Soybean Rust Genome Sequencing Project

Martha L. Posada-Buitrago[1], Jeffrey L. Boore[1] and Reid D. Frederick[2].
[1] DOE Joint Genome Institute. 2800 Mitchell Drive, Walnut Creek, CA 94598.
[2] USDA-ARS Foreign Disease-Weed Science Research Unit, 1301 Ditto Avenue, Fort Detrick, MD 21702.
mlposada-buitrago@lbl.gov

Soybean rust is caused by two closely related fungal pathogens, *Phakopsora pachyrhizi* and *P. meibomiae.* The Asian soybean rust pathogen *P. pachyrhizi* (ASR) is highly aggressive and is responsible for significant losses of soybean crop in Africa, Asia, Australia and South America (Figure 1). It was discovered for the first time in the continental United States in Louisiana in November 2004. During 2005, the presence of ASR was confirmed in 138 counties across nine southern states. ASR poses a significant threat to the U.S. soybean industry (17 billion dollar anually), depending on the severity and extent of subsequent outbreaks.

Currently, no commercial soybeans are resistant to ASR, and fungicides are generally recognized as the most effective means for controlling the disease. Very little is known about the molecular mechanisms involved in the soybean-rust interaction. In order to develop new strategies to control the disease, it is crucial to increase our understanding of the biology of the pathogen and the infection process.

Here, we present strategies and preliminary results from the *P. pachyrhizi* Genome Sequencing Project, including the complete mitochondrial genome sequence and the comparative analysis of expressed sequence tags (ESTs) generated from four specific-stages of *P. pachyrhizi.*

Figure 1. **A** and **B**. Infected soybean fields in Zimbabwe; **C**. A soybean field treated three times with fungicides, yellow line showing where the fungicide didn't reach the plants (Zimbabwe).

Initial Strategy

Random shotgun libraries:
General 3kb insert size in vector pUC18,
Mid-size 8-10kb insert in vector p21
Fosmid (40kb insert size) in pCC1FOS

cDNA libraries from different stages of *P.pachyrhizi*

Sequencers:
ABI3730, MegaBACE 4000

Informatics:
Reads processing by Phred
Reads assembly by Phrap, Jazz
Verification
Genome annotation

Several independent methods were used to estimate the genome size. Although there were considerable uncertainties associated with most of the methods used, they consistently yielded a genome size above 500 MB.

Table 1. Estimates of *P. pachyrhizi* genome size

Estimation Method	Genome Size
cDNA Coverage	720 Mb
All-Pairs Read Alignment	500-800 Mb
Gene Density	300-700 Mb
Shotgun Fosmid Coverage	600-950 Mb

Trace records (raw single-pass reads of DNA sequence) released by the DOE-JGI, available at the GenBank:
840.789 Mbp

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=170000

Fosmid Sequencing Strategy

Due to the genome's repetitive nature and its large size, the new sequencing strategy focused on random fosmids and selected fosmids. The selection was done by hybridization using probes designed to ESTs selected with high homology to pathogenicity related genes, important metabolic pathway genes or highly expressed genes during spore germination from *P. pachyrhizi.*

Random fosmids: Finishing at Stanford Genome Technology Center (SGTC). Finished 109, Incomplete 6.
Selected fosmids: Hybridizations at Lawrence Livermore National Laboratory (LLNL). Probes designed to 100 "genes". Already 65 selected. Finishing at SGTC: Sequencing 30, Finished 24.

Finished fosmid sequences are available at GenBank.

Mitochondrial Genome

Known mitochondrial genome sequences were compared to the entire set of genomic reads using the Blast algorithm. Potential mitochondrial sequences were assembled with the Phred Phrap Package. This resulted in single contig assembly for the *P. pachyrhizi* mitochondrial genome.

The complete nucleotide sequence of the mitochondrial (mt) genome was determined for *Phakopsora pachyrhizi.* This 32 kb genome contains the genes encoding ATP synthase subunits 6, 8, and 9 (atp6, atp8, and atp9), cytochrome oxidase subunits I, II, and III (cox1, cox2, and cox3), apocytochrome b (cob), reduced nicotinamide adenine dinucleotide ubiquinone oxireductase subunits (nad1, nad2, nad3, nad4, nad4L, nad5, and nad6), the large and small mitochondrial ribosomal RNAs (rrns and rrnl) and tRNAs for all amino acids.

