INTEGRATING TRANSCRIPTOMICS AND EXPRESSION ANALYSIS: THE SASRI APPROACH

Bernard A Potier, Robyn M Jacob and Dyfed Lloyd Evans

South African Sugarcane Research Institute

Private Bag X02, Mount Edgecombe 4300, South Africa

ISSCT joint Breeding & Germplasm and Molecular Biology Workshops
Saint Gilles Les Bains, Réunion Island, 1-5 June 2015

1925 - 2015
Unlocking the
potential of sugarcane

The context

Next Generation Sequencing technologies (NGS) are becoming cheaper and easier to produce,

Increased access to NGS data from various sources and organisms are becoming available,

Solid progresses are being made towards producing a template sequence for the sugarcane genomes,

SASRI has invested and committed efforts into producing and utilizing NGS data to deliver better varieties to the South African sugarcane industry,

Computing resources at SASRI are limited, there is a need to streamline the processes to avoid system crashes.

The biological needs

A better and deeper understanding of the complex molecular mechanisms underlying some biological systems is crucial to devise a better strategy to deal with them; we are certainly moving from a single gene analysis to a more holistic systems biology approach to study biological functions and molecular pathways. A few of the crucial targets for that kind of studies are:

-Pest resistance,
-Biotic/abiotic stresses,
-Sugar metabolism
-Other phenotypic traits

The what and how

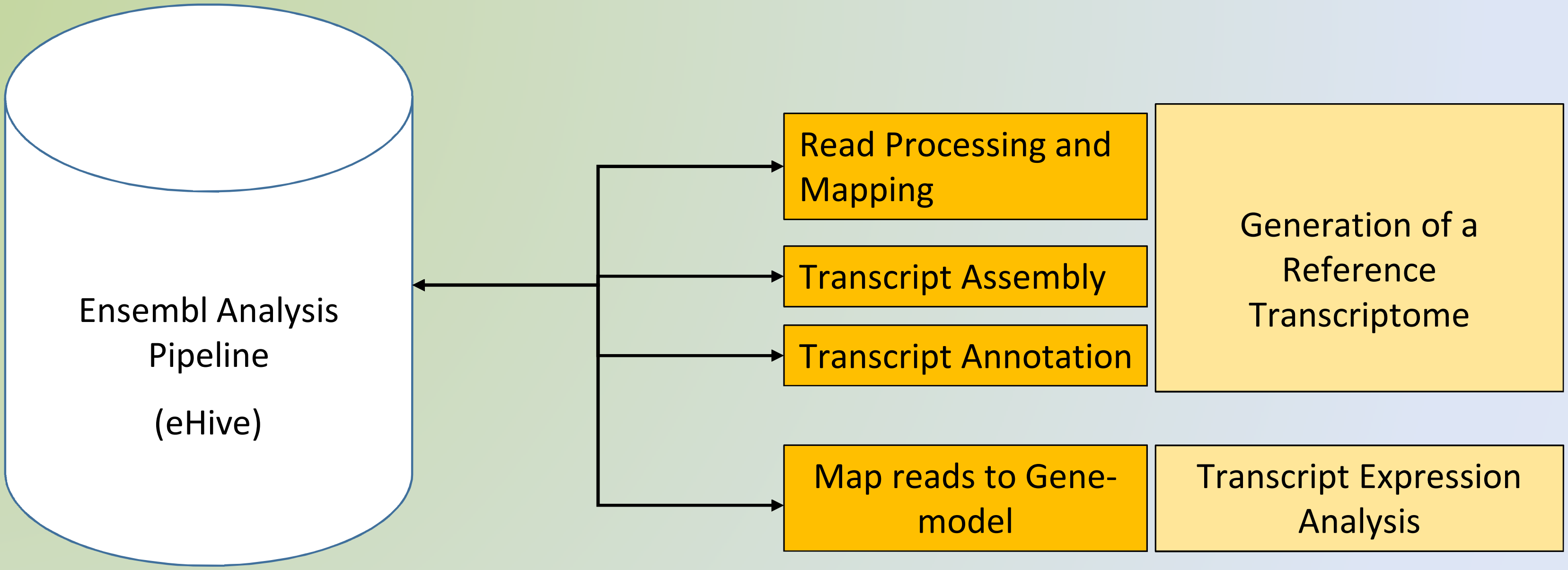
The analysis of gene expression is becoming a powerful tool to understand and ultimately modify specific traits and phenotypes (gene networks),

