

The Australian contribution to the R570 sugarcane genome sequence

Karen Aitken

CSIRO PLANT INDUSTRY
www.csiro.au

Challenges of polyploidy

Dramatic size increase

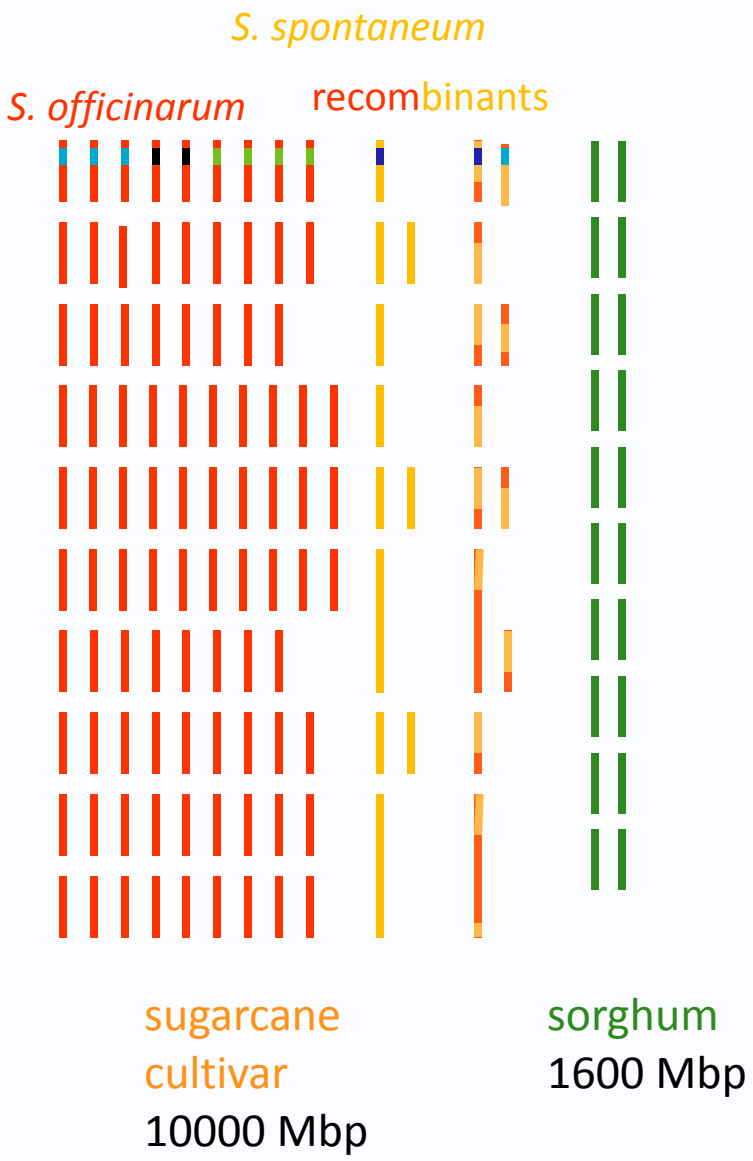
- Much more data required

Higher sequence similarity in functional regions

- Hampers molecular approaches
- Increased computational complexity of data analysis

