FUNCTIONAL GENOMICS OF THE GRAPE-*XYLELLA* INTERACTION: TOWARDS THE IDENTIFICATION OF HOST RESISTANCE DETERMINANTS (CUMULATIVE PROGRESS REPORT)

Project Leaders:
Douglas R. Cook
Francisco Goes-da-Silva
Hyunju Lim
Department of Plant Pathology
University of California
Davis, CA 95616

Cooperators:
M. Andrew Walker
Alberto Iandolino
Department of Viticulture and Enology
University of California
Davis, CA 95616

Reporting Period: The results reported here are from work conducted from April 2001 to October 2003.

INTRODUCTION

Pierce's disease (PD), caused by the bacterial pathogen *Xylella fastidiosa* (*Xf*), is one of the most destructive diseases of grapevines (Purcell and Hopkins, 1996). All genotypes of *Vitis vinifera* are susceptible to the PD pathogen and only certain non-vinifera species (e.g., *V. arizonica*), typically not suitable for wine production, are able to resist or tolerate this pathogen. Development of resistant varieties through classical breeding is complicated by the desire to retain varietal phenotypes in cultivated species, and by the generally poor agronomic properties (e.g., fruit quality) of these non-vinifera species. An alternative approach for developing disease resistant germplasm is to characterize the molecular basis of resistance and susceptibility in *Vitis* species, and to use this information to design rational strategies for crop protection.

In this project we are pursuing a genomics approach to identify transcriptional pathways that are correlated with susceptible or resistant interactions in *Vitis* species. The comparison of these two distinct interactions should reveal functional elements of the host resistance response, or conversely host functions that confer susceptibility. Such information will considerably increase our knowledge of the grape-*Xylella* interaction and potentially provide the basis for developing resistance to the PD pathogen in *V. vinifera*. A side benefit of these activities will be information that is either of direct relevance to Pierce's Disease (e.g., identification of novel *Xylella*-responsive promoters for gene regulation, improved molecular markers for traditional breeding of PD resistance, and alternative diagnostic methods based on host gene expression), or that is enabling to grape improvement generally (public EST databases, genome-wide molecular markers (e.g., SSR and SNPs), and a public grape oligonucleotide microarray).

OBJECTIVES

1. Construct and archive cDNA libraries from (a) infected and non-infected grape genotypes representing both susceptible *V. vinifera* and resistant/tolerant *Vitis* spp., and (b) a range of tissues and developmental stages to increase gene discovery.
2. Conduct sequencing reactions for 60,000 cDNA clones and submit cleaned, high quality sequence reads to the National Center for Biotechnology Information (NCBI).
3. Develop an analysis pipeline and web-accessible relational database for the grape transcriptome, including an annotated unigene set and the identification of candidate *Xylella*-induced genes.
4. Conduct transcriptional profiling to characterize host gene expression in susceptible and resistant/tolerant grape-*Xylella* interactions.

RESULTS AND CONCLUSIONS

A. EST sequencing

Susceptible *Vitis vinifera*. 120 plants of Cabernet Sauvignon were sampled over a six month period to identify infected and non-infected individuals. Tissue collected from these plants was subject to mRNA extraction, cDNA library construction, and sequencing of Expressed Sequence Tags (ESTs). Because symptom development in *Vitis vinifera* is a function of both bacterial infection and plant development, we sequenced cDNAs from plants collected early in the growing season when the pathogen could be detected but symptoms were not evident, as well as late in the growing season when symptoms were evident on infected plants. Eight cDNA libraries and >30,000 sequence reads were generated by the project.

Resistant *Vitis* germplasm

Working with Dr. Andrew Walker of the Department of Viticulture and Enology at UC Davis, we obtained tissue from resistant progeny of a cross between *Vitis rupestris* and *Vitis arizonica*. Two cDNA libraries were prepared, picked into 384-well plates, and archived at -80°C. 10,000 cDNA sequences were obtained and analyzed to assess variation between the *rupestris/arizonica* cross and the more extensive dataset of *Vitis vinifera* (see below).

Continued sequencing in *Vitis vinifera*

Nine additional cDNA libraries were constructed from *Vitis vinifera*, with the goal of expanding the number of characterized genes available for development of an oligonucleotide microarray. In total, 46,493 sequencing reactions were completed from these cDNA libraries, resulting in the submission of 36,966 sequences to the National Center for Biotechnology Information EST database.

