A GENOME-WIDE APPROACH TO PLANT-HOST PATHOGENICITY IN *XYLELLA FASTIDIOSA*: MULTIGENIC METHODS FOR IDENTIFYING STRAINS, FOR STUDYING THE ROLE OF INTER-STRAIN RECOMBINATION, AND FOR IDENTIFYING PATHOGENICITY CANDIDATE GENES

Project Leader:
Leonard Nunney
Department of Biology
University of California
Riverside, CA 92521

Cooperators:
Richard Stouthamer and Robert Luck
Department of Entomology
University of California
Riverside, CA 92521

Reporting Period: The results reported here are from work conducted September 31, 2004 to August 31, 2005.

ABSTRACT
We have developed a multilocus sequence typing (MLST) system for identifying the known pathovars of *Xylella fastidiosa* (including subsp. *fastidiosa* that causes Pierce's disease). This identification system is based on allelic variation at seven housekeeping genes (*holC*, *nuoL*, *leuA*, *gltT*, *cysG*, *petC* and *lacF*) and a public MLST database has been established at www.mlst.net (see Scally et al. 2005). This easily accessible system will allow for the rapid recognition of novel variants as they arise, since it can be continuously updated by researchers who have sequenced new isolates for the seven genes. We tested the effectiveness of the MLST system using 25 isolates of *Xylella fastidiosa* (*Xf*) from five different host plants: grapevine (Pierce's disease, PD), oleander (oleander leaf scorch, OLS), oak (oak leaf scorch, OAK), almond (almond leaf scorch, ALS) and peach (phony peach scorch, PP). An eBURST analysis identified six clonal complexes (CCs), using the grouping criterion that each member of a CC must be identical to at least one other member at five or more of the seven loci. These clonal complexes corresponded to phylogenetic clades that we had previously identified (Schuenzel et al. 2005), including subspecific clades of *fastidiosa* and *sandyi* (CC1 & 2), while CC3-5 defined host-specific sub-clades of the subsp. *multiplex*. CC6 identified a *multiplex*-like group characterized by a high frequency of inter-subspecific recombination. To begin to understand the potential role of recombination in the creation of new pathovars, we also used the MLST data (plus three additional loci) to estimate the relative contribution of recombination and mutation to the observed variability. Recombination between different alleles was estimated to give rise to 76% of the nucleotide changes and 31% of the allelic changes observed. However, sequence data also suggests that inter-subspecific recombination has started relatively recently. This new phenomenon may lead to an increased rate of pathovar formation.

INTRODUCTION
The availability of four genome sequences of *Xf* (two completely annotated and two non-annotated), allows us to exploit the extraordinary power of genomic research to investigate *Xf*'s genetic diversity. This diversity can provide information essential for understanding the plant-host specificity of the *Xf* subspecies. So far, the only form implicated in causing PD is subspecies *fastidiosa*, however a very real possibility exists that new pathovars may arise by recombination among the three North American subspecies. Sequence data from the *Xf* genomes suggests that recombination was historically rare; however this has never been quantified.

We have shown that the PD pathovar (subsp. *fastidiosa*) has very low sequence variability (Schuenzel et al. 2005). This suggests that the PD pathovar has been subject to intense selection- a result probably reflecting significant genetic constraint imposed by the grapevine host on the bacterium. Identifying this constraint is likely to lead to a mechanism for pathogen control.

Our first priority was to place the PD strain within a statistically robust phylogeny, extending earlier work defining the interrelationships of the plant-host strains of *Xf* (e.g. Hendson et al. 2001; Lin et al. 2005). Schaad et al. (2004) identified two North American subspecies based on DNA hybridization: subsp. *multiplex*, found on a range of hosts including almond, peach and plum, plus the PD pathovar, subspecies *fastidiosa* (initially named *piercei*). Using DNA sequence data, we added a third North American subsp. (*sandyi*) isolated originally from oleander and we estimated that these three subspecies have been separated for more than 15,000 years (Schuenzel et al. 2005).

Given a robust phylogeny, our challenge was to develop an effective method for identifying the known host pathovars. The "state-of-the-art" approach is to use MLST (multiple locus sequence typing) (Maiden et al. 1998). This technique has been applied primarily in the identification of pathovars of human pathogens, and a public database has been established at www.mlst.net, which is located at Imperial College, London and is funded by the Wellcome Trust.

Unambiguous identification of strains is of considerable importance for understanding the epidemiology of PD and the other plant diseases caused by this bacterium. Previously, this has been approached using a variety of DNA based methods (Banks et al. 1999; Hendson et at. 2001; Rodrigues et at. 2003; Meinhardt et al. 2003; Lin et al. 2005). As yet, it has not been established if results from these methods have a clear relationship to the true underlying phylogenetic relationships. The simple sequence repeat (SSR) approach of Lin et al. (2005) shows up high variability, useful for uniquely distinguishing isolates, however it relies on many (34) loci. In contrast, MLST methods rely on the allelic variability of just seven

housekeeping genes, and so measure variation more conservatively than SSR methods, but at a level designed to clearly identify significant phylogenetic pathovar groupings.

