

Toward a reference sequence of the gene-rich part of the sugarcane genome

Olivier GARSMEUR
CIRAD, Montpellier, France



cirad

Modern sugarcane cultivars

S. robustum
2n = 60,80-->200

Domestication

S. officinarum
2n = 10x=80



S. spontaneum
2n = 40 --> 128



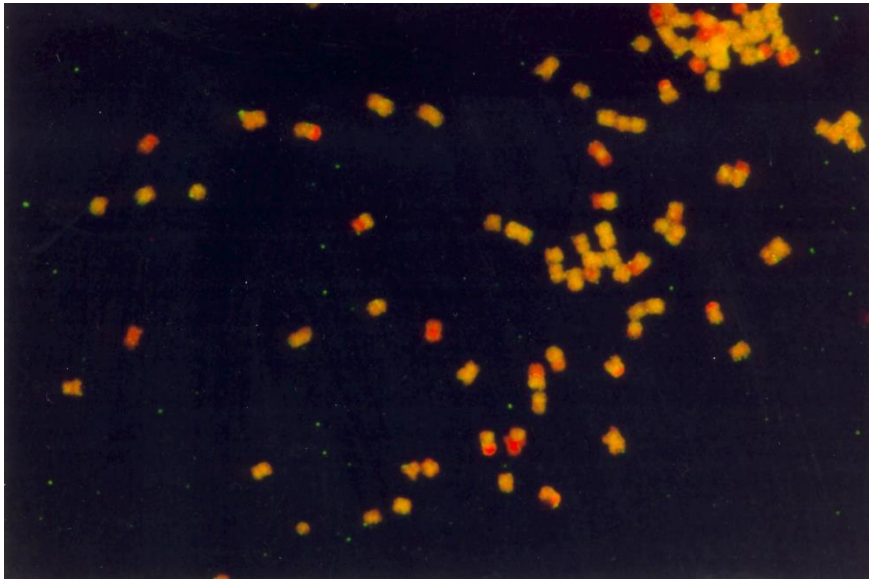
modern cultivars

2n ~ 110-130

Highly polyploid
Aneuploid
Interspecific

Global organization of the complex genome of sugarcane

Genomic in situ hybridization

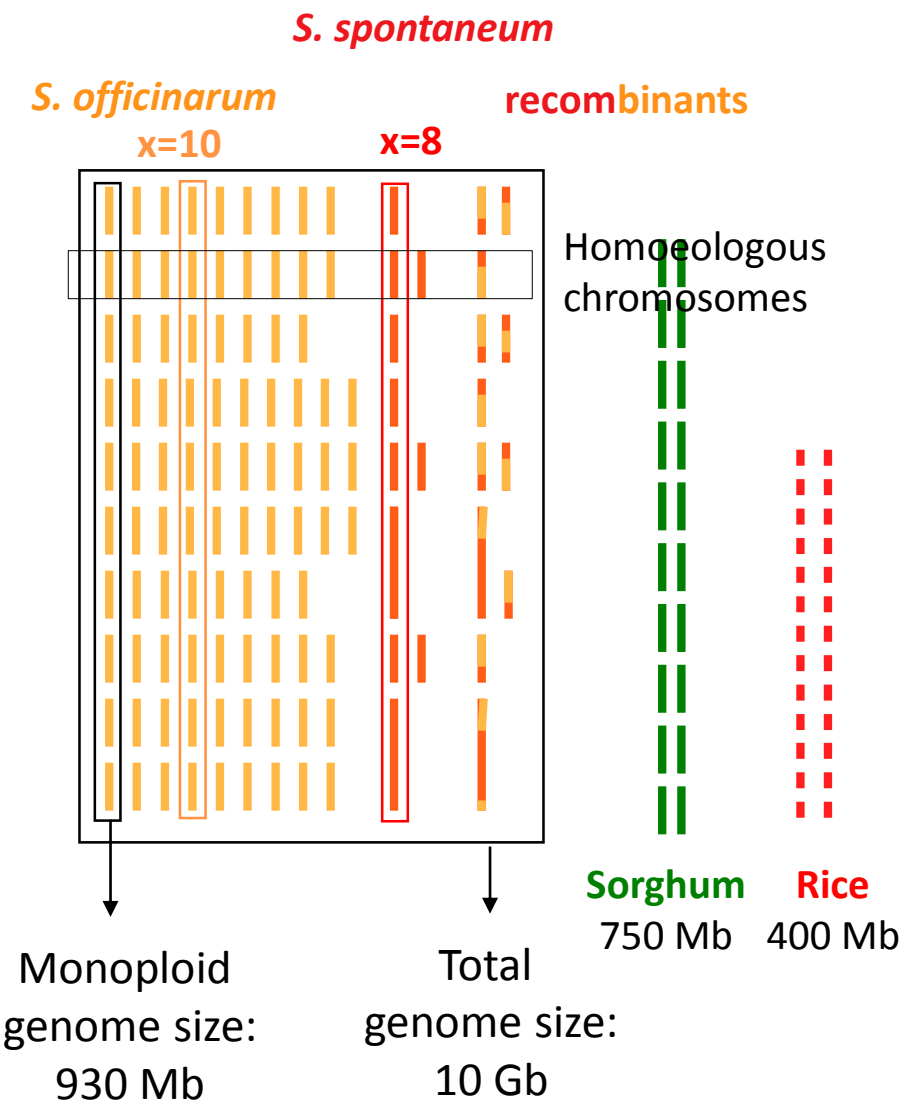


R570, 2n=ca 115

80 % S. officinarum

10% S. spontaneum

10% recombinants



(D'Hont et al, 1996; D'Hont et al, 1998; D'Hont 2005, Piperidis et al 2010)

