Unlocking the potential of sugarcane

SOUTH AFRICAN SUGARCANE RESEARCH INSTITUTE

AN INTEGRATIVE APPROACH TO GENERATING A REFERENCE TRANSCRIPTOME FOR SUGARCANE

Robyn Jacob

CHPC National Meeting 2017

South African Sugarcane Research Institute is a division of the South African Sugar Association

SASA

SOUTH AFRICAN SUGARCANE INDUSTRY STATISTICS

Sugarcane is the 2nd largest field crop by gross value.

Industry:

- 14 Sugar Mills
- 23 866 growers (94% SSG)

Area planted: ~370 000 ha

Production per annum*:

- Sugarcane: 14 to 20 million tons
- Sugar: 1.6 to 2.3 million tons

(DAFF: Abstract of Agricultural Statistics – 2017)

*Values over previous 5 years

THE RESEARCH INSTITUTE

Est. in 1925 - 90 years of research excellence

Staff complement

- Permanent ~ 540
- Contract/students ~150

Expertise

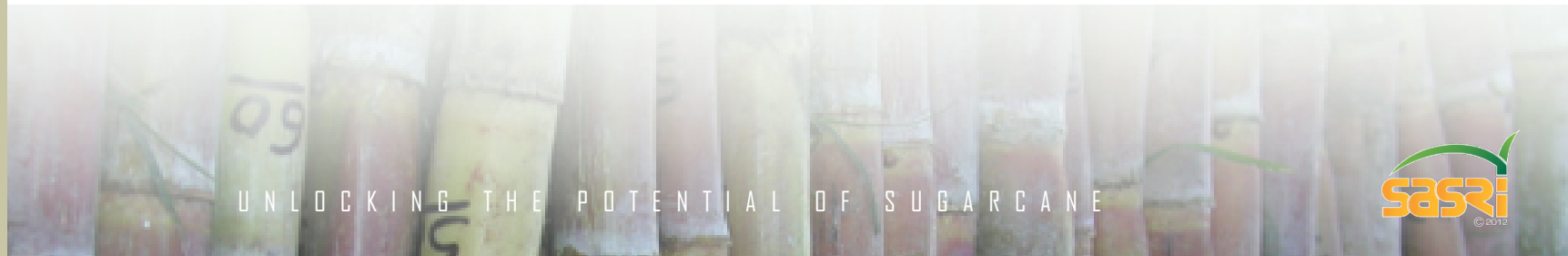
- Plant Breeding
- Molecular Biology and Biotechnology
- Crop Physiology
- Agronomy and Agronomic Modelling
- Entomology and Pathology
- Soils and Crop Nutrition
- Agricultural engineering
- Extension
- Biosecurity

Collaboration

- National and International Universities and Research Institutes

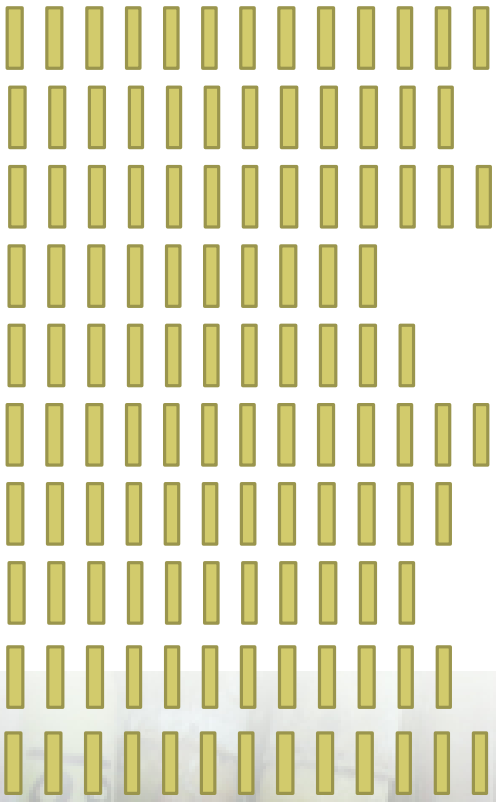
SASRI'S BIOINFORMATIC RESEARCH PROGRAM

- Whole plastome assembly, annotation and phylogenetics
- Sugarcane exome assembly and phylogenetics
- Transcript mapping, assembly and differential expression
- Alternate transcript discovery
- Full text searching (gene discovery, gene annotation, pathway assignment)
- Structural modelling, ligand binding, active site in silico mutagenesis
- Development of novel, low memory approaches to gene assembly and annotation
- Development of the only SNP mapping pipeline that works with polyploids
- Image analysis (primarily for automated karyotyping)



