Assembling and Finishing the Genome of CMS Sorghum Mitochondria using Long Span NGS Read Technology

Presenting Author
Robert R. Klein, USDA-ARS

Fertility Restoration & Cytoplasmic Male-sterility in Sorghum

• Understand interaction of nuclear-encoded fertility restorer genes & CMS-associated sequences of the mitochondrial genome
• Assist breeders in developing new CMS-female and male (restorer) inbreds

CMS Mitochondrial Genomes: Why de novo assembly?

• Genome expansion due to:
  – Large intragenic repeats
  – Repeated segments
  – Intron expansion
  – Incorporation of nuclear & plastid DNA
• Frequent recombination events & dynamic genome rearrangement
  – Rearrangements associated with CMS

Sorghum: Features of organelle genomes

• Multiple stretches of shared sequence
  • 20 segments ranging up to 6.4 kb at >95% identity in plastids & mitochondria
• Shared sequences contained within larger scale repeats
  • Mitochondrial ~16 & 4 kb direct repeat, & smaller repeat units
• Plastid genome contains >21 kb inverted repeat
  • includes 5.2 kb shared sequence region with mitochondria

Features of Organelle Genomes (continued)

• Presence of multiple subgenomic forms
  • Arise from recombination between repetitive elements
• Smaller circular subgenomes contain variable regions of the mitochondrial master circle genome

Why Mate Pair Data?

• Fragmented De novo Assembly
  Genome shredding by NGS process
• Repetitive Genomes
  Every contig begins & ends with a repeat
• Structural Variant Detection
  Indels/rearrangements are subtle
• Gap Closure & Genome Finishing
  Genomic context scrambled
De Novo Assembly Problems

Complex genomes contain $>10^6$ repeats $>500$bp long
Repeats longer than read length create gaps in the assembly
An assembler can not distinguish the repeats

Large Repeats, Shared Sequences Complicate Genome Closing

Multiple MP libraries are Ideal

Multiple MP libraries for Sorghum Organelle Assembly

- Enriched subcellular mitochondria fraction Percoll-purified from CMS inbred A$_1$Tx623
- Organelle DNA $>50$ kb isolated
  - 50 ug (low input DNA for multiple libraries)
- MP libraries produced by Lugicen
  - Fragment library
  - 8 kb
  - 20 kb
  - 40 kb fosmid

Ideal Mate Pair Library Technology

- High efficiency true mate pairs
- No chimeras (false mates)
- No bias, low redundancy
- Tight size distribution
- Full length reads from both sides
- Wide range of insert size choices
- Low input DNA
NxSeq Mate Pair Features

- Mate pair efficiency >90%
- Encrypted ChimeraCode™ detection eliminates false mates
- JunctionCode™ sequence identifies left & right mate pairs
- User defined mate pair libraries up to 20 kb
- Multiplexing, cross contamination control via barcodes

40 kb Fosmid Mate Pairs

Closing Large Genomes

40 kb Mate Pair Sequencing

- Mate pair sequencing of fosmid ends
  - Generates paired end sequences 40 kb apart
- pNGS FOS
  - New tool for massively parallel NGS analysis

Sorghum Mitochondria: Fosmid MP library insert size distribution

Span sizes for pNGS mate pair library

40 kb Inserts

NGS Mate Pairs (100–500 bp)
Conclusions

- A new paradigm for closing genomes has been developed.
- >90% efficient mate pair libraries.
- ChimeraCode™ identifies false junctions.
- JunctionCode™ identifies mate pair junctions.
- NxSeq NGS Mate Pair Technology enables accurate, economical assembly of BACS and genomes.

Collaborators

- David Jordan, Emma Mace University of Queensland, AU
- Ian Small, Western Australia University
- Lucigen Inc.
  - Mike Lodes
  - Scott Monsma
  - David Mead
- Patricia E. Klein & Mahnaz Kiani (RNASeq), Texas A&M University
- Funding provided by Australia Research Council & the USDA
40 kb NGS Mate Pair Fosmid

2-20 Long Mate Pair Kits

NxSeq Mate Pair Conversion