Table 1. Overview of the UC Davis *Vitis* sequencing project: 2001-2003.

Tissue	Total	Submitted	Funding
infected leaves	13206	9590	APHIS
healthy leaves	12764	8114	APHIS
berries*	21170	15237	ARS
cmpd bud	4992	4362	ARS
flower**	9984	8176	APHIS/CDFA
stem	5355	4700	CDFA
petiole	4992	4491	CDFA
Vitis spp	10752	6533	CDFA
Total	**83215**	**61203**	

*Berry libraries: Pre-veraison, Stage I; Pre-veraison, Stage II; Veraison; Post-veraison.
**Flower libraries: pre-anthesis and post anthesis.

B. Bioinformatics

We developed a laboratory information management system (LIMS) database in Oracle 8 to organize sequence information. The associated chromatogram pipeline downloads information from our ABI 3730XL sequencer nightly, processes chromatograms through a standard QC pipeline consisting of the PHRED and X-match algorithms, and generates NCBI submission files. Cleaned and trimmed FASTA format files are submitted to the NCBI GenBank EST database. We subsequently retrieve NCBI accession numbers and store the hyperlink in our Oracle database so that chromatograms can be automatically traced to the public record.

We have also developed an on-line Oracle relational database for analysis of the complete set of public *Vitis vinifera* EST data. The database begins with a contig assembly pipeline consisting of MegaBLAST and CAP3. The following analyses are performed on all EST singletons and contigs: (1) BLASTX against the NCBI non-redundant database to assess protein-coding similarity, (2) BLASTN against all *Vitis* data to identify close homologs (potential orthologs) from other *Vitis* species, and (3) identification of potential polymorphic sites as simple sequence repeats, including automated oligonucleotide primer design for genetic mapping. We have developed a controlled vocabulary that organizes genes according to three ontologies: (I) genotype (cultivar or species), (II) development (e.g., berries pre-veraison, or flower pre-anthesis), and (III) stress (e.g., biotic *Xylella fastidiosa*, or abiotic drought). A related query tool permits users to query the data for genes expressed in grape under specific conditions of genotype, development or stress. For example, one can construct a search to identify genes whose expression is enhanced in *Xylella*-infected tissue, or expressed in Stage II of berry development, or in pre-anthesis flowers. The database can also be queried based on unique identifiers (e.g., GenBank accession numbers), key words (e.g., chalcone synthase), at the nucleotide level based on BLASTn, or for specific simple sequence repeat motifs.

Because it is important to view all *Vitis* data simultaneously, we routinely download public *Vitis* data and incorporate them into our assemblies. As of October 15, 2003, the NCBI dbEST contains 135,601 *V. vinifera* sequences. The current *Vitis* unigene set consists of >25,000 predicted transcripts, including ~12,500 contigs. An important feature of many of these contigs is that they contain paired end reads from single clones, where sequence reads overlap. This feature of paired end reads is a consequence of our strategy to sequence both ends of every cDNA clone. The results of this strategy are (1) accurate 3' data to facilitate design of microarray oligonucleotides, and (2) experimental validation of transcript structure, with rapid identification of full-length or near-full length cDNA clones.

The current version of our *Vitis* EST database (http://cgf.ucdavis.edu) is based on the analysis of ~84K sequences available as of May 1, 2003. Our programmers are in the process of analyzing the October 1, 2003 data freeze that will constitute the basis of the Affymetrix microarrays. This new build of the database will include ESTs from the following species: 135K ESTs from *V. vinifera*, 6,500 ESTs from a genetic cross of *V. arizonica x V. rupestris*, and several thousand ESTs from a range of related *Vitis* species.

C. Genetic variation in Vitis: Analysis of diversity at the sequence level

One question that was raised by reviewers of our original proposal was whether EST data obtained from one variety of grape would be sufficiently representative of *V. vinifera* to develop a common microarray for use across the species. Similarly, reviewers questioned whether sequence from *V. vinifera* would be suitable for developing microarrays that would function in other *Vitis* species, in particular those with "resistance" to Pierce's Disease. To address these questions, we conducted sequence similarity analyses both within *V. vinifera* (between varieties) and between *V. vinifera* and other *Vitis* species.

