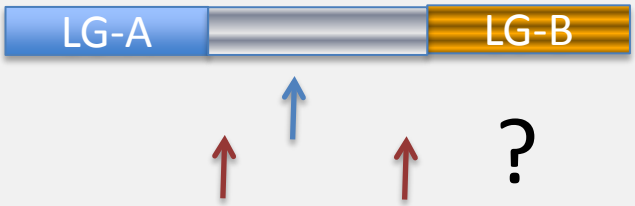
FALCON-deactivation of the control of repetion level at read extremities

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
36 197	1.45M	202 kb	4 832	36 kb	3.2

XRQ: First step of chimeric contig detection

Depuis les premiers essais d'intégration contigs/cartes génétiques sur l'assemblage 65x (ChrisGrassa été 2015) nous savons que nous avons des contigs chimériques

La carte génétique permet d'en détecter certains (mais pas tous) et la prédiction du ou des sites chimériques est difficile



➔Développement d'un logiciel qui analyse rapidement la consistance du DL le long des contigs

- permet d'expliquer 80% des régions détectées comme chimérique par la carte génétique
- génère des sites de coupures faux positifs

➔ On ne coupe pas dans une région: (i) confirmée par la carte génétique (ii) synténique avec PSC8

Sequence based *scaffolding / re-contiguing*

-Trimming of contig extremities using the repeated element library

- *Scaffolding/re-contuing* using various data
  - Comparison PSC8 vs XRQ (evidence from PSC8 of the contiguing of 2 XRQ contigs)
  - Use of 2 alternative assemblies CR2 corrected reads and different parameters (after DL inconsistency correction)

Merging only if gaps between contigs are DL-consistent

➔ 2725 accepted links

➔ 675 rejected

Contigs used for pseudomolecules production (4 nov 2015)

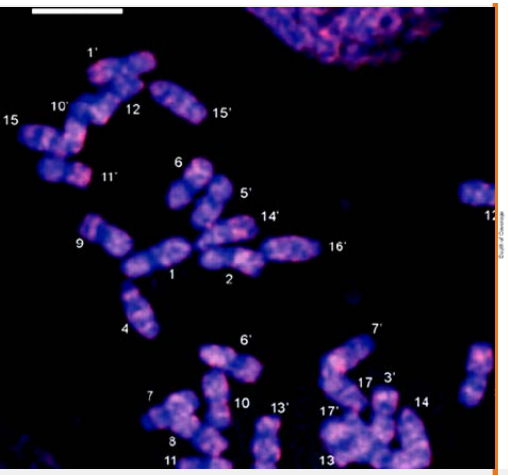
Assembly result

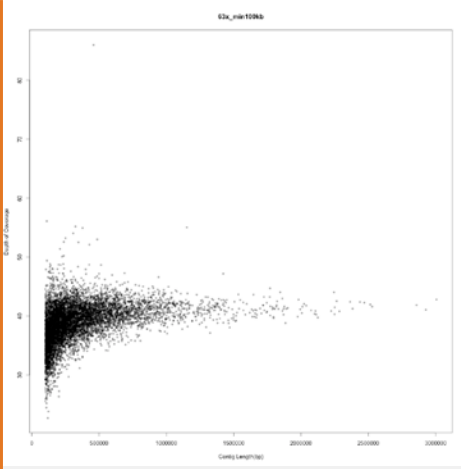
#contigs	LEN MAX	N50 BP	# > N50	MEDIAN	Gb
12 318	3.35 Mb	524 kb	1 684	120 kb	**2.93**

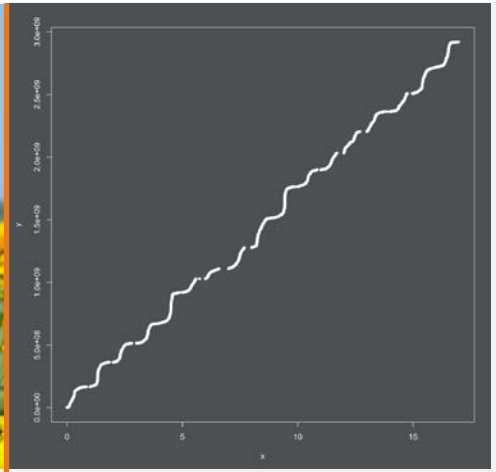
➔80% of genome in the contigs (No Ns)

