

THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Population Genomics of Adaptation in *Pseudomonas syringae*

Reuben W. Nowell

Thesis submitted for the degree of Doctor of Philosophy

Institute of Evolutionary Biology
School of Biological Sciences
University of Edinburgh

December 2014

Abstract

Horizontal gene transfer (HGT) and gene loss are important processes in the evolution of prokaryotic lineages. HGT involves the movement of genetic material between distantly related species, and can facilitate adaptation when gained genes confer advantageous phenotypes to recipient lineages. However, high levels of gene gain and loss are predicted to obfuscate patterns of vertical descent and homogenise nucleotide diversity across ecological and phylogenetic boundaries. Thus, a holistic understanding of the role of genome fluctuation in the emergence and maintenance of genetically and ecologically cohesive bacterial groups remains to be fully elucidated. In this thesis, I use the plant-associated bacterium *Pseudomonas syringae* as a model system to investigate the impact of HGT and gene loss on evolutionary processes such as adaptation, diversification and speciation. The Gram-negative Gammaproteobacterium *P. syringae* is an opportunistic plant pathogen, and has been used for decades as a model system with which to study the interaction between plants and their microbial pathogens. Recently, the diversification of lineages within this species has involved a number of host jumps onto a range of woody host plant species, resulting in the emergence of diseases such as bacterial canker of kiwi and bleeding canker of the European horse chestnut.

Using whole-genome sequence data and a range of comparative genomics and phylogenetics methods, I quantitatively reconstruct the history of gene gain and loss in *P. syringae* and show HGT to be the predominant evolutionary force in this species. Genomes of this species are under constant permutation, are subject to a highly diverse HGT genepool and show marked differences in patterns of codon usage between imported and core genes. I then generate additional genome data for 26 strains of *P. syringae* that are pathogenic on a range of different woody plants, and investigate the contribution from HGT to the adaptation of these strains into the woody niche. Using a method that accounts for the underlying phylogenetic relationships among *P. syringae* strains, I look for the correlated evolution between gained genes and the woody niche, and find that a substantial proportion of the genome is associated with this ecological niche. I then investigate the recent adap-

tation of *P. syringae* pv. *aesculi* onto the European horse chestnut, and show that a number of genomic events that include both homologous and non-homologous recombination are likely to have led to the evolution of this bacterium onto its host, where it has become the causal agent of the bleeding canker disease that is currently epidemic across much of northern and central Europe.

Overall, this thesis is an investigation into how HGT contributes to niche adaptation in *P. syringae*, and aims to further our understanding of the mechanisms that underlie bacterial evolution.

Lay Summary

This PhD project aims to further our understanding of how bacteria evolve to infect and cause disease in trees. Recently, a number of highly destructive and newly emerging diseases of tree species have been reported, including bacterial canker of kiwi, hazelnut decline and bleeding canker disease of European horse chestnut. These diseases are all caused by different strains of a bacterial species called *Pseudomonas syringae*, and are responsible for recent epidemics that have had substantial socio-economic impacts on the industries or amenity landscapes that they affect. A major component of this project was to sequence the genomes of a number of *P. syringae* strains that infect trees in order to access information regarding the genetic makeup of these strains. We have analysed these data using a number of comparative methods to ascertain differences between strains that might pinpoint the specific adaptations that have allowed *P. syringae* to live in trees. Furthermore, we can also use these data to tell us more about what evolutionary processes are important during this adaptive process. Thus, we can build a greater understanding of how bacteria evolve and adapt to cause novel diseases in previously uninfected tree species, and what might be the important evolutionary factors that underlie this process.

Acknowledgements

This thesis is the culmination of four years study, and thus, the list of people who have aided either it or me along the way is numerous. Firstly, I must thank my supervisors, Paul Sharp and Sarah Green, for establishing the project in the first place, for their help, advice and guidance, and for allowing me the freedom with which to pursue my own goals within this PhD. Special thanks to Paul, for putting up with all the things I didn't know. I am very grateful. I would also like to thank Bridget Laue, who patiently and kindly guided me through the laboratory stages of the project.

Here is a list of people who I feel, for one reason or another, deserve to see their name written here: Darren Obbard for help with genome assembly and for letting me use Basden; Richard Perry for fraternity; Georgios and Dom for proof-reading in their holidays; the entire collective of PhD students within IEB, particularly my own cohort, who have made the day to day a lot of fun; my parents, brother and sister, for... well, everything; and my friends (Corey, Robbie, Tiff, Theo, Hugo, Fife et al.), for remaining thus despite my poor showing over the last year.

There is a final person to whom the very existence of this thesis owes its gratitude, for surely I would have given up long ago had it not been for her kindness, affection and cajoling: Hannah, thank you very much.

Declaration

I declare that this thesis has been composed by myself and, except where otherwise stated, is entirely my own work. It has not been submitted for any other degree or qualification elsewhere.

—Reuben W. Nowell, April 21, 2015.

Contents

Abstract i

Lay Summary iii

Acknowledgements iv

Declaration v

1 Introduction 2
- 1.1 Forces that influence microbial evolution . . . 3
  - 1.1.1 The role of recombination in bacterial evolution . . . 4
  - 1.1.2 Ochman's spectre: HGT and the bacterial 'species' problem . 6
- 1.2 *Pseudomonas syringae*: a model system . . . 8
  - 1.2.1 The ecological diversity of *P. syringae* . . . 8
  - 1.2.2 Gene repertoire dynamics and phytoadaptation . . . 9
  - 1.2.3 Adaptation of *P. syringae* to the woody niche . . . 11
  - 1.2.4 The recent emergence of horse chestnut bleeding canker . . . 12
  - 1.2.5 Evolution of *Pseudomonas syringae* pv. *aesculi* . . . 13
- 1.3 Thesis outline . . . 15
  - 1.3.1 Aims . . . 15
  - 1.3.2 Chapters outline . . . 16

2 The extent of genome flux and its role in the differentiation of bacterial lineages 19
- 2.1 Abstract . . . 19
- 2.2 Introduction . . . 20
  - 2.2.1 Horizontal transfer and the bacterial species problem . . . 20
  - 2.2.2 *Pseudomonas syringae*: a model system . . . 22
  - 2.2.3 Objectives . . . 23
- 2.3 Materials & Methods . . . 24
  - 2.3.1 Genomic data . . . 24
  - 2.3.2 Assignment of orthology . . . 24
  - 2.3.3 Correction of errors associated with draft-genomes . . . 24
  - 2.3.4 Construction of core and pathovar-specific gene sets . . . 28
  - 2.3.5 Core genome phylogenetics . . . 29

2.3.6 Genome content analysis . . . 29
2.3.7 Reconstruction of gene gain and loss . . . 30
2.4 Results . . . 32
2.4.1 The core and pan genomes of *P. syringae* . . . 32
2.4.2 *Pseudomonas syringae* phylogenetics . . . 34
2.4.3 Gene gain and loss . . . 37
2.4.4 Source and functions of gained genes . . . 40
2.5 Discussion . . . 45
2.5.1 The extent of potential errors in draft genomes . . . 45
2.5.2 Phylogenetics of the core and flexible genomes . . . 46
2.5.3 Species definition in *P. syringae* . . . 48
2.5.4 Genome dynamics and niche adaptation in *P. syringae* . . . 49
2.5.5 Divergence via differential adaptation . . . 51

3 Synonymous codon usage in *Pseudomonas syringae* 54
3.1 Abstract . . . 54
3.2 Introduction . . . 55
3.3 Methods & Materials . . . 57
3.3.1 Sequence data . . . 57
3.3.2 Statistical analyses . . . 57
3.4 Results . . . 60
3.4.1 Variation in codon usage in *Pseudomonas syringae* . . . 60
3.4.2 A primary trend associated with the flexible genome . . . 63
3.4.3 A secondary trend associated with highly expressed genes . . . 66
3.4.4 A tertiary trend associated with GC skew . . . 67
3.5 Discussion . . . 71

4 Genome sequencing and assembly of 26 strains of *Pseudomonas syringae* isolated from woody hosts 75
4.1 Abstract . . . 75
4.2 Introduction . . . 75
4.3 Materials & Methods . . . 76
4.3.1 Bacterial strains . . . 76
4.3.2 Host information . . . 79
4.3.3 Growth conditions and DNA extraction . . . 80
4.3.4 Whole-genome sequencing, assembly and annotation . . . 81
4.3.5 Phylogenetic analyses and whole genome comparisons . . . 83
4.3.6 Characterisation of mobile genetic elements . . . 83
4.4 Results . . . 84
4.4.1 *P. syringae* assemblathon . . . 84
4.4.2 Assembly results . . . 87
4.4.3 Phylogenetics and whole-genome comparisons . . . 88
4.4.4 Mobile genetic elements . . . 95
4.5 Discussion . . . 97
4.5.1 Sampling across *P. syringae* diversity . . . 100

4.5.2 Genome characteristics . . . 101
4.5.3 *P. syringae* in the genomics era . . . 102

5 Population genomics of niche adaptation in *Pseudomonas syringae* 104
5.1 Abstract . . . 104
5.2 Introduction . . . 105
5.3 Materials & Methods . . . 106
5.3.1 Sequence data and orthology . . . 106
5.3.2 Gene occurrence matrix . . . 112
5.3.3 Phylogenetics . . . 112
5.3.4 Distribution of T3SEs and other virulence factors . . . 113
5.3.5 Statistical modelling of correlated evolution . . . 114
5.3.6 Hypothesis testing and null model . . . 115
5.4 Results . . . 117
5.4.1 Core and pan genomics . . . 117
5.4.2 Phylogenetics . . . 120
5.4.3 Homologous recombination . . . 123
5.4.4 Type III secretion system effector repertoires . . . 124
5.4.5 Phytotoxins and other non-secreted virulence factors . . . 128
5.4.6 Distribution of LR statistics . . . 131
5.4.7 Occurrence profile . . . 131
5.4.8 Functions of associated genes . . . 135
5.5 Discussion . . . 136
5.5.1 Genome dynamics in *P. syringae*: another look . . . 138
5.5.2 Evolution of *P. syringae* onto woody hosts . . . 140
5.5.3 HGT and niche adaptation in *P. syringae* . . . 142
5.5.4 An emerging model of *P. syringae* evolution . . . 144

6 Evolutionary adaptations of *Pseudomonas syringae* pv. *aesculi* to the European horse chestnut 146
6.1 Abstract . . . 146
6.2 Introduction . . . 147
6.3 Materials & Methods . . . 149
6.3.1 Sequence data . . . 149
6.3.2 Phylogenetics and reconstruction of gene gain and loss . . . 149
6.3.3 Whole-genome alignment . . . 152
6.3.4 Variant discovery and annotation . . . 153
6.4 Results . . . 154
6.4.1 Phylogenetics . . . 154
6.4.2 Gene gain along the *Pae* lineage . . . 155
6.4.3 Divergence analyses . . . 168
6.5 Discussion . . . 172
6.5.1 Evolution of PG3 pathovars onto woody hosts . . . 173
6.5.2 Gene gain, loss and rearrangement of the D-xylose operon . . . 174
6.5.3 Potential role of *hopV1* in the woody niche . . . 177

6.5.4 Acquisition of β-ketoadipate pathway coincides with expansion into the woody niche 178
6.5.5 A possible role for opine degradation? 179
6.5.6 Divergence and short-term rate of evolution for E-*Pae* 181

7 Perspectives **184**
7.1 The role of HGT in *P. syringae* adaptation 184
7.1.1 The *P. syringae* genome is labile in both time and space 184
7.1.2 HGT and niche adaptation in *P. syringae* 186
7.2 An emerging model of bacterial evolution? 187
7.3 Future directions 191

Appendix A **219**

“ *Though this be madness, yet there is method in't.* ”

Shakespeare, *Hamlet*, Act 2, scene 2



A European horse chestnut, *Aesculus hippocastanum*, on left, grows next to an Indian horse chestnut, *Aesculus indica*, in Edinburgh Botanical Gardens. R.W. Nowell, September, 2014.

Chapter 1

Introduction

PROKARYOTIC life is found in virtually every conceivable ecological niche on the planet. From deep below the earth's crust to high in the stratosphere, microbes from the Bacteria and the Archaea not only persist, but thrive in abundance, and contribute to a wealth of genetic and ecological diversity of which we have only scratched the merest of surfaces (Woese, 1987; Gold, 1992; Stetter et al., 1993; Sattler et al., 2001; Torsvik et al., 2002; Sogin et al., 2006). Out of this dazzling array come some immediate questions: how has the diversity of the microbial world arisen, and what are the forces of nature that sustain it? An explanation, in part, may be found by considering some of the major differences among the three domains of life, which may be dichotomously partitioned into the prokaryotes (bacteria and archaea) and the eukaryotes. Time, for example, has undoubtedly played a role. Ancestral prokaryotes probably arose approximately 3.8 billion years ago (about 2 billion years before the eukaryotes), and thus the geological timescale over which prokaryotic evolution has occurred may in part explain the extraordinary ecological diversity of these organisms (Woese and Fox, 1977; Knoll, 1992; Doolittle, 2000; Torsvik et al., 2002). In addition to this, however, there are also marked differences in the relative importance of the underlying mechanisms that bring about evolutionary change in prokaryotic and eukaryotic organisms. The scale and extent of such differences are only now beginning to be fully appreciated, assisted greatly by the ongoing "genomics revolution" that has had major impacts across the

field of evolutionary biology, particularly in the study of microbial evolution.

It is the overarching aim of this thesis, therefore, to add another piece to our understanding of the processes involved in prokaryotic evolution. To achieve this, the work presented here focuses on a single taxonomic group within the Gammaproteobacteria, the opportunistic plant-pathogen *Pseudomonas syringae*. Given the diversity of microbial life, it almost goes without saying that what is true for one bacterium is almost certainly not for others, let alone all. However, it is only by elucidating as much of the spectrum of diversity as possible that we can begin to understand the microbial world that exists around us. In this introductory chapter, I begin by discussing the context of this work in terms of what is known regarding the forces that influence microbial evolution. This is followed by a closer look at the organism that forms the focus of this thesis, *Pseudomonas syringae*, and a brief history of its use as a model organism for the study of plant–pathogen interactions. Finally, I provide a more detailed breakdown of both the overall structure of the thesis as well as the aims and objectives of the work contained herein.

1.1 Forces that influence microbial evolution

Bacteria reproduce clonally via binary fission, and are thus devoid of an obligate sexual reproductive stage as characterised by many higher organisms. Nonetheless, their genomes are subject to a variety of processes, including point mutation, genome rearrangement, and both homologous and non-homologous recombination, that result in the accumulation of both genetic and allelic variation within an otherwise clonal population of cells (Maynard Smith et al., 1993; Feil et al., 1999; Ochman et al., 2000; Spratt et al., 2001; Lawrence and Hendrickson, 2003; Didelot and Falush, 2007). Nucleotide substitution and other micro-evolutionary events are without doubt important factors in microbial evolution, and are the ultimate source of all extant genetic variation. Rates of nucleotide substitution have been found to vary widely among bacterial species (e.g., Ochman et al., 1999; Kuo and Ochman, 2009a), and also appear to be dependent on the time over which they are calculated (Ho et al.,

2011). Nonetheless, it has become increasingly apparent that the recombination of genetic material among lineages—that is, prokaryotic sex—is the major force impacting the evolution of the prokaryotes (Ochman et al., 2000; Woese, 2000; Spratt et al., 2001; Boto, 2010). In this section, I provide a brief overview of what is currently known regarding the causes and consequences of recombination on prokaryotic genome evolution, and highlight some of the major questions that remain.

1.1.1 The role of recombination in bacterial evolution

Recombination in bacteria may be defined as homologous, which requires some level of identity between the exchanged sequences, or non-homologous ('illegitimate'), which does not. Either may result when foreign DNA enters a recipient cell via one of three mechanisms: (a) conjugation, which involves the direct transfer of DNA from a donor to a recipient cell via conjugation pili, thus requiring cell-to-cell contact; (b) transduction, which involves the transfer of DNA via an intermediate bacteriophage vector; or (c) transformation, which involves the capture of 'naked' DNA from the surrounding environment, often from the lysis of another cell (Davison, 1999; Spratt et al., 2001; Didelot and Falush, 2007). The mechanism of homologous recombination (HR) requires a high degree of similarity between the incoming DNA and the host genome, and the rate of HR has been shown to decrease exponentially with respect to a decreasing level of sequence identity between donor and recipient (i.e., more distantly related species will recombine less) (Majewski and Cohan, 1998; Majewski, 2001; Thomas and Nielsen, 2005). However, there are no such restrictions on illegitimate recombination, which can occur via the mechanisms of either double-strand break repair (i.e., the joining together of DNA ends that is not contingent on homology between the incorporated sequence and the site of integration), or transposition of flanking regions such as phage integrases or mobile genetic elements that carry foreign DNA into a new host genome (Ochman et al., 2000; Thomas and Nielsen, 2005; Bowater and Doherty, 2006). Thus, while HR may act to shuffle genetic material among reasonably closely related species, illegitimate recombination can result in the horizontal transfer of DNA between more distantly

related lineages. Indeed, the importance of horizontal gene transfer (HGT) is not limited to the Bacteria or even the prokaryotes, and there is increasing evidence of the horizontal transfer of genetic material across the domains of the tree of life (Katz, 1999; Woese, 2000; Andersson et al., 2003; Rest and Mindell, 2003; Gophna et al., 2004; Watkins and Gray, 2006; Boto, 2010).

The advent of whole-genome sequencing technologies over the last two decades, corresponding with the increased ability to conduct comparative genomics analyses on these data, has provided an abundance of evidence that HGT is a significant force in the evolution of prokaryotic lineages, and has been shown to play a major role in the innovation, diversification, adaptation and possibly speciation of a wide range of different bacterial groups (Lawrence and Ochman, 1997; Lawrence, 1999; Makarova et al., 1999; Ochman et al., 2000; Ochman, 2001; Boucher et al., 2003; Lerat et al., 2005; Treangen and Rocha, 2011). In addition, comparative analyses have revealed a remarkably high rate of gene turnover among genes that are gained and lost, such that even very closely related strains can differ dramatically in their gene content (Lefébure and Stanhope, 2007; Kettler et al., 2007; Rocha, 2008; Touchon et al., 2009; Shapiro et al., 2012). Taken all together, the scale, extent and potential effects of HGT on bacterial evolution have led to calls for a revision of the current paradigm in prokaryotic evolution to one that incorporates HGT as a central force (e.g., Boto, 2010; Kumar et al., 2015).

However, while there exist plenty of specific examples of HGT within and among bacterial groups, what is less clear is the effect of such widescale HGT and recombination on evolutionary processes in the longer term. In particular, there is a major discontinuity between the high rates of HGT as observed through the comparison of whole genome sequences and the ability of prokaryotes to maintain both phylogenetically and ecologically stable groups, as they evidently do (e.g., Cohan, 2002; Doolittle and Papke, 2006; Fraser et al., 2009; Caro-Quintero and Konstantinidis, 2011). Thus, it is clear that a holistic understanding of microbial evolution—one that is able to incorporate high levels of HGT and recombination with the observed patterns of ecological and phylogenetic diversity—remains to be

synthesised. These problems, and the nascent models that seek to reconcile them, are discussed in more detail in the following section.

1.1.2 Ochman's spectre: HGT and the bacterial 'species' problem

The level of apparent fluidity of genes among genomes has led many researchers to question how it is possible for microbial populations to evolve and maintain ecologically cohesive groups, and the general validity of the species concept when applied to prokaryotes (e.g., Cohan, 2002; Doolittle and Papke, 2006; Hanage et al., 2005, 2006; Doolittle and Zhaxybayeva, 2009; Doolittle, 2012; Hanage, 2013). HGT is a major problem in the delineation of bacterial diversity into cohesive taxonomic units: a given genome is a mosaic of genes from potentially disparate sources, obfuscating any unambiguous patterns of vertical descent, while the acquisition of a relatively small number of genes may confer drastically different phenotypic qualities to otherwise identical populations. In a seminal paper published almost a decade ago, Ochman and colleagues (2005) describe HGT as a "spectre" over the species concept as applied to bacteria, and summarised the extent of the problem thus:

> Acknowledging the problems associated with classifying groups of asexual organisms into discrete species, the situation with bacteria is even worse. Bacteria reproduce asexually, yet they are also capable of obtaining genes from other organisms, even those of different kingdoms. Moreover, the amounts, types, and sources of imported genes can vary among lineages, allowing gene transfer to blur the boundaries of bacterial groups at every taxonomic level and in ways that are impossible to predict. And if patterns of vertical descent are obscured in varied and unknown ways, then the systematic classification of bacteria might not be possible (Ochman et al., 2005).

Having established the extent of the problem, however, these authors then envisaged an alternative hypothesis in which high rates of gene gain and loss may in fact facilitate, rather than obstruct, the formation of ecologically and phylogenetically differentiated clusters of lineages. The fundamental premise of Ochman et al.'s model is that HGT may facilitate the diversification of bacterial lineages via the process of differential adaptation, whereby gained genes that allow a recipient

lineage to expand into a novel niche may simultaneously facilitate the accumulation of divergence if the movement between niches co-occurs with a reduction in the rate of recombination between the derived lineage and its ancestral population (Lawrence and Ochman, 1998; Lawrence, 1999, 2002; Ochman et al., 2005).

More recent work has shown networks of horizontal transfer to be highly structured, and that the parameters of this organisation are determined (primarily) by the functional properties of horizontally transferred genes within the ecological context in which they are gained (Popa et al., 2011; Boucher et al., 2011; Smillie et al., 2011; Cadillo-Quiroz et al., 2012). In addition, other authors have suggested that the rate of gene turnover may be high enough to effectively decouple individual genes from their genomic backgrounds (Polz et al., 2013). These 'flexible' genes may be thought of as contributing to niche-specific gene pools that confer advantageous functions to recipient lineages within an ecologically-determined framework (Luo et al., 2011; Shapiro et al., 2012; Polz et al., 2013; Shapiro and Polz, 2014). An interesting extension of this idea is that, for sufficiently recombinogenic bacterial populations, it may be useful to consider individual genes (or, more likely, operons of genes) as having distinct niches in a manner analogous to whole organisms (Polz et al., 2013). These later findings add support to the divergence via differential adaptation model: ecologically structured niche-specific gene pools may provide pre-defined adaptive 'pathways' along which bacterial lineages may evolve, driven by the acquisition of niche-adaptive genes. The very same structure may simultaneously provide the basis of micro-geographic niche separation that potentiates physical barriers to continued homologous recombination between differentially adapting populations of bacteria, in a manner akin to the 'mosaic sympatry' model of speciation for eukaryotes (Polz et al., 2006; Mallet, 2008).

The work presented in this thesis is therefore an exploration of these ideas using a single group of bacterial lineages as a model system: those of the *Pseudomonas syringae* species complex. In the following section, I provide a summary of what is known about the evolution and ecology of this important group of bacteria.

1.2 *Pseudomonas syringae*: a model system

This thesis is an investigation into the role of genome fluctuation on bacterial evolution and adaptation using the plant-associated *Pseudomonas syringae* species complex as a model system. This bacterium resides within the Gammaproteobacteria, and is thus within the same class as a number of other well known bacterial genera such as *Vibrio* (a facultative anaerobe associated with the saltwater habitat), *Escherichia* (a common inhabitant of the mammalian gut), *Yersinia* (containing the causative agent of plague) and *Xanthomonas* (another plant pathogenic genus with a similar epidemiology as that of *P. syringae*) (Williams et al., 2010). The *Pseudomonas* genus itself is comprised of a number of environmental, plant and animal-associated species that can exist as saprophytes (on decaying plant matter), epiphytes (as commensals on the surfaces of plants), or within a range of soil, rock and water habitats. The genus contains a number of anthropocentrically important opportunistic pathogens, including *P. aeruginosa*, the causal agent of a range of conditions such as pneumonia and sepsis particularly of immunocompromised individuals (Bodey et al., 1983) and the plant pathogens *P. tolaasii* and *P. agarici*, both of which cause diseases of commercially grown mushroom species (Brodey et al., 1991; Young, 1970).

1.2.1 The ecological diversity of *P. syringae*

Lineages within *P. syringae* are both phylogenetically diverse and ecologically heterogeneous. Although the majority of *P. syringae* pathovars have been isolated from plants (and are therefore described as plant pathogens), recent work has shown that genetically diverse populations of *P. syringae* thrive in a multitude of environmental (non-plant) habitats, such as river headwaters, snowpack and even within clouds (Morris et al., 2008, 2009, 2010; Monteil et al., 2012, 2013). An important characteristic of many lineages within the *P. syringae* species complex is their ability to stimulate the formation of ice on plant surfaces at higher temperatures than would otherwise be the case (Maki et al., 1974). Ice nucleation may be beneficial to the bacterium through the release of nutrients by frost-damaged cells, and ice-nucleating *Pseu-*

domonas strains are considered the primary cause of frost damage on many crop species (Lindow, 1983; Hirano and Upper, 1995). As such, *P. syringae* may also play an important role in the water cycle as a natural cause of precipitation through 'cloud-seeding' (Morris et al., 2007, 2008).

The *P. syringae* species complex is divided into over 50 pathological variants (known as pathovars[1]) distributed into at least seven distinct phylogenetic groups (known as phylogroups) based on nucleotide sequence divergence of a small set of housekeeping genes (Sarkar and Guttman, 2004; Hwang et al., 2005; Parkinson et al., 2010; Berge et al., 2014). Pathovars of this species are the causal agents of a variety of blight, speck, spot and canker diseases on a range of economically important agricultural plant species, including tomato, bean, tobacco and kiwi (Hirano and Upper, 1990; O'Brien et al., 2011a; Mansfield et al., 2012). While the *P. syringae* species complex as a whole possesses a remarkably broad host range, individual lineages often show a high degree of host-specificity and are generally constricted to a small number of potential host species—when transferred to a non-susceptible plant, cells will either fail to proliferate or remain as epiphytes (Morris et al., 2000; Hunter and Taylor, 2006; Sarkar et al., 2006; Lindeberg et al., 2009).

1.2.2 Gene repertoire dynamics and phytoadaptation

The genetic and molecular interactions that occur between a *P. syringae* pathovar and its potential plant hosts are complex, and have been the subject of a great deal of research since the establishment of this species as a model system over a quarter of a century ago (Mansfield, 2009). As for many Gram-negative pathogenic bacteria, a key determinant of *P. syringae* pathogenicity is the canonical type III secretion system (T3SS) apparatus (Sarkar et al., 2006; Cunnac et al., 2009; Lindeberg et al., 2009, 2012). Related to the bacterial flagellum, the T3SS structure is composed of approximately 30 proteins and acts as a 'molecular syringe' to inject virulence factors known as type III secretion effector proteins (T3SEs) into host cells (Cornelis and Van Gijsegem, 2000; Saier, 2004; Lindeberg et al., 2009, 2012). Although *P. syringae*

[1] As such, the 'pathovar' represents a taxonomic delimitation at a level below that of the species, and is often named after the host genus from which a given *P. syringae* lineage has been isolated.

strains also secrete a range of other non-injected virulence factors, such as toxins and phytohormones (Bender et al., 1999), the nature of the interaction between T3SEs and the plant immune system is such that it is these molecules that are inferred to be the most important in the determination of factors such as virulence and host specificity (Cunnac et al., 2009; Lindeberg et al., 2009).

Effector proteins interact with the plant immune system on two levels, both of which elicit different evolutionary responses from the bacterium during the course of the resultant host–pathogen 'arms-race' (Ma et al., 2006; Lindeberg et al., 2012). The first level of plant defense is called pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI), which acts through the recognition of microbial-associated molecules such as flagellin and lipopolysaccherides that are detected by receptors on the surface of plant cells (Boller and Felix, 2009; Lindeberg et al., 2012). Injected T3SEs suppress PTI via the interference of vesicle trafficking within the host cell, resulting in the non-delivery of antimicrobials to the site of infection (Jones and Dangl, 2006; Lindeberg et al., 2012). The second level of host immunity involves the recognition of intracellular T3SEs by corresponding resistance proteins, resulting in a 'gene-for-gene' model of effector-triggered immunity (ETI) that manifests as a form of localized cell death known as the hypersensitive response (HR) in resistant plants (Jones and Dangl, 2006; Lindeberg et al., 2012).

One mechanism through which pathovars are able to evade the ETI response is through the rapid evolution of effector genes via mutation, recombination or loss, such that the novel effector is no longer recognised by the host immune system (Lindeberg et al., 2009, 2012). In addition, the acquisition of novel T3SEs via HGT may allow a recipient lineage to become pathogenic on a new host where that effector protein is not recognised. Thus, the complement and composition of the T3SE repertoire is frequently rearranged both within and among *P. syringae* lineages through the action of natural selection on recognition avoidance (Sarkar et al., 2006; Ma et al., 2006; Block and Alfano, 2011; Baltrus et al., 2011; Lindeberg et al., 2012). Both gene gain and gene loss are important in repertoire rearrangement: novel T3SEs can be acquired via horizontal transfer and domain recombination, while

redundant or 'betraying' effector genes (i.e., those recognised by the host) can be deleted through pseudogenization or episomal excision (Pitman et al., 2005; Ma et al., 2006; Lindeberg et al., 2006).

In addition to fluctuation within the T3SE repertoire, previous comparisons of *P. syringae* genome sequences have shown there to be substantial differences in gene content among lineages in other functional classes of genes, including virulence determinants such as toxins and phytohormones, as well as other genes that are hypothesised to have roles in the adaptation of these bacteria onto their respective host species (e.g., Green et al., 2010; Rodríguez-Palenzuela et al., 2010; Baltrus et al., 2011; Marcelletti et al., 2011). Thus, much of the work presented in this thesis is an investigation into the gain and loss of genes that have putative involvement in either pathogenicity or host adaptation. In particular, I focus on the adaptation of *P. syringae* to cause disease in woody plants, of which there have been a number of recent cases.

1.2.3 Adaptation of *P. syringae* to the woody niche

The diversification of lineages across the *P. syringae* species complex has involved a number of host jumps from herbaceous to woody host species. These transitions have resulted in the emergence of economically costly diseases such as bacterial canker of kiwi (Balestra et al., 2010), olive-knot disease (Rodríguez-Moreno et al., 2009), hazelnut decline (Scortichini, 2002) and bleeding canker of the European horse chestnut (Webber et al., 2008). The causal agents of these diseases share the ability to proliferate within the woody parts (i.e., the cambium, phloem and cortex) of their respective host species, thereby causing disease in these organs. Generally, strains that are pathogenic on woody hosts are distributed across the *P. syringae* phylogenetic tree, and thus represent multiple independent adaptations into the woody niche (Green et al., 2010; Mazzaglia et al., 2012; O'Brien et al., 2012).

One hypothesis for explaining the transition of *P. syringae* onto woody hosts is that these strains have gained a number of genes that are selectively advantageous within this particular environment. Consequently, both HGT and gene loss are likely

to be important evolutionary processes during the adaptation of *P. syringae* to the woody niche. Furthermore, it may be the case that the diversification of *P. syringae* onto its numerous different host species has been driven by an evolutionary process similar to the divergence via differential adaptation model discussed previously, whereby the genomes of individual lineages are 'tailored' to suit the requirements of specific plant habitats. Thus, the adaptation of lineages within the *P. syringae* species complex to the woody niche represents an excellent model system with which to explore these hypotheses further, and will advance our understanding of how processes such as HGT may be impacting the evolution of prokaryotic groups.

One such transition to the woody niche has resulted in the recent emergence and ongoing epidemic of 'bleeding canker' disease of the European horse chestnut tree. It is this particular transition, and the underlying evolutionary processes that have facilitated it, that forms the foundation of the work presented in this thesis. In the sections below I discuss briefly the history and importance of both the disease and its host, before providing further information on the research that has been performed on horse chestnut bleeding canker prior to the initiation of this PhD project.

1.2.4 The recent emergence of horse chestnut bleeding canker

The European horse chestnut, *Aesculus hippocastanum*, is an important tree species with a high amenity value in both urban and rural landscapes. First discovered growing natively in the Balkan peninsula in the late 16^{th} century, the horse chestnut is now commonly cultivated in parks, gardens, and tree-lined avenues across much of the temperate world (Howard, 1945; Lack, 2002). Often used as a shade tree in urban environments, the horse chestnut is a tree of high cultural importance and aesthetic value for its showy white candle-like flowers, large digitate leaves and distinctive 'conker' fruits (Lack, 2002; Lees et al., 2011).

Around 2002/03, a newly emergent and highly destructive disease of horse chestnut was first recorded in the south of England (Webber et al., 2008). Named horse chestnut bleeding canker, the disease attacks the woody parts of the main trunk

and branches of infected trees, causing necrosis of the phloem, cortex and cambium, rust-coloured oozing fissures in the bark ('bleeding' cankers; see FIG. 1.1a & b), foliar discolouration, crown dieback and, potentially, tree death (Webber et al., 2008; Green et al., 2009; Steele et al., 2010). By 2007, bleeding canker had spread across the UK, with levels of symptom prevalence recorded at 70, 42 and 36% for surveyed trees in England, Scotland and Wales respectively (Forestry Commission, 2008). In the decade since its first description, horse chestnut bleeding canker has become epidemic across much of north-west and central Europe, having been officially reported in Ireland (EPPO, 2011), Belgium (Bultreys et al., 2008), the Netherlands (de Keijzer et al., 2012), France (Bardoux and Rousseau, 2007), Germany (Schmidt et al., 2008), the Czech Republic (Mertelik et al., 2013), Norway (Talgo et al., 2012), and Hungary (EPPO, 2013).

Although previous incidences of bleeding canker on horse chestnut had been linked with *Phytophthora* spp., the causal agent of the current outbreak was shown to be a variant of the Gram-negative fluorescent Gammaprotobacterium *Pseudomonas syringae* (FIG. 1.1c) (Bultreys et al., 2008; Schmidt et al., 2008; Webber et al., 2008). Strains isolated from cankers on European *A. hippocastanum* were found to have identical gyrase B gene sequences to the current type strain for *P. syringae* pv. *aesculi* (strain NCPPB 3681), which was isolated from leaf spot lesions on an Indian horse chestnut (*Aesculus indica*) in the Himachal Pradesh region of northern India in 1969, over three decades prior to the European outbreak (Durgapal, 1971; Durgapal and Singh, 1980; Webber et al., 2008). On the basis of this identity, the European strains responsible for bleeding canker on *A. hippocastanum* were assigned to the same *aesculi* pathovar.

1.2.5 Evolution of *Pseudomonas syringae* pv. *aesculi*

The genome sequences for both Indian and European pv. *aesculi* (denoted I-*Pae* and E-*Pae* respectively) have been recently published (Green et al., 2010). These data confirmed a low level of divergence between European isolates and the Indian strain (∼0.05% substitutions per nucleotide site), and showed that three strains isolated

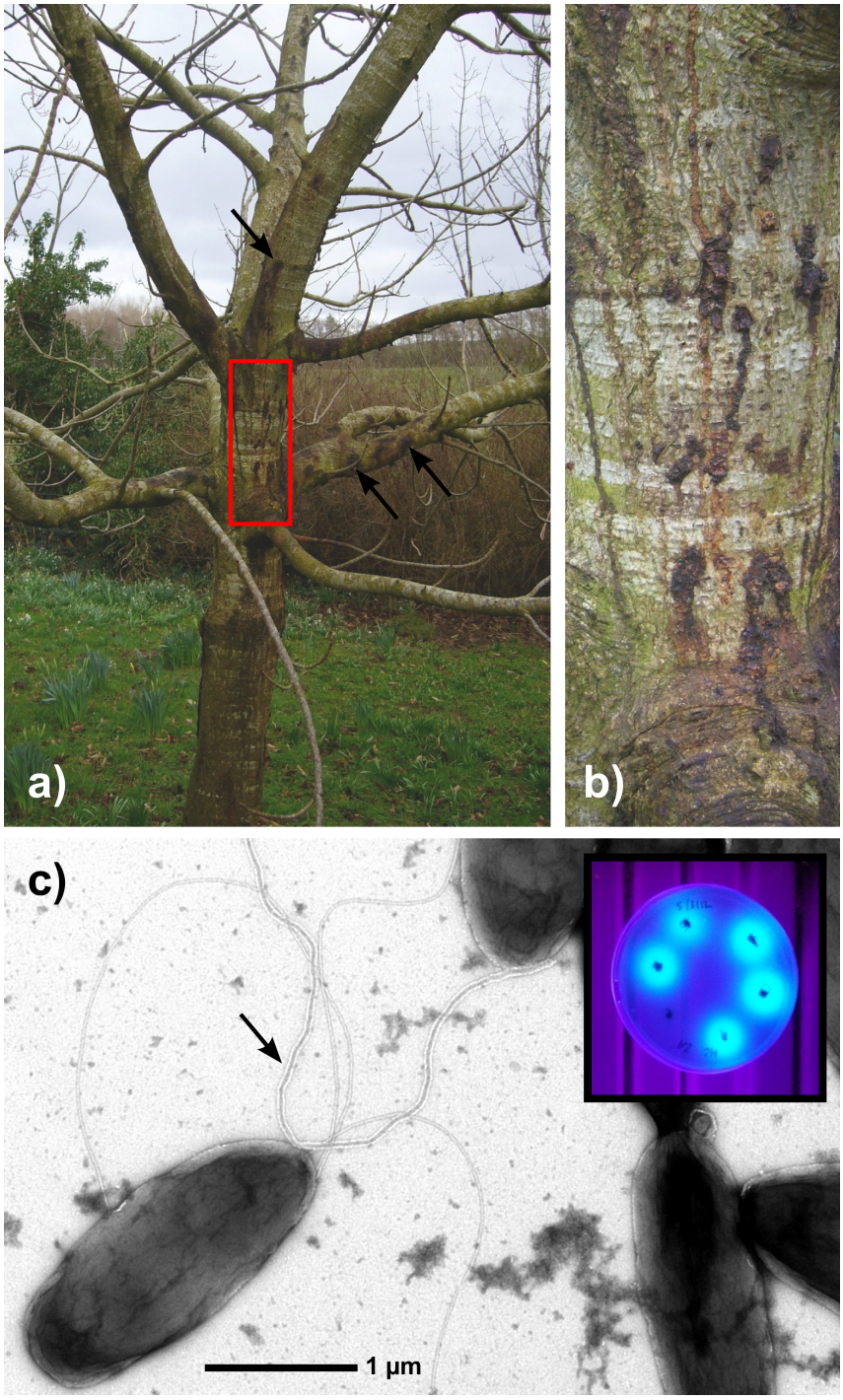


Figure 1.1: (a) Bleeding cankers evident on main stem and branches of a horse chestnut specimen growing in Ayrshire, Scotland; **(b)** Close-up of main stem (red box) showing suppuration of rusty-red coloured tree fluids from 'bleeding' cankers; **(c)** Transmission electron micrograph of pv. *aesculi* cells isolated from a diseased horse chestnut; note presence of multiple flagella. Scale bar represents 1 µm. Inset shows an example of the characteristic blue/green fluorescence of many *Pseudomonas syringae* strains under UV light. Photo credits: (a), (b) and inset are author's own (2012); (c) Bridget E. Laue (2012).

from across the UK (two in Scotland, one in England) were virtually identical at the nucleotide level (Green et al., 2010). This suggested a recent common ancestor for Indian and European pv. *aesculi*, and that the UK outbreak was most likely the result of a single introduction of the bacterium into Great Britain (Green et al., 2010).

In addition, phylogenetic analysis showed the clustering of both I-*Pae* and E-*Pae* within a monophyletyic clade containing a small number of other *P. syringae* pathogens of woody plant species, including olive (pv. *savastanoi*), apricot (pv. *morsprunorum*) and bayberry (pv. *myricae*), while comparisons with other sequenced *P. syringae* pathovars identified a number of genes putatively involved in the adaptation of *Pae* on *A. hippocastanum* (Green et al., 2010). These included genes involved in the catabolism of plant-derived aromatic compounds, the sequestration of iron and the ability to metabolise nitric oxide, that were most likely acquired relatively recently via the process of HGT.

Taken together, these findings suggest that bleeding canker is the result of a single and recent host jump by the *aesculi* lineage, from its natural host *A. indica*, where it causes only minor leaf-lesions, onto a novel host, *A. hippocastanum*, where it has become highly virulent within the woody parts of the tree. However, the evolutionary mechanisms that underlie this niche transition remain poorly understood. In addition, the relative rarity of bacterial diseases of tree hosts has resulted in a paucity of information regarding the genetic basis of bacterial pathogenesis within this environment. In the following section, I synthesise some of the questions that have arisen from this Introduction into a number of overarching aims for this thesis, and briefly outline the work contained within the following chapters.

1.3 Thesis outline

1.3.1 Aims

During this PhD project, I have used the plant-associated bacterium *Pseudomonas syringae* as a model system to investigate the contribution from horizontal gene transfer to the adaptive process in this species. I have generated novel genome

sequence data for 26 additional strains of *P. syringae*, and have applied a variety of comparative methods to specifically investigate the genomic basis of the adaptation of *P. syringae* onto woody host species. Thus, the overarching objective of this work is to add to our understanding of bacterial evolution through the elucidation of how HGT may contribute to this process. More specifically, I aim to elucidate the role of HGT in the adaptation of *P. syringae* onto woody hosts, and to further understand the genetic basis of the adaptation of *P. syringae* pv. *aesculi* onto the European horse chestnut.

1.3.2 Chapters outline

The structure of this thesis, in terms of the scope and focus of the questions that it asks in each chapter, progresses in a top-down manner. That is, I start with an investigation that asks broad and encompassing questions, and progressively 'zoom in' to look at more specific aspects of *P. syringae* evolution that therefore have a more limited scope. Thus, the first data chapter asks the fundamental question: simply, does HGT and gene loss occur in these lineages? And if so, what can we determine about the nature of this fluctuation? In contrast, the final chapter offers a detailed investigation of the genetic and genomic changes that may have facilitated the adaptation of *P. syringae* pv. *aesculi* onto its horse chestnut host. The following sections provide a brief synopsis of each chapter:

Chapter 2 Firstly, I analyse whole genome sequence data for 27 lineages of the *P. syringae* species complex and quantitatively reconstruct the evolutionary history of gene gain and loss for these lineages. I show that genome fluctuation is both extensive and rapid for this species, and that recently acquired genes are highly diverse in terms of both their functions and their putative source. In addition, I develop and discuss a method that is able to account for a range of errors that are prevalent in 'draft' genome sequence data, and investigate the dynamics of genome fluctuation in *P. syringae*.

Chapter 3 In this chapter, I analyse patterns of synonymous codon usage in *P. syringae*, using the completely sequenced genome of *P. syringae* pv. *phaseolicola* as a reference. This reveals that, while genomic G + C content is the major driver of codon preference in this species, a number of other factors also contribute to variation in codon usage among genes. The major factor is heterogeneity in the G + C content among genes, which is likely to be driven by the recent import, via HGT, of genes with anomalous %GC into the *phaseolicola* genome. This highlights the impact of high rates of HGT on codon usage in this species.

Chapter 4 Here, I present the sequencing, assembly and annotation of whole-genome sequence data for 26 additional strains of *P. syringae*, isolated from a range of woody host species. I show the phylogenetic placement of these strains within the *P. syringae* species complex *sensu lato*, and characterise the extent to which these genomes are comprised of mobile genetic elements such as plasmids and insertion sequences.

Chapter 5 In this chapter, I compare whole-genome data for 64 strains of *P. syringae* and investigate the adaptation of lineages onto woody hosts across the *P. syringae* phylogeny. I use a maximum likelihood method that accounts for the confounding effects of phylogeny to look for significant associations between gene presence and host preference (woody or non-woody) across genes that are subject to gain and loss among these genomes. I show that a substantial number of genes are associated with the woody niche, and that the HGT genepool available to *P. syringae* is essentially limitless. Overall, these results provide evidence of convergent gene gain associated with niche type, and suggest that a model of evolution similar to the hypothesis of divergence via differential adaptation (as discussed above) may be operating in *P. syringae*.

Chapter 6 Here, I reconstruct the history of gene gain specifically along the phylogenetic lineage leading to the causal agent of bleeding canker disease of the horse chestnut, *P. syringae* pv. *aesculi*. This analysis reveals a number of genetic and ge-

nomic changes that are putatively involved in the adaptation of *Pae* to the woody niche, and highlights the importance of both homologous and non-homologous recombination in this process. In addition, I show that the level of divergence among *Pae* isolates is very low, consistent with there being a single and recent introduction of this bacterium into the European continent.

Chapter 7 Finally, I provide a general discussion to this thesis. This summarises the major findings of my work and considers to what extent they contribute to our wider understanding of bacterial evolutionary processes, with particular reference to the ideas of divergence via differential adaptation and the contribution of HGT to evolutionary processes in bacteria.

Chapter 2

The extent of genome flux and its role in the differentiation of bacterial lineages

This chapter is published as: Nowell, R. W., Green, S., Laue, B. L. and Sharp, P. M. 2014. The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome Biology and Evolution* 6:1514–1529. See Appendix A for published manuscript and accompanying Supplementary Materials.

2.1 Abstract

Horizontal gene transfer (HGT) and gene loss are key processes in bacterial evolution. However, the role of gene gain and loss in the emergence and maintenance of ecologically differentiated bacterial populations remains an open question. Here, we use whole-genome sequence data to quantify gene gain and loss for 27 lineages of the plant-associated bacterium *Pseudomonas syringae*. We apply an extensive error-control procedure that accounts for errors in draft-genome data and greatly improves the accuracy of patterns of gene occurrence among these genomes. We demonstrate a history of extensive genome fluctuation for this species, and show that individual lineages could have acquired thousands of genes in the same period in which a 1%

amino-acid divergence accrues in the core genome. Elucidating the dynamics of genome fluctuation reveals the rapid turnover of gained genes, such that the majority of recently gained genes are quickly lost. Despite high observed rates of fluctuation, a phylogeny inferred from patterns of gene occurrence is similar to a phylogeny based on amino acid replacements within the core genome. Furthermore, the core genome phylogeny suggests that *P. syringae* should be considered a number of distinct species, with levels of divergence at least equivalent to those between recognised bacterial species. Gained genes are transferred from a variety of sources, reflecting the depth and diversity of the potential genepool available via HGT. Overall our results provide further insights into the evolutionary dynamics of genome fluctuation and implicate HGT as a major factor contributing to the diversification of *P. syringae* lineages.

2.2 Introduction

2.2.1 Horizontal transfer and the bacterial species problem

Bacterial genomes are labile entities, fluctuating in both size and content through time (Ochman et al., 2005; Hao and Golding, 2006; Touchon et al., 2009). Genome fluctuation is mediated by the counteracting processes of gene gain and gene loss. The addition of novel genes expands the functional diversity of the recipient genome, while genes whose presence is no longer required for survival can be deleted. The addition of new genes occurs primarily through horizontal (lateral) gene transfer (HGT), which can drive both phenotypic innovation and subsequent adaptation when the acquired genes confer new traits that allow diversification of that lineage into a novel environment (Ochman et al., 2000; Boucher et al., 2003; Lerat et al., 2005; Treangen and Rocha, 2011). However, the specific contribution of HGT towards other evolutionary processes, such as genomic differentiation and the formation of ecologically-defined population structure, from which speciation may result, remains an open question (Lawrence, 2002; Ochman et al., 2005; Polz et al., 2013).

High rates of gene exchange due to HGT can lead to extensive gene-repertoire

differences among otherwise closely related lineages, and can result in genes within the same genome having conflicting evolutionary histories (Feil et al., 2001b; Gogarten et al., 2002; Ochman et al., 2005; Shapiro et al., 2012). Thus, HGT is predicted to obfuscate patterns of vertically inherited differences between lineages, and should undermine the definition of bacterial 'species' as non-recombining taxonomic units (Gogarten et al., 2002; Doolittle and Papke, 2006; Doolittle and Zhaxybayeva, 2009). Yet despite overwhelming evidence of extensive HGT at virtually all phylogenetic levels, bacterial lineages generally do behave as distinct entities that often appear to have well-defined boundaries between taxonomic groups (Daubin et al., 2003; Ochman et al., 2005; Hanage et al., 2006; Caro-Quintero and Konstantinidis, 2011). These apparently conflicting observations have led to the proposition that the majority of HGT is transient and therefore does not impact the evolutionary trajectories of bacterial lineages (Hao and Golding, 2006; Kuo and Ochman, 2009b). Furthermore, recent work has suggested that HGT may play an important role in the evolution of population structure within bacterial communities, and may facilitate the speciation process (Ochman et al., 2005; Polz et al., 2013).

HGT may bring about divergence and differentiation if the diversifying effect of HGT is able to subdivide a population to the extent that speciation can occur (Lawrence and Ochman, 1998; Lawrence, 1999, 2002; Barraclough et al., 2012; Polz et al., 2013). Under this model, variation in the functions of acquired genes results in the differential adaptation of sub-populations of bacteria into correspondingly different ecological habitats or niches, thereby creating a physical barrier to continued exchange and allowing sequence divergence to accrue to the extent that homologous recombination is suppressed. Subsequent gene loss can enhance differentiation between lineages by limiting niche overlap between incipient species through the restriction of metabolic diversity of populations of bacteria (Lawrence, 1999). The innovative and diversifying effects of HGT are well documented, and as such, the exposure to a potentially huge pool of novel genes via HGT could be the mechanism that drives the early stages of this process. Thus, the forces of genome fluctuation, HGT and adaptation are inexorably intertwined and can, under the right conditions,

act in concert to promote both genomic differentiation and ecological diversification, from which speciation can result.

2.2.2 *Pseudomonas syringae*: a model system

Here, we examine the interplay between these evolutionary processes using the *Pseudomonas syringae* species complex as a model system. *P. syringae* is a remarkably diverse species found almost globally in a multitude of different environments (Hirano and Upper, 1990; Sarkar and Guttman, 2004). Although best known as a plant pathogen, it can exist as a plant commensal as well as a free living organism capable of thriving in non-agricultural habitats (Sarkar and Guttman, 2004; Morris et al., 2008, 2010). The species complex comprises over 50 pathovars (pathological variants) delimited into a number of phylogenetic groups (phylogroups) based on sequence divergence (Sarkar and Guttman, 2004; Hwang et al., 2005; Parkinson et al., 2010). Although as a species *P. syringae* has a broad host range, individual strains are generally confined to a limited set of potential host species (Sarkar et al., 2006; Lindeberg et al., 2009). Strains injected into a non-host plant will either simply fail to grow or will elicit an immune response from the plant that halts the progression of disease symptoms (Morris et al., 2000; Hunter and Taylor, 2006; Sarkar et al., 2006). Thus, individual strains have specific genetic repertoires that promote pathogenicity only when expressed in a particular environment, i.e., that of a compatible host species (Lin, 2005; Sarkar et al., 2006).

The species complex offers a unique opportunity to study the contribution of HGT to the early stages of differentiation between lineages for a number of reasons. Firstly, the radiation of *P. syringae* pathovars onto multiple different host species combined with the high degree of host-specificity of individual pathovars allows for an investigation into the role that HGT may have in the partitioning of the *P. syringae* species complex into different ecological niches, where in this case each host species can be viewed as a distinct niche environment to which its *P. syringae* pathovar is specifically adapted. Secondly, the *P. syringae* species complex is an example of a group of bacteria in the early- to mid-stages of differentiation, with

a number of clades (i.e. phylogroups) with a level of sequence divergence in core genes comparable to that of other 'young' bacterial species. Lastly, there exists a good representation of whole genome sequence data across at least three major phylogroups within the complex, allowing for the accurate elucidation of genome content variation across closely related but phenotypically distinct lineages.

2.2.3 Objectives

We utilise these data to quantify the extent of gene gain and loss within a phylogenetic framework for 27 lineages of the *P. syringae* species complex. Using available whole-genome sequence data and a suite of tools developed to limit the confounding effects of errors associated with draft genomes, we decompose the *P. syringae* pan-genome into its constituent core and flexible components and apply various phylogenomic methods to infer the evolutionary histories of both these entities. We reconstruct the most likely history of gene gain and loss within the framework of the evolutionary history of the core genome, and correlate the rate of gain and loss with phylogenetic depth to provide insights into the dynamics of genome fluctuation through time. We then classify gained genes based on both putative function and potential source to elucidate the functional diversity and the taxonomic breadth of donor species of gained genes. We use these results to ask if the divergence via differential adaptation model described above can explain the observed radiation of *P. syringae* pathovars onto different host species. If this were the case, we would expect to see the acquisition of genes along phylogenetic branches that correspond to host adaptation events. Of particular importance in this process are genes of the type III secretion system (T3SS). This 'molecular syringe' is a fundamental component of the *P. syringae* armoury that knocks down innate plant immune defences via the injection of type III effector proteins (T3SEs) into host cells, allowing for infection (Lindeberg et al., 2009, 2012). Thus, the repertoire of T3SEs possessed by a given pathovar is a key determinant of host specificity, and the gain of T3SEs may play an important role in the diversification of *P. syringae* lineages onto novel host species (Sawada et al., 1999; Ma et al., 2006; Lindeberg et al., 2009; Baltrus et al., 2011;

O'Brien et al., 2012).

2.3 Materials & Methods

2.3.1 Genomic data

All available P. syringae genome sequences (as of 4th May 2012) were downloaded from NCBI GenBank, including multiple strains of the same pathovar (but excluding multiple editions of the same strain). This yielded 27 strains of 18 pathovars from across the *P. syringae* species complex. Summary information for all genomes included in this analysis is given in TABLE 2.1. All protein sequences used in subsequent analyses were based on the published annotations for these genome sequences, also retrieved from NCBI GenBank.

2.3.2 Assignment of orthology

Orthologous relationships among proteins were determined using ORTHOMCL v2.0 (Li et al., 2003). ORTHOMCL uses Basic Local Alignment Search Tool (BLAST) followed by Markov clustering (MCL) to group proteins into putative orthologous groups (OGs). A minimum length cut-off of 50 residues was chosen as many sequences less than 50 were found to be protein fragments (see below). The all-versus-all BLASTp was performed with an *E*-value $\leq$ 1e$-$5, and the inflation index, which controls the threshold at which the MCL algorithm defines distinct OGs, was set at 1.5 (Li et al., 2003; Lefébure and Stanhope, 2007; Lefébure et al., 2010).

2.3.3 Correction of errors associated with draft-genomes

The resultant list of orthologous groups was subjected to a series of quality control procedures to minimise the confounding effects of errors such as misannotations and (in particular) the presence of protein fragments. Protein fragments are caused by the premature truncation of a protein-coding sequence (CDS) in the genome by the termination of a contig in the middle of an open reading frame. When a CDS spans a break between two adjacent contigs (i.e. a gap in a draft genome), it

Table 2.1: Strain information.

Pathovar	Strain	Identifier[a]	Host	Contigs[b]	CDS[c]	Reference
aceris	M302273PT	acer	*Acer* sp. (maple)	1,179	6,185	Baltrus et al. (2011)
actinidiae	M302091	actn	*Actinidia delicosa* (kiwi)	941	5,626	Baltrus et al. (2011)
aesculi	2250	aes2	*Aesculus hippocastanum* (European horse chestnut)	776	5,619	Green et al. (2010)
aesculi	NCPPB 3681	aesN	*Aesculus indica* (Indian horse chestnut)	841	5,649	Green et al. (2010)
aptata	DSM 50252	apta	*Beta vulgaris* (sugar beet)	3,776	6,368	Baltrus et al. (2011)
—	Cit7	cit7	*Citrus sinensis* (navel orange)	2,655	7,145	Baltrus et al. (2011)
glycinea	B076	glyB	*Glycine max* (soybean)	104	5,580	Qi et al. (2011)
glycinea	race 4	gly4	*G. max* (soybean)	108	5,207	Qi et al. (2011)
japonica	M301072PT	japa	*Hordeum vulgare* (barley)	4,661	8,796	Baltrus et al. (2011)
lachrymans	M301315	lacM	*Cucumis sativus* (cucumber)	791	6,691	Baltrus et al. (2011)
lachrymans	M302278PT	lacP	*C. sativus* (cucumber)	798	5,636	Baltrus et al. (2011)
maculicola	ES4326	Pcan	*Raphanus sativus* (radish)	878	6,101	Baltrus et al. (2011)
mori	M301020	mori	*Morus alba* (white mulberry)	3,414	7,544	Baltrus et al. (2011)
morsprunorum	M302280PT	mors	*Prunus domesticus* (European plum)	969	5,837	Baltrus et al. (2011)
oryzae	1_6	oryz	*Oryza sativa* (rice plant)	2,855	7,837	Reinhardt et al. (2008)
phaseolicola	1448A	phas	*Phaseolus vulgaris* (common bean)	3	5,172	Joardar et al. (2005)
pisi	1704B	pisi	*Pisum sativum* (pea)	5,099	9,160	Baltrus et al. (2011)
savastanoi	NCPPB 3335	sava	*Olea europaea* (olive tree)	403	4,898	Rodríguez-Palenzuela et al. (2010)
syringae	B728a	syrB	*P. vulgaris* (common bean)	1	5,089	Feil et al. (2005)
syringae	FF5	syrF	*Pyrus calleryana* (pear)	4,578	8,445	Sohn et al. (2012)
syringae	642	syr6	Small herbaceous plant	296	5,274	Clarke et al. (2010)
tabaci	ATCC 11528	taba	*Nicotiana tabacum* (tobacco plant)	1,405	6,467	Studholme et al. (2009)

Continued on next page

Table 2.1: Continued from previous page

Pathovar	Strain	Identifier	Host	Contigs	CDS	Reference
tomato	DC3000	tomD	*Solanum lycopersicum* (tomato)	3	5,619	Buell et al. (2003)
tomato	K40	tomK	*S. lycopersicum* (tomato)	582	5,853	Cai et al. (2011)
tomato	Max13	tomM	*S. lycopersicum* (tomato)	349	5,749	Unpublished
tomato	NCPPB 1108	tomN	*S. lycopersicum* (tomato)	304	5,619	Cai et al. (2011)
tomato	T1	tomT	*S. lycopersicum* (tomato)	122	5,702	Cai et al. (2011)

[a] Four-letter identifier used in this study; [b] The number of contiguous sequences given for each genome; [c] The number of proteins given for each genome.

is possible to get both the 5′ and 3′ portions of the split gene being annotated as distinct entities (with unique protein accession numbers) within the proteome of that genome.

The presence of these fragments causes error in the inference of orthology via two primary mechanisms. Firstly, fragments can cluster to form an orthologous group that is distinct from the group containing their full length true co-orthologs. This introduces an additional, spurious orthologous group to the analysis and can lead to an overestimate in the number of gains and/or losses. Secondly, protein translations of both the 5′ and 3′ fractions of a split CDS may cluster within the same OG, giving rise to multiple proteins from the same genome within a given orthologous group (i.e. apparent paralogs). This can lead to an underestimate of the core genome size when only genes in single-copy per genome are counted. In addition, fragments that do not cluster in any orthologous group can overestimate the number of lineage-specific proteins per pathovar. To account for these issues, a series of filters (written in PERL) were developed and are briefly discussed below.

To correct for entirely spurious orthologous groups, the consensus sequences of all 2,358 two-member OGs ('twosomes') were used to search a database consisting of the consensus sequences of all 10,320 OGs inferred by ORTHOMCL. Alignments were constructed using T-COFFEE v8.99 with default parameters (Notredame et al., 2000), and BLASTp was run at *E*-value $\leq$ 1e$-$5. Query sequences with a significant hit ($\geq$ 80% amino-acid identity over $\geq$ 80% query length) to another alignment were classified as spurious and the corresponding twosome OG was removed from the analysis. The process was repeated iteratively over increasing OG size classes (threesomes, foursomes etc.) until the method inferred an error rate of $<$ 0.05% in the OG dataset (at tensomes).

To correct for OGs containing multiple fragments of the same CDS, the alignments of all groups that had co-orthologous proteins from the same genome were subjected to further investigation. For alignments that had an average pairwise amino-acid identity of $\geq$ 80% and that showed evidence that the two paralogs were in fact fragments, only the longest fragment of any split protein was retained in the

alignment. A further 499 OGs were removed from all subsequent analyses due to poor alignment and/or ambiguous orthology.

Any sequences with homology to any known insertion-sequence related proteins were removed from the dataset by querying the 59 known *P. syringae* IS-elements (Bardaji et al., 2011a), downloaded from 'IS Finder' at `www-is.biotoul.fr/is.html` (Siguier et al., 2006b), against the combined input proteomic dataset (BLASTp, E-value $\leq$ 1e$-$10). Any OG that contained a sequence with $\geq$ 70% pairwise amino-acid identity was removed from subsequent analyses.

2.3.4 Construction of core and pathovar-specific gene sets

The resultant lists of orthologous groups from the ORTHOMCL analysis were decomposed into putative pan-, core- and lineage-specific gene sets using a further set of custom PERL scripts. Only core genes with no paralogs (i.e., 1-to-1 orthologs) were included in the core gene set. Since the exclusion of a single pathovar from an 'almost-core' OG necessarily excludes that gene from the core set as defined above, all 969 OGs that had a member protein in 26 of the 27 genomes were further scrutinised to ensure the non-existence of an ortholog in the non-participating genome, using tBLASTn to account for the possibility of annotation errors. Briefly, this filter queried each non-participating genome with the corresponding ortholog from the pv. *phaseolicola* str. 1448A (*phas*) genome, using tBLASTn (*E*-value $\leq$ 1e$-$5). All hits that had an 80% or greater amino acid identity over 80% or more of the query length were considered homologous. When pv. *phaseolicola* itself was the non-participating genome, the corresponding ortholog from pv. *tomato* str. DC3000 (*tomD*) was used.

All sequences that did not cluster to form an OG were compiled into an initial lineage-specific gene set. The presence of protein fragments (discussed above) is known to erroneously inflate this gene set (Lefébure and Stanhope, 2007; Lefébure et al., 2010; Richards et al., 2011). Thus, a protein was considered lineage-specific only if it returned no significant hits (other than to self) when queried against the combined genomic dataset, using tBLASTn (*E*-value $\leq$ 1e$-$10).

A binary matrix of gene presence or absence across all OGs and pathovars

was used to plot gene accumulation curves with a custom PERL script; for each comparison, genome input order was randomly permuted 100 times to acquire sample distributions.

2.3.5 Core genome phylogenetics

The genome-wide average evolutionary history for these *P. syringae* strains was inferred using a core-genome concatenation approach. For each core OG, protein sequences were aligned using T-COFFEE. All sites that contained a gap in any sequence were excluded. The presence of truncated sequences in many alignments meant that the exclusion of gap sites sometimes ruled out the entire alignment (i.e., when all alignment columns had at least one gap). Alignments with low identity ($<$ 30% identical sites), short length ($<$ 30 residues) or 100% identity were not included in the concatenation. The remaining 2,140 alignments were concatenated and a maximum likelihood tree was built with PHYML v. 3.0 (Guindon and Gascuel, 2003; Guindon et al., 2010) using the Le and Gascuel (Le and Gascuel, 2008) model of amino-acid substitution with four gamma-distributed rate categories (LG + Γ) and a combination of nearest-neighbor interchange (NNI) and subtree-prune-regraft (SPR) methods of tree topology search; 100 bootstrap replicates were performed to test topological robustness. A maximum likelihood phylogeny based on over 1 million core nucleotide sites was also constructed in PHYML, using a GTR + Γ model with four rate categories, NNI + SPR branch swapping and 100 bootstrap replicates.

Maximum likelihood phylogenies for all 2,140 individual core alignments were constructed at both the DNA and protein level, using the same parameters as above. Topological concordance amongst the set of phylogenies was examined by using the program CONSENSE from the PHYLIP package v3.5 (Felsenstein, 1993) to build a consensus tree (extended majority rule).

2.3.6 Genome content analysis

To estimate a phylogeny based on genome similarity in terms of gene content, a distance matrix was constructed where the distance measure between two taxa is

inversely proportional to the number of genes shared between them (Snel et al., 1999; Kettler et al., 2007; Georgiades et al., 2011). If gain and loss is assumed to affect the flexible genome only, the distance between genomes A and B (d_{AB}) can be calculated as:

$$d_{AB} = 1 - ((S_{AB} - C)/(T - C)) \tag{2.1}$$

where S_{AB} is the number of shared genes, C is the core genome size and T is the total number of proteins in the smaller of the two genomes. To correct for unobservable events along deeper branches in the phylogeny a Poisson correction procedure was applied:

$$d'_{AB} = -\ln(1 - d_{AB}) \tag{2.2}$$

A phylogeny was then inferred using the neighbor-joining method (Saitou and Nei, 1987) with 1000 bootstrap replicates sampling columns of the presence/absence matrix, and recording the consensus support over the bootstrap datasets at each node using consense.

Levels of synonymous (K_S) and nonsynonymous (K_A) divergence between all pairs of taxa within and between the major clades in the core-genome phylogeny were estimated for the same 2,140 core nucleotide alignments (over 1 million sites), using the method of Li (1993), implemented through the R package SEQINR (Charif and Lobry, 2007).

2.3.7 Reconstruction of gene gain and loss

The history of gene gain and loss was analysed using a stochastic mapping method implemented through the GLOOME server (`http://gloome.tau.ac.il/`) (Cohen et al., 2010; Cohen and Pupko, 2010, 2011). The GLOOME algorithm takes as input the phyletic matrix of gene-family presence and absence and a reference phylogenetic tree, and infers probabilities and expectations for all gain and loss events in a per site per branch manner. This method may be particularly suitable for the analysis of genomic datasets that are split into numerous contigs (i.e. draft genomes), since it does not require gene synteny information. In addition, other methods that reconcile

differences between a gene tree and a species tree to infer duplication, transfer and loss events often require fully binary trees—a condition that is often not met in comparative genomics analyses of closely related taxa where there may be poor phylogenetic resolution at the level of the individual gene. Stochastic mapping allows for the total number of gains and losses along a given phylogenetic branch to be calculated as well as the subsequent extraction of specific gain events that have a high associated probability. Probabilistic models of gene gain and loss are superior to parsimony based approaches because the rate of gain and loss is allowed to vary among OGs, and both topology as well as branch lengths are considered in the estimation of associated probabilities (Cohen and Pupko, 2010). The core-genome phylogeny was used as the reference tree, and the total number of gains for each branch was calculated as the probability of gain for each gene, summed across all genes and rounded to the nearest whole number. The procedure was repeated for losses. A branch-specific rate of gene gain/loss was also calculated as the ratio of the number of events (gains or losses) to the length of that branch, equivalent to the number of events per 1% amino acid sequence change of the core-genome.

Using the stochastic mapping results, genes with a gain probability of ≥ 0.8 were extracted on a branch-by-branch basis for functional categorisation. A cut-off of 0.8 was chosen since this was the lowest observed probability value with which lineage-specific genes were inferred to have been gained on the terminal branches of the tree; i.e., a more conservative cut-off value may have excluded some lineage-specific genes from the analysis. For each branch, a reference sequence from each gained OG was chosen randomly and functionally annotated using BLAST2GO v2.5.0 under default settings (Conesa et al., 2005). Gained sequences were initially classified into one of three (mutually exclusive) basic categories: 'functional protein' if the sequence was successfully annotated within BLAST2GO, 'hypothetical protein' if no function was ascribed and 'phage-related' if that sequence had matches to known phage-related proteins. Those sequences that were ascribed functions were further delimited based on their associated gene ontology (GO) terms. Only branches with ≥ 10 gains were analysed, and only GO categories with ≥ 3 sequences are

shown. The 'functional protein' category was chosen for further GO classification to ascertain the underlying functional diversity of gained sequences.

The potential source of recently transferred genes was investigated using a BLAST-based approach. Protein sequences for all genes gained along terminal branches (i.e. recent acquisitions) with a gain probability of ≥ 0.8 were queried against the NCBI non-redundant database using BLASTp (E-value $\leq$ 1e$-$20). For each query, the donor species was inferred to be the most similar match with the lowest BLAST E-value. This approach will miss potential donor species that are not represented within the database, but allows for an approximate estimation of the extent to which HGT occurs within phylogroups, between phylogroups or from other species.

To assess the extent to which effector genes of the T3SS comprise recently acquired genes, all terminally gained proteins were queried against a database of all currently described type III secretion system effectors (T3SEs) using BLASTp. The database was constructed using the T3SE sequence information provided at `www.pseudomonas-syringae.org` and by Wang et al. (2012) and consists of $>$ 1,700 T3SS-related protein products. Matches with $\geq$ 80% identity over $\geq$ 80% of the query length were inferred to be homologous to a T3SE.

2.4 Results

2.4.1 The core and pan genomes of *P. syringae*

We compared the genome sequences of 27 strains of *P. syringae*. Initially, these showed a combined total of 159,748 proteins, of which 148,254 ($\sim$93%) were clustered into a total of 10,320 inferred orthologous groups (OGs) by ORTHOMCL. The relative frequency across genomes of each OG size-class is shown before and after correction for errors (FIG. 2.1). Prior to correction, 11,791 genes ($\sim$53% of the initial estimate of the pan-genome) were inferred to be lineage-specific. After accounting for errors due (primarily) to the presence of protein fragments, this value was reduced by more than three-quarters to only 2,677 genes ($\sim$12% of the initial estimate of the

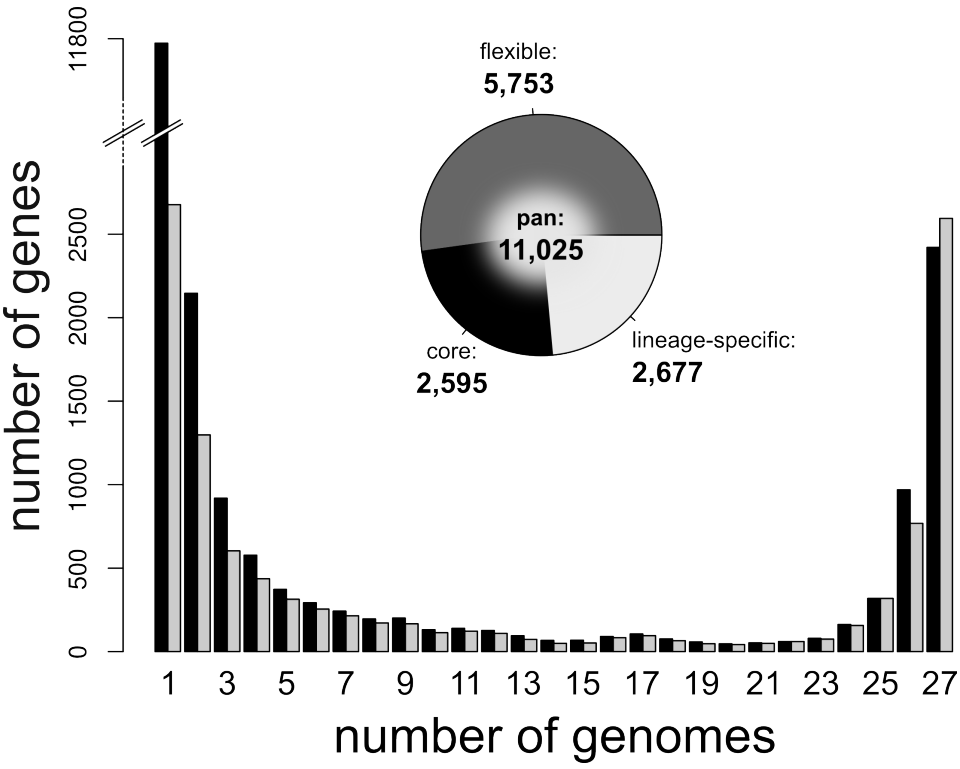


Figure 2.1: Frequency distribution of genes across genomes. Black bars represent the number of genes in each OG size category prior to the various correction procedures outlined in materials and methods; grey bars are corrected values. The pie chart shows the delineation of the *P. syringae* pan genome into its constituent entities: core (i.e., genes present in all strains), flexible (genes with a variable occurrence across strains) and lineage-specific (genes unique to individual strains).

pan-genome). The shape of the corrected distribution suggests that the majority ($\sim$75%) of genes of the pan-genome are found in either very few (≤ 3) or most (≥ 25) genomes, with a paucity of genes distributed in the mid-range of the distribution.

After correction, the size of the pan-genome for all 27 lineages was 11,025 genes (FIG. 2.2, upper panel). However the curve did not tend towards an asymptote, suggesting that a considerable amount of variation (in terms of novel genes) remains to be discovered with the addition of further genomes. The core genome of the entire *P. syringae* species complex was initially estimated to be 2,421 genes. Correction for missed core genes increased this value by 174 to 2,595 genes, or $\sim$52% of the average *P. syringae* genome content. Again, the shape of the curve suggests that the number of core genes will decrease further given additional genome sampling. We considered the effect of missed genes due to incomplete genome sequences by

running an analysis where 'almost core' genes (with a member in $n - 1$ genomes) were included. This increased the estimated core genome to 3,364 genes but had little effect on the shape of the accumulation curve, suggesting that the failure of the original curve to tend towards an asymptote is not due to missed genes. When analysed individually, pan genome estimates for phylogroups 1, 2 and 3 were 6,778, 7,436 and 7,790 genes respectively, and estimates for the core genome were 3,785, 3,481 and 3,477 genes respectively (FIG. 2.2, lower panel). These values are substantially lower (pan) and higher (core) when compared to the corresponding value under the appropriate number of genomes in the full analysis.

2.4.2 *Pseudomonas syringae* phylogenetics

The core genome phylogeny for *P. syringae* was estimated from 2,140 concatenated proteins totalling 351,468 aligned residues (FIG. 2.3). Pathovar *maculicola* str. ES4326 (*Pcan*) was defined as the outgroup to root the core-genome phylogeny, since it has been reported that this pathovar was originally misidentified and is in fact a strain of *Pseudomonas cannabina* (Baltrus et al., 2011), a closely related species outwith the *P. syringae* species complex (Parkinson et al., 2010). Bootstrap support of branching patterns was generally very high with only one partition showing less than 75% support. The phylogeny delineates the 26 ingroup lineages into four clades, in agreement with the phylogroup notation used previously (e.g. Sarkar and Guttman, 2004; Hwang et al., 2005). A phylogeny constructed from nucleotide sequence alignments of the same 2,140 genes produced a similarly well resolved tree with the same topology (not shown).