Variation between *Vitis vinifera* varieties

We analyzed the public data available as of May 13, 2003, composed of 37,902 ESTs from Cabernet Sauvignon (this project), 40,437 ESTs from Chardonnay, ESTs from 8,191 Shiraz, and 1,157 ESTs from other *Vitis vinifera* varieties. The genome of all grapes is highly heterozygous, and therefore allelic variation is common even within a single variety. To quantify allelic variation we conducted multiple sequence alignment of the public ESTs to give us the first detailed glimpse of genetic variation at the species level. In particular, we analyzed the structure of contigs that contained sequences from at least two different varieties, and where each variety was represented by multiple sequences within the contig. This practice allowed us to distinguish random sequencing error from systematic base pair variation (i.e., variation present in multiple sequences within a contig). Such systematic sequence variation is presumed to represent single nucleotide polymorphism (SNP), the most frequent type of natural genetic variation found within species. Two general themes emerged from this analysis. First, the majority of sequences from Cabernet Sauvignon, Chardonnay and Shiraz are 100% identical, and (when present) the level of allelic variation observed within a varietal genome is not substantially higher than variation between genomes (typically >99% identity).

Variation between *Vitis* species

To address the issue of DNA sequence variation between *V. vinifera* and other *Vitis* species, we compared the following data sets:

***V. vinifera* x *V. aestivalis*.** On a contract basis, the CA&ES Genomics Facility sequenced 2,500 cDNAs of *Vitis aestivalis*. These data were deposited at NCBI and were therefore available for comparison of sequence conservation with the much deeper collection of *Vitis vinifera* sequence information. BLASTn analyses and the corresponding pairwise alignments for these two species revealed that the vast majority of sequences possess identity in the range of 98-100%, indicating that the transcripts of these two genomes are remarkably well conserved.

***Vitis vinifera* and *V. arizonica/V. rupestris*.** A similar analysis was conducted on the contig information generated by our *V. arizonica* X *V. rupestris* sequencing effort. The vast majority of comparisons to *Vitis vinifera* have e-values of zero and sequence identities in the range of 98-100%.

The high levels of sequence similarity between *V. vinifera* and other *Vitis* species suggests that it will be possible to produce oligonucleotide arrays with comparable specificity for all of these genomes, using the *Vitis vinifera* data as the reference. Such cross-species use of oligonucleotide arrays are not uncommon in the genomics community, as recently applied in the comparison of human-chimps-monkeys and mice.

Identification of single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) in the grape genome(s)
Allelic variation is the basis of many types of genetic studies, including association genetics and genetic mapping, and a methodical listing of allelic variation should have utility for current projects that aim to map traits for PD resistance. The *Vitis vinifera* GeneIndex contains a SNP (single nucleotide polymorphism) link for all contigs where SNPs have been identified, and technologies such allele specific oligonucleotide (ASO) primer extension can be used to rapidly convert this information into gene-specific genetic markers. Equally useful, however, is a class of genetic variation termed Simple Sequence Repeat (SSR) markers. Thus, we have implemented an SSR identification pipeline to mine the grape unigene set for SSR motifs. The data is organized in an Oracle database, and is accessible via the "http://cgf.ucdavis.edu" web site. Both of these data types are collateral benefits of our funded EST project and should impact many areas of grape improvement, including the genetic analysis of PD resistance.

D. Analysis of the grape transcriptional response to pathogen challenge

The analyses described below are based on the analysis of combined data sets generated under this project and that of our collaborators at the University of Nevada-Reno, and other members of the grape genomics community. In total, 40% of the 135K *V. vinifera* ESTs and 100% of the sequencing focused on Pierce's Disease originated from this project.

Identification of *Xylella*-induced genes in *Vitis vinifera*.

We have identified several (~35 in total) genes that appear to be up-regulated in response to infection by *X. fastidiosa* (see table below). The most abundant contig (1007061) shares homology with a stress-related RNA from Arabidopsis, although the function is unknown in any system. Interestingly, this gene is upregulated in infected plants, prior to symptom development, making it a top candidate for an early and sensitive marker of Pierce's Disease. Some genes in the list have homology to proteins implicated in signaling during disease resistance, while others have been identified as pathogen responsive, or have been implicated in plant-insect interactions. After confirmation of the *Xylella*-specific transcription of such contigs (see below) we are poised to isolate the promoters from these genes from genomic DNA libraries. The potential application of such promoters to drive *Xyella*-induced and/or tissue specific expression of transgenes is planned as a topic of a future grant proposal.