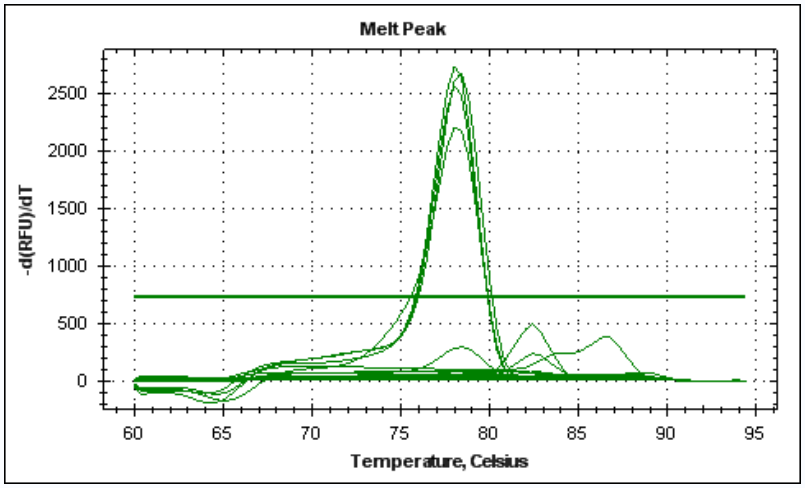
More functional complexity

- Potential exponential increase in gene-gene interactions

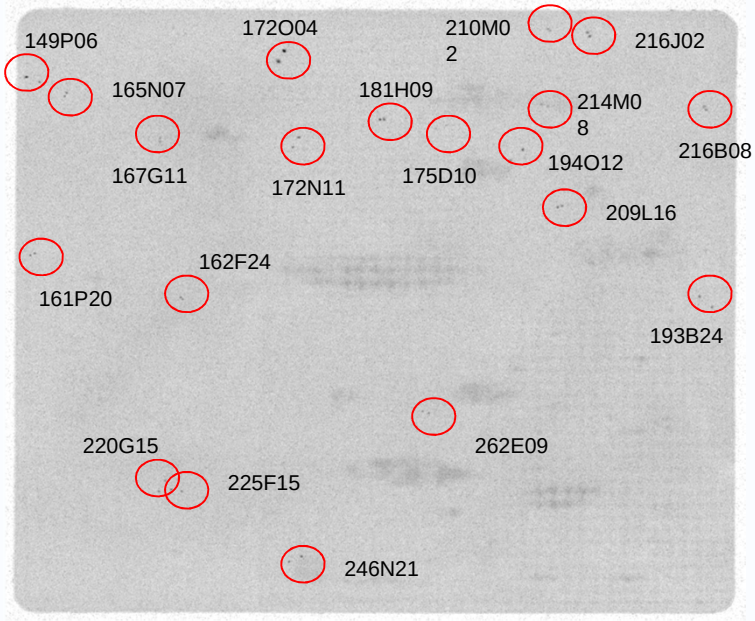


Generation of sugarcane genomic sequence -BAC sequencing

- Generation of a reference sequence by sequencing the minimum tilling path of BAC clones aligned to the sorghum genome
- Selected BAC clones that were linked to traits of agronomic importance by screening the BAC library using both the qPCR 3d pooling strategy and the macroarray hybridisation method using markers with highly significant and robust associations to traits



qPCR used to identify BAC clone

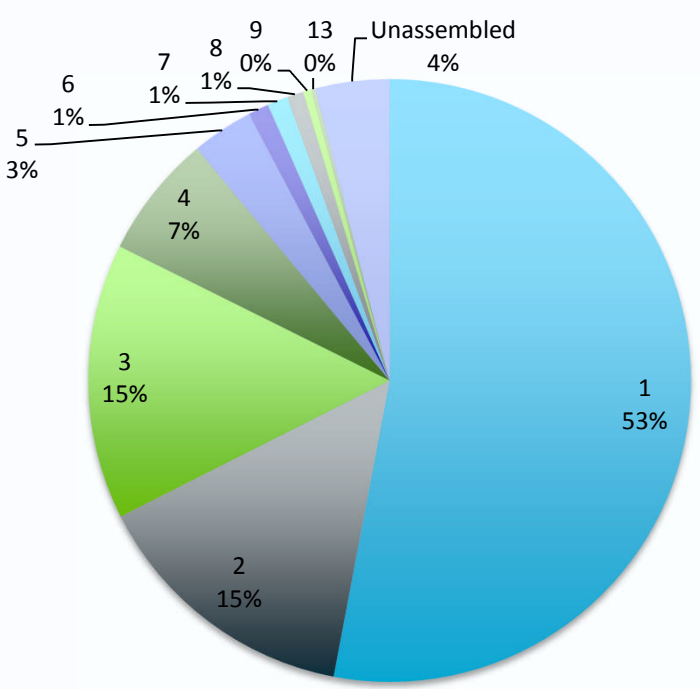


BAC sequencing and assembly

- Sequenced 452 BAC clones using the Illumina HiSeq2000 platform
  - Generated 500 to 10,000 fold coverage of each BAC
  - Assembled reads using a custom pipeline developed in collaboration with the University of Queensland

