



**EXPLORING THE PINE GENOME USING COT
FILTRATION AND 454 LIFE SCIENCES
MASSIVELY PARALLEL SHOTGUN SEQUENCING**

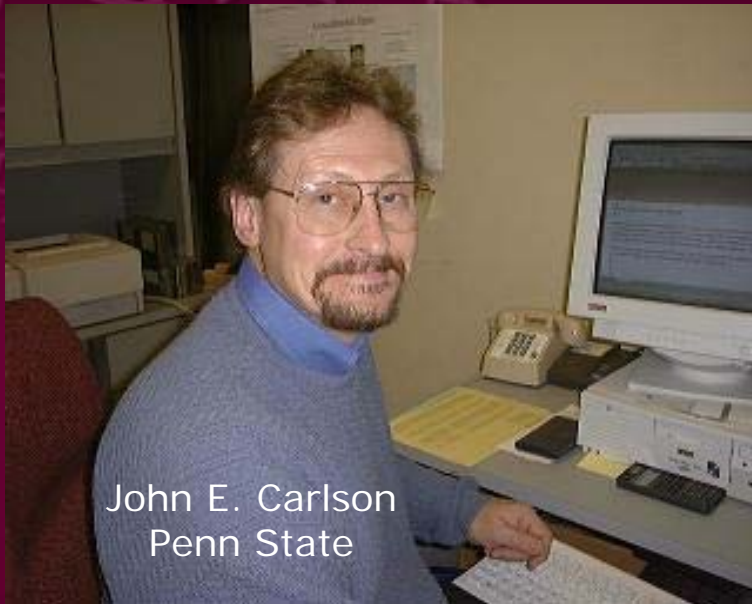
Daniel G. Peterson

Mississippi Genome Exploration Laboratory (MGEL)

Mississippi State University



COLLABORATORS



John E. Carlson
Penn State



MGEL – Philippe Chouvarine, Supaphan Thummasuwan, Dipaloke Mukherjee, Surya Saha, LaShonda Robertson, Annita Avery, & Zenaida Magbanua



Jim Leebens-Mack
Penn State

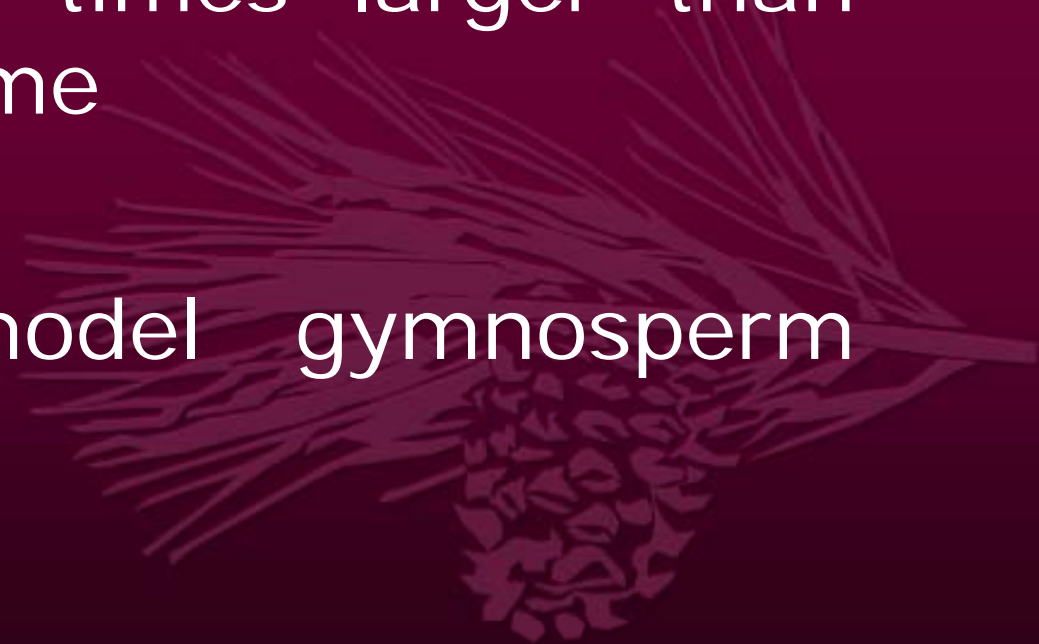
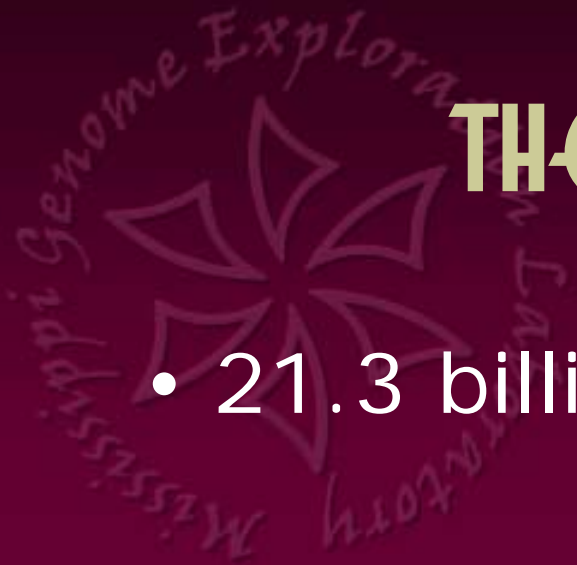
PINUS TAEDA L. (LOBLOLLY PINE)