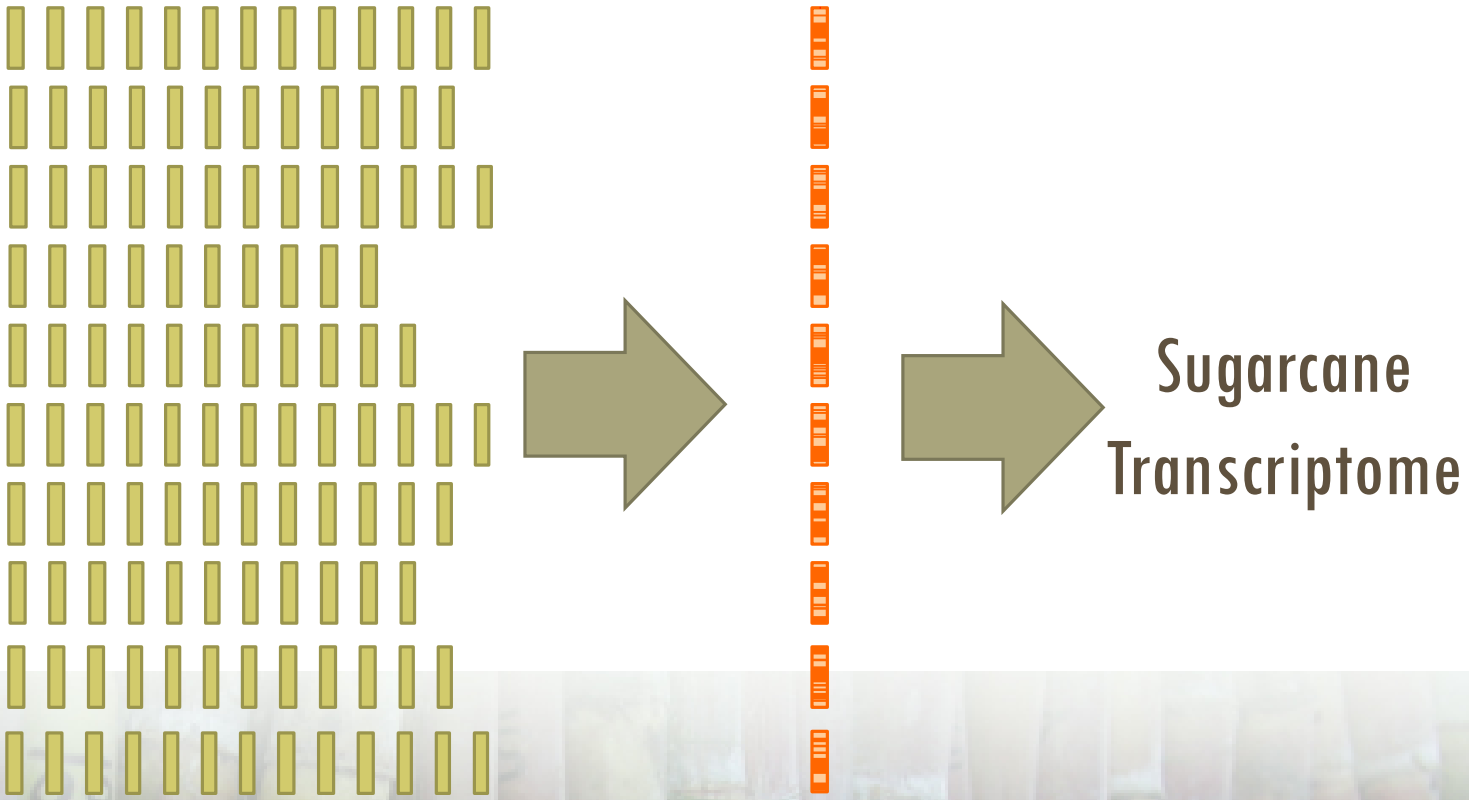
Modern Cultivated Sugarcane

- Complex genome
  - Highly polyploid and aneuploid (10+ alleles per gene)
  - Large total genome size (~10Gb)
  - 2n = 100 to 130 chromosomes