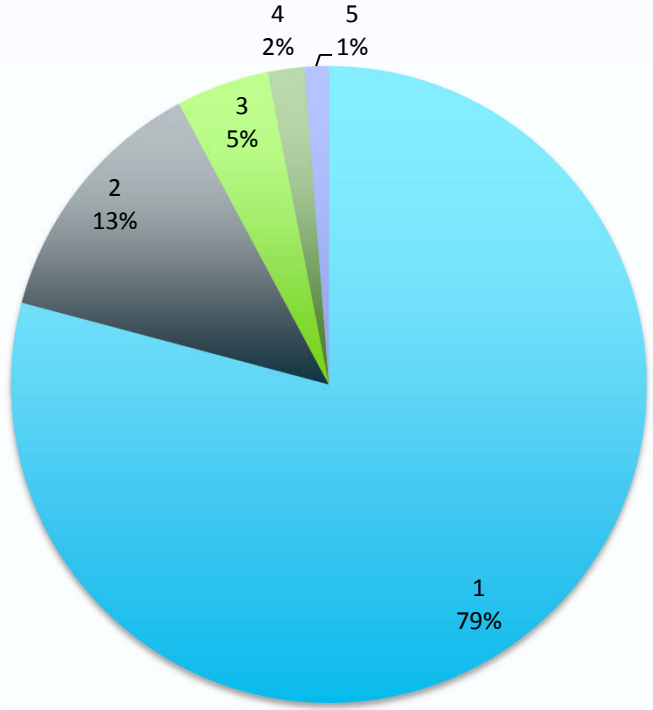
- Sequenced and assembled an additional 696 BAC clones through KeyGene using the Pacific Biosciences long-read platform
  - This platform has been shown to reduce the computational cost of the assembly process

Number of contiguous sequences assembled per BAC for Illumina and Key Gene sequence methods



Illumina Hi seq sequenced

- 68% of BAC clones assembled into 1 or 2 sequences

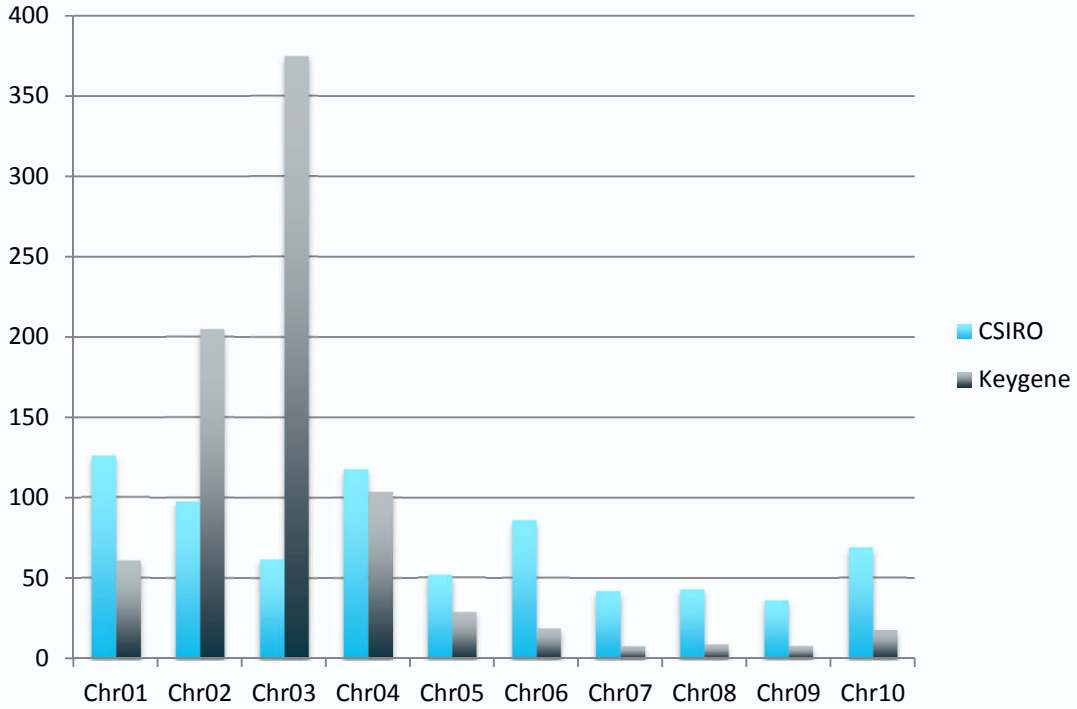


Pacific Bioscience long-read platform

- 92% of BAC clones assembled into 1 or 2 sequences

BAC sequence

- In total we have assembled 987 BAC clones
- Total assembled sequence length is 121 Mbp
- Approximately 17% of the monoploid sugarcane genome
- Aligned to sorghum