Based on the core genome, maximum pairwise K_A (the number of substitutions per nonsynonymous site) within phylogroups was calculated to be 0.018, 0.022 and 0.011 for phylogroups 1, 2 and 3 respectively. Between phylogroups, average estimates were 0.100, 0.094 and 0.067 for phylogroup 1 versus 2, 1 versus 3, and 2 versus 3 respectively. Maximum pairwise values for K_S (number of substitutions per synonymous site) within groups were 0.299, 0.312 and 0.124 for phylogroups 1, 2 and 3 respectively, while average K_S between groups were 1.584, 1.644 and 1.218

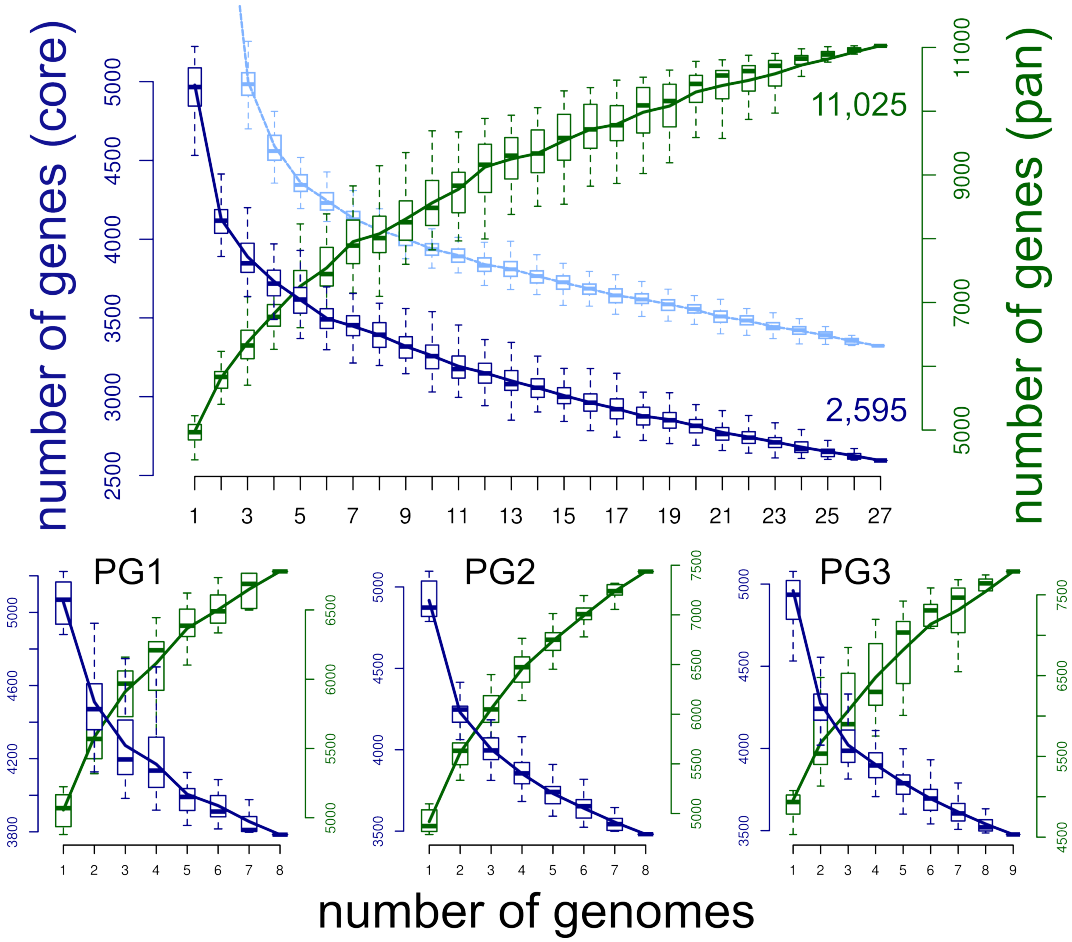


Figure 2.2: Core and pan genomes of *P. syringae*. The upper figure shows gene accumulation curves for the core (dark blue) and pan (green) genomes for the combined dataset (27 genomes). The light blue curve shows the core estimate when almost core genes (i.e., genes present in 26 genomes) are also included (comparisons of < 3 genomes are not shown). The dark blue axis on the left shows the number of genes of the core genome, and the green axis on the right shows the number of genes of the pan genome (note the different scales). Boxes show the interquantile range of 100 random permutations of genome input order for the given number of genomes in comparison; the horizontal line shows the median value. The lower figure shows the equivalent results when the data is partitioned into the three major phylogroups. PG1, PG2 and PG3 represent phylogroups 1, 2 and 3 respectively.

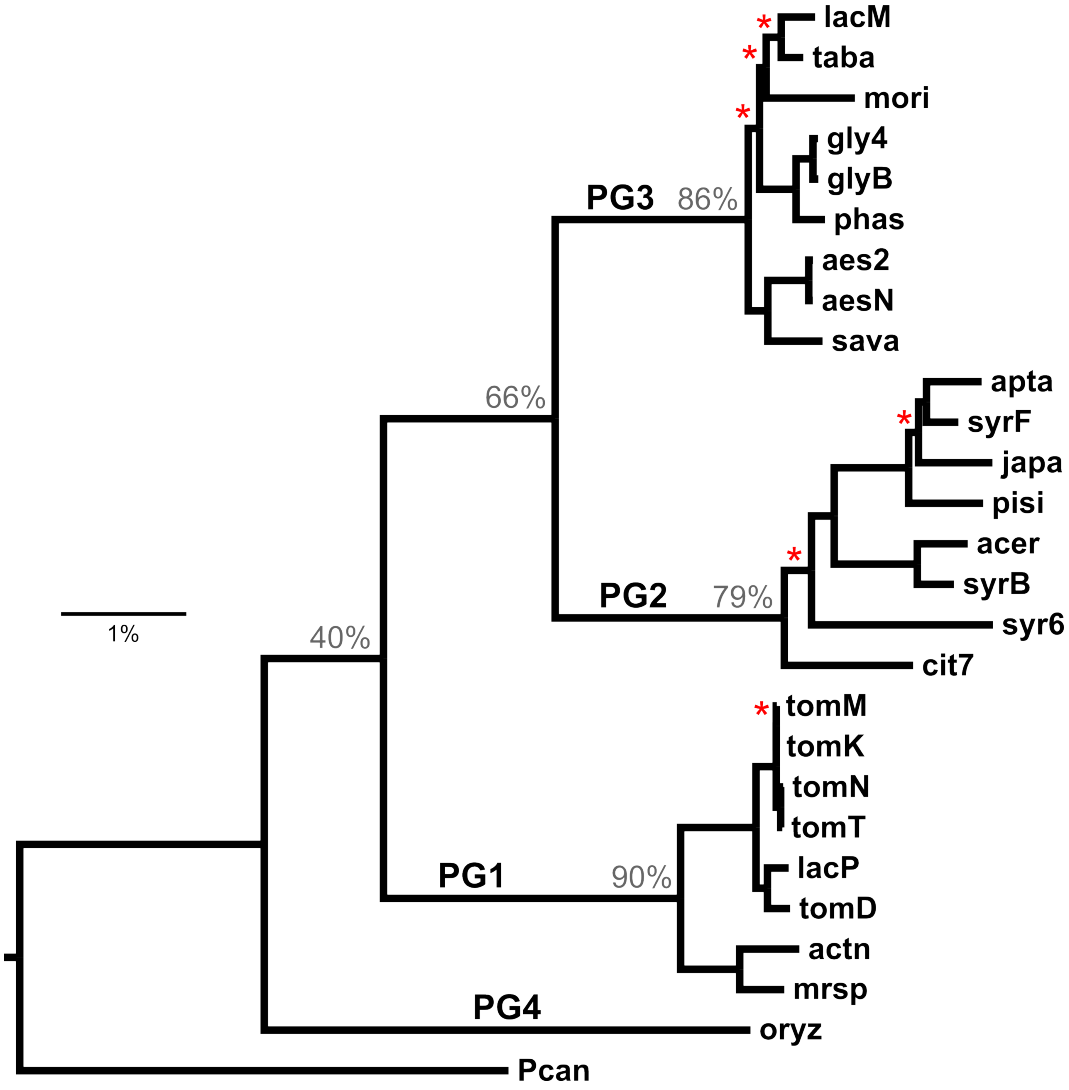


Figure 2.3: Core genome phylogeny for *P. syringae*, derived from 2,140 concatenated protein alignments. All internal branches have 100% bootstrap support except where indicated with asterisks. Consensus support from individual gene trees is shown with grey numbers on major branches, and phylogroups PG1 to PG4 are indicated. Tree is rooted with pv. *maculicola* str. ES4326 (*Pcan*). Scale bar indicates 1% amino acid replacements per site.

for phylogroup 1 versus 2, 1 versus 3 and 2 versus 3 respectively.

An analysis of consensus branching patterns within the set of individual core gene-trees was performed to assess the level of support for the topology of the concatenated phylogeny at the individual gene level. An extended majority consensus tree is shown in supplementary FIG. S1 (APPENDIX A). Consensus support for phylogroups 1, 2 and 3 was 90%, 79% and 86% respectively, and overall the consensus topology is in agreement with that of the core genome phylogeny at all partitions bar the relative placement of the two phylogroup 2 strains *cit7* and *syr6*; this discrepancy is reflected in the low bootstrap value (54%) at this node in the core-genome tree. Two-thirds of gene trees support the clustering of phylogroups 2 and 3, but only 40% support the position of PG1 as a sister to these. This appears to be due to uncertainty about the placement of PG4 (represented by *oryz*); indeed an earlier analysis based on four genes clustered PG4 with PG2 and PG3 (Hwang et al., 2005).

Relationships among strains based on genome content similarities were also analysed (FIG. 2.4). Overall, there is a high degree of topological congruence between the genome content analysis and the core genome phylogeny. Phylogroups 1, 2 and 3 are reproduced monophyletically with 100% bootstrap support, while there is also substantial agreement in branching patterns within phylogroups. An exception is the placement of the PG4 strain *oryz*, which clusters with PG2 in the genome-content analysis (red asterisk). The major difference between the core-genome and genome-content phylogenies can be seen in the relative branch lengths with respect to phylogenetic depth within the two trees. Even after correction for multiple hits, the deeper branches in the genome-content analysis are very short while the terminal branches are relatively extended.

2.4.3 Gene gain and loss

We used a stochastic mapping method to infer the number of gains and losses along all branches of the *P. syringae* core-genome phylogeny (FIG. 2.5). An underlying assumption of this method is that gain and loss events occur independently of one

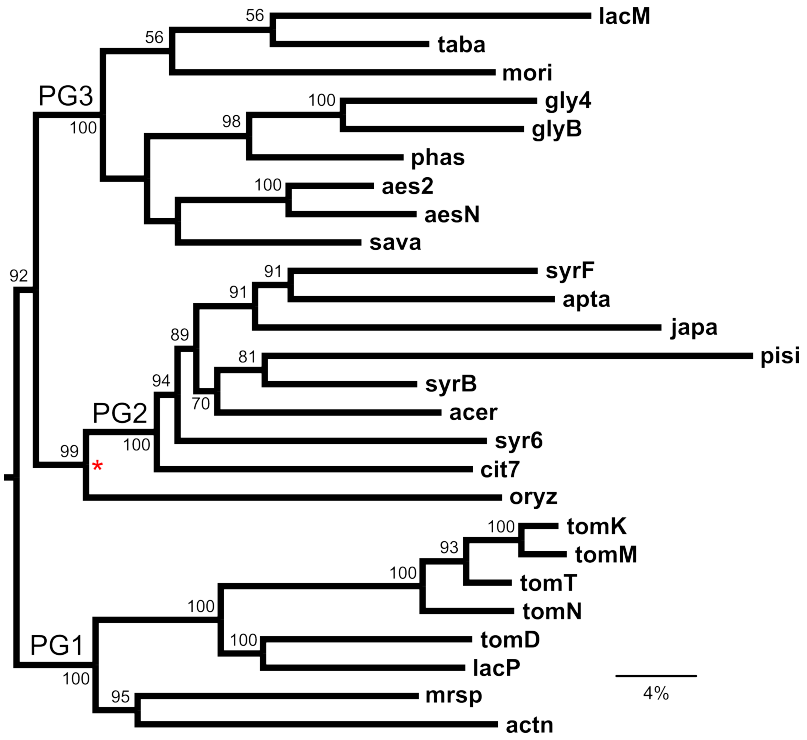


Figure 2.4: Genome content analysis. Numbers indicate percentage consensus support for that bifurcation from 1000 bootstrap replicates. Phylogroups PG1, PG2 and PG3 are indicated. The red asterisk shows the only incongruence relative to the core-genome phylogeny. Scale bar indicates 4% genomic dissimilarity, i.e. percentage difference in terms of number of genes shared. Tree is rooted with *Pcan* (not shown).

another. In reality this is unlikely to be the case, since many genes may be gained (or lost) in a single transfer event, e.g. the acquisition of a plasmid. Thus, the results shown here likely overestimate the number of individual transfer events involved.

There is considerable variation in gene gain and loss among branches. Generally, values are higher along terminal branches and decrease along branches deeper in the phylogeny. The pathovar with the highest inferred number of terminal gains was *lacM* (1,229), followed by *pisi* (683) and *oryz* (527). On average, 6.4% of a pathovar's total genome has been gained recently (subsequent to the divergence of a lineage from its nearest sister taxa), but this value increases to 13.2% when considering this proportion relative to the size of the flexible genome only. On terminal branches, lineages have lost 9.2% of the current genome size on average, and the three pathovars that have undergone the greatest number of losses are *oryz*

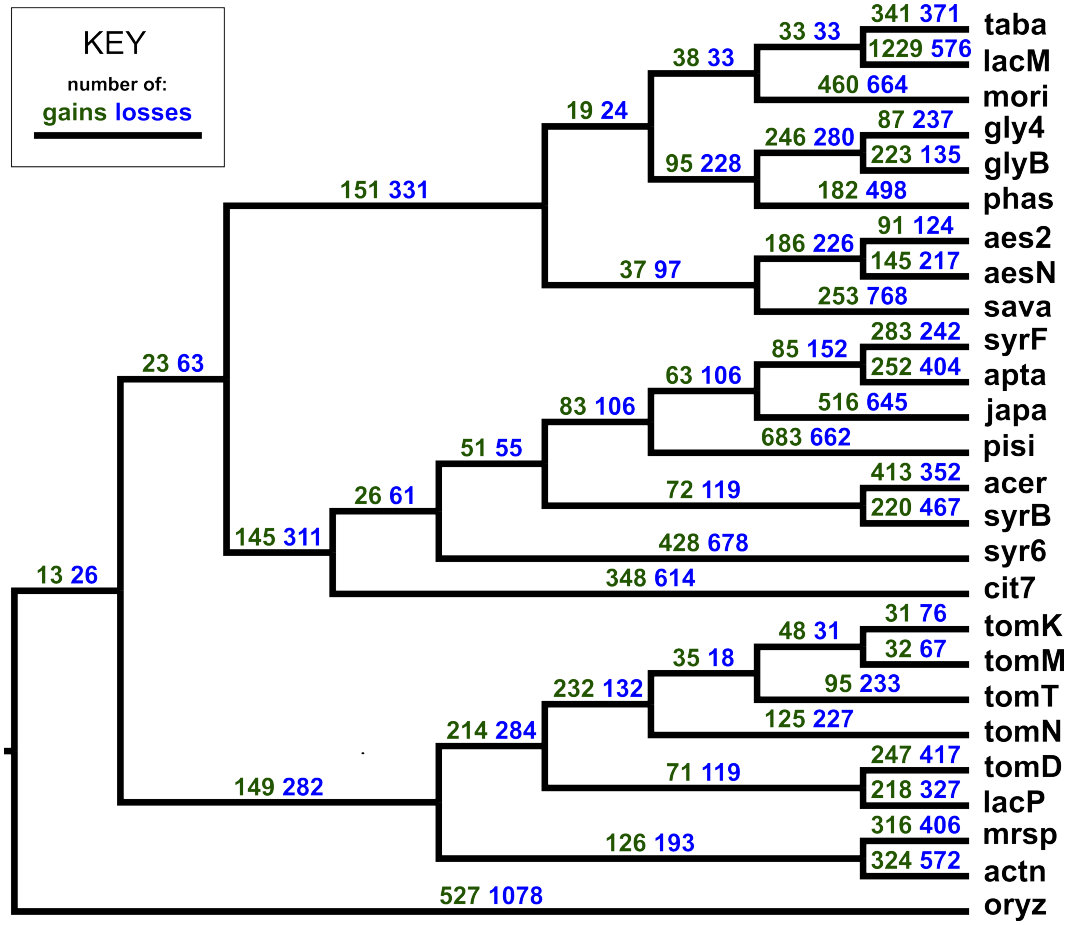


Figure 2.5: Gene gain and loss in *P. syringae*. Inferred numbers of gains (green) and losses (blue) are given above each branch. Lineage-specific genes are included in gains along terminal branches. Values in all cases are rounded to the nearest whole number of events. Topology is based on Fig. 2.3; branch lengths are not to scale.

(1,226), *sava* (768) and *syr6* (678).

High values of gain and loss for individual pathovars may be the result of either biologically relevant variation in genome content, or an artefact of phylogenetic sampling. For example, *P. syringae* strains are known to harbour plasmids and the exceptionally large number of recent gains in *lacM* can be explained by the discovery of a ∼1 Mb plasmid in this genome (Baltrus et al., 2011). Similarly, evidence for the presence of plasmids has been found for the *pisi* genome (Baltrus et al., 2011), although the exact size and gene content are not known. However, while *oryz* is also thought to contain plasmids (Reinhardt et al., 2008), the high number of gains and losses seen in this pathovar is also due to the relatively long branch that separates it from other pathovars; *oryz* has had more time to accumulate differences in genome content.

Comparisons between absolute numbers of gains and losses per branch can be misleading, however, since these values are subject to the effects of phylogenetic sampling. To account for this, we calculated a rate of gain and loss relative to the amount of amino-acid sequence divergence that has occurred along each branch of the core-genome phylogeny (supplementary FIG. S2; APPENDIX A). To investigate the evolutionary dynamics of gain and loss at different depths of the phylogenetic tree, we plotted these branch-specific rates against the phylogenetic depth of that branch, measured as the average distance from the midpoint of that branch to all descendent tips (FIG. 2.6). The apparent rates of gain and loss decrease exponentially with respect to increasing phylogenetic depth, reflecting the much higher levels of observable gain and loss along branches closer to the tips of the tree (i.e., more recent in time).

2.4.4 Source and functions of gained genes

The extent of HGT within and between phylogroups was assessed using BLAST to determine the most similar sequence for all genes gained along terminal branches across the phylogeny (FIG. 2.7a). In phylogroups 2 and 3, the primary source of gained genes appears to have been other pathovars within the same phylogroup. A

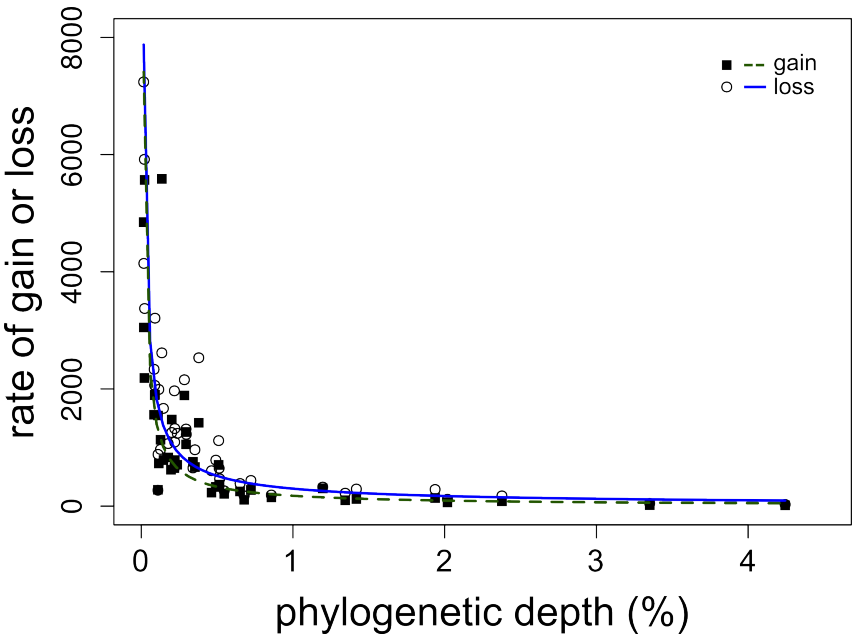


Figure 2.6: Rate of gene gain (filled squares) and loss (open circles) versus phylogenetic depth. Phylogenetic depth is calculated as the average distance (percent amino acid replacements per site) from the midpoint of each branch of the core genome phylogeny to its descendent tips; the *y*-axis shows the branch-specific rate of gain or loss expressed as the number of inferred events per 1% amino acid divergence of the core genome phylogeny.

significant contribution also comes from pathovars in other phylogroups, as well as from other *Pseudomonas* species outwith the *P. syringae* group. Taking phylogroups 1, 2 and 3 together, the 10 most frequent putative donor genera were (in rank order) *Pseudomonas*, *Burkholderia*, *Vibrio*, *Escherichia*, *Xanthomonas*, *Ralstonia*, *Yersinia*, *Salmonella*, *Acidovorax* and *Pectobacterium*. Transfers from these bacterial groups represent the great majority (∼86%) of the total number of recently transferred genes, and many contain species which are known to be associated with plant, water or soil habitats. For example, both *Ralstonia* and *Xanthomonas*, themselves much-studied pathogens of other plant species (Mansfield et al., 2012), appear as frequent donors of genetic material into the genomes of phylogroup 2 and 3 pathovars. In addition, *Pectobacterium* contains phytopathogenic species that cause disease in both herbaceous and woody hosts (Toth et al., 2003), *Acidovorax* are pathogens of cucurbits (Schaad et al., 2008), while *Polaromonas* species are often associated with water and can be found in alpine meltwater (Jeon et al., 2004; Margesin et al., 2012), a known

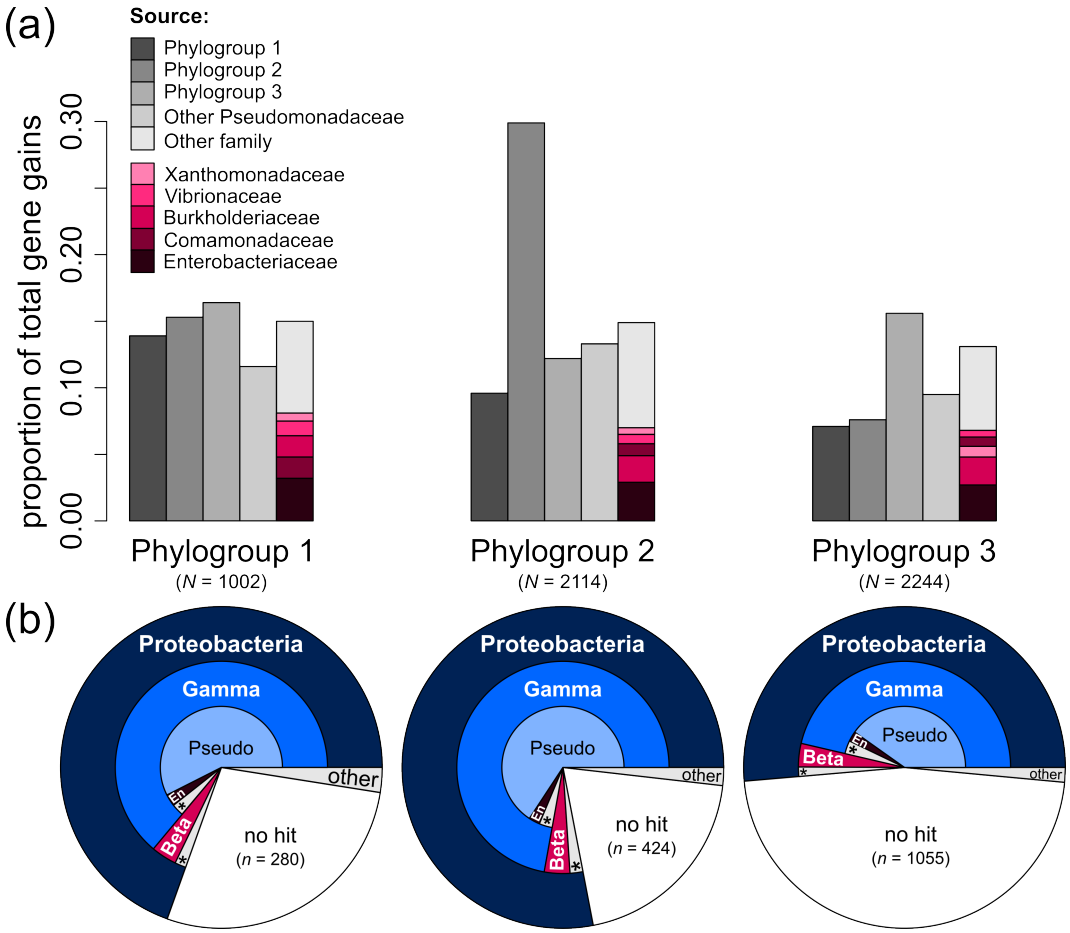


Figure 2.7: Sources of gained genes. **(a)** Bar charts detail the fine-scale sources of genes gained recently in phylogroups 1, 2 and 3, as identified by best BLAST hit (*E*-value $\leq$ 1e$-$20). The *x*-axis represents the proportion of total terminally acquired genes for each phylogroup. The category 'other family' is further decomposed to show the family of the top 5 donors for each phylogroup (pink shaded boxes). **(b)** Broader taxonomic classification of gained genes. 'Pseudo.', Pseudomonadaceae; 'En.', Enterobacteriaceae; asterisks denote 'other' categories.

habitat of *P. syringae* (Morris et al., 2008). Species belonging to the Burkholderiaceae, from the Betaproteobacteria, are among the top five putative donors from outwith the Pseudomonads for all three phylogroups, highlighting the potential contribution from relatively distant families.

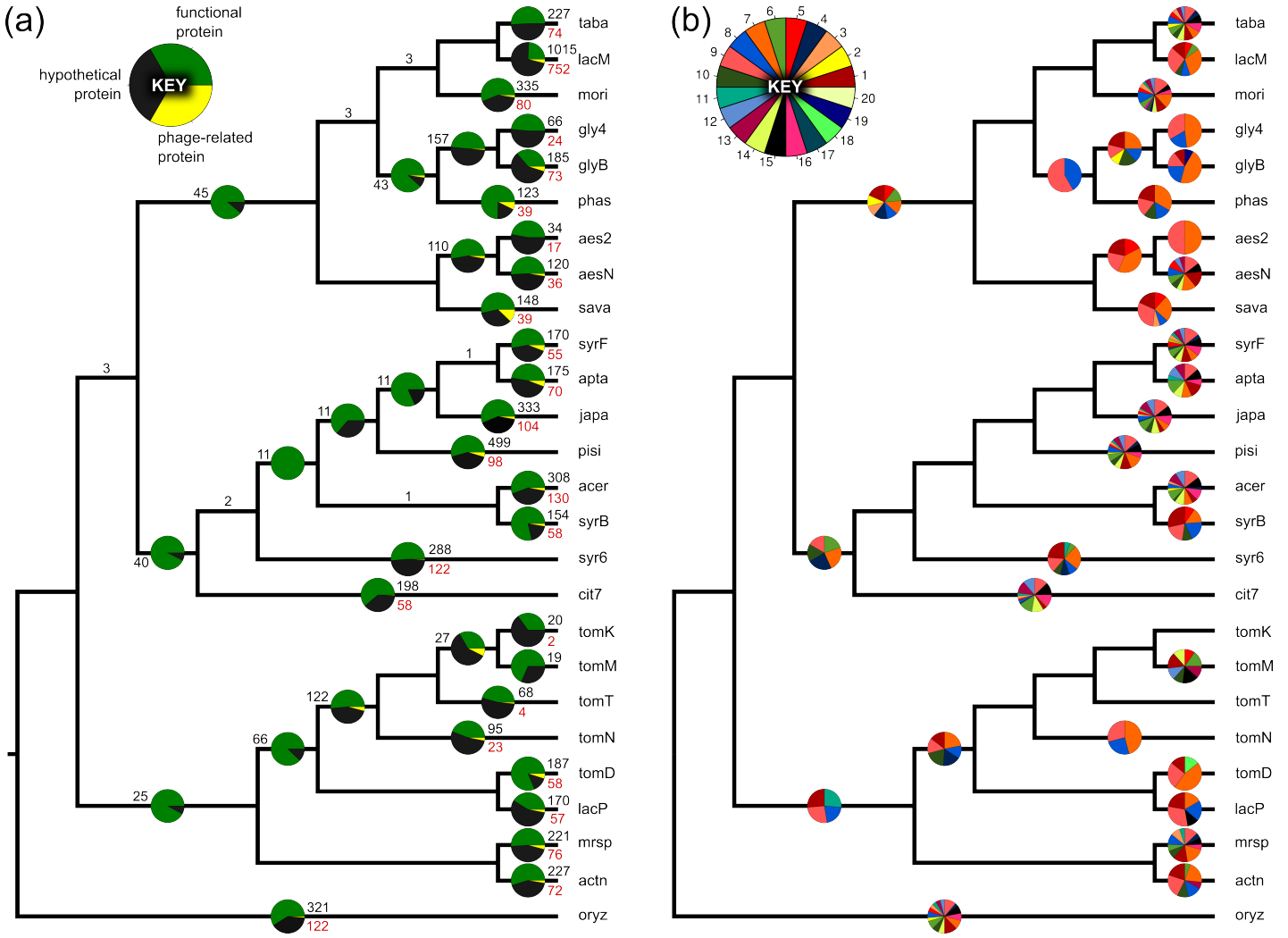
Further analysis of putative donor species from other genera reveals the phylogenetic breadth of potential HGT sources (FIG. 2.7b). For all phylogroups, the majority of genes are donated from species within the Gammaproteobacteria. A smaller proportion of genes are also acquired from more distant taxonomic groups such as

the Alpha- and Betaproteobacteria, and from other phyla such as the cyanobacteria. A significant fraction of genes in all phylogroups show 'no hit', indicating that many recently gained genes have not been previously sampled and are of unknown origin or function. The proportion of no hits in PG3 is approximately twice that relative to PG1 and PG2 due to the inflated number of genes specific to *P. syringae* pv. *lachrymans* str. M301315 (*lacM*), presumably encoded on the ∼1 Mb plasmid unique to this strain, the majority of which had no BLAST matches. The source of this large lineage-specific fragment of DNA remains to be elucidated.

Across all branches, the protein sequences of 6,378 genes gained with a probability of ≥ 0.8 were delimited into three basic categories based on gene ontology: 'functional protein', 'hypothetical protein' and 'phage-related protein' (FIG. 2.4.4a). Across all branches, 59.6%, 37.8% and 2.6% of gains were labelled as functional, hypothetical or phage-related respectively. A further categorisation of those sequences that were successfully annotated (i.e., 'functional proteins', as above) is provided in FIG. 2.4.4b.

Proteins inferred to be terminally acquired were queried against a database of known type III secretion system effectors (T3SEs) to ascertain the extent of recent transfer of this important family of genes. At the 1e−5 BLAST *E*-value threshold, this yielded 13, 34 and 36 hits for phylogroups 1, 2 and 3 respectively. However, applying the 80% identity over 80% query coverage homology rule narrowed this down to 8, 7 and 12 T3SEs for phylogroups 1, 2 and 3 respectively, accounting for a very small fraction (∼0.5%) of all recently gained genes. In addition, the terminal acquisition of T3SEs was not distributed evenly among pathovars. For example,

Figure 2.8 ***(following page)*:** Functional categorisation of gained genes. **(a)** Simple categorisation. Numbers above each branch indicate the total number of gains (≥ 0.8) for that branch; red numbers on terminal branches indicate the number of terminal gains that are lineage-specific, i.e. unique to the given lineage. Pies along each branch represent the classification of genes into one of three basic categories: 'functional protein' (green), 'hypothetical protein' (black) or 'phage-related protein' (yellow). Only branches with ≥ 10 gains are delimited into pies. The topology of the cladogram is based on the core-genome phylogeny (branch lengths not to scale). **(b)** Gene ontology categorisation (molecular function, GO level 3). Genes from (a) in the functional protein category are further characterised based on gene ontology (GO) information. Only GO terms with ≥ 3 sequences are shown. GO terms corresponding to the numbers in the key are given in supplementary Table S3 (Appendix A).


(a)
functional protein
hypothetical protein
KEY
phage-related protein
taba
lacM
mori
gly4
glyB
phas
aes2
aesN
sava
syrF
apta
japa
pisi
acer
syrB
syr6
cit7
tomK
tomM
tomT
tomN
tomD
lacP
mrsp
actn
oryz
(b)
KEY

within phylogroup 3, six were acquired along the terminal branch leading to *mori*, four along the branch leading to *lacM* and two along the branch leading to *phas*. Only one T3SE protein, HopBE1 from *acer* (accession EGH71060.1), was both terminally acquired and strain-specific.

2.5 Discussion

We have investigated both the scale and dynamics of genome fluctuation in *P. syringae*. We demonstrate a history of extensive gene gain and loss that can rapidly alter the gene-content of individual *P. syringae* lineages. Below, we discuss the wider implications of these results within the context of the role of HGT and gene loss in the differentiation and diversification of bacterial lineages, as well as the consequences of such fluctuation on the evolution of *P. syringae*. First, we consider some of the potential issues surrounding the use of whole-genome draft sequence data in comparative analyses that may have been previously underestimated in other studies.

2.5.1 The extent of potential errors in draft genomes

We applied a suite of quality control and data correction procedures to ensure the resultant patterns of orthology and gene occurrence were as accurate as possible. The confounding effects of protein fragments, caused when a contig terminates in the middle of a CDS, were of particular significance. Our observations emphasize the care required when drawing inferences from uncorrected patterns of orthology constructed using multiple (often hundreds of) contig draft genome sequences in conjunction with popular orthology methods such as ORTHOMCL.

Not accounting for this source of error can introduce considerable bias in a number of common comparative genomics analyses. For example, we estimated that ∼77% of inferred lineage-specific proteins and ∼13% of the pre-corrected number of orthologous groups could be apportioned to error. Even fundamental descriptive metrics for a genome, such as the number of proteins it encodes, can be substantially

overestimated by the presence of protein fragments. For example, the combined data correction procedures applied here reduce the number of proteins encoded by the *P. syringae* pv. *pisi* genome from 9,160 to 5,085, a reduction of ∼44%—almost one half of the original number. The mean reduction in proteome size across the 27 genomes in the present study is ∼17%. These observations simultaneously reveal the possible level of error in the annotation of genome sequences, and the potential for the use of orthology to reduce the problem.

Errors in the observed patterns of orthology among proteins can adversely affect the estimation of core and pan genome sizes, and can lead to the miscalculation of the level of paralogy and gene duplication within genomes. For example, the core genome of *P. syringae* was recently estimated at 1,856 genes, approximately 740 genes smaller than that reported here (O'Brien et al., 2012). Problems caused by protein fragments may be one reason for this discrepancy. In addition, a recent study on genome fluctuation in cyanobacteria suggests that a large number of paralogs within the genomes of certain groups of cyanobacteria may be indicative of a higher rate of gene duplication (Larsson et al., 2011). However, it is also possible that a large fraction of the apparent paralogs may be attributable to the presence of protein fragments.

2.5.2 Phylogenetics of the core and flexible genomes

We inferred the evolutionary relationships among pathovars based on both the evolution of the core genome and gene-repertoire similarity of the flexible genome. Repertoire similarity has been predicted to be a poor indicator of phylogenetic relatedness, where the rapidity of gene turnover is expected to blur inherited patterns of gene presence and absence (Hao and Golding, 2006; Vernikos et al., 2007; Touchon et al., 2009). Thus, topological discordance between phylogenies based on these two genome components is thought to be an indicator of extensive HGT among lineages (Shapiro et al., 2012). However, the congruence observed between the *P. syringae* core-genome and genome-content phylogenies suggests that, despite high rates of HGT and gene loss, an accurate and robust phylogenetic signal is recoverable from

patterns of presence and absence among genes of the flexible genome.

Varying degrees of topological congruence between the core and flexible components of bacterial genomes have been reported elsewhere. For example, an investigation of genome evolution in *Prochlorococcus* (Kettler et al., 2007) shows near perfect congruence, whereas studies on *Vibrio* (Shapiro et al., 2012) and *Rickettsia* (Georgiades et al., 2011) show near total discordance. Reasons for this variation may include differences in the rate and extent of genome fluctuation experienced by these groups, the level of divergence among lineages or perhaps ecological factors associated with these different species. However, both *Prochlorococcus* and *Vibrio* are ocean-dwelling species where the opportunity for HGT is likely to be relatively high, yet they show opposite patterns. Moreover, *Rickettsia* species are obligate intracellular parasites of eukaryotic hosts with presumably limited opportunity for gene gain. However, *Rickettsia* genomes are known to be shrinking in size (Georgiades et al., 2011), and so it may be the case that the genome-content topology is driven by large-scale differential gene loss as these species adapt to different hosts. Patterns of divergence also fail to correlate with the level of congruence. These groups of bacteria are generally highly similar, having 16S ribosomal RNA nucleotide identities ranging between ∼97% (*Prochlorococcus*) and 100% (*Vibrio*). Contrary to what might be expected, the group with the greatest 16S divergence (*Prochlorococcus*) shows the most congruence between phylogenies despite presumably having more time to accumulate differences in genome content. Furthermore, *P. syringae* and *Rickettsia* differ dramatically in the level of congruence despite having similar 16S identities (both ∼99%). In all cases, the lack of signal seems not to be the issue, since all groups show considerable variation in genome repertoires among lineages (Kettler et al., 2007; Georgiades et al., 2011; Shapiro et al., 2012). Certainly, *P. syringae* provides a clear example that a high level of genome fluctuation does not preclude a high degree of congruence between trees derived from the core genome and genome content comparisons.

2.5.3 Species definition in *P. syringae*

Based on over 1 million orthologous nucleotide sites of the core genome, we estimate that the three *P. syringae* phylogroups are as diverged from each other as other taxa classified as distinct species, or even genera. For example, K_A between the sister clades PG2 and PG3 is estimated to be 0.06 substitutions per site. This value is greater than that for a number of other species comparisons, including *Escherichia coli–Salmonella enterica* and *Neisseria gonorrhoeae–N. polysaccharea* (Ochman et al., 1999). K_A divergence for comparisons between phylogroups 1 and 2/3 are greater yet, and are comparable with K_A between two other '*bona fide*' *Pseudomonas* species, *P. aeruginosa* and *P. putida* (FIG. 2 of Ochman et al., 1999). Values of K_S between phylogroups are all > 1, suggesting that substitution at synonymous sites is approaching saturation. Although high K_S values are unreliable as point estimates, species-pairs in a number of other genera (including *Buchnera*, *Mycobacterium*, *Bacillus* and *Pseudomonas*) all have lower (< 1) estimates of K_S (Ochman et al., 1999). In addition, consensus analysis from the set of individual core gene trees shows a high percentage of core genes support the delineation of these lineages into three major groups (90%, 79% and 86% for phylogroups 1, 2 and 3 respectively), suggesting that the recombination of core alleles between phylogroups does not occur at a level sufficient to erode the boundaries between groups. As such, and notwithstanding the ongoing debate as to the validity of such terms (Lawrence, 2002; Doolittle and Papke, 2006; Caro-Quintero and Konstantinidis, 2011; Hanage, 2013), *P. syringae* phylogroups may be effectively described as nascent species.

There has been ongoing debate regarding the delimitation of the *P. syringae* complex into distinct species (e.g. Gardan et al., 1999; Sarkar and Guttman, 2004; Bull et al., 2010). Sarkar and Guttman (2004) suggest that due to the ecological similarity between pathovars across phylogroups, the sharing of core alleles among phylogroups and evidence of extensive HGT between phylogroups, *P. syringae* should remain as a single species. However, the results presented here suggest that these factors may not necessarily hinder speciation, since neither is sufficient to halt the accumulation of divergence from either the core or the flexible components

of the *P. syringae* pan genome. We suggest that the data presented in this study support the delineation of *P. syringae* into a number of distinct species based on the phylogroup partitioning that is already well established. The strains *gly4*, *glyB* and *sava* within phylogroup 3 have been previously referred to as strains of *Pseudomonas savastanoi* (e.g. Bull et al., 2010; Rodríguez-Palenzuela et al., 2010; Qi et al., 2011; Ramos et al., 2012); clearly, the other pathovars within PG3 (*taba*, *lacM*, *mori*, *phas*, *aes2* and *aesN*) should be distinguished by the same species name (although see Bull et al. (2010), who state that PG3 strains should be classified as *P. amygdali* if unification were to occur since *P. savastanoi* and others are later synonyms). The type strain for *P. syringae*, NCPPB 281, clusters in phylogroup 2 along with the complete genome *P. syringae* pv. *syringae* str. B728a (*syrB*) (Parkinson et al., 2010), while it appears that species names for phylogroups 1 and 4 have not yet been proposed.

2.5.4 Genome dynamics and niche adaptation in *P. syringae*

The quantification of gene gain and loss suggests that both have been extensive in the evolutionary history of the *P. syringae* species complex, and that the genomes of these pathovars may undergo considerable fluctuation even over relatively short evolutionary time-periods. A large fraction of the total gain and loss was observed along terminal branches leading to individual strains, in agreement with previous studies investigating genome dynamics in other bacterial species (Hao and Golding, 2006; Kettler et al., 2007; Lefébure and Stanhope, 2007; Touchon et al., 2009).

When absolute numbers of gene gains and losses are compared to the rate of sequence divergence in proteins encoded by the core-genome, the results are striking. At shallower phylogenetic depths (representative of terminal branches), we estimate that hundreds or thousands of genes may be gained (and lost) in the same time period as a 1% amino acid divergence. For example, the terminal branch leading to the complete genome of *P. syringae* pv. *phaseolicola* str. 1448A (*phas*) has gained 182 genes over a divergence of about 0.25% (equivalent to 728 genes over 1% divergence). Estimates of rates of gene loss are of at least the same magnitude. However, across the *P. syringae* complex as a whole, more than 50% of the genome

remains as core, despite divergence values up to nearly 10%. This suggests, in agreement with other studies, that the majority of gained genes are transient and are quickly discarded from recipient genomes (Tettelin et al., 2005; Hao and Golding, 2006; Kuo and Ochman, 2009b; Touchon et al., 2009). This leads to a perspective bias, where the rates of gene gain and loss both appear to decrease with increasing phylogenetic depth, because over longer time periods only a fraction of older gains have persisted and remain observable (Rocha, 2008).

Based on seven closely related genomes of the *Bacillus cereus* (Bc) group with divergences of the order of approximately 1% nucleotide substitutions per site, Hao and Golding (2006) estimated the recent rate of turnover to be such that a gene may be gained and lost $\sim$5 times during the time required to observe a single nucleotide substitution per site, and conclude that the rate of transfer is at least comparable to that of nucleotide replacement. It is clear that the estimated rate of gene turnover relative to sequence divergence is highly dependent on the phylogenetic depth over which it has been calculated. However, using a core genome tree based on nucleotide sequences, we calculate the rate for *P. syringae* terminal branches that are equivalent length to that of the Bc group to be perhaps four orders of magnitude greater. This highlights the predominance of HGT as an adaptive force for *P. syringae*, as well as the potential scale of the difference in rates of HGT between different bacterial groups.

We also observe an excess of gene losses relative to gene gains for the majority of branches of the core genome phylogeny. Taken at face value, this suggests a recent contraction of the *P. syringae* genome. However, given that extant strains of *P. syringae* have above-average sized genomes compared to other Gammaproteobacteria, it seems unlikely that they have undergone prolonged genome reduction. It has been demonstrated that methodological biases or data issues can overestimate the rate of either gain or loss (Zhaxybayeva et al., 2007; Hao and Golding, 2008); thus, it may be the case that our approach has either overestimated gains or underestimated losses.

Generally, the most frequent donors of genetic material appear to have been lineages within the same phylogroup as the recipient genome. Smaller contributions

have also come from the other phylogroups and from other Pseudomonads. This might be expected, since the rate of homologous recombination decreases exponentially with sequence divergence (Majewski, 2001), reducing the incidence of HGT when genes are carried into a new genome via the homologous recombination of flanking regions (Touchon et al., 2009; Polz et al., 2013). Genes arriving from outwith the *Pseudomonas* genus have been acquired from a diverse range of donor species, many of which are known to be soil-inhabiting, plant-associated or present in water systems. Gene ontology analysis shows that gained genes are drawn from a diverse range of functional categories, and are therefore likely to contribute a correspondingly diverse range of novel phenotypes to recipient lineages. *P. syringae* is well known for its ability to exist in a multitude of both plant and environmental niches, and the results presented here suggest that such ecological versatility is facilitated by high rates of gene gain into the flexible genome.

These results suggest an underlying structure to gene gain in *P. syringae* that is based on both phylogeny and ecology. An attractive extrapolation of such structuring of HGT is the hypothesis of 'niche-specific gene pools' that contribute to adaptation within an ecologically determined framework, via the transfer of genes that are selectively advantageous within a certain niche (Popa et al., 2011; Smillie et al., 2011; Polz et al., 2013). In the case of plant-pathogenic lineages of *P. syringae*, this framework may be determined by the ecology and population structure of the numerous plant host species on which these pathovars thrive. In this way, the expansion of *P. syringae* lineages onto a new host species may occur when newly acquired genes from other species that already inhabit that environment prove to be selectively advantageous to the recipient lineage, thus facilitating a host jump.

2.5.5 Divergence via differential adaptation

As the extent and prevalence of HGT has become increasingly evident, much discussion has focused on the implications of HGT for the phylogenetic reconstruction of prokaryotic lineages and the validity of the species definition. More recently however, models have emerged in which HGT can be an active facilitator (rather

than a inhibitor) of the evolution of population structure, ecological differentiation and speciation (Ochman et al., 2005; Shapiro et al., 2012; Polz et al., 2013). These ideas rely on the understanding that although rates of HGT and gene loss can be very high, the underlying dynamics of this fluctuation are highly structured in such a way that actively promotes the differentiation of sub-populations of bacteria into distinct genotypic clusters.

HGT and gene loss can directly facilitate the accumulation of divergence between bacterial populations via the differential adaptation of lineages into distinct ecological niches (Ochman et al., 2005; Polz et al., 2013). This mechanism may be operating in the evolution of *P. syringae*, and may explain the broad host range of the *P. syringae* species complex while maintaining the host-specificity of individual pathovars. The genomes of *P. syringae* show a turbulent history of genome fluctuation, with evidence of high levels of gene gain and loss in all lineages across the species complex. Gained genes are functionally diverse and appear to be gained from sources that are both phylogenetically and ecologically structured. The degree of host specificity shown by many individual pathovars suggests that their gene repertoires are specifically adapted to a given niche environment (Morris et al., 2000; Hunter and Taylor, 2006; Sarkar et al., 2006). It is likely that this genomic 'tailoring' is achieved through the continuous uptake of a large number of genes via HGT, but the retention of only those few niche-specific genes that confer an advantage in a given host environment.

In *P. syringae*, much attention has been focused on the contribution of effector genes of the type III secretion system (T3SEs) towards adaptation and pathogenicity. Although the distribution of the $\sim$57 currently described effector families across the *P. syringae species* complex has been well characterised, the relationship between the T3SE repertoire and adaptation is complex and remains to be fully elucidated (Sawada et al., 1999; Ma et al., 2006; Lindeberg et al., 2009; Baltrus et al., 2011; Lindeberg et al., 2012). For example, strains that are pathogenic on the same host species can have divergent effector complements (Almeida et al., 2009; Baltrus et al., 2011; O'Brien et al., 2012), suggesting that *P. syringae* lineages are able to adapt to a

given environment using multiple variations of effector repertoires (Lindeberg et al., 2012). The acquisition of novel T3SEs may be indicative of adaptive transitions into new niche environments (Sarkar et al., 2006). If T3SEs were solely responsible for the adaptive diversification of *P. syringae* pathovars, it might be expected to find a unique repertoire of T3SEs for each transition onto a different host species, and that recent host jumps correspond to recent gains of novel T3SE genes. However, we find that not all pathovars have recently gained T3SE genes, and that of those strains that show evidence of recent T3SE gain, only one was also specific to that lineage. Taken together, these observations suggest that host adaptation and the evolution of host-specificity may be facilitated by the acquisition of genes that are not directly involved in the T3SS, or by molecular adaptations in virulence genes that are already present within the genome.

When acquired genes confer the ability of that lineage to occupy a new ecological niche, such as a new host species, a novel disease can be the result. In the case of *P. syringae*, a number of newly emergent diseases have been reported in the last decade, including bleeding canker of horse chestnut, caused by *P. syringae* pv. *aesculi* (Webber et al., 2008), and bacterial canker of kiwi caused by pv. *actinidiae* (Balestra et al., 2010). Both of these pathovars have recently acquired genes involved in the catabolism of plant-derived aromatic compounds such as derivatives of lignin that are likely to be central in their ability to inhabit (and cause disease in) the woody parts of their respective host species (Green et al., 2010; Steele et al., 2010; Marcelletti et al., 2011). *Pseudomonas syringae* pv. *aesculi* has recently become epidemic across much of north-west Europe, and the success of this pathogen appears to be due, at least in part, to the ability of this species to acquire genes from disparate sources that has facilitated its adaptation into a distinct ecological niche.

Chapter 3

Synonymous codon usage in *Pseudomonas syringae*

3.1 Abstract

The genetic code is degenerate, allowing for the unequal usage of synonymous codons that encode the same amino acid. Codon usage bias may be driven by mutation or selection, and varies widely both among species and among genes within species. Here, we analyse variation in codon usage for *Pseudomonas syringae*, and find that overall codon usage is driven by mutational biases favouring GC-rich codons over AT-rich synonyms. Nonetheless, we also observe substantial heterogeneity in codon usage among genes encoded on the main chromosome of this strain. Using a multivariate statistical approach, we show that the major source of variation among genes is due to the presence of a relatively small number of non-core (flexible) genes that show unusually low G + C content relative to the genome-wide average, and are most likely the product of recent horizontal gene transfer (HGT). This finding is consistent with that previously observed for the closely related species *Pseudomonas aeruginosa*, where variation in codon usage among genes was hypothesised to be driven by HGT. In addition, we find two other sources of variation: genes that are likely to be highly expressed show different codon usage preferences presumably due of selection for translationally optimal codons in these

genes, and genes that are located on the leading strand show different codon usage preferences to those located on the lagging strand due to strand-specific mutational biases. Overall, our findings highlight the potential impact of HGT on patterns of codon usage in bacterial species.

3.2 Introduction

With four nucleotide bases and a tri-nucleotide genetic code, there are 64 possible codons from which DNA may be translated into polypeptide sequences. However, there exist only 20 canonical amino acids from which the majority of proteins are composed, and thus there is redundancy in the genetic code such that multiple distinct codons may give rise to the same amino acid. In addition, there is an abundance of evidence to show that synonymous codons for the same amino acid are not used equally, and that both the strength and direction of this bias varies considerably not only among species, but also among genes within the same genome (e.g., Grantham et al., 1980; Ikemura, 1985; Sharp et al., 1988, 1993, 2005).

The pattern of synonymous codon usage shown by a given gene is a product of both mutational and selective forces acting on it, as well as random genetic drift (Bulmer, 1991; Sharp and Li, 1986; Shields and Sharp, 1987). However, the relative importance of these evolutionary forces differs substantially when considering codon usage either among or within species. Among species, mutation is likely to be the major factor: heterogeneity in mutational biases can lead to drastic differences in the nucleotide composition of genomes (Sueoka, 1962), which in turn may influence the preferential usage of certain synonyms over others (Muto and Osawa, 1987; Chen et al., 2004). Furthermore, codon usage differences among bacteria are also driven by natural selection. Codon selection acts on the efficiency (and possibly the accuracy) of translation of highly expressed genes (Gouy and Gautier, 1982), and is a reflection of selective pressures on the rate of bacterial growth (Kurland, 1987; Dong et al., 1996). However, different species vary widely both in terms of which synonyms are selected for and in the overall strength of selection, presumably

due to differences in the ecology, life history and genome biology of the particular species in question (Sharp et al., 2005, 2010).

In addition, the analysis of whole genomes has revealed extensive heterogeneity in codon usage among genes within genomes. The majority of this variation is derived from three primary sources, the relative importance of which may differ across species. Firstly, there is variation in the usage of a subset of codons that are inferred to be translationally optimal, and are therefore under selection for this trait (as discussed above) (Gouy and Gautier, 1982). Secondly, there are extensive differences in G + C content at synonymously variable codon positions among genes, often due to the anomalous %GC of recently imported genes that have not yet acquired the codon usage patterns of the host genome (Lawrence and Ochman, 1998; Ochman et al., 2000; Sharp et al., 2005). Thirdly, the extent to which the species under question shows strand-specific nucleotide composition biases (i.e., GC- and/or TA-skew) may also influence the codon usage of genes, depending on their physical location on either the leading or lagging strand (with respect to chromosome replication) of the double stranded DNA molecule (Lobry, 1996; McLean et al., 1998; Lafay et al., 1999).

Patterns of synonymous codon usage have been previously elucidated for *Pseudomonas aeruginosa*, a close relative of *P. syringae* within the Gammaproteobacteria (Grocock and Sharp, 2002). This study suggested that codon usage differences among genes in this species is primarily influenced by G + C content variation among genes, and hypothesised that the horizontal transfer of genes with unusually low %GC compared to the average for this bacterium may be the underlying driver of this pattern. Here, we analyse patterns of codon usage in *Pseudomonas syringae*, for which the core genome has been previously determined (CHAPTER 2). The ability to designate genes as either core or non-core (i.e., 'flexible') allows for a more thorough investigation into the potential effects of HGT on codon usage: given that flexible genes are considered more likely to undergo HGT compared to core genes (Lawrence and Hendrickson, 2003; Ciccarelli et al., 2006; Sorek et al., 2007), we might expect a pattern to emerge whereby genes with anomalous %GC are also flexible.

Thus, we first investigate whether *P. syringae* and *P. aeruginosa* show similar trends in their codon usage, and then test the hypothesis of Grocock and Sharp (2002) by elucidating differences in codon usage patterns between core and flexible genes. We find that codon usage in *P. syringae* is largely mediated by its relatively high genomic G + C content (∼58%), such that the majority of preferred codons are enriched for G/C nucleotides. Using the core/flexible delineation as determined in CHAPTER 2, we also show clear differences in codon usage among core versus flexible genes, thus demonstrating the impact of HGT on patterns of codon usage in *P. syringae*.

3.3 Methods & Materials

3.3.1 Sequence data

A total of 4,985 coding sequences from the completed main chromosome (accession CP000058) of *P. syringae* pv. *phaseolicola* str. 1448A (Joardar et al., 2005) were downloaded from the GenBank database using the ACNUC program (Gouy et al., 1985). This strain was chosen for analysis as it possesses a completely sequenced (i.e., 'closed') genome sequence, a factor which is necessary for a comprehensive elucidation of codon usage that includes potential sources of variation such as strand bias. The effective number of codons (N_C, see below) could not be calculated for 14 sequences due to the absence of some amino acids; these genes were excluded from further analysis. Genes were classified as core or flexible based on the core genome as defined for 27 lineages of *P. syringae* in CHAPTER 2; thus, 2,578 genes for this strain were classified as core (52%), and 2,393 genes were classified as non-core.

3.3.2 Statistical analyses

The relative synonymous codon usage (RSCU) is a useful measure for comparing codon usage variation among genes that may differ in their amino acid composition, and is calculated by dividing the observed usage of each codon by that expected given the equal usage of all synonymous codons for a particular amino acid (Sharp and Li, 1986). Thus, an RSCU value of 1.0 for each of a set of synonyms would

indicate the absence of codon usage bias for that amino acid, whereas higher values indicate preferred usage. Absolute codon usages and RSCU values were calculated for all genes using the program CODONW v1.4.4 (Peden, 1999; available at: `http://sourceforge.net/projects/codonw/`).

Correspondence analysis (Greenacre, 1984) was then used to decompose the major axes of variation in codon usage among the 4,971 genes. Recently, the method of within-groups correspondence analysis (WCA) has been developed (Lobry et al., 2003) to account for the potential effects of amino acid composition bias among genes whilst maintaining the correct data structure for correspondence analyses (Perrière and Thioulouse, 2002; Suzuki et al., 2008). WCA plots the usage data for the 59 synonymous codons in a 41-dimensional hyperspace (according to the degrees of freedom left over from 59 synonymous codons across 18 amino acid groups) and elucidates the major axes that successively account for the most variation (Lobry et al., 2003). Causal factors that may be driving the major trends in codon usage may then be investigated by correlating the primary axes of variation from the WCA with a number of other genetic and genomic properties of these genes that are known to influence codon usage in other bacteria. WCA was performed using the ADE4 package (Charif et al., 2005) within the R statistical environment v3.1.0 (R Core Team, 2014).

Optimal codons were identified on the basis of statistical over-representation within a subset of 40 genes that are expected to be highly expressed in bacteria, as defined by Sharp et al. (2005). The frequency of codons within this highly expressed subset (HE) was compared to that within all genes using a χ^2 significance test, with sequential Bonferroni correction to account for the problem of multiple testing (Rice, 1989; Henry and Sharp, 2007). A number of other descriptors of codon usage bias for each gene were also calculated, primarily using CODONW, and are defined as follows.

$GC3_S$ Defined as the proportion of G + C nucleotides at the third position of synonymously variable codons (i.e., excluding Met and Trp, encoded by a single

codon each, and the three termination codons).

N_C The 'effective number of codons' is a general measure of the non-uniform use of synonymous codons for a given gene (Wright, 1990). Lower values of N_C indicate stronger levels of bias; a minimum value of 20 is possible under conditions of extreme bias such that only one codon is used for each amino acid, and a maximum of 61 is observed when all synonymous codons are used randomly. However, the influence of %GC content may cause this metric to be lower than 61 even in the absence of any codon usage bias due to selection; this is accounted for by comparing observed values of N_C to those expected under a range of %GC content values (Wright, 1990; Andersson and Sharp, 1996a).

F_{OP} The 'frequency of optimal codons' is a measure of the biased use of a subset of codons that are classified as being translationally optimal for the particular species in question (Ikemura, 1981; Andersson and Sharp, 1996b). For *P. syringae*, 19 optimal codons encoding 17 different amino acids were identified on the basis of their over-representation in a subset of 40 genes that are expected to be highly expressed across the Bacteria, including genes encoding the translation elongation proteins Tu (encoded by *tufA*), Ts (*tsf*) and G (*fusA*) and 37 ribosomal proteins (*rplA* to *rplF*, *rplI* to *rplT* and *rpsB* to *rpsT*) (Sharp et al., 2005). For each gene, F_{OP} is then calculated by taking the ratio of the occurrence of all optimal codons over the total occurrence of the 17 amino acids they encode (Andersson and Sharp, 1996b; Grocock and Sharp, 2002).

GC Skew Many bacterial genomes show an asymmetry in the nucleotide composition between the leading and the lagging strands of the DNA molecule, such that the leading strand is enriched for G/T while the lagging strand is enriched for C/A nucleotides (Lobry, 1996). Thus, GC-skew is a measure of the non-random distribution of nucleotides with respect to strand and can be calculated with: $(G3_S - C3_S)/(G3_S + C3_S)$, where $X3_S$ is the frequency of synonymously variable third codon position sites for the base 'X' (Lobry, 1996). This value is positive when

the number of G bases is greater than the number of C bases, as is the case for the GT-rich leading strand, but is negative when the number of C bases is greater than the number of G bases, as is the case on the CA-rich lagging strand. Thus, the genomic position of the origin and terminus of replication may be elucidated by taking the cumulative sum of $(G3_S - C3_S)/(G3_S + C3_S)$ across adjacent genes; the location of the origin occurs at the minimum of this sum, while the terminus occurs at the maximum (Grigoriev, 1998). The same logic may also be applied to TA-skew, using: $(T3_S - A3_S)/(T3_S + A3_S)$.

3.4 Results

3.4.1 Variation in codon usage in *Pseudomonas syringae*

Overall, the codon usage summed across *P. syringae* genes shows a bias towards the preferential usage of GC-rich codons (TABLE 3.1), as might be expected for this GC-rich bacterium. On average, the G + C content across all codon positions was 0.59, while the G+C content at the third position of synonymous codons ($GC3_S$) was higher still, at 0.68. Of the 18 amino acids for which multiple synonyms exist, the codon with the highest RSCU value was the most (or equal) GC-rich of the available synonyms for all but one (Glu). Nonetheless, variation in usage exists even between G- or C-ending synonyms for the same amino acid. For example, the codon CUG, encoding Leu, occurs approximately five times more frequently across all genes relative to the synonym CUC (TABLE 3.1). Thus, it is likely that the preferential usage of synonymous codons for some amino acids is due to factors over and above the influence of G + C content.

A preliminary investigation of codon usage heterogeneity may be made by plotting the effective number of codons (N_C) versus the $GC3_S$ content of each gene (FIG. 3.1). This shows that the majority of genes (about 87%) cluster within a $GC3_S$ range of 0.6 to 0.8 and an N_C range of 35 to 50. Generally, N_C values are below the expected curve for codon usage driven solely by %GC content, indicating an additional source of bias above that driven by mutational bias. In FIG. 3.1, blue-filled

Table 3.1: Relative synonymous codon usage (RSCU) in *P. syringae*.

		High[a]		All[b]				High		All	
		N	RSCU	*N*	RSCU			*N*	RSCU	*N*	RSCU
Phe	UUU	28	0.30	20,568	0.68	Ser	**UCU***	34	0.65	6,363	0.37
	UUC*	157	1.70	40,068	1.32		**UCC***	99	1.89	16,269	0.95
Leu	UUA	5	0.06	3,785	0.12		UCA	8	0.15	7,610	0.45
	UUG	52	0.63	28,758	0.91		UCG	72	1.37	23,625	1.38
	CUU	42	0.51	15,359	0.49	Pro	CCU	53	0.87	13,395	0.68
	CUC	47	0.57	23,768	0.75		CCC	11	0.18	16,348	0.83
	CUA	8	0.10	4,004	0.13		**CCA***	76	1.25	10,060	0.51
	CUG*	340	4.13	113,654	3.60		CCG	103	1.70	39,188	1.98
Ile	AUU	70	0.57	24,797	0.88	Thr	**ACU***	138	1.49	11,316	0.54
	AUC*	297	2.41	53,741	1.92		ACA	195	2.10	44,925	2.13
	AUA	3	0.02	5,601	0.20		ACC	14	0.15	8,561	0.41
Met	AUG	174	1.00	39,134	1.00		ACG	24	0.26	19,520	0.93
Val	**GUU***	260	1.61	17,484	0.59	Ala	**GCU***	313	1.75	24,581	0.56
	GUC	149	0.93	39,850	1.35		GCA	180	1.01	70,995	1.61
	GUA*	95	0.59	12,823	0.43		GCC	129	0.72	29,379	0.67
	GUG	140	0.87	48,247	1.63		GCG	93	0.52	51,169	1.16
Tyr	UAU	30	0.40	16,003	0.77	Cys	UGU	6	0.31	4,178	0.52
	UAC*	119	1.60	25,753	1.23		UGC	33	1.69	12,038	1.48
ter	UAA	29	2.18	1,298	0.78	ter	UGA	9	0.68	3,156	1.90
ter	UAG	2	0.15	531	0.32	Trp	UGG	31	1.00	23,418	1.00
His	CAU	31	0.47	16,108	0.86	Arg	**CGU***	363	3.88	27,523	1.58
	CAC*	102	1.53	21,140	1.14		CGC	184	1.97	49,175	2.81
Gln	**CAA***	83	0.71	20,063	0.54		CGA	2	0.02	7,053	0.40
	CAG	151	1.29	54,589	1.46		CGG	8	0.09	13,336	0.76
Asn	AAU	32	0.28	17,194	0.64	Ser	AGU	14	0.27	11,976	0.70
	AAC*	196	1.72	36,153	1.36		AGC	88	1.68	36,524	2.14
Lys	**AAA***	255	0.93	24,608	0.80	Arg	AGA	2	0.02	3,296	0.19
	AAG	292	1.07	37,102	1.20		AGG	2	0.02	4,440	0.25
Asp	GAU	103	0.64	36,234	0.80	Gly	**GGU***	290	1.96	31,860	1.00
	GAC*	218	1.36	54,403	1.20		GGC	290	1.96	71,632	2.24
Glu	**GAA***	277	1.26	50,862	1.08		GGA	5	0.03	7,712	0.24
	GAG	163	0.74	42,962	0.92		GGG	7	0.05	16,502	0.52

[a]RSCU values for a subset of 40 highly expressed (HE) genes; [b]RSCU values for all genes. Asterisks indicate those codons occurring at a significantly higher frequency relative to other synonyms for the same amino acid (in the HE subset).

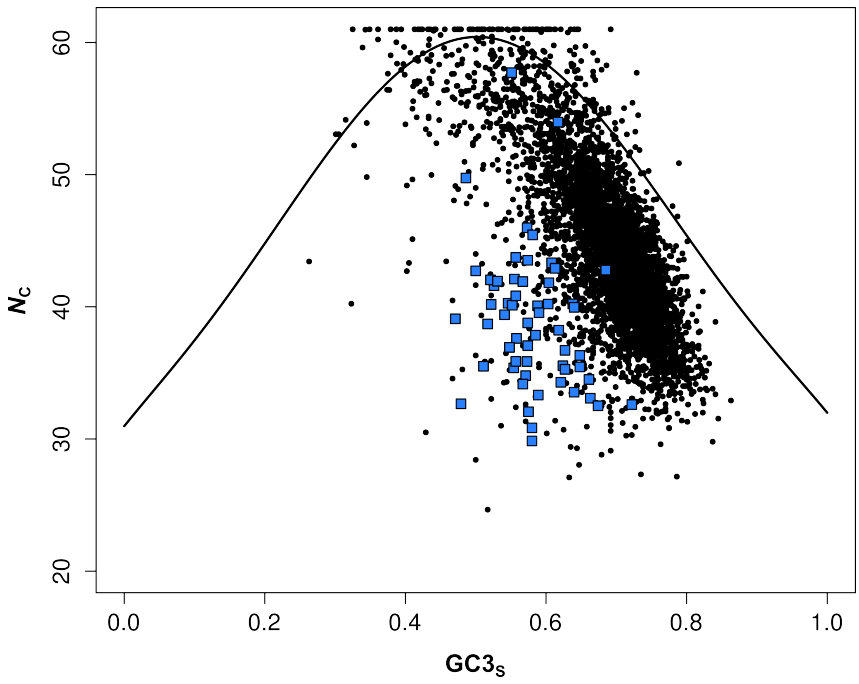


Figure 3.1: Codon usage heterogeneity across 4,971 *P. syringae* pv *phaseolicola* 1448A genes. The *y*-axis shows the effective number of codons (N_C), while the *x*-axis shows the proportion of G + C at the synonymously variable third position ($GC3_S$) for each gene. The solid line shows the expected N_C due to variation in %GC alone (Wright, 1990). A subset of 40 highly expressed genes are indicated with blue squares.

squares indicate a subset of genes known or expected to be highly expressed in all bacterial species (Sharp et al., 2005). These HE genes are generally found to the left of the main cluster, and therefore show a much lower expected N_C value for their given $GC3_S$. This suggests that genes within *P. syringae* may also show patterns of codon usage that are associated with translational selection.

Genes with anomalously low %GC compared to the rest of the genome may represent recent imports via horizontal gene transfer (HGT) (Lawrence and Ochman, 1998; Ochman et al., 2000). Genes with an N_C of 61 presumably lack any form of constraint on codon usage; in *P. syringae*, these genes predominantly encode proteins of undetermined function (e.g., 'hypothetical proteins') or are involved in transposition (such as transposases and prophage elements), and may therefore represent recent pseudogenes or annotation errors.

3.4.2 A primary trend associated with the flexible genome

Within-groups correspondence analysis of codon usage among *P. syringae* genes identifies a total of 41 trends that explain variation in codon usage among genes. Combined, the first four axes account for over a third of the total observed variation; individually, axes 1, 2, 3 and 4 explain 20.6, 7.2, 5.2 and 2.8% of the total variation, respectively.

The position of all genes on axes 1 and 2 is shown in FIG. 3.2, and shows the majority of genes to cluster near the origin. However, a spread of genes with lower co-ordinate values is evident, extending towards the left hand side of the plot. Axis 1 is highly correlated with $GC3_S$ (Pearson's correlation coefficient $r = 0.85$, $p \ll 0.0001$), such that genes further to the left of FIG. 3.2 have lower $GC3_S$ values (i.e., are more AT-rich). Among genes, the proportion of $GC3_S$ ranges from a minimum of 0.26 (encoding an a 'hypothetical protein') to a maximum of 0.86 (encoding an ABC transporter permease); however, the distribution of these values is skewed towards a higher G + C content (median = 0.70, interquartile range = 0.65–0.73). Thus, this trend is driven by a relatively small fraction of the total genes that are AT-rich.

In FIG. 3.2, core and flexible genes are indicated by colour, with core genes shown in red and flexible genes shown in black. This reveals a clear distinction in the relative positioning of core and flexible genes on axis 1: core genes cluster to the right, primarily within the main group, whereas the great majority of genes with lower values (i.e., a higher AT-richness) are flexible. Flexible genes are defined as those with variable occurrence across a set of 27 *P. syringae* lineages (Nowell et al., 2014), and are considered more likely to be subject to HGT than core genes (although no gene is 'immune' from transfer) (Lawrence and Hendrickson, 2003; Ciccarelli et al., 2006; Sorek et al., 2007). Thus, the major trend in codon usage in *P. syringae* is most likely driven by the presence of recently imported genes that show a reduced level of G + C content relative to the background level of this GC-rich bacterium.

Variation along axis 1 can also be visualised by focusing on the position of each of the 59 synonymous codons on axes 1 and 2 (FIG. 3.3). When coloured by the nucleotide at the third codon position, the trend associated with $GC3_S$ becomes

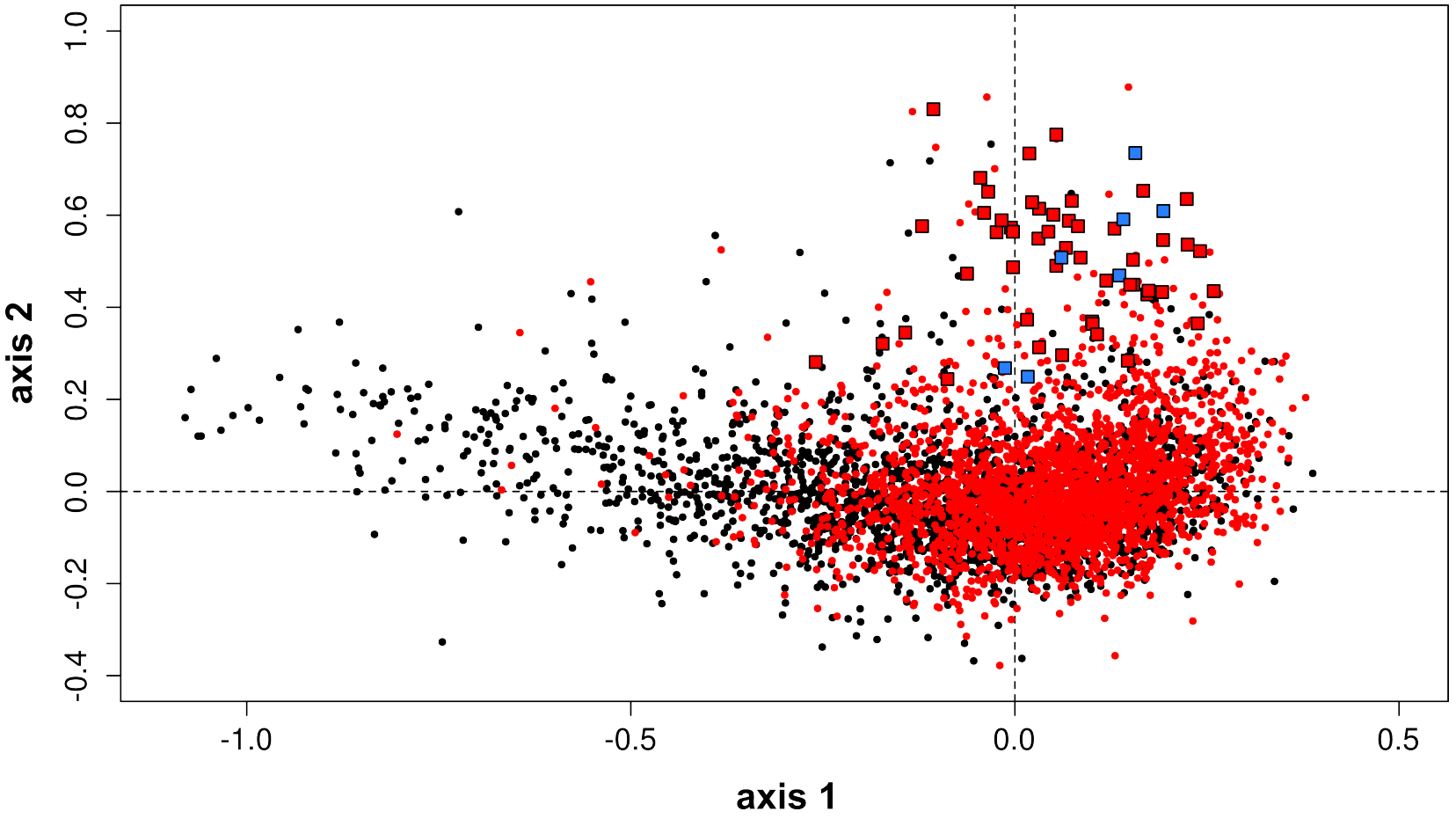


Figure 3.2: Axis 1 and 2 of the correspondence analysis for codon usage among 4,971 *P. syringae* genes. Genes of the core genome, as defined in Chapter 2, are shown in red; black indicates non-core (flexible) genes. Red-filled squares indicate highly expressed genes that are also members of the core genome, while highly expressed genes that are not defined as core are shown with blue squares as before.

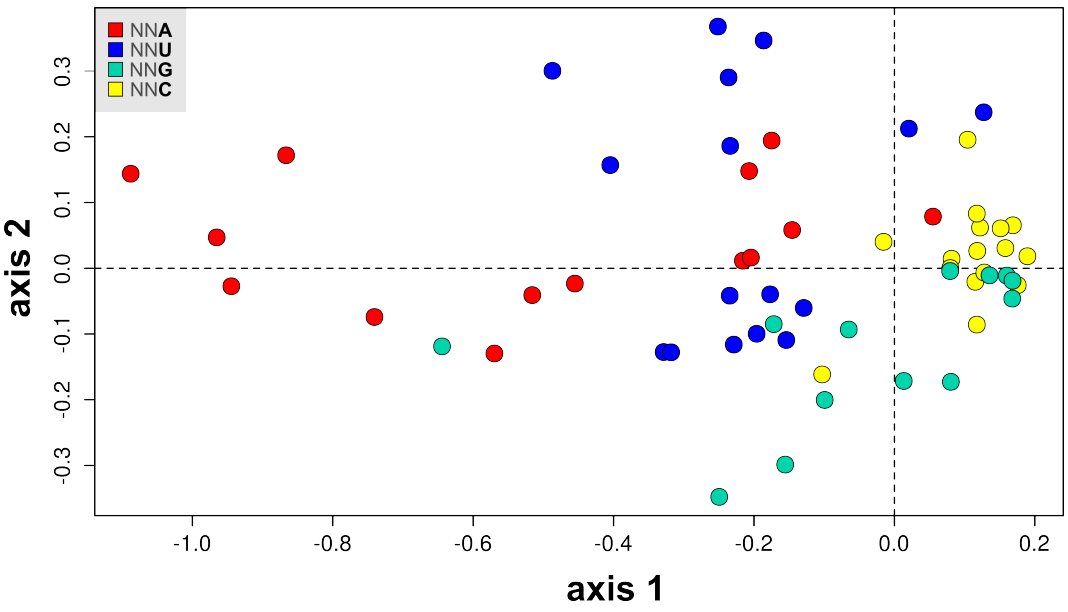


Figure 3.3: Distribution of codons along axes 1 and 2. Each point represents one of the 59 synonymously variable codons, coloured by the base at its third position: A, U, G and C-ending codons are coloured red, blue, teal and yellow, respectively.

evident: the majority of codons ending in either G or C (teal and yellow) are clustered tightly on the right hand side of axis 1, while codons ending in A or U (red and blue) occur predominantly on the left hand side, and exhibit a much greater degree of spread towards the lower extreme on axis 1.