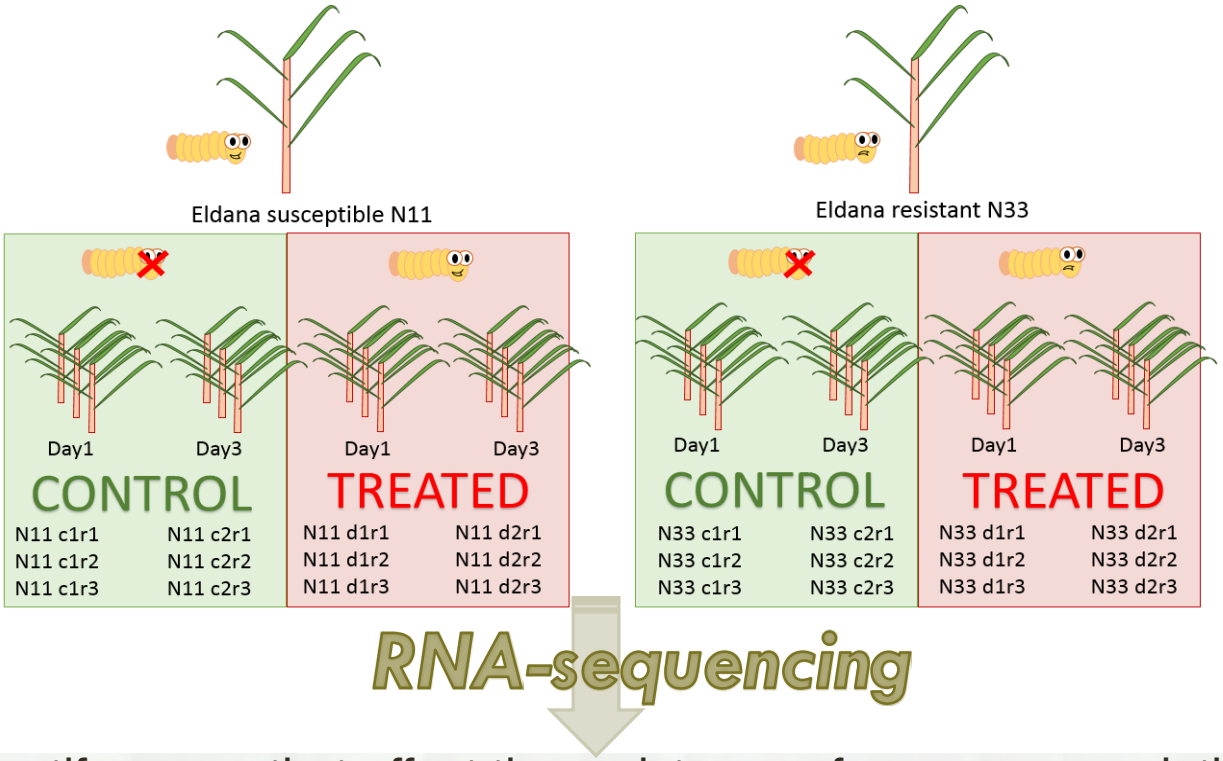
Modern Cultivated Sugarcane

- Complex genome
  - Highly polyploid and aneuploid (10+ alleles per gene)
  - Large total genome size (~10Gb)
  - 2n = 100 to 130 chromosomes