Whole genome shotgun sequencing strategy

•Parallel approach to BAC by BAC sequencing method using same variety R570

• Will be combined with BAC data in sequencing consortium

• Short-read Illumina sequences from a range of overlapping fragments

• Advanced computational methods to assemble sequence into scaffolds

Whole genome shotgun assembly

- Illumina short reads were generated as paired-end data spanning 180bp to 600bp
  - Enabled construction of high quality localised contiguous sequences
- Illumina long mate-pair libraries 2kbp to 32kbp
  - Scaffolding of these localised sequences into longer fragments
  - Ability to bridge repetitive DNA sequence
- Pacific Bioscience long read technology enables scaffolding to occur with reads up to 20kbp
  - Has the capacity to close gaps of scaffold assemblies
- BioNano provides physical mapping over very large distances up to 150kbp
  - Enables long distance scaffolding in sequence approaching chromosome pseudo-molecule status

Whole genome shotgun data sets

•complete genome sequence to a level of 40-fold

• represents approx. 500-fold monoploid genome coverage

• highlights sequence variation at each locus

LIBRARY	CURRENT SEQUENCE COVERAGE (x)	CURRENT PHYSICAL COVERAGE (x)
Illumina 180bp	12.14	10.93
Illumina 250bp	3.74	3.12
Illumina 300bp	11.06	16.59
Illumina 600bp	8.70	26.10
Illumina 2,000bp	2.83	31.44
Illumina 2,500bp	18.65	41.94
Illumina 4,500bp	18.64	76.25
Illumina 5,000bp	2.84	78.89
Illumina 7,500bp	1.21	45.38
Illumina 32,000bp	0.37	58.61
Pacbio 20,000bp	3.20	3.20
TOTAL	**83.38**	**320.24**

Assembly of whole genome shotgun data

- 4 High-Performance Computing facilities tested (up to 6 Tb)
- 3 Assembly algorithms compared
  - Velvet (public), AllPaths-LG (public), **Biokanga (CSIRO-developed)**
- Cumulative >6000 days of compute time & max. 1.5Tb RAM used

Total number assembled scaffolds	**4,083,536**
Total assembly Size (bp)	6,759,296,755
Total assembly N50 (bp)	3,186
Overall longest contig (bp)	71,417
Number contigs > 1 kbp	1,941,395 (48%)
Number contigs > 10 kbp	30,502

The first draft assembly represents ~67% of the complete polyploid R570 genome sequence

Assembly validation… (?)

CEGMA (Core Eukaryotic Genes Mapping Approach)

	#Prots	%Completeness	#Total	Average	% Orthology
Complete	171	68.95	489	2.86	78.36
Partial	224	90.32	876	3.91	91.52

- Partial hits to 224 proteins (90% complete), 2.8 copies on average
- Complete hits 176 proteins (69% complete), 3.9 copies on average

Comparisons (BLASTN, evalue $1e^{-20}$)

- Saccharum Gene Index
  - 116,078 (95.7%)

- Sugarcane Transcribed Unigenes (Cardosa-Silva, 2014)
  - 72,269 (97.1%)