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

Production of chromosome sequences







Sunflower Genome

- 17 chromosomes
- 3.6G base pairs

Genome assembly

- 12,318 Contigs
- 2.93 Gb

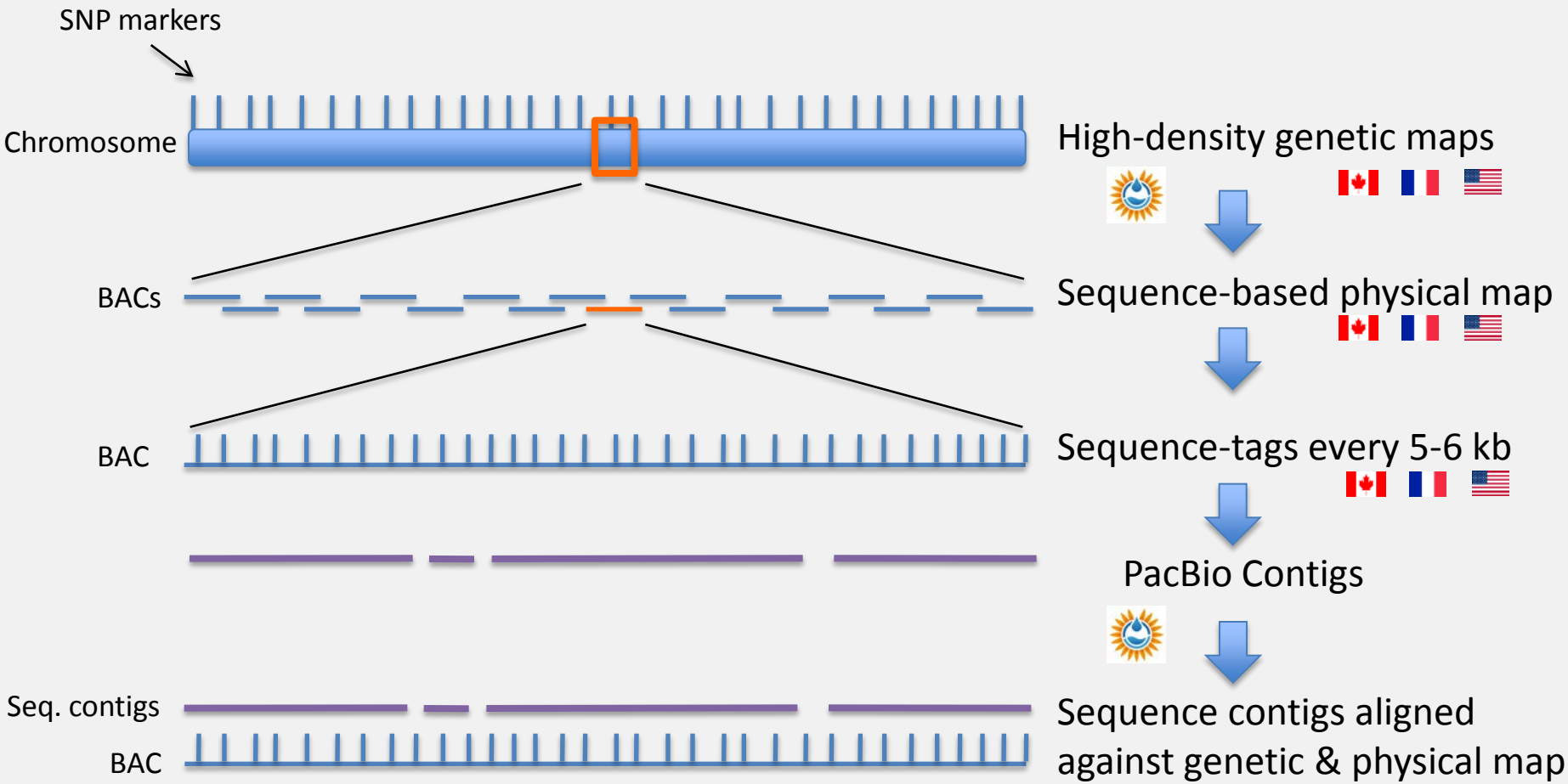
Chris Grassa

DIRECTLY FROM CONTIGS TO PSEUDO-MOLECULES

Reference Genome

- 17 pseudomolecules