Fine structure and evolution of the sugarcane genome

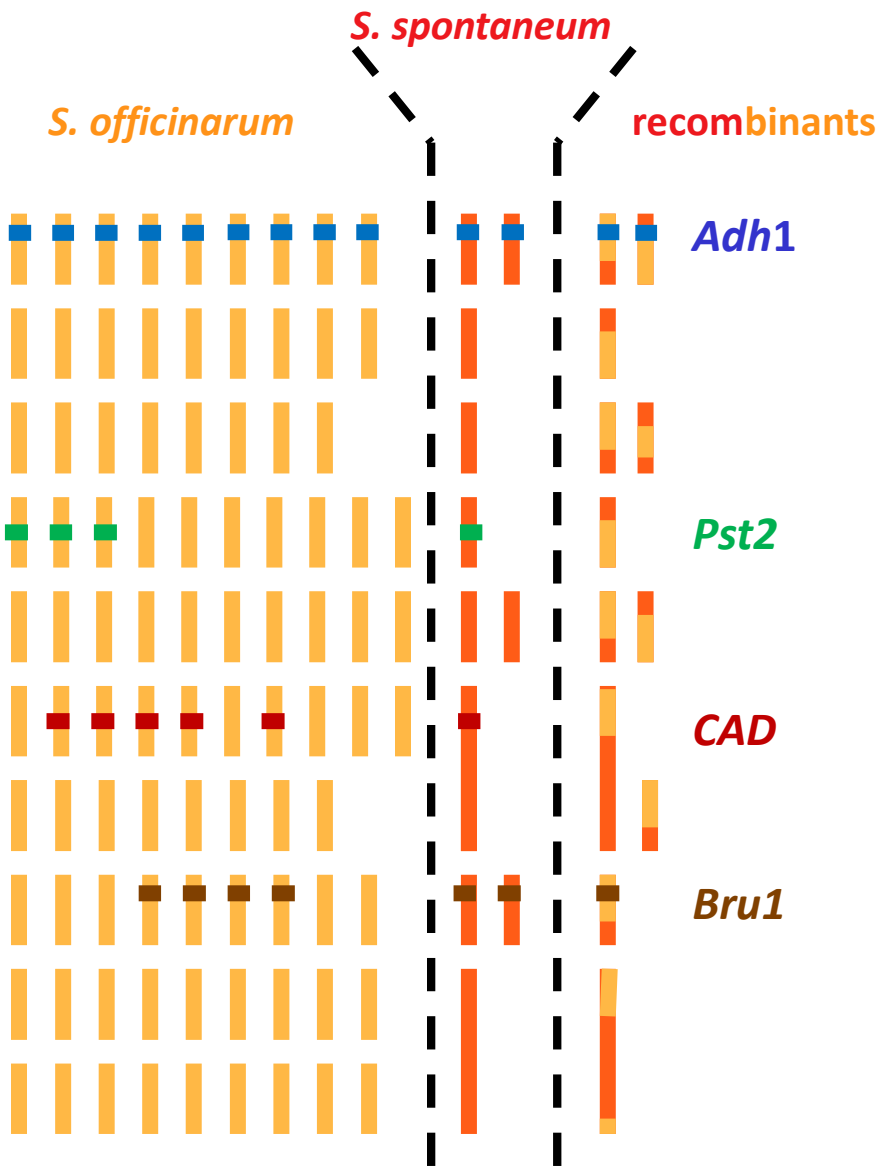
→ **Comparison of sugarcane homoeologous regions**

Thirteen hom(oe)ologous BAC clones corresponding to the ***Adh1*** region

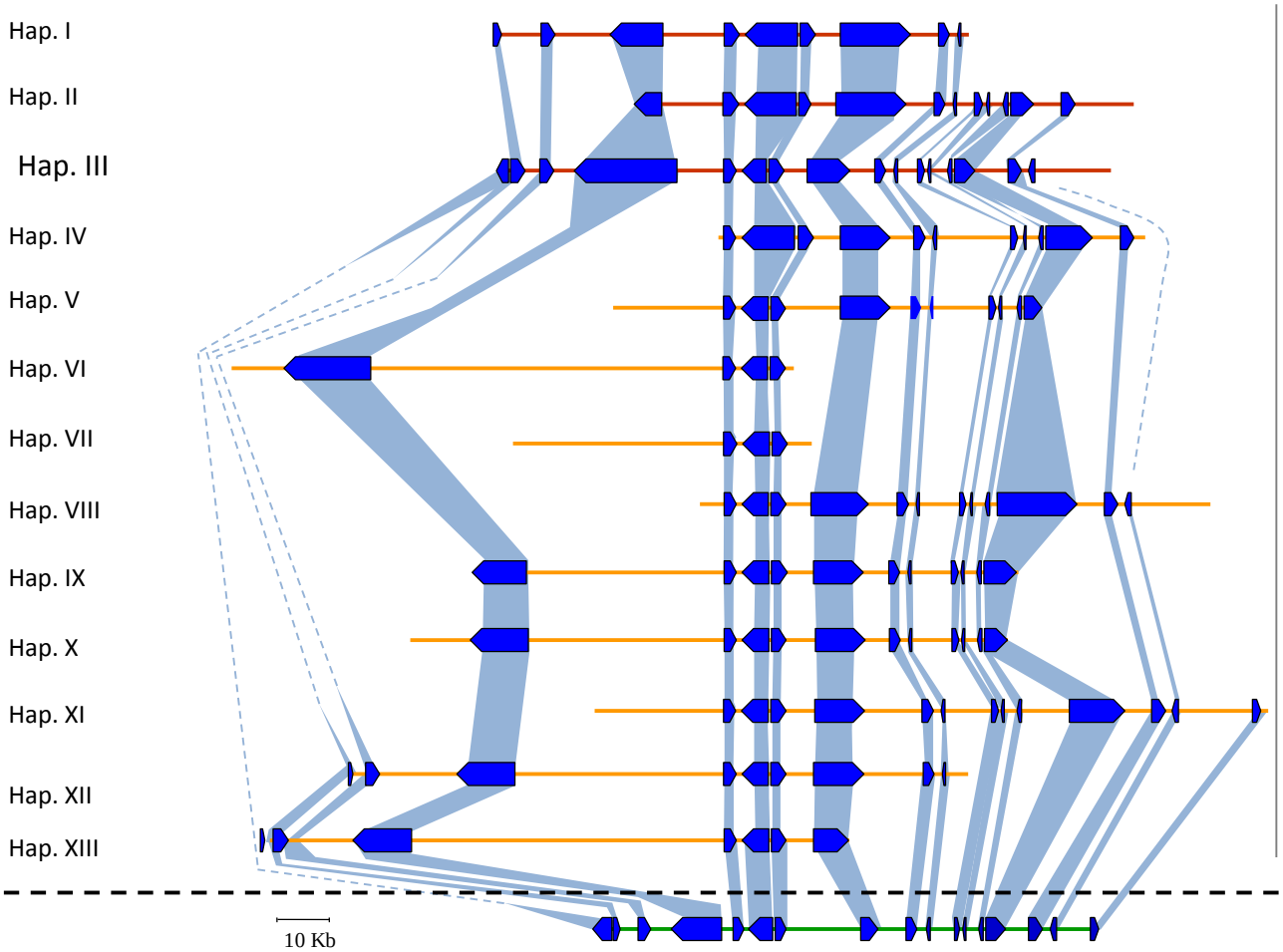
Four hom(oe)ologous BAC clones corresponding to the ***Pst2*** region