Development of Real-Time reverse transcriptase PCR for gene expression analyses and improved diagnostic tools for pathogen detection
Ultimately, detailed analysis of transcriptional responses will require methodical analysis by means of microarray studies, to be initiated in 2004. At the same time, the current list of putatively *Xylella*-induced genes may provide leads for further analysis by means of real time reverse transcriptase PCR. We are particularly interested in developing host transcripts that can serve as markers of *Xylella* infection; such markers may be more sensitive (e.g., expressed systemically in locally-

infected plants) than pathogen-based PCR primers for diagnosis of Pierce's Disease. This strategy has been referred as "transcriptional fingerprinting", and is considered to hold great potential for diagnostic analyses.

We have generated primers and probes for real-time reverse transcriptase PCR analysis of 20 candidate *Xylella*-induced transcripts. These primers and probes are being used to monitor gene expression of infected and non-infected plant samples collected from the Napa Valley of California in the months of July and September, 2003 and ultimately from plants grown in growth chamber conditions (a more controlled environment). Results from these analyses will be presented.

Transcriptional profiling

In grapes the development of transcriptional profiling tools that function broadly across all species of interest (i.e., susceptible *Vitis vinifera* and resistant *Vitis* spp.) has the potential to impact grape improvement on multiple fronts. In the case of Pierce's Disease we can anticipate the following outcomes: (1) transcriptional fingerprints for diseased plants may facilitate more reliable diagnosis, e.g., by detecting systemic responses in the host, (2) critical evaluation of the physiology of the host during symptom expression, examining long-standing but untested hypotheses such as (a) the relationship between water stress and disease, (b) the relationship between host development and disease symptom expression, or (c) how source-sink relationships change in infected versus healthy plants during fruit development, and (3) identifying transcriptional response pathways that are correlated with disease resistance, tolerance or susceptibility. Many of the genes induced in resistant interactions may be causal to resistance, for example the induction of insecticidal or anti-bacterial proteins and peptides, or the induction of host secondary metabolic pathways leading to anti-bacterial or insecticidal compounds, would be prime candidates for the development of resistance to *Xylella fastidiosa* in *Vitis vinifera*.

We anticipate having Affymetrix microarrays available in February 2004. Using the Affy arrays, we will initiate experiments to address the following questions: (1) which host genes are induced during resistant, tolerant and susceptible interactions with *Xylella fastidiosa*; (2) what is the relationship between water stress (drought) and infection by *Xylella fastidiosa*?

Genome-wide identification of transcripts in *Vitis vinifera*

As a preliminary effort towards transcriptional profiling we have used statistical tools to analyze gene expression in the public grape EST datasets. Our initial analysis focused on the 84,000 ESTs available in the National Center for Biotechnology Information (NCBI) EST database (DbEST) as of May 2003. These expressed sequence tags (ESTs) were representative of different cultivars, different stages of plant development, and tissue exposed to biotic (e.g., infection by *Xylella fastidiosa*) and abiotic (e.g., drought, salt and cold) stress factors. Correlation analysis was used to select 2821 contigs (preidcted genes) for clustering based on Euclidian Distance and Principal Component Analysis. As expected, both types of analyses revealed that gene expression profiles were highly predictive of plant development. Thus, gene expression profiles resolved leaves, from roots, from berries. Moreover, developmental stage was also highly correlated with gene expression, such that young leaves clustered together and separate from old leaves, while pre-veraison berries from Stage I and Stage II had were more similar to one another than to veraison or post veraison berries, etc.

In terms of Pierce's Disease, there are three potentially important trends in the data. First, the statistical analysis confirmed our previous selection of genes correlated with Pierce's Disease (described above and in the Figure below). Second, transcripts that cluster with putative *Xylella*-induced transcripts become strong candidates for further study, now involving dozens of transcripts. Third, principal component analysis (data not shown) suggests that on infected plants late in the season (when fruit would normally be maturing) there is a significant shift in gene expression, such that they are more similar to young leaves. This change may reflect a shift in source-sink relationships brought as a consequence of aborted berry development in infected individuals.

FUNDING AGENCIES

Funding for this project was provided by the USDA Animal and Plant Health Inspection Service, the USDA Agricultural Research Service, and the CDFA Pierce's Disease and Glassy-winged Sharpshooter Board.