- Most important crop plant in southeastern U.S.
- Most highly planted tree species in the world.
- Loblolly and other southern yellow pines grow on only 6% of U.S. forest lands but account for 58% of wood supply.



THE PINE GENOME

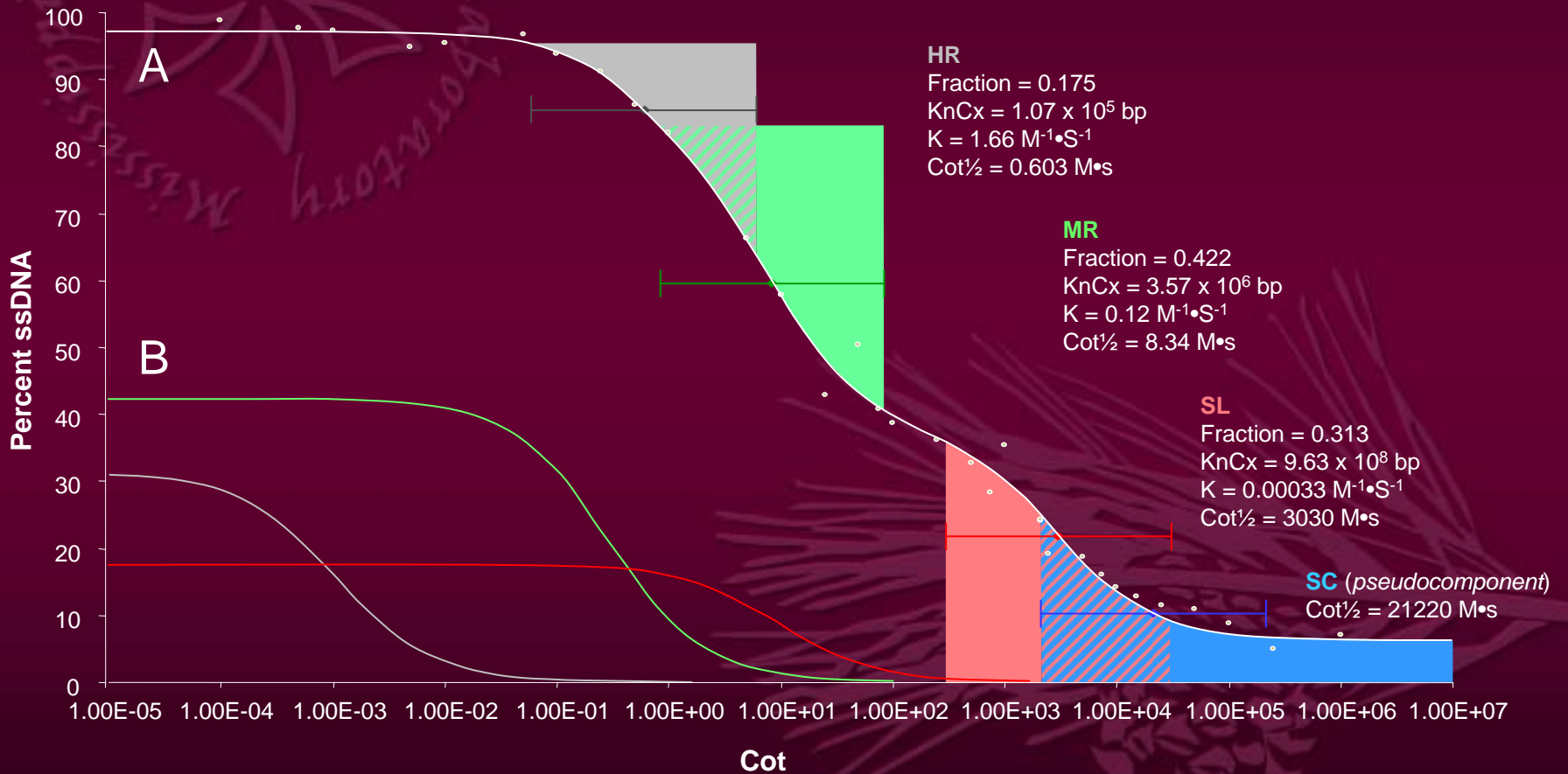
- 21.3 billion base pairs
- About seven times larger than human genome
- *De facto* model gymnosperm genome