We also note that seven genes within the HE subset, and a number of others with high values on axis 2, are designated as non-core genes on the basis of the core/flexible delineation made previously (Nowell et al., 2014, CHAPTER 2). These apparently flexible genes include essential proteins such as the translation elongation factors Tu and Ts, the ribosomal proteins L16 (*rplP*) and S16 (*rpsP*), and the β'-subunit of RNA polymerase (*rpoC*). Since bacterial proliferation cannot proceed without these genes, their apparent absence from certain lineages of *P. syringae* must represent assembly/annotation errors within the draft genomes from which these proteins are derived.

3.4.3 A secondary trend associated with highly expressed genes

Relative to the axis 1, the second axis accounts for a substantially lower amount, 7.2%, of total variation in codon usage among genes. Nonetheless, the positioning of the highly expressed genes towards the upper limit of axis 2 of FIG. 3.2 suggests that variation along this axis may be driven by selection on translationally optimal codons in highly expressed genes. This was further investigated by searching for significantly over-represented codons among the HE subset when compared to all genes, using χ^2-tests with a sequential Bonferroni correction. Based on this approach, we designate a total of 19 codons, encoding 17 amino acids, to be translationally optimal (shown in bold in TABLE 3.1).

For a given amino acid, one synonym may be translationally optimal over another if it is either (a) better recognised by the available tRNAs and thus utilised more frequently, or (b) corresponds to the more abundant tRNA species present in the cell. There is evidence that both these factors may be driving codon preferences in HE genes in *P. syringae*. For example, Phe is encoded by two codons, UUU and UUC, but only a single tRNA species (with the anticodon 5'-GAA-3') has been observed in bacteria (Sharp et al., 2005). Thus, although both UUU and UUC codons are recognised by this single tRNA species (via wobble), the exactly complementary UUC codon is always preferred in highly expressed genes, presumably due to its superior binding strength, and is thus the translationally optimal codon (Kanaya et al., 1999; Sharp et al., 2005). A similar situation is also found for the amino acids Ile (optimal codon AUC), Tyr (UAC) and Asn (AAC).

In addition, the usage of other codons appears to be driven by tRNA gene copy number variation within the pv. *phaseolicola* genome. The abundance of a specific tRNA within the cell is dictated largely by its gene copy number (Kanaya et al., 1999). Thus, for amino acids where copy number variation exists among tRNA species, the preferred codon tends to corresponds with the more abundant tRNA anticodon (Ikemura, 1985). This pattern is evident for a number of optimal codons in *P. syringae*. For example, there are four Pro tRNA genes encoded by the pv. *phaseolicola* genome sequence: three have the anticodon TGG and one has the anticodon GGG. The former

translates CCA well, but also CCG (through wobble), while the latter translates CCC well, but also CCU. Despite having no directly complementary tRNA available, the codon CCG has the highest RSCU value when all genes are considered, presumably as a result of the high %GC observed in *P. syringae*. The codon CCA is inferred to be translationally optimal among HE genes, however, and indeed is perfectly complementary to the most abundant tRNA species. Similarly, there are two copies of the Leu tRNA species with the anticodon CAG (compared to one each for the other five); correspondingly, CUG is inferred to be the optimal codon for this amino acid. The same pattern is also evident for Arg, where the most abundant tRNA (anticodon ACG) corresponds to the optimal codon among HE genes (CGU). It is interesting to note that the majority of optimal codons (11 of the 19) end in either U or A, a trend that is in the opposite direction to the overall mutational bias towards G/C in this species.

To gain an overview of this trend on axis 2, the frequency of these optimal codons (F_{OP}) was calculated across all genes by dividing the usage of each optimal codon by the total occurrence of the amino acid it encodes (see MATERIALS & METHODS). Values of F_{OP} ranged across genes from 0.18 to 0.76, with a median of 0.37. Plotting F_{OP} values versus gene coordinates for axis 2 showed these variables to be highly correlated ($r = 0.71$, $p \ll 0.0001$) (FIG. 3.4). Thus, the second largest cause of variation in codon usage among genes is the preferential usage of certain codons in highly expressed genes, most likely due to their translational efficiency and/or accuracy in these genes (Andersson and Kurland, 1990; Stoletzki and Eyre-Walker, 2007; Sharp et al., 2010).

3.4.4 A tertiary trend associated with GC skew

Many bacteria exhibit strand-specific composition biases, whereby the leading strand is enriched for G and/or T nucleotides while the lagging strand is enriched for C and/or A nucleotides. In circular bacterial genomes, the designation of the leading and lagging strands switches around the origin of replication (Ori) and again at the terminus of replication (Ter); thus, starting from *dnaA*, which is usually

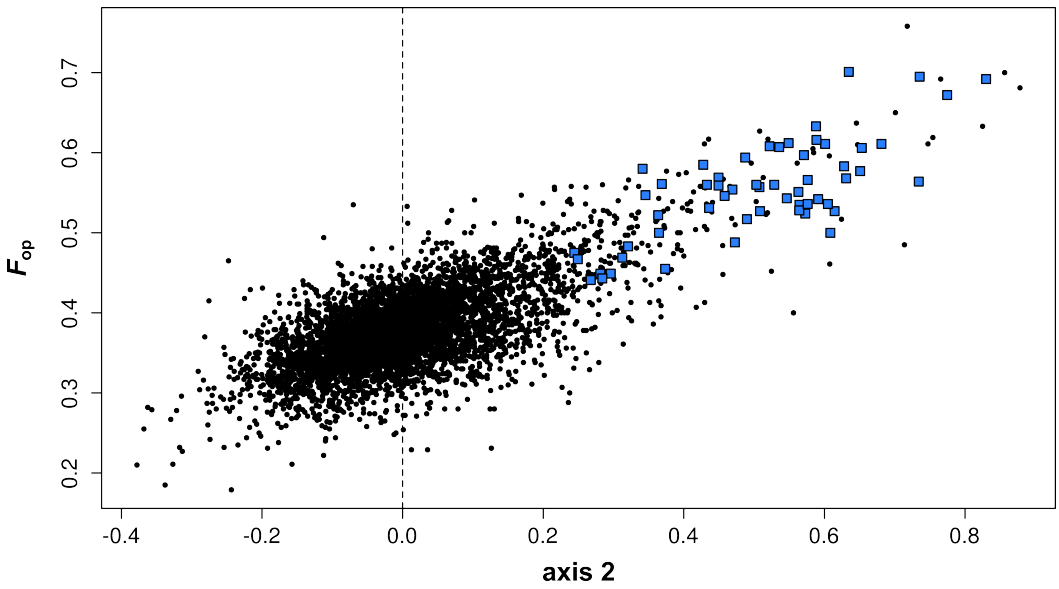


Figure 3.4: Relationship between axis 2 and the frequency of optimal codons (F_{OP}) within all genes (r = 0.71, $p \ll 0.0001$). Highly expressed genes are shown with blue squares as before.

located near to the Ori, the cumulative sum of GC- and TA-skew values across genes will increase in value as the terminus is approached and then decrease in value back towards the origin (Grigoriev, 1998).

We found variation along axis 3 to be highly correlated with the frequency at which G or T nucleotides occur at synonymously variable third codon positions (r = 0.83, $p \ll 0.0001$), suggestive of strand-specific base composition bias in this bacterium. This was confirmed by a cumulative skew diagram for *P. syringae* genes, given in FIG. 3.5. This shows the highest cumulative sum for both GC- and TA-skew to occur 2,637 genes downstream from *dnaA*, encompassing approximately 3.2 Mb (54%) of the 5.9 Mb main chromosome of this genome.

Based on this result, we classified the strand locations of all genes as leading or lagging for both the left and right replichores of the main chromosome, excluding 24 genes located around the inferred Ter where the designation was not clear, and plotted their axis 3 coordinate with respect to their position on the chromosome of this strain (FIG. 3.6). This reveals a clear pattern in codon usage based on strand: genes on the leading strand generally show higher coordinate values on axis 3, whereas genes on the lagging strand generally show lower (negative) coordinate

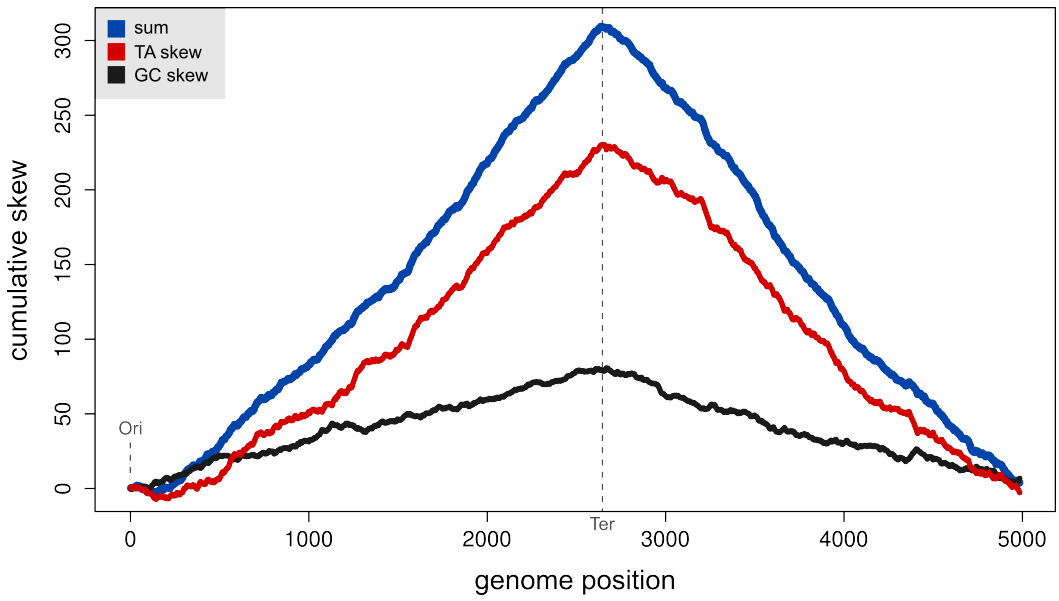


Figure 3.5: Cumulative skew diagram for pv. *phaseolicola*. GC-skew is shown in black; TA-skew in red, and their sum in blue. The first gene in the series is *dnaA*, which is usually located either at or near to the origin of replication in prokaryotes. The peak of the diagram indicates the location of the terminus, located approximately 2,637 genes downstream of *dnaA*.

values on axis 3. In addition, the direction of transcription for each gene is also indicated, where genes marked in red denote those transcribed in the forward direction for this genome sequence and blue denotes genes transcribed as the reverse complement. A clear switch in strand designation is evident around the terminus (Ter), whereby genes previously located on the leading strand in the right replichore suddenly are located on the lagging strand on the left replichore.

The position of the 59 synonymously variable codons on axes 3 and 4 is also shown in FIG. 3.7. When coloured by the nucleotide at the third codon position as before, a pattern emerges along axis 3 whereby codons ending in G or U cluster to the right of this axis, while codons ending in C or A are found to the left. Thus, while axis 3 only accounts for a relatively minor 5.2% of the total variation, we infer that the third source of variation in codon usage among *P. syringae* genes is due to the genomic position of each gene on either the leading or the lagging strand.

Finally, axis 4 explains only 2.8% of the total variation in codon usage among genes, and does not correlate with any other obvious measure of codon usage bias.

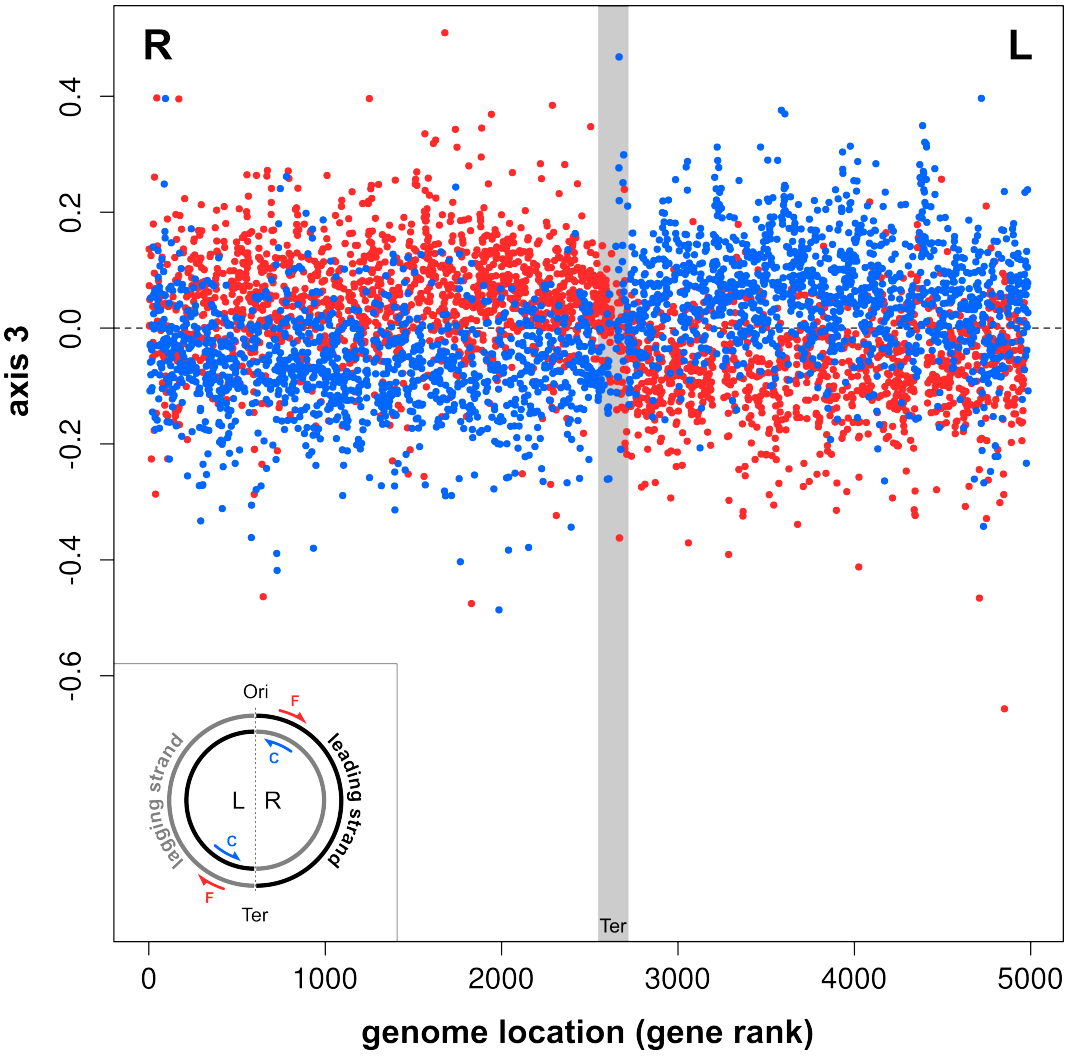


Figure 3.6: Strand bias in *P. syringae*. Here, the coordinate of each gene on the third axis of variation is plotted with respect to its genomic location (given as the ranked position of the gene in the genome where *dnaA* is at position 1). Genes marked in red are orientated in the forward direction in the genome sequence for this strain ('F', see inset), whereas genes marked blue are transcribed as the reverse complement ('C'). The inset shows the main chromosome in its circular form, and visualises the switching of the strand designation for F and C genes around the terminus of replication (Ter, the approximate position of which is indicated on the main plot by the shaded area) and again at the origin (Ori). For example, F-orientated genes are on the leading strand in the right replichore ('R') immediately upstream of Ter, but are on the lagging strand in the left replichore ('L') immediately downstream of Ter.

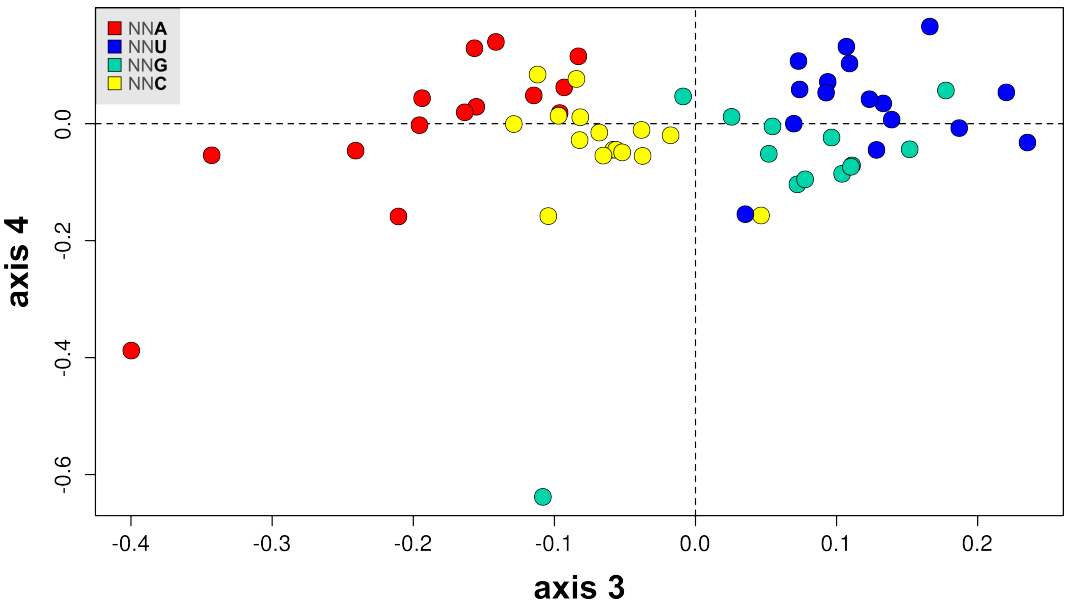


Figure 3.7: Distribution of codons along axes 3 and 4. Codons ending in A, U, G and C are respectively coloured red, blue, teal and yellow, as before.

This is consistent with codon usage studies on other bacteria, where typically only the first three (or fewer) axes have a meaningful biological interpretation (Suzuki et al., 2008). It is likely that variation along this axis is driven by the position of the two codons with particularly low values on axis 4 (FIG. 3.7). These correspond to the two Arg codons AGG and AGA, which are generally used much less frequently than other synonyms (CGN) for this amino acid.

3.5 Discussion

Here, we elucidate the major sources of variation in codon usage for *P. syringae* pv. *phaseolicola*. Given that the genome of this species is about 59% G + C, it is perhaps unsurprising that G + C content is the primary driver of codon usage bias, such that the majority of the most abundantly used synonyms are those that contain higher proportions of G and/or C nucleotides. It is also likely that the relative GC-richness of *P. syringae* is driven by mutational biases, given that the G + C content at less-constrained synonymously variable third codon positions ($GC3_S$) is about 10% higher than the genome-wide average. Thus, codon usage for *P. syringae*

is predominated by the mutational bias in favour of G/C nucleotides (i.e., the frequency of A/T to G/C mutations is greater than the reverse direction), and thus is similar to other bacterial species such as *Mycobacterium tuberculosis* (Andersson and Sharp, 1996b) and *Pseudomonas aeruginosa* (Grocock and Sharp, 2002).

Nonetheless, we detect substantial variation in codon usage patterns among genes within the pv. *phaseolicola* genome. The decomposition of this variation reveals three underlying explanatory mechanisms that have all been previously observed in other species. The major source of codon usage heterogeneity among genes was variation in the G + C content of individual genes. This factor accounted for approximately one fifth of the total observed variance, and was driven by the presence of a relatively small number of genes with unusually low %GC. When the distinction between core and flexible genes is superimposed onto these data, a clear pattern of association emerges whereby genes with low %GC are predominantly flexible. While an anomalous G + C content does not necessarily indicate a foreign origin for these genes (e.g., Koski et al., 2001), the fact that the majority of AT-rich genes are also flexible is strongly suggestive of their recent import into the pv. *phaseolicola* genome. If these genes were to persist over evolutionary timescales in this lineage, it would be expected that the G/C mutational bias experienced by the genome as a whole would lead to their amelioration towards the genome average %GC (Lawrence and Ochman, 1997, 1998). However, the analyses presented in CHAPTER 2 suggest that pv. *phaseolicola* (and all of *P. syringae*) experiences remarkably high rates of gene gain and loss. Continued HGT is likely to result in a constant influx of anomalous G+C genes into the *P. syringae* genome, and highlights the potential impact of HGT and genome fluctuation on codon usage patterns in this species.

We also detect two additional yet relatively minor sources of variation in codon usage preferences for pv. *phaseolicola*. Firstly, genes expressed at higher levels show differences in codon usage patterns relative to all genes. This likely represents selection for the usage of codons that are translationally optimal in these genes. Secondly, we also detect a strand-specific bias such that genes located on the leading strand

are enriched for G/T nucleotides, and therefore show a preference for synonymous codons which reflect this bias, while genes located on the lagging strand are enriched for A/C nucleotides, and therefore show the opposite pattern.

As such, the patterns of codon usage observed in pv. *phaseolicola* are similar to those shown by another GC-rich Pseudomonad, the opportunistic human pathogen *P. aeruginosa* PA01 (Grocock and Sharp, 2002). For this bacterium, Grocock and Sharp (2002) also inferred the major influence of codon usage to be G + C content driven by mutational biases, and also detected variation among *P. aeruginosa* genes caused by the anomolous G + C content of recently imported genes, differential codon usage among highly expressed genes and strand-specific base composition biases. Furthermore, these sources of variation occur in the same rank order of importance in *P. aeruginosa* as they do for *P. syringae*. Despite this, there exist notable differences in the direction of selection on translationally optimal codons, even between these two closely related species. Applying the same methodology here as was used for *P. aeruginosa* indicates a total of 15 optimal codons for this species; however, the identification of transationally optimal codons in this species is confounded by its extremely high G + C content ($GC3_S$ = 0.88). Nonetheless, it is clear that some differences do exist between the species. For example, the optimal codon for Pro in *P. syringae* is CCA, with an RSCU value among highly expressed genes of 1.25. In *P. aeruginosa*, however, this codon is very rarely used within the orthologous set of highly expressed genes, and has an RSCU of only 0.07. Thus, the overall patterns observed for *P. aeruginosa* and *P. syringae* suggests a commonality between the evolutionary forces that impact codon usage in these two species, while the differences in preferred codons among highly expressed genes highlights the rapidity at which different synonyms may be selected.

Our investigations also reveal a number of genes that are likely to be under selection for translational efficiency/accuracy (with values > 0.4 on axis 2 of FIG. 3.2), but are apparently missing from the core genome as defined previously. This includes genes whose products are essential for bacterial cellular functioning, such as the translation-elongation factors Tu and Ts, which facilitate the process of polypeptide

synthesis within the ribosomal complex (Stark et al., 1997), the ribosomal proteins L16, S16 and L34, and two cold shock proteins CapA and CapB, which are involved in countering the deleterious effects of a sudden downshift in temperature (Phadtare et al., 1999). These genes are all present in the complete genome sequence of pv. *phaseolicola*, but are classified as non-core on the basis of their absence from one of the other *P. syringae* lineages used in the delineation of the core and flexible genome components (Nowell et al., 2014). Using BLASTp (Altschul et al., 1997), we found these proteins to be truly missing from the proteome of the 'absent' strain. In all cases, however, tBLASTn analysis of the offending genome sequence suggested that these essential genes are in fact present, as would be expected. In many cases, the coding sequence was either truncated by the termination of a contig, or was frameshifted by the presence of small insertion-deletion mutations. Thus, we conclude that the apparent absence of these necessary genes from certain lineages is an artefact of assembly and/or annotation errors in the draft genome sequences from which these data are derived.

Chapter 4

Genome sequencing and assembly of 26 strains of *Pseudomonas syringae* isolated from woody hosts

4.1 Abstract

Compared to bacterial pathogens of herbaceous plants, there are relatively few available genome sequences of bacteria that infect woody host species. Here, we report the whole-genome sequencing, assembly and annotation of 26 strains of *Pseudomonas syringae* isolated from a range of woody hosts, including a number of large tree species. These data provide the basis for comparative analyses with existing genomes of *P. syringae*, from which insights into the evolutionary adaptation of this important pathogen onto woody hosts are gained.

4.2 Introduction

Pseudomonas syringae is an extensively studied facultative phytopathogen of a range of plant species, including commercially valuable crops such as tomato, bean, wheat

and tobacco (Hirano and Upper, 1990; Sarkar and Guttman, 2004; O'Brien et al., 2011b). The *P. syringae* species complex has been divided into over 50 pathovars (*patho*logical *var*iants) that are adapted to particular host species and are generally unable to cause disease when transferred to an alternative host (Hunter and Taylor, 2006; Sarkar et al., 2006).

The diversification of lineages within *P. syringae* has involved a number of shifts from herbaceous hosts onto various species of woody plant, resulting in the recent emergence of diseases such as bacterial canker of kiwi, hazelnut decline and bleeding canker of the European horse chestnut (Scortichini et al., 2002; Webber et al., 2008; Balestra et al., 2010). Currently, the genome sequence resource available for the study of these adaptive shifts is limited to a handful of genomes (e.g., Green et al., 2010; Marcelletti et al., 2011; O'Brien et al., 2012), and is far outweighed by the number of available genomes of pathovars isolated from herbaceous hosts.

Here, we report the draft genome sequences of 26 strains of 16 different pathovars from across the *P. syringae* species complex, which were isolated from a diverse range of woody host species. We characterise the complement and occurrence of mobile genetic elements among our assemblies, and identify the extent to which these strains may harbour putative plasmids. These data provide a platform from which to study the genomic basis of host adaptation in these important plant pathogenic bacteria.

4.3 Materials & Methods

4.3.1 Bacterial strains

Strains for sequencing were chosen to maximise the sampling of diversity both in terms of the strains themselves and their host species. Isolates were originally collected between 1958 to 2011 from a range of geographic locations across Europe, Japan and North America. Details of the year and location of isolation for all strains are given in TABLE 4.1. Thirteen strains were obtained from the culture collection held at the Collection Française de Bactéries Phytopathogènes (CFBP), Beaucouzé,

Table 4.1: Strain information.

Pathovar	Strain	Identifier[a]	Source[b]	Location	Year
aesculi	2113	aesc2113	NRS	UK	2006
aesculi	2279	aesc2279	NRS	Netherlands	2002
aesculi	2306	aesc2306	NRS	Belgium	2010
aesculi	2315	aesc2315	NRS	Netherlands	2006
aesculi	2329	aesc2329	NRS	Czech Republic	2011
aesculi	2336	aesc2336	NRS	Norway	2010
avii	CFBP 3648	avii3648	CFBP	France	1991
broussonetiae	CFBP 5140	brou5140	CFBP	Japan	1980
castaneae	CFBP 4217	cast4217	CFBP	Japan	1977
cerasicola	CFBP 6109	cera6109	CFBP	Japan	1995
daphniphylli	CFBP 4219	daph4219	CFBP	Japan	1981
dendropanacis	CFBP 3226	dend3226	CFBP	Japan	1979
eriobotryae	CFBP 2343	erio2343	CFBP	USA	1970
fraxini	CFBP 5062	frax5062	CFBP	Netherlands	1978
morsprunorum	2341	mors2341	J. Nemeth	Hungary	1988
morsprunorum	HRI-W 5261	mors5261	HRI-W	UK	1990
morsprunorum	HRI-W 5269	mors5269	HRI-W	UK	1990
myricae	CFBP 2897	myri2897	CFBP	Japan	1978
nerii	CFBP 5067	neri5067	CFBP	Spain	1979
papulans	CFBP 1754	papu1754	CFBP	Canada	1973
rhaphiolepidis	CFBP 4220	rhap4220	CFBP	Japan	1980
syringae	2339	syri2339	J. Nemeth	Hungary	1984
syringae	2340	syri2340	J. Nemeth	Hungary	1985
syringae	HRI-W 7872	syri7872	HRI-W	UK	2000
syringae	HRI-W 7924	syri7924	HRI-W	UK	2000
ulmi	CFBP 1407	ulmi1407	CFBP	Yugoslavia	1958

[a]Unique identifier allocated to each genome; [b]NRS, Forest Research, Northern Research Station, Midlothian, UK; CFBP, Collection Française de Bactéries Phytopathogènes, Beaucouzé, France; HRI-W, Horticultural Research International, Warwick, UK.

France, nine from the collection held at Northern Research Station (NRS), Midlothian, UK, and four from the Horticultural Research Institute, Warwick, UK (HRI-W). The host of isolation, associated disease symptoms and key references for each strain are given in TABLE 4.2. All strains infect the woody parts of their respective host species, and cause a range of diseases with symptoms including cankers, galls, knots and tissue necrosis.

It should be noted that the host range and pathogenicity of individual strains where not explicitly tested, but were inferred from individual pathovar designations and from the literature (see references in TABLE 4.2).

Table 4.2: Host of isolation, disease symptoms and references.

Identifier	Host	Disease/symptoms	Reference[a]
aesc2113	*Aesculus hippocastanum* (Horse chestnut)	Bleeding canker	This study
aesc2279	*Aesculus hippocastanum* (Horse chestnut)	Bleeding canker	This study
aesc2306	*Aesculus hippocastanum* (Horse chestnut)	Bleeding canker	This study
aesc2315	*Aesculus hippocastanum* (Horse chestnut)	Bleeding canker	This study
aesc2329	*Aesculus hippocastanum* (Horse chestnut)	Bleeding canker	This study
aesc2336	*Aesculus hippocastanum* (Horse chestnut)	Bleeding canker	This study
avii3648	*Prunus avium* (Sweet cherry)	Canker	Ménard et al. (2003)
brou5140	*Broussonetia kazinoki* (Paper mulberry)	Necrotic lesions on shoots	Takahashi et al. (1996)
cast4217	*Castanea crenata* (Japanese chestnut)	Canker	Höfte and Vos (2006)
cera6109	*Prunus yedoensis* (Yoshino cherry)	Galls	Kamiunten et al. (2000)
daph4219	*Daphniphyllum teigsmanni* (—)	Galls	Ogimi et al. (1990); Höfte and Vos (2006)
dend3226	*Dendropanax trifidus* (Ivy tree)	Galls	Ogimi et al. (1988); Höfte and Vos (2006)
erio2343	*Eriobotrya japonica* (Loquat)	Bud blight, twig canker	Höfte and Vos (2006)
frax5062	*Fraxinus excelsior* (Ash)	Canker	Höfte and Vos (2006); Ramos et al. (2012)
mors2341	*Prunus cerasus* (Sour cherry)	Twig necrosis	—
mors5261	*Prunus avium* (Sweet cherry cv. Roundel)	Canker	Vicente et al. (2004)
mors5269	*Prunus cerasus* (Sour cherry)	Canker	Vicente et al. (2004)
myri2897	*Myrica rubra* (Chinese bayberry)	Galls	Ogimi and Higuchi (1981); Höfte and Vos (2006)
neri5067	*Nerium oleander* (Oleander)	Galls, knots	Höfte and Vos (2006); Ramos et al. (2012)
papu1754	*Malus sylvestris* (Crab apple)	Blister canker	Höfte and Vos (2006)
rhap4220	*Rhaphiolepis umbellata* (—)	Galls	Ogimi et al. (1992); Höfte and Vos (2006)
syri2339	*Prunus avium* (Sweet cherry)	Shoot necrosis	—
syri2340	*Pyrus* sp. (Pear)	Shoot necrosis	—
syri7872	*Prunus domestica* (Plum cv. Opal)	Canker	Vicente et al. (2004)
syri7924	*Prunus cerasus* (Sour cherry)	Canker	Vicente et al. (2004)
ulmi1407	*Ulmus* sp. (Elm)	Leaf and shoot blight	Höfte and Vos (2006)

[a]Original reference detailing the host from which each strain was isolated and its associated disease symptoms.

4.3.2 Host information

Further information regarding the host plants for these bacterial pathogens is given in TABLE 4.3. All strains were isolated from plant species that are members of the 'core' eudicots (Bremer et al., 2009). Within this extremely broad and speciose group, 11 host species are from within the rosids, while three are placed in a sister group, the asterids. The rosid group contains some 7,000 species, divided into two primary groups called the fabids (including members of the Rosales and Fagales) and the malvids (including the Sapindales), while the Saxifragales order is unplaced within the rosids (Bremer et al., 2009; Burleigh et al., 2009; Wang et al., 2009).

Half of the strains presented here were isolated from members of the Rosales order, with the majority (11 strains) having hosts within the Rosaceae (the rose family), and one each from the Moraceae (mulberry family) and the Ulmaceae (elm family). These include all of the strains pathogenic on various stone fruit species such as cherry (*avii*, *morsprunorum* and *syringae* strains 2339 and 7924), plum (*syri7872*), pear (*syri2340*) and apple (*papulans*). Within the Rosaceae, the genera *Malus*, *Eriobotrya*, *Rhaphiolepidis* and *Pyrus* are closely related, within the same sub-tribe, while *Prunus* is more distantly related (Potter et al., 2007).

Two pathovars, *myricae* and *castanae*, were isolated from species within the Fa-

Table 4.3: Host information.

Host	Pathogens	Order, Family
A. hippocastanum	*aesculi* (×6)	Sapindales, Sapindaceae
Prunus spp.	*avii*, *morsprunorum* (×3), *syringae* (×3)	Rosales, Rosaceae
B. kazinoki	*broussonetiae*	Rosales, Moraceae
C. crenata	*castanae*	Fagales, Fagaceae
D. teigsmanni	*daphniphylli*	Saxifragales, Daphniphyllaceae
D. trifidus	*dendropanacis*	Apiales, Araliaceae
E. japonica	*eriobotryae*	Rosales, Rosaceae
F. excelsior	*fraxini*	Lamiales, Oleaceae
M. rubra	*myricae*	Fagales, Myricaceae
N. oleander	*nerii*	Gentianales, Apocynaceae
M. sylvestris	*papulans*	Rosales, Rosaceae
R. umbellata	*rhaphiolepidis*	Rosales, Rosaceae
Pyrus sp.	*syringae*	Rosales, Rosaceae
Ulmus sp.	*ulmi*	Rosales, Ulmaceae

gales, which, together with the Cucurbitales, forms a sister group to the Rosales. The horse chestnut is from the the Sapindales order within the fabids, and is thus only distantly related to either the Rosales or the Fagales, and we note that the *Castanae* and *Aesculus* genera, containing the Japanese chestnut and the horse chestnut respectively, are therefore not close relatives.

More distantly related yet, within the asterids, are the Lamiales (e.g., ash), the Gentianales (oleander), and the Apiales (ivy tree), from which pv.'s *fraxini*, *nerii* and *dendropanacis* were isolated.

4.3.3 Growth conditions and DNA extraction

Freeze-dried samples were revived by streaking onto King's B agar and incubated for ∼24 hours at room temperature. For each strain, a single colony was selected and grown overnight in 3 ml King's B broth for ∼12 hours, shaking at room temperature. Laboratory passage of strains was minimised as much as possible to mitigate against the loss of non-essential genes, although it should be noted that the total length of passage since the original isolation is not known.

Prior to whole-genome sequencing, strains were identified as *P. syringae* by PCR amplification and Sanger sequencing of a ∼600 bp fragment of the *rpoD* gene (F primer: 5'-TGAAGGCGARATCGAAAT-3'; R primer: 5'-YGCMGWCAGCTTYTGCT-3'). PCR products were aligned with other *rpoD* sequences available in GenBank and identified as *P. syringae* based on sequence similarity.

For each isolate, cells were harvested by centrifuging 1.5 ml of overnight culture at 5000 rpm for 5 minutes, discarding the supernatant and storing at −80°C. Genomic DNA was extracted using a DNeasy Plant Mini Kit (Qiagen), briefly as follows. Each pellet was resuspended in 400 µl Lysis Buffer AP1 and 4 µl RNase. After mixing by inverting the tube and incubation at 65°C for 10 minutes, 130 µl Buffer P3 was added to precipitate out unwanted cellular debris such as proteins, detergents and polysaccharides. We found that mixing by pipetting up and down followed by 2 or 3 sharp inversions of the tube resulted in a more successful separation of genomic DNA from the precipitates at this stage, giving a significantly increased final DNA

yield. Tubes were held on ice for at least 10 minutes then centrifuged at 14,000 rpm for 5 minutes. The lysate was transferred into a supplied QIAshredder Mini spin column and centrifuged at 14,000 rpm for 2 minutes. The flow-through was transferred into a new 1.5 ml centrifuge tube and 1.5 volumes of Buffer AW1 added and mixed by pipetting. About 650 µl was then transferred to a supplied DNeasy Mini spin column, centrifuged for 1 minute at 8,000 rpm and the flow-through discarded. 500µl Buffer AW2 was added and centrifuged at 8,000 rpm for 1 minute. Another 500 µl Buffer AW2 was added followed by centrifugation at 14,000 rpm for 2 minutes to wash and dry the column membrane. Finally, the column was transferred to a new 1.5 ml tube and 60 µl of Buffer AE (warmed to 65°C) pipetted directly onto the membrane, incubated for at least 5 minutes at room temperature and centrifuged at 8000 rpm for 1 minute to elute. A second elution was performed with a volume of 45 µl AE Buffer.

Final DNA concentrations were ascertained using a PicoGreen assay (Invitrogen), and ranged from 54–155 ng/µl in a final volume of 60 µl.

4.3.4 Whole-genome sequencing, assembly and annotation

Library preparation and whole-genome sequencing were performed by ARK Genomics (The Roslin Institute, Midlothian, UK). Briefly, paired-end libraries were prepared using the Illumina Nextera kit with an estimated average insert length of ∼270 bp. Libraries were multiplexed and run on a single lane of an Illumina MiSeq benchtop sequencer.

Prior to assembly, a quality-control pipeline was constructed to remove as many low-quality bases from the raw reads as possible. The removal of low-quality bases, which are likely to include sequencing errors, can simplify the de Brujin graphs used in many *de novo* assembly algorithms and can substantially improve the resultant assembly (Paszkiewicz and Studholme, 2010; Smeds and Künstner, 2011; Magoc et al., 2013). In addition, the short average insert length generated by the Nextera kit, in combination with the longer read-length now possible on the Illumina MiSeq platform, resulted in a large amount of contamination from artefactual sequences

such as adapter and barcode sequences due to the read length being longer than the fragment size in many cases.

Reads containing adapter contamination were identified and trimmed using a combination of CUTADAPT v1.2.1 (Martin, 2011) and TAGDUST v1.12 (Lassmann et al., 2009). Low quality base-pairs (quality-score threshold ≤ 25) were trimmed from the 3′ ends of reads using CONDETRI v2.2 (Smeds and Künstner, 2011). Reads with a high proportion ($> 80\%$) of low-quality bases throughout the read length were discarded, as were any reads containing base ambiguities ('Ns'). Quality reports were generated and manually inspected prior to assembly using FASTQC v0.10.1 (Andrews, 2010).

Genome assembly was initially trialled using a range of different *de novo* assemblers, including ABYSS v1.3.4 (Simpson 2009), IDBA v1.1.0 (Peng et al., 2012), RAY v2.1.0 (Boisvert et al., 2010), SPADES v2.4.0 (Bankevich et al., 2012) and VELVET v1.2.08 (Zerbino and Birney, 2008). We evaluated the performance of these programs by assembling the genomes of three strains, *syri2339*, *mors5261* and *aesc2113* from phylogroups 1, 2 and 3 respectively, and recording assembler performance and assembly information. For each of the three genomes, the programs were run with a number of different parameter combinations using the same set of cleaned input reads, and the 'best' assembly was chosen based on evaluation of N_{50} (the size of the contig at which 50% of the genome is covered by a contig of the same length or longer), total genome length and total number of contigs. Assembly statistics were calculated using the original 'Assemblathon' PERL script (Earl et al., 2011).

Based on the results from above, and in conjunction with another recently published evaluation of bacterial genome assemblers (Magoc et al., 2013), data for the remaining strains was assembled using a modified version of SPADES that allowed for an increased final *k*-mer value of 229. All genomes were annotated with the Rapid Annotation using Subsystem Technology (RAST) online server (Aziz et al., 2008; Overbeek et al., 2014).

4.3.5 Phylogenetic analyses and whole genome comparisons

The phylogenetic placement of these 26 strains within the *P. syringae* species complex was inferred using a multilocus concatenation approach. Following the MLST schema of Hwang et al. (2005), partial fragments for four housekeeping genes (*gap1*, *gltA*, *gyrB* and *rpoD*) were identified in each genome using BLAST (Altschul et al., 1997) and aligned using the Translational Align function within the program GENEIOUS v5.6.5 (Biomatters). These data were combined with the orthologous sequences from 100 additional *P. syringae* strains from across the *P. syringae* species complex, obtained from the online MLST database PAMDB (available at `www.pamdb.org`), resulting in a concatenated alignment 1,965 bp in length. Two *Pseudomonas cannabina* strains and *Pseudomonas fluorescens* Pf-5 were included as outgroups. A maximum likelihood phylogeny was then constructed with PHYML v3.0 (Guindon and Gascuel, 2003; Guindon et al., 2010), using a general time-reversible (GTR) model of nucleotide substitution, four substitution rate categories with an estimated gamma distribution (GTR + Γ), nearest-neighbor-interchange (NNI) topology searching and with 100 bootstrap replicates to assess topological support.

Variation in genome content among strains was visualised using the BLAST Ring Image Generator (BRIG) v0.95, which uses colour thresholds to visually summarise levels of BLAST similarity between sets of sequences (Alikhan et al., 2011). BRIG was run with similarity thresholds at 50% and 70%, a BLAST *E*-value threshold of 1e$-$5 and soft-masking turned off. Contig data for our 26 strains were compared to the three complete sequences of *P. syringae* pv. *tomato* DC3000 from phylogroup 1, pv. *syringae* B728a from phylogroup 2 and pv. *phaseolicola* 1448A from phylogroup 3.

4.3.6 Characterisation of mobile genetic elements

The extent to which our genomes are composed of mobile genetic elements (i.e., insertion sequences and plasmids) was characterised using a BLAST-based approach. Plasmid sequence was delimited from the chromosome if a given contig showed a minimum of 80% identity over at least 80% of its length (hereafter referred to as the 80/80 rule) to a database containing over 4,500 known plasmid sequences

(downloaded from GenBank on June 24, 2014) using BLASTn (*E*-value $\leq$ 1e$-$5). Since contigs consisting entirely of repeat elements (such as transposons) may be derived from both plasmid and chromosomal sequences, any contig with homology (at the 80/80 rule threshold) to the three 'finished' (i.e., completely sequenced with no gaps) chromosomes of pv.'s *phaseolicola* str. 1448A, *syringae* str. B728a or *tomato* str. DC3000 were not included in the final estimate of plasmid length.

To characterise the repertoire of insertion sequence (IS) elements present among our assemblies, 59 protein products encoded by 34 different IS elements were downloaded from `www-is.biotoul.fr` (Siguier et al., 2006b) and queried against the contig data for each strain using tBLASTn (*E*-value $\leq$ 1e$-$5), defining presence using the 80/80 rule as above. For IS elements encoding multiple proteins, presence was recorded for that strain if any one protein query passed the 80/80 rule. Patterns of presence and absence for different IS elements among strains were plotted with respect to their phylogeny using the package 'ape' v3.1-2 (Paradis et al., 2004) in R v3.1.0 (R Core Team, 2014).

4.4 Results

4.4.1 *P. syringae* assemblathon

The results from trialling a variety of different *de novo* assemblers on three representative strains of *P. syringae* are shown in TABLE 4.4. Assessing assembly quality is a non-trivial problem that has generated a lot of debate in the recent literature (e.g., Earl et al., 2011; Baker, 2012; Salzberg et al., 2012; Magoc et al., 2013). *De novo* assemblies are particularly problematic in that the content and composition of the 'true' genome, by definition, is not known. A commonly used metric for heuristically assessing genome assemblies is the assembly N_{50}, defined as the length of the contig at which 50% of the assembled genome is covered by a contig of equivalent size or larger (Earl et al., 2011; Baker, 2012). A large N_{50} is an indication that the assembler has been able to successfully combine smaller contigs into larger ones, thereby increasing the potential contiguity of the assembled genome (Baker, 2012).

Table 4.4: *P. syringae* assemblathon.

Assembler:	**ABySS**	**IDBA**	**Ray**[a]	**SPAdes**	**Velvet**
***syri2339* (PG1)**					
K-mer value	96	31 to 231	61	21 to 229	131
Coverage (X)[b]	**98**	45	80	**98**	84
Contigs	333	138	13,439	**69**	760
Contigs >500bp	151	75	98	**59**	245
N_{50}[c]	72,145	169,729	108,110	**173,289**	47,111
L_{50}[d]	27	14	19	**12**	44
Max contig	255,899	365,925	261,557	**434,602**	177,857
Genome size	6,115,507	6,145,718	5,975,805	6,122,834	**6,249,407**
%GC	59.3	59.2	59.2	59.2	59.2
CDS[e]	5,247	5,284	5,116	5,246	**5,363**
***mors5261* (PG2)**					
K-mer value	96	31 to 231	61	21 to 229	131
Coverage (X)	72	33	61	**73**	63
Contigs	957	513	10,877	**264**	822
Contigs >500bp	395	281	296	**221**	685
N_{50}	35,720	76,141	46,133	**83,133**	17,984
L_{50}	53	**27**	40	**27**	98
Max contig	162,358	**338,421**	182,644	258,299	95,242
Genome size	6,513,326	**6,558,799**	6,507,222	6,471,635	6,492,597
%GC	58.4	58.4	58.3	58.4	58.4
CDS	5,787	**6,051**	6,033	5,887	6,003
***aesc2113* (PG3)**					
K-mer value	96	31 to 231	61	21 to 229	151
Coverage (X)	92	43	74	93	**103**
Contigs	1,028	877	14,111	**330**	618
Contigs >500bp	426	365	414	**264**	521
N_{50}	38,647	**60,975**	50,437	60,900	19,979
L_{50}	51	**33**	38	34	76
Max contig	137,050	**195,810**	153,000	172,324	104,375
Genome size	6,253,515	**6,301,277**	6,219,588	6,165,885	5,074,855
%GC	58.2	58.1	58.2	58.3	58.0
CDS	5,690	5,818	**5,879**	5,644	4,715

Maximum / minimum values shown in bold; [a]Ray produces a large number of very small contigs; thus, N_{50} for this assembler is based on contigs >500 bp; [b]Coverage calculated as the ratio of the total number of bp in readset (after QC) to the assembled genome length; [c]The length of contig at which one half of the genome is covered by a contig of equivalent size or greater; [d]The number of contigs required to achieve N_{50}; [e]Total number of protein-coding sequences annotated by RAST.

For two of the three trial genomes (*syri2339* and *mors5261*), SPADES (Bankevich et al., 2012) produced assemblies with the highest N_{50}, while IDBA (Peng et al., 2012) produced the best N_{50} for *aesc2113*. Although there was often little difference between these two assemblers, they frequently achieved an N_{50} that was about 2–5 times higher than the other three programs. Both SPADES and IDBA are iterative de Bruijn graph assemblers that incrementally increase the *k*-mer value (the length of the substring used to construct the de Bruijn graph) during the assembly process. Iteration provides a solution to common assembly issues such as the 'gap problem' (where not every region in the genome is represented in the set of reads due to non-uniform or low coverage) and the 'branching problem' (where repeat regions or read errors result in multiple possible assembly solutions) (Peng et al., 2010). A small *k*-mer size alleviates the gap problem at the expense of the branching problem, and most assemblers, e.g. VELVET and ABYSS, choose a moderate *k*-mer value that represents a trade-off between these sources of error. However, the SPADES and IDBA algorithms construct an 'accumulated' de Bruijn graph that incorporates the advantages of both smaller *k*-mers (fewer gaps) and larger *k*-mers (fewer unresolved regions) into the final assembly (Peng et al., 2010; Bankevich et al., 2012; Peng et al., 2012). These algorithms may perform particularly well on data that has both high coverage and a longer read length, as is the case here.

To gain an insight into variation in gene-content among assemblies generated via different programs, we annotated coding regions (CDS) for each assembly using RAST as before. There were no clear patterns for the number of genes called using different assemblers—no single assembler consistently produced the genome with either the fewest or greatest number of coding regions (as annotated by RAST). It should be noted that the number of genes is not in itself a measure of the quality of that assembly and was not considered as such; for example, we have previously found that breaks in contigs that occur in the middle of CDSs can erroneously inflate the number of given genes in a draft genome (Nowell et al., 2014).

To discern any differences between SPADES and IDBA, we assembled all 26 strains using both programs. A comparison of N_{50} values showed similar overall

Table 4.5: N_{50} comparison.

Assembler	Min	Median	Mean	Max
IDBA	40,560	62,000	93,840	300,600
SPAdes	41,760	66,290	92,710	246,400

All values in bp.

results, with SPADES showing an increase in the median N_{50} of about 6 kb over that of IDBA (TABLE 4.5).

In addition, an extensive evaluation of *de novo* assemblers with specific reference to bacterial genome assembly has recently been carried out by Magoc et al. (2013). The authors used a range of programs, including ABYSS, SPADES and VELVET, to assemble 250 bp paired-end MiSeq data for strains of *Bacillus cereus*, *Rhodobacter sphaeroides*, *Mycobacterium abscessus* and *Vibrio cholerae*, all of which have a complete reference genome (unlike here). The comparison of these assemblies to a complete reference genome enabled the estimation of the number of assembly errors, defined as misjoins in contigs or scaffolds when compared to the complete, 'true' genome sequence. The authors conclude that the SPADES assembler performs well given a single library of long (i.e., 250 bp) reads at a reasonably high coverage (around 100 X), consistently producing low numbers of misassemblies, high values of N_{50} and a good gene-capture rate, defined as the number of protein-coding sequences (CDS) fully contained within contigs (Magoc et al., 2013). Thus, based on our own findings as well as those given by Magoc et al. (2013), the final assembly for all strains was generated using the SPADES algorithm (Bankevich et al., 2012).

4.4.2 Assembly results

Assembly information for all strains is given in TABLE 4.6. Assembly coverage, defined as the number of base pairs in the read set (after quality control) divided by the assembled genome size, was reasonably high, with a median value of 90.6 X and a range from 41.5 to 167.2 X. Genomes ranged in overall size from 5.62 to 6.47 Mb, with a median size of 6.19 Mb. N_{50} ranged from 41.8 to 246.4 kb with a median value of 66.3 kb. The number of contigs in a given assembly was more variable, although

all genomes were assembled into fewer than 400 contigs. Excluding small sequences < 500 bp, the number of contigs ranged from 59 (*syri2339*) to 333 (*daph4219*), with a median value of 229 contigs. Average G + C content was 58.4%, ranging from 57.8 to 59.3%, representing typical values for *P. syringae*. The number of coding sequences (CDS) annotated in these genomes ranged from 5,058 (*syri7872*) to 5,933 CDS (*ulmi1407*), with a median of 5,684 CDS, which are also typical values for gene content in *P. syringae* genomes (e.g., Baltrus et al., 2011).

4.4.3 Phylogenetics and whole-genome comparisons

We used an MLST concatenation approach to elucidate the phylogenetic position of the 26 newly sequenced strains within the *P. syringae* species complex. The maximum likelihood phylogeny for 128 members representing seven phylogroups of this diverse bacterial group is shown in FIG. 4.1. Generally, this phylogeny is consistent with previous analyses based on MLST data (e.g., Sarkar and Guttman, 2004; Hwang et al., 2005; Parkinson et al., 2010). The majority (19 of 26) of newly sequenced strains cluster in PG3, including pathovars *aesculi* (×6), *broussonetiae*, *castanae*, *cerasicola*, *daphniphylli*, *dendropanacis*, *eriobotryae*, *fraxini*, *morsprunorum* strains 2341 and 5269, *myricae*, *nerii*, *rhaphiolepidis* and *ulmi*. Five strains cluster in PG2, including pv.'s *syringae* (×4) and *papulans*, while two strains, pv.'s *avii* and *morsprunorum* strain 5261, cluster in PG1. The phylogenetic positioning of woody-host strains within PG1, PG2 and PG3 is discussed in detail later (CHAPTERS 5 and 6).

We used BRIG (Alikhan et al., 2011) to visualise and qualitatively assess genome content variation among the 26 newly sequenced genomes, using the three completed genomes of pv's. *tomato* DC3000 (Buell et al., 2003), *syringae* B728a (Feil et al., 2005) and *phaseolicola* 1448A (Joardar et al., 2005) as reference sequences (FIG. 4.4.3). This illustrates clearly the mosaic structure of *P. syringae* genomes. All three references show numerous unique genome regions, represented by transectional segments with no shading in any ring, as well as other regions with a more patchy distribution of presence in some strains but not others. Strain-specific regions for each reference frequently correspond to areas of high AT-richness, and most

Table 4.6: Assembly statistics.

Identifier	Reads	Cov (X)	Contigs (>500 bp)	N_{50}	L_{50}	Genome length (bp)	%GC	CDS	tRNA
aesc2113	4,785,760	93.2	330 (264)	60,900	34	6,165,885	58.26	5,644	66
aesc2279	3,643,588	68.4	322 (266)	59,035	36	6,198,893	58.28	5,688	66
aesc2306	2,656,716	120.4	291 (255)	66,232	32	6,233,619	58.27	5,734	65
aesc2315	4,676,544	87.9	289 (238)	60,957	34	6,132,449	58.29	5,623	66
aesc2329	3,379,080	113.9	319 (263)	60,902	32	6,270,659	58.26	5,797	65
aesc2336	4,993,386	144.2	288 (248)	66,242	32	6,205,117	58.27	5,717	66
avii3648	5,588,938	69.3	324 (256)	53,970	34	6,234,384	58.65	5,680	63
brou5140	2,828,564	77.2	359 (302)	43,556	42	6,107,623	57.80	5,784	63
cast4217	2,998,406	66.3	220 (176)	85,390	25	6,264,218	57.91	5,710	63
cera6109	2,601,214	104.4	353 (307)	41,757	38	5,747,038	58.33	5,415	63
daph4219	3,728,226	113.1	370 (333)	41,937	46	6,273,474	58.10	5,697	72
dend3226	4,402,796	79.7	219 (202)	66,340	24	5,874,436	58.08	5,334	58
erio2343	3,215,470	82.7	129 (119)	133,333	16	6,321,961	57.82	5,733	66
frax5062	3,331,228	77.7	331 (281)	68,855	28	6,265,151	57.94	5,723	63
mors2341	3,050,932	101.8	173 (152)	169,197	12	6,282,759	58.07	5,692	66
mors5261	3,976,422	72.9	264 (221)	83,133	27	6,471,635	58.44	5,887	65
mors5269	3,206,302	104.2	158 (116)	145,957	13	6,185,459	58.11	5,580	66
myri2897	4,110,878	67	204 (187)	98,368	19	6,029,350	58.19	5,421	68
neri5067	2,615,006	54.2	242 (220)	56,758	33	5,793,692	58.22	5,249	61
papu1754	2,053,796	41.5	174 (158)	92,708	20	6,181,181	58.84	5,705	67
rhap4220	1,889,286	97.3	292 (275)	42,737	40	5,615,841	58.26	5,159	62
syri2339	3,632,210	97.7	69 (59)	173,289	12	6,122,834	59.23	5,246	67
syri2340	3,996,704	167.2	98 (82)	246,416	10	6,182,129	59.13	5,354	69
syri7872	7,058,622	124	105 (73)	212,122	8	5,900,176	59.34	5,058	66
syri7924	4,955,082	53.7	130 (112)	123,509	18	6,240,460	59.11	5,478	66
ulmi1407	2,061,050	123.7	323 (296)	56,873	35	6,404,243	57.92	5,933	66

All genomes assembled using SPAdes; statistics defined as previously.

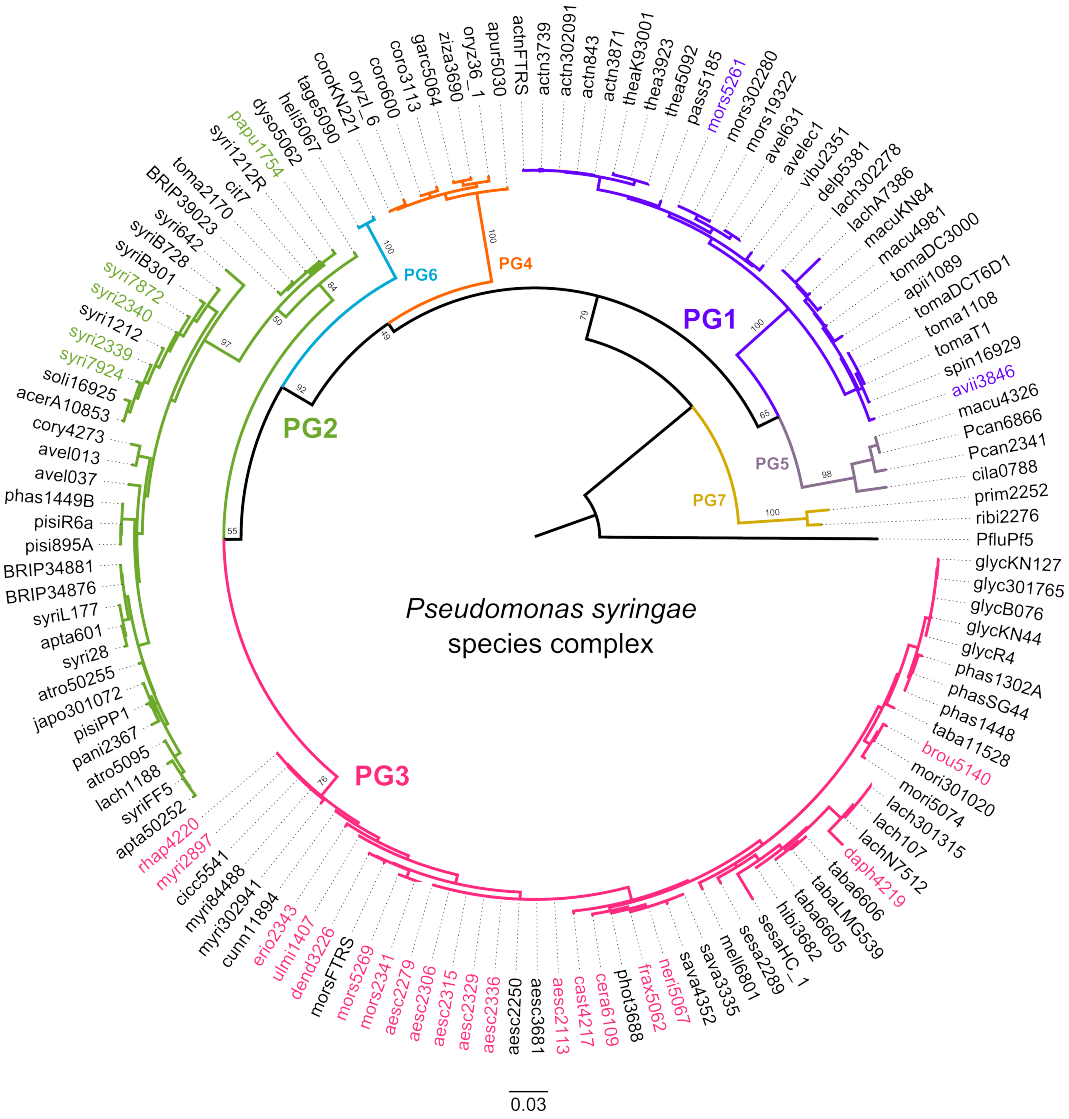


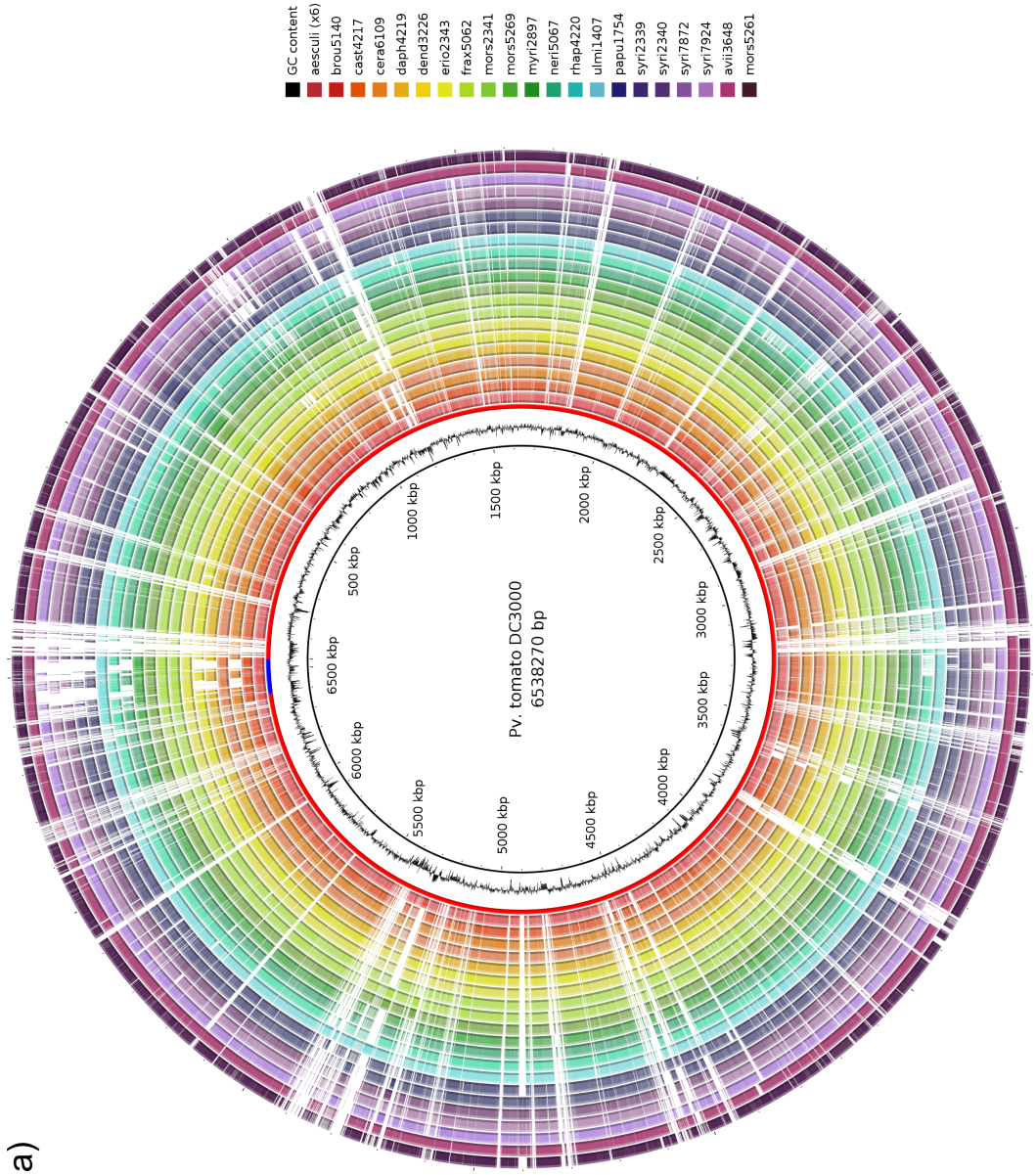
Figure 4.1: Phylogenetic position of newly sequenced strains within the *P. syringae* species complex *sensu lato*. Maximum likelihood phylogeny is based on the 1,965 bp concatenated alignment of partial sequences for *gyrB*, *rpoD*, *gap1* and *gltA*. Branch colours delimit strains into the major phylogroups (PG1–7, shown on branches). Coloured taxon names indicate the position of the 26 newly sequenced woody-host strains, and which phylogroup each falls into. Numbers on branches show bootstrap support, shown for major partitions only. Tree is rooted with *P. fluorescens* Pf-5 (*PfluPf5*); scale bar represents 0.03 nucleotide substitutions per site.

likely represent recent acquisitions into these genomes via horizontal gene transfer (HGT). The frequency of such regions in all three reference sequences highlights the contribution of HGT and gene loss to genome-content variation in *P. syringae* genomes.

The plasmids of pv.'s *tomato* DC3000 and *phaseolicola* 1448A are shown in blue in FIG. 4.4.3, and constitute regions of increased variability among the 26 strains. In particular, the four pv. *syringae* assemblies display an absence of homology to the two plasmids found in either the DC3000 or the 1448A strain. The plasmid content of these strains is discussed in more detail below, however, it is interesting to note that the assemblies of both PG1 and PG3 strains show similarity to plasmids in both the PG1 and PG3 references; whether these homologous sequences are also found on plasmids in our strains is not known.

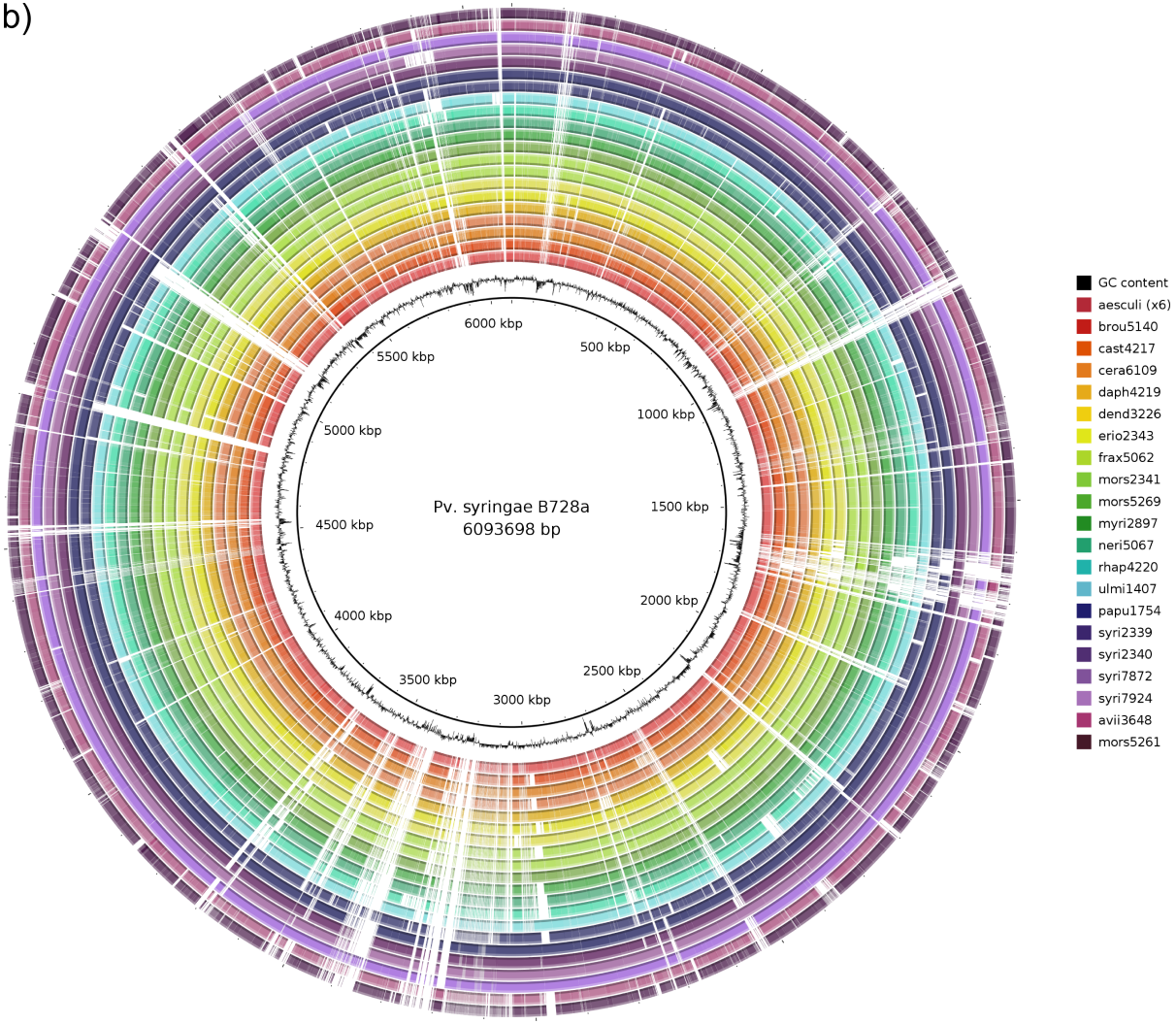
Finally, we note the relative lack of obvious phylogenetic signal in genome-content variation among strains. Although phylogroup-specific regions are apparent for all three comparisons, the majority of variation appears to be strain-specific—either regions gained in the reference that are absent from all other strains, or regions that have been lost from individual strains. This is in contrast with previous results, which showed genome-content variation at the level of the individual gene to be a good predictor of phylogeny in *P. syringae* (Nowell et al., 2014).

Figure 4.2 *(following page)*: Whole-genome comparisons between 26 newly sequenced *P. syringae* genomes and the three completed genomes of: **(a)** pv. *tomato* DC3000 (PG1); **(b)** pv. *syringae* B728a (PG2); and **(c)** pv. *phaseolicola* 1448A (PG3). Strains are arranged as per phylogroup, PG3 to PG1 from in to out; this order is also repeated within each key. %GC content of each reference genome is shown in black. For DC3000 and 1448A, the innermost ring demarks the main chromosome (red) from plasmid sequences (blue); B728a does not contain plasmids. Each coloured ring then shows the level of sequence identity between each genome and the reference (the six pv. *aesculi* strains are combined into a single ring). For each comparison, the three shades of that colour (white, light and dark) correspond to < 50, 51–69 and $\geq 70\%$ identity, respectively.

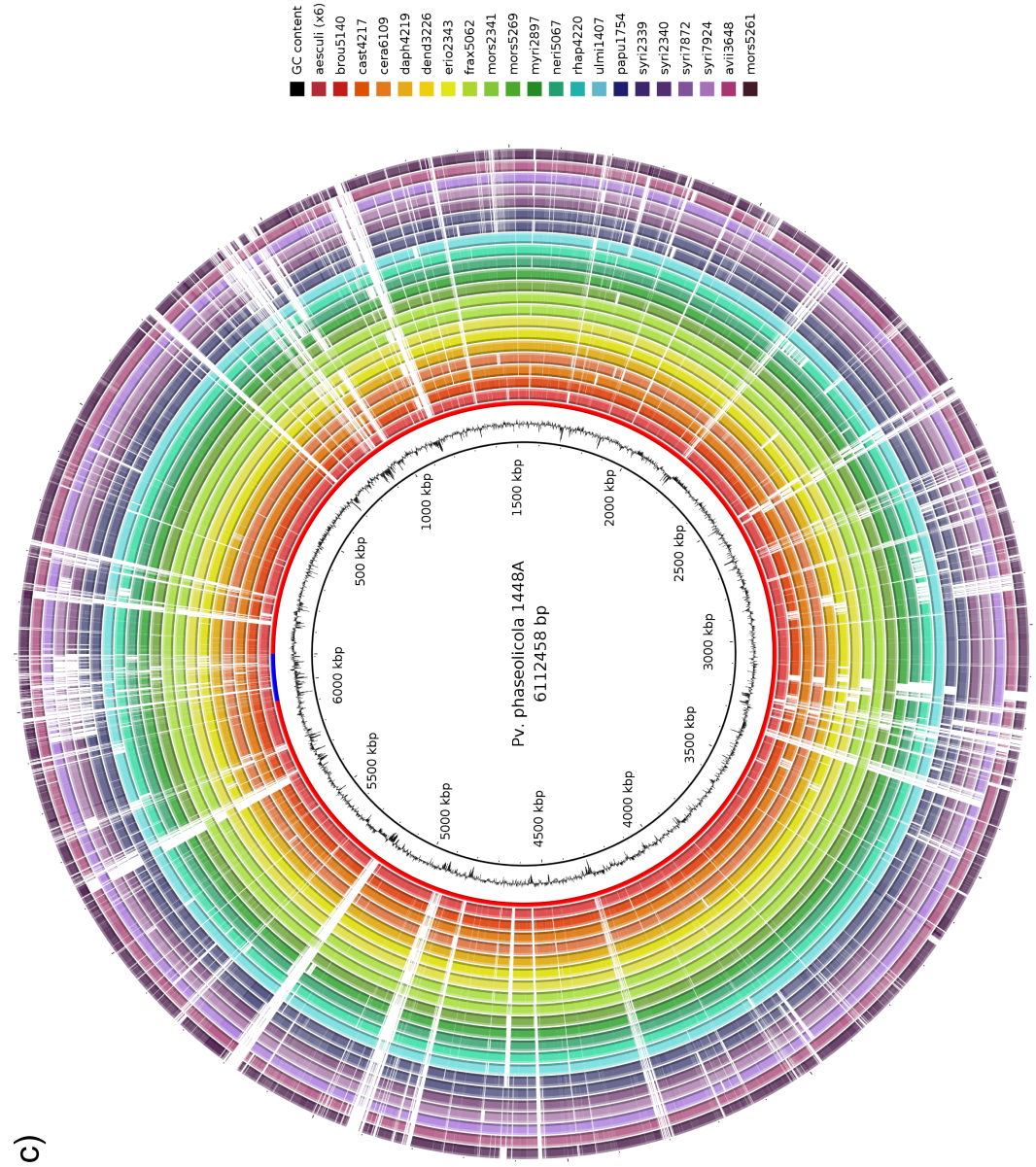

GC content
aesculi (x6)
brou5140
cast4217
cera6109
daph4219
dend3226
erio2343
frax5062
mors2341
mors5269
myri2897
neri5067
rhap4220
ulmi1407
papu1754
syri2339
syri2340
syri7872
syri7924
avii3648
mors5261
Pv. tomato DC3000
6538270 bp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp
5500 kbp
6000 kbp
6500 kbp

a)

b)


Pv. syringae B728a
6093698 bp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp
5500 kbp
6000 kbp
GC content
aesculi (x6)
brou5140
cast4217
cera6109
daph4219
dend3226
erio2343
frax5062
mors2341
mors5269
myri2897
neri5067
rhap4220
ulmi1407
papu1754
syri2339
syri2340
syri7872
syri7924
avii3648
mors5261

c)


GC content
aesculi (x6)
brou5140
cast4217
cera6109
daph4219
dend3226
erio2343
frax5062
mors2341
mors5269
myri2897
neri5067
rhap4220
ulmi1407
papu1754
syri2339
syri2340
syri7872
syri7924
avii3648
mors5261
Pv. phaseolicola 1448A
6112458 bp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp
5500 kbp
6000 kbp

4.4.4 Mobile genetic elements

Plasmids

The proportion of each genome likely to be comprised of plasmid sequences was estimated using a BLAST-based approach, and is shown in TABLE 4.7. Overall, putative plasmid sequences constituted between 0–2.4% of our assembled genomes, with a mean value of 1.2% (approximately equivalent to 80 kb of the average *P. syringae* genome length). In comparison, the completely sequenced pv.'s *phaseolicola* str. 1448A and *tomato* str. DC3000 each contain 2 plasmids each, comprising 3% and 2.2% of their total genome length respectively. Since this analysis excludes unsampled plasmid sequences (i.e., that were not present in the plasmid database downloaded from GenBank at the time of the analysis) and repeated sequences that may be found on both the chromosome and plasmids, the estimated proportion of plasmid sequence for our assemblies is probably an underestimate of the true value.

P. syringae pv. *syringae* str. B728a, in phylogroup 2, does not harbour any plasmids (Feil et al., 2005). Accordingly, we find no contigs with homology to known plasmids in the three PG2 strains *syri2339*, *syri2340* and *syri7924*, while the fourth (*syri7872*) shows only a small ∼2.5 kb region identified as plasmid using our criteria. However, strain *papu1754*, which also clusters within PG2, comprises at least 52 kb with homology to known plasmids.

