Towards Copy-Aware Assembly of the Sugarcane Genome

Gabriel R. A. Margarido
Cristina Pop
Bob Davidson
Glaucia M. Souza
David Heckerman
Outline

• Initial efforts

• Initial findings

• Directions we are headed
Introduction

• Traditional breeding facilitated by cloning

• Complex aneuploid and polyploid genome
  • 10 Gb in 100-130 chromosomes
    o 1 Gb monoploid genome
    o 6 to 12 copies each

• Interest in difference between homoeologues
  o Assembly software collapses SNP’s
Motivation

- Goal: assembly with chromosome copy sorting
Synthetic Genome

- *Sorghum bicolor*
  - Closest diploid species to sugarcane
    - 95% similarity
  - Sequenced genome
Synthetic Genome

Original Sorghum chromosome

Derived chromosome #1
  - Derived #1_1
  - Derived #1_2

Derived chromosome #2
  - Derived #2_1
  - Derived #2_2

Derived chromosome #3
  - Derived #3_1
  - Derived #3_2

Mutations
Synthetic Genome

• Rearrangements
  o Fusions
  o Duplications
  o Inversions
  o (Reciprocal) Translocations
• Hypothetical polyploid genome
  o 96 chromosomes
• Read simulation
  o Any desirable error model
Comparative Approach

• Comparative genome assembly
  o Alignment against reference
  o Layout identification
  o Contig formation

• Results (projected third-generation technology)
  o Broken and collapsed assembly
  o Short contigs
Ideas from RNA-Seq

- Alternative splicing
- Expression of both alleles in diploids
Trinity

- Read set
- Extend in k-mer space and break ties
- Overlap linear sequences by overlaps of k - 1 to build graph components
- De Bruijn graph (k = 5)
- Compacting
- Finding paths
- Compact graph with reads
- Extracting sequences
- Transcripts
Trinity Results

- Extremely short contigs
  - Smaller than average read length
- Theoretically promising method
- In current state, not appropriate
Current Technologies

- Read length
- Sequencing errors
  - 454 reads for initial assembly
  - Deep coverage (> 50X) Illumina data for SNP calling
Phasing in Humans

- Most assemblies are haploid
- SNP calling
- Phasing = chromosome sorting in sugarcane
- Current technologies are not enough for reliable phasing
- Current trend
  - Use of fosmids for individual sequencing of haploid segments
Hierarchical Assembly

- High fragment coverage and low read coverage
Acknowledgements

- Manju Shivanna
- Nir Friedman
- Carlos Hotta
- Jacob Kitzman