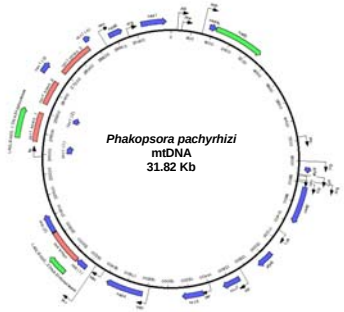

Figure 2. *P. pachyrhizi* complete mitochondrial genome.
Analysis and annotation using DOGMA Dual Organellar GenoMe Annotator (http:// bugmaster.jgi-psf.org/dogma).
tRNAscan-SE 1.21 (http:// www.genetics.wustl.edu/eddy/tRNAscan-SE/).
MacVector 7.1 (Accelrys). Blast algorithm.

Acknowledgements
USDA/ARS/FDWSRU: Christine L. Stone
DOE Joint Genome Institute: Harris Shapiro and Peter Brokstein
Lawrence Livermore National Laboratory: Laurie Gordon
Stanford Genome Technology Center: Jane Grimwood

This research is funded by USDA-ARS, DOE-LBNL and DHS-LLNL

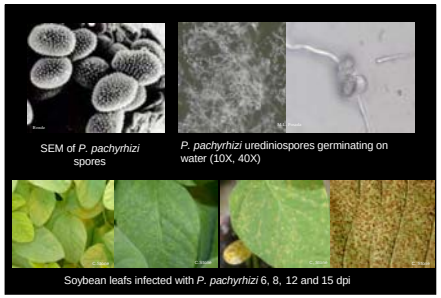

Figure 3. Four *P. pachyrhizi* unidirectional cDNA libraries from different stages were constructed in pSPORT1 (Invitrogen): Germinating urediniospores, resting urediniospores , hyphal growth (6-8 days post inoculation) and high sporulation (13-15 days post inoculation)

The raw sequences were processed by Phred to obtain quality scores. Redundancy was determined by comparing all EST sequences with one another using the BLAST algorithm (Altschul et al, 1997) in the individual libraries and the four libraries combined. The redundant ESTs were assembled by Phrap. The consensus sequences (organized in clusters) were queried against the current non-redundant protein databases ("nr") (EMBL, GenBank, Swissprot) using the BLAST algorithm (Altschul et al, 1997) and run through **Pfam** (Protein families database of alignments and HMMs [hidden Markov models] (The Wellcome Trust Sanger Institute, Cambridge, England) to further classify them into functional categories.

Table 2. Expressed sequence tags' statistics

Library	ESTs	cDNAs	Clusters	Consensus	Singlets
6-8 dpi	5986	4883	1272	1564	1525
13-15 dpi	6385	4248	1565	1926	869
Resting urediniospores	2124	1449	388	495	196
Germinating urediniospores	31442	17096	3112	4749	585
Total	45937	27676	6337	8734	3175

The cDNA clusters were classified into functional categories based on the BlastX hits and the Pfam hits, according to the Expressed Gene Anatomy database (EGAD, TIGR, Rockville, MD). Approximately 23 % of the cDNA clusters from the 6-8 dpi and 13-15 dpi libraries and 40% from the germinating and resting spores libraries show similarity to hypothetical proteins or proteins of unknown function. Several homologs to pathogenesis related proteins (PR proteins) and defense proteins were identified in the infected leaf tissue libraries (Apidaecin, Beta defensin, Thaumatin, etc). In the germinating urediniospores library several homologs to pathogenicity proteins were identified. All the libraries show a high percentage of metabolism related proteins.

The fosmid sequences from *P. pachyrhizi* will allow us to characterize relevant genes, specially those identified in the EST analysis with putative function related to pathogenicity and metabolism, identify regulatory sequences, etc .

The ESTs' consensus sequences were compared to the complete sequenced fosmids using the Blast algorithm and several genes were identified. The genes contained introns with sizes ranging between 50 nt to 150nt. The majority of the introns detected in this study had the canonical **5'GU. . .AG'3** donor-acceptor ss pairs, and the branch site consensus sequence **CURAY**.

The available genomic sequences and the ESTs represent an important collection that provides new molecular markers as microsatellites and SSR for diversity and phylogenetic studies. Further analysis of these data will also provide a new genome size estimation, information on gene and non-coding regions distribution and repeat families.

LBNL-59946 Poster

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.