New pathovars can develop via horizontal transfer of new genetic material or via changes in existing genes. Changes in existing areas can accumulate via mutation within a pre-existing variant, or via recombination among the known pathovars. Recombination in particular has great potential to rapidly create new and very different forms. *Xf* is generally assumed to be clonal, although we know that virally-mediated horizontal transfer of genes must occur since some regions of DNA are unique to one subspecies (Van Sluys et al. 2003). The possibility of homologous recombination is more difficult to detect, but Schuenzel et al. (2005) found evidence of such genetic exchange. Since recombination could lead to the very rapid evolution of novel pathogenic forms, we were interested in quantifying how much of the genetic variation seen within subspecies was likely to be due to recombination, as well as further examining the possibility of inter-subspecific transfer.

OBJECTIVES

1. To develop a multilocus sequence typing (MLST) system for identifying pathovars of *Xf*. Our objective was to develop an MLST method that unambiguously identifies the known host pathovars, and that can provide for the efficient recognition of new forms.
2. Estimate the frequency of recombination. Our objective was to measure the effect of homologous recombination. Genetic transfer can dramatically increase the rate of evolution, and rapidly create new host strains.

RESULTS

Objective 1: To develop a multilocus sequence typing (MLST) system for identifying pathovars of Xf

MLST is a recently devised method for identifying strains of bacteria based solely on nucleotide sequence differences across a small number of housekeeping genes (Maiden et al. 1998). Usually seven genes are used and each allele identified is given its own locus-specific number, so that each isolate characterized is represented by its sequence type (ST)- the set of seven numbers defining the alleles at each locus (Table 1). In contrast to prior DNA-based methods, MLST sequence data is unambiguous, can be easily interpreted and replicated between labs, and is generally made available on a public database. MLST typically has higher resolution than most previous methods, while avoiding the problems of excessive variability associated with use of microsatellite loci (i.e. SSRs).

Table 1. MLST allelic profiles of 25 *Xf* isolates based on seven housekeeping genes. The resulting 15 sequence types (STs) were divided into six clonal complexes (CCs).

Clonal Complex	Isolate (grouped by Sequence Type	Gene						
		holC	**nuoL**	**gltT**	**cysG**	**petC**	**leuA**	**lacF**
CC1	PD1,4,6,10,7, ALS5	1	1	1	1	1	1	1
	ALS11	1	1	1	1	1	4	1
	PD16	1	4	1	1	1	1	4
	PD14	1	1	1	4	1	1	1
CC2	OLS 8, 19,20, 21	2	2	2	2	2	2	2
	OLS 2	2	5	2	2	2	2	2
	OLS9	2	6	2	2	2	2	2
CC3	ALS13, 15	3	3	3	3	3	3	3
	ALS3	3	7	3	3	3	3	3
CC4	OAK17, 23	4	7	4	5	3	3	5
	OAK24	5	7	4	5	3	3	5
CC5	PP27	6	7	5	3	4	3	5
	PP28	6	7	6	3	4	5	5
CC6	ALS12	7	7	7	6	3	6	5
	ALS22	7	8	7	7	3	6	5

The MLST data is used to group closely related strains into clonal complexes (CCs). We adopted the definition of a clonal complex suggested by Feil et al. (2001): it is a group in which every member shares at least five identical alleles across the seven loci with at least one other genotype in the group.

We examined the effectiveness and robustness of our MLST method at detecting subspecies and plant-host pathovars by comparing the clonal complexes defined from the analysis of 25 isolates to our pre-existing phylogeny (Shuenzel et al. 2005). The strains used in the analysis originated from symptomatic individuals of five plant species, grape (PD), oleander (OLS), oak (OAK), almond (ALS) and peach (PP). The allelic profiles of these 25 strains produced 15 different sequence types (STs) (Table 1). The eBURST program (Feil et al. 2004) assigned these to six clonal complexes (CCs) (Figure 1). All

strains from grapevine plus two from almond (ALS5 and ALS11) formed the clonal complex 1 (CC1), corresponding to subsp. *fastidiosa*. Similarly, the six OLS strains of subsp. *sandyi* formed CC2. The MLST method divided isolates of the subsp. *multiplex* into three plant-host groups: ALS3, 13, 15 formed the complex CC3, and the OAK and PP strains formed groups CC4 and CC5. Finally, the strains ALS12 and ALS22 formed CC6. These last two isolates are unusual because they exhibit clear signs of recent inter-subspecific recombination (Schuenzel et al. 2005).

