Comparative analysis of expressed sequence tags (ESTs) generated from four specific-stages of *Phakopsora pachyrhizi*

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

The Asian soybean rust pathogen *Phakopsora. pachyrhizi* (ASR) is highly aggressive and is responsible for significant losses of soybean crop in Africa (Figure 1), Asia, Australia and South America. 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 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).

MATERIALS AND METHODS

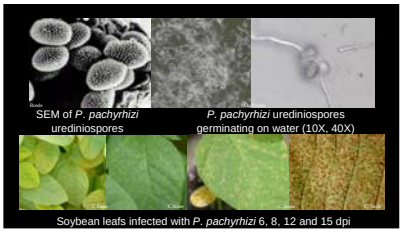
Fungal Isolate and Growth Conditions

P. pachyrhizi isolate Taiwan 72-1 was maintained on *Glycine max* cv. Williams at the USDA-ARS Foreign Disease-Weed Science Research Unit Plant Pathogen Containment Facility (BSL-3) at Ft. Detrick (Maryland, USA). Spores were germinated in distilled water for 16 hours in Pyrex dishes at RT (Figure 2A). The germinating spores were collected and frozen in liquid nitrogen. 14-17 day old soybean plants (*Glycine max* cv. Williams) were inoculated with *P. pachyrhizi* (isolate TW 72-1) urediniospores (75000spores/plant), kept in a dew chamber for 24 hours at 20 C. The plants were transferred to the greenhouse and kept at 25 C, 16/8h light. Infected leaf tissue was collected for 6, 7, 8, 13, 14 and 15 days post inoculation (dpi) (Figures 2B and 2C) and stored at –80 C).

cDNA Library Construction

mRNA was isolated from resting urediniospores, germinating urediniospores, infected leaf tissue 6-8 dpi and 13-15 dpi using the RNeasy and OLIGOTEX mRNA purification kits (Qiagen). Unidirectional cDNA libraries were constructed in plasmid pSPORT1 using the Superscript Plasmid System for cDNA Synthesis and Cloning Kit (Invitrogen).

Figure 2. Four *P. pachyrhizi* unidirectional cDNA libraries from different specific-stages were constructed in pSPORT1 (Invitrogen): Resting urediniospores, Germinating urediniospores, Hyphal growth (6-8 days post inoculation) and High sporulation (13-15 days post inoculation)



DNA Sequencing

Sequencing reactions were analyzed using an ABI Applied Biosystems 3730 DNA analyzer.

Data handling

The raw sequences were processed using the JGI EST Pipeline. Phred software was used to call the bases and generate quality scores. Vector, linker, adapter, poly-A/T, and other artifact sequences are removed using the Cross_match software (Ewing and Green 1998; Ewing et al. 1998), and an internally developed short pattern finder. Low quality regions of the read are identified using internally developed software which masks regions with a combined quality score of less than 15. The longest high quality region of each read is used as the EST. ESTs shorter than 150 bp are removed from the data set. ESTs containing common contaminants such as E. coli, common vectors, and sequencing standards are also removed from the data set. EST Clustering is performed ab-initio, based on alignments between each pair of trimmed, high quality ESTs. Pair-wise EST alignments are generated using the Malign software (Chapman, et. al., Unpublished), a modified version of the Smith-Waterman algorithm (Smith and Waterman, 1981), which was developed at the JGI for use in whole genome shotgun assembly. ESTs sharing an alignment of at least 98% identity, and 150 bp overlap are assigned to the same cluster. These are relatively strict clustering cutoffs, and are intended to avoid placing divergent members of gene families in the same cluster. ESTs that do not share alignments were assigned to the same cluster, if they were derived from the same cDNA clone.

EST cluster consensus sequences were generated by running the Phrap software (Ewing and Green 1998; Ewing et al. 1998) on the ESTs comprising each cluster in the individual libraries and the four libraries combined. The consensus sequences and the singlets from each library and the four libraries combined 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. The consensus sequences were also compared to the available fungal protein and EST databases, as well as the soybean ESTs available at the GenBank using the BLAST algorithm (Altschul et al, 1997) . All BLAST algorithms were run on March 2/2006

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

Table 1. 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
Ppac (all four libraries)	45937	27676	5801	8156	2730

Table 2. BLAST hits to different databases

	Percentage of consensus sequences blastx hits												Percentage of consensus sequences blastn hits		
Library (# Consensus+singlets)	nr	PHI	AN	BC	CG	FG	MG	NC	PC	RO	SC	SN	UM	est_others	COGEME	soybean ESTs
6-8 dpi (3091)	84.6	3.9	32.9	31.9	31	34.5	32.6	32.9	33.3	35	28.2	33.2	32	93.9	3.3	91.1
13-15 dpi (2795)	78.2	5.1	30.8	29.8	32.1	31.6	31.4	31.9	32	26.8	31.6	29.9		86.7	3.3	83.1
Germinating urediniospores (5334)	39.6	6.1	37.3	38.4	36.9	38.6	39.1	38.5	38.5	37	39.6	39.3	40.6	11.5	2.7	1.34
Resting urediniospores (691)	38.2	3	24.4	24.4	23.4	25.3	25.5	24.7	27.3	21.8	18.9	24.4	26.2	31.1	1	0.05
Ppac (10886)	64.9	5.3	33.9	34.2	33	35.2	34.3	34.5	39.91	34.5	27.9	35	35.1	29.5	2.1	42.7