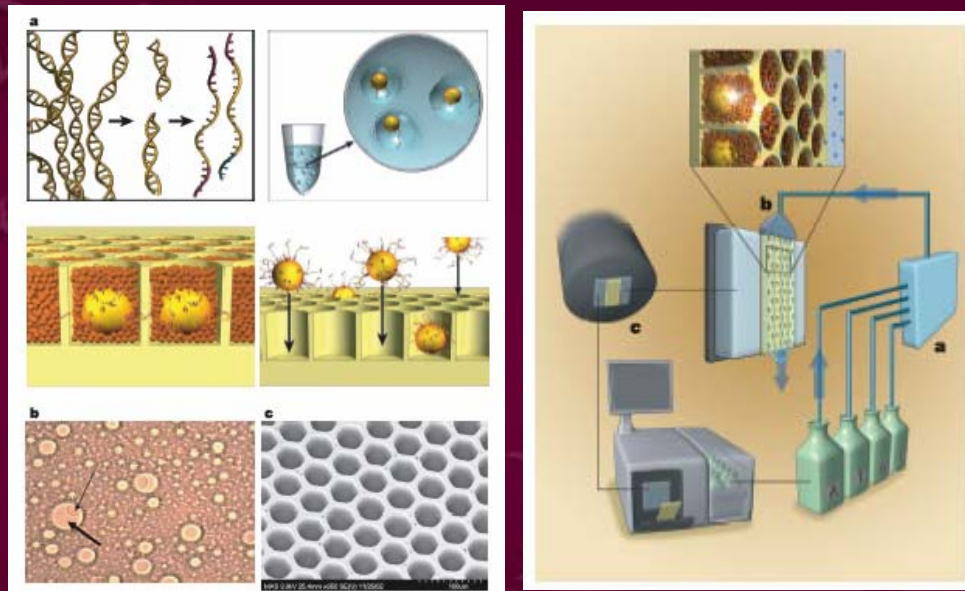
COT FILTRATION

- *Cot filtration* (CF), a technique rooted in the principles of DNA renaturation kinetics (*i.e.*, Cot analysis), is a means by which the repetitive DNA sequences that dominate many eukaryotic genomes can be separated from "gene-rich" single/low-copy sequences.
- CF is most accurately performed if fractionation is based upon the results of a Cot analysis.

