Discovery of novel transcripts and transcriptionally active *Phytophthora* specific genes using RNASeq data

Sucheta Tripathy[1], Lecong Zhou[1], Matt Dyer[2], Brett M. Tyler[1]

1 Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA-24061; 2 Applied Biosystems, Inc.

RNAseq technology provides an accurate and reproducible way of measuring gene expression data. It also has numerous other applications including transcript/isoform discovery, finding copy number variation, splice variants and small non coding RNA discovery. We present here the RNAseq analysis results of *P.sojae* mycelial and Soybean tissue infected with *P.sojae* libraries.

We have analyzed RNAseq data from *P.sojae* mycelia with 4 experimental replicates each having an average of approximately 18 million reads. We used Bowtie short read aligner for initial alignment of reads to the genome sequences. Our attempts of including Tophat junction mapper for predicting junctions failed because of the read length; Tophat is optimized for paired end reads of length >=75 bp. As a work-around to determine how many reads were lost in the junction regions, we mapped the reads to the predicted transcripts and the unigenes derived from *P.sojae* EST libraries. We finally assembled the mapped reads to the genome sequence using Cufflink. Approximately 55% of the reads mapped to the genome sequence, whereas only 45% could be aligned to the predicted transcripts. The 10% reads that did not have a match with the transcripts were assembled to form putative exons. We analyzed these exons and studied their expression patterns in *P.sojae* and *P.sojae* infected Soybean libraries. We discovered novel transcripts represented by ~3000 exons. We found 17 new crinklers, out of which 9 are isoforms of the existing ones and 8 are new. The most highly expressed exons that are not present in the existing gene models are found to be unique to *Phytophthora sojae* with no known similarities within NCBI's nr database. We have also curated the non-coding RNAs from the RNAseq data. Our transcriptomics database will temporarily store the assembled transcripts and raw aligned reads. Our browser is being upgraded to accommodate RNASeq viewing as well as text alignment viewing. Our transcriptomic database is available at http://est.vbi.vt.edu/

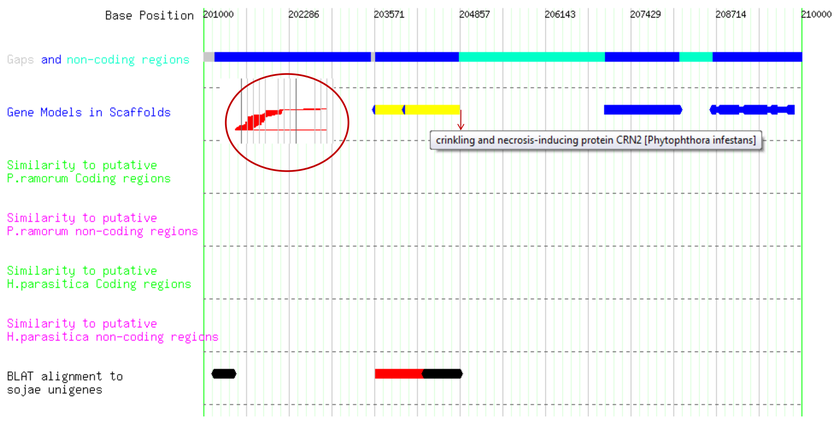
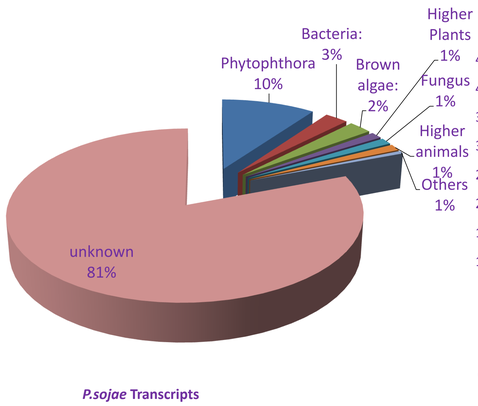


Fig – 1 : Missing crinkler gene in the present annotation in scaffold_44

Location	FPKM for PS library	FPKM for Infection Library	Transcript New or Isoform
scaffold_97:152408-152943	501.91	105.949	**New**
scaffold_313:1-196	183.163	42.0011	Isoform
scaffold_44:201166-201421	-	28.9385	Isoform
scaffold_13:348696-348891	-	19.9609	Isoform
scaffold_13:665370-665565	-	19.9609	Isoform
scaffold_7:3188-3433	16.488	5.95	Isoform
scaffold_1279:1141-1320	-	5.43	**New**
scaffold_1279:38-109	-	2.28	**New**
scaffold_48:53756-53804	-	5.06	Isoform
scaffold_95:41169-41356	-	3.469	**New**
scaffold_52:89052-89163	-	2.922	Isoform
scaffold_474:9385-9446	-	2.658	**New**
scaffold_70:91178-91299	-	1.3	**New**
scaffold_44:201165-201440	158.063	28.9385	Isoform
scaffold_9:225429-225479	3.19	-	**New**
scaffold_971:6679-6794	2.08443	-	Isoform
scaffold_189:22834-22926	**1.737**	-	**New**

Table – 1: Expression level of crinklers in *P. sojae* mycelial and infection libraries



P.sojae Transcripts

Fig – 2 : *P.sojae* expressed transcripts with similarities to different organisms

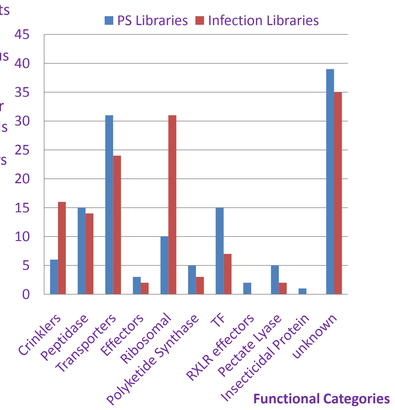


Fig – 3 : Comparative account of functional genes in mycelial and infection libraries

Results:

We assembled the reads that mapped to the *P. sojae* V1.0 genome into putative coding regions. After removing the overlaps with the existing annotations, we created a list of 3000 new exons. We co-ordinate merged these exons from different libraries into consensus sequences. Blastx of these novel exons with nr database reveals most of the highly expressed genes are conserved sequences and unique to *Phytophthora sp*. [fig -2]. Among several functional categories from infection as well as mycelial libraries, there was a overall increased number of crinkler genes in infection libraries [fig-3]. Three of the crinklers in mycelial libraries had several fold higher level of expression than crinklers from infection libraries [Table-1]. From the annotated dataset, senescence related protein(VMD id: 125126) had the highest level of expression followed by translocase (109355), followed by Avr1a(127179). Rest of the crinklers in infection libraries had moderate levels of expression.

Discussion:

Absence of highly expressed house keeping genes from the present annotation indicates the inability of gene finders in predicting organism specific gene models when trained with a related organism other than *Phytophthora*. This further indicates that the pool of highly expressed genes are highly conserved and does not share compositional and signal information with even related organisms. Expression profiling in mycelial and infection tissue reveal higher level of expressions in several infection and senescence related proteins, as expected. However, there is consistently higher level of expression in few crinklers in mycelial libraries than that of infected libraries which still remains to be addressed.

Acknowledgement:

This work was supported by grants from the National Research Initiative of the USDA National Institute of Food and Agriculture, number 2007-35600-18530 and from the US National Science Foundation, number MCB-0731969