Comparison to Sorghum transcripts

Biokanga BLITZ

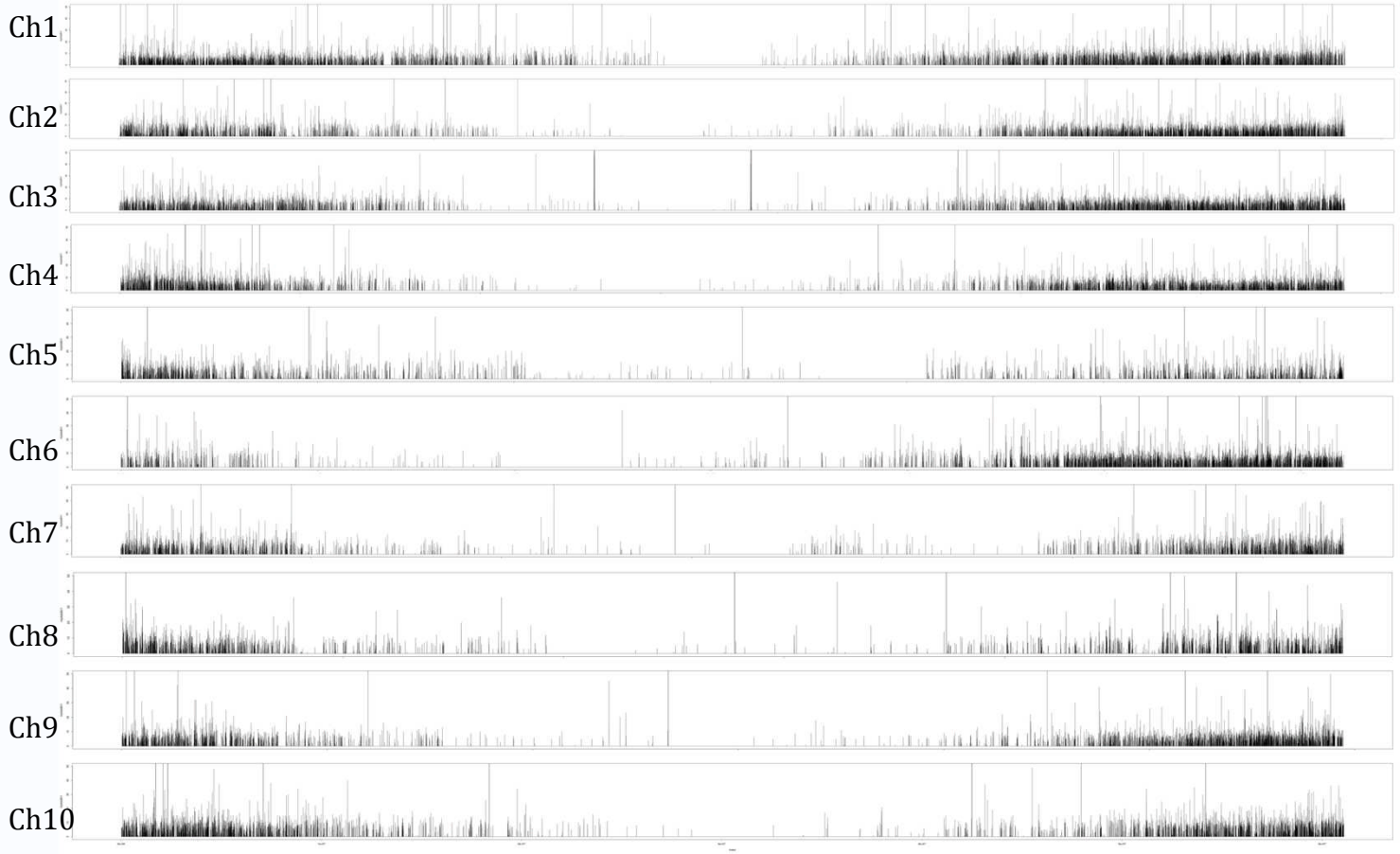
- BLAT-like alignment, developed for transcript/genome alignment
- Extracts intron-exon structures

Sorghum genome v2.1 annotated CDS

- Out of 33,032 transcripts, 29,249 (88.5%) were represented in assembly
- Non-overlapping alignments totaled ~152 Mbp
  - ~400% of Sorghum transcriptome size

- Can we observe allelic diversity in this assembly?