The proportion of total contigs comprising putative plasmid sequence ranged from 3.2–16% per genome, averaging ∼8% (discounting both complete genomes and PG2 strains with no inferred plasmid sequences). Thus, putative plasmid sequences constitute a much higher proportion of the total number of contigs in a given assembly than they do in terms of length, suggesting a high degree of fragmentation most likely due to the accumulation of repeat sequences such as transposable elements in plasmid genomes.

Table 4.7: Plasmid content.

Strain	Number of contigs[a]	Proportion contigs (%)	Total length (bp)	Proportion genome size (%)
aesc2113	40	12.1	113,984	1.8
aesc2279	23	7.1	97,100	1.6
aesc2306	42	14.4	113,341	1.8
aesc2315	11	3.8	90,495	1.5
aesc2329	52	16.3	153,256	2.4
aesc2336	22	7.6	106,964	1.7
avii3648	22	6.8	75,442	1.2
brou5140	27	7.5	106,091	1.7
cast4217	7	3.2	33,562	0.5
cera6109	32	9.1	120,493	2.1
daph4219	14	3.8	50,366	0.8
dend3226	9	4.1	24,210	0.4
erio2343	12	9.3	95,405	1.5
frax5062	35	10.6	100,735	1.6
mors2341	24	13.9	80,157	1.3
mors5261	12	4.5	70,021	1.1
mors5269	11	7.0	63,468	1.0
myri2897	15	7.4	104,132	1.7
neri5067	11	4.5	27,200	0.5
papu1754	7	4.0	51,933	0.8
phas1448A[b]	—	—	183,661	3.0
rhap4220	12	4.1	53,962	1.0
syri2339	0	0	0	0
syri2340	0	0	0	0
syri7872	6	5.7	2,541	∼0
syri7924	0	0	0	0
syriB728a[b]	—	—	0	0
tomaDC3000[b]	—	—	141,134	2.2
ulmi1407	33	10.2	122,487	1.9

[a]Number of contigs putatively derived from plasmids; [b]Pathovars with completely sequenced genomes are shown in bold for reference; *phas1448A* and *tomaDC3000* harbour two plasmids each; *syriB728a* has none.

Transposable elements

The distribution of insertion sequence (IS) elements among strains is shown in FIG. 4.3. It is not possible to accurately quantify the number of IS elements in unfinished 'draft' genomes, such as those reported here. Thus, we define only presence and absence for the proteins encoded by each element using the 80/80 rule as defined previously.

The 26 strains showed a rich diversity of IS elements both within and among strains, with evidence for the presence of at least 33 different elements from 10 IS families. Across strains, *syri7872* showed the least diversity of IS elements matches to proteins encoded by five different elements, while strain *avii3846* showed the presence of at least 23 different elements. Generally, the four pv. *syringae* strains in PG2 showed a paucity of IS elements relative to other strains in either PG1 and PG3. There is little evidence of phylogroup specificity of IS elements.

The abundance of transposable elements within *P. syringae* genomes is thought to be the primary difficulty in obtaining closed genome sequences for this species, despite high coverage (e.g., Qi et al., 2011). We investigated the potential reason for gaps in our genomes by using BLASTx to query a 200 bp portion from the termini of each contig of our 26 assemblies versus the non-redundant protein database *nr* (*E*-value $\leq$ 1e$-$5). Considering only the top hit for each query, we found that the number of annotations for contig termini that contained the words "transposase" or "transposition" ranged 16–74% among strains, and that this proportion was significantly positively correlated with the number of contigs in the assembly ($t = 5.1$, $p \ll 0.001$; FIG. 4.4). The PG2 pv. *syringae* strains cluster at the lower end of this relationship, showing both lower numbers of contigs and a lower proportion of hits to transposable elements (despite no drastic reduction in total assembly length).

4.5 Discussion

Here, we describe the sequencing, assembly and annotation of 26 strains of *P. syringae* isolated from a variety of woody hosts. Strains have been sampled from a wide

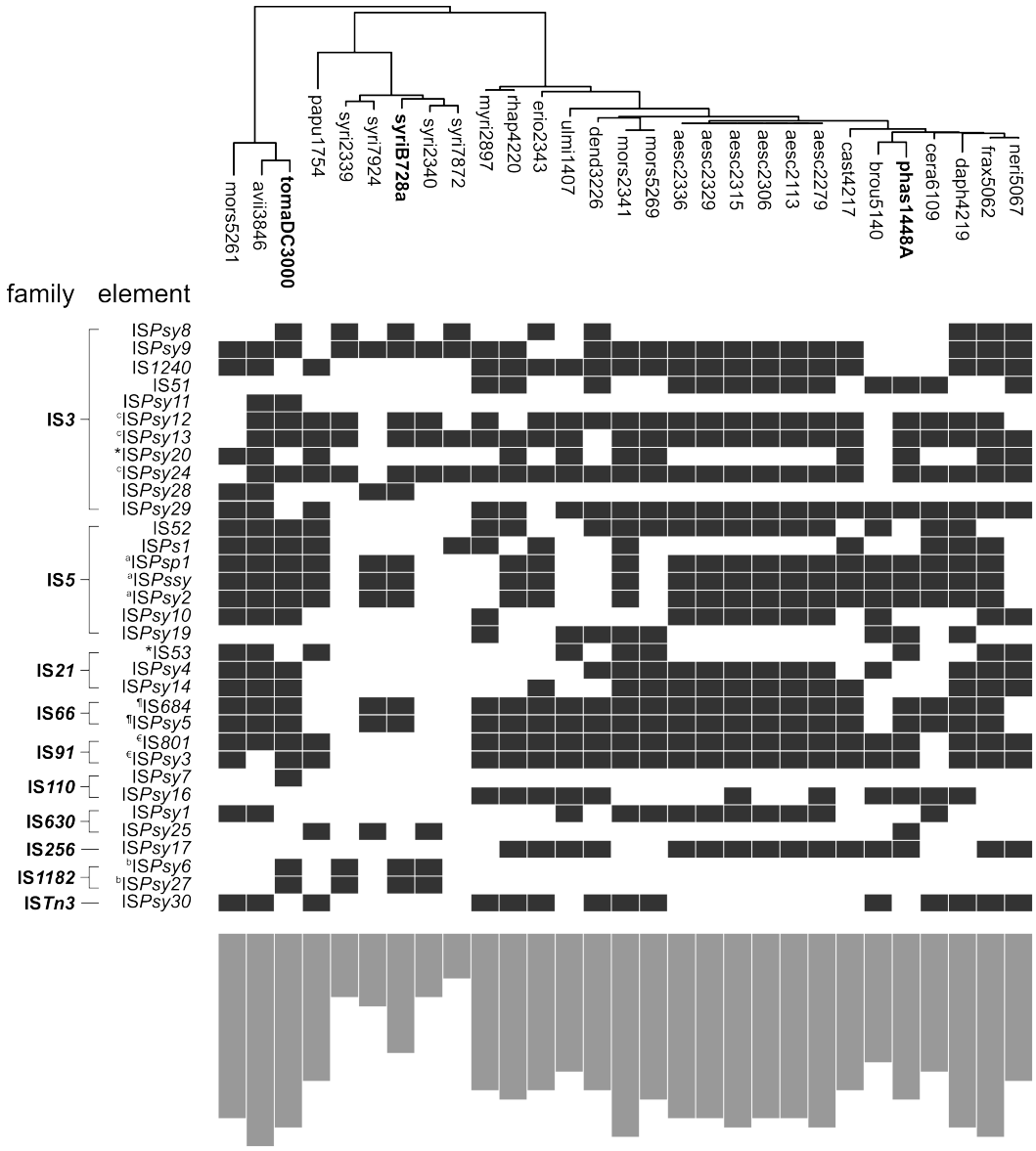


Figure 4.3: Distribution of IS elements across strains. The three complete genomes *tomaDC3000*, *syriB728a* and *phas1448A* are included for reference (shown in bold); phylogeny constructed from $\sim$ 6.5 kb alignment of the genes *gap1*, *gltA*, *gyrB* and *rpoD* for these 29 strains, using the same method as per Materials & Methods. Elements are arranged per family, as shown on the left. Superscript symbols indicate elements that are closely related to each other (80% identity over 80% sequence length). Bars at the bottom indicate the proportion of elements found in each strain. Presence is defined as a minimum of 80% identity over at least 80% of the query length in each case.

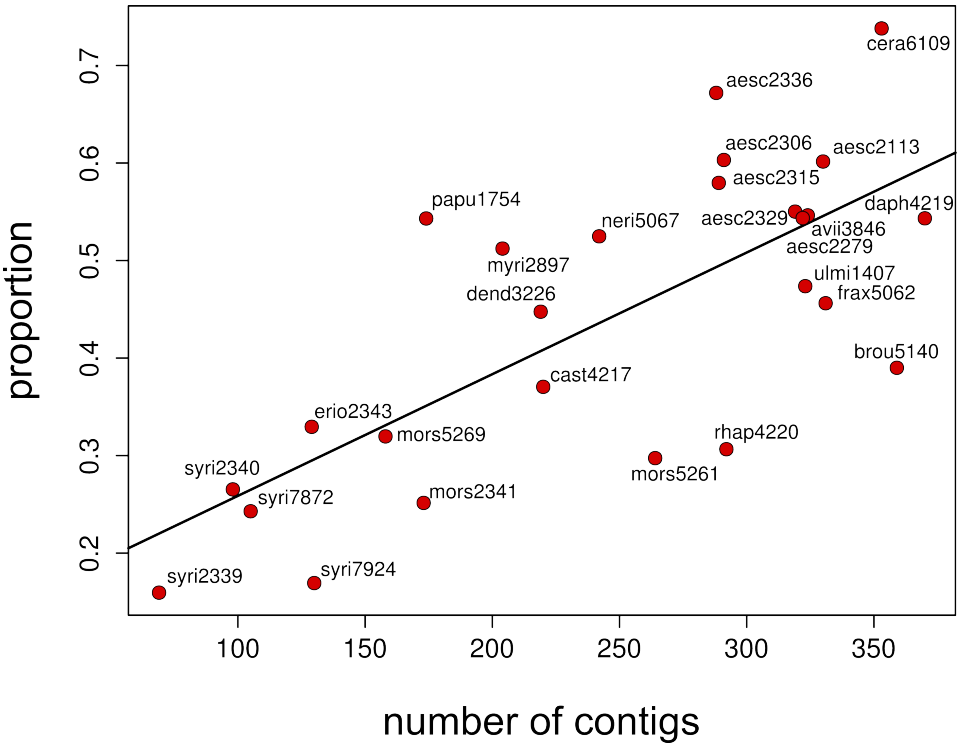


Figure 4.4: Relationship between the proportion of contig termini with hits to transposable elements (shown on y-axis) and the number of contigs in a given genome assembly (x-axis). The linear correlation is highly significant ($t = 5.15$, $p \ll 0.001$; correlation coefficient $= 0.505$).

range of geographic locations over a time period about 50 years, and have been drawn from across the three major phylogroups within the *P. syringae* species group. Thus, we consider these data to be a good representation of the potential diversity of woody-host lineages within *P. syringae*. Using these data, we hope to gain an insight into the evolutionary adaptations that have facilitated the movement of these bacteria into the woody niche, and, in particular, onto the European horse chestnut.

Below, we discuss some of the preliminary characteristics of these strains, such as the sampling distribution, host species distribution, and genome characteristics such as the presence of mobile elements. Finally, we consider the merits of whole-genome sequence data and note the unprecedented rate at which this technology is becoming available.

4.5.1 Sampling across *P. syringae* diversity

A high proportion ($\sim$ 27%) of strains originate from Japan and/or East Asia, sampled primarily by a single group of researchers over the latter part of the 20th century (Ogimi and Higuchi, 1981; Ogimi et al., 1988, 1990, 1992). It is unclear whether this represents a genuine increase in bacterial diseases of woody host species in this region at that time, which was met with a corresponding increase in the efforts to study them, or whether sampling biases or other factors have resulted in the over-representation of East Asian strains in the international culture collections such as NCPPB and CFBP.

The 26 strains reported here have been isolated from an extremely diverse range of plant host species found across the eudicots. Nonetheless, it is clear that members of the order Rosales are frequent hosts for *P. syringae* strains. The Rosales itself is a broad and speciose group, containing over 7,000 species delimited into about 260 genera. However, 11 strains of six distinct pathovars are pathogens of members of the Rosaceae family within the Rosales. In particular, species of the *Prunus* genus, which includes stone fruit species such as peach, plum, cherry, nectarine and apricot, appear vulnerable to attack by a phylogenetically diverse range of *P. syringae* pathovars. For example, pv. *morsprunorum* from both PG1 and PG3, pv. *avii*

(PG1) and pv. *syringae* strains 2339, 7872 and 7924 (PG2) were all isolated from either sweet cherry (*P. avium*) or its close relative the sour cherry (*P. cerasus*).

The ability of multiple, divergent lineages of *P. syringae* to cause disease on the same host has been previously investigated with regards to the independent evolution of two divergent pv. *avellanae* lineages onto the hazelnut, *Corylus avellana* (Scortichini et al., 2002, 2006; Wang et al., 2007). Pathovar *avellanae* is found in both PG1 and PG2, and a recent comparative genomics analysis found little evidence of convergent evolution in terms of the repertoire of virulence genes encoded by these divergent *avellanae* lineages (O'Brien et al., 2012). It will be interesting to ascertain if a similar lack of convergence is evident in the adaptation of woody-host strains at both at the niche level, in terms of adaptation into the woody environment, and at the host level, for example in the case of the adaptation of divergent strains onto *Prunus* spp.

4.5.2 Genome characteristics

As for other *P. syringae* genomes, the 26 strains represented here are rich in mobile genetic elements such as insertion sequences and plasmids. All strains show the presence of multiple different IS families, and all but four strains are likely to contain plasmids. The abundance of repeat elements explains, in part, the high frequency of genome rearrangements observed in *P. syringae* (Wozniak and Waldor, 2010), and is an underlying factor in the difficulty in obtaining complete genome sequences for this species (Baltrus et al., 2011; Qi et al., 2011). This is corroborated by the positive correlation between the number of contigs in a given draft assembly and the proportion of contig termini from that assembly with a match to an IS element. This indicates that, for some strains at least, the majority of contigs terminate within or near a CDS encoding a transposon protein. Thus, while a significant resequencing effort would be required to close the gaps in our draft assemblies, the fact that contigs terminate at regions where they might be expected to—that is, at repeat regions—suggests that our sequencing and assembly strategy has performed well in recovering these genomes. Despite this, it is worth acknowledging that gaps in

draft genomes preclude complete confidence when assigning gene absences, and it is also possible that some genes (particularly plasmid-borne) may have been lost during the passage of these strains in the laboratory.

We note that the four assemblies with the fewest gaps (*syri2339*, *syri2340*, *syri7872* and *syri7924*) showed none or very few contigs with homology to a database containing over 4,500 plasmid sequences. This agrees with the absence of plasmids in other pv. *syringae* genomes, such as the complete sequence of pv. *syringae* str. B728a (Feil et al., 2005). The lack of plasmids in these strains may explain the greater degree of completedness of these assemblies, since plasmids are often littered with repeated sequences and can be notoriously difficult to assemble (Siguier et al., 2006a; Bardaji et al., 2011b). Conversely, it is likely that the main chromosomes of the other strains, and thus the majority of their genic content, are covered by a relatively small number of highly contiguous sequences.

4.5.3 *P. syringae* in the genomics era

At the time of writing, the generation of these data represents an increase in the genomics resource available to the *P. syringae* research community of about 57% at the genome level, from 46 to 72, or about 65% at the pathovar level, from 20 to 33. However, the rate of advancement in the field of comparative genomics and genome analysis is such that we predict that our data will soon be subsumed by a plethora of new sequence data for *P. syringae*.

As such, it is impossible not to note the remarkable rate at which the field of bacterial genomics is progressing. The first complete bacterial genome was the 1.8 Mb sequence of the opportunistic human pathogen *Haemophilus influenzae*, achieved in 1995 by Fleischmann et al. (1995). By the year 2000 complete sequences for over 30 microbial species were available; by 2006 the number was over 300 (Binnewies et al., 2006). Today, the total number of bacterial genomes (complete plus draft) available for download from the Pathosystems Resource Integration Center (PATRIC, `http://patricbrc.org/`) stands at approximately 22,000, corresponding to a production rate of just under 100 genomes per month since the publication of

the *H. influenzae* sequence. In addition, the establishment of a competitive genomics industry has fuelled both the advancement of sequencing technologies and the reduction in sequencing costs. Thus, the rate of decrease of DNA sequencing costs is now predicted to be below that of Moore's law, and makes the US National Human Genome Research Institute's aim of a complete human genome for under $1,000 appear an inevitability (Hayden, 2014). Crucially, all these factors combine to allow for the possibility of data generation on this scale within the time and budget limitations of a single PhD project, a concept that was neither feasible nor realistic even a few years ago.

The *P. syringae* research community has not been slow to embrace the genomics revolution. Beginning with the 2003 publication of the complete sequence of the tomato pathogen *P. syringae* pv. *tomato* str. DC3000 (Buell et al., 2003), it continued with the subsequent production of genome data for the bean pathogens pv. *syringae* B728a and pv. *phaseolicola* 1448A (Feil et al., 2005; Joardar et al., 2005). These have remained as the only completely sequenced genomes within the *P. syringae* species complex until the release of two further complete genomes within the last year, that of *P. syringae* pv. *syringae* str. SM, isolated from wheat (A. Dudnik and R. Dudler, unpublished), and *P. cichorii* str. JBC1, isolated from soybean (B.-Y. Kim and Y.H Lee, also unpublished). In line with the global trend, however, the past decade has witnessed the production of a much larger number of genomes in 'draft' status, including investigations that have extensively sampled both across the *P. syringae* phylogeny (e.g., Baltrus et al., 2011; Monteil et al., 2013) and within individual pathovars (Mazzaglia et al., 2012; McCann et al., 2013).

Thus, the data generated here will contribute to the expanding genomic resource for this important group of plant-associated bacterial lineages, and will be used in subsequent work that aims to further our understanding of the genomic basis of microbial evolution and adaptation.

Chapter 5

Population genomics of niche adaptation in *Pseudomonas syringae*

5.1 Abstract

The diversification of lineages within the plant-associated bacterium *Pseudomonas syringae* has involved a number of recent adaptive shifts from herbaceous hosts onto various species of tree, resulting in the emergence of highly destructive diseases such as hazelnut decline and bleeding canker disease of the European horse chestnut. Given the history of extensive gene gain and loss in these bacteria, it is likely that the processes of horizontal gene transfer (HGT) and gene loss have played a central role in the phytoadaptation of *P. syringae* onto woody hosts. Here, we combine existing genome data for *P. syringae* with that for 26 strains isolated from a diverse range of woody host species, and use a maximum likelihood method that accounts for phylogenetic relatedness to test for correlations between host type and gene presence among genes of the flexible genome. We find evidence for a common set of genes associated with the woody niche, suggesting that divergent lineages of *P. syringae* tend towards similar genome compositions that may underpin their adaptation into

a common niche environment. Our analyses allow for a greater understanding of how the flexible genome contributes to host adaptation in *P. syringae*, and how the flexibility of 'niche-specific' genes, mediated through HGT, may facilitate bacterial evolution.

5.2 Introduction

The diversification of lineages within the *Pseudomonas syringae* species complex has involved a number of recent adaptive shifts from herbaceous hosts onto various species of tree, resulting in the emergence of diseases such as bacterial canker of kiwi (Tsuymu and Goto, 1989; Balestra et al., 2010), olive knot disease (Rodríguez-Moreno et al., 2009; Rodríguez-Palenzuela et al., 2010), hazelnut decline (Janse et al., 1996; Scortichini et al., 2002) and bleeding canker of the European horse chestnut (Webber et al., 2008; Green et al., 2009).

Previous chapters have shown evidence of extensive genome fluctuation in *P. syringae*, and have revealed the breadth and depth of the potential gene pool available via horizontal gene transfer (HGT). The magnitude of genome fluctuation is remarkable: individual lineages may be exposed to hundreds, perhaps even thousands, of new genes within the same time-frame as a 1% divergence accrues among protein sequences of the core genome (Nowell et al., 2014). Such large-scale repertoire shuffling allows *P. syringae* lineages to constantly trial novel combinations of genes, and suggests that successful lineages of virulent *P. syringae* pathovars may emerge when a strain is able to invade an environment in which its specific genome composition is selectively advantageous. In addition, it is now known that genetically diverse populations of *P. syringae* thrive in a multitude of environmental (i.e., non-plant) habitats, including leaf-litter, river headwaters, snow-pack and even within clouds (Morris et al., 2008, 2010; Monteil et al., 2012, 2013). Thus, the sobriquet of *P. syringae* as a plant pathogen is perhaps misleading insofar as a plant-pathogenic lifestyle may be the exception rather than the rule for these lineages. Given this naturally occurring reservoir of genetic diversity, Monteil et al. (2013)

have recently suggested an epidemic population structure for *P. syringae*, similar to other bacterial pathogens such as *Neisseria meningitidis* (Maynard Smith et al., 1993) and *Pseudomonas aeruginosa* (Pirnay et al., 2009), whereby clonal expansions of highly virulent lineages emerge from a frequently recombining and genetically diverse background population.

Taken together, these findings suggest that the flexible genomes of phytopathogenic *P. syringae* lineages are tailored to be selectively advantageous when expressed in a particular niche—that of a compatible host species—and implicate HGT and gene loss as key evolutionary mechanisms that facilitate the adaptation process.

In this chapter, we investigate the role of HGT in the adaptation of *P. syringae* into a niche environment that has been colonised multiple times during its evolutionary history—specifically, the woody organs of a woody host species. We have augmented the current genomic resource for *P. syringae* with draft genome sequences of 26 strains pathogenic on a range of other woody species (CHAPTER 4), and now use a comparative phylogenetics method that corrects for phylogenetic relatedness to look for statistical evidence of correlated evolution between genes of the flexible genome and host preference among 64 strains for which high quality whole-genome sequence data is available. By considering both shared and unique flexible genes within the context of a recurring niche invasion, we aim to build a better understanding of exactly how HGT and gene loss contribute to adaptation in this bacterium. Finally, we evaluate the evidence that the diversification of *P. syringae* lineages has been facilitated by the presence of highly structured, niche-specific gene pools that are organised within the ecological framework of the plant species that serve as hosts for these opportunistic plant pathogens.

5.3 Materials & Methods

5.3.1 Sequence data and orthology

Whole-genome data for 18 strains of *P. syringae* (covering nine distinct pathovars) were produced in the interim between CHAPTER 2 and the present work. Such is

the rate of progress in both DNA sequencing and genome assembly technologies, these more recent data are likely to be of a higher quality, with fewer base errors and more contiguous sequences, than those used in CHAPTER 2. Previous work has highlighted the potential for errors in draft genomes, while we also note the significant association between the level of fragmentation of a genome sequence and the number of putative errors (discussed later). Thus, we excluded from further analyses a number of older sequences that showed a high degree of fragmentation, including data for *P. syringae* pv.'s *aceris* M302273 (*acer* in CHAPTER 2), *mori* M301020 (*mori*), *oryzae* 1_6 (*oryz*), *pisi* 1704B (*pisi*) and *syringae* FF5 (*syrF*). In addition, we also excluded two further pv. *tomato* strains, K40 and Max13 (*tomK* and *tomM*) since these strains are highly similar to other pv. *tomato* strains retained in this work.

Thus, our final dataset consisted of 20 strains discussed in CHAPTER 2, 26 strains for which the genome sequencing and assembly is discussed in CHAPTER 4, and 18 strains for which data had been produced recently (downloaded from the National Center for Biotechnology Information (NCBI) GenBank as of June, 2014), yielding a total of 64 strains of 33 pathovars from across the *P. syringae* species complex. Information for all strains used in this study is shown in TABLE 5.1.

To account for potential variation in gene content due to differences and advances in annotation methodologies, all other strains were also re-annotated using RAST, with the exception of the three complete genome sequences of strains *P. syringae* pv. *tomato* str. DC3000, pv. *phaseolicola* str. 1448A and pv. *syringae* str. B728a, which have undergone substantial manual curation for CDS annotation (Buell et al., 2003; Feil et al., 2005; Joardar et al., 2005). The number of coding sequences (CDS) per strain is given in TABLE 5.1.

Proteins were clustered into orthologous groups using ORTHOMCL v2.0.9 (Li et al., 2003; Van Dongen, 2000). The ORTHOMCL pipeline performs an all-versus-all BLAST (E-value $\leq$ 1e$-$5) followed by Markov clustering (MCL) to determine clusters of orthologous proteins. An important parameter during the MCL stage is the inflation index, which controls cluster 'granularity', or the extent to which clusters are broken up. MCL was performed across a range of inflation parameters

Table 5.1: Strain information.

Pathovar	Strain	Identifier[a]	Host	Year[b]	Contigs	CDS[c]	Trait[d]	Reference[e]
actinidiae	MAFF 302091	actn302091	*Actinidia delicosa* (kiwi)	1984	941	5,169	W	Baltrus et al. (2011)
actinidiae	NCPPB 3739	actn3739	*Actinidia delicosa* (kiwi)	1984	815	5,283	W	Marcelletti et al. (2011)
actinidiae	NCPPB 3871	actn3871	*Actinidia delicosa* (kiwi)	1992	466	5,267	W	Marcelletti et al. (2011)
actinidiae	CRAFRU8.43	actn843	*Actinidia delicosa* (kiwi)	2008	585	5,513	W	Marcelletti et al. (2011)
aesculi	NRS 2113	aesc2113	*Aesculus hippocastanum* (European horse chestnut)	2006	330	5,644	W	This study
aesculi	NRS 2250	aesc2250	*Aesculus hippocastanum* (European horse chestnut)	2008	776	5,324	W	Green et al. (2010)
aesculi	NRS 2279	aesc2279	*Aesculus hippocastanum* (European horse chestnut)	2002	322	5,688	W	This study
aesculi	NRS 2306	aesc2306	*Aesculus hippocastanum* (European horse chestnut)	2010	291	5,734	W	This study
aesculi	NRS 2315	aesc2315	*Aesculus hippocastanum* (European horse chestnut)	2006	289	5,623	W	This study
aesculi	NRS 2329	aesc2329	*Aesculus hippocastanum* (European horse chestnut)	2011	319	5,797	W	This study
aesculi	NRS 2336	aesc2336	*Aesculus hippocastanum* (European horse chestnut)	2010	288	5,717	W	This study
aesculi	NRS 3681	aesc3681	*Aesculus indica* (Indian horse chestnut)	1979	841	5,293	W	Green et al. (2010)
aptata	DSM 50252	apta50252	*Beta vulgaris* (sugar beet)	1948	3,776	5,265	H	Baltrus et al. (2011)
atrofaciens	DSM 50255	atro50255	*Triticum aestivum* (wheat)	1974	669	5,040	H	Baltrus, D.A., et al. (unpublished)
atrofaciens	LMG 5095	atro5095	*Triticum aestivum* (wheat)	1974	1,007	5,160	H	Noh, Y.-H. and Cha, J.-S. (unpublished)

Continued on next page.

Table 5.1: Continued from previous page.

Pathovar	Strain	Identifier	Host	Year	Contigs	CDS	Trait	Reference
avellanae	ISPaVe037	avel037	*Corylus avellanae* (hazel)	1992	317	5,321	W	O'Brien et al. (2012)
avellanae	ISPaVe013	avel013	*Corylus avellanae* (hazel)	1992	191	5,172	W	O'Brien et al. (2012)
avellanae	BPIC631	avel631	*Corylus avellanae* (hazel)	1976	1,602	5,228	W	O'Brien et al. (2012)
avellanae	CRAFRUec1	avelec1	*Corylus avellanae* (hazel)	2003	547	5,160	W	Scortichini et al. (2013)
avii	CFBP 3846	avii3846	*Prunus avium* (cherry)	1991	324	5,680	W	This study
—	BRIP 34876	BRIP34876	*Hordeum vulgare* L. (barley)	1971	148	5,119	H	Gardiner et al. (2013)
—	BRIP 34881	BRIP34881	*Hordeum vulgare* L. (barley)	1971	157	5,136	H	Gardiner et al. (2013)
—	BRIP 39023	BRIP39023	*Hordeum vulgare* L. (barley)	1988	34	5,123	H	Gardiner et al. (2013)
brousonetiae	CFBP 5140	brou5140	*Broussonetia kazinoki* (paper mulberry)	1980	359	5,784	W	This study
castaneae	CFBP 4217	cast4217	*Castanea crenata* (Japanese chestnut)	1977	220	5,710	W	This study
cerasicola	CFBP 6109	cera6109	*Prunus yedoensis* (Yoshino cherry)	1995	353	5,415	W	This study
—	Cit7	cit7	*Citrus sinensis* (navel orange)	2008	2,655	5,321	H	Baltrus et al. (2011)
daphniphylli	CFBP 4219	daph4219	*Daphniphyllum teijsmanni* (—)	1981	370	5,697	W	This study
dendropanacis	CFBP 3226	dend3226	*Dendropanax trifidus* (ivy tree)	1979	219	5,334	W	This study
eriobotryae	CFBP 2343	erio2343	*Eriobotrya japonica* (loquat tree)	1970	129	5,733	W	This study
fraxini	CFBP 5062	frax5062	*Fraxinus excelsior* (ash tree)	1978	331	5,723	W	This study
glycinea	B076	glycB076	*Glycine max* (soybean)	2007	104	5,613	H	Qi et al. (2011)
glycinea	race 4	glycR4	*Glycine max* (soybean)	1977	108	5,314	H	Qi et al. (2011)
japonica	MAFF 301072	japo301072	*Hordeum vulgare* L. (barley)	1951	4,661	5,562	H	Baltrus et al. (2011)
lachrymans	MAFF 301315	lach301315	*Cucumis sativus* (cucumber)	1975	791	6,275	H	Baltrus et al. (2011)
lachrymans	MAFF 302278	lach302278	*Cucumis sativus* (cucumber)	1935	798	5,265	H	Baltrus et al. (2011)
maculicola	ES4326	macu4326	*Raphanus sativus* (radish)	1965	878	5,475	H	Baltrus et al. (2011)
morsprunorum	NRS 2341	mors2341	*Prunus cerasus* (wild cherry)	1988	173	5,692	W	This study
morsprunorum	MAFF 302280	mors302280	*Prunus domesticus* (European plum)	1977	969	5,338	H	Baltrus et al. (2011)

Continued on next page.

Table 5.1: Continued from previous page.

Pathovar	Strain	Identifier	Host	Year	Contigs	CDS	Trait	Reference
morsprunorum	HRI-W 5261	mors5261	*Prunus avium* (sweet cherry cv. Roundel)	1990	264	5,887	W	This study
morsprunorum	HRI-W 5269	mors5269	*Prunus cerasus* (sour cherry)	1990	158	5,580	W	This study
myricae	CFBP 2897	myri2897	*Myrica rubra* (Chinese bayberry)	1978	204	5,421	W	This study
nerii	CFBP 5067	neri5067	*Nerium oleander* (oleander)	1979	242	5,249	W	This study
panici	LMG 2367	pani2367	*Panicium miliaceum* (proso millet)	1963	148	5,154	H	Liu et al. (2012)
papulans	CFBP 1754	papu1754	*Malus sylvestris* (crab apple)	1973	174	5,705	W	This study
phasiolicola	1448A	phas1448A	*Phaseolus vulgaris* (common bean)	1985	3	5,172	H	Joardar et al. (2005)
pisi	PP1	pisiPP1	*Pisum sativum* (pea)	1978	256	5,157	H	(Baltrus et al., 2014)
rhaphiolepidis	CFBP 4220	rhap4220	*Rhaphiolepis umbellata* (yeddo hawthorn)	1980	292	5,159	W	This study
savastanoi	NCPPB 3335	sava3335	*Olea europaea* (olive tree)	1984	403	5,194	W	Rodríguez-Palenzuela et al. (2010)
syringae	1212	syri1212	*Pisum sativum* (pea)	—	338	5,324	H	(Baltrus et al., 2014)
syringae	NRS 2339	syri2339	*Prunus avium* (sweet cherry)	1984	69	5,246	W	This study
syringae	NRS 2340	syri2340	*Pyrus* sp. (pear)	1985	98	5,354	W	This study
syringae	642	syri642	Not stated[f]	2007	296	5,100	H	Clarke et al. (2010)
syringae	HRI-W 7872	syri7872	*Prunus domestica* (plum cv. Opal)	2000	105	5,058	W	This study
syringae	HRI-W 7924	syri7924	*Prunus cerasus* (sour cherry)	2000	130	5,478	W	This study
syringae	B301D-R	syriB301	*Pyrus communis* L. (pear flower)	1969	81	5,168	H	Dudnik and Dudler (2014)
syringae	B728a	syriB728a	*Phaseolus vulgaris* (snap bean)	1987	1	5,089	H	Feil et al. (2005)
tabaci	ATCC 11528	taba11528	*Nicotiana tabacum* (tobacco)	1905	1,405	5,432	H	Studholme et al. (2009)
tabaci	6605	taba6605	*Nicotiana tabacum* (tobacco)	1967	284	5,441	H	Studholme, D.J., et al. (unpublished)
theae	ICMP 3923	thea3923	*Camellia sinensis* (tea plant)	1974	378	5,633	W	Mazzaglia et al. (2012)

Continued on next page.

Table 5.1: Continued from previous page.

Pathovar	Strain	Identifier	Host	Year	Contigs	CDS	Trait	Reference
tomato	NCPPB 1108	toma1108	*Solanum lycopersicum* (tomato)	1961	304	5,467	H	Cai et al. (2011)
tomato	DC3000	tomaDC3000	*Solanum lycopersicum* (tomato)	1960	3	5,619	H	Buell et al. (2003)
tomato	T1	tomaT1	*Solanum lycopersicum* (tomato)	1986	122	5,583	H	Almeida et al. (2009)
ulmi	CFBP 1407	ulmi1407	*Ulmus* sp. (elm tree)	1958	323	5,933	W	This study

Source abbreviations: ATCC, American Type Culture Collection, Virginia, USA; CFBP, Collection Française de Bactéries Phytopathogènes, Beaucouzé, France; DSM, Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany; HRI-W, Horticultural Research International, Warwick, UK; ICMP, International Collection of Microorganisms from Plants, Auckland, New Zealand; LMG, BCCM/LMG Bacteria Collection, Ghent, Belgium; MAFF, Ministry of Agriculture, Forestry and Fisheries, Ibaraki, Japan; NCPPB, National Collection of Plant Pathogenic Bacteria, York, UK; NRS, Forest Research, Northern Research Station, Midlothian, UK. [a]Unique identifier used in this study; [b]Year of original isolation (if known); [c]Number of open reading frames as annotated by RAST; [d]Trait designation based on host type: H, herbaceous host; W, woody host (see section 5.3.5); [e]Genome sequence reference; [f]Given as ‘small herbaceous weed’ in Clarke et al. (2010).

from 1.2 to 4.8, choosing the final value, 1.5, which optimises the number of single-copy orthologous groups in all 64 strains (Lefébure and Stanhope, 2007; Swingley et al., 2008; Latysheva et al., 2012; Nowell et al., 2014).

The resultant list of putative orthologous groups (OGs) was subjected to a number of quality control procedures (see CHAPTER 2) to improve the inference of orthologous relationships among proteins and to correct for the confounding effects of common errors associated with draft genome sequences.

5.3.2 Gene occurrence matrix

The corrected list of OGs was converted into a binary matrix of gene occurrence across all 13,010 sites of the pan genome. For a given OG, 'presence' was recorded if a strain contributed a member protein to the OG and 'absence' if otherwise. Information regarding gene copy-number within OGs was not retained; i.e., cases where multiple genes from the same genome clustered into a single OG were simply counted as present. Gene accumulation curves were generated using a custom PERL script; for each comparison, genomes were randomly sampled 100 times (with replacement) to generate sample distributions, core genes that were not in single-copy per genome (n = 158) were excluded from the core curve but retained for the pan curve.

5.3.3 Phylogenetics

The evolutionary history of the *P. syringae* core genome was estimated from the concatenated alignment of 2,086 1-to-1 (single-copy) core orthologous genes. Nucleotide alignments were generated using T-COFFEE (Notredame et al., 2000) and concatenated using GENEIOUS v5.4 (Biomatters, available from `http://www.geneious.com/`). Any alignment column containing a gap was removed, yielding a final alignment over 1.15 Mb in length. A maximum likelihood phylogeny was constructed in RAXML v7.2.8 (Stamatakis, 2006), using the general time-reversible model of nucleotide substitution and allowing for rate heterogeneity among sites with four gamma-distributed rate categories (GTR + Γ). Phylogenetic robustness was assessed

with 100 bootstrap analyses.

The level of recombination among *P. syringae* strains was assessed using CLONALFRAME v1.2 (Didelot and Falush, 2007). This program assesses patterns of nucleotide substitutions along an alignment and differentiates between recombination and mutation, simultaneously estimating both the 'clonal frame' phylogeny (i.e., recombination-free) as well as the relative contributions of recombination and mutation to the observed overall nucleotide diversity. A whole-genome alignment (WGA) for a subset of 50 strains was generated using MUGSY v1.2.3 (Angiuoli and Salzberg, 2011) as it was not computationally possible to align highly fragmented genome sequences (despite access to a 94 GiB RAM machine). Thus, we trimmed the dataset by excluding genomes with high numbers of contigs (*apta50252* and *japo301072*) and strains of multiply sampled pathovars with low divergence from a sister strain based on the core genome phylogeny above (*actn302091*, *actn3871*, *aesc2250*, *aesc2279*, *aesc2306*, *aesc2315*, *aesc2329*, *aesc2336*, *atro5095*, *BRIP34876*, *taba11528* and *tomaT1*).

To check for parameter convergence, three instances of CLONALFRAME were run on the WGA of the remaining 50 strains; each for 20,000 MCMC generations with a burn-in of 10,000, sampling from the posterior sample every 10 generations. Convergence of resultant topologies was visually checked using FIGTREE v1.4.0 (available at: `http://tree.bio.ed.ac.uk/software/figtree/`), and population genetics parameters were extracted using the program PARSECF v1.3 (available at: `http://bellinghamresearchinstitute.com/software/index.html`).

5.3.4 Distribution of T3SEs and other virulence factors

Genomes were screened for the presence of type III effector genes (T3SEs) of the type III secretion system (T3SS), known to be a key virulence determinant in *P. syringae*. Sequence data for *P. syringae* T3SEs was downloaded from `www.pseudomonas-syringae.org` (August 16, 2013) and combined with a multi-species T3SE database compiled by Wang et al. (2012) to give a total of 1,729 T3SE proteins. These sequences were queried against the contig data for each strain using tBLASTn (*E*-value $\leq$ 1e$-$5), defining presence using the 80/80 rule as defined in CHAPTER 4. T3SEs

are known to undergo frequent domain restructuring and recombination; thus, we also recorded putatively truncated T3SEs if a hit showed a minimum of 80% identity over $<$ 80% query length.

Using a similar BLAST-based schema, we also screened for the presence of a range of other virulence-associated genes of known or suggested importance in *P. syringae* pathogenicity, including genes involved in the degradation of plant-derived aromatic compounds (Green et al., 2010; Mazzaglia et al., 2012; Bartoli et al., 2014), the metabolism of nitric oxide (Green et al., 2010) and a range of phytohormones and antimetabolite toxins (Bender et al., 1999; Hwang et al., 2005; Arrebola et al., 2011a). The search algorithms BLASTn, BLASTp or tBLASTn were used as applicable (*E*-value $\leq$ 1e$-$5), in most cases defining presence using the 80/80 rule unless stated otherwise.

5.3.5 Statistical modelling of correlated evolution

We modelled correlated evolution between two traits, host type and gene occurrence, using the DISCRETE module of the program BAYESTRAITS v2 (Pagel, 1994; Pagel and Meade, 2006). This method fits continuous-time Markov models to discrete binary data, and calculates the likelihood of two hierarchically nested evolutionary models, one in which two traits are allowed to evolve independently along a phylogenetic tree and another where the two traits evolve in a correlated (dependent) manner (Pagel, 1994; Barker and Pagel, 2005; Pagel and Meade, 2006).

Here, the two traits under consideration are host type and gene occurrence, and the maximum likelihood core genome phylogeny was used to infer evolutionary relationships among strains. We define host type as a discrete binary trait designated either 'herbaceous' or 'woody' ('H' and 'W' respectively), dependent on the ability of each individual strain to exist and proliferate (often causing disease) within the woody organs of its plant host species in its natural environment. The 'herbaceous' designation is defined loosely, and it is worth noting that the assignation of a given strain as 'H' does not necessarily require that the strain is a pathogen of a herbaceous host *sensu stricto*, only that it may be differentiated from 'W' strains by its inability

to live within the woody environment. The pathogenic capability of each strain was not tested explicitly; trait designation for host type was informed based on published literature for each strain and epidemiological information provided by the international culture collections where many of these strains originate. The host species and trait designation for host type for all strains are given in TABLE 5.1. Gene occurrence was defined as a discrete binary trait designated either '1' for gene presence or '0' for gene absence for all 13,010 genes of the *P. syringae* pan genome.

Two traits each with two states produce four possible combinations and a maximum of eight transition rate parameters that link them, represented in FIG. 5.1. In the independent model, the rate of transition in the second trait (e.g., from 0 $\rightarrow$ 1, described by the rate parameter q_{12}) does not depend on the state of the first trait (H or W), and thus rate q_{12} is equal to q_{34}. Since this applies to all traits and all states, it holds that $q_{12} = q_{34}$, $q_{21} = q_{43}$, $q_{13} = q_{24}$ and $q_{31} = q_{42}$, thus giving the independent model a maximum of four rate parameters. The dependent (correlated evolution) model, however, places no constraints on rate variation and therefore has a maximum of eight parameters. The dependent model will be a better fit to the data when the observed distribution of states of both traits across the phylogeny implies that a number of rates constrained to be equal in the independent model are in fact different (Pagel, 1994; Barker and Pagel, 2005). The model does not directly use the number of taxa with a given trait/state, but rather the number of transitions between traits and states across all nodes in the phylogeny.

5.3.6 Hypothesis testing and null model

The goodness-of-fit of the dependent versus the independent model can be compared using the likelihood ratio test:

$$\mathrm{LR} = -2(\log_e(H_0) - \log_e(H_1)) \tag{5.1}$$

where H_0 is the likelihood of the independent model and H_1 is the likelihood of the dependent (correlated evolution) model (Pagel, 1994). A custom PERL script was

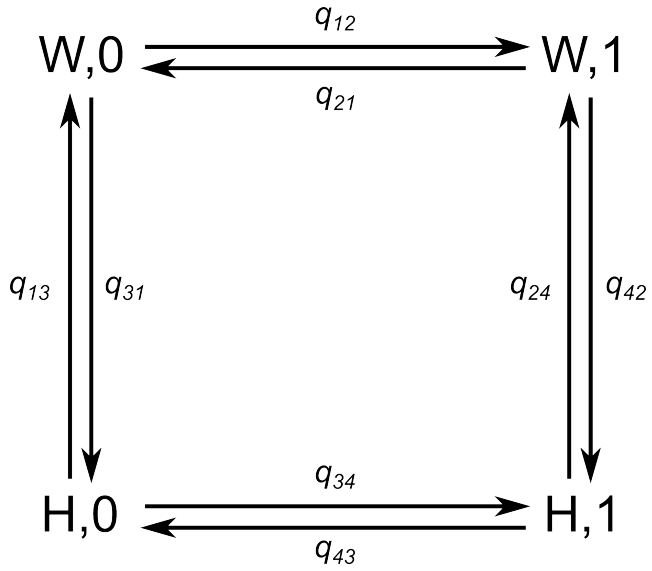


Figure 5.1: Transition rate parameters under the dependent model (H_1) of evolution for two traits, host-preference (H/W) and gene presence (1/0). 'H' and 'W' indicate herbaceous or woody host-preference while '1' and '0' indicate gene presence or absence, respectively. The independent model H_0 allows for the two traits to evolve independently on the phylogeny, such that the parameter q_{12}, which describes the rate of transition from $0 \rightarrow 1$ in trait 2, does not consider the background state (W or H) of trait 1. Thus, under H_0, q_{12} is equal to q_{34}, and similarly $q_{21} = q_{43}$, $q_{13} = q_{24}$ and $q_{31} = q_{42}$, resulting in a total of four q rate parameters. H_1 allows that each rate parameter between pairs of states may differ, and thus uses a total of eight parameters to describe the data. Thus, H_1 is favoured in cases where the constraints imposed under the null result in a poor fit to the observed data. Figure is adapted from Friedman and Barrett (2008).

written that ran both models and calculated the likelihood ratio statistic (LR) for each gene of the occurrence matrix. Genes that were present at either a very low or very high frequency among the 64 strains are unlikely to be contributing to a shared genetic basis for niche adaptation across the phylogeny as a whole; thus, genes that occurred in either fewer than five or greater than 59 genomes were excluded from further analysis. The pan genome for these strains showed 3,883 genes present in ≥ 5 and ≤ 59 genomes, thus our analysis resulted in a distribution of 3,883 LRs.

The magnitude of each LR describes the strength of the statistical evidence for correlated evolution, and follows a χ^2 distribution with degrees of freedom equalling the difference in the number of parameters between H_0 and H_1 (i.e., four; see FIG. 5.1). To account for the problem of multiple testing, however, we followed the approach of Barker and Pagel (2005), and constructed a null distribution of LRs that describes the random association between host preference and gene

presence. The construction of an empirically estimated null distribution negates the need for corrections such as Bonferroni adjustment as the null model should provide the expected distribution of LRs under the null hypothesis of no association between the two traits, given a large number of individual tests. This allows for the determination of critical *p*-value thresholds above which LR values should not be considered significant.

The null LR distribution was constructed by randomly permuting the gene occurrence data for each of the 3,883 tested genes a total of ten times, in each case calculating a new LR statistic. The phylogeny, the H/W trait designations for each taxon and the overall proportion of gene presence relative to absence (i.e., 1s to 0s) at each gene were held constant; only the occurrence profile was permuted (i.e., analogous to randomly shuffling the 1/0 column for each gene). A range of *p*-value thresholds were then derived directly from this null LR distribution.

5.4 Results

5.4.1 Core and pan genomics

Reannotation of 64 strains produced a total of 348,022 proteins, which were clustered into 11,200 initial groups by ORTHOMCL. Prior to correction, the *P. syringae* pan genome was estimated at 18,591 genes, consisting of 2,058 core and 16,533 flexible genes (of which 7,391 were lineage-specific). After applying the correction procedures outlined in CHAPTER 2, the size of the core genome was increased by 612 to 2,677 genes, or ∼48% of the total number of genes in an average *P. syringae* genome. Similarly, the number of inferred lineage-specific genes decreased from 7,391 to 2,384, a reduction of ∼68%. There is a significant positive relationship between the number of inferred errors in the initial core and lineage-specific estimates for a given genome sequence and its level of fragmentation (measured as number of contigs), highlighting the importance of sequence contiguity in draft genomes (FIG. 5.2). However, it is possible that these data are non-independent if there is a phylogenetic effect on either the number of contigs or the number of errors per genome; thus, we

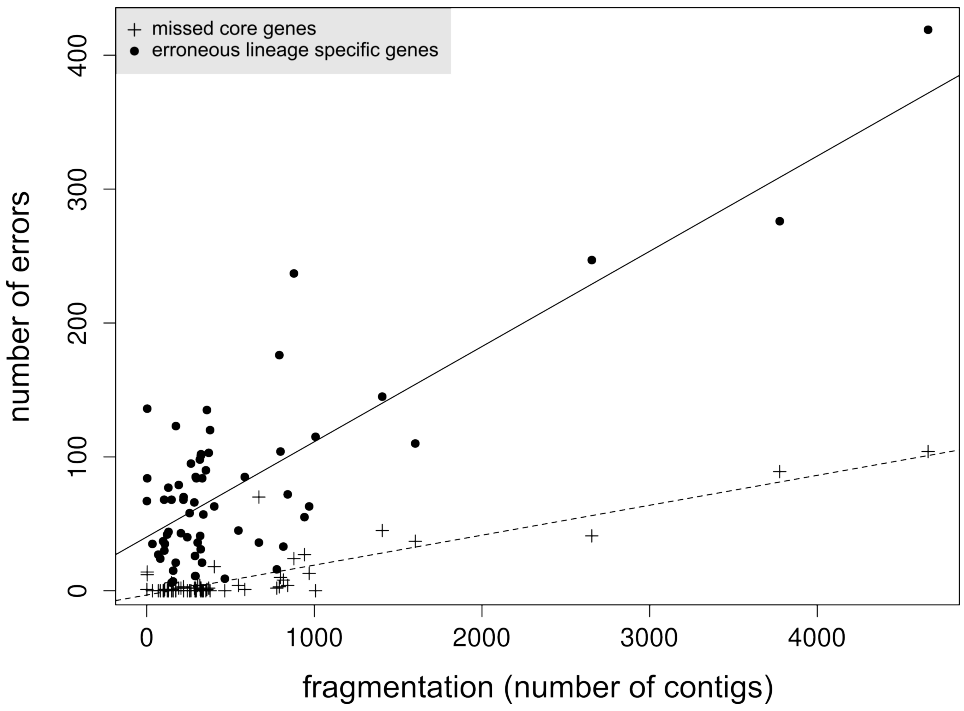


Figure 5.2: Relationship between the level of genome fragmentation (number of contigs) and inferred number of errors in draft genome sequences. Crosshairs represent missed core genes (t = 14.21, p < 2e−16; correlation coefficient = 0.76); filled circles represent genes erroneously inferred to be lineage-specific (t = 10.78, p = 8e−16; correlation coefficient = 0.65).

express caution when considering the significance of these results.

Gene accumulation curves for the core and pan genomes are shown in FIG. 5.3. This shows an estimated pan genome size of 13,010 genes and an estimated core of 2,677 when all 64 strains are considered. Surprisingly, this core estimate is about 80 genes larger than that estimated using 27 lineages of *P. syringae*, 20 of which are included in the present analysis (Nowell et al., 2014). Possible reasons for this anomalous result are discussed later. Pathovar *maculicola* str. ES4326 (*macu4326*), although often labelled as *P. syringae*, is thought to have been originally misidentified and is actually a strain of *Pseudomonas cannabina* (Baltrus et al., 2011). While *P. cannabina* is probably an outgroup to members of the loosely defined *P. syringae* 'species complex' (Parkinson et al., 2010), the exclusion of this strain from the core genome estimate increases its size only marginally, by 33 genes, to a value of 2,710.

The shape of the pan genome curve describes the rate at which new genes (i.e.,

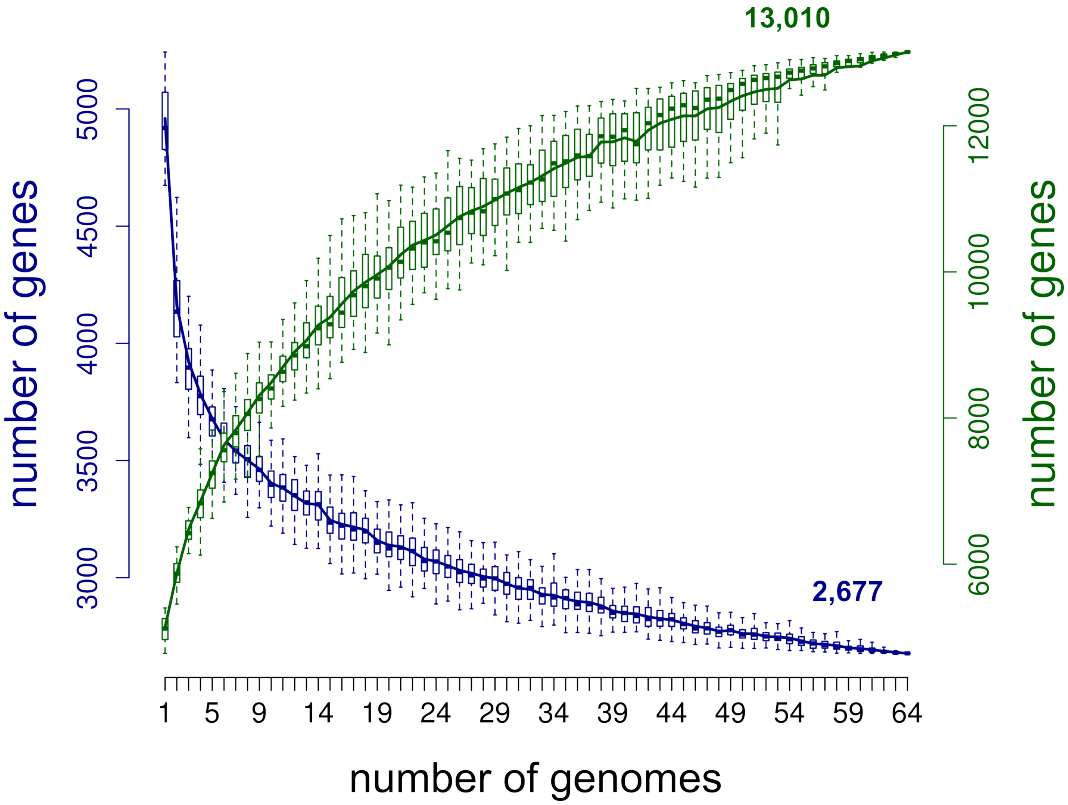


Figure 5.3: Gene accumulation curves for the core (blue scale on left-hand axis) and pan (green scale on right-hand axis) genomes of the 64 lineages of the *P. syringae* species complex used in this study. Note different scales for core and pan axes. The *x*-axis shows the number of genomes at each comparison point; box-and-whisker plots show the distribution of values for 100 random samples of genomes (with replacement).

previously unsampled) are discovered as a function of the number of genomes compared. An important property of this function is whether it approaches an upper limit, beyond which the addition of extra genomes would yield no genic novelty (Tettelin et al., 2005; Hogg et al., 2007; Rasko et al., 2008; Hiller et al., 2007). If the pan genome does tend to an asymptote, the pan genome is described as 'closed'; alternatively, there may be no limit to genic novelty, and the pan genome is 'open' (Tettelin et al., 2008; Medini et al., 2005). We ascertained this property for the *P. syringae* pan genome by fitting a power law function known as Heaps' law to the pan genome data. Heaps' law was originally applied to the rate at which a vocabulary (number of novel words) increases with respect to the number of texts considered, but has recently been applied to bacterial genome dynamics where

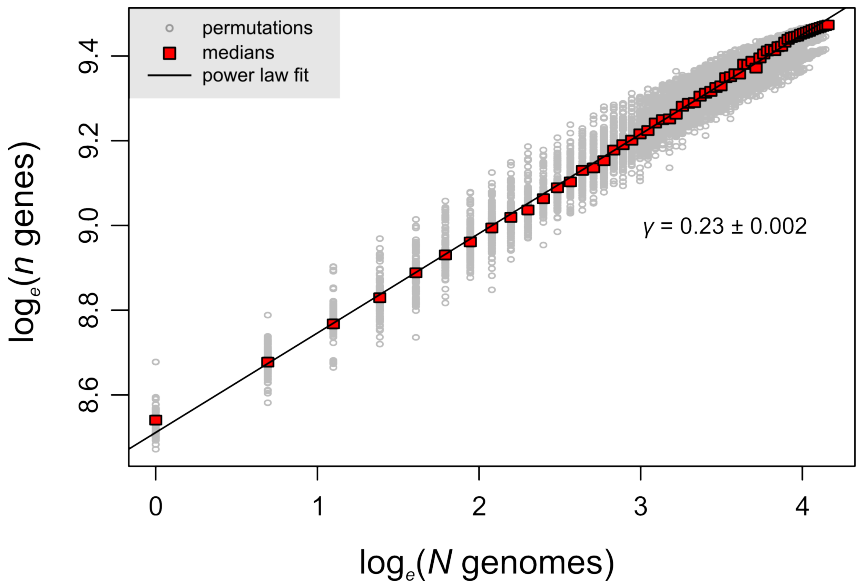


Figure 5.4: Heaps' law as applied to *P. syringae*. The *x*-axis shows the log number of genomes compared; *y*-axis shows log number of distinct genes in the pan genome. Gray circles are permuted values; red boxes show median values for each comparison point with a linear fit of the form: $\log_e(n) = \gamma \log_e(N) + \log_e(\alpha)$, equivalent to $n = \alpha N^{\gamma}$ on untransformed data. The value of the exponent γ is shown $\pm$ 95% confidence interval.

'words' are equivalent to genes and 'texts' are genomes (Heaps, 1978; Tettelin et al., 2008). As applied to pan genome analyses, Heaps' law states that the number n of distinct genes is proportional to a sub-linear power law of the number N of genomes compared, such that $n \sim N^{\gamma}$ (Tettelin et al., 2008). If $\gamma < 0$, the pan genome will reach a plateau and is closed; if $0 < \gamma < 1$, the pan genome conforms to Heaps' law and is 'unbounded', allowing for the discovery of unlimited genic novelty with respect to the number of genomes considered (Tettelin et al., 2008).

The least-squares fit of the power law $n = \alpha N^{\gamma}$ to median values of distinct genes is shown in FIG. 5.4. This shows $\gamma = 0.23 \pm 0.002$ (95% confidence interval), suggesting an open pan genome for these strains.

5.4.2 Phylogenetics

The core genome phylogeny for these strains was estimated from the 1.15 Mb concatenated nucleotide alignment of 2,086 1-to-1 orthologous genes using maximum likelihood, and is shown in FIG. 5.5. This shows the clear and well-supported

partitioning of these strains into three clusters, corresponding to phylogroups 1, 2 and 3 as per Sarkar and Guttman (2004) and others, with PG2 and PG3 as sister groups to the basal PG1.

A number of patterns are evident when scrutinising the distribution of strains across the phylogeny. Firstly, strains isolated from woody hosts are found in all three phylogroups. Pathovars *actinidiae* (pathogenic on kiwi), *theae* (tea), *morsprunorum* (cherry), *avellanae* (hazel) and *avii* (also cherry) are found within PG1; pv.'s *avellanae* (hazel), *syringae* (*Prunus* spp.) and *papulans* (crab apple) are members of PG2, while finally pv.'s *morsprunorum* (cherry), *rhapiolepedis* (yeddo hawthorn), *myricae* (bayberry), *eriobotryae* (loquat), *dendropanacis* (ivy tree), *savastanoi* (olive), *nerii* (oleander), *fraxini* (ash), *daphniphylli* (*Daphniphyllum* sp.), *castanae* (Japanese chestnut), *aesculi* (horse chestnut) *ulmi* (elm), *cerasicola* (cherry) and *broussonetiae* (paper mulberry) are all found within PG3. Thus, PG3 contains the majority of diversity for strains of woody hosts, and evolutionary relationships of woody-host strains within this phylogroup, including patterns of divergence among pv. *aesculi* strains, are discussed in more detail in CHAPTER 6.

Secondly, it is evident that woody-host strains are not monophyletic within any phylogroup. This indicates multiple transitions onto woody hosts have occurred both across the *P. syringae* species complex as a whole and within individual phylogroups. Assuming that the common ancestor of these lineages was not a pathogen of a woody-host species, the placement of strains across the phylogeny suggests at least seven transitions into the woody niche. Two major transitions are evident: one in PG3, which contains all of the PG3 strains with the exception of pv. *broussonetiae* (discussed in CHAPTER 6), and another in PG1, involving pv.'s *avellanae*, *morsprunorum*, *theae*, and *actinidiae*. The pathotype strain for pv. *morsprunorum*, M302278, is found within the PG1 cluster. Interestingly, this strain is considered inappropriate as a pathotype strain as it is thought to be non-pathogenic on cherry (Gardan et al., 1999; Bull et al., 2010). If *mors302278* is indeed avirulent, its placement represents a transition back to a 'non-woody' phenotype within this cluster.

In addition, the direction of some transitions is not clear from this phylogeny.

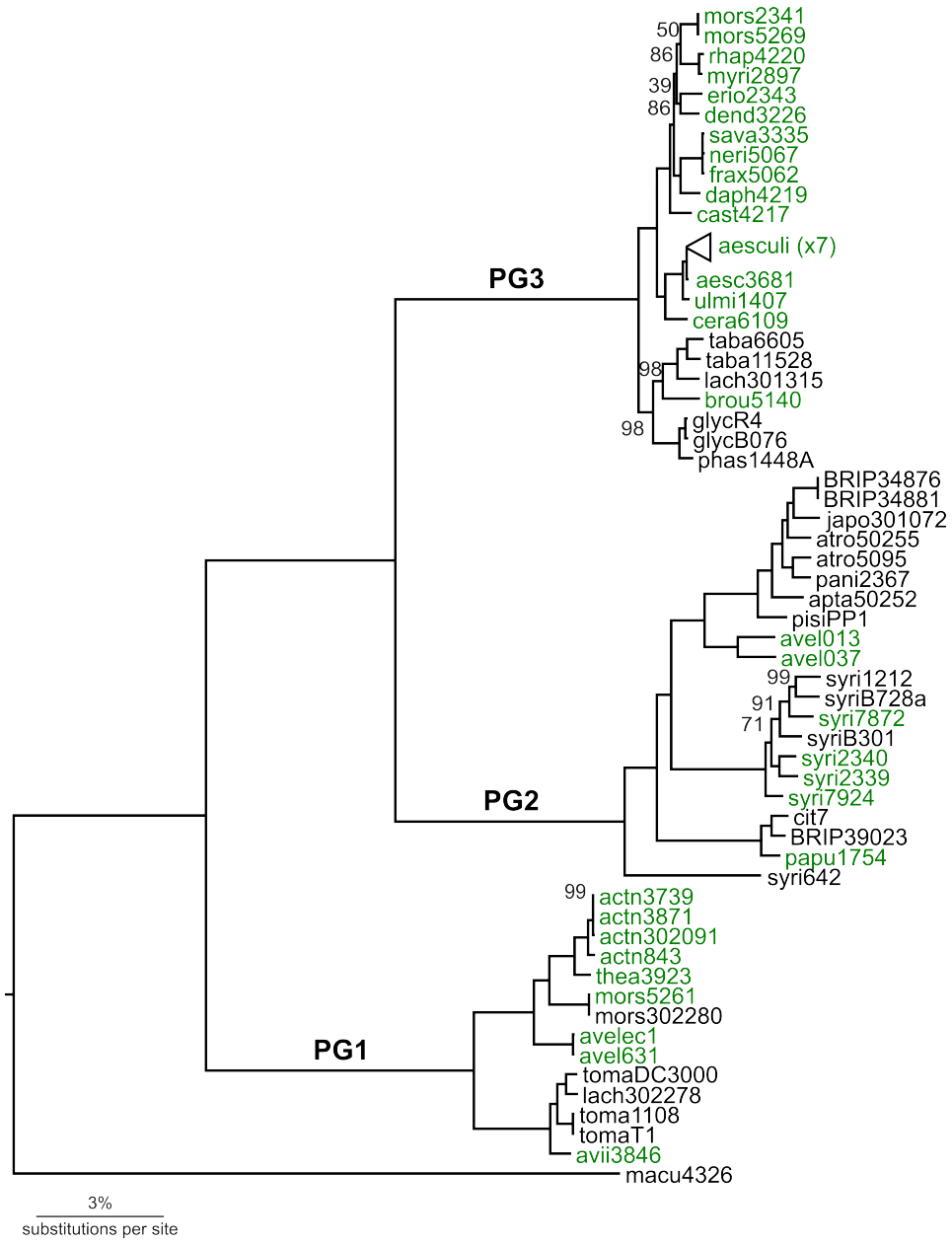


Figure 5.5: Maximium likelihood phylogeny for 64 strains of the *P. syringae* species complex. All nodes have 100% bootstrap support except where indicated. Taxon names in green are strains isolated from woody hosts (designated 'W' in Table 5.1). Major phylogroups PG1, PG2 and PG3 are shown on branches. Tree is rooted with *P. syringae* pv. *maculicola* str. ES4326 (*macu4326*); scale bar indicates 0.03 substitutions per site.

For example, the seven pv. *syringae* strains within PG2 cluster to form a reasonably diverse yet monophyletic clade; three are designated as non-woody-host strains while four are pathogens of various stonefruit (*Prunus*) spp. While the exact numbers are an artefact of phylogenetic sampling within this clade, we note that the two branches at the base of this cluster lead to three woody-host strains, suggesting that the ancestral pv. *syringae* lineage within PG2 may have also been a pathogen of a *Prunus* host and that the switches onto herbaceous hosts occurred later.

Other transitions to the woody niche are likely to have occurred on the terminal branch to pv. *broussonetiae* in PG3, in the common ancestor to the hazel pathogens *avel013* and *avel037* in PG2, the terminal branch to pv. *papulans* (pathogenic on crab apple) also in PG2, and finally, along the terminal branch to the cherry pathogenic strain *avii3846* in PG1.

Thirdly, we observe at least three separate cases where strains designated within the same pathovar cluster in different phylogroups. Strains of pv. *avellanae* are found in PG1 and PG2, while pv.'s *morsprunorum* and *lachrymans* both have representatives within PG1 and PG3. Previously, divergent lineages of pv. *morsprunorum* have been classified into two distinct 'races', with race 1 corresponding to strains within PG3 and race 2 corresponding to strains within PG1 (Gardan et al., 1999; Ménard et al., 2003). Thus, the strains *mors5269* and *mors2341* are the first reported genome sequences for race 1 *morsprunorum*, and confirm the polyphyly of strains within this pathovar. High levels of divergence within the pv.'s *avellanae* and *lachrymans* have been discussed elsewhere (Baltrus et al., 2011; O'Brien et al., 2012).

5.4.3 Homologous recombination

The level of homologous recombination among core-genome sequences was also assessed for a subset of 50 *P. syringae* strains, using the program CLONALFRAME (Didelot and Falush, 2007). Homologous recombination between divergent lineages may act to obfuscate patterns of vertical inheritance (Gogarten et al., 2002; Doolittle and Papke, 2006), but can also contribute to rates of horizontal transfer when foreign DNA is 'carried' into a recipient genome as the result of recombination at

Table 5.2: Rates of recombination and mutation.

Run	ρ_S	θ_S	r	m	r/m
1	1.832e−2	4.981e−2	618,757	184,759	3.349 (3.327, 3.370)
2	1.802e−2	4.981e−2	613,953	184,759	3.323 (3.303, 3.342)
3	1.789e−2	4.981e−2	609,150	184,759	3.297 (3.282, 3.313)

The parameters ρ_S and θ_S are defined as the recombination rate and mutation rate per site, respectively; r and m are the number of substitutions due to recombination or mutation, respectively; r/m is thus the relative contribution of recombination versus mutation to total nucleotide diversity. Values in parenthesis are lower and upper 95% confidence bounds.

homologous flanking regions (discussed in Polz et al., 2013).

Output parameters from the three CLONALFRAME analyses are given in TABLE 5.2. Convergence among runs was assessed using the Gelman-Rubin convergence diagnostic, which did not exceed the critical threshold of 1.2 for any parameter among runs (Didelot and Falush, 2007). The per-site mutation rate (θ_S) was estimated at approximately 5%, compared to a per-site recombination rate (ρ_S) of about 1.8%. The number of nucleotide substitutions present in the core genome alignment due to recombination (r) was estimated to be of the order of 6.1e+5, compared to the contribution from mutation (m) of 1.8e+5. The perhaps counter-intuitive result that $r > m$ despite $\rho_S < \theta_S$ arises because multiple substitutions may result from a single recombination event, such that a lower per-site recombination rate does not preclude this process contributing the greater proportion of nucleotide differences to the overall observed divergence, relative to that of mutation.

Thus, the ratio of the number of substitutions due to recombination to the number of substitutions due to mutation (r/m) ranged from 3.30 to 3.35, suggesting that recombination contributes approximately three times as much to nucleotide diversity than does mutation, for these strains of *P. syringae*.

5.4.4 Type III secretion system effector repertoires

The distribution of type III secretion system effectors (T3SEs) across these lineages is given in FIG. 5.4.4. Overall, patterns of T3SE occurrence are highly variable across these strains, reflective of the rapid turnover of the T3SE repertoire in *P. syringae* (Lindeberg et al., 2006). It should be noted that strain *syri642* is known to lack the

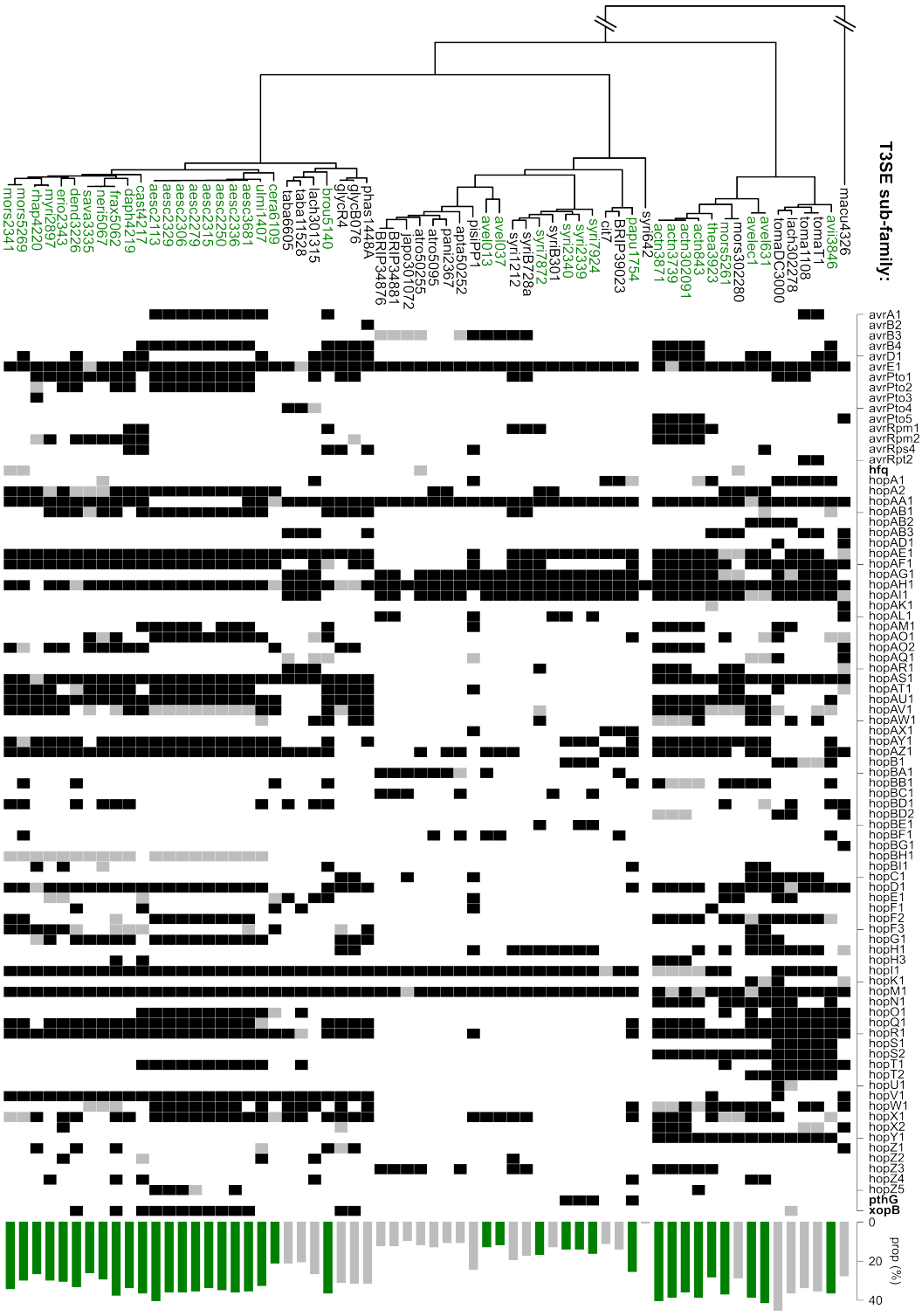
canonical type III secretion system (T3SS) (Clarke et al., 2010), and shows homology to only a single T3SE (HopAH1) in this analysis. Our search schema was relatively conservative, counting a match as as a minimum of 80% identity over at least 80% query length (the '80/80' rule), and differentiated between full-length and partial hits.

Partial hits may be biologically relevant, such as in cases where an effector CDS is chimæric through recombination or has been interrupted by the insertion of a mobile element, or may be an artefact of genome data in draft format, where a T3SE gene has been truncated by a gap. Effector truncation is known to occur frequently in *P. syringae* through domain rearrangements and recombination (Pitman et al., 2005; Ma et al., 2006); nonetheless the majority of hits shown in FIG. 5.4.4 are full-length and are therefore likely to encode functional T3SEs.

Counting full-length hits as 1 and partial hits as 0.5, the strain encoding the most T3SEs is the complete sequence of the tomato pathogen *tomaDC3000*, with a repertoire of 40.5 effectors. Second largest is *avel631* with 37 effectors, followed by *actn302091* and *actn3871* each with 36, then *actn843* and *actn3739* each with 34.5. Conversely, the strains encoding the fewest effectors (excepting *syr642*) are *japo301072* with nine, *pani2367* and *atro5095* each with 10, and *atro50252*, *avel037* and *cit7* each with 11. In agreement with previous analyses (e.g., Baltrus et al., 2011), strains within PG2 show a paucity of T3SEs relative to the other two phylogroups. The average repertoire size for PG2 is 13 effectors, compared to an average of 33 and 27 for PG1 and PG3 respectively.

When *syri642* is considered, there are no T3SEs ubiquitous across the 64 lineages. Discounting this anomalous strain, however, shows AvrE1 and HopM1 to be present

Figure 5.6 *(following page)*: Distribution of type III secretion system effectors (T3SEs) across the *P. syringae* species complex. Black boxes indicate presence (at least 80% identity over at least 80% query length); grey boxes indicate possible truncation (at least 80% identity over less than 80% query length); presence supersedes truncation in cases where both are found. T3SE names are given on right—genes designated with the same letter are within the same family, numbers indicate sub-families. Strains are arranged as per the core genome phylogeny (*cf.* Fig. 5.5). T3SEs in bold are from outwith the *P. syringae* species complex. Bar chart shows the proportion of T3SEs (out of a total of 88) present for each strain. Green taxa and bars represent strains pathogenic on woody hosts as before.