- +CP MT

Chromosome production strategy



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

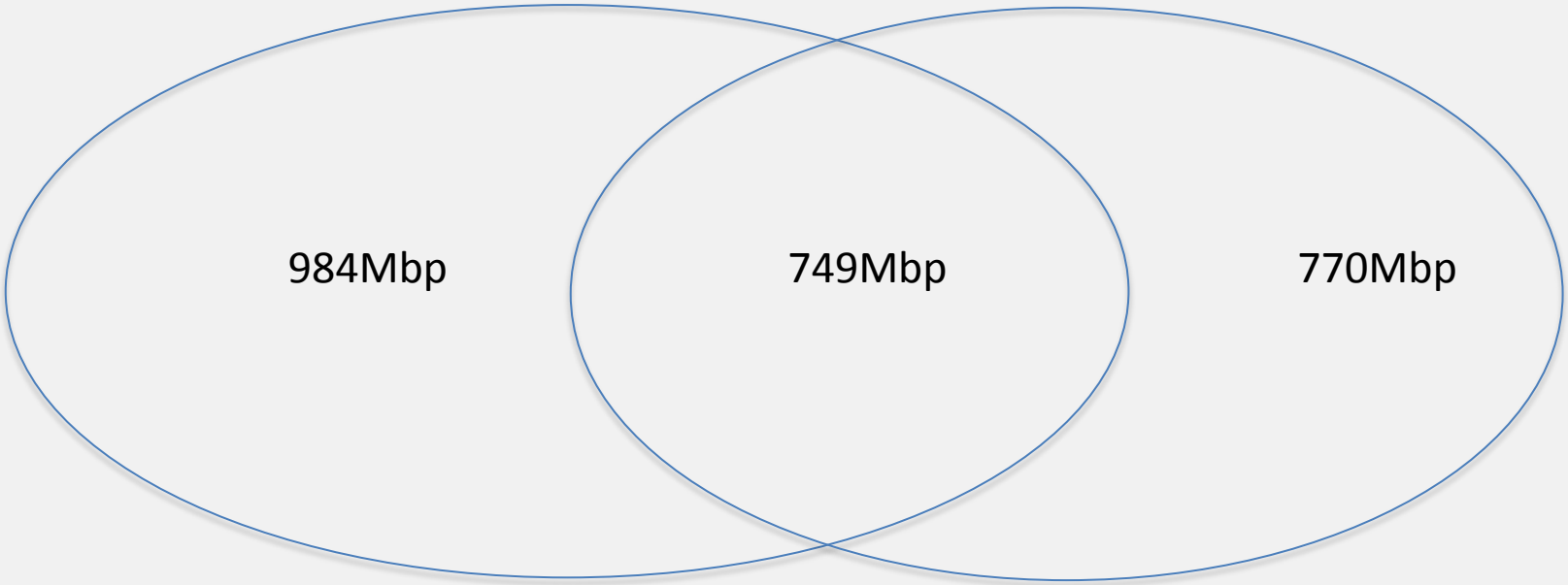
UBC
(Grassa)

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

Complementarity of genetic maps

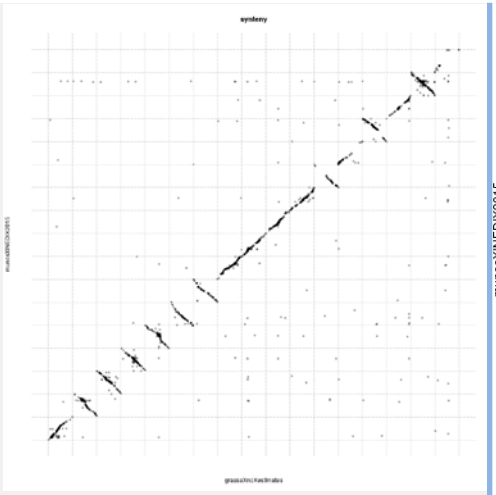
UBC map
#contigs: 12 209
bp placed: 1 733 Mb