PINE COT CURVE



THE 454 GS20



Margulies et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

THE 454 GS20

- 20 million nucleotides in 4.5 hours
- No cloning involved
- Can sequence a microbial genome in a day
- Read length is primary drawback (ca. 100 bp after trimming in our experience)
- Some question about accuracy when reading long, single nucleotide stretches

PINE 454 SEQUENCING TO DATE

- Total of 100.6 million nucleotides
- 1 million reads
- Mean read length = 100 bp

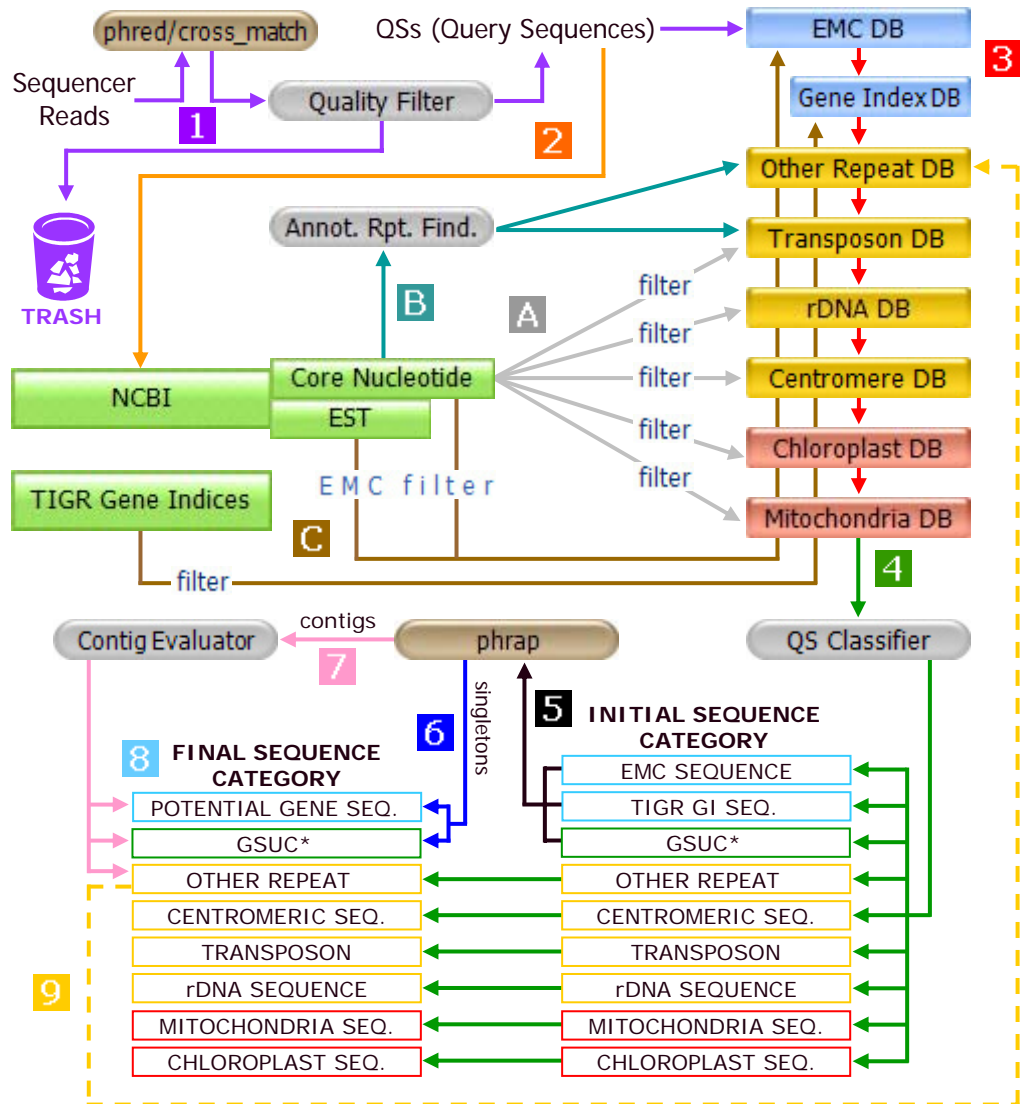


SEQUENCING TO DATE

- 28 million nucleotides of random genomic DNA
- 9.5 million nucleotides of SL DNA
- 22 million nucleotides of TS DNA
- 20 million nucleotides of MR DNA
- 22 million nucleotides of HR DNA