RNA-seq in sugarcane

- Perform a RNASeq experiment to identify resistance mechanisms to *Eldana saccharina*
  - *de novo* transcriptome assembly
  - Hybrid approach: Reference transcriptome + *de novo* transcriptome



Identify genes that affect the resistance of sugarcane varieties to the stalk borer, eldana

Unlocking the potential of sugarcane

AN INTEGRATIVE APPROACH TO GENERATING A REFERENCE TRANSCRIPTOME FOR SUGARCANE

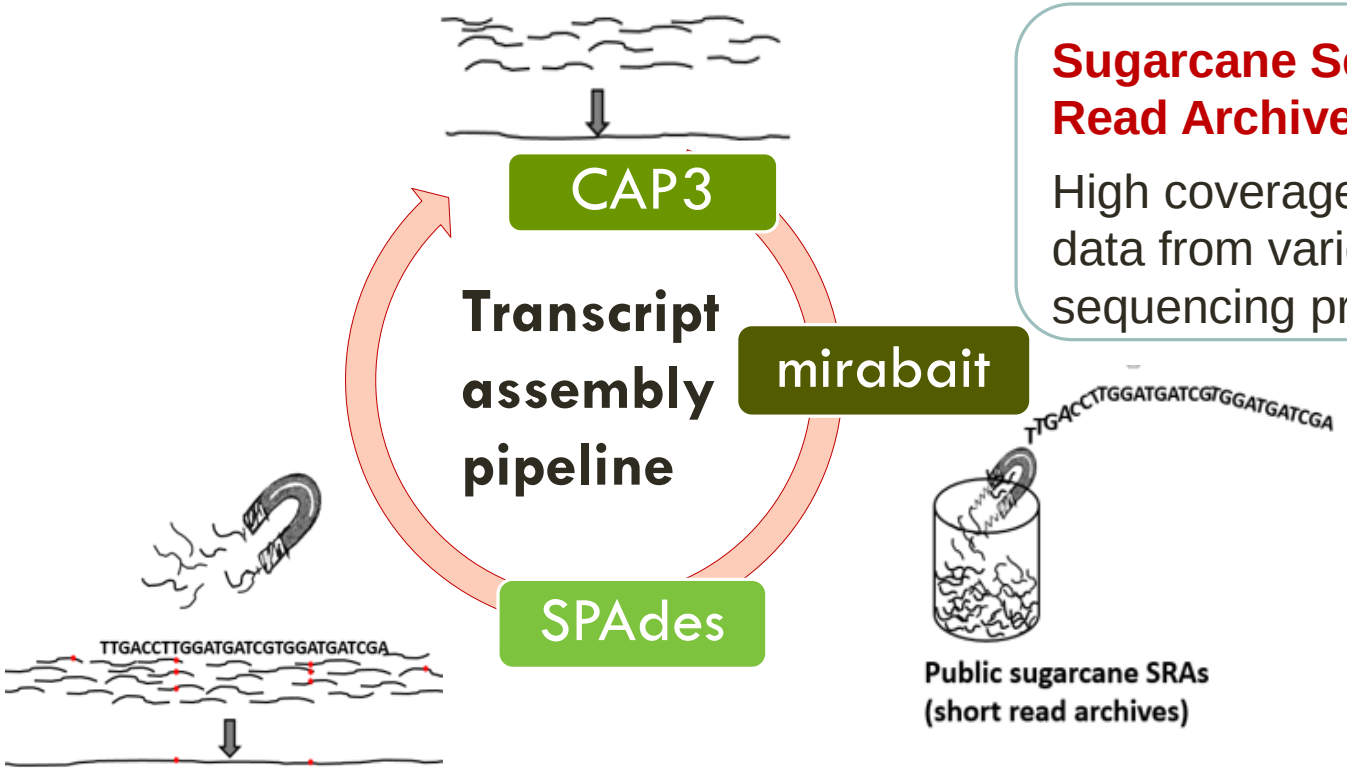
SASRI

Towards a Sugarcane Transcriptome
-The transcript assembly pipeline

Sugarcane ESTs:

ESTs from different organs and tissues at various developmental stages.