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



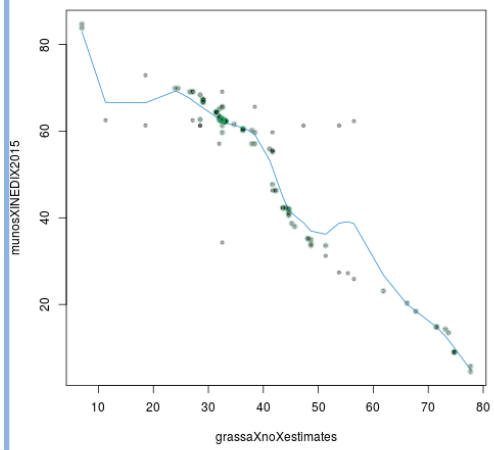
Other maps (INRA 2014, UGA, USDA): 415Mbp

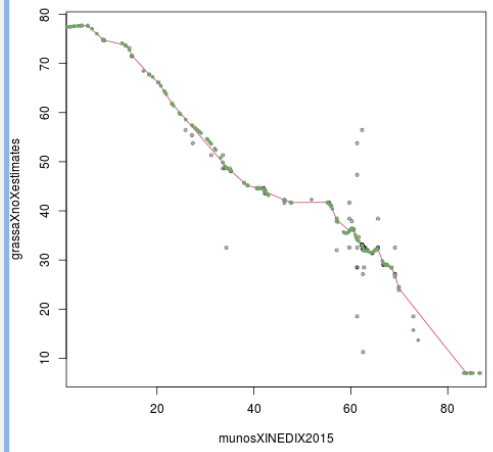
Build Consensus Map Units



Raw Maps

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





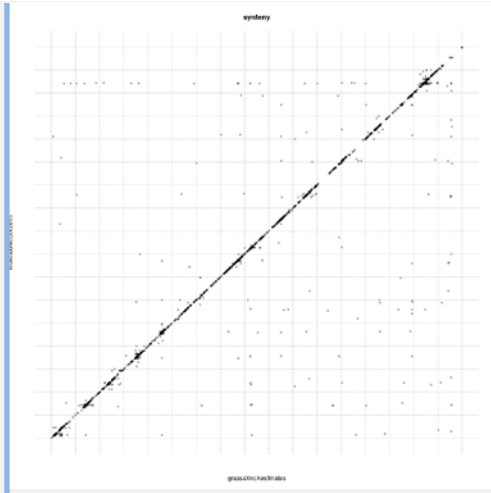
Machine-learn consensus units:

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

2) Drop outliers

3) Train model

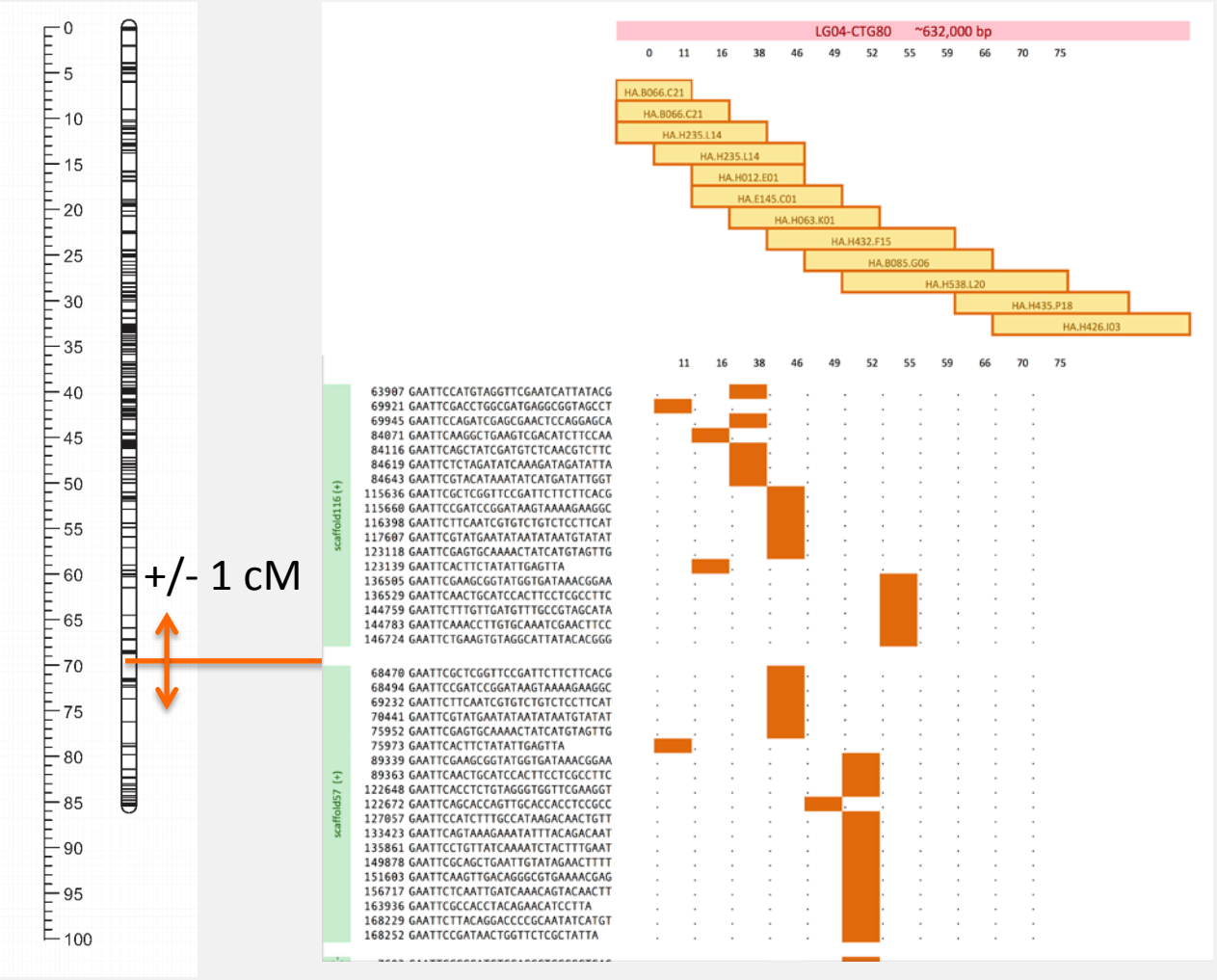
4) Predict positions in consensus units



Consensus Map

- Near colinearity
- **Common map units**

Physical Map Scaffolding



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

Polishing of pseudo-molecules

- **Quiver**

  (correction of contigs based on raw seqeucing data)