Six hom(oe)ologous BAC clones corresponding to the ***CAD*** region

Seven hom(oe)ologous BAC clones corresponding to the ***Bru1*** region



Analysis of sugarcane homeologs and comparison with sorghum



Sugarcane homoeologous BAC clones

Sorghum Orthologous region

→ Gene structure conservation among sugarcane homeologs
→ All hom(oe)o-alleles predicted functional
→ High colinearity with sorghum

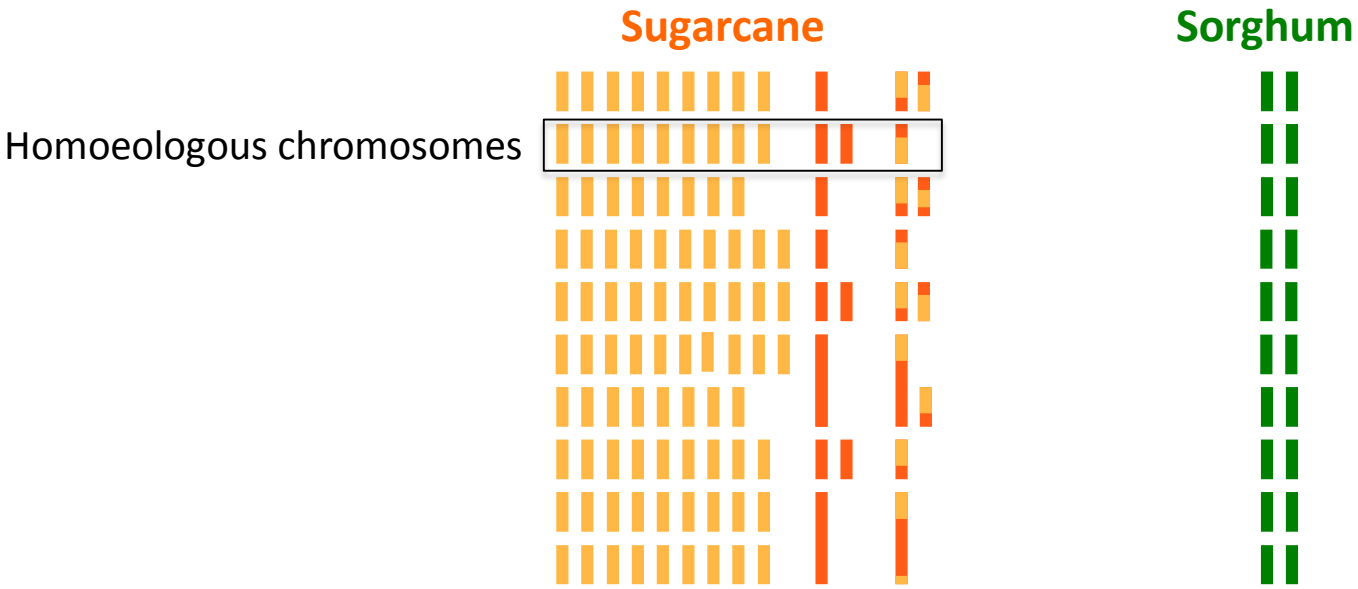
Garsmeur *et al*. New Phytologist 2010
Charron *et al.* submitted

Sequencing sugarcane genome is very challenging

The sugarcane genome poses challenges that have not been addressed in any prior sequencing project.

The main difficulties reside in

- its **high polyploidy**, with a complete set of homoeologous chromosomes predicted to range from **10 to 12 copies** and its **high level of heterozygosity**



Assembly of a whole genome sequence very challenging through a classical shotgun sequencing approach

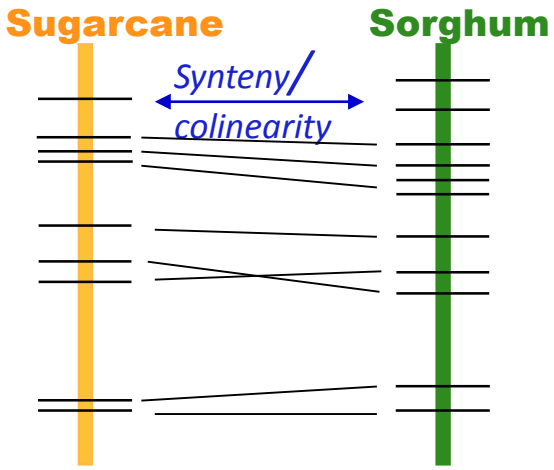
→ Alternative approach

Sugarcane sequencing strategy based on synteny with sorghum

- Genes are conserved among sugarcane hom(oe)ologous chromosomes

→ if we could sequence a set of BACs representing **one monoploid genome**, it would represents a **very useful reference sequence**

- High colinearity between sugarcane and sorghum



- Grivet et al 94,
- Dufour et al, 1996, 1997
- Guimaraes et al 1997
- Ming et al 1998,
-Asnaghi et al, 2001
-Jannoo et 2007
-Garsmeur et al 2011

→We should be able to use sorghum to identify a core set of sugarcane BACs representing **one monoploid genome**

→Mainly the genes are conserved → we focus on the gene-rich part of the genome

BAC selection through Whole Genome Profiling (WGP) Technology

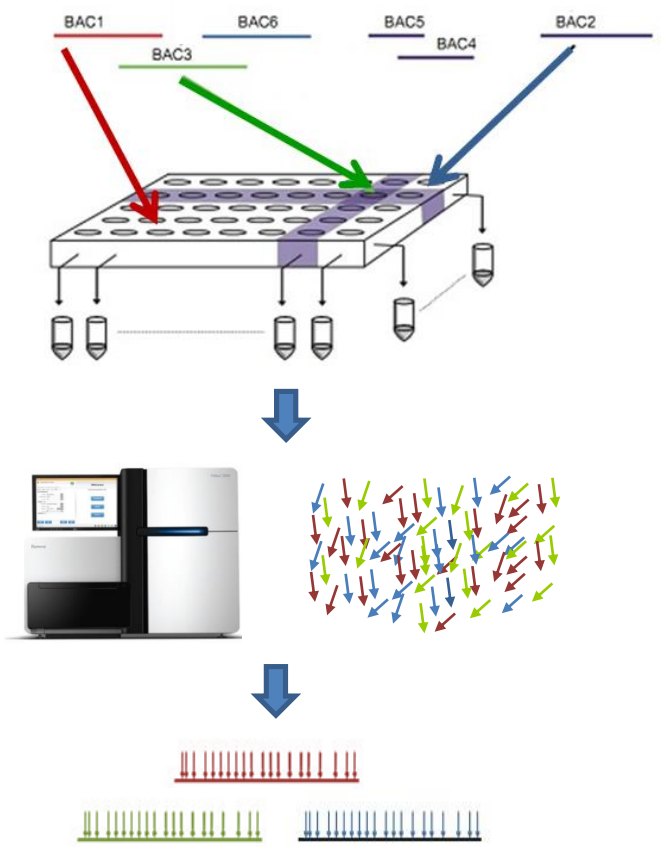
WGP technology generates short sequence tags from the terminal ends of restriction fragments from pooled BACs