T3SE sub-family:
macu4326
avii3846
tomaT1
toma1108
lach302278
tomaDC3000
avel631
avelec1
mors302280
mors5261
thea3923
actn843
actn302091
actn3739
actn3871
syri642
BRIP39023
cit7
Papu1754
syri1212
syriB728a
syri7872
syriB301
syri2340
syri2339
syri7924
avel037
avel013
pisiPP1
apta50252
Pan23367
atro5095
atro50255
japo3011072
BRIP34881
BRIP34876
phas1448A
glycB0076
glycR4
lach301315
brou5140
taba6605
taba71528
cera6109
ulmi1407
aesc3681
aesc2336
aesc2250
aesc2315
aesc2279
aesc2306
aesc2329
aesc2173
cast4217
daph4219
frax5062
neri5067
sava3335
dend3226
erio2343
myri2897
rhap4220
mors5269
mors2341
avrA1
avrB2
avrB3
avrB4
avrD1
avrE1
avrPto1
avrPto2
avrPto3
avrPto4
avrPto5
avrRpm1
avrRpm2
avrRps4
avrRpt2
hfq
hopA1
hopA2
hopAA1
hopAB1
hopAB2
hopAB3
hopAD1
hopAE1
hopAF1
hopAG1
hopAH1
hopAI1
hopAK1
hopAL1
hopAM1
hopAO1
hopAO2
hopAQ1
hopAR1
hopAS1
hopAT1
hopAU1
hopAV1
hopAW1
hopAX1
hopAY1
hopAZ1
hopB1
hopBA1
hopBB1
hopBC1
hopBD1
hopBD2
hopBE1
hopBF1
hopBG1
hopBH1
hopBI1
hopC1
hopD1
hopE1
hopF1
hopF2
hopF3
hopG1
hopH1
hopH3
hopI1
hopK1
hopM1
hopN1
hopO1
hopQ1
hopR1
hopS1
hopS2
hopT1
hopT2
hopU1
hopV1
hopW1
hopX1
hopX2
hopY1
hopZ1
hopZ2
hopZ3
hopZ4
hopZ5
pthG
xopB
0
20
40
prop (%)

in all of the remaining lineages in at least partial form. HopAH1 is also present at a very high frequency, and is apparently missing only from *rhap4220*. A more detailed search of the *rhap4220* assembly showed this gene to be completely absent at any level of identity. HopAA1 and HopAE1 occur in over 80% strains, while HopAZ1, HopAF1, HopR1, HopAS1, HopD1, HopX1, HopAY1, HopQ1, HopAU1, HopAI1 and HopAG1 all occur in at least half of these lineages. Thus, the majority of T3SEs (73 of the total of 89 presented here) are found in less than half of the strains, highlighting the patchy distribution of these genes among genomes.

At the other end of the spectrum, four effectors were found to be lineage-specific at the 80/80 threshold used here. AvrB2 has been shown to occur variably among strains of pv.'s *phaseolicola* and *glycinea* (Qi et al., 2011; Baltrus et al., 2012); here, we find this effector only in *phas1448A*. AvrPto3 is recorded within GenBank as being sequenced from pv. *myricae* M302941 (accession number GQ227720.1; Nguyen et al., 2010), but was found only in *rhap4220* in our analysis. Finally, HopBG1 is present only in the outgroup strain *macu4326*.

A total of seven T3SEs, including AvrPto3, HopBE1, HopBH1, PthG, HopBI1, HopH3 and HopZ5, were encoded exclusively by woody host strains, although only the latter three of these were found across more than one phylogroup. The average number of effectors encoded by woody-host strains (n = 38) is 27, while the average encoded by non-woody-host strains (n = 26) is lower, at 19 effectors. However, the phylogenetic non-independence of these data makes the significance of this difference difficult to ascertain.

We also found evidence for the presence of a number of effectors with homologs from species outwith the *P. syringae* species complex (shown in bold in FIG. 5.4.4). Hfq is a small (102 amino acid) RNA chaperone protein with a role in the regulation of the T3SS encoded by *Escherichia coli* O157:H7 str. EDL933 (Kendall et al., 2011). Partial homologs of *E. coli hfq* are found in the strains *mors302280*, *atro50255*, *mors2341* and *mors5269*; however, further investigation showed that a more diverged copy (that falls marginally short of the 80/80 criterion) of this gene is present in all *P. syringae* lineages.

PthG is an 488 residue effector from *Pantoea agglomerans* pv. *gypsophilae* (Ezra et al., 2004) with good hits found only in the PG2 strains *syri2339*, *syri2340*, *syri7924* and *papu1754* (but not in *syriB728a* or *syri7872*). PthG has no identity to any T3SEs already described for *P. syringae*. XopB is encoded by *Xanthomonas campestris* pv. *vesicatoria* (Schulze et al., 2012), and was present in *glycR4*, *glycB076*, *cast4217*, *dend3226*, *frax5062* and the eight *aesculi* strains, all from PG3, while a partial hit was also detected in the distantly related PG1 strain *lach302278*. However, hits at just under the 80/80 rule are found in at least 28 other strains of *P. syringae*, while XopB also shows a high level of similarity (88% over 84% coverage) to HopD1. Thus, the evolutionary origins of XopB homologs within *P. syringae* is not clear.

5.4.5 Phytotoxins and other non-secreted virulence factors

The distribution of a selected range of known and suggested virulence factors is shown in FIG. 5.7. Overall, patterns of presence and absence are clearer than that shown by genes of the T3SS, and often correspond to phylogeny.

The β-ketoadipate pathway is found only in the large cluster of woody-host strains within PG3, and the monophyletic clade consisting of pv.'s *actinidiae*, *theae* and *morsprunorum* in PG1. The significance of this operon in the adaptation of *P. syringae* to the woody niche is covered in detail in CHAPTER 6. Another gene cluster also with putative involvement in adaptation to the woody environment, the protocatechuate 4,5-deoxygenase pathway, was found to be unique to the eight pv. *aesculi* strains.

The coronatine phytotoxin, which is a mimic of the plant hormone jasmonic acid and acts to promote entry of the bacterium into a plant host by causing stomata to open (Melotto et al., 2006), was found variably in PG1 and PG3 strains. Genes involved in the production of the recently described mangotoxin, isolated from *P. syringae* pv. *syringae* strains pathogenic on mango trees (Arrebola et al., 2003, 2011b), were encoded ubiquitously across strains.

Phaseolotoxin inhibits the normal functioning of the urea cycle *in planta*, causing a deficiency in intracellular arginine that results in the characteristic chlorosis of

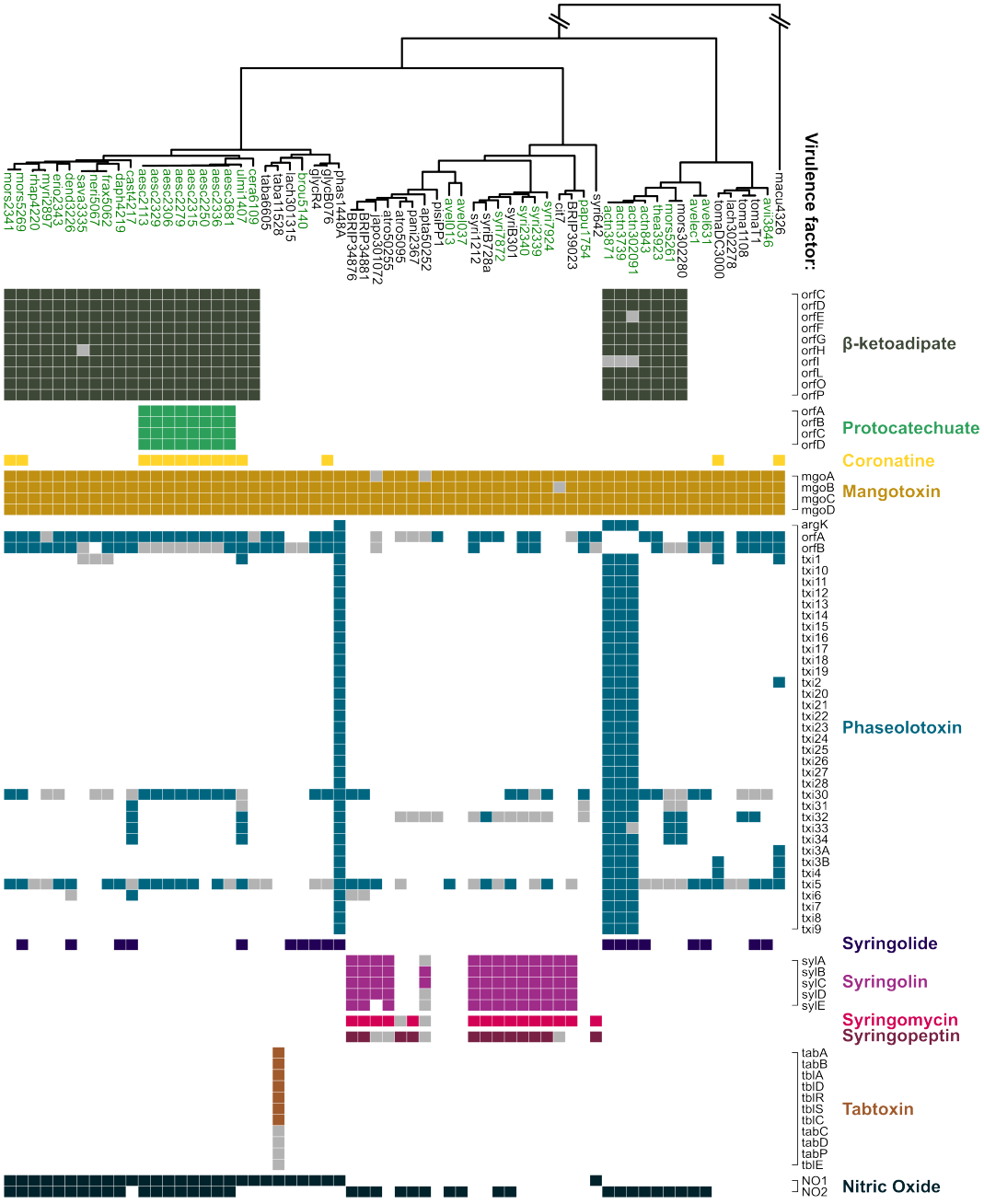


Figure 5.7: Distribution of known and suggested virulence factors across the *P. syringae* species complex. Genes within operons are arranged into coloured blocks; grey boxes indicate the presence of a partial hit for that gene (minimum 80% identity but less than 80% query coverage). Strains are arranged as per the core genome phylogeny as before.

infected regions (Bender et al., 1999). The 37 genes of the phaseolotoxin gene cluster (also known as the *argK-tox* cluster) is encoded in its entirety only by *phas1448* (PG3); however the majority of the pathway (35/37 genes) is also found in *actn3739*, *actn3871* and *actn302091* (but not *actn843*) in PG1. Other strains show variable occurrence for both full-length and partial hits to a smaller numbers of genes within the operon. The gene cluster encoding tabtoxin, responsible for the characteristic chlorotic 'halo' typical of wildfire disease of tobacco caused by *P. syringae* pv. *tabaci* (Uchytil and Durbin, 1980; Arrebola et al., 2011b), was found uniquely in *taba11528* and was not recovered in *taba6605* or any other strains.

The toxins syringomycin and syringopeptin contribute to virulence by forming pores in plant cell membranes, resulting in the free movement of ions across the membrane, the disruption of the cell's electrical potential and cell death (Takemoto et al., 1991; Scholz-Schroeder et al., 2001), whereas the proteasome inhibitor syrygolin toxin acts to suppress plant stomatal defences allowing entry of the bacterium into plant tissues (Schellenberg et al., 2010). Syringolin was found in all PG2 strains bar *atro5095*, *pisiPP1*, *avel013*, *avel037*, *papu1754* and *syri642* within this group. Syringomycin and syringopeptin were more prevalent, and were absent from only the strains *pisiPP1*, *avel013*, *avel037* and *papu1754*.

Syringolide, encoded by the *avrD* gene, is a water-soluble glycolipid that elicits a hypersensitive (HR) immune response in soybean cultivars carrying the *Rpg4* resistance gene, and thus plays a role in 'gene-for-gene' host–pathogen interactions (Keen et al., 1990; Ji et al., 1998). The *avrD* gene is encoded variably across a number of PG1 and PG3 strains but was not found in any PG2 strains.

Finally, the occurrence of two nitric oxide metabolism genes was also recorded. NO metabolism may play a role in bacterial defence against plant immune systems (Delledonne et al., 1998; Green et al., 2009). The first was ubiquitous across PG3 strains, and was also found in strain *syri642* in PG2. The second was more common, and was present in a number of strains from across the three phylogroups.

5.4.6 Distribution of LR statistics

In addition to elucidating the presence and absence of specific genes as above, we also looked for the correlated evolution between transitions into the woody niche and gene presence among all genes of the flexible genome that were present in 5–59 genomes (i.e., that were at neither very low nor very high frequency among lineages).

Our approach uses the program BAYESTRAITS (Pagel, 1994; Pagel and Meade, 2006) to calculate the likelihood of two hierarchically nested models, one in which transitions between host type (herbaceous and woody; H/W) occurs independently with respect to gene presence, and another, in which the transition into the woody niche co-occurs with the acquisition of a given gene (or indeed the converse, in which the transition onto a herbaceous host co-occurs with gene absence). Statistical support for correlated evolution was assessed via a likelihood ratio test, and the distribution of likelihood ratio statistics (LRs) for 3,883 genes of the flexible genome was compared to an empirically estimated null distribution of LRs over the same genes to ascertain significance (FIG. 5.8).

The shape of the LR distribution compared to the null suggests an excess of genes with an LR that exceeds the p-value critical thresholds (i.e., a shifting of the LR distribution to the right with respect to the null, overlayed in blue). Numbers of genes exceeding each threshold are shown in TABLE 5.3, along with the expected number of Type I (false positive) errors under the null. This shows that, of the 3,883 tested sites of the *P. syringae* flexible genome, 899 have an LR value that exceeds the $p \leq 0.05$ significance level. The expected number of false positives is 194; thus, there are 705 genes—corresponding to ∼18% of tested genes, or ∼7% the total *P. syringae* flexible genome—over and above that expected under the null hypothesis that show a significant association with strains that exist in the woody niche.

5.4.7 Occurrence profile

To gain a better understanding of what these data look like with respect to the phylogeny for these strains, we plotted the occurrence profile of all genes exceeding the p

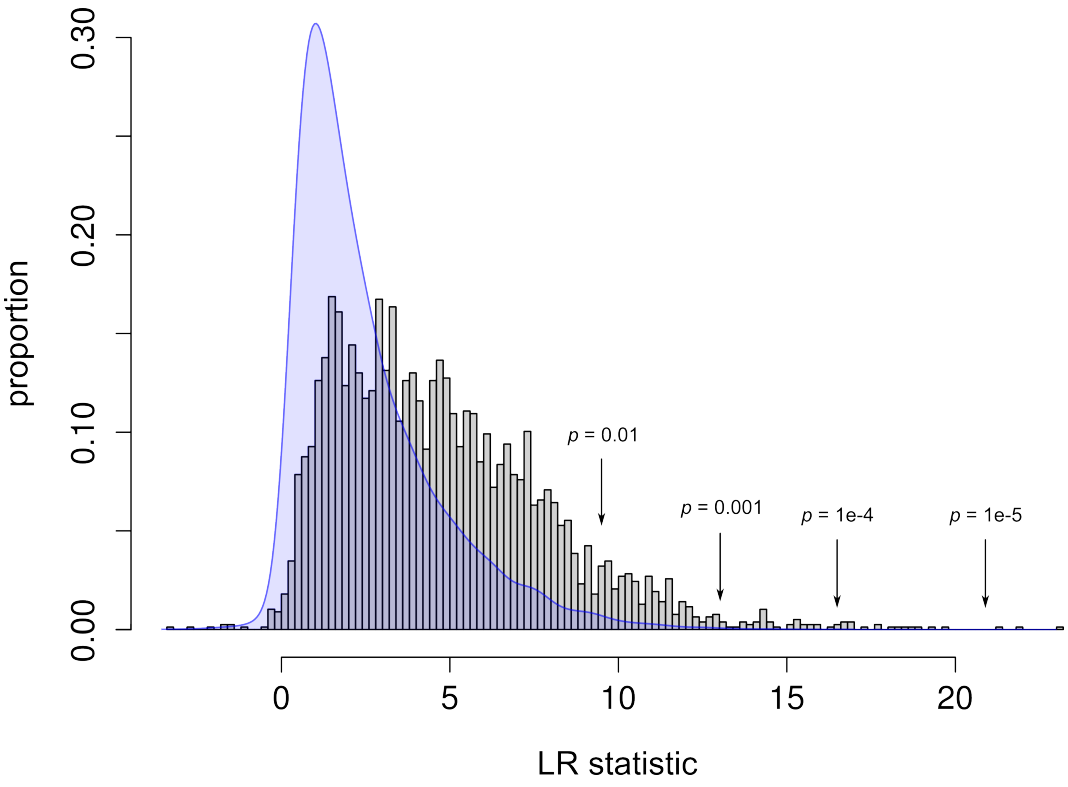


Figure 5.8: Gray bars show the distribution of likelihood ratio statistics for 3,883 tested sites of the *P. syringae* flexible genome measuring the support for correlated evolution between host-preference and gene presence. The estimated null is overlayed in blue; arrows indicate *p*-value thresholds based on the null.

Table 5.3: Number of genes at significance thresholds.

***P*-value**	**LR value**	**Expected[a]**	**Observed[b]**	**Proportion (%)**	
				Tested[c]	**Flexible[d]**
0.05	6.78	194	899	18.15	6.82
0.01	9.50	39	296	6.62	2.49
0.001	13.02	4	59	1.42	0.53
0.0001	16.50	<1	20	0.51	0.19
0.00001	20.89	$\ll 1$	3	0.08	0.03

[a]Expected number of Type I (false positive) errors under null; [b]Observed number of genes with an LR exceeding *p*-value threshold; [c]Proportion, out of 3,883 tested genes, of observed minus expected; [d]Proportion, out of total flexible genome (10,333 genes), of observed minus expected.

≤ 0.001 significance threshold, corresponding to the top 59 genes most-significantly associated with presence in woody-host strains (of which we expect approximately four to be Type I errors; see TABLE 5.3) (FIG. 5.4.7).

A number of patterns are immediately evident. Firstly, there are no genes that show a perfect correspondence with strains adapted to the woody niche. In addition, there were relatively few cases where gene presence was recorded only in woody-host strains—the majority of genes (47 of the 59) were also present in at least one herbaceous strain. For example, the most significant gene (LR = 23.1) was found in *avii3846*, *avel631*, *avel013*, *mors302278*, *mors5261*, *thea3923*, *actn3739*, *actn3871* (all PG1), *papu1754*, *syri2339*, *syri2340*, *syri7924* (PG2), and *phas1448A*, *brou5140*, *ulmi1407*, *cast4217*, *daph4219*, *frax5062*, *neri5067*, *dend3226*, *erio2343*, *rhap4220*, *mors5269* and *mors2341* (PG3). Thus, even considering the H/W ambiguity regarding the non-pathogenic *mors302278*, this gene is also found in strain *phas1448A*, the causal agent of halo blight of the common bean (*Phaseolus vulgaris*).

However, this gene is present in strains within all seven of the likely transitions from H $\rightarrow$ W across the phylogeny (discussed in section 5.4.2), and is therefore associated with all seven independent transitions into the woody niche in *P. syringae*. Out of the top 10 most significant genes based on LR magnitude, this was the only gene found in all seven transitions. In rank order, these top 10 genes were found in strains within seven, four, five, four, one, four, five, four, five and five inferred transition events. In the case of the gene associated with only one transition into the woody niche, this gene was the only one from the top ten to be present exclusively in strains from woody hosts. Looking down the rows of occurrence profiles shown in FIG. 5.4.7, it is clear that the majority of genes are present in multiple independent transition events, resulting in an overall association between woody hosts (green

Figure 5.9 *(following page)*: Occurrence profile for 59 genes significantly associated with the woody niche. Shading shows significance level: dark, medium and light shading corresponding to $p \leq 0.00001$, $p \leq 0.0001$ and $p \leq 0.001$ respectively. Genes are ordered from left to right corresponding to the magnitude of the LR statistic (decreasing significance). Strains are arranged according to the core genome phylogeny as before. The bar chart shows the proportion of genes (out of 59) present for each strain. Green taxa and bars represent strains pathogenic on woody hosts as before.


macu4326
avii3846
tomaT1
toma1108
lach302278
tomaDC3000
avelec1
avel631
mors302280
mors5261
thea3923
actn8423
actn392091
actn3739
actn3871
syri642
BRIP39023
cit7
papu1754
syri7872
syriB728a
syri1212
syriB301
syri2340
syri2339
syri7924
avel013
avel037
pisiPP1
apta50252
pani2367
atro50955
atro50255
apoc301072
BRIP34881
BRIP34876
phas1448A
glycR4
glycB076
lach301315
taba11528
taba6605
brou5140
aesc3681
ulmi1407
cera6109
aesc2236
aesc2250
aesc2315
aesc2279
aesc2306
aesc2329
aesc2113
cast4217
daph4219
frax5062
neri5067
sava3335
dend3226
erio2343
myri2897
rhap4220
mors5269
mors2341
LR > 20.89 (p < 1e-5)
LR > 16.5 (p < 1e-4)
LR > 13.02 (p < 0.001)
gene LR ranking
1
5
10
15
20
25
30
35
40
45
50
55
59
prop (%)
0
50

taxa in the phylogeny) and gene presence (green boxes in the matrix). This is summarised by the proportion of genes present in woody versus non-woody strains, represented by the bar chart at the bottom of the figure. Overall, woody-host strains possess 33 of the 59 genes (56%), compared to about 18 (30%) in non-woody strains.

5.4.8 Functions of associated genes

The 59 genes significant at the 0.001 threshold were annotated using both BLAST and BLAST2GO to elucidate potential gene functions. Only one returned no hit when queried against the non-redundant protein database (*nr*). 18 genes (∼30%), including five of the top 10, were annotated as hypothetical proteins. These ranged in length from 51 to 229 amino acid residues, and only a handful showed BLAST matches outwith the *P. syringae* species complex; thus, we infer the majority are likely to be annotation errors. A further 10 genes (∼17%) were described as having functions related to either transposition or conjugal transfer.

Annotations for the remaining 30 genes are shown in TABLE 5.4. Two proteins, including the top-most significant site with an LR = 23.1, have similarity to known type III secretion effector proteins (HopAY1 and HopAO1), while six proteins are involved in the uptake, transport or utilisation of urea. Four genes were found to encode proteins with enzymatic functions in eight different metabolic pathways, as defined using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). Pathways included the metabolism of purines and pyrimidines, metabolism of the amino acids alanine, aspartate and glutamate and lipopolysaccharide biosynthesis. In addition, 4-oxalocrotonate tautomerase (gene #23) and muconate cycloisomerase (gene #26) both have roles in the degradation of a number of aromatic compounds, including benzene, toluene and xylene, which are constituents of extracts from wood such as pine tar.

Physical linkage among these 59 genes was also assessed, using the *myri2897* as a reference as this strain showed the highest proportion of gene presence among the 64 lineages. Although a small number of genes showed close physical proximity to one another in the *myri2897* genome, the only operon of note contained five

genes involved in the utilisation of urea, *amiR*, *urtA*, *urtB*, *urtE* and *ureG*, which was located near the termini of a 158 kb contig in the *myri2897* assembly along with a number of other genes with functions in the utilisation of this compound.

5.5 Discussion

We have shown previously that the genomes of *P. syringae* lineages experience a high level of permutation, such that many hundreds of genes may be gained and lost over an evolutionary time period equivalent to a 1% divergence between core proteins (CHAPTER 2; Nowell et al., 2014). High levels of horizontal gene transfer (HGT) and gene loss can facilitate bacterial adaptation when the acquisition of 'niche-transcending' genes allow a recipient lineage to adapt into a novel environment, followed by the subsequent loss of other genes that are no longer required (Lawrence, 1999; Ochman, 2001; Ochman et al., 2005). The extent of genome fluctuation in *P. syringae* implies that this process has played a major role during the evolution of these lineages.

Here, we have elucidated the contribution of HGT to the adaptation of *P. syringae* into a specific niche environment, that of the woody parts of the various plant species that serve as hosts for these bacteria. We investigate genome dynamics using data for 64 lineages of *P. syringae*, and corroborate previous findings of extensive recombination in both the core and flexible genome components of these strains. We reconstruct the most likely evolutionary history for *P. syringae* using a core-genome phylogenetics approach, and characterise the distribution of a number of virulence determinants known to be important for pathogenesis across this phylogeny. We then assess evidence for the correlated evolution between host type and gene presence for genes of the flexible genome, using a method that is able to account for the phylogenetic non-independence of these data. Below, we discuss our main findings within the context of other recent work on *P. syringae* evolution, and suggest an emergent model of adaptation for *P. syringae* in which HGT and gene loss play central roles.

Table 5.4: Annotations of genes at 0.001 threshold.

#[a]	LR[b]	Id (%)[c]	Cov (%)[d]	Description
1	23.1	97	79	Type III effector HopAY1
7	18.7	97	98	Urease accessory protein UreG
8	18.4	97	100	CopG family transcripitonal regulator
11	17.7	100	100	Urea ABC transporter permease UrtB
13	17.3	94	52	Glutamate/aspartate transport ATP-binding protein
16	16.9	27	48	Choline dehydrogenase
17	16.8	84	83	Alpha/beta hydrolase family protein
18	16.8	100	100	ATP-dependent DNA helicase RecG
22	16.3	97	97	Acetyltransferase
23	16.0	97	100	4-oxalocrotonate tautomerase
26	15.8	97	99	Muconate cycloisomerase
27	15.6	95	99	Putative transcriptional regulator
29	15.4	99	100	Histidinol phosphatase
31	15.3	19	100	Glycoside hydrolase
35	14.7	87	93	N-acetyltransferase GCN5
36	14.6	99	100	Urea ABC transporter substrate binding protein UrtA
37	14.6	98	100	Urea uptake/utilization operon regulator AmiR
38	14.5	99	98	Nitrogen assimilation control protein
40	14.3	100	100	HPr kinase
42	14.3	99	100	Macrolide efflux protein
43	14.3	92	100	Methyl-accepting chemotaxis protein
44	14.2	99	100	Urea ABC transporter ATPase protein UrtE
45	14.2	92	99	Type III effector HopAO1
48	14.2	100	100	Anaerobically-induced outer membrane porin OprE precursor
49	14.1	99	100	Lipid A biosynthesis, N-terminal
50	14.0	97	100	Calcium-binding protein
51	13.9	100	98	AraC family transcriptional regulator
56	13.2	62	97	Sugar-binding protein
57	13.2	100	100	Urea ABC transporter permease UrtB[e]

Proteins annotated as 'hypothetical', mobilome-related or with no hit are omitted; [a]Number corresponds to rank in Fig. 5.4.7, determined by descending LR magnitude; [b]Likelihood ratio statistic value; [c]Percent identity; [d]Percent query coverage; [e]Divergent from #11.

5.5.1 Genome dynamics in *P. syringae*: another look

We estimate the core genome for 64 lineages of *P. syringae* to be 2,677 genes. Surprisingly, this is approximately 80 genes higher than that estimated in a previous analysis that used 27 genomes, 20 of which are also included here (Nowell et al., 2014). We consider possible explanations for this result. Firstly, these estimates are not based on entirely overlapping samples; that is, the smaller sample (27 genomes) is not represented entirely in the larger sample (64 genomes). Sampling effects lead to extensive variation around any given point estimate in gene accumulation analyses (such as that shown in FIG. 5.3), and can have a big effect on the final core value for a group of genomes. For example, the sample distribution for the comparison of 27 genomes in the current analysis shows a core estimate that ranges from a minimum of 2,849 to a maximum of 3,211—a difference of 362 genes.

In addition to this, however, it is likely that errors associated with whole-genome data have resulted in an underestimation of the core genome value in previous analyses (e.g., Baltrus et al., 2011; O'Brien et al., 2012; Nowell et al., 2014). Here, we show a positive relationship between the level of fragmentation in draft genome sequences and the number of inferred errors (missed core genes and erroneous lineage-specific genes; FIG. 5.2) contained within them. Missing or truncated genes due to annotation errors and/or fragmented sequence data can impact the determination of orthologous relationships among proteins, which may in turn lead to the underestimation of core and pan genome sizes. We note that a number of genomes used in prior studies were highly fragmented, sometimes consisting of thousands of contigs (e.g., pv. *pisi* str. 1704B). In the present study, the increased availability of genome data for *P. syringae* has allowed for the selective exclusion of poor-quality sequences, and thus we suggest that the current estimates of the core and pan genome sizes (2,677 and 13,010 genes respectively) are a more accurate representation of reality for the 64 lineages analysed here.

To investigate this further, we re-ran the gene accumulation analysis using a total of 63 genomes that did contain all 27 of those used in Nowell et al. (2014), and estimated the core genome of these strains to be 2,158 genes, about 440 genes lower

than that given in Nowell et al. (2014), and in line with what would be expected. Thus, we consider it likely that both sampling effects and errors associated with draft genome data are contributing to the unexpected difference in core genome estimates between these analyses.

We also investigate two other aspects of genome repertoire dynamics in *P. syringae* that have important implications for our understanding of how these bacteria evolve. Firstly, we infer that the core genome of *P. syringae* is highly recombinogenic, with homologous recombination estimated to contribute to nucleotide divergence threefold over that from point mutations ($r/m \approx 3$). Since our analysis contains strains from a diverse range of geographic locations and has not attempted to correct for potential population structure within the 50 genomes used, this is also likely to be an underestimate of the true value (Cohan, 2002; Vos and Didelot, 2009; Wiedenbeck and Cohan, 2011). These results are in contrast with a previous analysis regarding the extent of homologous recombination in *P. syringae* (Sarkar and Guttman, 2004). Using seven housekeeping genes sampled from 60 strains of *P. syringae*, Sarkar and Guttman (2004) estimate the relative diversifying effect of recombination versus mutation (r/m) to be approximately 0.25, and thus suggest a highly clonal population structure for *P. syringae* with little exchange of genetic material among lineages.

Rates of homologous recombination are known to vary greatly among bacterial species (e.g., Maynard Smith et al., 1993; Feil et al., 2001a; Vos and Didelot, 2009). A recent comparative analysis has shown the ratio r/m to range from 0.02 for the commensal species *Leptospira interrogans* (from the Spirochaetes) to 63.6 for the obligate pathogen *Flavobacterium psychrophilum* (from the Bacteroidetes), across a total of 49 different bacterial groups (Vos and Didelot, 2009). The authors also include an estimate for *P. syringae*, based on the sequence data of Sarkar and Guttman (2004), and calculate the relative effect of recombination to be 1.5 (1.1–2.0, 95% CI) for this species. Thus our estimate, which is based on the whole-genome alignment of 50 strains of *P. syringae*, is approximately twice the estimate of Vos and Didelot (2009) and an order of magnitude over that of Sarkar and Guttman (2004), and suggests a history of recombination within the core genomes of *P. syringae* lineages that is at

odds with the definition of *P. syringae* as a highly clonal species.

Secondly, we apply Heaps' law to the rate of gene accumulation in *P. syringae*, and show that the pan genome for this species appears to be unbounded, or 'open' (Medini et al., 2005; Tettelin et al., 2005, 2008). Thus, the addition of new *P. syringae* genomes into comparative analyses is likely to continue to yield novel genetic material that was previously unsampled in this species. The openness of the *P. syringae* pan genome implies that the scale of the HGT genepool available to *P. syringae* genomes may be very large indeed.

5.5.2 Evolution of *P. syringae* onto woody hosts

The *P. syringae* core genome phylogeny presented here consists of 64 lineages, 38 of which are pathogens of woody host species. Their placement on the phylogeny suggests that multiple transitions into the woody niche have occurred in the evolutionary history of this species, both across the phylogeny as a whole and within individual phylogroups (PGs). The majority of pathovars isolated from woody hosts cluster to form a well supported monophyletic clade within PG3, suggesting that a transition into the woody niche occurred in the ancestor to these lineages, followed by its diversification onto a range of different woody host species. Another major transition is evident in PG1, invlolving the pv.'s *avellanae*, *morsprunorum*, *theae* and *actinidiae*. Potential similarities in the genetic bases of these two transitions is presented in CHAPTER 6, and is discussed in detail therein.

We investigated the distribution of a number of virulence determinants across the core genome phylogeny. In particular, effector proteins of the type III secretion system (T3SS) are known to play a central role in pathogenicity and virulence in *P. syringae* (Sarkar et al., 2006; Cunnac et al., 2009; Lindeberg et al., 2012). We find a highly variable pattern of occurrence for type III secretion effectors (T3SEs) that is reflective of the high rates of gain and loss experienced by these genes (Ma et al., 2006; Lindeberg et al., 2009), and is in broad agreement with previous studies investigating the phylogenetic distribution of T3SE repertoires across *P. syringae* (e.g., Sarkar et al., 2006; Baltrus et al., 2011; Bartoli et al., 2014).

We detect seven T3SEs that are encoded exclusively by strains of woody hosts, but only three of these, *hopBI1*, *hopH3* and *hopZ5*, are found across multiple phylogroups. Of these, *hopBI1* is also found in the rice pathogen pv. *oryzae* strain 1_6, from phylogroup 4, which is not included in our comparative analysis. HopH3 is encoded by pv. *acinidiae* in PG1 and pv.'s *castanae* and *fraxini* in PG3, while HopZ5 is found in *actn843* in PG1 and pv. *aesculi* in PG3. The precise molecular functions of HopH3 and HopZ5 are unknown. However, the convergent acquisition of HopH3 and HopZ5 by divergent lineages is suggestive of a role in adaptation onto woody hosts, although further work would be required to confirm this. We also detected the presence of a T3SE that has not been previously observed in *P. syringae* and is not similar to any other T3SEs encoded by this species. This putatively novel T3SE was found in the PG2 strains *syri2339*, *syri2340*, *syri7924* and *papu1754*, and shares a high level of sequence identity (92% amino acid similarity) to the PthG effector from the gall-forming bacterium *Pantoea agglomerans* pv. *gypsophilae*, a pathogen of *Gypsophila paniculata* L. (from the carnation family) (Ezra et al., 2000, 2004). Although the exact function of the PthG homolog in *P. syringae* remains to be elucidated, its presence in these strains highlights the role of HGT in the expansion of T3SE repertoires, and illustrates the ecological context in which such transfer events are likely to occur.

The distribution of a range of non-secreted virulence determinants across the 64 *P. syringae* strains was also in general agreement with the literature (e.g., Hwang et al., 2005; Arrebola et al., 2011a; Baltrus et al., 2011). Of significant note, however, was the occurrence profile for genes of the catechol branch of the β-ketoadipate pathway. This operon has a function in the degradation of plant-derived aromatic compounds such as lignin derivatives, and has been suggested as a potentially important adaptation of pv. *savastanoi* onto olive (Rodríguez-Palenzuela et al., 2010), pv. *aesculi* onto the horse chestnut (Green et al., 2010) and pv. *actinidiae* onto kiwi (Mazzaglia et al., 2012; Bartoli et al., 2014). We show this operon to be present exclusively in strains representing the two major transitions of *P. syringae* lineages onto woody hosts in both PG1 and PG3. Again, we offer an in-depth analysis of the potential importance of this operon in the adaptation of these strains to

the woody niche in CHAPTER 6. However, it is interesting to note the absence of these genes from the PG1 hazelnut pathogens *avelec1* and *avel631*. These strains form a monophyletic cluster with strains *mors5261*, *mors302278*, *thea3923* and the four *actinidiae* strains within PG1, implying that the β-ketoadipate pathway was either lost from the ancestral *avellanae* lineage or gained after the divergence of this pathovar from the others.

5.5.3 HGT and niche adaptation in *P. syringae*

Using a method that accounted for the potentially confounding effects of phylogeny, we looked for the correlated evolution between strains that were isolated from woody host species and gene presence among the *P. syringae* flexible genome. Our results revealed a substantial proportion of the flexible genome ($\sim$7%, or about 700 genes) showed a significant association with strains that exist in the woody environment. This suggests that, for a certain fraction of the flexible genome at least, patterns of gain and loss are neither random nor strictly inherited (i.e., phylogenetically congruent); rather, they follow associations based on the ecological characteristics of these lineages—namely, the ability or otherwise to exist in the woody niche. An interpretation of this result is that strains inhabiting similar ecological niches require the same, or similar, sets of functions that are encoded by the same, or similar, sets of genes, in order to proliferate. Given the malleability and rapid turnover of *P. syringae* genomes, this suggests a convergent 'tailoring', achieved via the processes of HGT and gene deletion, of the flexible genomes of *P. syringae* lineages that is determined within the ecological context of the environment in which they live.

We have made a number of assumptions during the course of our analysis that are worth discussing here. Firstly, the degree to which these strains may be accurately partitioned based on a simple woody versus non-woody (i.e., herbaceous) host-type distinction is not clear. It was not practically feasible to individually assay these strains for their ability to cause disease in the woody parts of their respective host species, and thus we defined host type primarily using the published literature

and information associated with each strain. While the literature is extensive for many of these strains, the precise pathogenic capabilities for some was nonetheless difficult to ascertain. Moreover, although a high degree of host-specificity is a common feature for many *P. syringae* lineages, it is possible that some strains exhibit niche overlap at the H/W distinction used here, and would be able to inhabit both niches if given the opportunity. In these cases, associations between gene presence and the niche to which such a strain has been allocated are unlikely to be driven by the H/W distinction used here. Niche overlap has in fact been demonstrated for a number of *P. syringae* lineages isolated from environmental (i.e., non-plant) habitats, that were shown to be capable of causing bacterial canker of kiwi when inoculated onto this plant (Bartoli et al., 2014).

One obvious corollary of our result is that genes that show a significant association with a given niche do so because they confer a selective advantage when expressed in that particular environment. However, analysis of the functional properties of the proteins encoded by the most significant of these genes did not provide a particularly clear or obvious picture as to how they may be functionally contributing to adaptation to the woody niche. We note the presence of a number of genes involved in the utilisation of urea, suggesting an important role for the acquisition of nitrogen in these bacteria. In addition, two enzymes (muconate cycloisomerase and 4-oxalocrotonate tautomerase) have roles in the degradation of wood-derived compounds such as xylene and toluene. A muconate cycloisomerase is encoded within the β-ketoadipate pathway discussed previously, and in addition to the PG1 and PG3 woody host clades discussed above, is also found in the PG2 strains *avel013*, *avel037*, *BRIP34876*, *BRIP34881* and *japo301072*. However, in these latter strains this gene is not part of the same β-ketoadipate operon that is found in pathovars such as *aesculi* (Green et al., 2010) and *actinidiae* (Mazzaglia et al., 2012). The fact that this gene is found in the two PG2 hazelnut pathogens, *avel013* and *avel037*, is likely to have increased its significance relative to other members of the β-ketoadipate operon.

Despite these limitations, which are likely to obfuscate patterns of association

between gene presence and host-type and therefore drive down the level of significance, we observe a clear signal of association from these data at the genome-wide level. When these strains are defined by the fairly broad ecological distinction of woody versus non-woody host types, the occurrence profile of certain genes is dictated not by phylogeny, but by ecology. This finding implicates the flexible genome, and the evolutionary forces that influence it, as being central in the evolution of these bacteria to their niche habitats.

5.5.4 An emerging model of *P. syringae* evolution

P. syringae has been shown to be ubiquitous in a range of environmental (i.e., non-plant) habitats, and that these populations are both larger and more diverse than those living as plant-pathogens (Morris et al., 2008, 2010; Monteil et al., 2012, 2013). Moreover, strains isolated from environmental habitats are capable plant pathogens, often expressing a similar degree of virulence but a broader host range than strains isolated from the host plant itself (Monteil et al., 2013; Bartoli et al., 2014). For example, Bartoli et al. (2014) screened environmental strains of *P. syringae* that were genetically similar to pv. *actinidiae*, the causal agent of bacterial canker in kiwi, for the presence of the β-ketoadipate operon involved in the degradation of plant-derived aromatic compounds, and showed that lineages encoding this operon were in fact capable of causing bacterial canker on kiwi, despite having been isolated from non-agricultural habitats such as snow pack and water (Bartoli et al., 2014). The converse has also been shown to be possible: for example, cells of pv. *aesculi*, the causal agent of bleeding canker disease on horse chestnut, were found to be viable on horse chestnut after almost a year in sterile soil (Laue et al., 2014).

These results have lead to the suggestion that *P. syringae*, far from being a clonal, endemic plant pathogen as has been previously suggested (e.g., Sarkar and Guttman, 2004), is highly recombinogenic, has a population structure more similar to epidemic bacteria such as *Neisseria meningitidis*, and is likely to have spent the majority of its evolutionary history in environmental habitats (Morris et al., 2010; Monteil et al., 2013). Under this hypothesis, highly virulent plant-pathogenic lineages arise

from a highly diverse and freely recombining background population existing in environmental habitats; successful pathogens of monocultured crop plants then may expand clonally and disseminate globally (Monteil et al., 2013).

Overall, our findings agree with the model put forward by Monteil et al. (2013). Using 64 whole-genome sequences from across the *P. syringae* species complex, we show that recombination has been the predominant diversifying force among core genes in these bacteria, and that the scope for horizontal import of novel genes encoding novel functions is essentially unbounded. In addition, we find evidence that a subset of genes of the flexible genome show a significant association with an ecological niche that has been independently invaded by divergent lineages of *P. syringae*. We suggest that high levels of gene turnover have led to a plasticity of the flexible genome that allows for its tailoring to suit a range of different niche habitats, including both plant and non-plant environments. In this way, successful (i.e., virulent) crop pathogens arise when a lineage comes into contact with a plant species in which its given repertoire of genes and alleles confers to it a selective advantage when expressed on that particular host species.

Crucially, we propose high rates of HGT and recombination, in conjunction with the depth and diversity of the HGT genepool, may be sufficient to be able to transfer certain niche-transcending genes, such as those involved in the ability of *P. syringae* to proliferate within the woody tissues of woody host plants, across phylogenetic boundaries and into other lineages that may subsequently utilise them to invade a similar niche in a different host. In this way, this work highlights the direct importance of HGT in the evolution of plant-pathogenic lineages of *P. syringae*.

Chapter 6

Evolutionary adaptations of *Pseudomonas syringae* pv. *aesculi* to the European horse chestnut

6.1 Abstract

Pseudomonas syringae pathovar *aesculi* (*Pae*) is the causal agent of bleeding canker disease of the European horse chestnut (*Aesculus hippocastanum*). Recent analysis of the *Pae* genome has shed light on its evolutionary history, and highlighted a number of genes that may be important in the ability of this bacterium to proliferate in the woody parts of its host. The genome sequences of six additional *Pae* strains, sampled from across northern Europe, have been presented in CHAPTER 4. Here, we analyse these sequences within the context of their phylogenetic position within the *P. syringae* species complex to gain further insight into the genetic and genomic basis of adaptation in *Pae*. We apply a comparative genomics approach within a phylogenetic framework to elucidate the history of gene acquisition along the lineage leading to the *Pae* pathovar, and scrutinise these genes for their potential role in the adaptation of *P. syringae* to the woody niche. As well as highlighting a number of genes of potential further interest, our analyses reveal a genome rearrangement

that is exclusively shared by distantly related *P. syringae* strains that are pathogenic on different woody hosts, and suggest a role for recombination in the convergent evolution of these lineages. In addition, we find very little nucleotide diversity between European strains of *Pae*, suggesting a single and recent introduction of the bacterium into the continent. We estimate the divergence among *Pae* strains to be about 1×10^{-6}, and show that this level of divergence is not compatible with a recent study that infers a log-linear relationship between the rate of nucleotide substitution and the time since the last common ancestor across different groups of bacteria.

6.2 Introduction

Bleeding canker disease of the European horse chestnut (*Aesculus hippocastanum*) was first reported in the UK around the turn of the millennium (Webber et al., 2008), and has subsequently become epidemic over much of northern Europe. The causal agent of bleeding canker was found to be a lineage within the *Pseudomonas syringae* species complex, named *P. syringae* pv. *aesculi* (*Pae*) (Bultreys et al., 2008; Schmidt et al., 2008; Webber et al., 2008). The disease attacks the woody parts of the tree, proliferating within the cortex, cambium and phloem (Green et al., 2009; Steele et al., 2010), and is thus unusual in that there are very few described diseases of woody tissues that are caused by bacteria (Green et al., 2010).

Analysis of the genome sequence of *Pae* has shed light on a number of aspects of its evolution onto the horse chestnut (Green et al., 2010). Firstly, the level of divergence among three *Pae* strains isolated from locations across the UK was found to be extremely low, indicating a recent introduction of this bacterium into the UK. Secondly, *Pae* was shown to cluster within phylogroup 3 (PG3) of the *P. syringae* species complex, and is closely related to another strain of *Pae* that was isolated from the leaves of an Indian horse chestnut (*Aesculus indica*) approximately 30 years ago. Furthermore, both European and Indian strains of *Pae*, denoted E-*Pae* and I-*Pae* respectively, clustered with a number of other *P. syringae* pathovars pathogenic on woody host species to form a monophyletic clade within PG3, suggesting that there

may have been a single transition into the woody niche in the evolutionary history of PG3 *P. syringae*. Green et al. (2010) also identified a number of genes putatively involved in the adaptation of *Pae* onto the woody parts of its tree host, including genes involved in the degradation of plant-derived aromatic compounds that have been discussed previously (e.g., CHAPTER 5), the sequestration of iron, and the metabolism of antimicrobial toxin nitric oxide.

In this chapter, we review our understanding of the adaptation of *Pae* onto the horse chestnut in light of the much larger genome sequence resource that is now available. Firstly, the phylogenetic position of *Pae* is considered in the context of other pathogens of woody hosts from across the *P. syringae* species complex. Then, we use gene orthology data generated in CHAPTER 5 to reconstruct the history of gene gain and loss along the phylogenetic lineage leading to the *aesculi* pathovar, and assess the functions of acquired genes in relation to their potential role in pathogenesis in the woody niche. We report a range of genetic and genomic changes that have putative involvement in the adaptation of *P. syringae* to the woody niche, and that have occurred at various points in the evolutionary history of *Pae*; together, these findings provide a comprehensive overview of the adaptive history of the *Pae* lineage. Interestingly, we also show evidence of a recombination event between ancestral lineages of PG1 and PG3 pathovars of woody hosts, suggestive of a possible convergence of these strains on a common functional basis for proliferation within the woody niche.

In addition, we quantify the level of nucleotide divergence among six E-*Pae* genome sequences isolated from five locations across northern Europe, and estimate this value to be of the order of 1×10^{-6} substitutions per site. Using these data, we look for patterns of genomic differences that may indicate potential routes of transmission across Europe. We also infer the short-term rate of nucleotide substitution for E-*Pae*, and compare it to that of other bacterial groups with similar levels of divergence. Overall, this chapter offers a detailed investigation of the genetic and genomic basis for the evolution of *P. syringae* into the woody niche, and sheds further light on the adaptation of *Pae* onto the European horse chestnut.

6.3 Materials & Methods

6.3.1 Sequence data

The pathovar and strain designations, host species, location and year of isolation and reference information for all strains used in this chapter are shown in TABLE 6.1. The phylogenetic relationship of *P. syringae* pv. *mori* to pv. *aesculi* was of specific interest in this study; thus, pv. *mori* str. 301020 (Baltrus et al., 2011) is included in our analyses despite previous work showing the genome sequence for this strain to be of poor quality (see CHAPTER 2).

Genome sequencing, assembly and annotation for 19 previously unpublished strains of *P. syringae* is discussed in detail in CHAPTER 4. Publicly available whole-genome sequence data for the remaining *P. syringae* strains were downloaded from the National Center for Biotechnology Information (NCBI) on 4 June 2014.

6.3.2 Phylogenetics and reconstruction of gene gain and loss

Evolutionary relationships among *P. syringae* pathovars were assessed using a core genome concatenation approach outlined previously (see CHAPTER 2). The core genome of pv. *mori* str. 301020 was determined by querying the 2,086 core genes identified in CHAPTER 5 against the contig data for pv. *mori* using BLAST (*E*-value $\leq$ 1e$-$5) (Altschul et al., 1997). Hits for 2,061 genes were extracted and groups of core orthologs were realigned using T-COFFEE v11.0 (Notredame et al., 2000). Alignment columns containing gaps were excluded, resulting in a concatenated alignment $\sim$1.08 Mb in length. A maximum likelihood phylogeny was constructed using RAXML v7.8.2 (Stamatakis, 2006), implementing the general time reversible model of substitution with gamma distributed rate variation among sites (GTR + Γ). Reliability of the tree topology was assessed using 100 bootstrap replicates.

We used the gene occurrence data determined in CHAPTER 5 to infer the history of gene acquisition along the phylogenetic lineage leading to the *aesculi* pathovar. Gains and losses were mapped onto the core genome phylogeny using GLOOME (Cohen et al., 2010; Cohen and Pupko, 2010), and branch-specific gene gains with

Table 6.1: Strain information for *P. syringae* pathovars found in phylogroup 3.

Pathovar	Strain	Identifier[a]	Host	Location[b]	Year[c]	Reference[d]
aesculi	NRS 2113	aesc2113	*Aesculus hippocastanum* (European horse chestnut)	UK	2006	This study
aesculi	NRS 2250	aesc2250	*Aesculus hippocastanum* (horse chestnut)	UK	2008	Green et al. (2010)
aesculi	NRS 2279	aesc2279	*Aesculus hippocastanum* (horse chestnut)	Netherlands	2002	This study
aesculi	NRS 2306	aesc2306	*Aesculus hippocastanum* (horse chestnut)	Belgium	2010	This study
aesculi	NRS 2315	aesc2315	*Aesculus hippocastanum* (horse chestnut)	Netherlands	2006	This study
aesculi	NRS 2329	aesc2329	*Aesculus hippocastanum* (horse chestnut)	Czech Republic	2011	This study
aesculi	NRS 2336	aesc2336	*Aesculus hippocastanum* (horse chestnut)	Norway	2010	This study
aesculi	NCPPB 3681	aesc3681	*Aesculus indica* (Indian horse chestnut)	India	1979	Green et al. (2010)
brousonetiae	CFBP 5140	brou5140	*Broussonetia kazinoki* $\times$ *B. papyrifera* (paper mulberry)	Japan	1980	This study
castaneae	CFBP 4217	cast4217	*Castanea crenata* (Japanese chestnut)	Japan	1977	This study
cerasicola	CFBP 6109	cera6109	*Prunus yedoensis* (Yoshino cherry)	Japan	1995	This study
daphniphylli	CFBP 4219	daph4219	*Daphniphyllum teijsmanni* (—)	Japan	1981	This study
dendropanacis	CFBP 3226	dend3226	*Dendropanax trifidus* (ivy tree)	Japan	1979	This study
eriobotryae	CFBP 2343	erio2343	*Eriobotrya japonica* (loquat tree)	USA	1970	This study
fraxini	CFBP 5062	frax5062	*Fraxinus excelsior* (ash tree)	Netherlands	1978	This study
glycinea	B076	glycB076	*Glycine max* (soybean)	—	2007	Qi et al. (2011)
glycinea	race 4	glycR4	*Glycine max* (soybean)	—	1977	Qi et al. (2011)

Continued on next page.

Table 6.1: Continued from previous page.

Pathovar	Strain	Identifier	Host	Location	Year	Reference
lachrymans	MAFF 301315	lach301315	*Cucumis sativus* (cucumber)	Japan	1975	Baltrus et al. (2011)
mori	MAFF 301020	mori301020	*Morus alba* (white mulberry)	Japan	1966	Baltrus et al. (2011)
morsprunorum	NRS 2341	mors2341	*Prunus cerasus* (wild cherry)	Hungary	1988	This study
morsprunorum	HRI-W 5269	mors5269	*Prunus cerasus* (wild cherry)	UK	1990	This study
myricae	CFBP 2897	myri2897	*Myricae rubra* (Chinese bayberry)	Japan	1978	This study
nerii	CFBP 5067	neri5067	*Nerium oleander* (oleander)	Spain	1979	This study
phaseolicola	1448A	phas1488A	*Phaseolus vulgaris* (common bean)	—	1985	Joardar et al. (2005)
rhaphiolepidis	CFBP 4220	rhap4220	*Rhaphiolepis umbellata* (yeddo hawthorn)	Japan	1980	This study
savastanoi	NCPPB 3335	sava3335	*Olea europaea* (olive tree)	France	1984	Rodríguez-Palenzuela et al. (2010)
tabaci	ATCC 11528	taba11528	*Nicotiana tabacum* (tobacco)	—	1905	Studholme et al. (2009)
tabaci	6605	taba6605	*Nicotiana tabacum* (tobacco)	Japan	1967	Studholme, D.J., et al. (unpublished)
ulmi	CFBP 1407	ulmi1407	*Ulmus* spp. (elm tree)	Yugoslavia	1958	This study

NB: This table shows the same data as Table 5.1 (Chapter 5), reproduced here for ease of reference. Source abbreviations: ATCC, American Type Culture Collection, Virginia, USA; CFBP, Collection Française de Bactéries Phytopathogènes, Beaucouzé, France; HRI-W, Horticultural Research International, Warwick, UK; MAFF, Ministry of Agriculture, Forestry and Fisheries, Ibaraki, Japan; NCPPB, National Collection of Plant Pathogenic Bacteria, York, UK; NRS, Northern Research Station, Midlothian, UK. [a]Unique identifier used in this study (if required); [b]Location of isolation (if known); [c]Year of original isolation (if known); [d]Genome sequence reference.

a probability of gain > 0.8 were extracted using custom PERL scripts. Sequences were functionally annotated at the protein level using BLAST and/or BLAST2GO v2.7.2 (Conesa et al., 2005). Nucleotide data for individual genes were aligned using GENEIOUS v5.6.2 (Biomatters, available from `http://geneious.com`), and phylogenies were constructed using PHYML v3.0 (Guindon and Gascuel, 2003; Guindon et al., 2010), using the GTR + Γ model of nucleotide substitution with 100 bootstrap replicates.

Where applicable, nonsynonymous (K_A) and synonymous (K_S) rates of nucleotide substitution were estimated using the method of Li (1993), implemented through the R package SEQINR v2.0-3 (Charif and Lobry, 2007) within the R v3.1.0 statistical environment (R Core Team, 2014).

6.3.3 Whole-genome alignment

Nucleotide variation within pv. *aesculi* was assessed using the six recently sequenced pv. *aesculi* strains: *aesc2113*, *aesc2279*, *aesc2306*, *aesc2315*, *aesc2329* and *aesc2336*. The strain *aesc2250* was excluded from the variant analysis as preliminary investigations suggested a higher rate of sequencing errors in this genome. The assembled genome of pv. *aesculi* str. 2336 (*aesc2336*) was chosen as the reference sequence, and sequenced readsets from the other five strains were aligned to it using BWA v0.7.9a (Li and Durbin, 2009). Duplicate reads, arising from PCR amplification during library preparation, were removed using RMDUP v0.1.19 from the SAMTOOLS package (Li et al., 2009). This excluded between 1.4 to 2.5% of mapped read pairs from the set of alignments.

Repeat regions, such as insertion sequence (IS) elements, can result in alignment errors during read mapping and subsequent false-positive variant calls. Repeat regions are often characterised by aligned reads with a low mapping quality (MQ) score, indicating that the alignment algorithm has found multiple possible locations for a given read. We characterised such regions in the reference genome by aligning the reads from *aesc2336* to its own assembly as above, and extracting all alignments with a probability of read mismatch $> 0.001\%$ (MQ ≤ 50) using FILTER v2.3.0

from the BAMTOOLS package (Barnett et al., 2011). At this threshold, regions of low mapping quality accounted for a total of approximately 370 kb of the 6.2 Mb reference assembly (∼6% total assembly length). As might be expected, repeat regions were almost always found at the termini of larger contigs, and frequently covered the majority of smaller contigs that are most likely comprised entirely of repeated sequence. Putatively variant sites that fell within a repeat region as defined above were excluded from all further analyses.

6.3.4 Variant discovery and annotation

We used the Genome Analysis Toolkit's (GATK) UNIFIEDGENOTYPER program v3.2-2 (McKenna et al., 2010) to perform single nucleotide polymorphism (SNP) and insertion/deletion (indel) discovery across all genome alignments simultaneously, setting the ploidy option to one. GATK standard hard filters were applied to the resultant variant calls, as per the GATK 'Best Practices' recommendations (DePristo et al., 2011; Auwera et al., 2013). In addition, we required each putative SNP or indel to have a per-sample alternative genotype count of at least five, an overall depth of no more than 300 (i.e., excluding sites with excessive coverage), and an alternative genotype proportion of at least 95% (calculated as the number of reads that agree with the variant genotype divided by the total read depth for any site called as alternative to the reference for a given sample).

Variants that passed the filtering criteria were extracted using a combination of VCFTOOLS v0.1.12a and VCFINTERSECT (Danecek et al., 2011; Garrison and Marth, 2012), and SNP annotation was performed using SNPEFF v3.6 (Cingolani et al., 2012). All variant sites and their position in the reference genome were visually checked using TABLET v1.14.04.10 (Milne et al., 2010) and/or the Integrative Genomics Browser (IGV) v2.3.34 (Thorvaldsdóttir et al., 2012).

6.4 Results

6.4.1 Phylogenetics

The core genome phylogeny for *P. syringae* is given in FIG. 6.1. As expected, the six recently sequenced E-*Pae* strains (*aesc2113*, *aesc2279*, *aesc2306*, *aesc2315*, *aesc2329* and *aesc2336*) cluster with the previously described pv. *aesculi* str. 2250 (*aesc2250*) within phylogroup 3 (PG3) (Green et al., 2010); thus, we focus on the evolutionary relationships among strains within PG3.

The phylogeny shows the majority (13 out of a total of 14) of PG3 pathovars from woody hosts clustering within a single, monophyletic clade with 100% bootstrap support (denoted 'woody clade' in FIG. 6.1). Within this, two similarly well-supported sister clades are evident. The first clade, denoted 'savastanoi', contains 10 pathovars isolated from a range of different host species, including pv.'s *rhaphiolepidis* (Yeddo hawthorn), *myricae* (bayberry), *eriobotryae* (loquat), *morsprunorum* (cherry), *dendropanacis* (ivy tree), *savastanoi* (olive), *nerii* (oleander), *fraxini* (ash), *daphniphylli* (Yuzuri-ha tree), and *castaneae* (Japanese sweet chestnut). The second clade, denoted 'aesculi', clusters both E-*Pae* and I-*Pae* with pv. *ulmi*, isolated from an elm tree in the former Yugoslavia in 1958, and pv. *cerasicola*, isolated from a Yoshino cherry in Japan in 1995. The patristic distance (i.e., sum of branch lengths) from the base of the E-*Pae* strains to pv. *ulmi* is 0.0075 nucleotide substitutions per site (subs/site), while pv. *cerasicola* is an additional 0.0081 subs/site diverged. Although the monophyly of E-*Pae* strains shows 100% bootstrap support, the branching patterns within the E-*Pae* clade show negligible support and thus cannot be resolved at this level.

An exception to the monophyly of woody-host strains within PG3 is the placement of pv. *broussonetiae*, isolated from shoot lesions on a paper mulberry (*Broussonetia kazinoki* $\times$ *B. papyrifera*) in Japan in 1980 (Takahashi et al., 1996). This pathovar is closely related to pv. *mori* str. 301020, the causal agent of bacterial leaf-spot of white mulberry (*Morus alba*) (Kazempour et al., 2006; Green et al., 2010). Both *broussonetiae* and *mori* cluster with a number of pathogens of herbaceous hosts within the 'phaseolicola' clade, which is a sister to the woody clade, such as pv.'s *phaseolicola*

and *glycinea* (both bean pathogens), *lachrymans* (cucumber) and *tabaci* (tobacco).

6.4.2 Gene gain along the *Pae* lineage

The evolutionary history of gene gain for *P. syringae* pv. *aesculi* is shown in FIG. 6.2. This quantifies the number of well-supported (i.e., with a probability of gain $\geq$ 0.8) gene acquisitions for each branch along the phylogenetic lineage leading to the *aesculi* pathovar.

At this threshold, gene gains are evident on all branches with the exception of the deep branch that separates phylogroup 1 from phylogroups 2 and 3. For the remaining branches, gained sequences are delimited into three simple categories, defined as 'functional' (if that protein was ascribed a function on the basis of BLAST and gene ontology analyses), 'hypothetical' (if the protein was annotated as a hypothetical protein, or had no hit) or 'phage-associated' (if the protein had some function relating to the transposition of phage sequences). Across all branches (475 gene gains), the proportions of functional, hypothetical and phage-associated proteins were 50, 44 and 6%, respectively. These proportions were roughly reflected for the majority of individual branches. For four out of the six branches analysed, the majority of gene gains were defined as functional. The two branches where this was not the case were the branch delimiting pv.'s *aesculi*, *ulmi* and *cerasicola* from the other woody-host pathovars, and the branch leading to the E-*Pae* strains—here, gains were predominated by hypothetical proteins. The proportion of phage-related genes was generally higher along shallower branches near to the tips of the phylogeny.

Branch-specific lists of gains were scrutinised for genes that may have a direct functional involvement in the adaptation of *Pae* (and other woody-host pathovars) to cause disease in the woody organs of their respective host species. Specific examples are discussed in detail below.

Degradation of D-xylose

Our reconstruction shows the gain of a gene encoding a 278 amino-acid protein with 100% identity to xylose isomerase (XI) at the root of all PG3 pathovars. XI

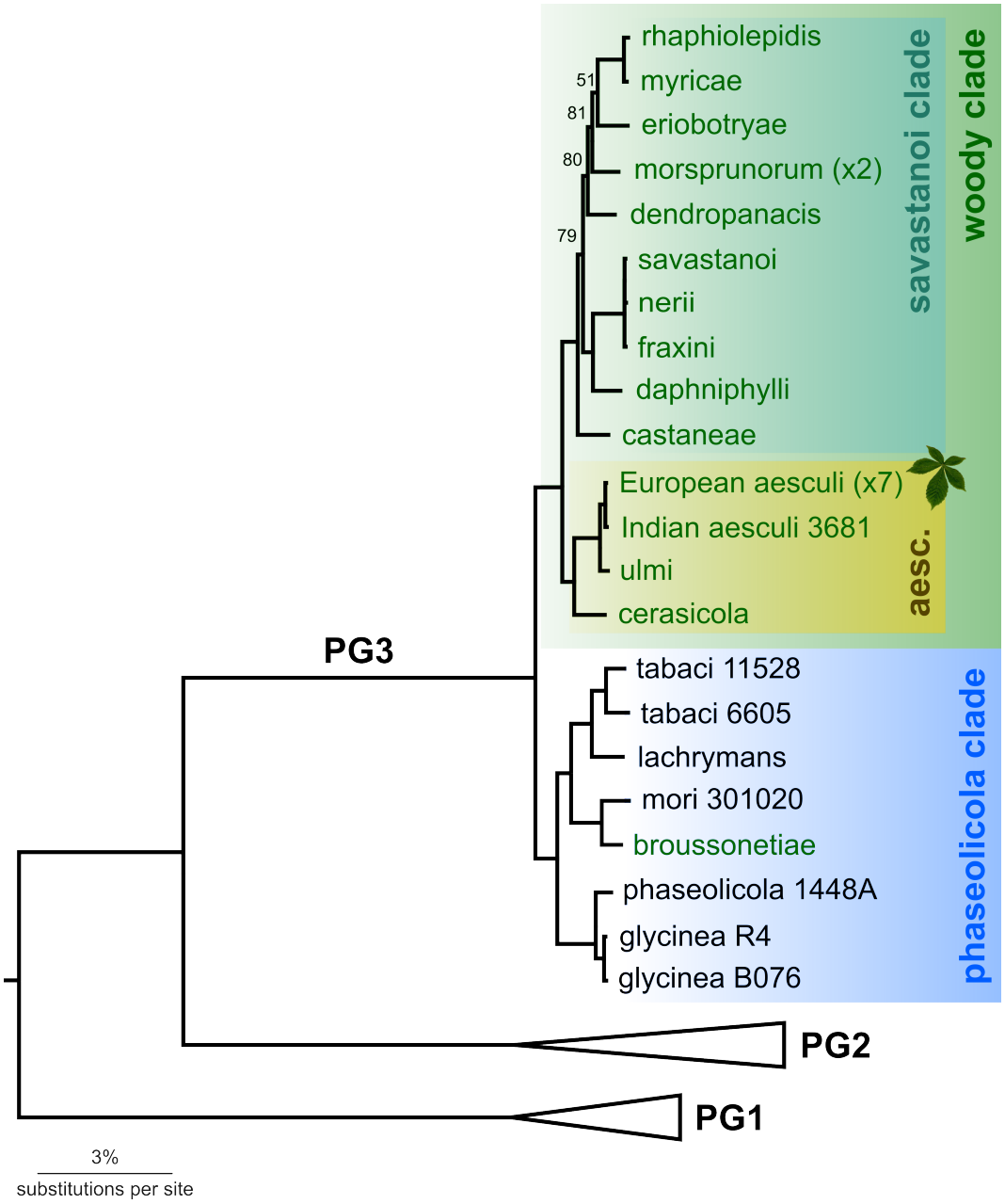


Figure 6.1: Evolutionary relationships among *P. syringae* pathovars based on the concatenated alignment of 2,061 core genes. PG1, PG2 and PG3 denote phylogroups 1, 2 and 3 respectively; details are shown for PG3 only. Taxa shown in green denote strains that cause disease in the woody parts of their respective host species. The placement of seven strains of European pv. *aesculi*, causal agent of horse chestnut bleeding canker, is marked with a leaf symbol. All bootstrap values are 100% except where indicated. Scale bar represents 0.03 nucleotide substitutions per site. Colours show clades discussed in the main text; 'aesc.' is an abbreviation of the aesculi clade. Tree is rooted with pv. *maculicola* str. ES4326 (not shown).

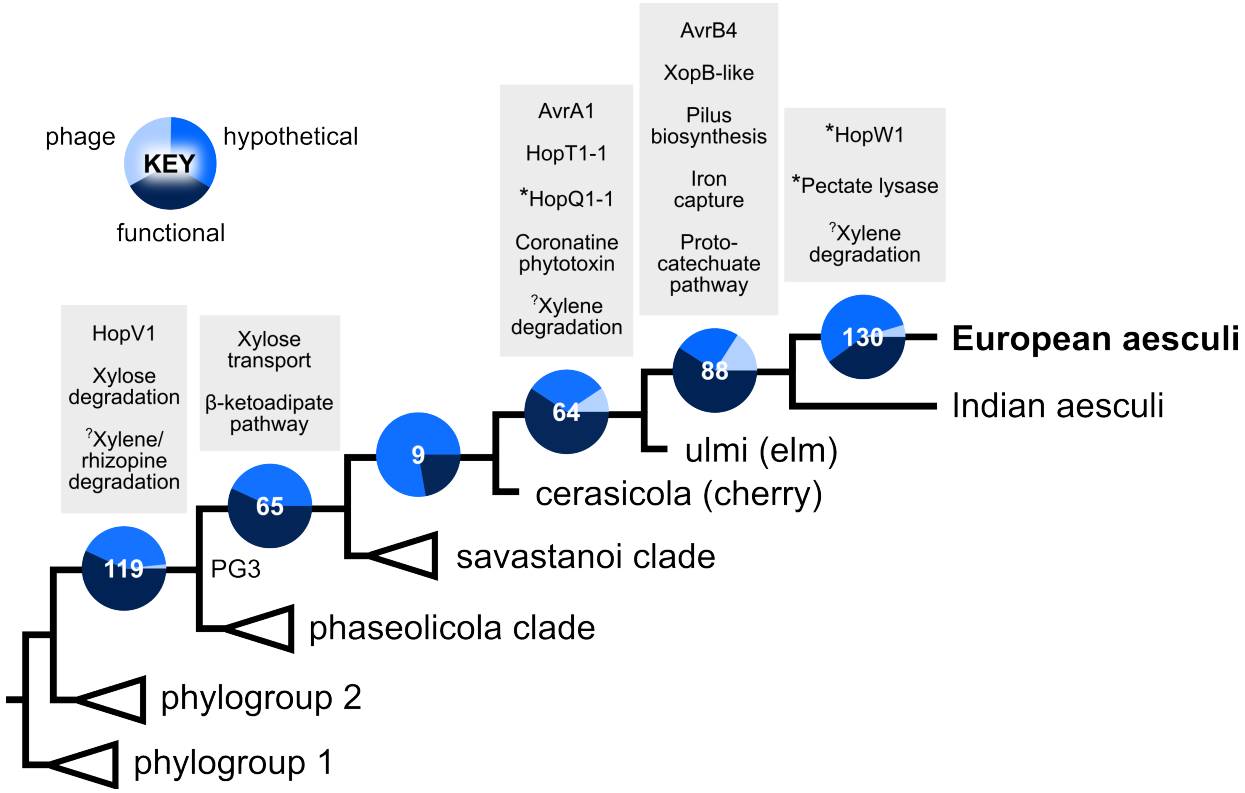


Figure 6.2: Gene gain along the phylogenetic lineage leading to *P. syringae* pv. *aesculi*. The number of well-supported gene gains is indicated for each branch, delimited into three basic categories: 'functional', 'hypothetical' or 'phage-related' (see key). Genes/functions of specific interest with respect to their putative involvement in the adaptation of *P. syringae* into the woody niche are written above each branch. Asterisks denote partial or truncated genes; question marks denote an incomplete pathway or where the gain-of-function is unclear. Topology is based on the core genome phylogeny; branch lengths not drawn to scale.

plays a role in the incorporation of D-xylose into the pentose phosphate pathway (PPP), thus allowing for the utilisation of these environmentally abundant sugars as a carbon source (Stephens et al., 2007; Bettiga et al., 2008). BLAST analysis showed the gene encoding PG3 XI (hereafter referred to as $xylA_2$) to be present in all PG3 strains, but also in the relatively distantly related pv.'s *actinidiae* and *theae* in PG1. The nearest BLAST match to this protein outwith the *P. syringae* species complex was to *Pseudomonas fluorescens* (83% amino acid identity over 100% length), with hits also identified in other members of the Pseudomonadaceae as well as a number of *Burkholderia* and *Rhizobium* spp. from outwith the Gammaproteobacteria.

Mapping of $xylA_2$ to the *aesc2336* assembly showed it to be located near the terminus of a ∼33 kb contig with no evidence of homology to known plasmid sequences. Moreover, this revealed that $xylA_2$ is not part of the xylose degradation operon that is ubiquitous across the *P. syringae* species complex. This operon also contains a xylose isomerase gene, denoted $xylA_1$, but these two genes are not similar: $xylA_2$ is 160 codons shorter than $xylA_1$, and alignment of $xylA_2$ to $xylA_1$ shows a low level of identity (∼15%) at the amino acid level. The structure of the xylose degradation operon in *P. syringae* is shown in FIG. 6.4.2a.

To further understand the evolution history of $xylA_2$ in relation to the *xyl* operon already found within *P. syringae*, we constructed maximum likelihood phylogenies for genes at both loci (FIG. 6.4.2b). The gene phylogeny for $xylA_1$ is in broad topological agreement with that for the core genome for these strains, and the level of divergence between $xylA_1$ homologs from phylogroups 1, 2 and 3 is similar to that of other core genes (FIG. 6.4.2b*i*). A notable exception to the 'core-like' evolution of genes at the *xyl* locus was *xylH*, encoding the permease component of the D-xylose transport system. This gene showed the monophyly of PG1 and PG3 homologs, with PG2 homologs basal to this group (FIG. 6.4.2b*ii*). The phylogeny for $xylA_2$ shows 100% bootstrap support for the partitioning of PG1 and PG3 $xylA_2$ homologs (FIG. 6.4.2b*iii*). However, the divergence between PG1 and PG3 $xylA_2$ homologs is notably reduced relative to $xylA_1$: the maximum patristic distance between PG1 and PG3 for $xylA_1$ is 0.249 substitutions per site, which is comparable to the 'genome-wide

Table 6.2: Patterns of divergence for $xylA_1$ and $xylA_2$.

	Distance[a]	K_A[b]	K_S[b]	K_A/K_S	%GC	
					PG1	PG3
$xylA_1$	0.249	0.028	0.631	0.044	59.1	59.0
$xylA_2$	0.068	0.020	0.118	0.174	57.7	58.5
Core genome	0.215	0.094	1.644	0.057	59.9	59.2

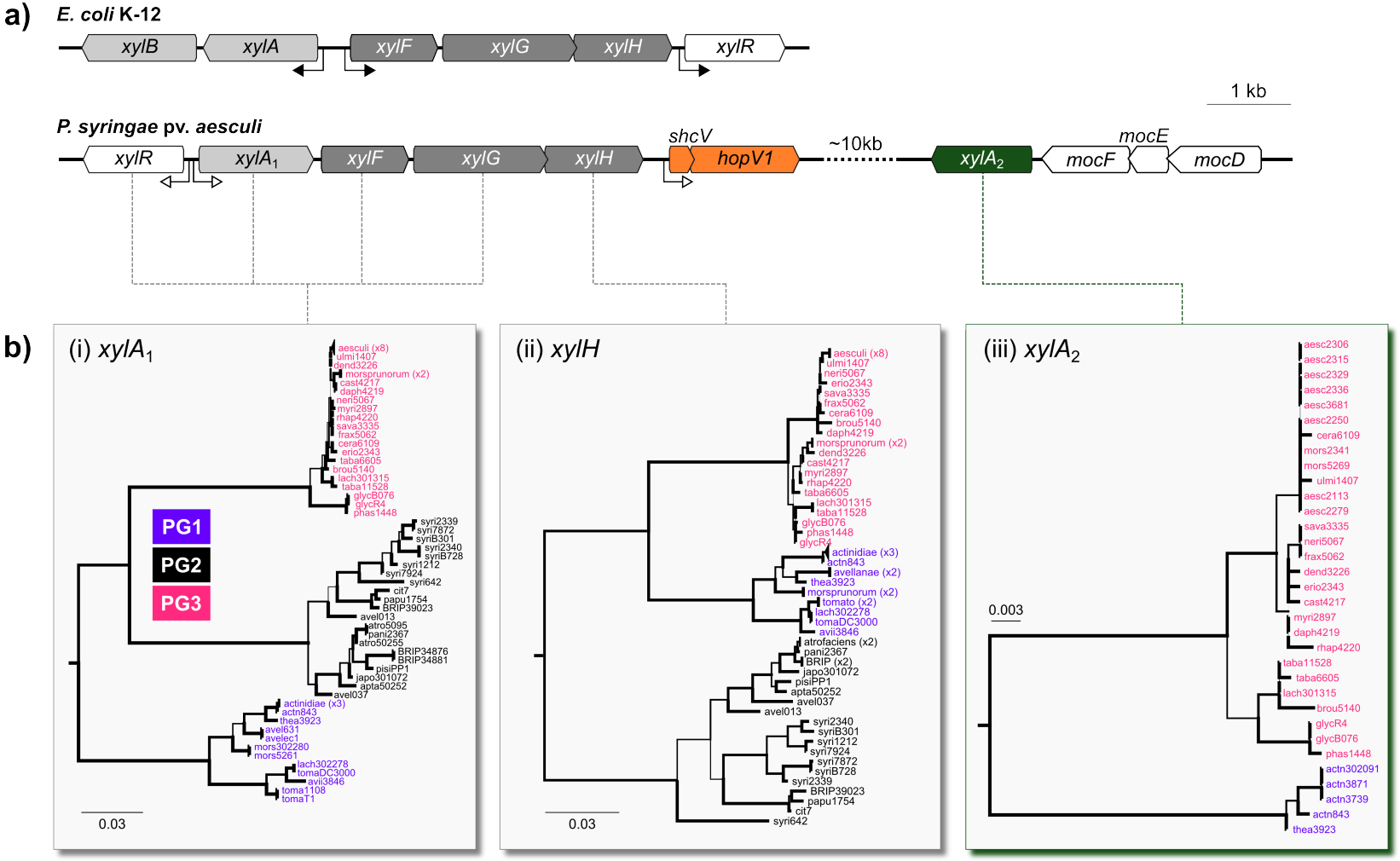
[a]Total (patristic) distance (measured as substitutions per site); [b]Values of K_A and K_S are the average of all pairwise comparisons between PG1 and PG3 homologs for each gene. K_A/K_S is then the ratio of these pairwise averages.

average' value of 0.215 subs/site (as represented by the same distance calculated for the core genome phylogeny). The equivalent distance for $xylA_2$, however, is over three times smaller, at 0.068 subs/site (TABLE 6.2).