3 400 192 insertions

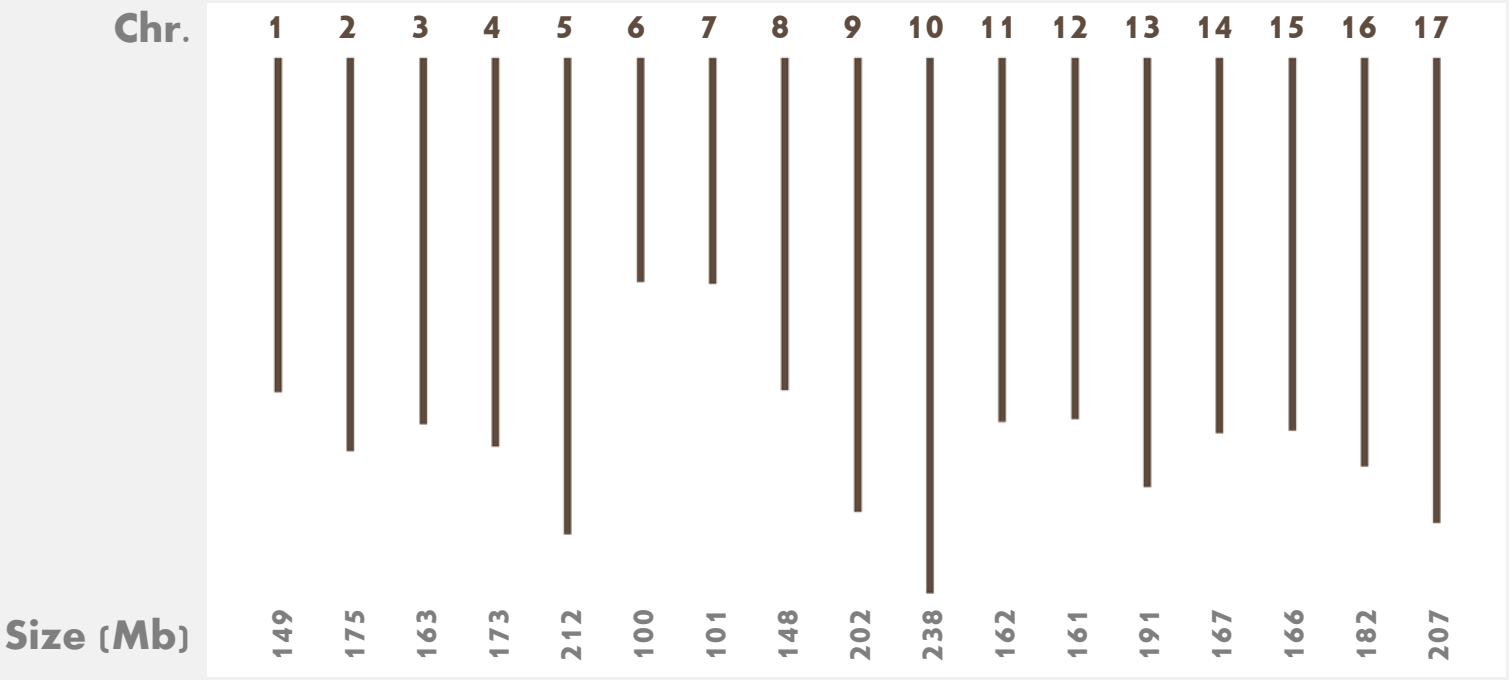
746 419 deletions

2 487 308 substitutions

0.2% corrections

Sunflower reference genome

3 027 Mb (3.38 % de N)



Mitochondrion

301 kb

Chloroplast

151 kb

98.5% of contigs in the chromosomes

Scaffolding result:
98.5% of contigs in the scaffolds

LG	Mb	# contigs	Oriented Mb	# oriented contigs	% Oriented bp	% # oriented contigs
1	149	493	109	240	73%	49%
2	175	597	122	289	70%	48%
3	163	575	114	273	70%	47%
4	173	610	115	295	66%	48%
5	212	725	149	352	70%	49%
6	100	334	59	135	59%	40%
7	101	322	58	112	57%	35%
8	148	525	108	258	73%	49%
9	202	711	144	362	71%	51%
10	238	820	178	429	75%	52%
11	162	605	91	230	56%	38%
12	161	568	113	280	70%	49%
13	191	637	146	340	76%	53%
14	167	680	114	338	68%	50%
15	166	537	117	262	70%	49%
16	182	678	126	303	69%	45%
17	207	745	152	379	73%	51%
CP	301					
MT	151					
SUM	2 894	10 162	2 014	4 877	**70%**	48%

Raw description of the final reference sequence

*** Statistiques Générales** apres suppression des contigs inclus dans des contigs plus gros

NUM 1531
MIN 518
MAX 246315975
N50 BP 178899001
N50 NUM 8
N90 BP 153252196
N90 NUM 15
MEAN 1977768
MEDIAN 15169
BP 3027963057 (3.38 % de N)