SHOTGUN SEQUENCE ANALYSIS PIPELINE (SSAP)

- Initial *training* of pipeline performed using manually annotated *Sorghum bicolor* CF sequences.
- Assigns each sequence to one of eight different functional/descriptive categories
- Allows classification of hundreds of thousands of sequence reads per week
- Once trained with genomic sequence data it can be used to efficiently calculate gene/repeat enrichment or reduction afforded by CF or other reduced-representation techniques.



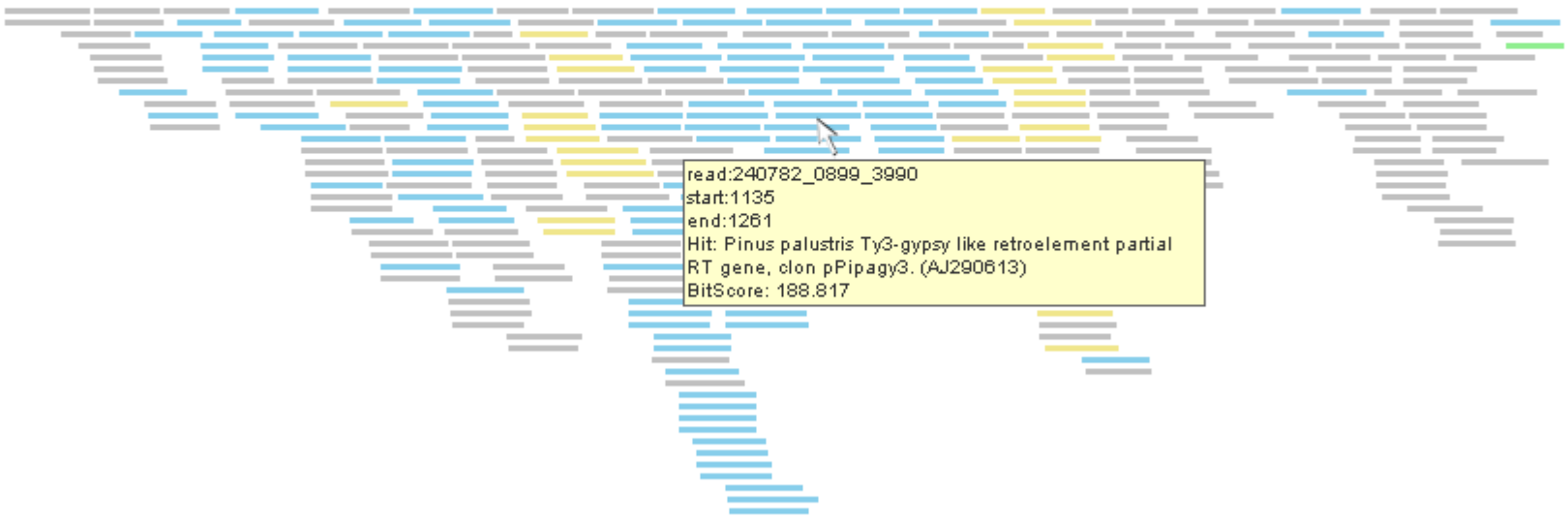
*GSUC = GENOMIC SEQUENCE OF UNKNOWN CHARACTER

INITIAL EXPERIMENT

- Analyzed 28 Mb of random pine genomic DNA (through initial categorization stage)
- Used *Phrap* (default parameters) to analyze sequence
- 28,483 contigs produced

CONTIG CHART

Centromere Chloroplast Mitochondria rDNA Transposon Other Repeats TIGR GIs EMC Undetermined



Contig 39662

2284 bases long

341 query seq. with a total length of 37,354 nt

Almost every sequence classified as a transposon showed its strongest hit to a gymnosperm sequence