(A)

(B)

(C)

(D)



(A) 20,736 BAC clones from R570 Sugarcane BAC library
providing ~ 2x coverage of the monoploid genome

(B) BAC pooling, DNA extraction and restriction
BAC DNA is extracted after pooling each row and each column (2D pools). Pools of BAC are restricted using EcoRI/MseI enzymes

(C) Template preparation and sequencing
Restricted BAC pools are amplified after ligation of barcoded adaptors for pool identification. Pooled DNA is sequenced using Illumina Genome Analyser II. The sequencing produce 50bp sequence tags.

(D) Decovolution
Using barcoded adaptors, tags are assigned to individual BAC. Around 30 to 50 tags per BAC are produced.

→ Anchoring of the produced BAC sequence tags onto the sorghum genome

Sugarcane BAC WGP-tags production and mapping on sorghum

- Sequence tags production

→ 701,066 Tags (50 bp) were produced from the **20,736 BACs**
→ **455,656 uniq tags** (some identical tag sequences correspond to different BACs)
→ represent an average of **37 tags / BACs**

- Mapping of tags sequences onto the sorghum genome

→ 222,745 tags = ~ 50% do NOT aligned onto sorghum genome
→ 141,958 tags = ~ 30% aligned more than one time
→ 90,953 tags = ~ 20% aligned exactly 1 time

- Only tags aligning exactly one time are first used to anchor BAC onto the sorghum. A BAC is anchored onto the sorghum genome if **>3 tags / BAC map in a window of 300 kb on the sorghum genome**

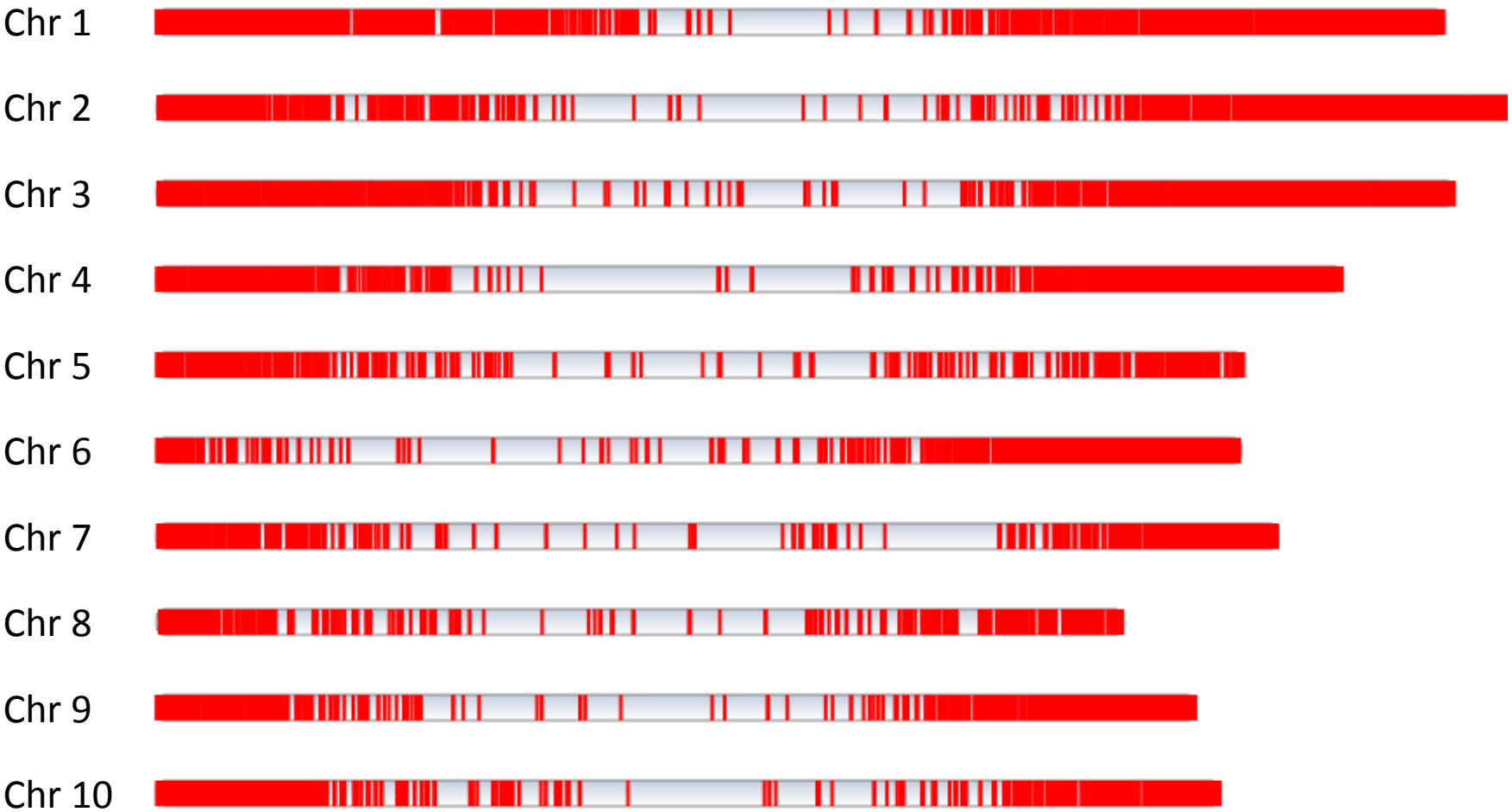
- Tags that mapped to multi-site are re-anchored in accordance to the location of unique tags. This step allows the extension of the coverage of the sorghum target sites