There are however some challenges that need to be tackled in order to extract the best informative value from those studies.
One of them is to devise tools to best analyze the data, especially in the context of a plant crop without (yet) a reference genome,

SASRI is developing its own end-to-end transcriptomics analysis pipeline, to investigate (*i.a.)* differential gene expressions,

The pipeline will be modular (use of available open source tools), will integrate all currently available data from model and non-model organisms to first build a reference set (gene model).

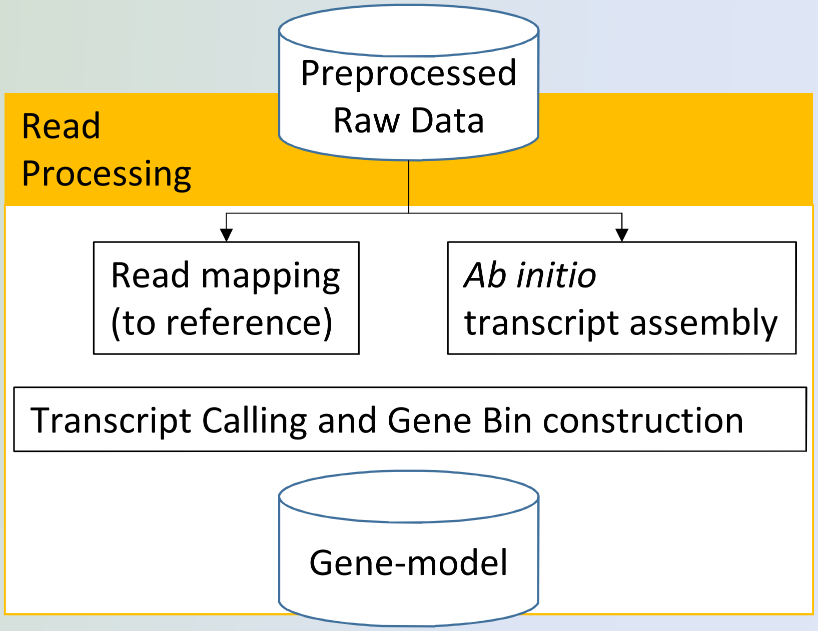
Limited memory and storage mean that the pipeline must be streamlined. The time factor is also crucial in this process.


Ensembl Analysis Pipeline
(eHive)
Read Processing and Mapping
Transcript Assembly
Transcript Annotation
Generation of a Reference Transcriptome
Map reads to Gene-model
Transcript Expression Analysis

The Ensembl Analysis pipeline will be extended to include these modular processes. The Ensembl pipeline is not only able to control and monitor parallelized tasks but also permits the addition of custom algorithms and components into the pipeline *via* an object oriented programming interface. All sequences, analyses, pre-processed analyses and output data are stored as serialized objects in a relational database system.

