

Soybean Rust Genome Sequencing Project

Martha L. Posada-Buitrago¹, Jeffrey L. Boore¹ and Reid D. Frederick².

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.

mposada-buitrago@lbl.gov



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Asian soybean rust, caused by *Phakopsora pachyrhizi*, is responsible for significant losses of soybean crop in Africa, Asia, Australia and South America. No commercially available soybean cultivars are resistant to all isolates of *P. pachyrhizi* and very little is known about the molecular mechanisms involved in the soybean-rust interaction. Currently, in areas where soybean rust is present, fungicides are used to control the rust, but they can be expensive. 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*.

Initial Genome Project 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
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)
<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 115 random fosmids and 100 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 11, Finished 17, Unfinished 6. Finished fosmid sequences are available at the GenBank.

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...AG3' donor-acceptor ss pairs, and the branch site consensus sequence CURAY.

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 oxidoreductase 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.

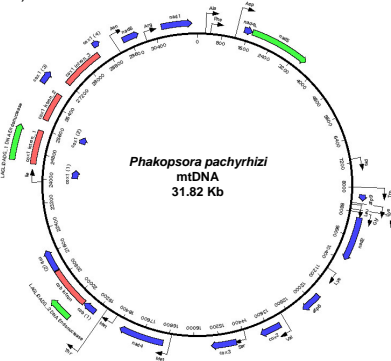


Figure 1. *P. pachyrhizi* complete mitochondrial genome. Analysis and annotation using DOGMA Dual Organellar GenoMe Annotator (<http://bugmaster.jgi-psf.org/dogma>). tRNAscan-SE 1.21 (<http://www.genetics.wustl.edu/eddy/tRNAscan-SE/>). MacVector 7.1 (Accelrys). Blast algorithm.

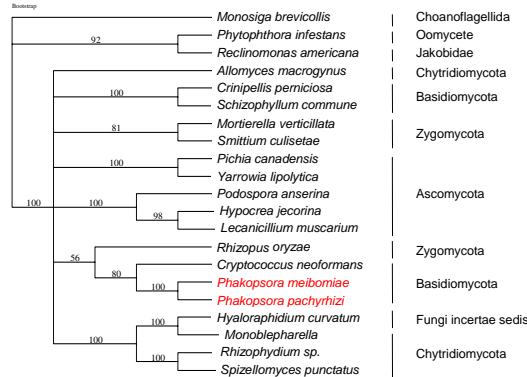


Figure 2. Phylogenetic analysis: 1296 amino acid positions from seven mitochondrial-encoded proteins (cob, cox1, cox2, cox3, nad1, nad4 and nad5) were analyzed for 21 taxa, including 18 species from all fungal phyla and *Monosiga brevicollis*, *Phytophthora infestans* and *Reclinomonas americana* as outgroups. Parsimony-bootstrap support was calculated from 100 replicates using Paup 4.0b10.

This research is funded by USDA-ARS and DOE-LBNL

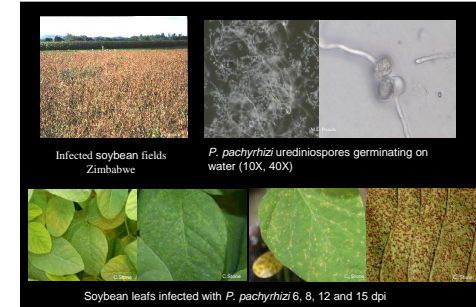


Figure 3. Four *P. pachyrhizi* unidirectional cDNA libraries from different stages were constructed in pSPORT1 (Invitrogen): Germinating urediniospores, resting urediniospores, resting urediniospores, hyphal growth (6-8 days post inoculation) and high sporulation (13-15 days post inoculation)

Table 2. Expressed sequence tags' statistics

Library	ESTs	cDNAs	Clusters	Consensus	Singlets
6-8 dpi	6100	5374	1154	1278	1827
13-15 dpi	6023	4610	1291	1387	1356
Resting urediniospores	2295	1762	393	455	335
Germinating urediniospores	29601	18638	2686	3394	2142
Total	44019	30244	5524	6514	5660

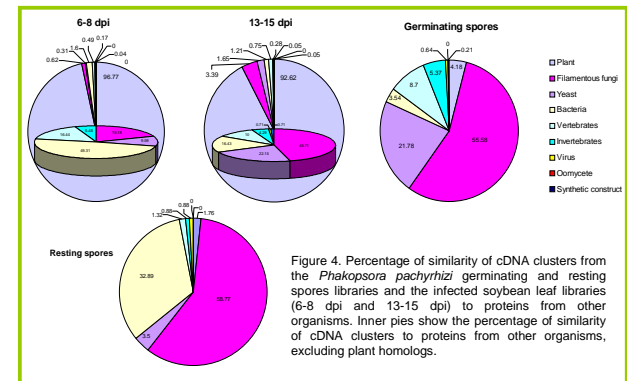


Figure 4. Percentage of similarity of cDNA clusters from the *Phakopsora pachyrhizi* germinating and resting spores libraries and the infected soybean leaf libraries (6-8 dpi and 13-15 dpi) to proteins from other organisms. Inner pies show the percentage of similarity of cDNA clusters to proteins from other organisms, excluding plant homologs.

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 (Apidaein, 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.

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