→ 113,257 tags corresponding to 11,732 BACs were mapped on the sorghum genome

Distribution of the sugarcane BAC in the sorghum genome

Sorghum

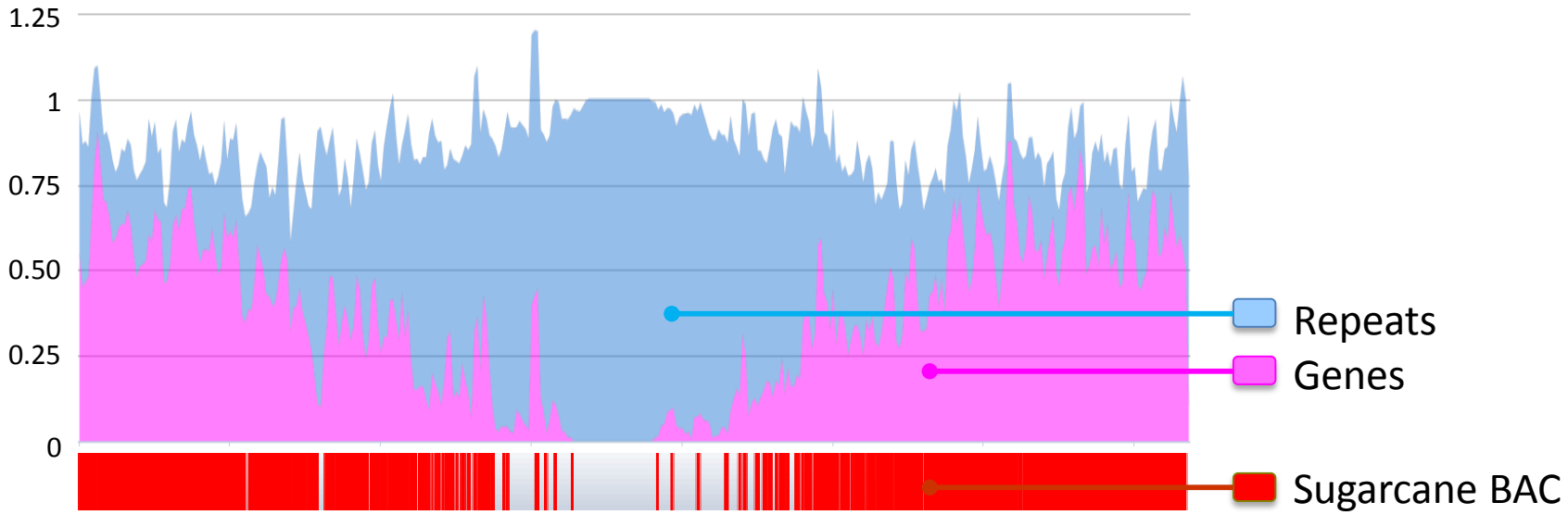
Red bars represent BAC positions on the sorghum chromosomes



→11,732 R570 sugarcane BACs anchored onto the sorghum
→Mostly distributed in distal chromosomal regions

Sugarcane BACs are anchored in sorghum gene-rich regions

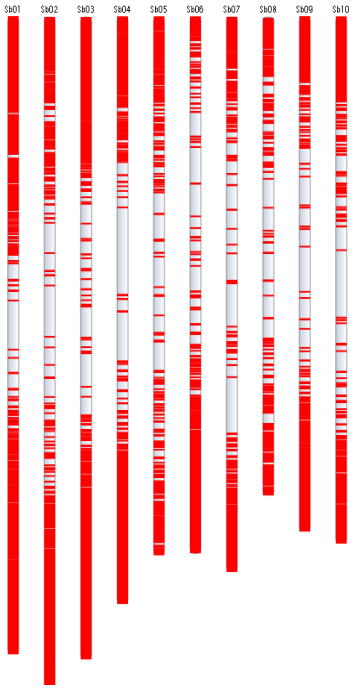
Example: Sorghum Chromosome 1



→BAC mostly distributed in gene-rich regions

Minimum tiling path (MTP) of sugarcane BACs

MTP = minimum set of BACs to be sequenced to obtain the best coverage of chromosomes.