* Détail NON chr0

HanXRQChr01 len=153905722
HanXRQChr02 len=180951721
HanXRQChr03 len=168485022
HanXRQChr04 len=178899001
HanXRQChr05 len=219053782
HanXRQChr06 len=103887801
HanXRQChr07 len=103871911
HanXRQChr08 len=153252196
HanXRQChr09 len=209865144
HanXRQChr10 len=246315975
HanXRQChr11 len=168460249
HanXRQChr12 len=166489593
HanXRQChr13 len=197258317
HanXRQChr14 len=174509413
HanXRQChr15 len=171247636
HanXRQChr16 len=188620566
HanXRQChr17 len=214723238
HanXRQMT len=301004
HanXRQCP len=151101

NUM 19=17+CP+MT
MIN 151101
MAX 246315975
N50 BP 178899001
N50 NUM 8
N90 BP 153252196
N90 NUM 15
MEAN 157907862
MEDIAN 171247636
BP 3000249392 (3.41% de N)

Raw description of the final reference sequence

* General statistics

NUM	1 531
MIN	518
MAX	246 315 975
N50 BP	178 899 001
N50 NUM	8
N90 BP	153 252 196
N90	15
MEAN	1 977 768
MEDIAN	15 169
BP	3 027 963 057 (3.38 % of Ns)

* Pseudomolecules statistics

NUM	19=17+CP+MT
MIN	151 101
MAX	246 315 975
N50 BP	178 899 001
N50 NUM	8
N90 BP	153 252 196
N90 NUM	15
MEAN	157 907 862
MEDIAN	171 247 636
BP	3 000 249 392 (3.41% of Ns)

Raw description of the final reference sequence

* General statistics

NUM	1 531
MIN	518
MAX	246 315 975
N50 BP	178 899 001
N50 NUM	8
N90 BP	153 252 196
N90	15
MEAN	1 977 768
MEDIAN	15 169
BP	3 027 963 057 (3.38 % of Ns)

* Chr0 statistics: 1512 scaffolds (mainly contigs)

NUM	1 512
MIN	518
MAX	152 385
N50 BP	24 794
N50 NUM	324
N90 BP	10 127
N90 NUM	974
MEAN	18 329
MEDIAN	15 018
BP	27 713 665 (0.01% de N)

Gene content and genome annotation

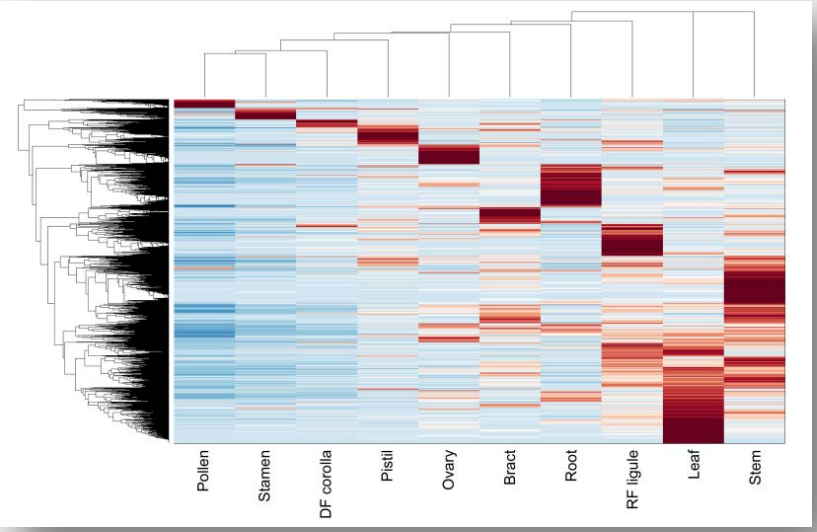
98% of transcripts mapped on pseudo-molecules

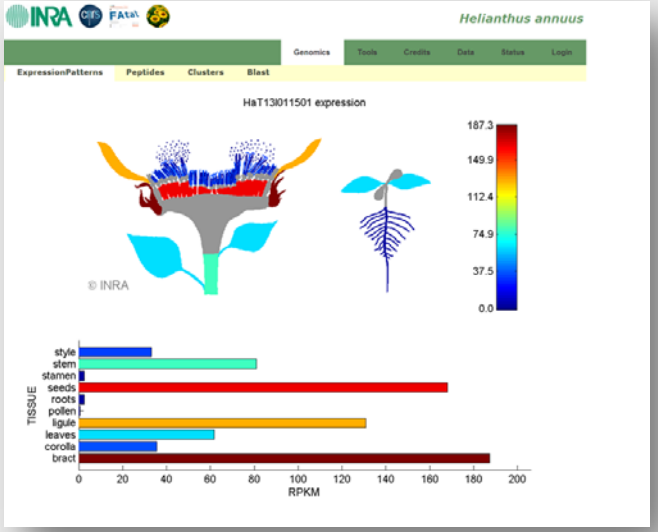
39 RNA-Seq librairies on the sequenced genotype (XRQ)