Further investigation showed this difference to be attributable primarily to synonymous site divergence (TABLE 6.2). While the pairwise average rate of non-synonymous substitution (K_A) between PG1 and PG3 homologs for both genes was approximately 0.02, the rate of synonymous substitution (K_S) for $xylA_2$ was $\sim$5 times reduced relative to $xylA_1$ (0.118 and 0.631 respectively). Correspondingly, the K_A/K_S ratio for $xylA_2$ was estimated to be 0.174, approximately four times greater than that for $xylA_1$. %GC content values for $xylA_1$ and $xylA_2$ in both PG1 and PG3 are all similar to the average genome-wide value of 59% G + C (TABLE 6.2).

We also investigated the surrounding genomic neighbourhood for PG1 and

Figure 6.3 *(following page)*: Structure and evolution of the xylose degradation operon in *P. syringae*. **(a)** Comparison with *Escherichia coli* K-12 shows multiple gene rearrangements within this region, including the loss of the xylulokinase gene (*xylB*), which is located elsewhere in the *P. syringae* genome, and the rearrangement of *xylAFGH* into a single transcriptional unit. Light grey indicates metabolic enzymes, dark grey indicates transport components, the regulatory gene *xylR* is shown in white. The position of the T3SE gene *hopV1*, immediately downstream of the *xylRAFGH* operon, is shown in orange. Confirmed and predicted promoter regions are indicated by filled and open arrows respectively. The layout for K-12 is adapted from Song and Park (1997). **(b) (i)** Individual gene phylogenies for *xylR*, $xylA_1$, *xylF* and *xylG* are similar to that of the core genome ($xylA_1$ is shown). **(ii)** The phylogeny for *xylH*, however, suggests that PG1 and PG3 share a more recent common ancestor for this gene. **(iii)** Gene phylogeny for the alternative xylose isomerase ($xylA_2$), which is located on a different contig in the *Pae* assembly but maps to a position $\sim$10 kb from the *xylRAFGH* operon in the completely sequenced chromosome of pv. *phaseolicola* 1448A. Taxa are coloured blue, black and pink corresponding to PG1, PG2 and PG3 respectively. Branch thicknesses are drawn relative to bootstrap support (thicker indicates higher support) for that branch. Scale bars represent 0.03 subs/site for (i) and (ii), and 0.003 subs/site for (iii). Trees are rooted with *macu4326* (not shown) for (i) and (ii), whereas (iii) is midpoint rooted.


a)
E. coli K-12
xylB
xylA
xylF
xylG
xylH
xylR
1 kb
P. syringae pv. aesculi
xylR
xylA1
xylF
xylG
xylH
shcV
hopV1
~10kb
xylA2
mocF
mocE
mocD
b)
(i) xylA1
PG1
PG2
PG3
0.03
(ii) xylH
0.03
(iii) xylA2
0.003

PG3 homologs of the $xylA_2$ gene. In both PG1 and PG3 genomes, $xylA_2$ is located immediately downstream of the *mocDEF* operon, and forms a single transcriptional unit with these genes (see FIG. 6.4.2a). Genes of the *mocDEF* operon are also inferred to have been gained in the ancestor to PG3, and are similarly encoded only by PG3 strains and pv.'s *theae* and *actinidiae* in PG1. BLAST analysis of the entire *mocDEF* operon (4.8 kp) shows a nearest hit outwith PG1 and PG3 *P. syringae* to other pseudomonad species such as *P. fluorescens* (73% nucleotide identity), *P. poae* (73%) and *P. cichorii* (70%). Partial coverage was observed for a number of rhizobial species, such as *Mesorhizobium opportunistum* (64% identity over 33% query length) and *Neorhizobium galegae* (63% identity over 29% query).

It was not possible to ascertain the gene neighbourhood upstream of $xylA_2$ for E-*Pae* assemblies, as in all cases for this pathovar the $xylA_2$ ORF was located near ($\sim$700 bp) to a contig terminus. In the complete genome of the PG3 strain *phas1448*, however, $xylA_2$ is found $\sim$10.5 kb from the xylose degradation operon (containing $xylA_1$); for *thea3923* in PG1 (which was the only PG1 strain with a contig that spanned this region), this distance was $\sim$6.5 kb.

In addition to the xylose isomerase gained in the ancestor to PG3, two genes involved in the transport of D-xylose across the cell membrane were inferred to have been acquired in the ancestor of the monophyletic cluster of woody-host pathovars in PG3 (i.e., the 'woody-host' clade in FIG. 6.1). The first, denoted $xylG_2$, shows a high degree of similarity to the ATP-binding component of an ATP-binding cassette (ABC) xylose transporter encoded by *P. fluorescens* (86% identity), with good matches to *Pseudomonas putida*, *Pseudomonas fragi* and *Burkholderia* and *Azospirillum* spp. also identified. The second gene, denoted $xylH_2$, is similar to genes encoding the permease component of an ABC sugar transporter found in a similar range of species as for $xylG_2$. Alignment to the *aesc2336* assembly shows that these genes are also distinct from the *xylRAFGH* operon discussed above, and show only weak amino acid identity ($\sim$43 and $\sim$38% respectively) to the $xylG_1$ and $xylH_1$ homologs within the *xylRAFGH* cluster. Both $xylG_2$ and $xylH_2$ were restricted to the PG3 woody clade (i.e., were not found in *P. syringae* strains in other phylogroups), but showed

variable occurrence within this group; $xylG_2$ was not detected in the strains *mors2341*, *mors5269*, *sava3335*, *daph4219*, *ulmi1407* and *aesc3681* (the Indian strain of pv. *aesculi*), while $xylH_2$ was absent from *sava3335*, *daph4219*, *ulmi1407* and *aesc3681*. Both $xylG_2$ and $xylH_2$ were located on the same contig as a number of genes involved in the uptake and utilisation of sucrose, as previously discussed by Green et al. (2010).

Type III secretion system effectors

We find evidence for the gain of at least six effector proteins of the type III secretion system (T3SS) in the evolutionary history of the *aesculi* pathovar. The acquisition of the type III secretion effector (T3SE) gene *hopV1* was inferred to have occurred along the branch ancestral to PG3. BLAST analysis revealed this T3SE to be ubiquitous among PG3 strains, but was also found in pv.'s *theae* and *tomato* str. DC3000 in PG1, and pv. *maculicola* str. ES4326 (the outgroup). Phylogenetic analysis of *hopV1* showed the pv. *theae* homolog to cluster within the PG3 clade, suggesting the recent transfer of this gene from a PG3 lineage into the pv. *theae* genome (FIG. 6.4). Alignment of *hopV1* to the *aesc2336* assembly showed this gene to be inserted immediately downstream of the D-xylose degradation operon *xylRAFGH*, discussed previously (see FIG. 6.4.2a).

Three T3SEs, *avrA1*, **hopQ1-1* and *hopT1-1*, were inferred to have been gained on the ancestral branch to pv.'s *ulmi* and *aesculi*. The *avrA1* gene and the partial **hopQ1-1* map to a ∼8 kb contig and are embedded among a number of mobile elements. The partial **hopQ1-1* is a 138 bp fragment (∼10%) aligning to the middle of the full-length *hopQ1-1*, the latter being distributed across phylogroups 1 and 3 (but not PG2). The *avrA1* gene encodes a 906 amino acid protein, and is present in full-length in all eight *aesculi* strains, *ulmi1407*, *brou5140*, and *tomaT1* and *toma1108* from PG1.

hopT1-1 is found on a ∼4.5 kb contig also littered with transposase genes. Full-length *hopT1-1* homologs are found in all eight *aesculi* strains, *cast4217*, *ulmi1407* and *taba11528* from PG3, *avii3846*, *mors5261*, *tomaT1*, *toma1108* and *tomaDC3000* from PG1, and the outgroup *macu4326*.

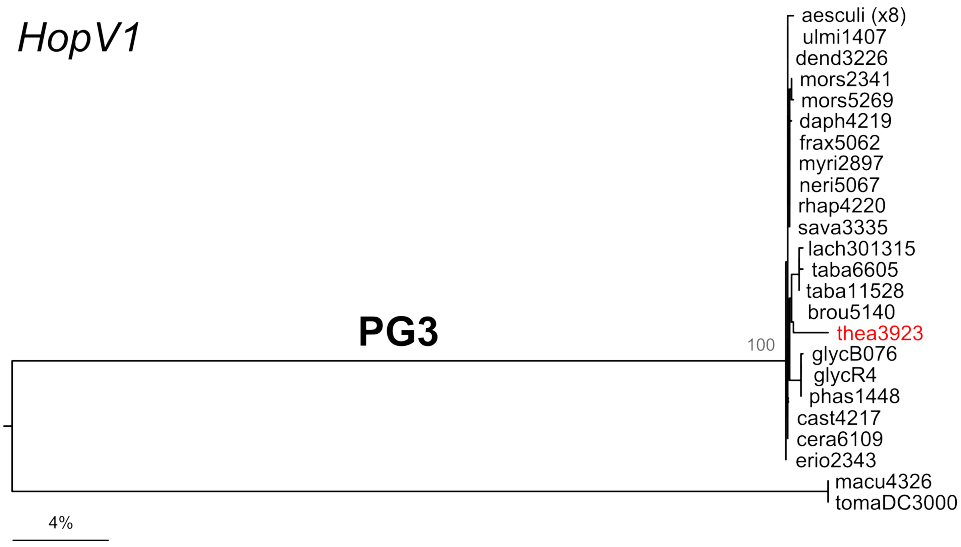


Figure 6.4: *HopV1* gene phylogeny. The PG1 strain *thea3923* is highlighted in red. Grey numbers indicate bootstrap support. Tree is midpoint rooted; scale bar indicates 0.04 substitutions per nucleotide site.

avrB4 and '*xopB*-like' were inferred to have been gained in the ancestor to both E-*Pae* and I-*Pae*. The *avrB4* gene aligns to a ∼9 kb contig in *aesc2336* containing a number of transposases and hypothetical proteins, and is encoded by pv.'s *aesculi*, *broussonetiae* and *mori* in PG3, and pv.'s *avii* and *actinidiae* in PG1. The *xopB*-like gene encodes a 831 amino acid peptide which is annotated as a hypothetical protein in the PG3 pv.'s *aesculi*, *broussonetiae*, *daphniphylli* and *dendropanacis* in which it is found. However, XopB-like shows partial similarity (∼34% amino acid identity over ∼44% query length) to the type III secretion system effector XopB characterised in *Xanthomonas* spp.

Finally, **hopW1* is inferred to have been gained on the branch leading to the seven E-*Pae* strains. This gene aligns to the 5′ end of the full length *hopW1* ortholog, but has been disrupted by a frameshift mutation leading to the introduction of a premature stop codon in the CDS. The partial **hopW1* is found in E-*Pae*, *savastanoi*, *fraxini* and *nerii* from PG3, and pv.'s *actinidiae*, *theae*, *morsprunorum* and *avellanae* from PG1, although alternative fragments of *hopW1* were detected in a range of pathovars from across the *P. syringae* species complex.

Degradation of plant-derived aromatic compounds

The ability of the *aesculi* lineage to catabolise plant-derived aromatic compounds via both the catechol branch of the β-ketoadipate pathway and the protocatechuate 4,5-dioxygenase pathway has been previously implicated as a potential adaptation of *Pae* to the woody parts of its horse chestnut host (Green et al., 2010). Previous work has highlighted the variation in the occurrence of these pathways across the *P. syringae* species complex (see CHAPTER 5); here, we formalise these results within the context of the gain of these pathways along the phylogenetic lineage leading to *Pae*.

Our reconstruction suggests that the β-ketoadipate pathway was gained relatively early in the evolution of PG3 *P. syringae* strains, along the branch leading to the woody-host clade. Genes comprising this operon are found in all PG3 woody-host strains, and pv.'s *actinidiae*, *theae* and both *morsprunorum* strains in PG1. Phylogenetic analysis of the ∼7.5 kb concatenated alignment of these 10 genes shows the well-supported partitioning of these homologs into clusters that correspond to phylogroups 1 and 3 of the core genome phylogeny (FIG. 6.4.2). As for $xylA_2$, however, the observed level of divergence between PG1 and PG3 homologs was substantially lower compared to genes of the core genome (maximum patristic distance of 0.131 subs/site versus a genome-wide average of 0.215). The β-ketoadipate operon is located in the middle of a large contig (∼109 kb) in the *aesc2336* assembly; thus, we looked for a phylogenetic signal of recombination in this region by constructing individual gene phylogenies for a number of core genes in the immediate upstream and downstream locations of this gene cluster (FIG. 6.4.2). Similarly to that shown by the flanking regions of the $xylA_2$/*mocDEF* insertion, tree topologies for genes immediately upstream and downstream show the clustering of PG1 with PG3, whereas phylogenies for loci further away show a core-like tree topology (FIG. 6.4.2).

In contrast to the β-ketoadipate operon, genes involved in the degradation of protocatechuate via the protocatechuate 4,5-dioxygenase pathway were inferred to have been gained more recently, in the ancestor of both E-*Pae* and I-*Pae*. The presence of the four genes that comprise this operon is a unique characteristic of pv. *aesculi*

among lineages of the *P. syringae* species complex, while outwith *P. syringae*, the best-BLAST hits for these proteins are to other members of the Pseudomonadaceae, such as *P. putida* and *P. fragi*.

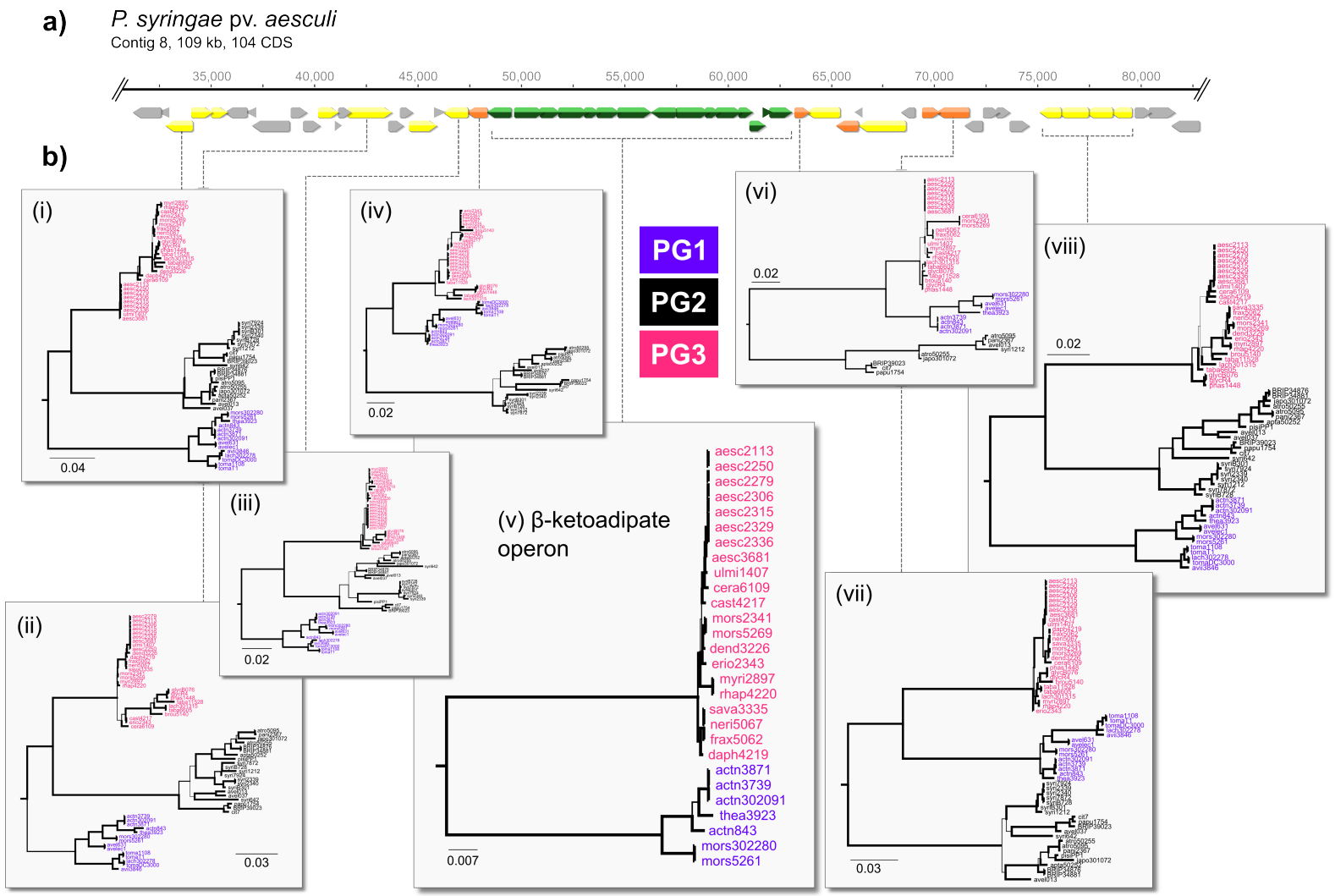
In *Pae*, both the β-ketoadipate and protocatechuate operons reside near to genes involved in the sequestration of iron. For example, the β-ketoadipate operon has been inserted near to genes encoding the EfeUOB transporter complex, involved in the acquisition of both ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron (Große et al., 2006; Miethke et al., 2013). In PG1 strains that possess the β-ketoadipate genes, the *efeUOB* operon is located elsewhere in the genome, suggesting a different insertion point to that of PG3 strains. Similarly, the protocatechuate genes are located near to a series of genes involved in the sequestration of ferric iron via the secretion of the enterobactin siderophore. While *efeUOB* is common to all *P. syringae*, the enterobactin biosynthesis operon is also specific to the *aesculi* pathovar.

Other virulence factors

A range of other genes with functions in virulence or metabolic pathways that may have facilitated the adaptation of *P. syringae* to the woody niche were also inferred to have been acquired at various points in the evolutionary history of pv. *aesculi*.

For example, the coronatine phytotoxin, a structural mimic of the plant hormone jasmonic acid (JA), was inferred to have been gained in the ancestor to pv.'s *aesculi* and *ulmi*. The genetic basis for this complex toxin involves a suite of approximately 20 genes spread over two non-contiguous genome regions, which may be plasmid-borne or on the main chromosome depending on the pathovar (e.g., in pv. *glycinea*,

Figure 6.5 *(following page)*: Phylogenetic history of the β-ketoadipate operon. **(a)** Part of the $\sim$109 kb contig from *aesc2336* containing the β-ketoadipate gene cluster (green). Yellow demarks individual genes with a phylogenetic history that is congruent with that of the core genome phylogeny; orange demarks genes where there is phylogenetic discordance. Grey indicates genes for which phylogenies were not estimated. **(b)** Selected gene phylogenies. Strains from the three phylogroups are shown in blue, black and pink for PG1, PG2 and PG3 respectively (as for Fig. 6.4.2). All phylogenies are rooted with the outgroup strain *macu4326* (not shown), except for trees (v) and (vi) for which a outgroup sequence was not available; these trees were midpoint rooted. Branch thicknesses are drawn relative to bootstrap support (thicker indicates higher support) for that branch. Scale bars represent nucleotide substitutions per site; note different scales for each phylogeny.


a)
P. syringae pv. aesculi
Contig 8, 109 kb, 104 CDS
35,000
40,000
45,000
50,000
55,000
60,000
65,000
70,000
75,000
80,000
b)
PG1
PG2
PG3
(i)
0.04
(ii)
0.03
(iii)
0.02
(iv)
0.02
(v) β-ketoadipate operon
aesc2113
aesc2250
aesc2279
aesc2306
aesc2315
aesc2329
aesc2336
aesc3681
ulmi1407
cera6109
cast4217
mors2341
mors5269
dend3226
erio2343
myri2897
rhap4220
sava3335
neri5067
frax5062
daph4219
actn3871
actn3739
actn302091
thea3923
actn843
mors302280
mors5261
0.007
(vi)
0.02
(vii)
0.03
(viii)
0.02

coronatine genes are located on plasmid p4180A whereas in pv. *tomato* str. DC3000 they are found on the chromosome) (Ma et al., 1991; Bender et al., 1999). Coronatine biosynthesis genes were found in pv.'s *aesculi*, *ulmi*, *morsprunorum* and *glycinea* str. B076 in PG3, pv. *tomato* str. DC3000 in PG1 and the more distantly related pv. *maculicola* str. ES4326.

A ∼6.5 kb contig with genes resembling the *pilLNOPQ* operon encoded by *P. fluorescens* Pf-5 was gained on the ancestral branch to both E-*Pae* and I-*Pae*. These genes are inferred to have a function in the biosynthesis of mating pili (Mavrodi et al., 2009), and are found in a range of *P. syringae* strains from all three phylogroups: pv. *actinidiae*, *toma1108* and *tomaT1* from PG1, pv.'s *japonica*, *panici*, *BRIP34876*, *BRIP34881*, *BRIP39023*, *atro5095*, *avel013*, *cit7*, *syri2340* and *syriB728a* from PG2 and pv.'s *aesculi* and *tabaci* in PG3.

In addition, two genes with putative roles in the degradation of the plant-derived compounds toluene and xylene were identified. These included a short-chain dehydrogenase gained in the common ancestor to pv.'s *aesculi* and *ulmi*, and a gene encoding the enzyme 4-oxalocrotonate tautomerase (4-OT) in the ancestor of the European *aesculi* strains. 4-OT is part of an inducible set of enzymes encoded by the TOL plasmid of *P. putida*, and expression of these genes allows carrying bacteria to utilise toluene and xylene as their sole source of energy (Chen et al., 1992). However, there was no evidence of homology between E-*Pae* contigs and the TOL plasmid (pWW0), and none of the remaining genes of the xylene degradation pathway encoded on the TOL plasmid were found among the set of *P. syringae* genomes analysed.

A pectate lyase precursor gene was also inferred to have been gained in the ancestor to E-*Pae*. Pectate lyase is an enzyme involved in the degradation of plant cell walls and is a known virulence factor in other pseudomonads such as *P. viridiflava* and *P. fluorescens*. The pectate lyase acquired in the ancestor of E-*Pae* aligned to the 5′ end of the *pelS* gene, previously characterised in the cucumber pathogen pv. *lachrymans* (Bauer and Collmer, 1997). However, a 32-bp insertion into the E-*Pae* *pelS* gene has resulted in the truncation of this gene; thus, the PelS encoded by E-*Pae*

Table 6.3: Pairwise nucleotide differences between E-*Pae* strains.

	aesc2113	aesc2279	aesc2306	aesc2315	aesc2329	aesc2336
aesc2113	—					
aesc2279	5	—				
aesc2306	2	3	—			
aesc2315	8	9	6	—		
aesc2329	5	6	3	9	—	
aesc2336	6	7	4	10	7	—

is likely to be non-functional.

6.4.3 Divergence analyses

We also investigated the extent of nucleotide divergence among six E-*Pae* strains each sampled from a different geographic locations across northern Europe.

In total, we detected 40 well-supported variant sites across 5.84 Mb of aligned genome sequence. Of these, 29 were single nucleotide polymorphisms (SNPs) while the remaining 13 were insertion/deletions (indels) ranging from 1 to 36 bp in length. Excluding variant sites located on contigs with homology to plasmid sequences reduced these values to 18 SNPs and 11 indels. Pairwise nucleotide differences between E-*Pae* strains are shown in TABLE 6.3. Values range from 2 to 10, with the least diverged pair being *aesc2113* and *aesc2306*, isolated from the UK and Belgium respectively, and the most diverged pair being *aesc2315* and *aesc2336*, from the Netherlands and Norway respectively.

To elucidate the potential phylogeographic signal of these data, the homologous nucleotide site for these 18 SNPs was extracted from two previously sequenced *Pae* strains (Green et al., 2010), *aesc2250* (UK) and *aesc3681* (India). However, these combined data showed very little evolutionary structure that might indicate potential routes of transmission of *Pae* strains across Europe, and a phylogeny constructed from the observed distances showed a star-like topology that is typical of a lack of phylogenetic signal (FIG. 6.6).

Overall, only two polymorphisms were shared among the eight strains: *aesc2250* and *aesc2113*, both from the UK, share a non-synonymous G $\rightarrow$ T substitution (resulting in the amino acid change glutamine to histidine) within a sensor histidine

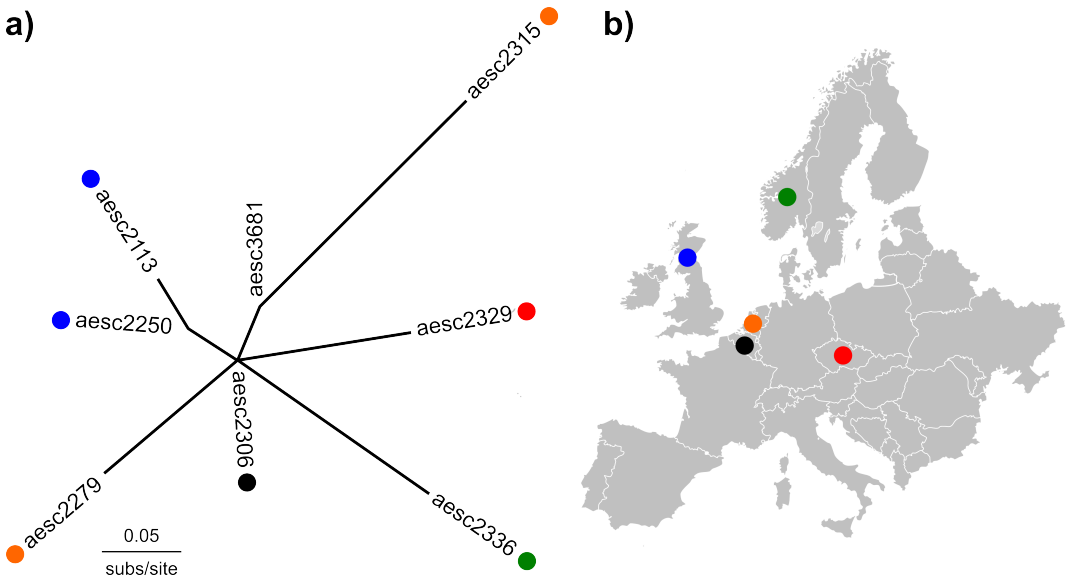


Figure 6.6: Lack of phylogenetic signal among *Pae* strains. **(a)** Star-like phylogeny shows little evolutionary structure among strains. Colours indicate location of sampling indicated on the map of Europe shown in **(b)**; blue, orange, red, green and black represent the UK, the Netherlands, the Czech Republic, Norway and Belgium, respectively. Strain *aesc3681* was originally isolated from Northern India.

kinase gene, while *aesc2315*, from the Netherlands, shares a A → C substitution with *aesc3681*, from India, resulting in a glutamine to proline amino acid change within a LysR family transcriptional regulator protein. Full gene effects of these SNPs are shown in TABLE 6.4. As might be expected for bacteria, where the majority of the genome is coding, all but two (∼89%) variants occur in coding regions (CDS). Of these, (∼81%) are non-synonymous, resulting in missense mutations. No detected variants resulted in the introduction of a premature stop codon into the CDS. The transition to transversion ratio (Ti/Tv) was 0.64.

For highly similar sequences with a recent common ancestor, as is the case here, the nucleotide substitution rate can be approximated by simply dividing the number of differences between a pair of sequences by the alignment length (known as the *p*-distance). Using a denominator of 5.73 Mb in all cases (i.e., the filtered genome length minus the total length of putative plasmid sequences in *aesc2336*), estimated values for pairs of E-*Pae* strains ranged from 0.35–1.75$\times 10^{-6}$, with an average of 1.05$\times 10^{-6}$ subs/site.

To estimate the time to the most recent common ancestor (TMRCA) for E-*Pae*, we

Table 6.4: Gene effects for 18 SNPs among E-*Pae* strains.

#	Contig	Length	Pos	Ref[a]	Alt[b]	Effect[c]	Type	Codon[d]	AA[e]	Protein affected
1	Contig_4	137,344	83,910	G	T	Non-syn	Missense	caG/caT	Q1845H	Sensor histidine kinase
2	Contig_5	134,135	103,223	C	A	Non-syn	Missense	Ctg/Atg	L169M	DAHP synthase
3	Contig_10	97,342	78,834	T	C	Non-syn	Missense	Ttt/Ctt	K257L	Putrescine ABC transporter PotF
4	Contig_13	94,571	9,173	A	C	Non-syn	Missense	cAg/cCg	Q170P	LysR family transcriptional regulator
5	Contig_16	87,661	14,374	G	T	Non-syn	Missense	tgG/tgT	W83C	Hypothetical protein
6	Contig_16	87,661	43,806	T	C	Non-syn	Missense	Tgg/Cgg	W291R	Amino acid permease protein LivM
7	Contig_18	86,779	77,386	G	A	Non-syn	Missense	gGc/gAc	G515D	DNA gyrase subunit A
8	Contig_18	86,779	77,387	C	A	Syn	Silent	ggC/ggA	—	DNA gyrase subunit A
9	Contig_25	73,525	70,022	A	G	Non-syn	Missense	atA/atG	I282M	TldE/TldD proteolytic complex
10	Contig_33	64,975	28,756	G	T	Syn	Silent	ggG/ggT	—	Alpha-ketoglutarate transporter
11	Contig_34	61,959	18,454	G	A	Non-syn	Missense	Gaa/Aaa	E24K	Butyryl-CoA dehydrogenase
12	Contig_39	54,220	7,533	G	A	Syn	Silent	caG/caA	—	Two-component sensor PilS
13	Contig_44	46,444	14,542	G	T	Non-syn	Missense	aGc/aTc	S66I	Malate dehydrogenase
14	Contig_57	37,559	6,974	G	C	Non-syn	Missense	cGg/cCg	R224P	Hypothetical protein
15	Contig_68	33,248	20,898	G	T	Non-syn	Missense	Gcc/Tcc	A297S	Glyceraldehyde-3-phosphate dehydrogenase
16	Contig_70	32,511	16,171	G	T	Intergenic	—	—	—	—
17	Contig_77	25,261	14,346	A	C	Intergenic	—	—	—	—
18	Contig_84	22,152	1,225	G	A	Non-syn	Missense	Gcc/Acc	A652T	Chromosome segregation ATPase

[a]State of site in reference genome (*aesc2336*); [b]State of site in variant genome; [c]Non-syn, non-synonymous; Syn, synonymous; [d]Codon change, capital letter indicates variant site; [e]Amino-acid change.

looked for a correlation between the root-to-tip divergence and the year of sampling for each strain, inferring the consensus of each SNP site to be the ancestral (root) state. However, we detected no significant time signal in these data ($t = -0.45$, $p = 0.66$); thus, it is not possible to measure the TMRCA using this approach.

A well-known phenomenon in DNA studies is the apparent decrease in the estimated rate of nucleotide substitution and the length of time over which this rate is calculated (reviewed in Ho et al., 2011). Recently, Comas et al. (2013) compiled estimates of substitution rates and divergence times from eight different bacterial groups, and suggested a negative log-linear relationship between the observed rate and the TMRCA for each group (the species analysed were: *Yersinia pestis* (Bos et al., 2011), *Streptococcus pneumoniae* (Croucher et al., 2011), *Staphylococcus aureus* (Harris et al., 2010), *Vibrio cholerae* (Mutreja et al., 2011), *Shigella sonnei* (Holt et al., 2012), *Clostridium difficile* (He et al., 2013), *Helicobacter pylori* (Kennemann et al., 2011) and *Mycobacterium tuberculosis* (Comas et al., 2013)). Thus, we investigated if our estimate of short-term divergence among *Pae* strains fits with this prediction.

Although our data did not allow for an estimated TMRCA for *Pae*, we can nonetheless calculate the expected substitution rate across a range of possible TMRCA times given our estimate of divergence of 1.05×10^{-6} subs/site, using the molecular clock prediction:

$$\gamma = K/2t \tag{6.1}$$

where γ represents the rate of nucleotide substitutions measured as subs/site/year, K is the estimated divergence in subs/site, and t is the time since divergence. With closely related strains, as is the case for E-*Pae*, the impact of multiple substitutions per site is likely to be negligible and is ignored. With a range of t values from 10 to 100,000 years, we predict the log-linear decrease in observed substitution rate with respect to divergence time to follow the form: $\log_{10}(Y) = -\log_{10}(X) - 6.28$, where $Y = \gamma$ and $X = 2t$ (FIG. 6.7). Using this approach, we find that our estimate of divergence among *Pae* strains is inconsistent with the relationship inferred by Comas and colleagues. For example, if we assume that the E-*Pae* strains shared a common ancestor 10 years ago, at approximately the time of the first reported cases

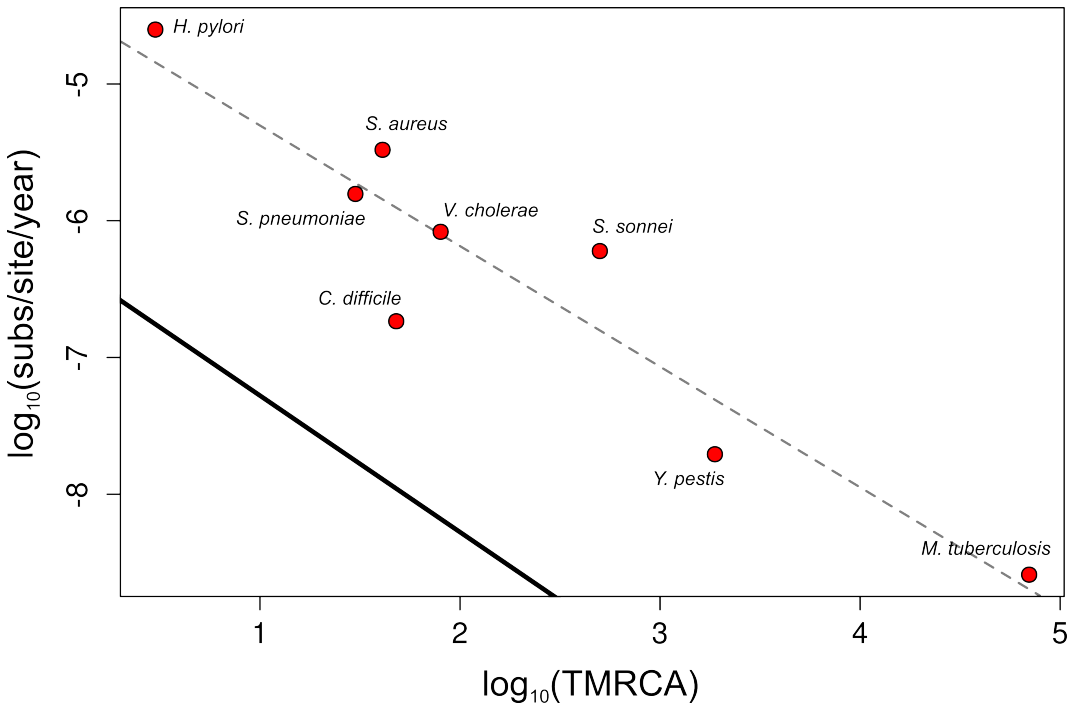


Figure 6.7: Relationship between substitution rate (y-axis) and divergence time (x-axis) among bacteria. Dates and rates for eight species are shown in red, as compiled by Comas et al. (2013). The grey dashed line shows the inferred log-linear relationship for these data ($t = -6.5$, $p < 0.0001$; $R^2 = 0.88$). The bold line shows the predicted relationship between the substitution rate and divergence time specifically for E-*Pae*. Note both axes are drawn on a $\log_{10}$ scale.

of bleeding canker in Europe, the rate estimate would be 5.25×10^{-8} subs/site/year; similarly, if the time to the shared a common ancestor were 100 or 1000 years ago, the rate estimates would be 5.25×10^{-9} and 5.25×10^{-10} subs/site/year respectively. The dependence of the rate estimate on the TMRCA is shown by the bold line in FIG. 6.7. Notably, there is no reasonable date that would yield a rate compatible with the conclusions of Comas et al.; whatever the actual time of divergence among the E-*Pae* strains, the rate estimated is almost two orders of magnitude slower than the relationship suggested by Comas et al.

6.5 Discussion

In this chapter, we investigate some of the key genomic changes that have facilitated the adaptation of *P. syringae* into the woody niche, and, specifically, pv. *aesculi* (*Pae*)

onto the European horse chestnut. Our 'phylogeny-aware' comparative method allows not only for the identification of putative adaptations to this niche, but also, crucially, of *when* these changes are most likely to have occurred in the evolutionary history of the *Pae* lineage. Thus, our approach identifies a number of novel genomic innovations that are implicated in the adaptation of *Pae* to the woody niche, and offers new insights into the genetic and genomic bases of niche expansion in *P. syringae*. Below, we summarise our findings and discuss the wider implications of these results within the context of bacterial evolution and adaptation.

6.5.1 Evolution of PG3 pathovars onto woody hosts

A previous analysis of the evolutionary relationships among *P. syringae* pathovars identified a monophyletic group of woody-host pathogens within PG3, including pv.'s *aesculi*, *savastanoi*, *myricae*, *morsprunorum* and *mori* (Green et al., 2010). These pathovars are the causal agents of various gall, knot or canker diseases in the woody parts of their respective host species, with the exception of pv. *mori*, which causes leaf-spot lesions in the white mulberry (*Morus alba*). Pathovar *mori* was shown to cluster at the base of this clade, prompting the suggestion that the evolution of PG3 *P. syringae* onto the woody parts of a host plant may have occurred in a step-wise manner—first with the ability to colonise the leaves of a woody host (mulberry), then onto the woody organs themselves (Green et al., 2010).

Our phylogeny, based on the concatenated alignment of more than 2,000 core genes, significantly increases the sampling depth of woody-host strains within this group and is thus able to further our understanding of the evolution of PG3 pathovars onto woody hosts. There appear to have been at least two transitions onto woody hosts within PG3: the first is represented by the well-supported 'woody clade' and contains the majority of sampled woody-host strains that are pathogenic on a diverse range of host species, while the second involves a later, relatively minor transition involving only the *brousonnetiae* pathovar. These switches are likely to have occurred at different times, have resulted in very different patterns of subsequent diversity, and are apparently not underpinned by the same genetic basis

(discussed in more detail later). In addition, the core genome phylogeny shows pv. *mori* to cluster with a number of pathogens of herbaceous hosts. This suggests that, for the majority of PG3 woody-host strains at least, the movement of the ancestral lineage into the woody niche was not preceded by the ability of this lineage to colonise the leaves of an *M. alba* host.

However, the phylogeny shows pv. *mori* to be closely related to the only PG3 woody-host strain that is not grouped within the woody-clade, pv. *broussonetiae*, which was isolated from a necrotic lesion on the shoots of another species of mulberry (*Broussonetia kazinoki*). Thus, a step-wise mechanism of evolution may have occurred in the adaptation of these strains onto members of the Moraceae: first onto the leaves of *M. alba*, then into the woody parts of the related *B. kazinoki*. Thus it may be the case that the ancestor to the woody clade followed a similar evolutionary trajectory, but that the descendants of early leaf-colonising lineages have either become extinct or have simply not been sampled.

6.5.2 Gene gain, loss and rearrangement of the D-xylose operon

Our results implicate the utilisation of D-xylose as a potentially important adaptation in woody-host *P. syringae* pathovars. This compound (from the Latin: 'wood sugar') is an environmentally abundant pentose (five carbon) sugar, and is the primary constituent of hemicellulose xylan, itself a major component of both hard- and softwoods (Jeffries, 1983). Thus, the ability to utilise D-xylose as a carbon source is likely to be selectively advantageous for bacterial pathogens that cause cell damage and necrosis in the woody parts of their host species. D-xylose degradation has been described in a number of other bacterial species, including *Escherichia coli* (Lawlis et al., 1984), *Bacillus* spp. (Rygus et al., 1991; Scheler et al., 1991), *Lactobacillus pentosus* (Lokman et al., 1991) and *Caulobacter crescentus* (Stephens et al., 2007). In this simple pathway, xylose isomerase (XI) converts D-xylose directly to D-xylulose, which is then phosphorylated to yield D-xylulose-5-phosphate, an intermediate in the pentose phosphate pathway (see Figure 1 of Bettiga et al., 2008). Thus, D-xylose can be utilised as a source of energy in these bacteria, and the pathway is actively

induced in the presence of xylose (Lawlis et al., 1984; Lokman et al., 1991).

We infer the gain of a number of genes involved in both the transport and isomerisation of D-xylose at various points along the lineage leading to *Pae*. Crucially, these genes are not part of the D-xylose degradation operon, *xylRAFGH*, that is present in all *P. syringae* pathovars regardless of host type. The additional *xyl* genes are highly diverged from their homologs within the *xylRAFGH* operon, and thus are unlikely to have arisen via a recent duplication of this region. Rather, we infer these genes to have been imported via horizontal gene transfer (HGT) from a lineage outwith the *P. syringae* species complex. On the basis of BLAST identity, potential donors include other members of the Pseudomonadaceae such as *P. fluorescens*, *P. putida* and *P. fragi*, as well as other plant-associated species such as *Burkholderia*, *Azospirillum* and *Rhizobium* spp. Although the specific selective advantages of these imported *xyl* genes have not been determined for *Pae*, we speculate that their presence may allow for an increase in either the rate or efficiency of D-xylose utilisation in the woody environment.

The alternative xylose isomerase gene ($xylA_2$) is encoded by both PG3 strains and the relatively distantly related PG1 pathovars *actinidiae* and *theae*, causal agents of bacterial canker of kiwi (*Actinidia* spp.) and shoot blight of tea (*Camellia sinensis*) respectively. Phylogenetic analysis of this gene revealed that, while the partitioning of PG1 and PG3 $xylA_2$ homologs is unambiguous, the level of sequence divergence between groups is greatly reduced relative to genes of the core genome ($K_S = 0.12$ for $xylA_2$, compared to $K_S = 0.63$ for $xylA_1$ and an average $K_S \gg 1$ across all core genes). The reduced divergence of $xylA_2$ is unlikely to be driven by selectional constraints, since K_A for this gene is approximately equivalent to that for $xylA_1$. Thus, the observed level of divergence is incongruous with the evolutionary scenario of a single acquisition of $xylA_2$ into the ancestor of all three *P. syringae* phylogroups, followed by its retention in PG3 and certain PG1 lineages, and loss from others in which the gene is absent (e.g., the ancestor to PG2).

Furthermore, a number of putatively gained genes, including $xylA_2$, *hopV1* and the mannopine degradation genes *mocDEF*, mapped to locations either at or near

to the *xylRAFGH* operon. This pattern was observed in both PG1 and PG3 strains. Detailed phylogenetic analysis of the *xylRAFGH* operon revealed marked differences in the evolutionary histories of these genes. While *xylR*, $xylA_1$, *xylF* and *xylG* all showed a 'core genome-like' evolutionary history, the final gene of the operon, *xylH*, showed the opposite pattern: here, PG1 strains clustered as a monophyletic clade with PG3 strains, with PG2 as the basal group. This pattern is indicative of a recombination event between the ancestral PG1 and PG3 lineages, such that the time to coalescence for PG1 and PG3 *xylH* homologs is much shorter than the genome-wide average.

Thus, we hypothesise that the presence of these genes in both PG3 strains and pv.'s *actinidiae* and *theae* from PG1 is most likely explained by the gain of a single contiguous sequence encoding (at least) $xylA_2$, *mocDEF* and possibly *hopV1* in either the ancestor to PG1 or the ancestor to PG3, followed by the inter-phylogroup transfer into the other lineage. The genetic evidence of this HGT event is found in the topological discordance of the *xylH* gene, and the transfer most likely involved the homologous recombination of this locus. HGT involving the homologous recombination of flanking regions is known to contribute to O-antigen diversity in Gram-negative bacteria such as *Vibrio splendidus* (Wildschutte et al., 2010), but has perhaps not been considered a major driver of genome-content differences among bacterial lineages (Polz et al., 2013). Given that all pathovars within PG1, including those that no longer encode the $xylA_2$ gene, cluster as a sister group to PG3 at the *xylH* locus, the transfer event is likely to have occurred prior to either of these groups' subsequent diversification into the extant taxa observed today.

In addition, the rearrangement of genes in and around the *xylRAFGH* operon in *P. syringae* has resulted in two structural changes of significant note, when compared to that found in *E. coli*. Firstly, the gene encoding the enzyme xylulokinase (*xylB*) has been translocated to a different location in the genome. In *Pae*, *xylB* is found on a different contig; in the complete sequence of *phas1448A*, it is located > 250 kb from the remainder of the operon. The phylogeny for *xylB* agrees with that of the core genome, and thus we infer this rearrangement predates the differentiation of

P. syringae into phylogroups 1, 2 and 3. The impact of the translocation of *xylB* on xylose utilisation in *P. syringae* is unclear. In *Pae*, *xylB* is located downstream from an AraC-family transcriptional regulator of polyol utilisation, in a cluster of genes including polyol permease and a fructokinase. Common polyols include the sugar alcohol xylitol, which is an intermediate in an alternative xylose metabolic pathway (Bettiga et al., 2008). Thus, it may be the case that *xylB* and the *xylRAFGH* operon are co-expressed under similar environmental conditions, namely in the presence of xylose/xylitol, despite the physical distance between these genes. Secondly, rearrangements within the operon have resulted in the reorientation of the *xylAFGH* genes into a single transcriptional unit, with the regulatory gene *xylR* orientated in the opposite direction.

6.5.3 Potential role of *hopV1* in the woody niche

The simultaneous acquisition of the type III effector gene *hopV1* with $xylA_2$ potentiates its involvement as a virulence factor in pathovars of woody hosts. For both PG3 and pv. *theae*, we find *hopV1*, along with the gene encoding its chaperone molecule ShcV, inserted immediately downstream of, and in the same orientation as, the *xylH* gene of the xylose degradation operon. The rearrangement of the *xylAFGH* genes into a single transcriptional unit, in conjunction with the insertion of *shcV/hopV1* immediately downstream, suggests that *hopV1* may be co-expressed with the inducement of the xylose operon—i.e., in the presence of D-xylose. This mechanism may be selectively advantageous if the HopV1 effector contributes to pathogenicity in xylose-rich environments, such as the woody tissues of an infected woody-host plant. Interestingly, although a full-length *hopV1* homolog was not found in the genomes of the four *actinidae* strains analysed here, we did find some evidence for the presence of partial fragments, suggesting that *hopV1* has been recently deleted from the *actinidiae* lineage. The gain, loss, and rearrangement of T3SE genes is known to occur frequently within *P. syringae* lineages, and the complement of encoded T3SEs in a given pathovar is likely to contribute to both its pathogenicity and host range (Sarkar et al., 2006; Cunnac et al., 2009; Lindeberg et al., 2009, 2012).

We observe that T3SEs gained in the *Pae* lineage are often found on contigs that are littered with mobile genetic elements, and probably comprise either plasmids or pathogenicity islands. Thus, we propose that the loss of *hopV1* from the ancestral *actinidiae* genome may have played a role in the adaptative shift of this lineage from tea (*Camellia*) to the related kiwi (*Actinidia*).

6.5.4 Acquisition of β-ketoadipate pathway coincides with expansion into the woody niche

The ability of *P. syringae* pathovars such as *aesculi*, *savastanoi* and *actinidiae* to utilise plant-derived aromatic compounds via the β-ketoadipate pathway has been previously implicated as an important adaptation in these bacteria (Green et al., 2010; Rodríguez-Palenzuela et al., 2010; Marcelletti et al., 2011). A very recent investigation into the epidemic potential of environmental strains of *P. syringae* has highlighted the importance of this operon for pathogenicity on kiwi (Bartoli et al., 2014). Our findings agree with this assessment: we show the gain of the β-ketoadipate pathway to occur on the branch leading to the woody clade within PG3, and, with the exception of pv. *broussonetiae*, delimits pathogens of woody hosts from pathogens of herbaceous hosts within PG3. Thus, we suggest that the acquisition of these genes was perhaps the major factor that facilitated the niche expansion of *P. syringae* into the woody environment, and may have been the causal factor in the subsequent diversification of this lineage onto a range of woody-host species.

Bartoli et al. (2014) also construct a phylogeny of the β-ketoadipate operon for pv.'s *actinidiae*, *theae*, *aesculi*, *morsprunorum* and *savastanoi*, and show the partitioning of PG1 and PG3 homologs into distinct clusters (as is the case here). These authors suggest that, given the topological congruence of this tree to the species phylogeny and a number of other factors (e.g., the lack of nucleotide composition differences and the absence of nearby mobile elements such as phage integrases or insertion sequences), the presence of this operon in both PG1 and PG3 strains is likely to be due to a single gain in the ancestor to the *P. syringae* species complex. However, our results show a reduced level of divergence between PG1 and PG3 homologs at this

locus that indicates a more recent common ancestor for these genes, relative to the genome-wide average. Furthermore, we find evidence of phylogenetic discordance at genes flanking the β-ketoadipate cluster that is indicative of recombination in these regions. In particular, we observe the clustering of PG1 and PG3 as sister groups similar to the pattern observed at the $xylA_2$ locus. Thus, in contrast to Bartoli et al. (2014), we conclude that the β-ketoadipate pathway was most probably gained subsequent to the division of phylogroups 1, 2 and 3, either by the ancestor to the PG3 woody clade or the ancestor to the *actinidiae* clade, followed by the horizontal transfer into the other lineage. The direction of the inter-phylogroup transfer remains unclear.

6.5.5 A possible role for opine degradation?

We also infer the gain of a number of genes that may encode an additional pathway for the catabolism of plant-derived aromatic compounds in *P. syringae*, including an 4-oxalocrotonate tautomerase gene (4-OT) and the *mocDEF* operon. The *moc* oxygenase system, encoded by MocDEF, has been characterised for the rhizobial species *Sinorhizobium meliloti* and *Rhizobium leguminosarum* (Bahar et al., 1998, 2000), and is part of a larger pathway (*mocCABRDEF*) involved in the utilisation of rhizopine. Rhizopines are unique inositol compounds that are synthesised by nitrogen-fixing bacteroids within the root nodules of plants, as part of the rhizobia–legume symbiotic interaction (Saint et al., 1993; Murphy et al., 1995; Oke and Long, 1999). In addition, the action of MocDEF is thought to be similar to the initial stages of the degradation of aromatic hydrocarbons such as toluene, benzene and xylene via the action of the *tod*, *bnz* and *xyl* pathways encoded by the TOL plasmid in *Pseudomonas putida* (Bahar et al., 2000; Suzuki et al., 1991). However, we find the MocDEF proteins encoded by *Pae* are more similar to the *moc* system of *Rhizobium leguminosarum* than they are to the oxygenase components of the TOL-encoded pathways in *P. putida*.

Rhizopines are classed as opine-like compounds (Murphy et al., 1988). Other members of this family of molecules include mannopine, agropine and rideopine, which are produced by plant gall-tumour cells upon infection and transformation by

crown gall-inducing *Agrobacterium* spp. (Kim and Farrand, 1996; Bahar et al., 2000). The production of opines by the tumour cells is central in the interaction between the *Agrobacterium* pathogen and its host, as these compounds are then utilised as a source of carbon and energy by the bacterium (Kim and Farrand, 1996). In addition, Yamada et al. (1986) have previously reported a high degree of homology between IS elements within pv. *savastanoi* strains and those found on plasmid pTi15955 of *A. tumefaciens*, and suggest a recent common origin for the tumour-inducing genes in both pv. *savastanoi* and *A. tumefaciens*.

However, the *moc* genes encoded by *Pae* are highly diverged from those involved in the catabolism of opines in *Agrobacterium* spp. In addition, we found no evidence to suggest that PG1 or PG3 strains encoded homologs of *mocCABR*, the other genes in the rhizobial opine degradation pathways. Thus, the exact function of the *mocDEF* operon encoded by *P. syringae* remains unclear. It may be that MocDEF allows for the utilisation of opine molecules, either rhizopine or others, produced by other plant-pathogenic bacterial species that are commonly found within the woody niche. Alternatively, MocDEF may contribute to the initial stages of the degradation of other plant-derived aromatic hydrocarbons, such as toluene and/or xylene, both of which are compounds associated with wood (e.g., both are found in wood oils) (Kandyala et al., 2010). We note a number of other genes inferred to have been gained later along the lineage leading to *Pae* that are also inferred to have functions in the degradation of benzene/toluene/xylene (BTX). Since gained genes that do not confer a selective advantage are usually rapidly deleted (Rocha, 2008; Touchon et al., 2009), the MocDEF genes were presumably selectively retained within the lineage for some function. Thus, one possible interpretation may be that the acquisition of the MocDEF oxygenase system allowed *P. syringae* to utilise alternative sources of carbon—such as BTX or rhizopine—in the woody niche, and that additional genes that confer additional selective advantages are being co-opted into this system as and when their acquisition occurs.

6.5.6 Divergence and short-term rate of evolution for E-*Pae*

We have analysed nucleotide diversity among six strains of E-*Pae* sampled from a diverse range of locations across northern Europe, and spanning approximately nine years. These strains differ at only 18 well-supported nucleotide sites across almost 6 Mb of aligned genome, representing an average pairwise divergence of about 1×10^{-6} substitutions per site. This estimate is approximately an order of magnitude higher than that found among four strains isolated from across the UK (Green et al., 2010), but nonetheless represents an extremely low level of divergence, and strongly suggests that European populations of the *aesculi* pathovar have arisen from a single and recent introduction into the continent.

We found no significant correlation between year of sampling and divergence for our E-*Pae* sequences, indicating a lack of time signal in these data that precluded an estimate of the approximate time to the most recent common ancestor (TMRCA) for E-*Pae*. Similarly, the lack of phylogeographical structure among these SNPs has meant that it was not possible to infer potential routes of transmission across Europe from these data. These findings are perhaps unsurprising, given the low sample size, the extremely low levels of divergence between samples, and the relatively large and approximate unit of time used for estimating sampling date (year). In addition, it is possible that mutations accumulated during laboratory passage may confound the relationship between time and divergence. Other bacterial species vary in the strength of association between time and divergence—for example, *Clostridium difficile* (He et al., 2013) and *Streptococcus pneumoniae* (Croucher et al., 2011) show good associations, whereas a recent study on the evolutionary origins of *Mycobacterium tuberculosis* found no correlation between time of sampling and divergence (Comas et al., 2013). Thus, it remains to be seen whether this method can be used to date the emergence of *P. syringae* lineages given a greater depth of sampling over longer time periods, or if further sampling may reveal a greater diversity that might allow for robust phylogeographical analyses.

There is a well-documented discrepancy between the estimated rate of nucleotide substitution and the timescale over which this rate is measured, the causes of which

are yet to be fully elucidated (Ho et al., 2005; Howell et al., 2008; Gibbs et al., 2010; Ho et al., 2011). This results in a negative exponential decrease in the relationship between the substitution rate and the TMRCA, such that rates estimated over short periods are often many orders of magnitude higher than those measured over longer, geological timescales. A recent study has suggested a log-linear relationship exists between the estimated rate of substitution and the TMRCA across different bacterial groups at differing levels of divergence (Comas et al., 2013). We elucidated the relationship between nucleotide substitution rate and divergence time for E-*Pae*, and found that it did not fit well with the prediction made by Comas et al.: given their model, the observed divergence among E-*Pae* strains implies a time since their shared ancestor of about 10^{-15} years ago, approximately equivalent to 5 nanoseconds—a result that is quite clearly incongruous with the evolutionary history of E-*Pae*.

Furthermore, we note an interesting caveat in the methodological approach used by Comas et al. (2013). These authors correlate substitution rates with the TMRCA for eight species; however the calculation of a given substitution rate uses the TMRCA estimate as the denominator (e.g., see EQUATION 6.1), and are thus non-independent: larger TMRCA values are necessarily associated with lower estimates of rate values since the response variable is divided by the predictor variable. On a log-log plot, this is always likely to yield a statistically significant association with a negative slope parameter even if divergence is in fact entirely random with respect to TMRCA. Thus, it is possible that simple variation in levels of divergence rates among different groups of bacteria, perhaps in turn driven by variation in mutation rates, may be driving spurious relationships between the rate of substitution and the divergence time when analysed across different bacterial species.

“ ‘We could have a conversation,’ said the mattress. ‘Would you like that?’

It was a large mattress, and probably one of quite high quality. Very few things actually get manufactured these days, because in an infinitely large universe such as, for instance, the one in which we live, most things one could possibly imagine and a lot of things one would rather not, grow somewhere. ”

Douglas Adams, *Life, the Universe and Everything*, 1982

Chapter 7

Perspectives

THIS THESIS is an investigation into the role of horizontal gene transfer in the evolution of the plant-associated bacterium *Pseudomonas syringae*. In this final chapter, I summarise and discuss the overall findings of this work, and place them into the wider context of what is known regarding both bacterial evolution generally and *P. syringae* evolution specifically. I also discuss potential additional work that would shed further light on the processes that impact prokaryotic evolution.

7.1 The role of HGT in *P. syringae* adaptation

7.1.1 The *P. syringae* genome is labile in both time and space

One of the major findings of this thesis regards the level of plasticity of the *P. syringae* genome. Our results show evidence for extensive variation in genome content both within and between *P. syrinage* lineages. Approximately 50% of genes in a given *P. syringae* genome are flexible—that is, show variable occurrence across lineages—while closely related (sister) strains may differ by up to hundreds of genes. Much of this variation is found at the tips of the phylogeny, comprising genes that are specific to individual lineages; the level of observable fluctuation decreases with respect to increasing phylogenetic depth. This pattern suggests a dynamism to genome fluctuation whereby the majority of gains have a short 'genome-life' within

the recipient lineage, and, once gained, are rapidly lost again. The structure and rapidity of gene turnover is such that patterns of gene presence and absence are in fact congruent with patterns of nucleotide substitutions within *P. syringae* core genes, despite any given snapshot of genome content evolution revealing extensive differences in gene repertoires among lineages. This suggests that, for these bacteria at least, genes that do persist within lineages are subsequently transmitted in a vertical, tree-like manner.

Generally, the patterns of genome fluctuation observed in *P. syringae* are consistent with those reported in other bacterial groups, such as *Streptococcus* (Lefébure and Stanhope, 2007), *Prochlorococcus* (Kettler et al., 2007), *Vibrio* (Shapiro et al., 2012) and *Escherichia* (Rocha, 2008; Touchon et al., 2009). A notable exception, however, is the congruence between phylogenies constructed from patterns of nucleotide substitution versus gene occurrence, as discussed above. The level of discordance between these two entities appears to vary depending on the species under scrutiny (e.g., Snel, 2001; Georgiades et al., 2011; Kettler et al., 2007; Shapiro et al., 2012; Nowell et al., 2014). Whether this variation is due simply to methodological differences between studies, or is reflective of the underlying ecology, biology or evolutionary history of the species under consideration, is a question that remains to be answered.

The quantification of gene gain and loss shows that individual *P. syringae* genomes may permute their gene repertoires many hundreds of times within the same timescale of a 1% divergence among core proteins. In addition, the pan genome of *P. syringae* is 'open', such that the analysis of further genomes is likely to continue to yield novel genetic variation. Furthermore, we show that genes of the core genome also recombine frequently, and calculate that recombination contributes approximately three times as much to nucleotide diversity than does point mutation. These results imply an important role for both homologous and non-homologous (illegitimate) recombination in *P. syringae* genome evolution, both of which can contribute to genome fluctuation. We note that the majority of *P. syringae* strains are likely to harbour plasmid sequences, and all are rich in a variety of transposable elements. Both of these genome properties are likely to contribute to the fluctuation

of *P. syringae* genomes: the gain and loss of plasmids, which may encode hundreds of genes, can result in rapid and drastic differences in gene content between lineages, while repeated elements distributed across genomes may facilitate homologous recombination and directly contribute to HGT when genes become embedded within transposable elements.

7.1.2 HGT and niche adaptation in *P. syringae*

We investigated the potential contribution from such high rates of gene gain and loss to the adaptive process in *P. syringae*. A total of 64 lineages of *P. syringae* were delimited into two groups based on a broad and relatively loosely defined ecological characteristic: the ability, or otherwise, to cause disease in the woody parts of a host plant species. We then applied a method that looks for the statistically supported correlated evolution between the presence of genes of the flexible genome and this ecological trait, using a method that was able to account for the confounding effects of phylogenetic relatedness among strains. We show that a substantial proportion of the flexible genome (approximately 7%, or 705 genes, at the type I error rate of 0.05) are significantly associated with the woody niche. Patterns of presence and absence are highly variable within this subset of 'woody-niche' genes, however; there are no specific loci which are perfectly associated with niche-type, or that are apparently unequivocally required for existence within either habitat. That patterns of association are between gene occurrence and niche are evident suggests a clear role for the flexible genome in the adaptation of *P. syringae* to the woody environment, and implicates both HGT and gene loss as the evolutionary mechanisms through which these patterns are mediated. However, what is also clear is that the adaptive process itself is 'messy': the convergent evolution of highly diverged *P. syringae* lineages onto woody hosts can occur via many different pathways of gene gains and losses. Divergent signatures of gene gain among woody-host pathovars may be driven by differences in the host species themselves, such that different sets of genes are required for pathogenicity on different host species. Alternatively, it may be the case that the stochastic nature of HGT has resulted in the invasion of the

woody niche being functionally underpinned by different sets of genes. Previous analyses on the evolution of highly diverged *P. syringae* lineages onto the hazel tree (*Corylus* sp.) have shown little evidence of genomic convergence (O'Brien et al., 2012); similarly, we find little evidence of replicated gene gains among divergent lineages that are adapted onto the same host, such as is the case for pv. *morsprunorum* strains present within PG1 and PG3. These observations would support the latter scenario, although further work is required to fully confirm this hypothesis.

Nonetheless, a closer inspection of the genomic adaptations that have occurred along the phylogenetic lineage leading to phylogroup 3 woody-host lineages shows there to be clear parallels in the evolution of two of the most destructive bacterial pathogens of woody plants to have emerged within recent times: bleeding canker of the European horse chestnut, caused by pv. *aesculi* (found within PG3), and bacterial canker of the kiwi, caused by pv. *actinidiae* (found within PG1). We confirm that the presence of a number of genes involved in the degradation of plant-derived aromatic compounds such as the wood-sugar D-xylose are found uniquely in the clades in which these pathogens reside, and highlight a number of genome rearrangement events that may be contributing to pathogenicity in these strains. Moreover, we show that these distantly related lineages show evidence for a much more recent exchange of genetic material at the loci that are directly implicated in the adaptation of these strains into the woody parts of their respective host species. This incontrovertibly demonstrates the importance of HGT and recombination in the process of adaptation in *P. syringae*, and its role in the evolution of pathogenicity in this species.

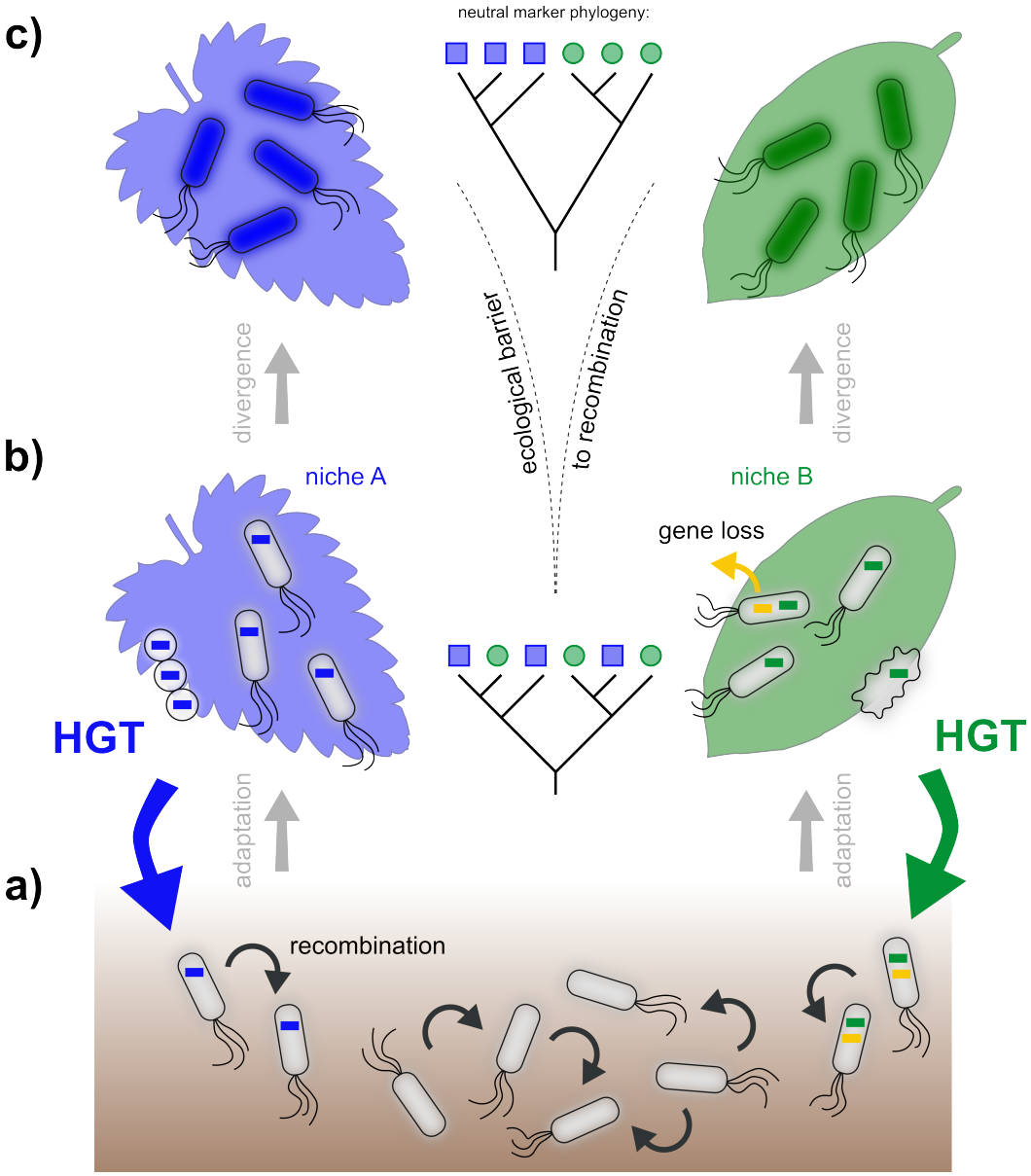
7.2 An emerging model of bacterial evolution?

Within book three of his five-part 'trilogy' *The Hitchhiker's Guide to the Galaxy*, Douglas Adams imagines a universe where, due to its infinite vastness, all conceivable forms of life (including talking mattresses) may be found somewhere—one merely has to find them (Adams, 1980). In some respects, Adams' vision holds true for emergent hypotheses regarding microbial evolutionary genomics: the impact of

extensive HGT and recombination within and among prokaryotic populations may be such that, over evolutionary timescales, it is possible for an extraordinarily large number of different combinations of genes to be trialed under different environmental conditions. In this sense, although rampant HGT and gene loss may appear a 'madness' of sorts, the prokaryotes (and, to a lesser extent, the eukaryotes also) are finding a method from it: these organisms utilise their genomic dynamism to quickly take advantage of new opportunities for adaptation as and when they arise. In this section, I re-examine an emergent model for the role of HGT and recombination in prokaryotic evolution that is synthesised from a number of hypotheses from the literature, and discuss how the results from the analysis of *P. syringae* add to this model.

Both Ochman et al. (2005) and more recently, Polz et al. (2013), have put forward models that consider the potential for HGT to actively promote bacterial diversification by its ability to facilitate differential adaptation. In FIG. 7.2, we show this divergence via differential adaptation model (DDA) as it might be applied to *P. syringae*. Here, plant-pathogenic lineages arise from a highly diverse and recombinogenic background population when they acquire, via HGT, niche-adaptive genes that are selectively advantageous in the environment of the host plant. These genes may encode virulence factors such as T3SEs or phytotoxic compounds, or genes that allow the bacterium to utilise other resources that are specific to a given plant environment. Importantly, the source of the niche-adaptive genes may have been a species that is already adapted to that environment by virtue of the presence of the

Figure 7.1 *(following page)*: Divergence via differential adaptation as applied to *P. syringae*. **(a)** Free-living and highly diverse populations of *P. syringae* exist in a multitude of environmental habitats, and recombine frequently via both homologous recombination and HGT. **(b)** The acquisition of niche-adaptive genes via HGT facilitates the movement of recipient lineages into different niche habitats, in this case represented by the gain of blue and green loci that are selectively advantageous on the plant hosts of the same colour. Genes that are no longer required in the new niche are deleted from the genome (e.g., the yellow locus). A phylogeny of neutral markers of the six lineages shown over the two plant habitats (blue and green) would not show genetic differentiation at this early stage. **(c)** The movement of these lineages onto their respective host plants results in the formation of a physical barrier to continued homologous recombination between sub-populations that is defined by the ecology of the plant hosts. Over time, this leads to an accumulation of nucleotide divergence between sub-populations and the formation of genetic clusters on the phylogeny.


c)
neutral marker phylogeny:
divergence
ecological barrier
to recombination
divergence
b)
niche A
niche B
gene loss
HGT
HGT
adaptation
adaptation
a)
recombination

very same genes—thus, high rates of gene turnover results in a fluidity of flexible genes such that it is the genes themselves that can be considered niche-specific, rather than either the donor or the recipient genomes (Polz et al., 2013). The movement of a given lineage into its plant niche results in the suppression of homologous recombination between lineages, initially due to physical barriers defined by the ecological context of the niche itself, then latterly also from the accumulation of nucleotide divergence, to which the rate of homologous recombination is inversely proportional (Majewski, 2001). Thus, over time, lineages in ecologically distinct niches will diverge as they continue to differentially adapt into their respective niche habitats, eventually forming genetically distinct units of diversity represented by clusters within a phylogenetic tree.

So how does *P. syringae* fit into this picture? Within the context of the adaptation of *P. syringae* into the woody niche, we show that: (1) the genomes of *P. syringae* are highly recombinogenic; (2) there are significant associations between gene gains and the woody niche; and (3) that despite this, the majority of transitions into the woody niche have occurred via the gain of different sets of genes: that is, lineages appear to be following different adaptive 'pathways' of gene acquisition to convergently evolve into a similar niche habitat. The patterns of genome dynamics seen for *P. syringae* are consistent with the model of divergence via differential adaptation; thus, we propose that an evolutionary scenario similar to the DDA model may be operating in this species, and can be used to further our understanding of the patterns of divergence and diversity observed in this important agricultural pathogen.

Overall, these ideas place the forces that influence genome fluctuation, namely HGT and gene deletion, at the very heart of the formation and maintenance of microbial population structure: far from being a destructive, homogenising force, HGT and gene loss can facilitate the evolution of genetically and ecologically cohesive units of diversity via processes such as the DDA model discussed above, and thus play central roles in microbial innovation, adaptation, diversification and ultimately speciation.

7.3 Future directions

This work illustrates the importance of HGT in the evolution of *P. syringae*, and thus may be considered an analysis of the contribution from the flexible genome towards evolutionary processes such as adaptation and differentiation in this species. However, the flexible genome is only half of the story, and it is likely that other evolutionary mechanisms, such as nucleotide substitutions, insertion/deletions and genome rearrangements, also play a role in the adaptation of *P. syringae* lineages to different environments. Thus, it remains to be seen whether *P. syringae* strains that are adapted to the woody niche also share variant sites among alleles of core genes that could also explain their convergence into this environment. The sharing of alleles may be driven by either recombination among shared genes or convergent evolution at individual nucleotide sites (i.e., homoplasy), and sites in significant association with the woody niche could be discovered using the genome-wide association (GWA) method that is commonly used in the analysis of human genetic diseases (e.g., Hirschhorn and Daly, 2005; Burton et al., 2007; Scott et al., 2007; McCarthy et al., 2008). The GWA approach has not been applied extensively to bacteria, primarily due to the problems associated with disentangling the confounding effect of phylogeny from that of causal variation (Falush and Bowden, 2006). However, a number of methods that analyse patterns in either the *k*-mer distribution (Sheppard et al., 2013) or phylogenetic information (Shapiro et al., 2012) of the core genome have recently arisen, and have been applied to the evolution of pathogenicity in *Campylobacter* and early events in the differentiation of populations of *Vibrio* respectively.

These methods may be applicable to the analysis of niche adaptation in *P. syringae*, and, in addition to the potential discovery of core genes that may be functionally important for survival in the woody environment, would allow for an estimation of the relative contribution from the flexible versus the core genome towards adaptation of *P. syringae* to the woody niche. The elucidation of the relative importance of these genome components would provide a fascinating insight into the genome dynamics of adaptation in *P. syringae*, and would add an important facet to our understanding

of the evolutionary process in bacterial species.

Bibliography

ADAMS, D. 1980. Life, the Universe and Everything. The Hitchhiker's Guide to the Galaxy. Pan Macmillan.

ALIKHAN, N.-F., PETTY, N., BEN ZAKOUR, N., AND BEATSON, S. 2011. BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics* 12:402.

ALMEIDA, N. F., YAN, S., LINDEBERG, M., STUDHOLME, D. J., SCHNEIDER, D. J., CONDON, B., LIU, H., VIANA, C. J., WARREN, A., EVANS, C., KEMEN, E., MACLEAN, D., ANGOT, A., MARTIN, G. B., JONES, J. D., COLLMER, A., SETUBAL, J. C., AND VINATZER, B. A. 2009. A draft genome sequence of *Pseudomonas syringae* pv. *tomato* T1 reveals a type III effector repertoire significantly divergent from that of *Pseudomonas syringae* pv. *tomato* DC3000. *Molecular Plant-Microbe Interactions* 22:52–62.

ALTSCHUL, S. F., MADDEN, T. L., SCHÄFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., AND LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

ANDERSSON, J. O., SJÖGREN, Å. M., DAVIS, L. A., EMBLEY, T. M., AND ROGER, A. J. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Current Biology* 13:94–104.

ANDERSSON, S. AND KURLAND, C. 1990. Codon preferences in free-living microorganisms. *Microbiological Reviews* 54:198–210.

ANDERSSON, S. G. AND SHARP, P. M. 1996a. Codon usage and base composition in *Rickettsia prowazekii*. *Journal of Molecular Evolution* 42:525–536.

ANDERSSON, S. G. AND SHARP, P. M. 1996b. Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* 142:915–925.

ANDREWS, S. 2010. FastQC: A quality control tool for high throughput sequence data. Available at: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

ANGIUOLI, S. V. AND SALZBERG, S. L. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27:334–342.

ARREBOLA, E., CAZORLA, F. M., DURÁN, V. E., RIVERA, E., OLEA, F., CODINA, J. C., PÉREZ-GARCIA, A., AND DE VICENTE, A. 2003. Mangotoxin: A novel antimetabolite toxin produced by *Pseudomonas syringae* inhibiting ornithine/arginine biosynthesis. *Physiological and Molecular Plant Pathology* 63:117–127.

ARREBOLA, E., CAZORLA, F. M., PÉREZ-GARCÍA, A., AND VICENTE, A. D. 2011a. Chemical and metabolic aspects of antimetabolite toxins produced by *Pseudomonas syringae* pathovars. *Toxins* 3:1089–1110.

ARREBOLA, E., CAZORLA, F. M., PÉREZ-GARCÍA, A., AND VICENTE, A. D. 2011b. Genes involved in the production of antimetabolite toxins by *Pseudomonas syringae* pathovars. *Genes* 2:640–660.

AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D., THIBAULT, J., ET AL. 2013. From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* pp. 11–10.