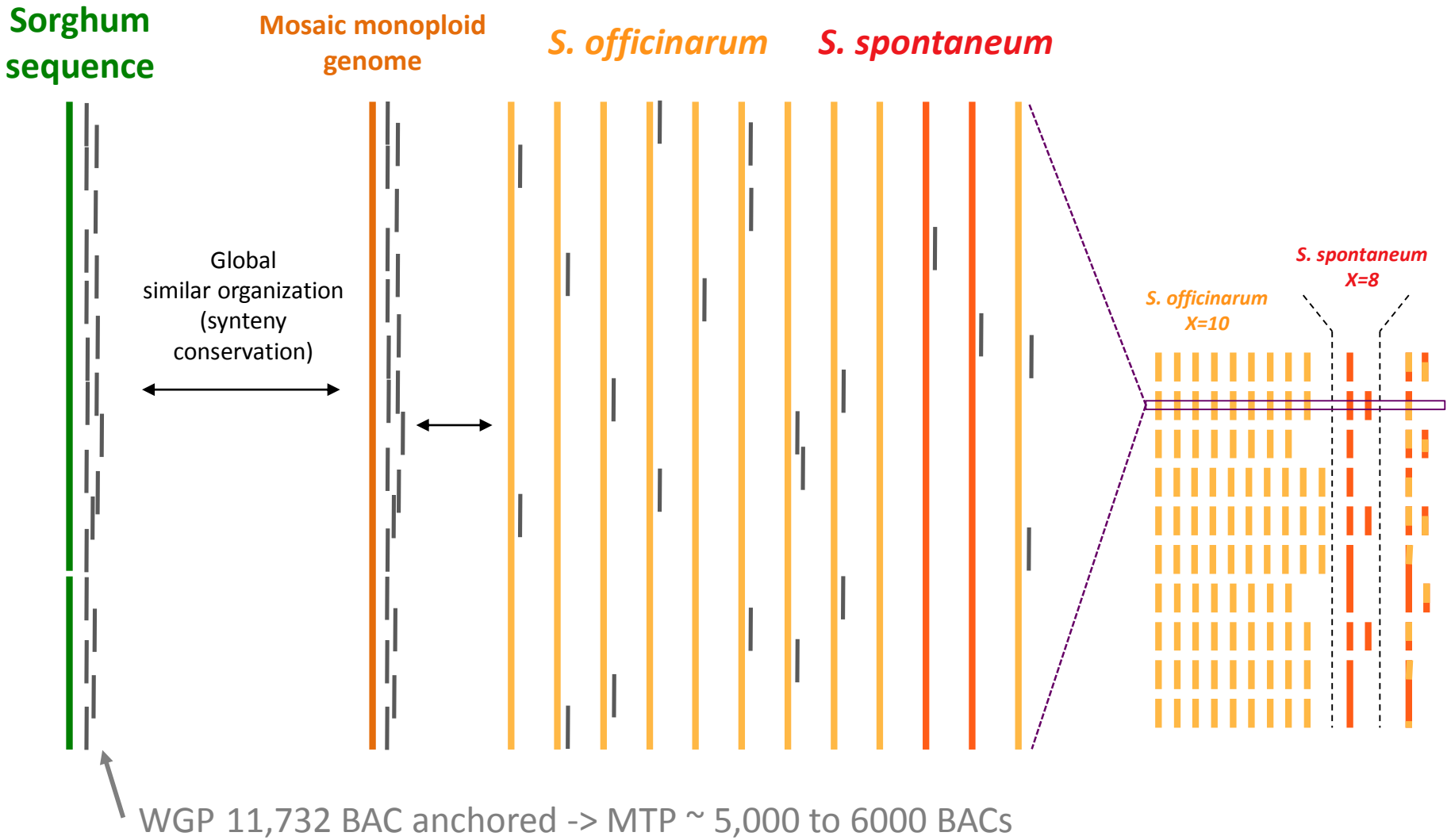


sorghum	number of genes on the sorghum chromosome	numbers of genes covered by sugarcane BACs	% of genes covered by sugarcane BACs	optimized subset of BACs to sequence (MTP)
chr1	6518	4726	73	838
chr2	5170	3480	67	655
chr3	5335	3760	70	695
chr4	4287	2921	68	539
chr5	2722	1564	57	332
chr6	3356	2257	67	446
chr7	2727	1771	65	371
chr8	2315	1352	58	306
chr9	3142	2109	67	441
chr10	3376	2065	61	406
	38948	26005	67	**5029**

The MTP includes around 800 BAC sequences that were already sequenced by SASRI, CSIRO, CIRAD, USP and QAAFI

→ **MTP 5,000 to 6,000 BACs**
→ **The MTP covers <u>at least</u> 26,005 sorghum genes (70%)**

Sugarcane BAC sequencing strategy



→ The MTP of R570 BAC represent a "mosaic" of the gene-rich part of one monoploid genome of sugarcane

High quality results obtained with PacBio BAC sequencing

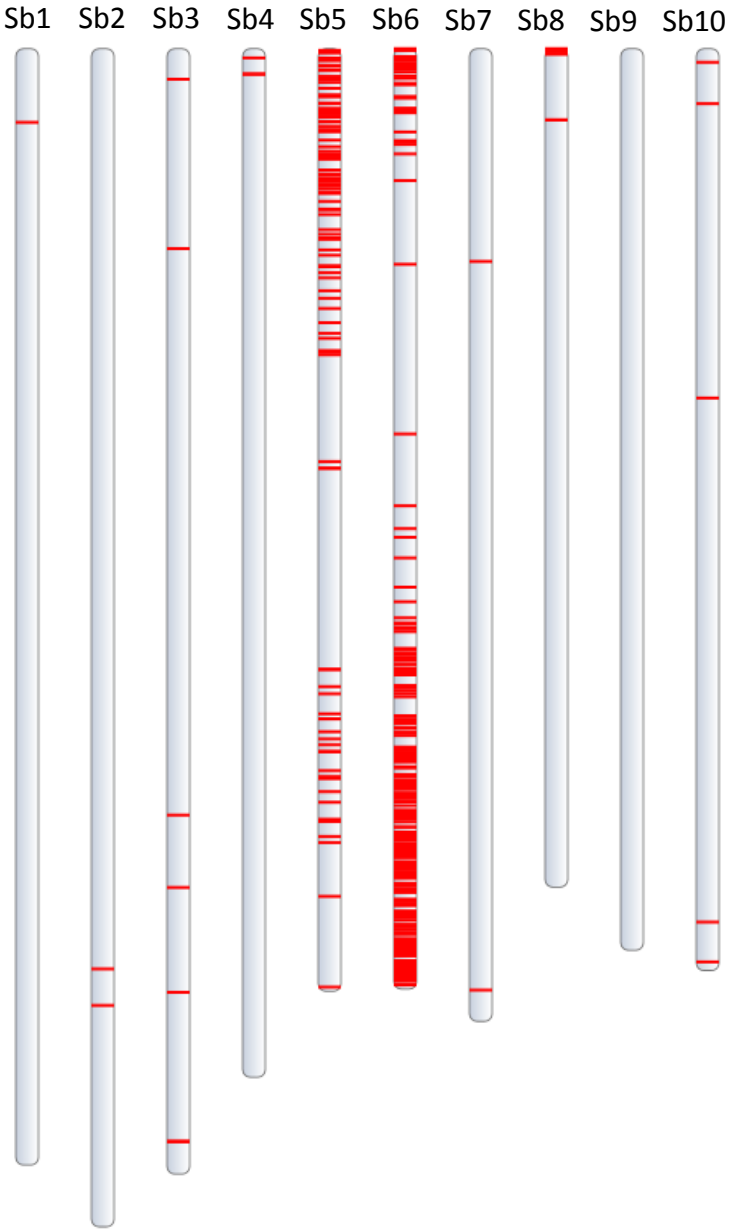
Pilot sequencing project to test the efficiency of BAC quality/assembly and their selection through WGP

456 BACs anchored on sorghum chr6 and part of chr5 were sequenced using PacBio RS II technology

PacBio sequence reads (pool of 24 BACs) were assembled and linked to BAC using WGP-tags.

BAC sent for sequencing	456		
BAC sequenced	452	**98%**	
Mean coverage of contigs	**75 X**		
			Mean size of ctgs (Kb)
BAC assembled in 1 contig	387	**85%**	113.3
BAC assembled in 2 contigs	31	7%	64.8
BAC assembled in 3 contigs	5	1%	42.3

Validation of the BAC selection through Whole Genome Profiling



448 sequenced sugarcane BACs from chr6 and part of 5 initially anchored through WGP were re-anchored onto the sorghum

→ The pilot project allowed to confirm that

- BAC anchoring through WGP was efficient
- PacBio sequencing with 75X coverage resulted in high quality BAC sequences mostly assembled in single contigs

The sequencing was continued with BAC anchored on sorghum chr 2, 3, 4 and end of 5 (CIRAD, SASRI, CSIRO, ICSB)
To date around 2500 BAC have been sequenced