~285 138 ESTs



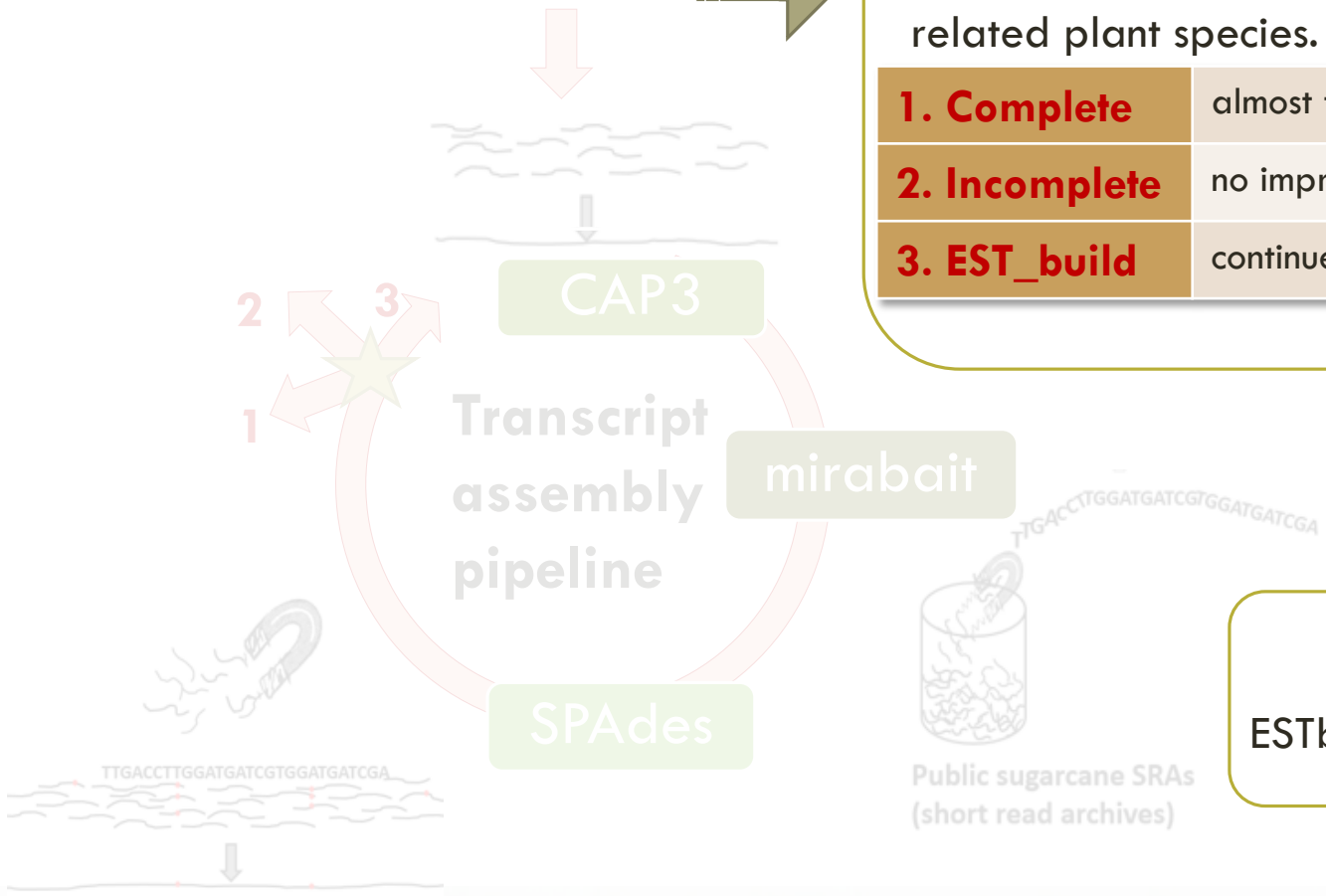
Sugarcane Sequence Read Archives (SRA's):

High coverage short read data from various sequencing projects

1. CAP3: **Huang and Madan** (1999) Genome Res **9**:868-877.
2. mirabait: **Chevreux et al.** (1999) GCB **99**:45-56.
3. SPAdes: **Bankevich et al.** (2012) J Comput Biol **19**:455-477.