BLAST hits with expected value E < 1e-5 for proteins and E< 1e-35 for nucleotides (ESTs) were considered to show significant similarity to the entries in the different databases.Protein databases: "nr": non-redundant protein database (GenBank). PHI: PHI-base Pathogen Host Interactions protein database (http://www.phi-base.org); PC: Phanerochaete chrysosporium (JGI: http://genome.jgi-psf.org/Phchr1/Phchr1.info.html). AN: Aspergillus niger; BC: Botrytis cinerea, CG: Chaetomium globosum, FG: Fusarium graminearum, MG: Magnaporthe grisea, NC: Neurospora crassa, RO: Rhizopus oryzae, SC: Sclerotinia sclerotiorum, SN: Stegonospora nodorum and UM: Ustilago maydis from the Broad Institute (http://www.broad.mit.edu/annotation/fgi/). EST databases: "est_others" (ESTs from organisms other than human and mouse GenBank). COGEME: Phytopathogenic Fungi and Oomycete EST Database Version 1.5 (http://cogeme.ex.ac.uk/). Soybean ESTs (GenBank).

Most of the hits to the fungal protein databases correspond to hypothetical, predicted and unnamed proteins, or proteins of unknown function. New revisions on the finished and ongoing fungal genome projects and databases will provide invaluable information and tools for the study of new and challenging fungi, such as obligate parasites as *Phakopsora pachyrhizi*

The 5801 clusters identified shared similarity to a wide variety of other organisms (Fig. 3). The infected leaf libraries show a very high percentage of putative plant genes, and the difference between them is due to the presence of more *Phakopsora pachyrhizi* mycelia in the 13-15 dpi tissue, also shown by the percentage of hits to the soybean EST database. The germinating spores (GS) library shows a high percentage of filamentous fungi and yeast putative genes, 55.58% and 21,78 % respectively. The GS library also shows a higher percentage of similarity to vertebrate and invertebrate genes than to plant genes. The resting spores library (RS) also shows a higher percentage of fungal putative genes.

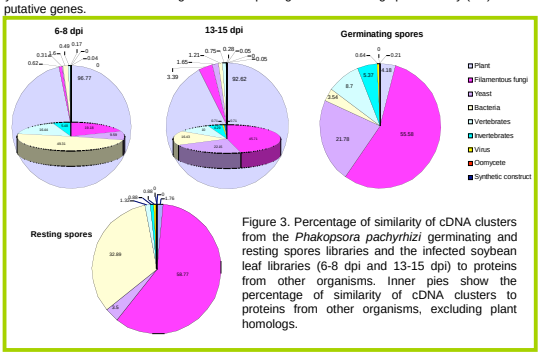


Figure 3. Percentage of similarity of cDNA clusters from the *Phakopsora pachyrhizi* germinating and resting spores libraries and the infected soybean leaf libraries (6-8 dpi and 13-15 dpi) to proteins from other organisms. Inner pies show the percentage of similarity of cDNA clusters to proteins from other organisms, excluding plant homologs.

When clustered together (Ppac), the four libraries shared only 4 clusters. The germinating spores (GS) and the resting spores (RS) library shared 169 clusters; the GS and the 6-8dpi shared 26 clusters, the GS, the 13-15dpi and the GS shared 99 clusters, and the 6-8 and the 13-15dpi libraries shared 444 clusters. The low number of clones shared by the four libraries is most likely due to the little mass of mycelia present in the infected leaf, specially the 6-8 dpi time point. The high number of cDNAs shared by the infected leaf libraries are related to house-keeping genes, and not surprisingly the most redundant clones in the infected leaf libraries show similarity to genes related to photosynthesis.

The ESTs' consensus sequences and singlets 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.

CONCLUSIONS

1. Only the 39.6% of the cDNA consensus sequences/singlets from germinating spores and the 38.2 % from the resting spores conseq/singlets appeared to be related to previously characterized genes. The 84.6 % and 78.2 % of the cDNA clusters from infected leaf tissue from 6-8 dpi and 13-15 dpi appeared to be related to previously characterized genes.
2. Putative functions such as primary metabolism (amino acids, carbohydrates, lipids), RNA and protein metabolism, cell structure, growth and morphology, cell division, stress response, transport and permeases, DNA repair and secondary metabolism could be assigned to many cDNA clusters based on similarity to sequences in the non-redundant protein databases and the Pfam (Protein families database of alignments and HMMs.
3. The relative low percentage of BlastX hits and the high percentage of hypothetical proteins or proteins with unknown function in the germinating and the resting spores libraries indicate the limited amount of information on fungal gene expression and genomics and show the urgent need for more studies.

Acknowledgements

USDA/ARS/FDWSRU: Christine L. Stone

DOE Joint Genome Institute: EST Pipeline Developer Peter Brokstein

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.