Generation of a Reference Transcriptome

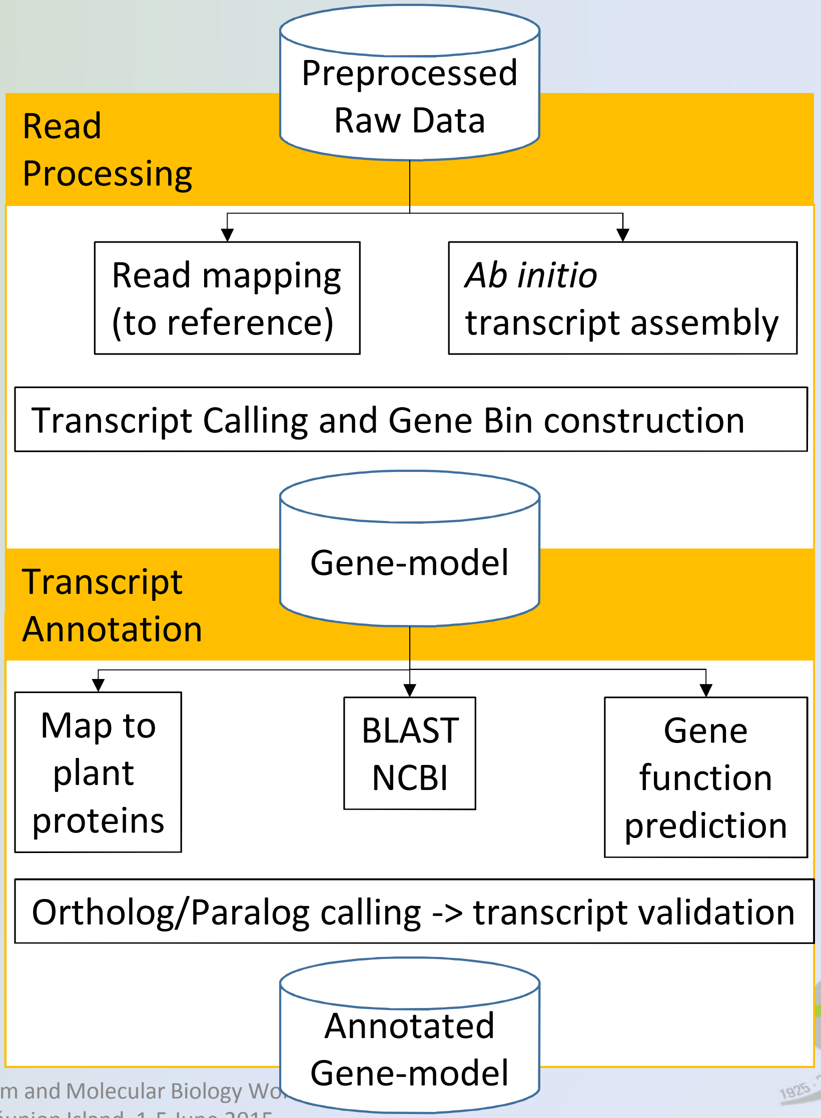
- Read Processing of publicly available data
  - Transcripts identified through
    - direct mapping of reads to suitable reference plant genomes, gene sets and transcriptomes
    - *Ab initio* assembly of reads
  - Gene-model generation: final transcript calling and gene binning


Preprocessed Raw Data
Read Processing
Read mapping (to reference)
Ab initio transcript assembly
Transcript Calling and Gene Bin construction
Gene-model