Towards a Sugarcane Transcriptome
-The transcript assembly pipeline

~285 138
ESTs



Decision Tree

EST_build transcript compared to a reference set of genes from closely related plant species.

1. Complete	almost full-length match
2. Incomplete	no improvement in builds
3. EST_build	continue building

28 244
ESTbuild_transcripts

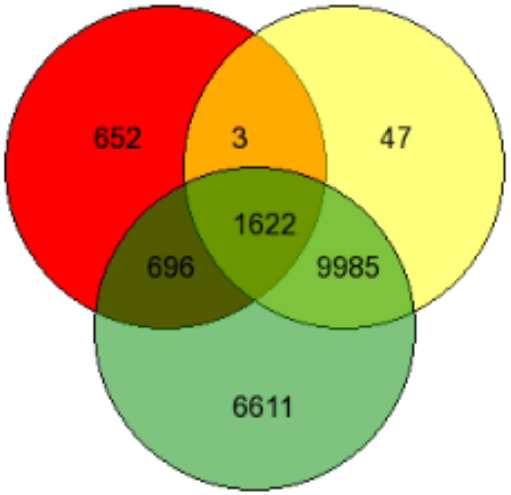
Towards a Sugarcane Transcriptome
-Further validation of ESTbuild_transcripts

28 244
ESTbuild_transcripts

SP80 synthetic long reads
LASTZ
CAP3
BLAT
exonerate

Sorghum transcripts
LASTZ

Sorghum CDS
LASTZ

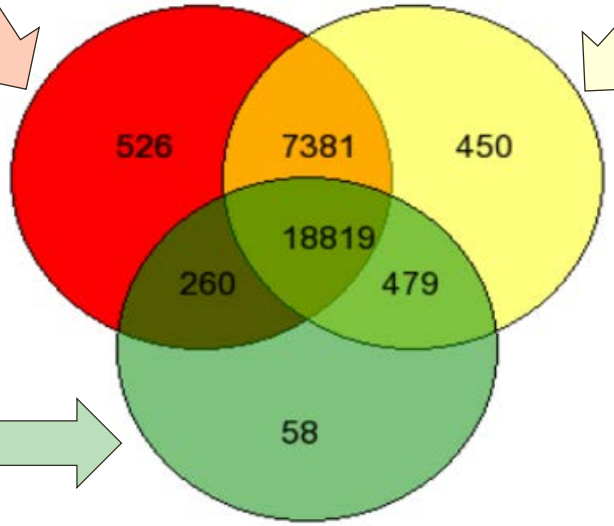


19 616
Reference transcript builds

Towards a Sugarcane Transcriptome

RNA-seq data from stem tissue of local sugarcane varieties
Remove rRNA
Remove plastomes
Trim adapters
Trim low quality
Trinity *de novo* assembly
LASTZ

RNA-seq data from leaf, root and stem tissues of sugarcane variety R570
Remove rRNA
Remove plastomes
Trim adapters
Trim low quality
Trinity *de novo* assembly
LASTZ



Reference transcript batch1
19 616

27 973 Reference transcript builds

Current Status of Reference Transcriptome

- ESTbuild_transcripts: 27 973 (99%) were validated using an additional data source

But what about those genes not represented in the sugarcane ESTs ?

- Currently using the transcript assembly pipeline to build synthetic sugarcane transcripts of
  - all sorghum CDS not represented by any sugarcane EST
  - The sugarcane SRA's and RNA-seq data are being used in the build, to determine the sugarcane version of the sorghum gene
- Annotation of the transcripts

CONCLUSIONS

- We have developed a novel transcript assembly
- The approach taken maximises the use of available data
- It is highly accurate and effective
- The methodology is less memory intensive than pure *de novo* assembly and is well suited for distributed computing and CHPC architecture

ACKNOWLEDGEMENTS