Organ-specific expression (10 organs)

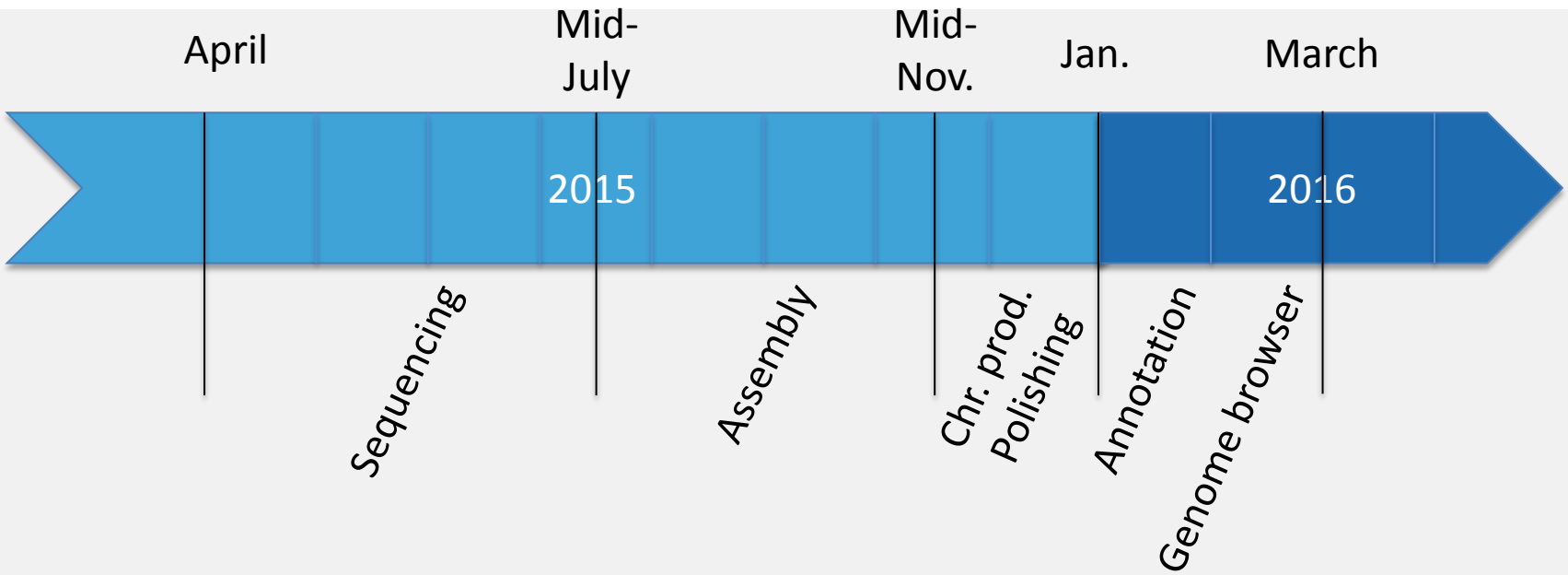
Abiotic stress response: drought, osmotic stress, salt stress

Hormone regulation: (9 hormones in roots and leaves)





TIMELINE



March 2016

- Release to International Sunflower Consortium
- Evaluation of overall quality

In 2016

- Publication and public release (early access on demand)

Post-Doctoral position available

Contact:

stephane.munos@toulouse.inra.fr

SUNRISE

UNE CULTURE POUR LE FUTUR

Thank you for your attention

www.sunrise-project.fr

@SUNRISE_France

Funding

Partnership