Generation of a Reference Transcriptome

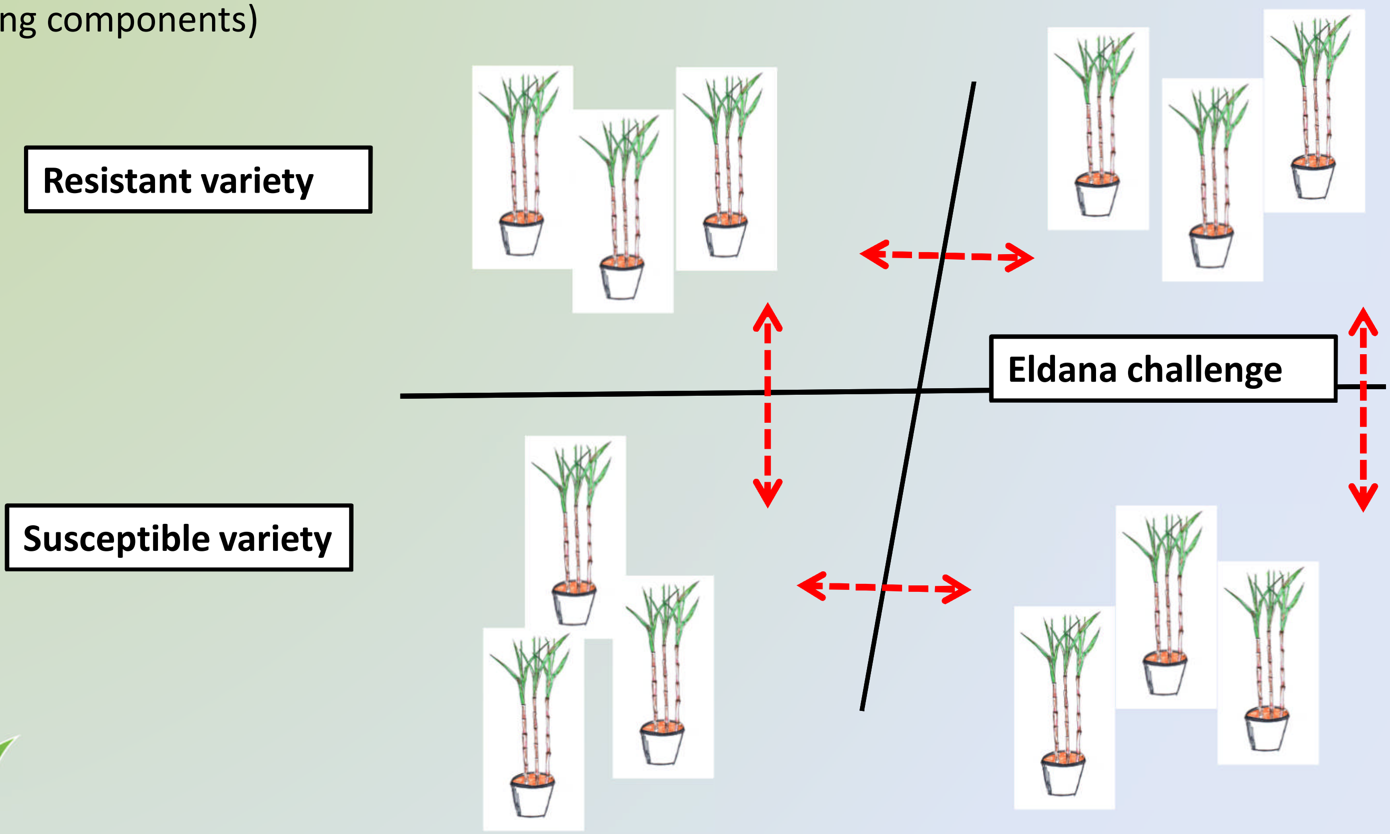
- Read Processing of publicly available data
  - Transcripts identified through
    - direct mapping of reads to suitable reference plant genomes, gene sets and transcriptomes
    - *Ab initio* assembly of reads
  - Gene-model generation: final transcript calling and gene binning

- Transcript Annotation
  - Map to known plant genes at protein level
  - BLAST against NCBI databases
  - Gene function prediction by presence of domains and motifs


Preprocessed Raw Data
Read Processing
Read mapping (to reference)
Ab initio transcript assembly
Transcript Calling and Gene Bin construction
Gene-model
Transcript Annotation
Map to plant proteins
BLAST NCBI
Gene function prediction
Ortholog/Paralog calling -> transcript validation
Annotated Gene-model

Test case: Identification of Eldana resistance mechanisms in sugarcane nodes of a resistant variety.

The expected outcome will be a set of genes differentially regulated in the stem upon boring by Eldana larvae and potential gene candidates involved in resistance (combined temporal responses of many interacting components)