Distribution of allelic diversity across the sorghum genome

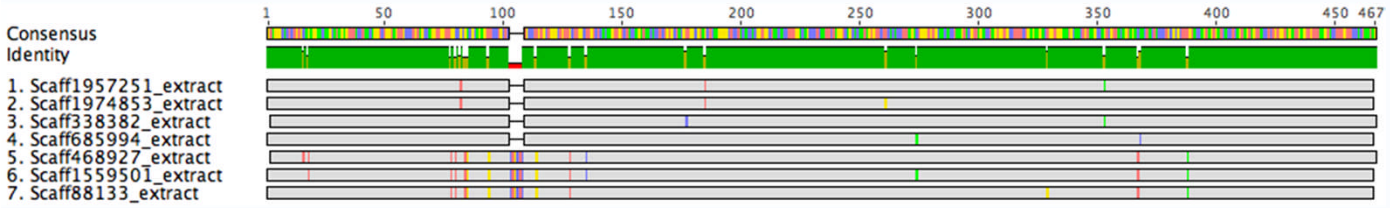


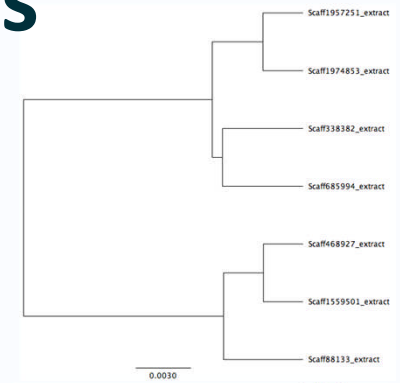
Average depth of 10 alleles

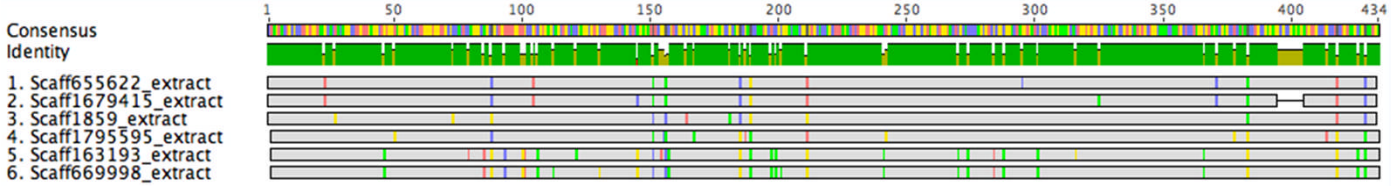
Allelic diversity

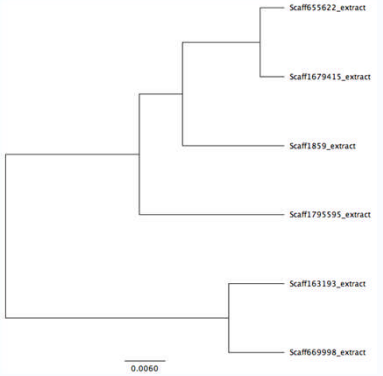
- All genes that were longer than 300bp across the genome
- At least 5 alleles
- Analysed to determine whether the gene copies originated from *S. officinarum* or *S. spontaneum*

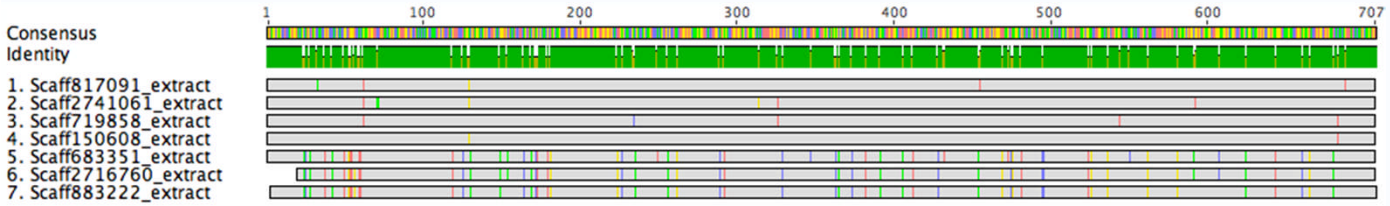
Comparison to Sorghum transcripts













The results indicate that it could be possible to determine the contributions of alleles from parental species