The six clonal complexes identified corresponded to six statistically significant clades in our original maximum likelihood tree (Figure 2, see also Schuenzel et al. 2005). Thus our MLST system identified meaningful phylogenetic groupings. Note however that the phenetic tree derived from the MLST relationships can be misleading; note the incorrect position of OLS strains in the UPGMA tree (Figure 1). This same problem can be seen in the SSR-based tree of Lin et al. (2005).

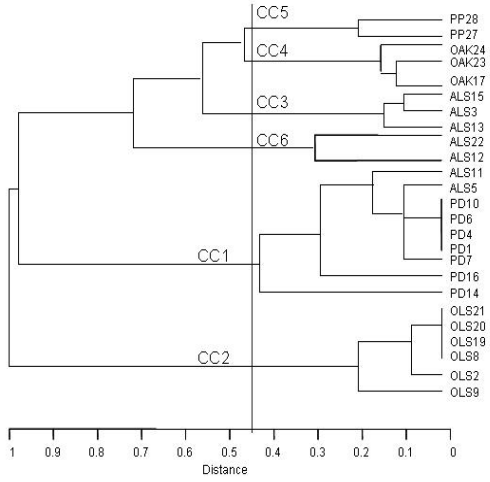

Figure 1. Dendrogram showing the relationships among the clonal complexes based on the unweighted pair-group method with arithmetic averages (UPGMA). The dashed line indicates region demarcating the designation of six clonal complexes.

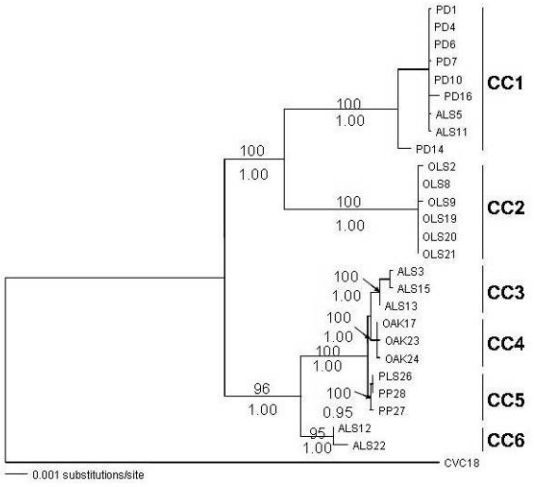

Figure 2. Maximum likelihood phylogeny of *Xf* strains based on 9,307 base pairs. Numbers above and below branches refer to bootstrap support and Bayesian posterior probabilities respectively. The phylogenetic placement of the clonal complexes (CCs) is also shown.

Objective 2: Estimate the frequency of recombination

Several methods have been suggested for estimating the frequency of recombination. One method uses MLST data directly (Feil et al. 2001). In this method, the alleles within a clonal complex are compared to the alleles of its "ancestral" sequence type. (The ancestral sequence type is the ST within the CC that has the largest number of STs that differ at just a single locus). Allelic differences due to a single base pair are used to approximate the number of point mutations, whereas those differing at multiple sites are assumed to reflect recombination events. The role of recombination, relative to mutation, in creating clonal diversity can then be measured in two ways: by the ratio of recombination to mutation (r/m) per allele (where each event results in one change) and by the ration per nucleotide (where each mutation results in one change, but each recombination results in more than one change). This method implicitly assumes that detectable recombination is between (and not within) clonal complexes. We also estimated recombination between clonal complexes by looking directly for congruence between alleles in different clonal complexes both by using the DnaSP software (see Betran et al. 1997) and by visual inspection of the data.

Table 2. Number of recombination events estimated from the sample of 25 isolates, each sequenced for 10 genes (the seven MLST loci plus *rfbD*, *nuoN*, and *pilU*). Three different methods were used (see text).

Method	No of recom-bination events	Genes featuring recombination events
Visual inspection	4	holC, cysG (2), pilU
DnaSP	5	holC, cysG (2), pilU, leuA
MLST	10	holC, cysG (2), pilU, leuA, rfbD, nuoL (4)

All methods identified four clear examples of recombination among the clonal complexes, and the DnaSP and MLST methods identified an additional event in the *leuA* gene (Table 2). An additional five potential examples of homologous recombination were identified by the MLST method alone (mainly in the *nuoL* gene). These examples involved 2-4 base pair changes and additional sampling of isolates would help determine if these are true recombination events, or examples of

multiple point mutations within the same gene within a clonal complex. The total of 10 events detected by MLST involved 71 base pair substitutions compared to 10 allelic changes due to single base pair substitutions. This converts to an estimate of 31% of new alleles arising from recombination and of 76% of DNA base changes within a clonal complex due to recombination. In comparison to many other bacteria, *Xf* has low recombination to point mutation ratio both per allele (0.46:1) and per nucleotide (3.23:1). For example, *Streptococcus pneumoniae* (8.9:1, 61.0:1) and *Neisseria meningitidis* (4.75:1, 100.0:1) (Feil et al. 2001) have ratios shifted more than 10 fold in favor of recombination. However, the clear phylogenetic separation of the clonal complexes, and particularly of the three subspecies, combined with our current estimates, suggests that recombination in *Xf* may be increasing in frequency, possibly due to the effects of agriculture and new insect vectors mixing previously separated subspecies.