CONTIG CHART

Centromere Chloroplast Mitochondria rDNA Transposon Other Repeats TIGR GIs EMC Undetermined



Contig 39501

1480 bases long

78 query seq. with a total length of 8127 bp

0.41% of sequences were recognized as chloroplast DNA. From these we have assembled a rough draft of the loblolly pine chloroplast genome – see J. Carlson's talk (W122) at Forest Trees Workshop

CONTIG CHART

Centromere Chloroplast Mitochondria rDNA Transposon Other Repeats TIGR GIs EMC Undetermined



Contig 39605

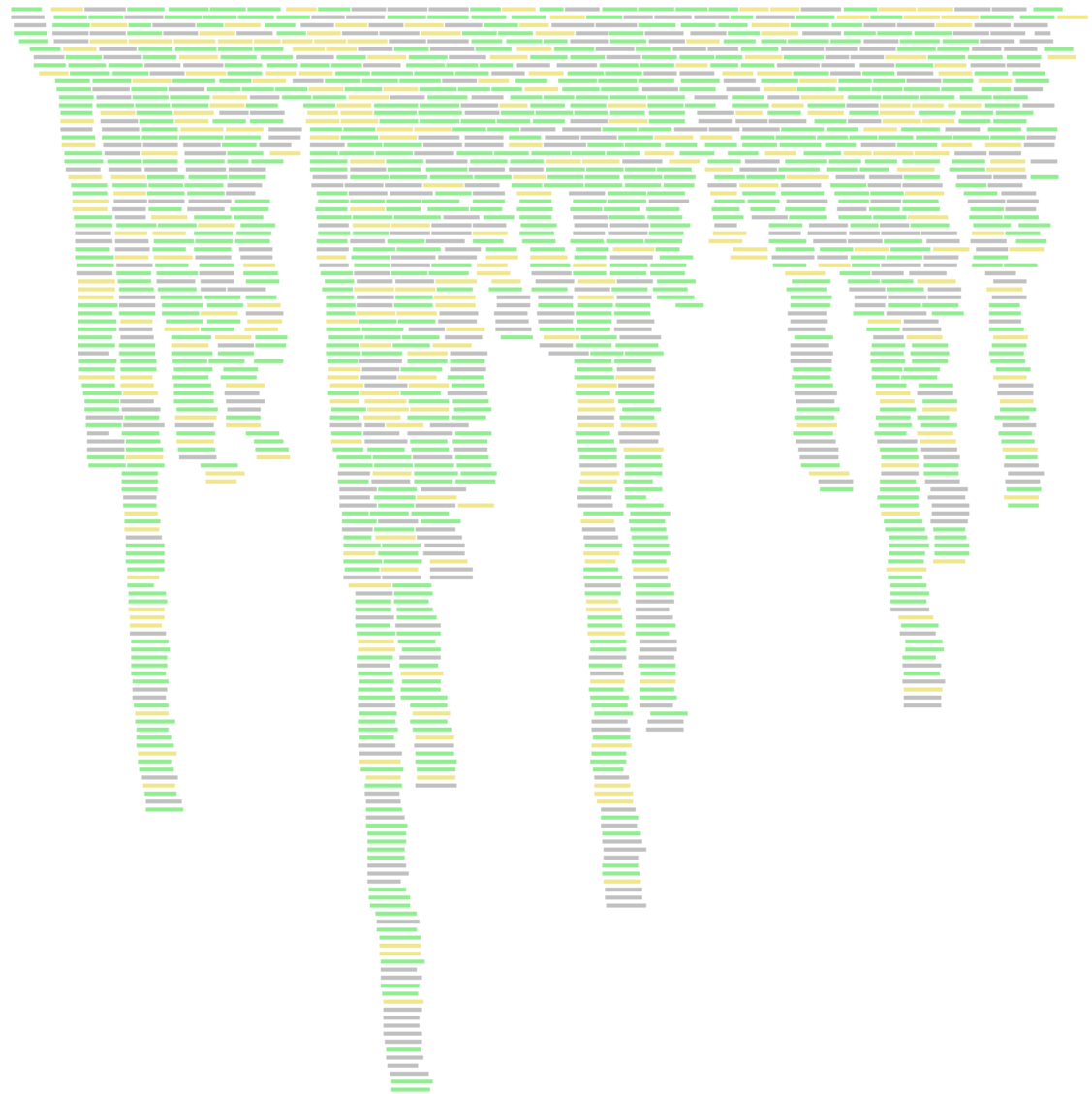
1399 bases long

134 query seq. with a total length of 13,632 nt

Many contigs dominated by *genomic sequences of unknown character* (GSUC)

CONTIG CHART

Centromere Chloroplast Mitochondria rDNA Transposon Other Repeats TIGR GIs EMC Undetermined