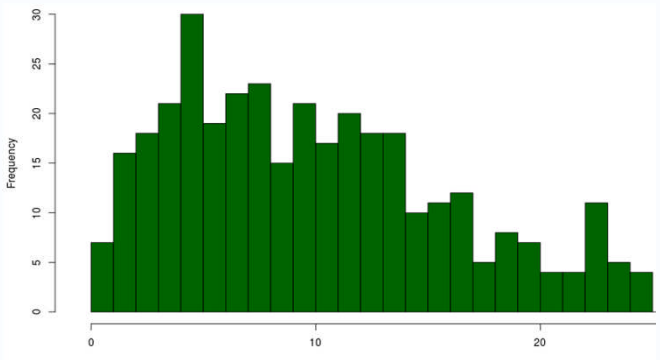
WGS assembly vs BAC assembly

WGS assembly can be used to improve BAC assembly

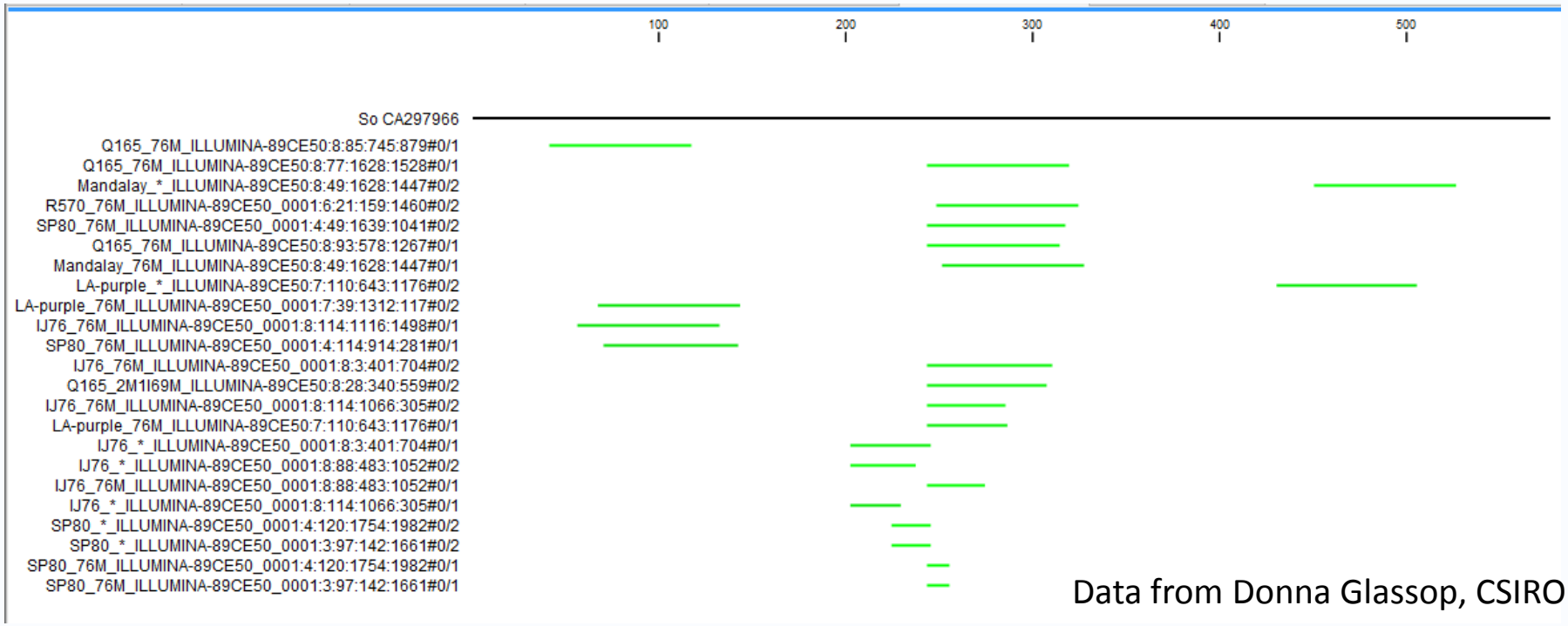
- BAC assembly = template for allele diversity analysis in WGS



- Average WGS coverage on 664 BACs orthologous to *S. bicolor* chromosome 6 has a mode of ~8x



Allelic variants of FT identified from whole genome shotgun assembly



Data from Donna Glassop, CSIRO

Now enables analysis of allele distribution and role in signalling

Summary

- The present assembly does not include the PacBio long read data and work is being carried out to combine this with the current assembly
- To help improve the assembly a physical map of R570 is being generated using BioNano Genomics optical mapping technologies
- The R570 mitochondrial genome sequence has been assembled
  - 9 large contigs
- The current preliminary assembly of R570 has been used to extract genomic context information for sucrose accumulation genes and flowering genes. We see this as a major resource for further sugarcane research.
- The current assembly has also been used to validate all alleles of genes in the sequenced BAC clones.

Acknowledgements

Team members

- Paul Berkman
- Jai Perroux
- Jingchuan Li
- Rosanne Casu
- Anne Rae