CONCLUSIONS

1. We have established a MLST system for *Xf* based on seven housekeeping genes. The database is publicly available at www.mlst.net, an Imperial College London website supported by the Wellcome Trust.
2. The MLST system groups North American *Xf* isolates into six clonal complexes (CCs) that correspond to the six statistically-supported host-related clades identified using phylogenetic methods. Any isolate can be classified within this framework s a previously recognized sequence type or as novel. If it is a new sequence type then it may be within one of the already identified clonal complexes, or recognized as defining a new clonal complex. This will allow host shifts and/or the emergence of new pathovars to be easily tracked
3. We estimated that 31% of new alleles arose from recombination, and that 76% of DNA base changes within a clonal complex arose from inter-complex recombination. These results suggest that the possibility of novel pathovars arising by recombination is high.

REFERENCES

Banks, D, R Albibi, J Chen, O Lamikanra, R. Jarret, BJ Smith. 1999. Specific detection of *Xylella fastidiosa* Pierce's disease strains. Curr. Microbiol. 39:85-88.

Betran, E., J. Rozas, A. Navarro and A. Barbadilla. 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146:88–99.

Feil, E.J., E.C. Holmes, D.E. Bessen, M.-S. Chan, N.P.J. Day, M.C. Enright, R. Goldstein, D.W. Hood, A. Kalia, C.E. Moore, J. Zhou and B.G. Spratt. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA. 98:182-187.

Feil, E.J., B.C. Li, D.M. Aanensen, W.P. Hanage and B.G. Spratt. 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 186:1518-1530.

Hendson, M, AH. Purcell, D Chen, C Smart, M Guilhabert, B Kirkpatrick. 2001. Genetic diversity of Pierce's disease strains and other pathotypes of *Xylella fastidiosa*. Appl. Environ. Microbiol. 67: 895-903.

Lin, H, E.L. Civerolo, R. Hu, S. Barros, M. Francis, and M. A. Walker. 2005. Multilocus Simple Sequence Repeat Markers for Differentiating Strains and Evaluating Genetic Diversity of *Xylella fastidiosa.* Appl. Env. Micobiol. 71: 4888-4892.

Maiden, MCJ, JA Bygraves, E Feil, G Morelli, JE Russell, et al. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Nat. Acad. Sci. 95: 3140-3145.

Meinhardt LW, MPMA Ribeiro, HD Coletta, CK Dumenyo, SM Tsai, CD Bellato. 2003. Genotypic analysis of *Xylella fastidiosa* isolates from different hosts using sequences homologous to the *Xanthomonas* rpf genes. Mol. Plant Path. 4: 327-335

Rodrigues JLM, ME Silva-Stenico, JE Gomes, JRS Lopes, SM Tsai. 2003. Detection and diversity assessment of *Xylella fastidiosa* in field-collected plant and insect samples by using 16S rRNA and gyrB sequences. Appl Env. Microbiol. 69: 4249-4255.

Scally, M., E. L. Schuenzel, R. Stouthamer, and L. Nunney. 2005. A multilocus sequence type system for the plant pathogen *Xylella fastidiosa*, and the relative contribution of recombination versus point mutation to clonal diversity. Appl. Environ. Microbiol. In Press.

Schaad, N.W., E. Pastnikova, G. Lacey, M. Fatmi, C.J. Chang. 2004. *Xylella fastidiosa* supspecies: *Xf* subsp piercei, subsp. nov., *Xf* subsp. multiplex subsp. nov., and *Xf* subsp. pauca subsp. nov. Syst. Appl. Microbiol. 27: 290-300; & 763-763 (correction).

Scheunzel, EL, M. Scally, R. Stouthamer, L. Nunney. 2004. A multi-gene phylogenetic study of clonal diversity and divergence in North American strains of the plant pathogen *Xylella fastidiosa*. Appl. Env. Microbiol. 71: 3832-3839.

Van Sluys MA, MC de Oliveira, CB Monteiro-Vitorello, CY Miyaki, LR Furlan, *et al*. 2003. Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J. Bacteriol. 185: 1018-1026.

FUNDING AGENCIES

Funding for this project was provided by the University of California Pierce's Disease Grant Program.