AZIZ, R. K., BARTELS, D., BEST, A. A., DEJONGH, M., DISZ, T., EDWARDS, R. A., FORMSMA, K., GERDES, S., GLASS, E. M., KUBAL, M., MEYER, F., OLSEN, G. J., OLSON, R., OSTERMAN, A. L., OVERBEEK, R. A., MCNEIL, L. K., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G. D., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A., AND ZAGNITKO, O. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9:75.

BAHAR, M., DE MAJNIK, J., SAINT, C. P., AND MURPHY, P. J. 2000. Conservation of a Pseudomonad-like hydrocarbon degradative ferredoxin oxygenase complex involved in rhizopine catabolism in *Sinorhizobium meliloti* and *Rhizobium leguminosarum* bv. *viciae*. *Journal of Molecular Microbiology and Biotechnology* 2:257–259.

BAHAR, M., DE MAJNIK, J., WEXLER, M., FRY, J., POOLE, P. S., AND MURPHY, P. J. 1998. A model for the catabolism of rhizopine in *Rhizobium leguminosarum* involves a ferredoxin oxygenase complex and the inositol degradative pathway. *Molecular Plant-Microbe Interactions* 11:1057–1068.

BAKER, M. 2012. *De novo* genome assembly: What every biologist should know. *Nature Methods* 9:333–337.

BALESTRA, G. M., RENZI, M., AND MAZZAGLIA, A. 2010. First report of bacterial canker of *Actinidia deliciosa* caused by *Pseudomonas syringae* pv. *actinidiae* in Portugal. *New Disease Reports* 22:10.

BALTRUS, D. A., DOUGHERTY, K., BECKSTROM-STERNBERG, S. M., BECKSTROM-STERNBERG, J. S., AND FOSTER, J. T. 2014. Incongruence between multi-locus sequence analysis (MLSA) and whole-genome-based phylogenies: *Pseudomonas syringae* pathovar *pisi* as a cautionary tale. *Molecular Plant Pathology* 15:461–465.

BALTRUS, D. A., NISHIMURA, M. T., DOUGHERTY, K. M., BISWAS, S., MUKHTAR, M. S., VICENTE, J., HOLUB, E. B., AND DANGL, J. L. 2012. The molecular basis of host specialization in bean pathovars of *Pseudomonas syringae*. *Molecular Plant-Microbe Interactions* 25:877–888.

BALTRUS, D. A., NISHIMURA, M. T., ROMANCHUK, A., CHANG, J., MUKHTAR, M., CHERKIS, K., ROACH, J., GRANT, S. R., JONES, C. D., AND DANGL, J. L. 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. *PLoS Pathogens* 7:e1002132.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A., AND PEVZNER, P. A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19:455–477.

BARDAJI, L., AÑORGA, M., JACKSON, R. W., MARTÍNEZ-BILBAO, A., YANGUAS-CASÁS, N., AND MURILLO, J. 2011a. Miniature transposable sequences are frequently mobilized in the bacterial plant pathogen *Pseudomonas syringae* pv. *phaseolicola*. *PLoS ONE* 6:e25773.

BARDAJI, L., PÉREZ-MARTÍNEZ, I., RODRÍGUEZ-MORENO, L., RODRÍGUEZ-PALENZUELA, P., SUNDIN, G. W., RAMOS, C., AND MURILLO, J. 2011b. Sequence and role in virulence of the three plasmid complement of the model tumor-inducing bacterium *Pseudomonas savastanoi* pv. *savastanoi* NCPPB 3335. *PLoS ONE* 6:e25705.

BARDOUX, S. AND ROUSSEAU, P. 2007. Le dépérissement bactérien du marronnier. *Phytoma – La Défense des Végétaux* 605:22–23.

BARKER, D. AND PAGEL, M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology* 1:e3.

BARNETT, D. W., GARRISON, E. K., QUINLAN, A. R., STRÖMBERG, M. P., AND MARTH, G. T. 2011. BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27:1691–1692.

BARRACLOUGH, T. G., BALBI, K. J., AND ELLIS, R. J. 2012. Evolving concepts of bacterial species. *Evolutionary Biology* 39:148–157.

BARTOLI, C., LAMICHHANE, J. R., BERGE, O., GUILBAUD, C., VARVARO, L., BALESTRA, G. M., VINATZER, B. A., AND MORRIS, C. E. 2014. A framework to gauge the epidemic potential of plant pathogens in environmental reservoirs: the example of kiwifruit canker. In press.

BAUER, D. W. AND COLLMER, A. 1997. Molecular cloning, characterization, and mutagenesis of a *pel* gene from *Pseudomonas syringae* pv. *lachrymans* encoding a member of the *Erwinia chrysanthemi* PelADE family of pectate lyases. *Molecular Plant-Microbe Interactions* 10:369–379.

BENDER, C. L., ALARCÓN-CHAIDEZ, F., AND GROSS, D. C. 1999. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiology and Molecular Biology Reviews* 63:266–292.

BERGE, O., MONTEIL, C. L., BARTOLI, C., CHANDEYSSON, C., GUILBAUD, C., SANDS, D. C., AND MORRIS, C. E. 2014. A user's guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. *PloS ONE* 9:e105547.

BETTIGA, M., HÄHN-HAGERDAL, B., AND GORWA-GRAUSLUND, M. F. 2008. Comparing the xylose reductase/xylitol dehydrogenase and xylose isomerase pathways in arabinose and xylose fermenting *Saccharomyces cerevisiae* strains. *Biotechnology for Biofuels* 1:16–16.

BINNEWIES, T. T., MOTRO, Y., HALLIN, P. F., LUND, O., DUNN, D., LA, T., HAMPSON, D. J., BELLGARD, M., WASSENAAR, T. M., AND USSERY, D. W. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & Integrative Genomics* 6:165–185.

BLOCK, A. AND ALFANO, J. R. 2011. Plant targets for *Pseudomonas syringae* type III effectors: virulence targets or guarded decoys? *Current Opinion in Microbiology* pp. 1–8.

BODEY, G. P., BOLIVAR, R., FAINSTEIN, V., AND JADEJA, L. 1983. Infections caused by *Pseudomonas aeruginosa*. *Review of Infectious Diseases* 5:279–313.

BOISVERT, S., LAVIOLETTE, F., AND CORBEIL, J. 2010. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology* 17:1519–1533.

BOLLER, T. AND FELIX, G. 2009. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology* 60:379–406.

BOS, K. I., SCHUENEMANN, V. J., GOLDING, G. B., BURBANO, H. A., WAGLECHNER, N., COOMBES, B. K., MCPHEE, J. B., DEWITTE, S. N., MEYER, M., SCHMEDES, S., ET AL. 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478:506–510.

BOTO, L. 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences* 277:819–827.

BOUCHER, Y., CORDERO, O. X., TAKEMURA, A., HUNT, D. E., SCHLIEP, K., BAPTESTE, E., LOPEZ, P., TARR, C. L., AND POLZ, M. F. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global vibrio cholerae populations. *mBio* 2:e00335–10–e00335–10.

BOUCHER, Y., DOUADY, C. J., PAPKE, R. T., WALSH, D. A., BOUDREAU, M. E. R., NESBØ, C. L., CASE, R. J., AND DOOLITTLE, W. F. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics* 37:283–328.

BOWATER, R. AND DOHERTY, A. J. 2006. Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genetics* 2:e8.

BREMER, B., BREMER, K., CHASE, M., FAY, M., REVEAL, J., SOLTIS, D., SOLTIS, P., AND STEVENS, P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* .

BRODEY, C. L., RAINEY, P. B., TESTER, M., AND JOHNSTONE, K. 1991. Bacterial blotch disease of the cultivated mushroom is caused by an ion channel forming lipodepsipeptide toxin. *Molecular Plant–Microbe Interactions* 4:407–411.

BUELL, C. R., JOARDAR, V., LINDEBERG, M., SELENGUT, J., PAULSEN, I. T., GWINN, M. L., DODSON, R. J., DEBOY, R. T., DURKIN, A. S., KOLONAY, J. F., MADUPU, R., DAUGHERTY, S., BRINKAC, L., BEANAN, M. J., HAFT, D. H., NELSON, W. C., DAVIDSEN, T., ZAFAR, N., ZHOU, L., LIU, J., YUAN, Q., KHOURI, H., FEDOROVA, N., TRAN, B., RUSSELL, D., BERRY, K., UTTERBACK, T., VAN AKEN, S. E., FELDBLYUM, T. V., D'ASCENZO, M., DENG, W.-L., RAMOS, A. R., ALFANO, J. R., CARTINHOUR, S., CHATTERJEE, A. K., DELANEY, T. P., LAZAROWITZ, S. G., MARTIN, G. B., SCHNEIDER, D. J., TANG, X., BENDER, C. L., WHITE, O., FRASER, C. M., AND COLLMER, A. 2003. The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Proceedings of the National Academy of Sciences* 100:10181–10186.

BULL, C., MANCEAU, C., LYDON, J., KONG, H., VINATZER, B. A., AND FISCHER-LE SAUX, M. 2010. Comprehensive list of names of plant pathogenic bacteria. *Systematic Applied Microbiology* 33:105–115.

BULMER, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

BULTREYS, A., GHEYSEN, I., AND PLANCHON, V. 2008. Characterization of *Pseudomonas syringae* strains isolated from diseased horse-chestnut trees in Belgium, pp. 283–293. *In Pseudomonas syringae* Pathovars and Related Pathogens–Identification, Epidemiology and Genomics. Springer.

BURLEIGH, J. G., HILU, K. W., AND SOLTIS, D. E. 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. *BMC Evolutionary Biology* 9:61.

BURTON, P. R., CLAYTON, D. G., CARDON, L. R., CRADDOCK, N., DELOUKAS, P., DUNCANSON, A., KWIATKOWSKI, D. P., MCCARTHY, M. I., OUWEHAND, W. H., SAMANI, N. J., ET AL. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678.

CADILLO-QUIROZ, H., DIDELOT, X., HELD, N. L., HERRERA, A., DARLING, A., RENO, M. L., KRAUSE, D. J., AND WHITAKER, R. J. 2012. Patterns of gene flow define species of thermophilic archaea. *PLoS Biology* 10:e1001265.

CAI, R., LEWIS, J., YAN, S., LIU, H., CLARKE, C. R., CAMPANILE, F., ALMEIDA, N. F., STUDHOLME, D. J., LINDEBERG, M., SCHNEIDER, D., ZACCARDELLI, M., SETUBAL, J. C., MORALES-LIZCANO, N. P., BERNAL, A., COAKER, G., BAKER, C., BENDER, C. L., LEMAN, S., AND VINATZER, B. A. 2011. The plant pathogen *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection to evade tomato immunity. *PLoS Pathogens* 7:e1002130.

CARO-QUINTERO, A. AND KONSTANTINIDIS, K. T. 2011. Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* 14:347–355.

CHARIF, D. AND LOBRY, J. R. 2007. SeqinR 1.0-2: A contributed package to the R Project for statistical computing devoted to biological sequences retrieval and analysis, pp. 207–232. *In* Structural Approaches to Sequence Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg.

CHARIF, D., THIOULOUSE, J., LOBRY, J., AND PERRIÈRE, G. 2005. Online synonymous codon usage analyses with the ADE4 and SEQINR packages. *Bioinformatics* 21:545–547.

CHEN, L. H., KENYON, G., CURTIN, F., HARAYAMA, S., BEMBENEK, M., HAJIPOUR, G., AND WHITMAN, C. 1992. 4-oxalocrotonate tautomerase, an enzyme composed of 62 amino acid residues per monomer. *Journal of Biological Chemistry* 267:17716–17721.

CHEN, S. L., LEE, W., HOTTES, A. K., SHAPIRO, L., AND MCADAMS, H. H. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences* 101:3480–3485.

CICCARELLI, F. D., DOERKS, T., VON MERING, C., CREEVEY, C. J., SNEL, B., AND BORK, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.

CINGOLANI, P., PLATTS, A., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X., AND RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6:80–92.

CLARKE, C. R., CAI, R., STUDHOLME, D. J., GUTTMAN, D. S., AND VINATZER, B. A. 2010. *Pseudomonas syringae* strains naturally lacking the classical *P. syringae hrp/hrc* locus are common leaf colonizers equipped with an atypical type III secretion system. *Molecular Plant-Microbe Interactions* 23:198–210.

COHAN, F. M. 2002. What are bacterial species? *Annual Reviews in Microbiology* 56:457–487.

COHEN, O., ASHKENAZY, H., BELINKY, F., HUCHON, D., AND PUPKO, T. 2010. GLOOME: Gain loss mapping engine. *Bioinformatics* 26:2914–2915.

COHEN, O. AND PUPKO, T. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Molecular Biology and Evolution* 27:703–713.

COHEN, O. AND PUPKO, T. 2011. Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony—a simulation study. *Genome Biology and Evolution* 3:1265–1275.

COMAS, I., COSCOLLA, M., LUO, T., BORRELL, S., HOLT, K. E., KATO-MAEDA, M., PARKHILL, J., MALLA, B., BERG, S., THWAITES, G., ET AL. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genetics* 45:1176–1182.

CONESA, A., GOTZ, S., GARCIA-GOMEZ, J. M., TEROL, J., TALON, M., AND ROBLES, M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676.

CORNELIS, G. R. AND VAN GIJSEGEM, F. 2000. Assembly and function of type III secretory systems. *Annual Reviews in Microbiology* 54:735–774.

CROUCHER, N. J., HARRIS, S. R., FRASER, C., QUAIL, M. A., BURTON, J., VAN DER LINDEN, M., MCGEE, L., VON GOTTBERG, A., SONG, J. H., KO, K. S., ET AL. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* 331:430–434.

CUNNAC, S., LINDEBERG, M., AND COLLMER, A. 2009. *Pseudomonas syringae* type III secretion system effectors: repertoires in search of functions. *Current Opinion in Microbiology* 12:53–60.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., ET AL. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

DAUBIN, V., MORAN, N. A., AND OCHMAN, H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* 301:829–832.

DAVISON, J. 1999. Genetic exchange between bacteria in the environment. *Plasmid* 42:73–91.

DE KEIJZER, J., VAN DEN BROEK, L. A., KETELAAR, T., AND VAN LAMMEREN, A. A. 2012. Histological examination of horse chestnut infection by *Pseudomonas syringae* pv. *aesculi* and non-destructive heat treatment to stop disease progression. *PloS ONE* 7:e39604.

DELLEDONNE, M., XIA, Y., DIXON, R. A., AND LAMB, C. 1998. Nitric oxide functions as a signal in plant disease resistance. *Nature* 394:585–588.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., ET AL. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43:491–498.

DIDELOT, X. AND FALUSH, D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.

DONG, H., NILSSON, L., AND KURLAND, C. G. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *Journal of Molecular Biology* 260:649–663.

DOOLITTLE, W. F. 2000. Uprooting the tree of life. *Scientific American* 282:90–95.

DOOLITTLE, W. F. 2012. Population genomics: How bacterial species form and why they don't exist. *Current Biology* 22:R451–R453.

DOOLITTLE, W. F. AND PAPKE, R. T. 2006. Genomics and the bacterial species problem. *Genome Biology* 7:116.

DOOLITTLE, W. F. AND ZHAXYBAYEVA, O. 2009. On the origin of prokaryotic species. *Genome Research* 19:744–756.

DUDNIK, A. AND DUDLER, R. 2014. Genome and transcriptome sequences of *Pseudomonas syringae* pv. *syringae* B301D-R. *Genome Announcements* 2:e00306–14.

DURGAPAL, J. 1971. A preliminary note on bacterial diseases of temperate plants in India. *Indian Phytopathology* 24:379–382.

DURGAPAL, J. AND SINGH, B. 1980. Taxonomy of pseudomonads pathogenic to horsechestnut, wild fig and wild cherry in India. *Indian Phytopathology* 33:533–535.

EARL, D., BRADNAM, K., ST JOHN, J., DARLING, A., LIN, D., FASS, J., YU, H. O. K., BUFFALO, V., ZERBINO, D. R., DIEKHANS, M., NGUYEN, N., ARIYARATNE, P. N., SUNG, W. K., NING, Z., HAIMEL, M., SIMPSON, J. T., FONSECA, N. A., BIROL, I., DOCKING, T. R., HO, I. Y., ROKHSAR, D. S., CHIKHI, R., LAVENIER, D., CHAPUIS, G., NAQUIN, D., MAILLET, N., SCHATZ, M. C., KELLEY, D. R., PHILLIPPY, A. M., KOREN, S., YANG, S. P., WU, W., CHOU, W. C., SRIVASTAVA, A., SHAW, T. I., RUBY, J. G., SKEWES-COX, P., BETEGON, M., DIMON, M. T., SOLOVYEV, V., SELEDTSOV, I., KOSAREV, P., VOROBYEV, D., RAMIREZ-GONZALEZ, R., LEGGETT, R., MACLEAN, D., XIA, F., LUO, R., LI, Z., XIE, Y., LIU, B., GNERRE, S., MACCALLUM, I., PRZYBYLSKI, D., RIBEIRO, F. J., YIN, S., SHARPE, T., HALL, G., KERSEY, P. J., DURBIN, R., JACKMAN, S. D., CHAPMAN, J. A., HUANG, X., DERISI, J. L., CACCAMO, M., LI, Y., JAFFE, D. B., GREEN, R. E., HAUSSLER, D., KORF, I., AND PATEN, B. 2011. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Research* 21:2224–2241.

EPPO 2011. *European Plant Protection Organisation Reporting Service*. European Plant Protection Organisation, Paris, France.

EPPO 2013. *European Plant Protection Organisation Reporting Service*. European Plant Protection Organisation, Paris, France.

EZRA, D., BARASH, I., VALINSKY, L., AND MANULIS, S. 2000. The dual function in virulence and host range restriction of a gene isolated from the pPATH$_{Ehg}$ plasmid of *Erwinia herbicola* pv. *gypsophilae*. *Molecular Plant-Microbe Interactions* 13:683–692.

EZRA, D., BARASH, I., WEINTHAL, D. M., GABA, V., AND MANULIS, S. 2004. *pthG* from *Pantoea agglomerans* pv. *gypsophilae* encodes an avirulence effector that determines incompatibility in multiple beet species. *Molecular Plant Pathology* 5:105–113.

FALUSH, D. AND BOWDEN, R. 2006. Genome-wide association mapping in bacteria? *Trends in Microbiology* 14:353–355.

FEIL, E. J., HOLMES, E. C., BESSEN, D. E., CHAN, M.-S., DAY, N. P., ENRIGHT, M. C., GOLDSTEIN, R., HOOD, D. W., KALIA, A., MOORE, C. E., ET AL. 2001a. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences* 98:182–187.

FEIL, E. J., HOLMES, E. C., BESSEN, D. E., CHAN, M. S., DAY, N. P., ENRIGHT, M. C., GOLDSTEIN, R., HOOD, D. W., KALIA, A., MOORE, C. E., ZHOU, J., AND SPRATT, B. G. 2001b. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences* 98:182–187.

FEIL, E. J., MAIDEN, M., ACHTMAN, M., AND SPRATT, B. G. 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution* 16:1496–1502.

FEIL, H., FEIL, W. S., CHAIN, P., LARIMER, F., DIBARTOLO, G., COPELAND, A., LYKIDIS, A., TRONG, S., NOLAN, M., GOLTSMAN, E., THIEL, J., MALFATTI, S., LOPER, J. E., LAPIDUS, A., DETTER, J. C., LAND, M., RICHARDSON, P. M., KYRPIDES, N. C., IVANOVA, N., AND LINDOW, S. E. 2005. Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. *Proceedings of the National Academy of Sciences* 102:11064–11069.

FELSENSTEIN, J. 1993. {PHYLIP}: phylogenetic inference package, version 3.5 c.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J.-F., DOUGHERTY, B. A., MERRICK, J. M., ET AL. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science* 269:496–512.

FORESTRY COMMISSION 2008. *Report on the national survey to assess the presence of bleeding canker disease of horse chestnut trees in Great Britain*. Forestry Commission, Edinburgh, UK.

FRASER, C., ALM, E. J., POLZ, M. F., SPRATT, B. G., AND HANAGE, W. P. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323:741–746.

FRIEDMAN, J. AND BARRETT, S. C. 2008. A phylogenetic analysis of the evolution of wind pollination in the angiosperms. *International Journal of Plant Sciences* 169:49–58.

GARDAN, L., SHAFIK, H., BELOUIN, S., BROCH, R., GRIMONT, F., AND GRIMONT, P. 1999. DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (ex Sutic and Dowson 1959). *International Journal of Systematic Bacteriology* 49:469–478.

GARDINER, D. M., STILLER, J., COVARELLI, L., LINDEBERG, M., SHIVAS, R. G., AND MANNERS, J. M. 2013. Genome sequences of *Pseudomonas* spp. isolated from cereal crops. *Genome Announcements* 1:e00209–13.

GARRISON, E. AND MARTH, G. 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints* 1207:3907.

GEORGIADES, K., MERHEJ, V., EL KARKOURI, K., RAOULT, D., AND PONTAROTTI, P. 2011. Gene gain and loss events in *Rickettsia* and *Orientia* species. *Biology Direct* 6:6.

GIBBS, A., FARGETTE, D., GARCIA-ARENAL, F., AND GIBBS, M. 2010. Time–the emerging dimension of plant virus studies. *Journal of General Virology* 91:13–22.

GOGARTEN, J. P., DOOLITTLE, W. F., AND LAWRENCE, J. G. 2002. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* 19:2226–2238.

GOLD, T. 1992. The deep, hot biosphere. *Proceedings of the National Academy of Sciences* 89:6045–6049.

GOPHNA, U., CHARLEBOIS, R. L., AND DOOLITTLE, W. F. 2004. Have archaeal genes contributed to bacterial virulence? *Trends in Microbiology* 12:213–219.

GOUY, M. AND GAUTIER, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* 10:7055–7074.

GOUY, M., GAUTIER, C., ATTIMONELLI, M., LANAVE, C., AND DI PAOLA, G. 1985. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Computer Applications in the Biosciences* 1:167–172.

GRANTHAM, R., GAUTIER, C., GOUY, M., MERCIER, R., AND PAVE, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research* 8:197–197.

GREEN, S., LAUE, B., FOSSDAL, C. G., A'HARA, S. W., AND COTTRELL, J. E. 2009. Infection of horse chestnut (*Aesculus hippocastanum*) by *Pseudomonas syringae* pv. *aesculi* and its detection by quantitative real-time PCR. *Plant Pathology* 58:731–744.

GREEN, S., STUDHOLME, D. J., LAUE, B. E., DORATI, F., LOVELL, H., ARNOLD, D., COTTRELL, J. E., BRIDGETT, S., BLAXTER, M., HUITEMA, E., THWAITES, R., SHARP, P. M., JACKSON, R. W., AND KAMOUN, S. 2010. Comparative genome analysis provides insights into the evolution and adaptation of *Pseudomonas syringae* pv. *aesculi* on *Aesculus hippocastanum*. *PLoS ONE* 5:e10224.

GREENACRE, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London.

GRIGORIEV, A. 1998. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research* 26:2286–2290.

GROCOCK, R. J. AND SHARP, P. M. 2002. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* 289:131–139.

GROSSE, C., SCHERER, J., KOCH, D., OTTO, M., TAUDTE, N., AND GRASS, G. 2006. A new ferrous iron-uptake transporter, EfeU (YcdN), from *Escherichia coli*. *Molecular Microbiology* 62:120–131.

GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W., AND GASCUEL, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307–321.

GUINDON, S. AND GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52:696–704.

HANAGE, W. P. 2013. Fuzzy species revisited. *BMC biology* 11:41.

HANAGE, W. P., FRASER, C., AND SPRATT, B. G. 2005. Fuzzy species among recombinogenic bacteria. *BMC biology* 3:6.

HANAGE, W. P., FRASER, C., AND SPRATT, B. G. 2006. Sequences, sequence clusters and bacterial species. *Philosophical Transactions of the Royal Society B: Biological Sciences* 361:1917–1927.

HAO, W. AND GOLDING, G. B. 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research* 16:636–643.

HAO, W. AND GOLDING, G. B. 2008. High rates of lateral gene transfer are not due to false diagnosis of gene absence. *Gene* 421:27–31.

HARRIS, S. R., FEIL, E. J., HOLDEN, M. T., QUAIL, M. A., NICKERSON, E. K., CHANTRATITA, N., GARDETE, S., TAVARES, A., DAY, N., LINDSAY, J. A., ET AL. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474.

HAYDEN, E. C. 2014. Technology: the $1,000 genome. *Nature* 507:294–5.

HE, M., MIYAJIMA, F., ROBERTS, P., ELLISON, L., PICKARD, D. J., MARTIN, M. J., CONNOR, T. R., HARRIS, S. R., FAIRLEY, D., BAMFORD, K. B., ET AL. 2013. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nature Genetics* 45:109–113.

HEAPS, H. S. 1978. Information retrieval: Computational and theoretical aspects. Academic Press, Inc.

HENRY, I. AND SHARP, P. M. 2007. Predicting gene expression level from codon usage bias. *Molecular Biology and Evolution* 24:10–12.

HILLER, N. L., JANTO, B., HOGG, J. S., BOISSY, R., YU, S., POWELL, E., KEEFE, R., EHRLICH, N. E., SHEN, K., HAYES, J., BARBADORA, K., KLIMKE, W., DERNOVOY, D., TATUSOVA, T., PARKHILL, J., BENTLEY, S. D., POST, J. C., EHRLICH, G. D., AND HU, F. Z. 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: Insights into the pneumococcal supragenome. *Journal of Bacteriology* 189:8186–8195.

HIRANO, S. AND UPPER, C. 1990. Population biology and epidemiology of *Pseudomonas syringae*. *Annual Review of Phytopathology* 28:155–177.

HIRANO, S. AND UPPER, C. 1995. Ecology of ice nucleation-active bacteria. *Biological ice nucleation and its applications* pp. 41–61.

HIRSCHHORN, J. N. AND DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6:95–108.

HO, S. Y., LANFEAR, R., BROMHAM, L., PHILLIPS, M. J., SOUBRIER, J., RODRIGO, A. G., AND COOPER, A. 2011. Time-dependent rates of molecular evolution. *Molecular Ecology* 20:3087–3101.

HO, S. Y., PHILLIPS, M. J., COOPER, A., AND DRUMMOND, A. J. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular biology and evolution* 22:1561–1568.

HÖFTE, M. AND VOS, P. 2006. Plant pathogenic *Pseudomonas* species. *Plant-Associated Bacteria* pp. 507–533.

HOGG, J. S., HU, F. Z., JANTO, B., BOISSY, R., HAYES, J., KEEFE, R., POST, J. C., AND EHRLICH, G. D. 2007. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biology* 8:R103.

HOLT, K. E., BAKER, S., WEILL, F.-X., HOLMES, E. C., KITCHEN, A., YU, J., SANGAL, V., BROWN, D. J., COIA, J. E., KIM, D. W., ET AL. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics* 44:1056–1059.

HOWARD, A. L. 1945. The horse-chestnut tree (*Aesculus hippocastanum*). *Nature* 155:521–522.

HOWELL, N., HOWELL, C., AND ELSON, J. L. 2008. Time dependency of molecular rate estimates for mtdna: this is not the time for wishful thinking. *Heredity* 101:107–108.

HUNTER, P. J. AND TAYLOR, J. D. 2006. Patterns of interaction between isolates of three pathovars of *Pseudomonas syringae* and accessions of a range of host and nonhost legume species. *Plant Pathology* 55:46–53.

HWANG, M., MORGAN, R. L., SARKAR, S. F., WANG, P. W., AND GUTTMAN, D. S. 2005. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. *Applied and Environmental Microbiology* 71:5182.

IKEMURA, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology* 151:389–409.

IKEMURA, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* 2:13–34.

JANSE, J. D., ROSSI, P., ANGELUCCI, L., SCORTICHINI, M., DERKS, J. H. J., AKKERMANS, A. D. L., DE VRIJER, R., AND PSALLIDAS, P. G. 1996. Reclassification of *Pseudomonas syringae* pv. *avellanae* as *Pseudomonas avellanae* (spec. nov, the bacterium causing canker of Hazelnut (*Corylus avellana* L.). *Systematic Applied Microbiology* 19:589–595.

JEFFRIES, T. W. 1983. Utilization of xylose by bacteria, yeasts, and fungi. Springer.

JEON, C. O., PARK, W., GHIORSE, W. C., AND MADSEN, E. L. 2004. *Polaromonas naphthalenivorans* sp. nov., a naphthalene-degrading bacterium from naphthalene-contaminated sediment. *International Journal of Systematic and Evolutionary Microbiology* 54:93–97.

JI, C., BOYD, C., SLAYMAKER, D., OKINAKA, Y., TAKEUCHI, Y., MIDLAND, S. L., SIMS, J. J., HERMAN, E., AND KEEN, N. 1998. Characterization of a 34-kda soybean binding protein for the syringolide elicitors. *Proceedings of the National Academy of Sciences* 95:3306–3311.

JOARDAR, V., LINDEBERG, M., JACKSON, R. W., SELENGUT, J., DODSON, R., BRINKAC, L., DAUGHERTY, S., DEBOY, R., DURKIN, A., AND GIGLIO, M. 2005. Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. *Journal of Bacteriology* 187:6488.

JONES, J. D. AND DANGL, J. L. 2006. The plant immune system. *Nature* 444:323–329.

KAMIUNTEN, H., NAKAO, T., AND OSHIDA, S. 2000. *Pseudomonas syringae* pv. *cerasicola*, pv. nov., the causal agent of bacterial gall of cherry tree. *Journal of General Plant Pathology* 66:219–224.

KANAYA, S., YAMADA, Y., KUDO, Y., AND IKEMURA, T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238:143–155.

KANDYALA, R., RAGHAVENDRA, S. P. C., AND RAJASEKHARAN, S. T. 2010. Xylene: An overview of its health hazards and preventive measures. *Journal of Oral and Maxillofacial Pathology* 14:1.

KANEHISA, M. AND GOTO, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28:27–30.

KATZ, L. A. 1999. The tangled web: Gene genealogies and the origin of Eukaryotes. *The American Naturalist* 154:S137–S145.

KAZEMPOUR, N. M., KAMRAN, E., AND TALESHI, S. S. 2006. An identification of the bacterial agents of mulberry leaf blight, guilan province. *Iranian Journal of Agricultural Sciences* .

KEEN, N., TAMAKI, S., KOBAYASHI, D., GERHOLD, D., STAYTON, M., SHEN, H., GOLD, S., LORANG, J., THORDAL-CHRISTENSEN, H., DAHLBECK, D., ET AL. 1990. Bacteria expressing avirulence gene d produce a specific elicitor of the soybean hypersensitive reaction. *Molecular Plant-Microbe Interactions* 3:112–121.

KENDALL, M. M., GRUBER, C. C., RASKO, D. A., HUGHES, D. T., AND SPERANDIO, V. 2011. Hfq virulence regulation in enterohemorrhagic *Escherichia coli* o157: H7 strain 86-24. *Journal of Bacteriology* 193:6843–6851.

KENNEMANN, L., DIDELOT, X., AEBISCHER, T., KUHN, S., DRESCHER, B., DROEGE, M., REINHARDT, R., CORREA, P., MEYER, T. F., JOSENHANS, C., ET AL. 2011. *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences* 108:5033–5038.

KETTLER, G. C., MARTINY, A. C., HUANG, K., ZUCKER, J., COLEMAN, M. L., RODRIGUE, S., CHEN, F., LAPIDUS, A., FERRIERA, S., JOHNSON, J., STEGLICH, C., CHURCH, G. M., RICHARDSON, P., AND CHISHOLM, S. W. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics* 3:e231.

KIM, K.-S. AND FARRAND, S. K. 1996. Ti plasmid-encoded genes responsible for catabolism of the crown gall opine mannopine by *Agrobacterium tumefaciens* are homologs of the T-region genes responsible for synthesis of this opine by the plant tumor. *Journal of Bacteriology* 178:3275–3284.

KNOLL, A. H. 1992. The early evolution of eukaryotes: a geological perspective. *Science* 256:622–627.

KOSKI, L. B., MORTON, R. A., AND GOLDING, G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Molecular Biology and Evolution* 18:404–412.

KUMAR, N., LAD, G., GIUNTINI, E., KAYE, M. E., UDOMWONG, P., SHAMSANI, N. J., YOUNG, J. P. W., AND BAILLY, X. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biology* 5:140133.

KUO, C.-H. AND OCHMAN, H. 2009a. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biology Direct* 4:35.

KUO, C.-H. AND OCHMAN, H. 2009b. The fate of new bacterial genes. *FEMS Microbiology Reviews* 33:38–43.

KURLAND, C. 1987. Strategies for efficiency and accuracy in gene expression. *Trends in Biochemical Sciences* 12:126–128.

LACK, H. W. 2002. The discovery and rediscovery of the horse chestnut. *Arnoldia* 61:15–19.

LAFAY, B., SHARP, P. M., LLOYD, A. T., MCLEAN, M. J., DEVINE, K. M., AND WOLFE, K. H. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Research* 27:1642–1649.

LARSSON, J., NYLANDER, J. A., AND BERGMAN, B. 2011. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evolutionary Biology* 11:187.

LASSMANN, T., HAYASHIZAKI, Y., AND DAUB, C. O. 2009. Tagdust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25:2839–2840.

LATYSHEVA, N., JUNKER, V. L., PALMER, W. J., CODD, G. A., AND BARKER, D. 2012. The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics* 28:603–606.

LAUE, B., STEELE, H., AND GREEN, S. 2014. Survival, cold tolerance and seasonality of infection of European horse chestnut (*Aesculus hippocastanum*) by *Pseudomonas syringae* pv. *aesculi*. *Plant Pathology* .

LAWLIS, V., DENNIS, M., CHEN, E., SMITH, D., AND HENNER, D. 1984. Cloning and sequencing of the xylose isomerase and xylulose kinase genes of *Escherichia coli*. *Applied and Environmental Microbiology* 47:15–21.

LAWRENCE, J. G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology* 2:519–523.

LAWRENCE, J. G. 2002. Gene transfer in bacteria: Speciation without species? *Theoretical Population Biology* 61:449–460.

LAWRENCE, J. G. AND HENDRICKSON, H. 2003. Lateral gene transfer: when will adolescence end? *Molecular Microbiology* 50:739–749.

LAWRENCE, J. G. AND OCHMAN, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution* 44:383–397.

LAWRENCE, J. G. AND OCHMAN, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences* 95:9413–9417.

LE, S. Q. AND GASCUEL, O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* 25:1307–1320.

LEES, D. C., LACK, H. W., ROUGERIE, R., HERNANDEZ-LOPEZ, A., RAUS, T., AVTZIS, N. D., AUGUSTIN, S., AND LOPEZ-VAAMONDE, C. 2011. Tracking origins of invasive herbivores through herbaria and archival DNA: the case of the horse-chestnut leaf miner. *Frontiers in Ecology and the Environment* 9:322–328.

LEFÉBURE, T., PAVINSKI BITAR, P. D., SUZUKI, H., AND STANHOPE, M. J. 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biology and Evolution* 2:646–655.

LEFÉBURE, T. AND STANHOPE, M. J. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biology* 8:R71.

LERAT, E., DAUBIN, V., OCHMAN, H., AND MORAN, N. A. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology* 3:e130.

LI, H. AND DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R., ET AL. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

LI, L., STOECKERT, C. J., AND ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13:2178–2189.

LI, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* 36:96–99.

LIN, N. 2005. An avrPto/avrPtoB mutant of *Pseudomonas syringae* pv. *tomato* DC3000 does not elicit Pto-mediated resistance and is less virulent on tomato. *Molecular Plant-Microbe Interactions* .

LINDEBERG, M., CARTINHOUR, S., MYERS, C. R., SCHECHTER, L. M., SCHNEIDER, D. J., AND COLLMER, A. 2006. Closing the circle on the discovery of genes encoding Hrp regulon members and type III secretion system effectors in the genomes of three model *Pseudomonas syringae* strains. *Molecular Plant-Microbe Interactions* 19:1151–1158.

LINDEBERG, M., CUNNAC, S., AND COLLMER, A. 2009. The evolution of *Pseudomonas syringae* host specificity and type III effector repertoires. *Molecular Plant Pathology* 10:767–775.

LINDEBERG, M., CUNNAC, S., AND COLLMER, A. 2012. *Pseudomonas syringae* type III effector repertoires: last words in endless arguments. *Trends in Microbiology* .

LINDOW, S. 1983. The role of bacterial ice nucleation in frost injury to plants. *Annual Review of Phytopathology* 21:363–384.

LIU, H., QIU, H., ZHAO, W., CUI, Z., IBRAHIM, M., JIN, G., LI, B., ZHU, B., AND XIE, G. L. 2012. Genome sequence of the plant pathogen *Pseudomonas syringae* pv. *panici* LMG 2367. *Journal of Bacteriology* 194:5693–5694.

LOBRY, J. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution* 13:660–665.

LOBRY, J. R., CHESSEL, D., ET AL. 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *Journal of Applied Genetics* 44:235–262.

LOKMAN, B. C., VAN SANTEN, P., VERDOES, J. C., KRÜSE, J., LEER, R. J., POSNO, M., AND POUWELS, P. H. 1991. Organization and characterization of three genes involved in D-xylose catabolism in *Lactobacillus pentosus*. *Molecular and General Genetics* 230:161–169.

LUO, C., WALK, S. T., GORDON, D. M., FELDGARDEN, M., TIEDJE, J. M., AND KONSTANTINIDIS, K. T. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences* 108:7200–7205.

MA, S., MORRIS, V., AND CUPPELS, D. 1991. Characterization of a DNA region required for production of the phytotoxin coronatine by *Pseudomonas syringae* pv. *tomato*. *Molecular Plant–Microbe Interactions* 4:69–74.

MA, W., DONG, F. F. T., STAVRINIDES, J., AND GUTTMAN, D. S. 2006. Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS Genetics* 2:e209.

MAGOC, T., PABINGER, S., CANZAR, S., LIU, X., SU, Q., PUIU, D., TALLON, L. J., AND SALZBERG, S. L. 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 29:1718–1725.

MAJEWSKI, J. 2001. Sexual isolation in bacteria. *FEMS Microbiology Letters* 199:161–169.

MAJEWSKI, J. AND COHAN, F. M. 1998. The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* 148:13–18.

MAKAROVA, K. S., ARAVIND, L., GALPERIN, M., GRISHIN, N. V., TATUSOV, R. L., WOLF, Y. I., AND KOONIN, E. V. 1999. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research* 9:608–628.

MAKI, L. R., GALYAN, E. L., CHANG-CHIEN, M.-M., AND CALDWELL, D. R. 1974. Ice nucleation induced by *Pseudomonas syringae*. *Applied Microbiology* 28:456–459.

MALLET, J. 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363:2971–2986.

MANSFIELD, J., GENIN, S., MAGORI, S., CITOVSKY, V., SRIARIYANUM, M., RONALD, P., DOW, M., VERDIER, V., BEER, S. V., MACHADO, M. A., TOTH, I., SALMOND, G., AND FOSTER, G. D. 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology* 13:614–629.

MANSFIELD, J. W. 2009. From bacterial avirulence genes to effector functions via the *hrp* delivery system: an overview of 25 years of progress in our understanding of plant innate immunity. *Molecular Plant Pathology* 10:721–734.

MARCELLETTI, S., FERRANTE, P., PETRICCIONE, M., FIRRAO, G., AND SCORTICHINI, M. 2011. *Pseudomonas syringae* pv. *actinidiae* draft genomes comparison reveal strain-specific features involved in adaptation and virulence to *Actinidia* species. *PLoS ONE* 6:e27297.

MARGESIN, R., SPRÖER, C., ZHANG, D.-C., AND BUSSE, H.-J. 2012. *Polaromonas glacialis* sp. nov. and *Polaromonas cryoconiti* sp. nov., isolated from alpine glacier cryoconite. *International Journal of Systematic and Evolutionary Microbiology* 62:2662–2668.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10–12.

MAVRODI, D. V., LOPER, J. E., PAULSEN, I. T., AND THOMASHOW, L. S. 2009. Mobile genetic elements in the genome of the beneficial rhizobacterium *Pseudomonas fluorescens* Pf-5. *BMC Microbiology* 9:8.

MAYNARD SMITH, J., SMITH, N. H., O'ROURKE, M., AND SPRATT, B. G. 1993. How clonal are bacteria? *Proceedings of the National Academy of Sciences* 90:4384–4388.

MAZZAGLIA, A., STUDHOLME, D. J., TARATUFOLO, M. C., CAI, R., ALMEIDA, N. F., GOODMAN, T., GUTTMAN, D. S., VINATZER, B. A., AND BALESTRA, G. M. 2012. *Pseudomonas syringae* pv. *actinidiae* (PSA) isolates from recent bacterial canker of kiwifruit outbreaks belong to the same genetic lineage. *PLoS ONE* 7:e36518.

MCCANN, H. C., RIKKERINK, E. H., BERTELS, F., FIERS, M., LU, A., REES-GEORGE, J., ANDERSEN, M. T., GLEAVE, A. P., HAUBOLD, B., WOHLERS, M. W., ET AL. 2013. Genomic analysis of the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* provides insight into the origins of an emergent plant disease. *PLoS Pathogens* 9:e1003503.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P., AND HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9:356–369.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M., ET AL. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297–1303.

MCLEAN, M. J., WOLFE, K. H., AND DEVINE, K. M. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution* 47:691–696.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V., AND RAPPUOLI, R. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development* 15:589–594.

MELOTTO, M., UNDERWOOD, W., KOCZAN, J., NOMURA, K., AND HE, S. Y. 2006. Plant stomata function in innate immunity against bacterial invasion. *Cell* 126:969–980.

MÉNARD, M., SUTRA, L., LUISETTI, J., PRUNIER, J. P., AND GARDAN, L. 2003. *Pseudomonas syringae* pv. *avii* (pv. nov.), the causal agent of bacterial canker of wild cherries (*Prunus avium*) in France. *European Journal of Plant Pathology* 109:565–576.

MERTELIK, J., KLOUDOVA, K., PANKOVA, I., KREJZAR, V., AND KUDELA, V. 2013. Occurrence of horse chestnut bleeding canker caused by *Pseudomonas syringae* pv. *aesculi* in the Czech Republic. *Forest Pathology* 43:165–167.

MIETHKE, M., MONTEFERRANTE, C. G., MARAHIEL, M. A., AND VAN DIJL, J. M. 2013. The *Bacillus subtilis* EfeUOB transporter is essential for high-affinity acquisition of ferrous and ferric iron. *Molecular Cell Research* 1833:2267–2278.

MILNE, I., BAYER, M., CARDLE, L., SHAW, P., STEPHEN, G., WRIGHT, F., AND MARSHALL, D. 2010. Tablet—next generation sequence assembly visualization. *Bioinformatics* 26:401–402.

MONTEIL, C. L., CAI, R., LIU, H., MECHAN LLONTOP, M. E., STUDHOLME, D. J., MORRIS, C. E., AND VINATZER, B. A. 2013. Nonagricultural reservoirs contribute to emergence and evolution of *Pseudomonas syringae* crop pathogens. *New Phytologist* 199:800–811.

MONTEIL, C. L., GUILBAUD, C., GLAUX, C., LAFOLIE, F., SOUBEYRAND, S., AND MORRIS, C. E. 2012. Emigration of the plant pathogen *Pseudomonas syringae* from leaf litter contributes to its population dynamics in alpine snowpack. *Environmental Microbiology* 14:2099–2112.

MORRIS, C. E., BARDIN, M., KINKEL, L. L., MOURY, B., NICOT, P. C., AND SANDS, D. C. 2009. Expanding the paradigms of plant pathogen life history and evolution of parasitic fitness beyond agricultural boundaries. *PLoS Pathogens* 5:e1000693.

MORRIS, C. E., GLAUX, C., LATOUR, X., GARDAN, L., SAMSON, R., AND PITRAT, M. 2000. The relationship of host range, physiology, and genotype to virulence on cantaloupe in *Pseudomonas syringae* from cantaloupe blight epidemics in france. *Phytopathology* 90:636–646.

MORRIS, C. E., KINKEL, L. L., XIAO, K., PRIOR, P., AND SANDS, D. C. 2007. Surprising niche for the plant pathogen *Pseudomonas syringae*. *Infection, Genetics and Evolution* 7:84–92.

MORRIS, C. E., SANDS, D. C., VANNESTE, J. L., MONTARRY, J., OAKLEY, B., GUILBAUD, C., AND GLAUX, C. 2010. Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. *mBio* 1:e00107–10–e00107–20.

MORRIS, C. E., SANDS, D. C., VINATZER, B. A., GLAUX, C., GUILBAUD, C., BUFFIÈRE, A., YAN, S., DOMINGUEZ, H., AND THOMPSON, B. M. 2008. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. *The ISME Journal* 2:321–334.

MURPHY, P. J., HEYCKE, N., TRENZ, S. P., RATET, P., DE BRUIJN, F. J., AND SCHELL, J. 1988. Synthesis of an opine-like compound, a rhizopine, in alfalfa nodules is symbiotically regulated. *Proceedings of the National Academy of Sciences* 85:9133–9137.

MURPHY, P. J., WEXLER, W., GRZEMSKI, W., RAO, J. P., AND GORDON, D. 1995. Rhizopines—their role in symbiosis and competition. *Soil Biology and Biochemistry* 27:525–529.

MUTO, A. AND OSAWA, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences* 84:166–169.

MUTREJA, A., KIM, D. W., THOMSON, N. R., CONNOR, T. R., LEE, J. H., KARIUKI, S., CROUCHER, N. J., CHOI, S. Y., HARRIS, S. R., LEBENS, M., ET AL. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477:462–465.

NGUYEN, H. P., YEAM, I., ANGOT, A., AND MARTIN, G. B. 2010. Two virulence determinants of type iii effector avrpto are functionally conserved in diverse *Pseudomonas syringae* pathovars. *New Phytologist* 187:969–982.

NOTREDAME, C., HIGGINS, D. G., AND HERINGA, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302:205–217.

NOWELL, R. W., GREEN, S., LAUE, B. E., AND SHARP, P. M. 2014. The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome Biology and Evolution* 6:1514–1529.

O'BRIEN, H. E., DESVEAUX, D., AND GUTTMAN, D. S. 2011a. Next-generation genomics of *Pseudomonas syringae*. *Current Opinion in Microbiology* pp. 1–7.

O'BRIEN, H. E., THAKUR, S., GONG, Y., FUNG, P., ZHANG, J., YUAN, L., WANG, P. W., YONG, C., SCORTICHINI, M., AND GUTTMAN, D. S. 2012. Extensive remodeling of the *Pseudomonas syringae* pv. *avellanae* type III secretome associated with two independent host shifts onto hazelnut. *BMC Microbiology* 12:141.

O'BRIEN, H. E., THAKUR, S., AND GUTTMAN, D. S. 2011b. Evolution of plant pathogenesis in *Pseudomonas syringae*: A genomics perspective. *Annual Review of Phytopathology* 49:269–289.

OCHMAN, H. 2001. Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. *Science* 292:1096–1099.

OCHMAN, H., ELWYN, S., AND MORAN, N. 1999. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences* 96:12638–12643.

OCHMAN, H., LAWRENCE, J. G., AND GROISMAN, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.

OCHMAN, H., LERAT, E., AND DAUBIN, V. 2005. Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Sciences* 102 Suppl 1:6595–6599.

OGIMI, C. AND HIGUCHI, H. 1981. Bacterial gall of Yamamomo (*Myrica rubra* S. et Z.) caused by *Pseudomonas syringae* pv. *myricae* pv. nov. *Annals Phytopathological Society of Japan* 47:443–448.

OGIMI, C., HIGUCHI, H., AND TAKIKAWA, Y. 1988. Bacterial gall disease of Kakuremino (*Dendropanax trifidus* Mak.) caused by *Pseudomonas syringae* pv. *dendropanacis* pv. nov. *Annals of the Phytopathological Society of Japan* 54.

OGIMI, C., KAWANO, C., HIGUCHI, H., AND TAKIKAWA, Y. 1992. Bacterial gall disease of sharinbai (*Rhaphiolepis umbellata*) caused by *Pseudomonas syringae* pv. *rhaphiolepidis* pv. nov. *Journal of the Japanese Forestry Society* 74:308–313.

OGIMI, C., KUBO, Y., HIGUCHI, H., AND TAKIKAWA, Y. 1990. Bacterial gall diseases of himeyuzuriha (*Daphniphyllum teijsmanni* Z.) caused by *Pseudomonas syringae* pv. *daphniphylli* pv. nov. *Journal of the Japanese Forestry Society* 72:17–22.

OKE, V. AND LONG, S. R. 1999. Bacteroid formation in the *Rhizobium*–legume symbiosis. *Current Opinion in Microbiology* 2:641–646.

OVERBEEK, R., OLSON, R., PUSCH, G. D., OLSEN, G. J., DAVIS, J. J., DISZ, T., EDWARDS, R. A., GERDES, S., PARRELLO, B., SHUKLA, M., VONSTEIN, V., WATTAM, A. R., XIA, F., AND STEVENS, R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* 42:D206–D214.

PAGEL, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255:37–45.

PAGEL, M. AND MEADE, A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist* 167:808–825.

PARADIS, E., CLAUDE, J., AND STRIMMER, K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

PARKINSON, N., BRYANT, R., BEW, J., AND ELPHINSTONE, J. 2010. Rapid phylogenetic identification of members of the *Pseudomonas syringae* species complex using the *rpoD* locus. *Plant Pathology* 60:338–344.

PASZKIEWICZ, K. AND STUDHOLME, D. J. 2010. *De novo* assembly of short sequence reads. *Briefings in Bioinformatics* 11:457–472.

PEDEN, J. F. 1999. Analysis of Codon Usage, Ph.D. Thesis. University of Nottingham.

PENG, Y., LEUNG, H. C., YIU, S.-M., AND CHIN, F. Y. 2010. IDBA—a practical iterative de Bruijn graph *de novo* assembler. *In* Research in Computational Molecular Biology, pp. 426–440. Springer.

PENG, Y., LEUNG, H. C. M., YIU, S. M., AND CHIN, F. Y. L. 2012. IDBA-UD: A *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428.

PERRIÈRE, G. AND THIOULOUSE, J. 2002. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Research* 30:4548–4555.

PHADTARE, S., ALSINA, J., AND INOUYE, M. 1999. Cold-shock response and cold-shock proteins. *Current Opinion in Microbiology* 2:175–180.

PIRNAY, J.-P., BILOCQ, F., POT, B., CORNELIS, P., ZIZI, M., VAN ELDERE, J., DESCHAGHT, P., VANEECHOUTTE, M., JENNES, S., PITT, T., ET AL. 2009. *Pseudomonas aeruginosa* population structure revisited. *PLoS ONE* 4:e7740.

PITMAN, A. R., JACKSON, R. W., MANSFIELD, J. W., KAITELL, V., THWAITES, R., AND ARNOLD, D. L. 2005. Exposure to host resistance mechanisms drives evolution of bacterial virulence in plants. *Current Biology* 15:2230–2235.

POLZ, M. F., ALM, E. J., AND HANAGE, W. P. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics* 29:170–175.

POLZ, M. F., HUNT, D. E., PREHEIM, S. P., AND WEINREICH, D. M. 2006. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 361:2009–2021.

POPA, O., HAZKANI-COVO, E., LANDAN, G., MARTIN, W., AND DAGAN, T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research* 21:599–609.

POTTER, D., ERIKSSON, T., EVANS, R. C., OH, S., SMEDMARK, J., MORGAN, D. R., KERR, M., ROBERTSON, K. R., ARSENAULT, M., DICKINSON, T. A., ET AL. 2007. Phylogeny and classification of Rosaceae. *Plant Systematics and Evolution* 266:5–43.

QI, M., WANG, D., BRADLEY, C. A., AND ZHAO, Y. 2011. Genome sequence analyses of *Pseudomonas savastanoi* pv. *glycinea* and subtractive hybridization-based comparative genomics with nine Pseudomonads. *PLoS ONE* 6:e16451.

R CORE TEAM 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RAMOS, C., MATAS, I. M., BARDAJI, L., ARAGÓN, I. M., AND MURILLO, J. 2012. *Pseudomonas savastanoi* pv. *savastanoi*: some like it knot. *Molecular Plant Pathology* 13:998–1009.

RASKO, D. A., ROSOVITZ, M. J., MYERS, G. S. A., MONGODIN, E. F., FRICKE, W. F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N. R., CHAUDHURI, R., HENDERSON, I. R., SPERANDIO, V., AND RAVEL, J. 2008. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology* 190:6881–6893.

REINHARDT, J. A., BALTRUS, D. A., NISHIMURA, M. T., JECK, W. R., JONES, C. D., AND DANGL, J. L. 2008. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Research* 19:294–305.

REST, J. S. AND MINDELL, D. P. 2003. Retroids in archaea: phylogeny and lateral origins. *Molecular Biology and Evolution* 20:1134–1142.

RICE, W. R. 1989. Analyzing tables of statistical tests. *Evolution* pp. 223–225.

RICHARDS, V. P., LANG, P., BITAR, P. D. P., LEFÉBURE, T., SCHUKKEN, Y. H., ZADOKS, R. N., AND STANHOPE, M. J. 2011. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. *Infection, Genetics and Evolution* pp. 1–13.

ROCHA, E. P. 2008. Evolutionary patterns in prokaryotic genomes. *Current Opinion in Microbiology* 11:454–460.

RODRÍGUEZ-MORENO, L., JIMÉNEZ, A. J., AND RAMOS, C. 2009. Endopathogenic lifestyle of *Pseudomonas savastanoi* pv. *savastanoi* in olive knots. *Microbial Biotechnology* 2:476–488.

RODRÍGUEZ-PALENZUELA, P., MATAS, I. M., MURILLO, J., LÓPEZ-SOLANILLA, E., BARDAJI, L., PÉREZ-MARTÍNEZ, I., RODRÍGUEZ-MOSKERA, M. E., PENYALVER, R., LÓPEZ, M. M., QUESADA, J. M., BIEHL, B. S., PERNA, N. T., GLASNER, J. D., CABOT, E. L., NEENO-ECKWALL, E., AND RAMOS, C. 2010. Annotation and overview of the *Pseudomonas savastanoi* pv. *savastanoi* NCPPB 3335 draft genome reveals the virulence gene complement of a tumour-inducing pathogen of woody hosts. *Environmental Microbiology* .

RYGUS, T., SCHELER, A., ALLMANSBERGER, R., AND HILLEN, W. 1991. Molecular cloning, structure, promoters and regulatory elements for transcription of the *Bacillus megaterium* encoded regulon for xylose utilization. *Archives of Microbiology* 155:535–542.

SAIER, M. H. 2004. Evolution of bacterial type III protein secretion systems. *Trends in Microbiology* 12:113–115.

SAINT, C., WEXLER, M., MURPHY, P., TEMPE, J., AND TATE, M. 1993. Characterization of genes for synthesis and catabolism of a new rhizopine induced in nodules by *Rhizobium meliloti* rm220-3: extension of the rhizopine concept. *Journal of Bacteriology* 175:5205–5215.

SAITOU, N. AND NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.

SALZBERG, S. L., PHILLIPPY, A. M., ZIMIN, A., PUIU, D., MAGOC, T., KOREN, S., TREANGEN, T. J., SCHATZ, M. C., DELCHER, A. L., ROBERTS, M., MARCAIS, G., POP, M., AND YORKE, J. A. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* 22:557–567.

SARKAR, S. F., GORDON, J. S., MARTIN, G. B., AND GUTTMAN, D. S. 2006. Comparative genomics of host-specific virulence in *Pseudomonas syringae*. *Genetics* 174:1041–1056.

SARKAR, S. F. AND GUTTMAN, D. S. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology* 70:1999–2012.

SATTLER, B., PUXBAUM, H., AND PSENNER, R. 2001. Bacterial growth in supercooled cloud droplets. *Geophysical Research Letters* 28:239–242.

SAWADA, H., SUZUKI, F., MATSUDA, I., AND SAITOU, N. 1999. Phylogenetic analysis of *Pseudomonas syringae* pathovars suggests the horizontal gene transfer of argK and the evolutionary stability of hrp gene cluster. *Journal of Molecular Evolution* 49:627–644.

SCHAAD, N. W., POSTNIKOVA, E., SECHLER, A., CLAFLIN, L. E., VIDAVER, A. K., JONES, J. B., AGARKOVA, I., IGNATOV, A., DICKSTEIN, E., AND RAMUNDO, B. A. 2008. Reclassification of subspecies of *Acidovorax avenae* as *A. Avenae* (manns 1905) emend., *A. cattleyae* (pavarino, 1911) comb. nov., *A. citrulli* (schaad et al., 1978) comb. nov., and proposal of *A. oryzae* sp. nov. *Systematic and Applied Microbiology* 31:434–446.

SCHELER, A., RYGUS, T., ALLMANSBERGER, R., AND HILLEN, W. 1991. Molecular cloning, structure, promoters and regulatory elements for transcription of the *Bacillus licheniformis* encoded regulon for xylose utilization. *Archives of Microbiology* 155:526–534.

SCHELLENBERG, B., RAMEL, C., AND DUDLER, R. 2010. *Pseudomonas syringae* virulence factor syringolin a counteracts stomatal immunity by proteasome inhibition. *Molecular Plant-Microbe Interactions* 23:1287–1293.

SCHMIDT, O., DUJESIEFKEN, D., STOBBE, H., MORETH, U., KEHR, R., AND SCHRÖDER, T. 2008. *Pseudomonas syringae* pv. *aesculi* associated with horse chestnut bleeding canker in Germany. *Forest Pathology* 38:124–128.

SCHOLZ-SCHROEDER, B. K., HUTCHISON, M. L., GRGURINA, I., AND GROSS, D. C. 2001. The contribution of syringopeptin and syringomycin to virulence of *Pseudomonas syringae* pv. *syringae* strain b301d on the basis of *sypA* and *syrB1* biosynthesis mutant analysis. *Molecular Plant-Microbe Interactions* 14:336–348.

SCHULZE, S., KAY, S., BÜTTNER, D., EGLER, M., ESCHEN-LIPPOLD, L., HAUSE, G., KRÜGER, A., LEE, J., MÜLLER, O., SCHEEL, D., ET AL. 2012. Analysis of new type III effectors from *Xanthomonas* uncovers XopB and XopS as suppressors of plant immunity. *New Phytologist* 195:894–911.

SCORTICHINI, M. 2002. Bacterial canker and decline of European hazelnut. *Plant Disease* 86:704–709.

SCORTICHINI, M., MARCELLETTI, S., FERRANTE, P., AND FIRRAO, G. 2013. A genomic redefinition of *Pseudomonas avellanae* species. *PloS ONE* 8:e75794.

SCORTICHINI, M., MARCHESI, U., ROSSI, M., AND DI PROSPERO, P. 2002. Bacteria associated with hazelnut (*Corylus avellana* L.) decline are of two groups: *Pseudomonas avellanae* and strains resembling *P. syringae* pv. *syringae*. *Applied and Environmental Microbiology* 68:476–484.

SCORTICHINI, M., NATALINI, E., AND MARCHESI, U. 2006. Evidence for separate origins of the two *Pseudomonas avellanae* lineages. *Plant Pathology* 55:451–457.

SCOTT, L. J., MOHLKE, K. L., BONNYCASTLE, L. L., WILLER, C. J., LI, Y., DUREN, W. L., ERDOS, M. R., STRINGHAM, H. M., CHINES, P. S., JACKSON, A. U., ET AL. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341–1345.

SHAPIRO, B. J., FRIEDMAN, J., CORDERO, O. X., PREHEIM, S. P., TIMBERLAKE, S. C., SZABO, G., POLZ, M. F., AND ALM, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* 336:48–51.

SHAPIRO, B. J. AND POLZ, M. F. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology* 22:235–247.

SHARP, P. M., BAILES, E., GROCOCK, R. J., PEDEN, J. F., AND SOCKETT, R. E. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research* 33:1141–1153.

SHARP, P. M., COWE, E., HIGGINS, D. G., SHIELDS, D. C., WOLFE, K. H., AND WRIGHT, F. 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research* 16:8207–8211.

SHARP, P. M., EMERY, L. R., AND ZENG, K. 2010. Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:1203–1212.

SHARP, P. M. AND LI, W.-H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* 24:28–38.

SHARP, P. M., STENICO, M., PEDEN, J. F., AND LLOYD, A. T. 1993. Codon usage: mutational bias, translational selection, or both? *Biochemical Society Transactions* 21:835.

SHEPPARD, S. K., DIDELOT, X., MERIC, G., TORRALBO, A., JOLLEY, K. A., KELLY, D. J., BENTLEY, S. D., MAIDEN, M. C., PARKHILL, J., AND FALUSH, D. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proceedings of the National Academy of Sciences* 110:11923–11927.

SHIELDS, D. C. AND SHARP, P. M. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Research* 15:8023–8040.

SIGUIER, P., FILÉE, J., AND CHANDLER, M. 2006a. Insertion sequences in prokaryotic genomes. *Current Opinion in Microbiology* 9:526–531.

SIGUIER, P., PÉROCHON, J., LESTRADE, L., MAHILLON, J., AND CHANDLER, M. 2006b. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research* 34:D32–D36.

SMEDS, L. AND KÜNSTNER, A. 2011. ConDeTri—a content dependent read trimmer for Illumina data. *PLoS ONE* 6:e26314.

SMILLIE, C. S., SMITH, M. B., FRIEDMAN, J., CORDERO, O. X., DAVID, L. A., AND ALM, E. J. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–244.

SNEL, B. 2001. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Research* 12:17–25.

SNEL, B., BORK, P., AND HUYNEN, M. A. 1999. Genome phylogeny based on gene content. *Nature Genetics* 21:108–110.

SOGIN, M. L., MORRISON, H. G., HUBER, J. A., WELCH, D. M., HUSE, S. M., NEAL, P. R., ARRIETA, J. M., AND HERNDL, G. J. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* 103:12115–12120.

SOHN, K. H., JONES, J. D. G., AND STUDHOLME, D. J. 2012. Draft genome sequence of *Pseudomonas syringae* pathovar *Syringae* strain FF5, causal agent of stem tip dieback disease on ornamental pear. *Journal of Bacteriology* 194:3733–3734.

SONG, S. AND PARK, C. 1997. Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *Journal of Bacteriology* 179:7025–7032.

SOREK, R., ZHU, Y., CREEVEY, C. J., FRANCINO, M. P., BORK, P., AND RUBIN, E. M. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.

SPRATT, B. G., HANAGE, W. P., AND FEIL, E. J. 2001. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current Opinion in Microbiology* 4:602–606.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

STARK, H., RODNINA, M. V., RINKE-APPEL, J., BRIMACOMBE, R., WINTERMEYER, W., AND VAN HEEL, M. 1997. Visualization of elongation factor Tu on the *Escherichia coli* ribosome. *Nature* 389:403–406.

STEELE, H., LAUE, B. E., MACASKILL, G. A., HENDRY, S. J., AND GREEN, S. 2010. Analysis of the natural infection of European horse chestnut (*Aesculus hippocastanum*) by *Pseudomonas syringae* pv. *aesculi*. *Plant Pathology* 59:1005–1013.

STEPHENS, C., CHRISTEN, B., FUCHS, T., SUNDARAM, V., WATANABE, K., AND JENAL, U. 2007. Genetic analysis of a novel pathway for d-xylose metabolism in caulobacter crescentus. *Journal of Bacteriology* 189:2181–2185.

STETTER, K., HUBER, R., BLÖCHL, E., KURR, M., EDEN, R., FIELDER, M., CASH, H., AND VANCE, I. 1993. Hyperthermophilic archaea are thriving in deep north sea and alaskan oil reservoirs. *Nature* 365:743–745.

STOLETZKI, N. AND EYRE-WALKER, A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Molecular Biology and Evolution* 24:374–381.

STUDHOLME, D. J., IBANEZ, S., MACLEAN, D., DANGL, J. L., CHANG, J. H., AND RATHJEN, J. P. 2009. A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of *Pseudomonas syringae* pathovar *tabaci* 11528. *BMC Genomics* 10:395.

SUEOKA, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences* 48:582.

SUZUKI, H., BROWN, C. J., FORNEY, L. J., AND TOP, E. M. 2008. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Research* 15:357–365.

SUZUKI, M., HAYAKAWA, T., SHAW, J., REKIK, M., AND HARAYAMA, S. 1991. Primary structure of xylene monooxygenase: similarities to and differences from the alkane hydroxylation system. *Journal of Bacteriology* 173:1690–1695.

SWINGLEY, W. D., BLANKENSHIP, R. E., AND RAYMOND, J. 2008. Integrating markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Molecular Biology and Evolution* 25:643–654.

TAKAHASHI, K., NISHIYAMA, K., AND SATO, M. 1996. *Pseudomonas syringae* pv. *broussonetiae* pv. nov., the causal agent of bacterial blight of paper mulberry (*Broussonetia kazinoki* $\times$ *B. papyrifera*). *Annals of the Phytopathological Society of Japan* 62:17–22.

TAKEMOTO, J. Y., ZHANG, L., TAGUCHI, N., TACHIKAWA, T., AND MIYAKAWA, T. 1991. Mechanism of action of the phytotoxin syringomycin: a resistant mutant of *Saccharomyces cerevisiae* reveals an involvement of Ca^{2+} transport. *Journal of General Microbiology* 137:653–659.

TALGO, V., SPIES PERMINOW, J., SLETTEN, A., BRURBERG, M. B., HERRERO, M., STROMENG, G., AND STENSVAND, A. 2012. Fungal and bacterial diseases on horse chestnut in Norway. *Journal of Agricultural Extension and Rural Development* 4:256–258.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J. B., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R., AND FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102:13950–13955.

TETTELIN, H., RILEY, D., CATTURO, C., AND MEDINI, D. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* 11:472–477.

THOMAS, C. M. AND NIELSEN, K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* 3:711–721.

THORVALDSDÓTTIR, H., ROBINSON, J. T., AND MESIROV, J. P. 2012. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* p. bbs017.

TORSVIK, V., ØVREÅS, L., AND THINGSTAD, T. F. 2002. Prokaryotic diversity–magnitude, dynamics, and controlling factors. *Science* 296:1064–1066.

TOTH, I. K., BELL, K. S., HOLEVA, M. C., AND BIRCH, P. R. 2003. Soft rot erwiniae: from genes to genomes. *Molecular Plant Pathology* 4:17–30.

TOUCHON, M., HOEDE, C., TENAILLON, O., BARBE, V., BAERISWYL, S., BIDET, P., BINGEN, E., BONACORSI, S., BOUCHIER, C., BOUVET, O., CALTEAU, A., CHIAPELLO, H., CLERMONT, O., CRUVEILLER, S., DANCHIN, A., DIARD, M., DOSSAT, C., KAROUI, M. E., FRAPY, E., GARRY, L., GHIGO, J. M., GILLES, A. M., JOHNSON, J., LE BOUGUÉNEC, C., LESCAT, M., MANGENOT, S., MARTINEZ-JÉHANNE, V., MATIC, I., NASSIF, X., OZTAS, S., PETIT, M. A., PICHON, C., ROUY, Z., RUF, C. S., SCHNEIDER, D., TOURRET, J., VACHERIE, B., VALLENET, D., MÉDIGUE, C., ROCHA, E. P. C., AND DENAMUR, E. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics* 5:e1000344.

TREANGEN, T. J. AND ROCHA, E. P. C. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics* 7:e1001284.

TSUYMU, S. AND GOTO, M. 1989. *Pseudomonas syringae* pv. *actinidiae* pv. nov.: the causal bacterium of canker of kiwifruit in Japan. *Annals of the Phytopathological Society of Japan* 55:437–444.

UCHYTIL, T. AND DURBIN, R. 1980. Hydrolysis of tabtoxins by plant and bacterial enzymes. *Experientia* 36:301–302.

VAN DONGEN, S. M. 2000. Graph clustering by flow simulation. *PhD thesis, University of Utrecht* .

VERNIKOS, G. S., THOMSON, N. R., AND PARKHILL, J. 2007. Genetic flux over time in the *Salmonella* lineage. *Genome Biology* 8:R100.

VICENTE, J., ALVES, J., RUSSELL, K., AND ROBERTS, S. 2004. Identification and discrimination of *Pseudomonas syringae* isolates from wild cherry in England. *European Journal of Plant Pathology* 110:337–351.

VOS, M. AND DIDELOT, X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal* 3:199–208.

WANG, H., MOORE, M. J., SOLTIS, P. S., BELL, C. D., BROCKINGTON, S. F., ALEXANDRE, R., DAVIS, C. C., LATVIS, M., MANCHESTER, S. R., AND SOLTIS, D. E. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences* 106:3853–3858.

WANG, P. W., MORGAN, R. L., SCORTICHINI, M., AND GUTTMAN, D. S. 2007. Convergent evolution of phytopathogenic Pseudomonads onto hazelnut. *Microbiology* 153:2067–2073.

WANG, Y., HUANG, H., SUN, M., ZHANG, Q., AND GUO, D. 2012. T3DB: an integrated database for bacterial type III secretion system. *BMC Bioinformatics* 13:66.

WATKINS, R. F. AND GRAY, M. W. 2006. The frequency of eubacterium-to-eukaryote lateral gene transfers shows significant cross-taxa variation within amoebozoa. *Journal of Molecular Evolution* 63:801–814.

WEBBER, J. F., PARKINSON, N. M., ROSE, J., STANFORD, H., COOK, R. T. A., AND ELPHINSTONE, J. G. 2008. Isolation and identification of *Pseudomonas syringae* pv. *aesculi* causing bleeding canker of horse chestnut in the UK. *Plant Pathology* 57:368–368.

WIEDENBECK, J. AND COHAN, F. M. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews* 35:957–976.

WILDSCHUTTE, H., PREHEIM, S. P., HERNANDEZ, Y., AND POLZ, M. F. 2010. O-antigen diversity and lateral transfer of the *wbe* region among *Vibrio splendidus* isolates. *Environmental Microbiology* 12:2977–2987.

WILLIAMS, K. P., GILLESPIE, J. J., SOBRAL, B. W. S., NORDBERG, E. K., SNYDER, E. E., SHALLOM, J. M., AND DICKERMAN, A. W. 2010. Phylogeny of Gammaproteobacteria. *Journal of Bacteriology* 192:2305–2314.

WOESE, C. R. 1987. Bacterial evolution. *Microbiological Reviews* 51:221.

WOESE, C. R. 2000. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences* 97:8392–8396.

WOESE, C. R. AND FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* 74:5088–5090.

WOZNIAK, R. A. AND WALDOR, M. K. 2010. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology* 8:552–563.

WRIGHT, F. 1990. The 'effective number of codons' used in a gene. *Gene* 87:23–29.

YAMADA, T., LEE, P., AND KOSUGE, T. 1986. Insertion sequence elements of *Pseudomonas savastanoi*: nucleotide sequence and homology with *Agrobacterium tumefaciens* transfer DNA. *Proceedings of the National Academy of Sciences* 83:8263.

YOUNG, J. 1970. Drippy gill: A bacterial disease of cultivated mushrooms caused by *Pseudomonas agarici* n. sp. *New Zealand Journal of Agricultural Research* 13:977–990.

ZERBINO, D. R. AND BIRNEY, E. 2008. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research* 18:821–829.

ZHAXYBAYEVA, O., NESBO, C. L., AND DOOLITTLE, W. F. 2007. Systematic overestimation of gene gain through false diagnosis of gene absence. *Genome Biology* 8:402.

Appendix A

This appendix contains published material arising from this thesis:

Nowell, R. W., Green, S., Laue, B. L. and Sharp, P. M. 2014. The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome Biology and Evolution* 6:1514–1529.

The online Supplementary Materials associated with this publication are also provided.

GENOME BIOLOGY AND EVOLUTION

SMBE

GBE

The Extent of Genome Flux and Its Role in the Differentiation of Bacterial Lineages

Reuben W. Nowell[1,2,*], Sarah Green[2], Bridget E. Laue[2], and Paul M. Sharp[1,3]

[1]Institute of Evolutionary Biology, University of Edinburgh, United Kingdom

[2]Forest Research, Centre for Ecosystems, Society and Biosecurity, Roslin, Midlothian, United Kingdom

[3]Centre for Immunity, Infection and Evolution, University of Edinburgh, United Kingdom

*Corresponding author: E-mail: r.w.nowell@sms.ed.ac.uk.

Accepted: June 4, 2014

Data deposition: All alignments are available upon request from the authors.

Abstract

Horizontal gene transfer (HGT) and gene loss are key processes in bacterial evolution. However, the role of gene gain and loss in the emergence and maintenance of ecologically differentiated bacterial populations remains an open question. Here, we use whole-genome sequence data to quantify gene gain and loss for 27 lineages of the plant-associated bacterium *Pseudomonas syringae*. We apply an extensive error-control procedure that accounts for errors in draft genome data and greatly improves the accuracy of patterns of gene occurrence among these genomes. We demonstrate a history of extensive genome fluctuation for this species and show that individual lineages could have acquired thousands of genes in the same period in which a 1% amino acid divergence accrues in the core genome. Elucidating the dynamics of genome fluctuation reveals the rapid turnover of gained genes, such that the majority of recently gained genes are quickly lost. Despite high observed rates of fluctuation, a phylogeny inferred from patterns of gene occurrence is similar to a phylogeny based on amino acid replacements within the core genome. Furthermore, the core genome phylogeny suggests that *P. syringae* should be considered a number of distinct species, with levels of divergence at least equivalent to those between recognized bacterial species. Gained genes are transferred from a variety of sources, reflecting the depth and diversity of the potential gene pool available via HGT. Overall, our results provide further insights into the evolutionary dynamics of genome fluctuation and implicate HGT as a major factor contributing to the diversification of *P. syringae* lineages.

Key words: bacterial genome fluctuation, horizontal gene transfer, adaptation, speciation, *Pseudomonas syringae*.

Introduction

Bacterial genomes are labile entities, fluctuating in both size and content through time (Ochman et al. 2005; Hao and Golding 2006; Touchon et al. 2009). Genome fluctuation is mediated by the counteracting processes of gene gain and gene loss. The addition of novel genes expands the functional diversity of the recipient genome, whereas genes whose presence is no longer required for survival can be deleted. The addition of new genes occurs primarily through horizontal (lateral) gene transfer (HGT), which can drive both phenotypic innovation and subsequent adaptation when the acquired genes confer new traits that allow diversification of that lineage into a novel environment (Ochman et al. 2000; Boucher et al. 2003; Lerat et al. 2005; Treangen and Rocha 2011). However, the specific contribution of HGT toward other evolutionary processes, such as genomic differentiation and the formation of ecologically defined population structure, from which speciation may result, remains an open question (Lawrence 2002; Ochman et al. 2005; Polz et al. 2013).

High rates of gene exchange due to HGT can lead to extensive gene repertoire differences among otherwise closely related lineages and can result in genes within the same genome having conflicting evolutionary histories (Feil et al. 2001; Gogarten et al. 2002; Ochman et al. 2005; Shapiro et al. 2012). Thus, HGT is predicted to obfuscate patterns of vertically inherited differences between lineages and should undermine the definition of bacterial "species" as nonrecombining taxonomic units (Gogarten et al. 2002; Doolittle and Papke 2006; Doolittle and Zhaxybayeva 2009). However, despite overwhelming evidence of extensive HGT at virtually all phylogenetic levels, bacterial lineages generally do behave as distinct entities that often appear to have well-defined

© The Author(s) 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

boundaries between taxonomic groups (Daubin et al. 2003; Ochman et al. 2005; Hanage et al. 2006; Caro-Quintero and Konstantinidis 2011). These apparently conflicting observations have led to the proposition that the majority of HGT is transient and therefore does not impact the evolutionary trajectories of bacterial lineages (Hao and Golding 2006; Kuo and Ochman 2009). Furthermore, recent work has suggested that HGT may play an important role in the evolution of population structure within bacterial communities and may facilitate the speciation process (Ochman et al. 2005; Polz et al. 2013).