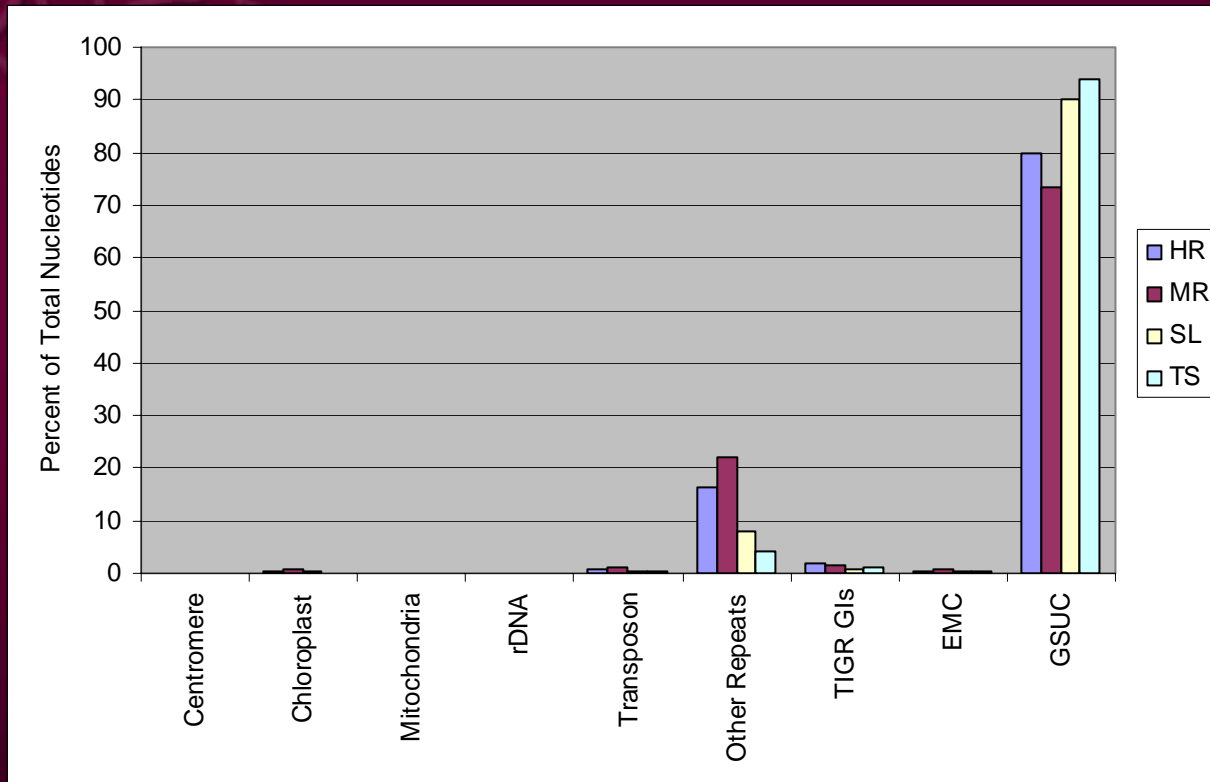


Contig 39667

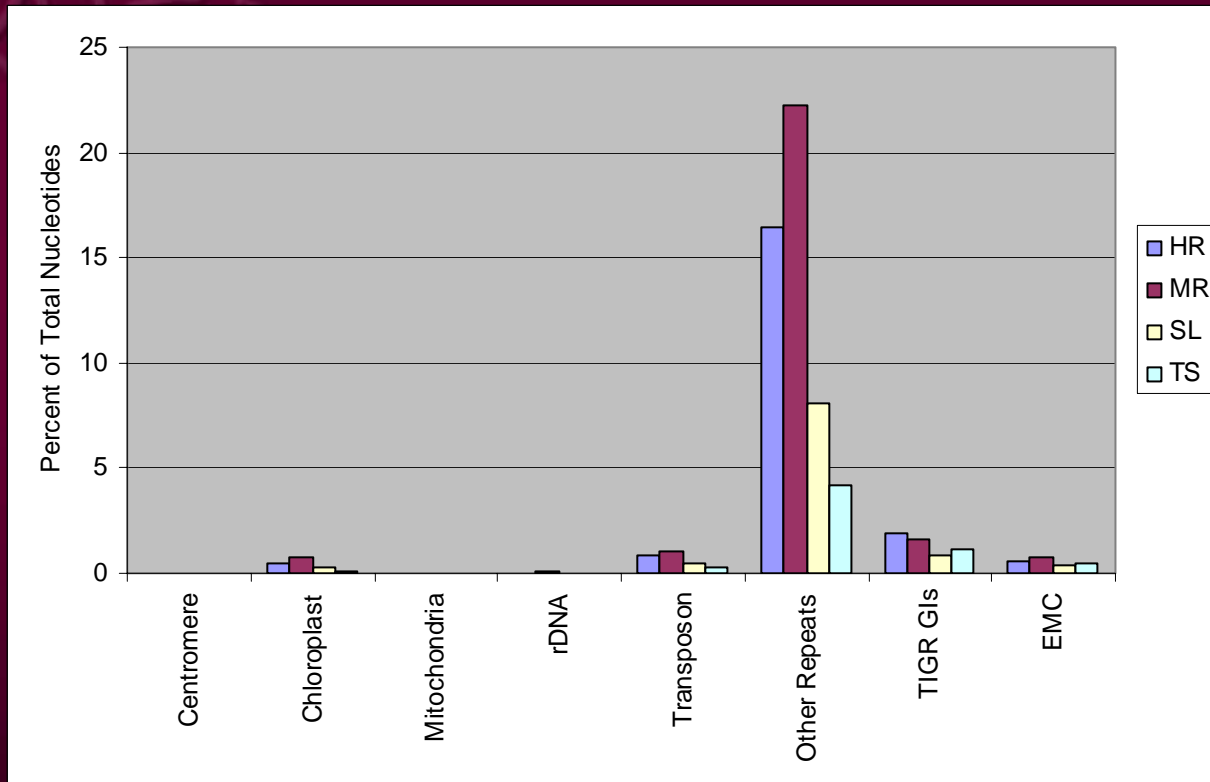
4292 bases long

1531 query seq.
with a total length
of 219,495 bp

TRIAL RUN WITH 5-10K READS FROM EACH CATEGORY



WITHOUT GSUC



IN THE FOLLOWING MONTHS...

- Refine SSAP
- Add HR, MR, SL, and TS sequences into contig analysis
- Analyze contigs showing most abundant hits

IN THE FOLLOWING MONTHS...

- HR, MR, SL, and random genomic DNA clones have been picked and are slated for sequencing via ABI3730. New data will be integrated with 454 sequence.
- Characterize potential retroelements/MITEs (Z. Magbanua)
- Contig, SSAP, and new ABI3730 data should allow study of the evolution of repeat families

CONTIG CHART

Centromere Chloroplast Mitochondria rDNA Transposon Other Repeats TIGR GIs EMC Undetermined



Contig 39666

5898 bases long

837 query seq. with a total length of 90,180 nt

THANKS TO...

- National Science Foundation Plant Genome Research Program – DBI-0421717
- Mississippi State University
- International Paper
- John, Paul, George, and Ringo