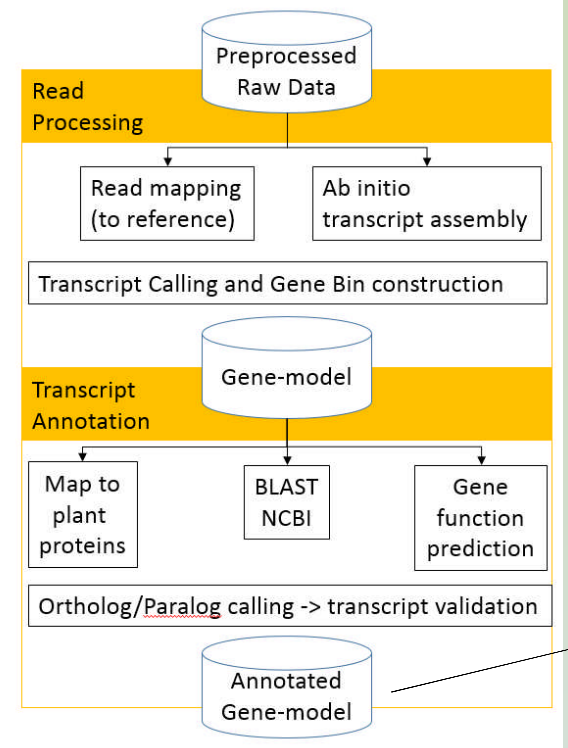
Resistant variety

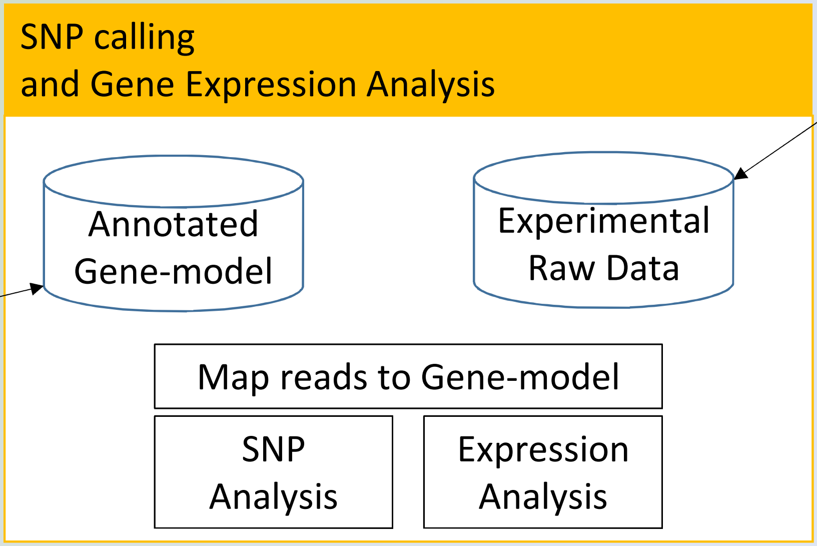
Susceptible variety

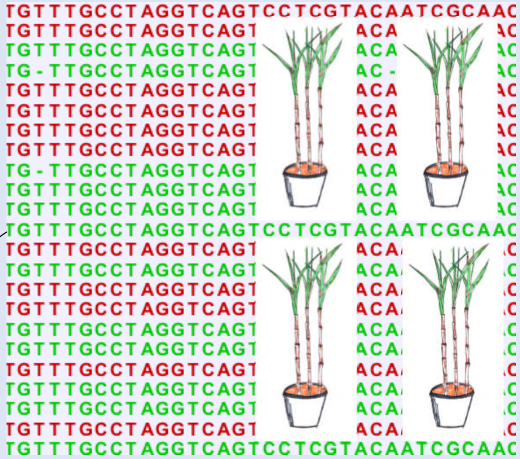
Eldana challenge

Transcript Expression Analysis

This pipeline will be an end-to-end analysis pipeline for dealing with transcriptome annotation and expression analysis in the absence of a reference genome


Preprocessed Raw Data
Read Processing
Read mapping (to reference)
Ab initio transcript assembly
Transcript Calling and Gene Bin construction
Gene-model
Transcript Annotation
Map to plant proteins
BLAST NCBI
Gene function prediction
Ortholog/Paralog calling -> transcript validation
Annotated Gene-model





More specifically, expected outcomes of SASRI-derived initial experiments:

-Gene expression quantification (depending on the RNA-seq technology that is used, normalization and experimental procedures)

-Transcripts annotation

-Multigene transcript families and their associated SNPs

-Alternate transcripts identification

Biological interpretation

-Identification of possible candidate genes

-Rare or unique transcripts involved

Key points of the SASRI pipeline:

-Integrate all currently available data from model and non-model organisms to generate a reference

-Perform transcriptomics analyses in the absence of a reference genome

-As more data (transcriptomics/genomic) become available, the reference gene-model improves

-Modular, both in terms of software and data sets

Bernard **Dyfed** **Robyn**

sasri ©2015

ISSCT joint Breeding & Germplasm and Molecular Biology Workshops
Saint Gilles Les Bains, Réunion Island, 1-5 June 2015

1925 - 2015
Unlocking the potential of sugarcane