Testing method to anchor BAC sequences to sugarcane chromosomes

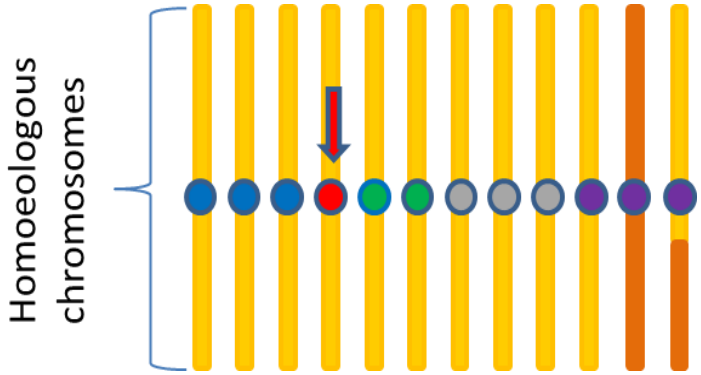
BAC were selected based on synteny conservation with sorghum.
→ essential to anchor them on sugarcane chromosomes

Anchoring BAC to chromosomes via a high density genetic map

Main difficulties in genetic mapping of sugarcane :

- Only simplex markers (present only on 1 out of the 10-12 homoeologous chromosomes) can be used genetic mapping

- Simplex markers are rare and unevenly distributed along the genome



→ At least 100X coverage is necessary to distinguish single dose SNP from sequencing errors

→ The development of high density genetic map requires the genotyping of very large number of loci

Testing Genotyping by sequencing (GBS) on sugarcane

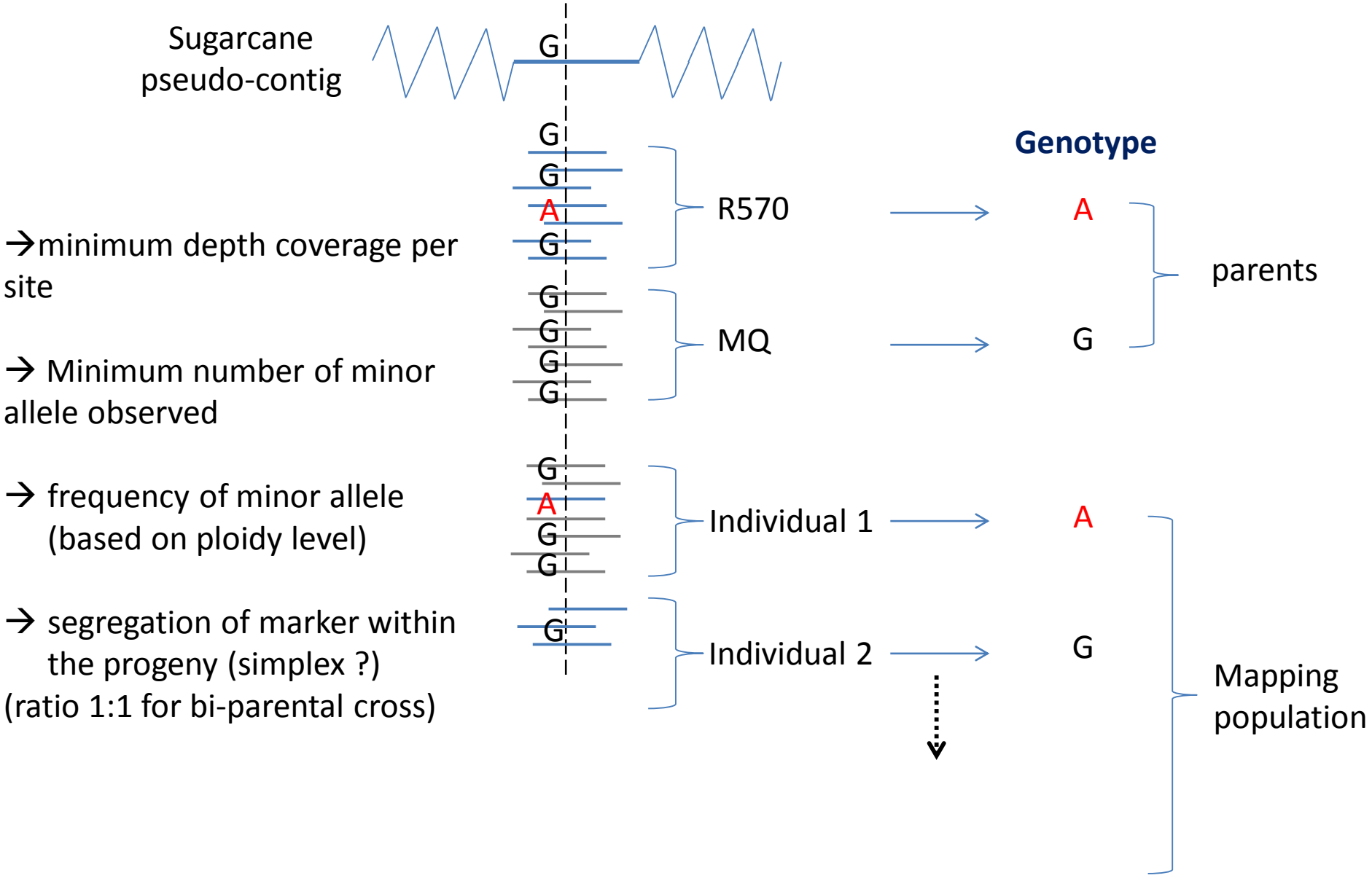
- GBS on a bi-parental cross (R570 x MQ76-53) and mapping population of 94 individuals

DNA	independent samples	illumina reads (100 bp)	coverage / SNP position
R570	8	21,000,000	~100 X
MQ	8	21,000,000	~100 X
population	2	400,000,000 ~4,000,000 / individual	~ 20 X

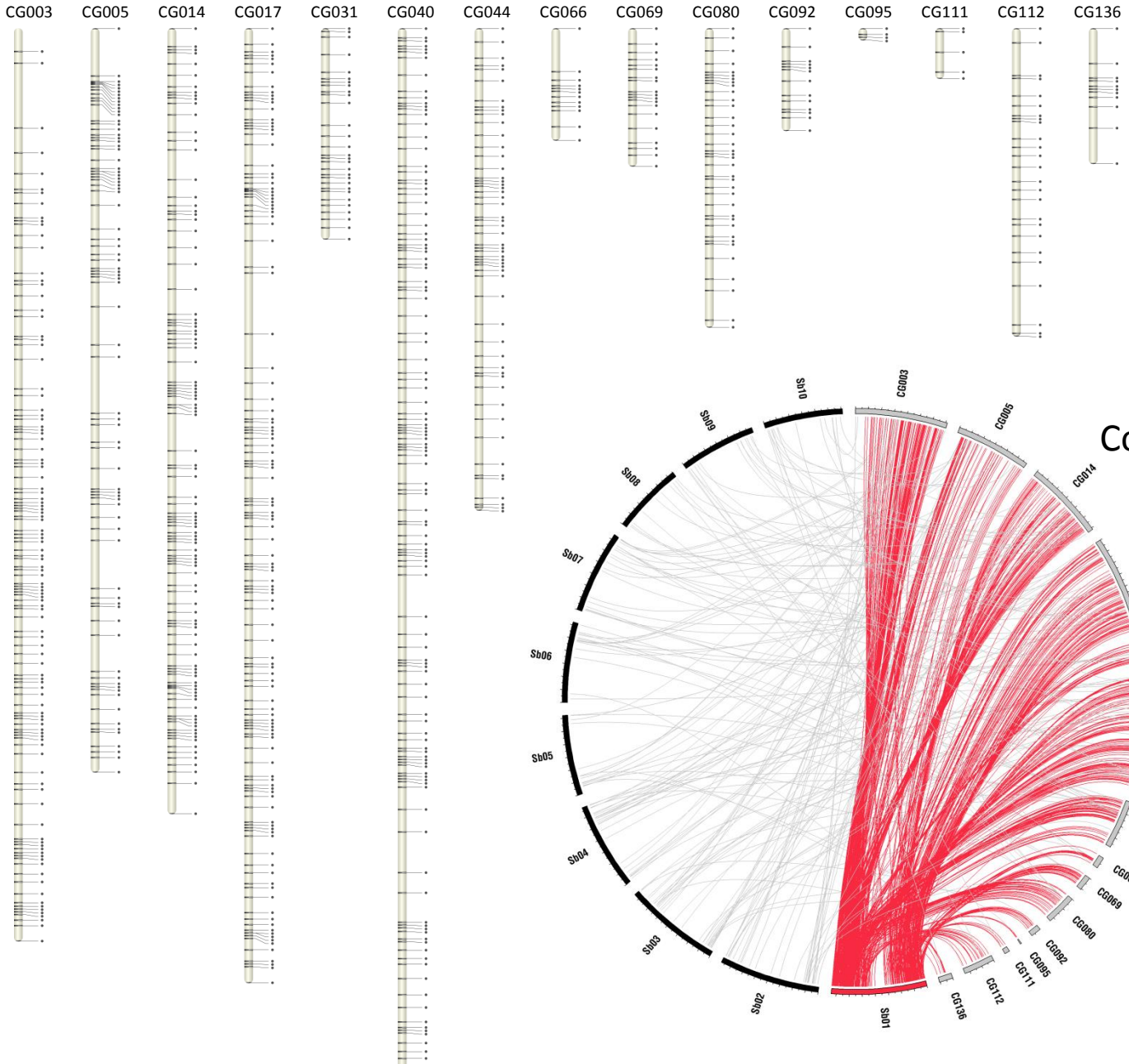
→ Independent sampling of DNA was performed before library construction to reduce potential read depth biased introduced by PCR

Bioinformatics tools for SNP discovery are not adapted for polyploids

We developed bioinformatics tools to identify "good simplex SNP markers" adapted to polyploid context



→ A R570 Genetic map comprising ~ 7,000 SNP has been built



Example:
R570 SNP map
HG I

Comparison with sorghum
→ syntenic to chr1

Conclusions

→ **WGP was efficent to select a minimum tiling path of BAC to be sequenced representing a "mosaic" of one monoploid (basic) genome of sugarcane**

→ **75X PACBio sequencing resulted in high quality BAC sequences and 85% were assembled in single contigs (~110 000 bp)**

→ **We have sequenced 2,500 BACs from the MTP corresponding to 5 out of the 10 basic chromosomes (half of BAC sequencing task is now achieved)**

→ **We tested "high throughput genotyping by sequencing " on sugarcane and developed bioinformatics tools to identify simplex SNP markers**

→ **We built a R570 genetic map comprising more than 7000 SNP markers**

→ **A web portal is currently being developed to make sequences and analyses available through a exploitable form to the community (similar to the existing "banana genome hub" that we have developed for the banana : banana-genome.cirad.fr**


Sugarcane Genome Hub
A genomic resource for sugarcane
Home
BAC sequencing
View
Edit
Devel
Organism Details
Entry: Sh142J21.annot.EG6.embl
Selected feature: bases 1788 amino acids 595
transposon
gene (7 exons)
Transposon En/Spm-like (CACTA)
source 1 126547
CDS 520 2395 Sh142J21g010 putative shrunken seed protein
CDS 5209 8201 c Sh142J21g040 CHP
repeat_region 7771 43555 Ty3-gypsy retrotransposable element similar to maggy like retrotransposon
repeat_region 17873 36854 retrotransposon LTR unclassified
List of annotations
>SEQUENCE
GAACATGCTACTTAATGTTAATGTGTTTGGCAGCACTACTTGACTTGCTT
TTTAGTTGCATAGAGAATTGCTGGTAAATTATTGCAATATATTGTAAGTG
GCTTAAATATGTTGTCAACCACTCTTAGTCTAGGAATTAAATTCTTAAGC
CATACATCCTTCCCCATAGCCTGGCGACGGCTACTGTTATTCAGCTGAGT
GTCTATCTTCGACACCTCTACTCCATCGGCGCTCTCCTTGTTGGTGGCGT
CACCAAATCAAATGCCAGGTTGACTTTGATCCACCTCTTCTTTCACTCCC
CTTCATGGCTTCCATCTCATGTAGCCTGGATGGTGGTACATGGAGCCTGG
ATGTGCTTCAACTGGTGACGCCCCCAACTTGAAGGGAGGTCGTCATCTCA
GGTGGGCTTCGGTTTGAGTGCATCTTCCCACCTTTTGTCCCTATTTTGGT
TGTGCTCCCTCAGGTGGCGGGCAACAAAAGAGCGGCATAGCCACGATGGG
Sh_205L01_1
Sh_251B10_1
Sh_032J13_1
Sh_253J21_1
PB455_unitig_21-Sh_227B16
Protein Coding Gene Models
Select Tracks
Supprimer le surlignage

FRANCE

Angélique D'Hont
Olivier Garsmeur
Carine Charron
Gaetan Droc
Stéphanie Bocs
Guillaume martin

AUSTRALIA

Karen Aitken
Paul Berkman

Robert Henry

SOUTH AFRICA

Bernard Potier
Derek Watt

BRAZIL

Marie-Anne Van Sluys

Edwin van der Vossen
Dennis Woltinge
Jan van Oeveren
Rudie Antonise

Bandie Harrison

International Consortium for Sugarcane Biotechnology

CHARCA, EEAOC, SRA,CTC, CENICANA,CINCAE,CEGICANA,
VSI, MSIRI, SASRI, Mitr Phol, FSCL, HARC, ASCL, RGVSG