HGT may bring about divergence and differentiation if the diversifying effect of HGT is able to subdivide a population to the extent that speciation can occur (Lawrence and Ochman 1998; Lawrence 1999, 2002; Ochman et al. 2005; Barraclough et al. 2012; Polz et al. 2013). Under this model, variation in the functions of acquired genes results in the differential adaptation of subpopulations of bacteria into correspondingly different ecological habitats or niches, thereby creating a physical barrier to continued exchange and allowing sequence divergence to accrue to the extent that homologous recombination is suppressed. Subsequent gene loss can enhance differentiation between lineages by limiting niche overlap between incipient species through the restriction of metabolic diversity of populations of bacteria (Lawrence 1999). The innovative and diversifying effects of HGT are well documented, and as such, the exposure to a potentially huge pool of novel genes via HGT could be the mechanism that drives the early stages of this process. Thus, the forces of genome fluctuation, HGT, and adaptation are inexorably intertwined and can, under the right conditions, act in concert to promote both genomic differentiation and ecological diversification, from which speciation can result.

Here, we examine the interplay between these evolutionary processes using the *Pseudomonas syringae* species complex as a model system. *Pseudomonas syringae* is a remarkably diverse species found almost globally in a multitude of different environments (Hirano and Upper 1990; Sarkar and Guttman 2004). Although best known as a plant pathogen, it can also exist as a plant commensal or a free living organism capable of thriving in nonagricultural habitats (Sarkar and Guttman 2004; Morris et al. 2008, 2010). The species complex comprises over 50 pathovars (pathological variants) delimited into a number of phylogroups based on sequence divergence (Sarkar and Guttman 2004; Hwang et al. 2005; Parkinson et al. 2010). Although as a species *P. syringae* has a broad host range, individual strains are generally confined to a limited set of potential host species (Sarkar et al. 2006; Lindeberg et al. 2009). Strains injected into a nonhost plant will either simply fail to grow or will elicit an immune response from the plant that halts the progression of disease symptoms (Morris et al. 2000; Hunter and Taylor 2006; Sarkar et al. 2006). Thus, individual strains have specific genetic repertoires that promote pathogenicity only when expressed in a particular environment, that is, that of a compatible host species (Lin and Martin 2005; Sarkar et al. 2006).

The *P. syringae* species complex offers a unique opportunity to study the contribution of HGT to the early stages of differentiation between lineages for a number of reasons. First, the radiation of *P. syringae* pathovars onto multiple different host species combined with the high degree of host specificity of individual pathovars allows for an investigation into the role that HGT may have in the partitioning of the *P. syringae* species complex into different ecological niches, where in this case each host species can be viewed as a distinct niche environment to which its *P. syringae* pathovar is specifically adapted. Second, the *P. syringae* species complex is an example of a group of bacteria in the early- to midstages of differentiation, with a number of clades (i.e., phylogroups) with a level of sequence divergence in core genes comparable with that of other "young" bacterial species. Finally, there exists a good representation of whole-genome sequence data across at least three major phylogroups within the complex, allowing for the accurate elucidation of genome content variation across closely related but phenotypically distinct lineages.

We utilize these data to quantify the extent of gene gain and loss within a phylogenetic framework for 27 lineages of the *P. syringae* species complex. Using available whole-genome sequence data and a suite of tools developed to limit the confounding effects of errors associated with draft genomes, we decompose the *P. syringae* pan genome into its constituent core and flexible components and apply various phylogenomic methods to infer the evolutionary histories of both these entities. We reconstruct the most likely history of gene gain and loss within the framework of the evolutionary history of the core genome and correlate the rate of gain and loss with phylogenetic depth to provide insights into the dynamics of genome fluctuation through time. We then classify gained genes based on both putative function and potential source to elucidate the functional diversity and the taxonomic breadth of donor species of gained genes. We use these results to ask if the divergence via differential adaptation model described above can explain the observed radiation of *P. syringae* pathovars onto different host species. If this were the case, we would expect to see the acquisition of genes along phylogenetic branches that correspond to host adaptation events. Of particular importance in this process are genes of the type III secretion system (T3SS). This "molecular syringe" is a fundamental component of the *P. syringae* armory that knocks down innate plant immune defenses via the injection of type III secretion effector proteins (T3SEs) into host cells, allowing for infection (Lindeberg et al. 2009, 2012). Thus, the repertoire of T3SEs possessed by a given pathovar is a key determinant of host specificity, and the gain of T3SEs may play an important role in the diversification of *P. syringae* lineages onto novel host species (Sawada et al. 1999; Ma et al. 2006; Lindeberg et al. 2009; Baltrus et al. 2011; O'Brien et al. 2012).

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution

SMBE

Materials and Methods

Genomic Data

All available *P. syringae* genome sequences (as of May 4, 2012) were downloaded from National Center for Biotechnology Information (NCBI) GenBank, including multiple strains of the same pathovar (but excluding multiple editions of the same strain). This yielded 27 strains of 18 pathovars from across the *P. syringae* species complex. Summary information for all genomes included in this analysis is given in table 1 and supplementary table S1, Supplementary Material online. All protein sequences used in subsequent analyses were based on the published annotations for these genome sequences, also retrieved from NCBI GenBank.

Assignment of Orthology

Orthologous relationships among proteins were determined using OrthoMCL version 2.0 (Li et al. 2003). OrthoMCL uses Basic Local Alignment Search Tool (BLAST) followed by Markov clustering (MCL) to group proteins into putative orthologous groups (OGs). A minimum length cutoff of 50 residues was chosen as many sequences less than 50 were found to be protein fragments (see below). The all-versus-all BLASTp was performed with an *E* value $\leq 1 \times 10^{-5}$, and the inflation index, which controls the threshold at which the MCL algorithm defines distinct OGs, was set at 1.5 (Li et al. 2003; Lefébure and Stanhope 2007; Lefébure et al. 2010).

Correction of Errors Associated with Draft Genome Sequence Data

The resultant list of OGs was subjected to a series of quality control procedures to minimize the confounding effects of errors such as misannotations and (in particular) the presence of protein fragments. Protein fragments are caused by the premature truncation of a protein-coding sequence (CDS) in the genome by the termination of a contig in the middle of an open reading frame. When a CDS spans a break between two adjacent contigs (i.e., a gap in a draft genome), it is possible to get both the 5′- and 3′-portions of the split gene being annotated as distinct entities (with unique protein accession numbers) within the proteome of that genome.

The presence of these fragments causes error in the inference of orthology via two primary mechanisms. First, fragments can cluster to form an OG that is distinct from the group containing their full-length true co-orthologs. This introduces an additional, spurious OG to the analysis and can lead to an overestimate in the number of gains and/or losses. Second, protein translations of both the 5′- and 3′-fractions of a split CDS may cluster within the same OG, giving rise to multiple proteins from the same genome within a given OG (i.e., apparent paralogs). This can lead to an underestimate of the core genome size when only genes in single copy per genome are counted. In addition, fragments that do not cluster in any OG can overestimate the number of lineage-specific proteins per pathovar. To account for these issues, a series of filters (written in Perl) were developed and are briefly discussed below.

To correct for entirely spurious OGs, the consensus sequences of all 2,358 2-member OGs ("twosomes") were used to search a database consisting of the consensus sequences of all 10,320 OGs inferred by OrthoMCL. Alignments were constructed using T-Coffee version 8.99 with default parameters (Notredame et al. 2000), and BLASTp was run at *E* value $\leq 1 \times 10^{-5}$. Query sequences with a significant hit ($\geq$80% amino acid identity over $\geq$80% query length) to another alignment were classified as spurious and the corresponding twosome OG was removed from the analysis. The process was repeated iteratively over increasing OG size classes (threesomes, foursomes, etc.) until the method inferred an error rate of $<$0.05% in the OG data set (at tensomes).

To correct for OGs containing multiple fragments of the same CDS, the alignments of all groups that had co-orthologous proteins from the same genome (i.e., apparent paralogs) were subjected to further investigation. For alignments that had an average pairwise amino acid identity of $\geq$80% and that showed evidence that the two paralogs were in fact fragments, only the longest fragment of any split protein was retained in the alignment. A further 499 OGs were removed from all subsequent analyses due to poor alignment and/or ambiguous orthology.

Any sequences with homology to any known insertion sequence-related proteins were removed from the data set by querying the 59 known *P. syringae* IS elements (Bardaji et al. 2011), downloaded from "IS Finder" at www-is.biotoul.fr/ (last accessed April 24, 2013), against the combined input proteomic data set (BLASTp, *E* value $\leq 1 \times 10^{-10}$). Any OG that contained a sequence with $\geq$70% pairwise amino acid identity was removed from subsequent analyses.

Construction of Core and Lineage-Specific Gene Sets

The resultant lists of OGs from the OrthoMCL analysis were decomposed into putative pan-, core-, and lineage-specific gene sets using a further set of custom Perl scripts. Only core genes with no paralogs (i.e., 1-to-1 orthologs) were included in the core gene set. Because the exclusion of a single pathovar from an "almost-core" OG necessarily excludes that gene from the core set as defined above, all 969 OGs that had a member protein in 26 of the 27 genomes were further scrutinized to ensure the nonexistence of an ortholog in the nonparticipating genome, using tBLASTn to account for the possibility of annotation errors. Briefly, this filter queried each nonparticipating genome with the corresponding ortholog from the pv. *phaseolicola* str. 1448A (*phas*) genome, using tBLASTn (*E* value $\leq 1 \times 10^{-5}$). All hits that had an 80% or greater amino acid identity over 80% or more of the query

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

Table 1
Summary Information for 27 *Pseudomonas syringae* Pathovars Used in this Study

Pathovar	Strain	Tag[a]	Species of Isolation	Contigs[b]	Proteins[c]
aceris	M302273PT	acer	*Acer* sp. (maple)	1,179	6,185
actinidiae	M302091	actn	*Actinidia delicosa* (kiwi fruit)	941	5,626
aesculi	2250	aes2	*Aesculus hippocastanum* (European horse chestnut)	776	5,619
aesculi	NCPPB 3681	aesN	*Aesculus indica* (Indian horse chestnut)	841	5,649
aptata	DSM 50252	apta	*Beta vulgaris* (sugar beet)	3,776	6,368
"cit7"	Cit7	cit7	*Citrus sinensis* (navel orange)	2,655	7,145
glycinea	B076	glyB	*Glycine max* (soybean)	104	5,580
glycinea	race 4	gly4	*G. max* (soybean)	108	5,207
japonica	M301072PT	japa	*Hordeum vulgare* (barley)	4,661	8,796
lachrymans	M301315	lacM	*Cucumis sativus* (cucumber)	791	6,691
lachrymans	M302278PT	lacP	*C. sativus* (cucumber)	798	5,636
maculicola	ES4326	Pcan	*Raphanus sativus* (radish)	878	6,101
mori	M301020	mori	*Morus alba* (white mulberry)	3,414	7,544
morsprunorum	M302280PT	mrsp	*Prunus domesticus* (European plum)	969	5,837
oryzae	1_6	oryz	*Oryza sativa* (rice plant)	2,855	7,837
phaseolicola	1448A	phas	*Phaseolus vulgaris* (common bean)	3	5,172
pisi	1704B	pisi	*Pisum sativum* (pea)	5,099	9,160
savastanoi	NCPPB 3335	sava	*Olea europaea* (olive tree)	403	4,898
syringae	B728a	syrB	*P. vulgaris* (common bean)	1	5,089
syringae	FF5	syrF	*Pyrus calleryana* (ornamental pear)	4,578	8,445
syringae	642	syr6	Unspecified	296	5,274
tabaci	ATCC 11528	taba	*Nicotiana tabacum* (tobacco plant)	1,405	6,467
tomato	DC3000	tomD	*Solanum lycopersicum* (tomato)	3	5,619
tomato	K40	tomK	*S. lycopersicum* (tomato)	582	5,853
tomato	Max13	tomM	*S. lycopersicum* (tomato)	349	5,749
tomato	NCPPB 1108	tomN	*S. lycopersicum* (tomato)	304	5,619
tomato	T1	tomT	*S. lycopersicum* (tomato)	122	5,702

Note.—More information, including genome references, can be found in supplementary table S1, Supplementary Material online.
[a]Four-letter identifier used in this study.
[b]The number of contiguous sequences given for each genome.
[c]The number of proteins given for each genome.

length were considered homologous. When pv. *phaseolicola* itself was the nonparticipating genome, then the corresponding ortholog from pv. *tomato* str. DC3000 (*tomD*) was used.

All sequences that did not cluster to form an OG were compiled into an initial lineage-specific gene set. The presence of protein fragments (discussed above) is known to erroneously inflate this gene set (Lefébure and Stanhope 2007; Lefébure et al. 2010; Richards et al. 2011). Thus, a protein was considered lineage specific only if it returned no significant hits (other than to self) when queried against the combined genomic data set, using tBLASTn (*E* value $\leq 1 \times 10^{-10}$).

A binary matrix of gene presence or absence across all OGs and pathovars was used to plot gene accumulation curves with a custom R script; for each comparison, genome input order was randomly permuted 100 times to acquire sample distributions.

Core Genome Phylogenetics

The genome-wide average evolutionary history for these *P. syringae* strains was inferred using a core genome concatenation approach. For each core OG, protein sequences were aligned using T-Coffee. All sites that contained a gap in any sequence were excluded. The presence of truncated sequences in many alignments meant that the exclusion of gap sites sometimes ruled out the entire alignment (i.e., when all alignment columns had at least one gap). Alignments with low identity (<30% identical sites), short length (<30 residues), or 100% identity were not included in the concatenation. The remaining 2,140 alignments were concatenated, and a maximum-likelihood tree was built with PhyML version 3.0 (Guindon and Gascuel 2003; Guindon et al. 2010) using the Le and Gascuel (2008) model of amino acid substitution with four gamma-distributed rate categories (LG + G) and a combination of nearest-neighbor interchange (NNI) and subtree-prune-regraft (SPR) methods of tree topology search; 100 bootstrap replicates were performed to test topological robustness. A maximum-likelihood phylogeny based on over 1 million core nucleotide sites was also constructed in PhyML, using a GTR + G model with four rate categories, NNI + SPR branch swapping, and 100 bootstrap replicates.

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Maximum-likelihood phylogenies for all 2,140 individual core alignments were constructed at both the DNA and protein level, using the same parameters as above. Topological concordance among the set of phylogenies was examined by using the program "consense" from the Phylip package (Felsenstein 1981) to build a consensus tree (extended majority rule).

Genome Content Analysis

To estimate a phylogeny based on genome similarity in terms of gene content, a distance matrix was constructed where the distance measured between two taxa is inversely proportional to the number of genes shared between them (Snel et al. 1999; Kettler et al. 2007; Georgiades et al. 2011). If gain and loss are assumed to affect the flexible genome only, the distance between genome A and B (d_{AB}) can be calculated as follows: $d_{AB} = 1-((S_{AB} - C)/(T - C))$, where S_{AB} is the number of shared genes, C is the core genome size, and T is the total number of proteins in the smaller of the two genomes. To correct for unobservable events along deeper branches in the phylogeny, a Poisson correction procedure was applied as follows: $d'_{AB} = -\ln(1 - d_{AB})$. A phylogeny was then inferred using the neighbor-joining method (Saitou and Nei 1987) with 1,000 bootstrap replicates sampling columns of the presence/absence matrix, and recording the consensus support over the bootstrap data sets at each node using consense.

Levels of synonymous (K_S) and nonsynonymous (K_A) divergence between all pairs of taxa within and between the major clades in the core-genome phylogeny were estimated for the same 2,140 core nucleotide alignments (over 1 million sites), using the method of Li (1993), implemented through the R package "SeqinR" (Charif and Lobry 2007).

Reconstruction of Gene Gain and Loss

The history of gene gain and loss was analyzed using a stochastic mapping method implemented through the GLOOME server (http://gloome.tau.ac.il/, last accessed October 23, 2013) (Cohen et al. 2010; Cohen and Pupko 2010, 2011). The GLOOME algorithm takes as input the phyletic matrix of gene-family presence and absence and a reference phylogenetic tree, and infers probabilities and expectations for all gain and loss events in a per site per branch manner. This method may be particularly suitable for the analysis of genomic data sets that are split into numerous contigs (i.e., draft genomes), because it does not require gene synteny information. In addition, other methods that reconcile differences between a gene tree and a species tree to infer duplication, transfer, and loss events often require fully binary trees—a condition that is often not met in comparative genomic analyses of closely related taxa where there may be poor phylogenetic resolution at the level of the individual gene. Stochastic mapping allows for the total number of gains and losses along a given phylogenetic branch to be calculated, as well as the subsequent extraction of specific gain events that have a high associated probability. Probabilistic models of gene gain and loss are superior to parsimony-based approaches because the rate of gain and loss is allowed to vary among OGs, and both topology and branch lengths are considered in the estimation of associated probabilities (Cohen and Pupko 2010). The core genome phylogeny was used as the reference tree, and the total number of gains for each branch was calculated as the probability of gain for each gene, summed across all genes and rounded to the nearest whole number. The procedure was repeated for losses. A branch-specific rate of gene gain/loss was also calculated as the ratio of the number of events (gains or losses) to the length of that branch, equivalent to the number of events per 1% amino acid sequence change of the core genome.

Using the stochastic mapping results, genes with a gain probability of $\geq$0.8 were extracted on a branch-by-branch basis for functional categorization. A cutoff of 0.8 was chosen because this was the lowest observed probability value with which lineage-specific genes were inferred to have been gained on the terminal branches of the tree, that is, a more conservative cutoff value may have excluded some lineage-specific genes from the analysis. For each branch, a reference sequence from each gained OG was chosen randomly and functionally annotated using Blast2GO version. 2.5.0 under default settings (Conesa et al. 2005). Gained sequences were initially classified into one of three (mutually exclusive) basic categories: "functional protein" if the sequence was successfully annotated within Blast2GO, "hypothetical protein" if no function was ascribed, and "phage related" if that sequence had matches to known phage-related proteins. Those sequences that were ascribed functions were further delimited based on their associated gene ontology (GO) terms. Only branches with $\geq$10 gains were analyzed, and only GO categories with $\geq$3 sequences are shown. The functional protein category was chosen for further GO classification to ascertain the underlying functional diversity of gained sequences.

The potential source of recently transferred genes was investigated using a BLAST-based approach. Protein sequences for all genes gained along terminal branches (i.e., recent acquisitions) with a gain probability of $\geq$0.8 were queried against the NCBI nonredundant database using BLASTp (E value $\leq 1 \times 10^{-20}$). For each query, the donor species was inferred to be the most similar match with the lowest BLAST E value. This approach will miss potential donor species that are not represented within the database but allows for an approximate estimation of the extent to which HGT occurs within phylogroups, between phylogroups or from other species.

To assess the extent to which effector genes of the T3SS comprise recently acquired genes, all terminally gained proteins were queried against a database of all currently described type III secretion system effector proteins (T3SEs)

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

using BLASTp. The database was constructed using the T3SE sequence information provided at www.pseudomonas-syringae.org (last accessed August 13, 2013) and by Wang et al. (2012) and consists of >1,700 T3SS-related protein products. Matches with ≥80% identity over ≥80% of the query length were inferred to be homologous to a T3SE.

Results

The Core and Pan Genomes of *P. syringae*

We compared the genome sequences of 27 strains of *P. syringae*. Initially, these showed a combined total of 159,748 proteins, of which 148,254 (~93%) were clustered into a total of 10,320 inferred OGs by OrthoMCL. The relative frequency across genomes of each OG size class is shown before and after correction for errors (fig. 1). Prior to correction, 11,791 genes (~53% of the initial estimate of the pan genome) were inferred to be lineage specific. After accounting for errors due (primarily) to the presence of protein fragments, this value was reduced by more than three-quarters to only 2,677 genes (~12% of the initial estimate of the pan genome). The shape of the corrected distribution suggests that the majority (~75%) of genes of the pan genome are found in either very few (≤3) or most (≥25) genomes, with a paucity of genes distributed in the mid range of the distribution.

After correction, the size of the pan genome for all 27 lineages was 11,025 genes (fig. 2, upper panel). However, the curve did not tend toward an asymptote, suggesting that a considerable amount of variation (in terms of novel genes) remains to be discovered with the addition of further genomes. The core genome of the entire *P. syringae* species complex was initially estimated to be 2,421 genes. Correction for missed core genes increased this value by 174 to 2,595 genes or approximately 52% of the average *P. syringae* genome content. Again, the shape of the curve suggests that the number of core genes will decrease further given additional genome sampling. We considered the effect of missed genes due to incomplete genome sequences by running an analysis where almost core genes (with a member in $n-1$ genomes) were included. This increased the estimated core genome to 3,364 genes but had little effect on the shape of the accumulation curve, suggesting that the failure of the original curve to tend toward an asymptote is not due to missed genes. When analyzed individually, pan genome estimates for phylogroups 1, 2, and 3 were 6,778, 7,436, and 7,790 genes, respectively, and estimates for the core genome were 3,785, 3,481, and 3,477 genes, respectively (fig. 2, lower panel). These values are substantially lower (pan) and higher (core) when compared with the corresponding value under the appropriate number of genomes in the full analysis.

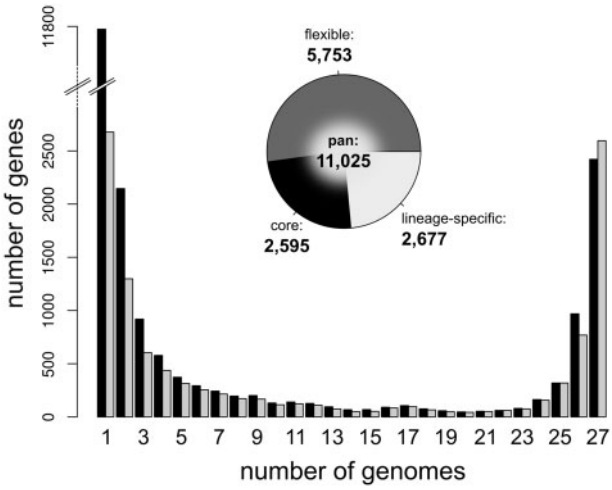


Fig. 1.—Frequency distribution of genes across genomes. Black bars represent the number of genes in each OG size category prior to the various correction procedures outlined in Materials and Methods; gray bars are corrected values. The pie chart shows the delineation of the *Pseudomonas syringae* pan genome into its constituent entities: Core (i.e., genes present in all strains), flexible (i.e., genes with a variable occurrence across strains), and lineage specific (i.e., genes unique to individual strains).

Pseudomonas syringae Phylogenetics

The core genome phylogeny for *P. syringae* was estimated from 2,140 concatenated proteins totaling 351,468 aligned residues (fig. 3). Pathovar *maculicola* str. ES4326 (*Pcan*) was defined as the outgroup to root the core genome phylogeny, because it has been reported that this pathovar was originally misidentified and is in fact a strain of *P. cannabina* (Baltrus et al. 2011), a closely related species outwith the *P. syringae* species complex (Parkinson et al. 2010). Bootstrap support of branching patterns was generally very high with only one partition showing less than 75% support. The phylogeny delineates the 26 ingroup lineages into four clades, in agreement with the phylogroup notation used previously (Sarkar and Guttman 2004; Hwang et al. 2005). A phylogeny constructed from nucleotide sequence alignments of the same 2,140 genes produced a similarly well-resolved tree with the same topology (not shown).

Based on the core genome, maximum pairwise values for K_A (the number of substitutions per nonsynonymous site) within phylogroups were calculated to be 0.018, 0.022, and 0.011 for phylogroups 1, 2, and 3, respectively. Between phylogroups, average estimates were 0.100, 0.094, and 0.067 for phylogroup 1 versus 2, 1 versus 3, and 2 versus 3, respectively. Maximum pairwise values for K_S (number of substitutions per synonymous site) within groups were 0.299, 0.312, and 0.124 for phylogroups 1, 2, and 3 respectively, whereas average K_S between groups were 1.584, 1.644, and 1.218 for phylogroup 1 versus 2, 1 versus 3, and 2 versus 3, respectively.

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution

SMBE

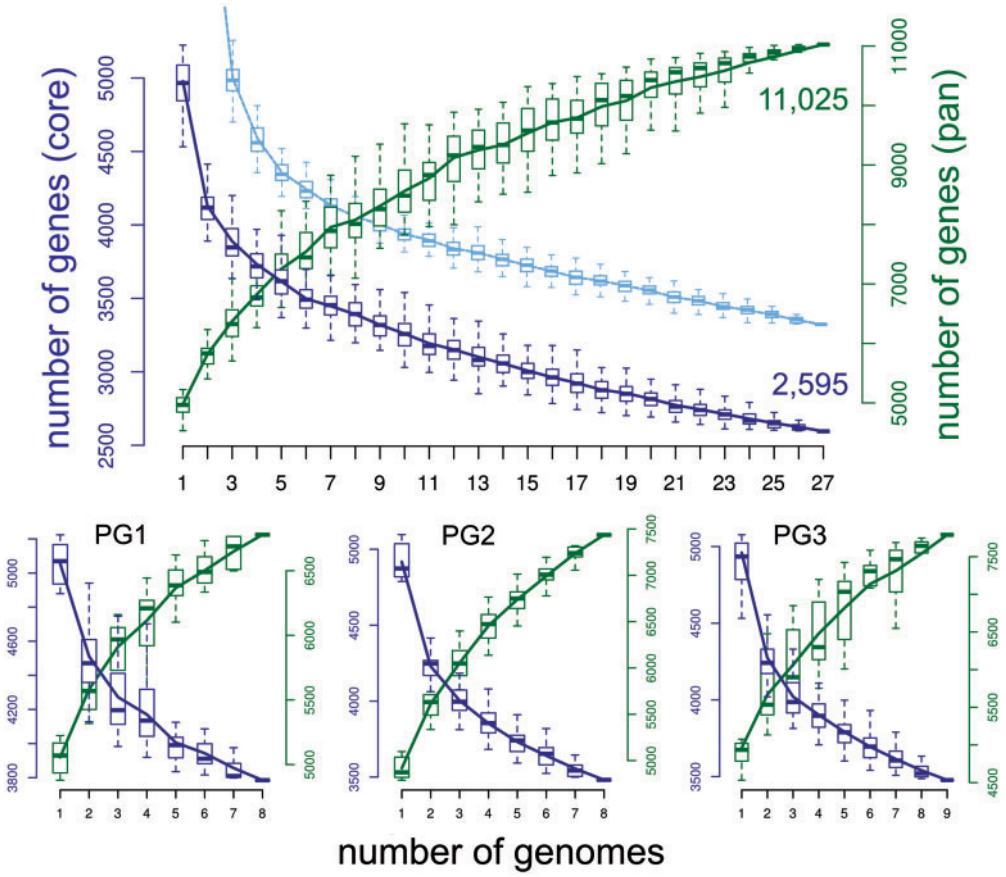


Fig. 2.—Core and pan genomes of *Pseudomonas syringae*. The upper figure shows gene accumulation curves for the core (dark blue) and pan (green) genomes for the combined data set (27 genomes). The light blue curve shows the core estimate when almost core genes (i.e., genes present in 26 genomes) are also included (comparisons of <3 genomes are not shown for this curve). The dark blue axis on the left shows the number of genes of the core genome, and the green axis on the right shows the number of genes of the pan genome (note the different scales). Boxes show the interquartile range of 100 random permutations of genome input order for the given number of genomes in comparison; the horizontal line shows the median value. The lower figure shows the equivalent results when the data are partitioned into the three major phylogroups. PG1, PG2, and PG3 represent phylogroups 1, 2, and 3, respectively.

An analysis of consensus branching patterns within the set of individual core gene trees was performed to assess the level of support for the topology of the concatenated phylogeny at the individual gene level. An extended majority consensus tree is shown in supplementary figure S1, Supplementary Material online. Consensus support for phylogroups 1, 2, and 3 was 90%, 79%, and 86%, respectively, and overall, the consensus topology is in agreement with that of the core genome phylogeny at all partitions bar the relative placement of the two phylogroup 2 pathovars *cit7* and *syr6*; this discrepancy is reflected in the low bootstrap value (54%) at this node in the core genome tree. Two-thirds of gene trees support the clustering of phylogroups 2 and 3 but only 40% support the position of PG1 as a sister to these. This appears to be due to uncertainty about the placement of PG4 (represented by *oryz*); indeed, an earlier analysis based on four genes clustered PG4 with PG2 and PG3 (Hwang et al. 2005).

Relationships among pathovars based on genome content similarities were also analyzed (fig. 4). Overall, there is a high degree of topological congruence between the genome content analysis and the core genome phylogeny. Phylogroups 1, 2, and 3 are reproduced monophyletically with 100% bootstrap support, while there is also substantial agreement in branching patterns within phylogroups. An exception is the placement of the PG4 strain *oryz*, which clusters with PG2 in the genome-content analysis (red asterisk). The major difference between the core genome and genome content phylogenies can be seen in the relative branch lengths with respect to phylogenetic depth within the two trees. Even after correction for multiple hits, the deeper branches in the genome-content analysis are very short, whereas the terminal branches are relatively extended.

Gene Gain and Loss

We used a stochastic mapping method to infer the number of gains and losses along all branches of the *P. syringae* core genome phylogeny (fig. 5). An underlying assumption of

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

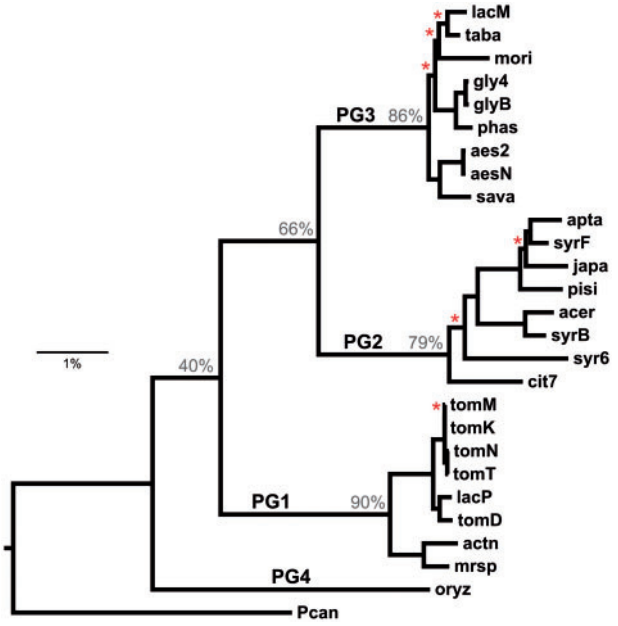


Fig. 3.—Core genome phylogeny for *Pseudomonas syringae*, derived from 2,140 concatenated protein alignments. All internal branches have 100% bootstrap support except where indicated with asterisks. Consensus support from individual gene trees is shown with gray numbers on major branches, and phylogroups PG1–PG4 are indicated. Tree is rooted with pv. *maculicola* str. ES4326 (*Pcan*). Scale bar indicates 1% amino acid replacements per site.

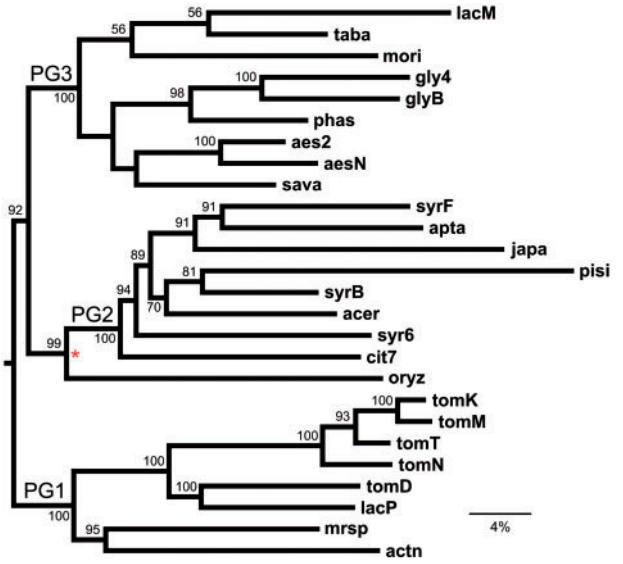


Fig. 4.—Genome content analysis. Numbers indicate percentage consensus support for that bifurcation from 1,000 bootstrap replicates. Phylogroups PG1, PG2, and PG3 are indicated. The red asterisk shows the only incongruence relative to the core genome phylogeny. Scale bar indicates 4% genomic dissimilarity, that is, percentage difference in terms of number of genes shared. Tree is rooted with *Pcan* (not shown).

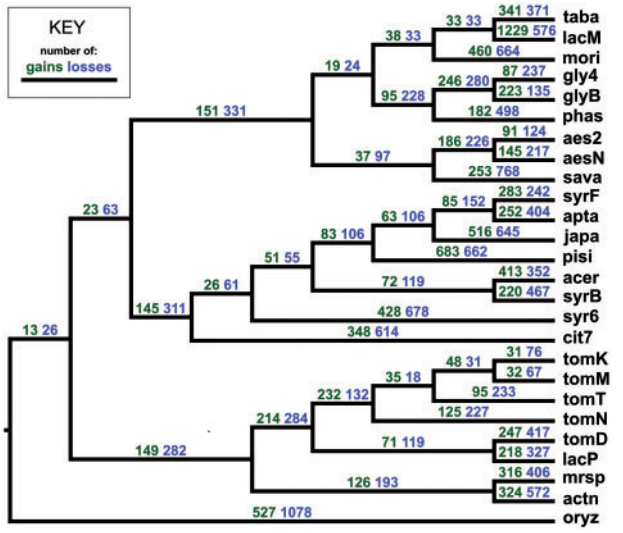


Fig. 5.—Gene gain and loss in *Pseudomonas syringae*. Inferred numbers of gains (green) and losses (blue) are given above each branch. Lineage-specific genes are included in gains along terminal branches. Values in all cases are rounded to the nearest whole number of events. Topology is based on figure 3; branch lengths are not to scale.

this method is that gain and loss events occur independently of one another. In reality, this is unlikely to be the case, because many genes may be gained (or lost) in a single transfer event (e.g., the acquisition of a plasmid). Thus, the results shown here likely overestimate the number of individual transfer events involved.

There is considerable variation in gene gain and loss among branches. Generally, values are higher along terminal branches and decrease along branches deeper in the phylogeny. The pathovar with the highest inferred number of terminal gains was *lacM* (1,229), followed by *pisi* (683) and *oryz* (527). On average, 6.4% of a pathovar's total genome has been gained "recently" (i.e., subsequent to the divergence of a lineage from its nearest sister taxa), but this value increases to 13.2% when considering this proportion relative to the size of the flexible genome only. On terminal branches, lineages have lost 9.2% of the current genome size on average, and the three pathovars that have undergone the greatest number of losses are *oryz* (1,226), *sava* (768), and *syr6* (678).

High values of gain and loss for individual pathovars may be the result of either biologically relevant variation in genome content or an artifact of phylogenetic sampling. For example, *P. syringae* strains are known to harbor plasmids and the exceptionally large number of recent gains in *lacM* can be explained by the discovery of an approximately 1 Mb plasmid in this genome (Baltrus et al. 2011). Similarly, evidence for the presence of plasmids has been found for the *pisi* genome (Baltrus et al. 2011), although the exact size and gene content are not known. However, although *oryz* is thought to contain

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution

SMBE

plasmids (Reinhardt et al. 2008), the high number of gains and losses seen in this pathovar is also due to the relatively long branch that separates it from other pathovars; *oryz* has had more time to accumulate differences in genome content.

However, comparisons between absolute numbers of gains and losses per branch can be misleading, because these values are subject to the effects of phylogenetic sampling. To account for this, we calculated a rate of gain and loss relative to the amount of amino acid sequence divergence that has occurred along each branch of the core genome phylogeny (supplementary fig. S2, Supplementary Material online). To investigate the evolutionary dynamics of gain and loss at different depths of the phylogenetic tree, we plotted branch-specific rates against the phylogenetic depth of each branch, measured as the average distance from the midpoint of that branch to all descendent tips (fig. 6). The apparent rates of gain and loss decrease exponentially with respect to increasing phylogenetic depth, reflecting the much higher levels of observable gain and loss along branches closer to the tips of the tree (i.e., more recent in time).

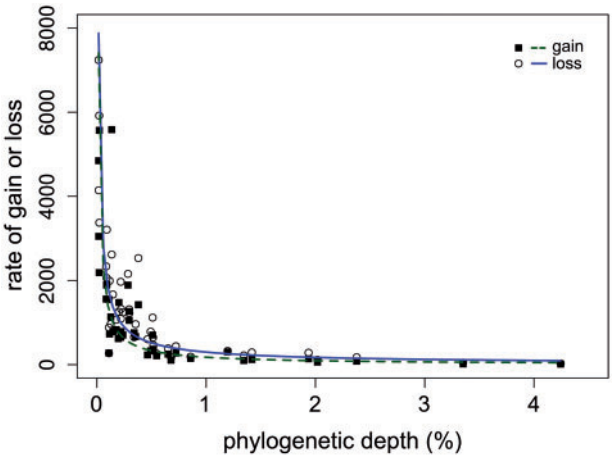


Fig. 6.—Rate of gene gain (filled squares) and loss (open circles) versus phylogenetic depth. Phylogenetic depth is calculated as the average distance (percent amino acid replacements per site) from the midpoint of each branch of the core genome phylogeny to its descendent tips; the *y* axis shows the branch-specific rate of gain or loss expressed as the number of inferred events per 1% amino acid divergence of the core genome phylogeny.

Source and Functions of Gained Genes

The extent of HGT within and between phylogroups was assessed using BLAST to determine the most similar sequence for all genes gained along terminal branches across the phylogeny (fig. 7). In phylogroups 2 and 3, the primary source of gained genes appears to have been other pathovars from within the same phylogroup. A significant contribution also comes from pathovars in other phylogroups, as well as from other *Pseudomonas* species outwith the *P. syringae* group. Taking phylogroups 1, 2, and 3 together, the ten most frequent putative donor genera were (in rank order) *Pseudomonas*, *Burkholderia*, *Vibrio*, *Escherichia*, *Xanthomonas*, *Ralstonia*, *Yersinia*, *Salmonella*, *Acidovorax*, and *Pectobacterium*. Transfers from these bacterial groups represent the great majority (~86%) of the total number of recently transferred genes, and many contain species that are known to be associated with plant, water, or soil habitats. For example, both *Ralstonia* and *Xanthomonas*, themselves much-studied pathogens of other plant species (Mansfield et al. 2012), appear as frequent donors of genetic material into the genomes of phylogroup 2 and 3 pathovars. In addition, *Pectobacterium* contains phytopathogenic species that cause disease in both herbaceous and woody hosts (Toth et al. 2003), *Acidovorax* are pathogens of cucurbits (Schaad et al. 2008), whereas *Polaromonas* species are often associated with water and can be found in alpine meltwater (Jeon et al. 2004; Margesin et al. 2012), a known habitat of *P. syringae* (Morris et al. 2008). Species belonging to the Burkholderiaceae, from the β-proteobacteria, are among the top five putative donors from outwith the Pseudomonads for all three phylogroups, highlighting the potential contribution from relatively distant families.

Further analysis of putative donor species from other genera reveals the phylogenetic breadth of potential HGT sources (fig. 7). For all phylogroups, the majority of genes are donated from species within the γ-proteobacteria. A smaller proportion of genes are also acquired from more distant taxonomic groups such as the α- and β-proteobacteria and from other phyla such as the cyanobacteria. A significant fraction of genes in all phylogroups show "no hit," indicating that many recently gained genes have not been previously sampled and are of unknown origin or function. The proportion of no hits in PG3 is approximately twice that relative to PG1 and PG2 due to the inflated number of genes specific to *P. syringae* pv. *lachrymans* str. M301315 (*lacM*), presumably encoded on the 1 Mb plasmid unique to this strain, the majority of which had no BLAST matches. The source of this large lineage-specific fragment of DNA remains to be elucidated.

Across all branches, the protein sequences of 6,378 genes gained with a probability of $\geq$0.8 were delimited into three basic categories based on OG: Functional protein, hypothetical protein, and phage-related protein (supplementary fig. S3*a*, Supplementary Material online). Across all branches, 59.6%, 37.8%, and 2.6% of gains were labeled as functional, hypothetical, or phage related, respectively. A further categorization of those sequences that were successfully annotated (i.e., functional proteins, as above) is provided in supplementary figure S3*b*, Supplementary Material online.

Proteins inferred to be terminally acquired were queried versus a database of known T3SEs to ascertain the extent of recent transfer of this important family of genes. At the 1×10^{-5} BLAST threshold, this yielded 13, 34, and 36 hits

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

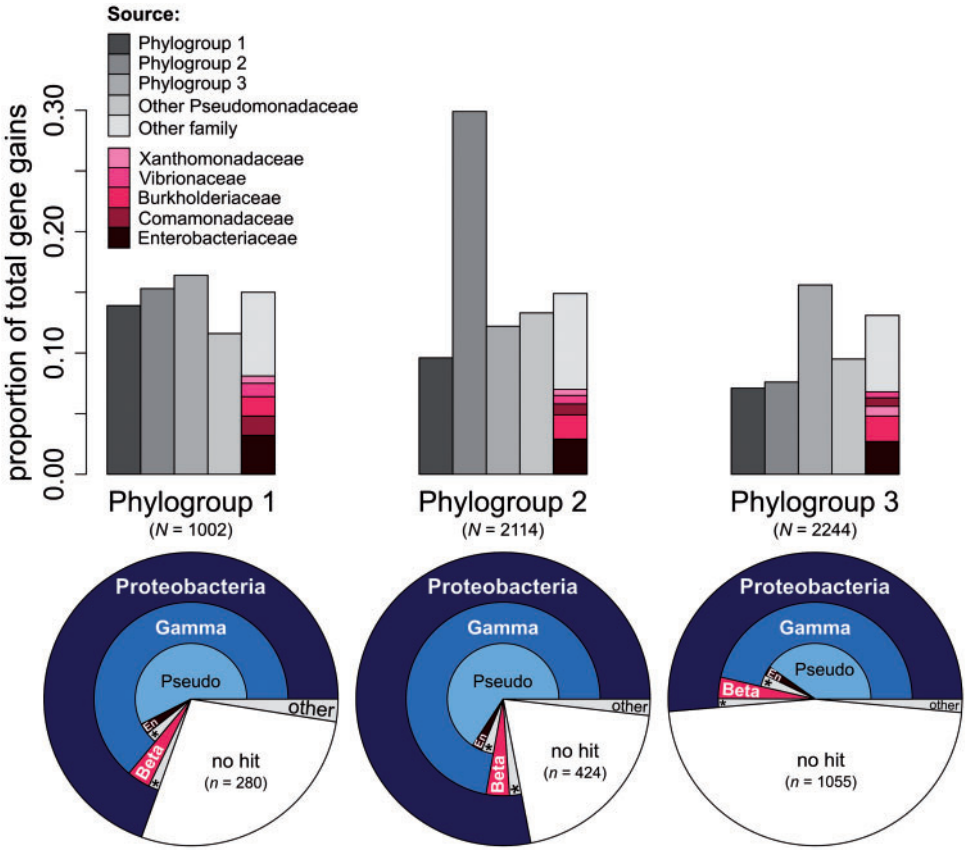


Fig. 7.—Sources of gained genes. Bar charts detail the fine-scale sources of genes gained recently in phylogroups 1, 2, and 3, as identified by best BLAST hit (*E* value $\leq 1 \times 10^{-20}$). The *y* axis represents the proportion of total terminally acquired genes for each phylogroup. The category "other family" is further decomposed to show the family of the top five donors for each phylogroup (pink shaded boxes). Pie charts below show broader taxonomic classification of gained genes. Pseudo, Pseudomonadaceae; En, Enterobacteriaceae; asterisks denote "other" categories.

for phylogroups 1, 2, and 3 respectively. However, applying the 80% identity over 80% query coverage homology rule narrowed this down to 8, 7, and 12 T3SEs for phylogroups 1, 2, and 3, respectively, accounting for a very small fraction (~0.5%) of all recently gained genes. The terminal acquisition of T3SEs was not distributed evenly among pathovars. For example, within phylogroup 3, six were acquired along the terminal branch leading to *mori*, four along the branch leading to *lacM*, and two along the branch leading to *phas*. Only one T3SE protein, EGH71060.1 from *acer*, was both terminally acquired and strain specific.

Discussion

We have investigated both the scale and dynamics of genome fluctuation in *P. syringae*. We demonstrate a history of extensive gene gain and loss that can rapidly alter the gene content of individual *P. syringae* lineages. Below, we discuss the wider implications of these results within the context of the role of HGT and gene loss in the differentiation and diversification of bacterial lineages, as well as the consequences of such fluctuation on the evolution of *P. syringae*. First, we consider some of the potential issues surrounding the use of whole-genome draft sequence data in comparative analyses that may have been previously underestimated in other studies.

The Extent of Potential Errors in Draft Genomes

We applied a suite of quality control and data correction procedures to ensure the resultant patterns of orthology and gene presence/absence were as accurate as possible. The confounding effects of protein fragments, caused when a contig terminates in the middle of a CDS, were of particular significance. Our observations emphasize the care required when drawing inferences from uncorrected patterns of orthology constructed using multiple (often hundreds of) contig draft genome sequences in conjunction with popular orthology methods such as OrthoMCL.

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution

SMBE

Not accounting for this source of error can introduce considerable bias in a number of common comparative genomic analyses. For example, we estimated that approximately 77% of inferred lineage-specific proteins and approximately 13% of the precorrected number of OGs could be apportioned to error. Even fundamental descriptive metrics for a genome, such as the number of proteins it encodes, can be substantially overestimated by the presence of protein fragments. For example, the combined data correction procedures applied here reduce the number of proteins encoded by the *P. syringae* pv. *pisi* genome from 9,160 to 5,085, a reduction of approximately 44% (i.e., almost one-half of the original number). The mean reduction in proteome size across the 27 genomes in this study is approximately 17%. These observations simultaneously reveal the possible level of error in the annotation of genome sequences and the potential for the use of orthology to reduce the problem.

Errors in the observed patterns of orthology among proteins can adversely affect the estimation of core- and pan-genome sizes and can lead to the miscalculation of the level of paralogy and gene duplication within genomes. For example, the core genome of *P. syringae* was recently estimated at 1,856 genes, approximately 740 genes smaller than that reported here (O'Brien et al. 2012). Problems caused by protein fragments may be one reason for this discrepancy. In addition, a recent study on genome fluctuation in cyanobacteria suggests that a large number of paralogs within the genomes of certain groups of cyanobacteria may be indicative of a higher rate of gene duplication (Larsson et al. 2011). However, it is also possible that a large fraction of the apparent paralogs may be attributable to the presence of protein fragments.

Phylogenetics of the Core and Flexible Genomes

We inferred the evolutionary relationships among pathovars based on both the evolution of the core genome and gene repertoire similarity of the flexible genome. Genome content similarity has been predicted to be a poor indicator of phylogenetic relatedness, where the rapidity of gene turnover is expected to blur inherited patterns of gene presence and absence (Hao and Golding 2006; Vernikos et al. 2007; Touchon et al. 2009). Thus, topological discordance between phylogenies based on these two genome components is thought to be an indicator of extensive HGT among lineages (Shapiro et al. 2012). However, the congruence observed between the *P. syringae* core genome and genome content phylogenies suggests that, despite high rates of HGT and gene loss, an accurate and robust phylogenetic signal is recoverable from patterns of presence and absence among genes of the flexible genome.

Varying degrees of topological congruence between the core and flexible components of bacterial genomes have been reported elsewhere. For example, an investigation of genome evolution in *Prochlorococcus* (Kettler et al. 2007) shows near perfect congruence, whereas studies of other bacteria, for example, *Vibrio* (Shapiro et al. 2012) and *Rickettsia* (Georgiades et al. 2011), show near total discordance. Reasons for this variation may include differences in the rate and extent of genome fluctuation experienced by these groups, the level of divergence among lineages, or perhaps ecological factors associated with these different species. However, both *Prochlorococcus* and *Vibrio* are ocean-dwelling species where the opportunity for HGT is presumably relatively high, yet they show opposite patterns. Moreover, *Rickettsia* species are obligate intracellular parasites of eukaryotic hosts with presumably relatively limited opportunity for gene gain. However, *Rickettsia* genomes are known to be shrinking in size (Georgiades et al. 2011), and so it may be the case that the genome content topology is driven by large-scale differential gene loss as these species adapt to different hosts. Patterns of divergence also fail to correlate with the level of congruence. These groups of bacteria are generally highly similar, having 16S ribosomal RNA nucleotide identities ranging between approximately 97% (*Prochlorococcus*) and 100% (*Vibrio*). Contrary to what might be expected, the group with the greatest 16S divergence (*Prochlorococcus*) shows the most congruence between phylogenies despite presumably having more time to accumulate differences in genome content. Furthermore, *P. syringae* and *Rickettsia* differ dramatically in the level of congruence despite having similar 16S identities (both ~99%). In all cases, the lack of signal seems not to be the issue, because all groups show considerable variation in genome repertoires among lineages (Kettler et al. 2007; Georgiades et al. 2011; Shapiro et al. 2012). Certainly, *P. syringae* provides a clear example that a high level of genome fluctuation does not preclude a high degree of congruence between trees derived from the core genome and genome content comparisons.

Species Definition in *P. syringae*

On the basis of over 1 million orthologous nucleotide sites of the core genome, we estimate that the three *P. syringae* phylogroups are as diverged from each other as other taxa classified as separate species or even genera. For example, K_A between the sister clade phylogroups 2 and 3 is estimated to be 0.06 substitutions per site. This value is greater than that for a number of other species comparisons, including *Escherichia coli–Salmonella enterica* and *Neisseria gonorrhoeae–N. polysaccharea* (Ochman et al. 1999). K_A divergence for comparisons between phylogroups 1 and 2/3 are greater yet and are comparable with K_A between two other "bona fide" *Pseudomonas* species, *P. aeruginosa* and *P. putida* (see fig. 2 of Ochman et al. 1999). Values of K_S between phylogroups are all >1, suggesting that substitution at synonymous sites is approaching saturation. Although high K_S values are unreliable as point estimates, species pairs in a number of other genera (including *Buchnera*, *Mycobacterium*,

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

Bacillus, and *Pseudomonas*) all have lower (<1) estimates of K_S (Ochman et al. 1999). In addition, consensus analysis from the set of individual core gene trees shows that a high percentage of core genes support the delineation of these lineages into three major groups (90%, 79%, and 86% for phylogroups 1, 2, and 3, respectively), suggesting that the recombination of core alleles between phylogroups does not occur at a level sufficient to erode the boundaries between groups. As such, and notwithstanding the ongoing debate as to the validity of such terms (Lawrence 2002; Doolittle and Papke 2006; Caro-Quintero and Konstantinidis 2011; Hanage 2013), phylogroups may be effectively described as nascent species.

There has been ongoing debate regarding the delimitation of the *P. syringae* complex into distinct species (Gardan et al. 1992; Sarkar and Guttman 2004; Bull et al. 2010). Sarkar and Guttman (2004) suggest that due to the ecological similarity between pathovars across phylogroups, the sharing of core alleles among phylogroups, and evidence of extensive HGT between phylogroups, *P. syringae* should remain as a single species. However, the results presented here suggest that these factors may not necessarily hinder speciation, because neither is sufficient to halt the accumulation of divergence from either the core or the flexible components of the *P. syringae* pan genome. We suggest that the data presented in this study support the delineation of *P. syringae* into a number of distinct species based on the phylogroup partitioning that is already well established. The strains *gly4*, *glyB*, and *sava* within phylogroup 3 have been previously referred to as strains of *P. savastanoi* (Bull et al. 2010; Rodríguez-Palenzuela et al. 2010; Qi et al. 2011; Ramos et al. 2012); clearly, the other pathovars within phylogroup 3 (*taba*, *lacM*, *mori*, *phas*, *aes2,* and *aesN*) should be classified in the same species (also see Bull et al. 2010, who state that phylogroup 3 strains should be classified as *P. amygdali* if unification occurs, because *P. savastanoi* and others are later synonyms). The type strain for *P. syringae*, NCPPB 281, clusters in phylogroup 2 along with the complete genome *P. syringae* pv. *syringae* str. B728a (*syrB*) (Parkinson et al. 2010), whereas it appears that species names for phylogroups 1 and 4 have not yet been proposed.

Genome Repertoire Dynamics and Niche Adaptation in *P. syringae*

The quantification of gene gain and loss suggests that both have been extensive in the evolutionary history of the *P. syringae* species complex and that the genomes of these pathovars may undergo considerable fluctuation even over relatively short evolutionary time periods. A large fraction of the total gain and loss was observed along terminal branches leading to individual pathovars, in agreement with previous studies investigating genome dynamics in other bacterial species (Hao and Golding 2006; Kettler et al. 2007; Lefébure and Stanhope 2007; Touchon et al. 2009).

When absolute numbers of gene gains and losses are compared with the rate of sequence divergence in proteins encoded by the core genome, the results are striking. At shallower phylogenetic depths (representative of terminal branches), we estimate that hundreds or thousands of genes may be gained (and lost) in the same time period as a 1% amino acid divergence. For example, the terminal branch leading to the complete genome of *P. syringae* pv. *phaseolicola* str. 1448A (*phas*) has gained 182 genes over a divergence of about 0.25% (equivalent to 728 genes over 1% divergence). Estimates of rates of gene loss are of at least the same magnitude. However, across the *P. syringae* complex as a whole, more than 50% of the genome remains as core, despite divergence values up to nearly 10%. This suggests, in agreement with other studies, that the majority of gained genes are transient and are quickly discarded from recipient genomes (Tettelin et al. 2005; Hao and Golding 2006; Kuo and Ochman 2009; Touchon et al. 2009). This leads to a perspective bias, where the rates of gene gain and loss both appear to decrease with increasing phylogenetic depth, because over longer time periods only a fraction of older gains have persisted and remain observable (Rocha 2008).

Based on seven closely related genomes of the *Bacillus cereus* group with divergences of the order of approximately 1% nucleotide substitutions per site, Hao and Golding (2006) estimated the recent rate of turnover to be such that a gene may be gained and lost approximately 5 times during the time required to observe a single-nucleotide substitution per site and conclude that the rate of transfer is at least comparable with that of nucleotide replacement. It is clear that the estimated rate of gene turnover relative to sequence divergence is highly dependent on the phylogenetic depth over which it has been calculated (fig. 6). However, using a core genome tree based on nucleotide sequences, we calculate the rate for *P. syringae* terminal branches that are equivalent in length to that of the *B. cereus* group to be perhaps four orders of magnitude greater. This highlights the predominance of HGT as an adaptive force for *P. syringae*, as well as the potential scale of the difference in rates of HGT between different bacterial groups.

We also observe an excess of gene losses relative to gene gains for the majority of branches of the core genome phylogeny. Taken at face value, this suggests a recent contraction of the *P. syringae* genome. However, given that extant strains of *P. syringae* have above-average sized genomes compared with other γ-proteobacteria, it seems unlikely that they have undergone prolonged genome reduction. It has also been demonstrated that methodological biases or data issues can overestimate the rate of either gain or loss (Zhaxybayeva et al. 2007; Hao and Golding 2008).

Generally, the most frequent donors of genetic material appear to have been lineages within the same phylogroup as the recipient genome. Smaller contributions have also

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution

SMBE

come from the other phylogroups and from other Pseudomonads. This might be expected because the rate of homologous recombination decreases exponentially with sequence divergence (Majewski 2001), reducing the incidence of HGT when genes are carried into a new genome via the homologous recombination of flanking regions (Touchon et al. 2009; Polz et al. 2013). Genes arriving from outwith the *Pseudomonas* genus have been acquired from a diverse range of donor species, many of which are known to be soil inhabiting, plant associated, or present in water systems. GO analysis shows that gained genes are drawn from a diverse range of functional categories and are therefore likely to contribute a correspondingly diverse range of novel phenotypes to recipient lineages. *Pseudomonas syringae* is well known for its ability to exist in a multitude of both plant and environmental niches, and the results presented here suggest that such ecological versatility is facilitated by high rates of gene gain into the flexible genome.

These results suggest an underlying structure to gene gain in *P. syringae* that is based on both phylogeny and ecology. An attractive extrapolation of such structuring of HGT is the hypothesis of "niche-specific gene pools" that contribute to adaptation within an ecologically determined framework, via the transfer of genes that are selectively advantageous within a certain niche (Popa et al. 2011; Smillie et al. 2011; Polz et al. 2013). In the case of *P. syringae*, this framework may be determined by the ecology and population structure of the numerous plant host species on which these pathovars thrive. In this way, the expansion of *P. syringae* lineages onto a new host species may occur when newly acquired genes from other species that already inhabit that environment prove to be selectively advantageous to the recipient lineage, thus facilitating a host jump.

Divergence via Differential Adaptation in *P. syringae*

As the extent and prevalence of HGT has become increasingly evident, much discussion has focused on the implications of HGT for the phylogenetic reconstruction of prokaryotic lineages and the validity of the species definition. More recently, however, models have emerged in which HGT can be an active facilitator, rather than a restrictor, of the evolution of population structure, ecological differentiation, and speciation (Ochman et al. 2005; Shapiro et al. 2012; Polz et al. 2013). These ideas rely on the understanding that although rates of HGT and gene loss can be very high, the underlying dynamics of this fluctuation are highly structured in such a way that they actively promote the differentiation of subpopulations of bacteria into distinct genotypic clusters.

HGT and gene loss can directly facilitate the accumulation of divergence between bacterial populations via the differential adaptation of lineages into distinct ecological niches (Ochman et al. 2005; Polz et al. 2013). This mechanism may be operating in the evolution of *P. syringae* and may explain the broad host range of the *P. syringae* species complex while maintaining the host specificity of individual pathovars. *Pseudomonas syringae* genomes show a turbulent history of genome fluctuation, with evidence of high levels of gene gain and loss in all lineages across the species complex. Gained genes are functionally diverse and appear to be gained from sources that are both phylogenetically and ecologically structured. The degree of host specificity shown by many individual pathovars suggests that their gene repertoires are specifically adapted to a given niche environment (Morris et al. 2000; Hunter and Taylor 2006; Sarkar et al. 2006). It is likely that this genomic "tailoring" is achieved through the continuous uptake of a large number of genes via HGT, but the retention of only those few niche-specific genes that confer an advantage in a given host environment.

In *P. syringae*, much attention has been focused on the contribution of effector genes of the type III secretion system (T3SS) toward adaptation and pathogenicity. Although the distribution of the approximately 57 currently described effector families across the *P. syringae* species complex has been well characterized, the relationship between the type 3 secretion effector (T3SE) repertoire and adaptation is complex and remains to be fully elucidated (Sawada et al. 1999; Ma et al. 2006; Lindeberg et al. 2009, 2012; Baltrus et al. 2011; O'Brien et al. 2012). For example, strains that are pathogenic on the same host species can have divergent effector complements (Almeida et al. 2009; Baltrus et al. 2011; O'Brien et al. 2011), suggesting that *P. syringae* lineages are able to adapt to a given environment using multiple variations of effector repertoires (Lindeberg et al. 2012). The acquisition of novel T3SEs may be indicative of adaptive transitions into new niche environments (Sarkar et al. 2006). If T3SEs were solely responsible for the adaptive diversification of *P. syringae* pathovars, it might be expected to find a unique repertoire of T3SEs for each transition onto a different host species and that recent host jumps correspond to recent gains of novel T3SE genes. However, we find that not all pathovars have recently gained T3SE genes and that of those strains that show evidence of recent T3SE gain, only one was also specific to that lineage. Taken together, these observations suggest that host adaptation and the evolution of host specificity may be facilitated by the acquisition of genes that are not directly involved in the T3SS or by molecular adaptations in virulence genes that are already present within the genome.

When acquired genes confer the ability of that lineage to occupy a new ecological niche, such as a new host species, a novel disease can be the result. In the case of *P. syringae*, a number of newly emergent diseases have been reported in the last decade, including bleeding canker of horse chestnut, caused by *P. syringae* pv. *aesculi* (Webber et al. 2008), and bacterial canker of kiwi caused by pv. *actinidiae* (Balestra et al. 2010). Both of these pathovars have recently acquired genes involved in the catabolism of plant-derived aromatic compounds such as derivatives of lignin that are likely to be central

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

in their ability to inhabit (and cause disease in) the woody parts of their respective host species (Green et al. 2010; Steele et al. 2010; Marcelletti et al. 2011). *Pseudomonas syringae* pv. *aesculi* has recently become epidemic across much of north-west Europe, and the success of this pathogen appears to be due, at least in part, to the ability of this species to acquire genes from disparate sources that facilitate adaptation into distinct ecological niches.

Supplementary Material

Supplementary figures S1–S3 and table S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a CASE studentship from the Biotechnology and Biological Sciences Research Council in partnership with Forest Research to R.W.N.

Literature Cited

Almeida NF, et al. 2009. A draft genome sequence of *Pseudomonas syringae* pv. *tomato* T1 reveals a type III effector repertoire significantly divergent from that of *Pseudomonas syringae* pv. *tomato* DC3000. Mol Plant Microbe Interact. 22:52–62.

Balestra GM, Renzi M, Mazzaglia A. 2010. First report of bacterial canker of *Actinidia deliciosa* caused by *Pseudomonas syringae* pv. *actinidiae* in Portugal. New Dis Rep. 22:10.

Baltrus DA, et al. 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. PLoS Pathog. 7:e1002132.

Bardaji L, et al. 2011. Miniature transposable sequences are frequently mobilized in the bacterial plant pathogen *Pseudomonas syringae* pv. *phaseolicola*. PLoS One 6:e25773.

Barraclough TG, Balbi KJ, Ellis RJ. 2012. Evolving concepts of bacterial species. Evol Biol. 39:148–157.

Boucher Y, et al. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet. 37:283–328.

Bull CT, et al. 2010. Comprehensive list of names of plant pathogenic bacteria, 1980–2007. J Plant Pathol. 92:551–592.

Caro-Quintero A, Konstantinidis KT. 2011. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 14:347–355.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Roman HE, Bastolla U, Porto M, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag. p. 207–232.

Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine. Bioinformatics 26:2914–2915.

Cohen O, Pupko T. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol Biol Evol. 27:703–713.

Cohen O, Pupko T. 2011. Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony—a simulation study. Genome Biol Evol. 3:1265–1275.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. Science 301:829–832.

Doolittle WF, Papke RT. 2006. Genomics and the bacterial species problem. Genome Biol. 7:116.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. Genome Res. 19:744–756.

Feil EJ, et al. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A. 98:182–187.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

Gardan L, Bollet C, Ghorrah MA, Grimont F, Grimont P. 1992. DNA relatedness among the pathovar strains of *Pseudomonas syringae* subsp. *savastanoi* Janse (1982) and proposal of *Pseudomonas savastanoi* sp. nov. Int J Syst Bacteriol. 42:606–612.

Georgiades K, Merhej V, El Karkouri K, Raoult D, Pontarotti P. 2011. Gene gain and loss events in *Rickettsia* and *Orientia* species. Biol Direct. 6:6.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19:2226–2238.

Green S, et al. 2010. Comparative genome analysis provides insights into the evolution and adaptation of *Pseudomonas syringae* pv. *aesculi* on *Aesculus hippocastanum*. PLoS One 5:e10224.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hanage WP. 2013. Fuzzy species revisited. BMC Biol. 11:41.

Hanage WP, Fraser C, Spratt BG. 2006. Sequences, sequence clusters and bacterial species. Philos Trans R Soc B Biol Sci. 361:1917–1927.

Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16:636–643.

Hao W, Golding GB. 2008. High rates of lateral gene transfer are not due to false diagnosis of gene absence. Gene 421:27–31.

Hirano SS, Upper CD. 1990. Population biology and epidemiology of *Pseudomonas syringae*. Annu Rev Phytopathol. 28:155–177.

Hunter PJ, Taylor JD. 2006. Patterns of interaction between isolates of three pathovars of *Pseudomonas syringae* and accessions of a range of host and nonhost legume species. Plant Pathol. 55:46–53.

Hwang MSH, Morgan RL, Sarkar SF, Wang PW, Guttman DS. 2005. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. Appl Environ Microbiol. 71:5182–5191.

Jeon CO, Park W, Ghiorse WC, Madsen EL. 2004. *Polaromonas naphthalenivorans* sp. nov., a naphthalene-degrading bacterium from naphthalene-contaminated sediment. Int J Syst Evol Microbiol. 54:93–97.

Kettler GC, et al. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet. 3:e231.

Kuo C-H, Ochman H. 2009. The fate of new bacterial genes. FEMS Microbiol Rev. 33:38–43.

Larsson J, Nylander JA, Bergman B. 2011. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol. 11:187.

Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol. 2:519–523.

Lawrence JG. 2002. Gene transfer in bacteria: speciation without species? Theor Popul Biol. 61:449–460.

Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A. 95:9413–9417.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.

Lefébure T, Bitar PD, Suzuki H, Stanhope MJ. 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol Evol. 2:646–655.

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Lefébure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8:R71.
Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Li W-H. 1993. Unbiased estimation of the rates of synonymous and non-synonymous substitution. J Mol Evol. 36:96–99.
Lin N, Martin GB. 2005. An *avrPto*/avrPtoB mutant of *Pseudomonas syringae* pv. *tomato* DC3000 does not elicit Pto-mediated resistance and is less virulent on tomato. Mol Plant Microbe Interact. 18: 43–51.
Lindeberg M, Cunnac S, Collmer A. 2009. The evolution of *Pseudomonas syringae* host specificity and type III effector repertoires. Mol Plant Pathol. 10:767–775.
Lindeberg M, Cunnac S, Collmer A. 2012. *Pseudomonas syringae* type III effector repertoires: last words in endless arguments. Trends Microbiol. 20: 199–208.
Ma W, Dong FF, Stavrinides J, Guttman DS. 2006. Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. PLoS Genet. 2:e209.
Majewski J. 2001. Sexual isolation in bacteria. FEMS Microbiol Lett. 199: 161–169.
Mansfield J, et al. 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. Mol Plant Pathol. 13:614–629.
Marcelletti S, Ferrante P, Petriccione M, Firrao G, Scortichini M. 2011. *Pseudomonas syringae* pv. *actinidiae* draft genomes comparison reveal strain-specific features involved in adaptation and virulence to *Actinidia* species. PLoS ONE 6:e27297.
Margesin R, Spröer C, Zhang D-C, Busse H-J. 2012. *Polaromonas glacialis* sp. nov. and *Polaromonas cryoconiti* sp. nov., isolated from alpine glacier cryoconite. Int J Syst Evol Microbiol. 62:2662–2668.
Morris CE, et al. 2000. The relationship of host range, physiology, and genotype to virulence on cantaloupe in *Pseudomonas syringae* from cantaloupe blight epidemics in France. Phytopathology 90: 636–646.
Morris CE, et al. 2008. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. ISME J. 2:321–334.
Morris CE, et al. 2010. Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. MBio 1: e00107–10.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302: 205–217.
O'Brien HE, Desveaux D, Guttman DS. 2011. Next-generation genomics of *Pseudomonas syringae*. Curr Opin Microbiol. 14:24–30.
O'Brien HE, et al. 2012. Extensive remodeling of the *Pseudomonas syringae* pv. *avellanae* type III secretome associated with two independent host shifts onto hazelnut. BMC Microbiol. 12:141.
Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proc Natl Acad Sci U S A. 96:12638–12643.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.
Ochman H, Lerat E, Daubin V. 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A. 102(Suppl. 1):6595–6599.
Parkinson N, Bryant R, Bew J, Elphinstone J. 2010. Rapid phylogenetic identification of members of the *Pseudomonas syringae* species complex using the *rpoD* locus. Plant Pathol. 60:338–344.
Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 29:170–175.
Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 21: 599–609.
Qi M, Wang D, Bradley CA, Zhao Y. 2011. Genome sequence analyses of *Pseudomonas savastanoi* pv. *glycinea* and subtractive hybridization-based comparative genomics with nine pseudomonads. PLoS One 6: e16451.
Ramos C, Matas IM, Bardaji L, Aragón IM, Murillo J. 2012. *Pseudomonas savastanoi* pv. *savastanoi*: some like it knot. Mol Plant Pathol. 13: 998–1009.
Reinhardt JA, et al. 2008. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. Genome Res. 19:294–305.
Richards VP, et al. 2011. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. Infect Genet Evol. 11:1263–1275.
Rocha EP. 2008. Evolutionary patterns in prokaryotic genomes. Curr Opin Microbiol. 11:454–460.
Rodríguez-Palenzuela P, et al. 2010. Annotation and overview of the *Pseudomonas savastanoi* pv. *savastanoi* NCPPB 3335 draft genome reveals the virulence gene complement of a tumour-inducing pathogen of woody hosts. Environ Microbiol. 12:1604–1620.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Sarkar SF, Gordon JS, Martin GB, Guttman DS. 2006. Comparative genomics of host-specific virulence in *Pseudomonas syringae*. Genetics 174: 1041–1056.
Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl Environ Microbiol. 70:1999–2012.
Sawada H, Suzuki F, Matsuda I, Saitou N. 1999. Phylogenetic analysis of *Pseudomonas syringae* pathovars suggests the horizontal gene transfer of *argK* and the evolutionary stability of *hrp* gene cluster. J Mol Evol. 49:627–644.
Schaad NW, et al. 2008. Reclassification of subspecies of *Acidovorax avenae* as *A. avenae* (Manns 1905) emend., *A. cattleyae* (Pavarino, 1911) comb. nov., *A. citrulli* (Schaad et al., 1978) comb. nov., and proposal of *A. oryzae* sp. nov. Syst Appl Microbiol. 31: 434–446.
Shapiro BJ, et al. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51.
Smillie CS, et al. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244.
Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nat Genet. 21:108–110.
Steele H, Laue BE, MacAskill GA, Hendry SJ, Green S. 2010. Analysis of the natural infection of European horse chestnut (*Aesculus hippocastanum*) by *Pseudomonas syringae* pv. *aesculi*. Plant Pathol. 59: 1005–1013.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102:13950–13955.
Toth IK, Bell KS, Holeva MC, Birch PRJ. 2003. Soft rot erwiniae: from genes to genomes. Mol Plant Pathol. 4:17–30.
Touchon M, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5: e1000344.
Treangen TJ, Rocha EP. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7: e1001284.
Vernikos GS, Thomson NR, Parkhill J. 2007. Genetic flux over time in the *Salmonella* lineage. Genome Biol. 8:R100.

Genome Biology and Evolution SMBE

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Wang Y, Huang H, Sun M, Zhang Q, Guo D. 2012. T3DB: an integrated database for bacterial type III secretion system. BMC Bioinformatics 13:66.
Webber JF, et al. 2008. Isolation and identification of *Pseudomonas syringae* pv. *aesculi* causing bleeding canker of horse chestnut in the UK. Plant Pathol. 57:368–368.
Zhaxybayeva O, Nesbø CL, Doolittle WF. 2007. Systematic overestimation of gene gain through false diagnosis of gene absence. Genome Biol. 8: 402–407.

Associate editor: Tal Dagan

Downloaded from http://gbe.oxfordjournals.org/ at Edinburgh University on July 1, 2014

Genome Biology and Evolution SMBE

Table S1.—Complete information for 27 *P. syringae* pathovars used in this study.

Pathovar name	Strain	Tag*	Host species	Genome size (Mb)	*n* proteins†	Genome status (*n* contigs)	Reference
aceris	M302273PT	acer	*Acer* Sp. (maple), USA	6.183	6185	Draft (1179)	Baltrus et al. (2011)
actinidiae	M302091	actn	*Actinidia delicosa* (kiwi fruit), Japan	5.849	5626	Draft (941)	Baltrus et al. (2011)
aesculi	2250	aes2	*Aesculus hippocastanum* (European horse chestnut), UK	5.926	5619	Draft (776)	Green et al. (2010)
aesculi	NCPPB 3681	aesN	*Aesculus indica* (Indian horse chestnut), India	5.836	5649	Draft (841)	Green et al. (2010)
aptata	DSM 50252	apta	*Beta vulgaris* (sugar beet), USA	6.243	6368	Draft (3776)	Baltrus et al. (2011)
'*cit7*'	Cit7	cit7	*Citrus sinensis* (navel orange), unknown	6.075	7145	Draft (2655)	Baltrus et al. (2011)
glycinea	B076	glyB	*Glycine max* (soybean), USA	6.236	5580	Draft (104)	Qi et al. (2011)
glycinea	race 4	gly4	Not explicitly stated	5.905	5207	Draft (108)	Qi et al. (2011)
japonica	M301072PT	japa	*Hordeum vulgare* (barley), Japan	6.381	8796	Draft (4661)	Baltrus et al. (2011)
lachrymans	M301315	lacM	*Cucumis sativus* (cucumber), Japan	7.727	6691	Draft (791)	Baltrus et al. (2011)
lachrymans	M302287PT	lacP	*Cucumis sativus* (cucumber), USA	5.895	5636	Draft (798)	Baltrus et al. (2011)
maculicola	ES4326	Pcan‡	*Raphanus sativus* (radish), unknown	6.221	6101	Draft (878)	Baltrus et al. (2011)
mori	M301020	mori	*Morus alba* (white mulberry), Japan	6.393	7544	Draft (3414)	Baltrus et al. (2011)
morsprunorum	M302280PT	mrsp	*Prunus domesticus* (European plum), unknown	6.039	5837	Draft (969)	Baltrus et al. (2011)
oryzae	1_6	oryz	*Oryza sativa* (rice plant), unknown	6.704	7837	Draft (2855)	Reinhardt et al. (2008)
phaseolicola	1448A	phas	*Phaseolus vulgaris* (common bean), Ethiopia	6.112	5172	Complete (1 chr., 2 plasmids)	Joardar et al. (2005)
pisi	1704B	pisi	*Pisum sativum* (pea), France	6.521	9160	Draft (5099)	Baltrus et al. (2011)
savastanoi	NCPPB 3335	sava	*Olea europaea* (olive tree), France	5.759	4898	Draft (403)	Rodriguez-Palenzuela et al. (2010)
syringae	B728a	syrB	Snap bean leaflet, unknown	6.093	5089	Complete (1 chr.)	Feil et al. (2005)
syringae	FF5	syrF	*Pyrus calleryana* (ornamental pear), USA	5.659	8445	Draft (4578)	Sohn et al. (2012)
syringae	642	syr6	Isolated from a "small newly emerging weedy plant" (Clarke *et al.*, 2010), USA	5.809	5274	Draft (296)	Clarke et al. (2010)
tabaci	ATCC 11528	taba	*Nicotiana tabacum* (tobacco plant), unknown	6.068	6467	Draft (1405)	Studholme et al. (2009)
tomato	DC3000	tomD	*Solanum lycopersicum* (tomato), Guernsey (Channel Is., UK)	6.538	5619	Complete (1 chr., 2 plasmids)	Buell et al. (2003)
tomato	K40	tomK	*Solanum lycopersicum* (tomato), unknown	6.154	5853	Draft (582)	Cai et al. (2011)
tomato	Max13	tomM	*Solanum lycopersicum* (tomato), France	6.105	5749	Draft (349)	Vinatzer et al. (unpublished)
tomato	NCPPB 1108	tomN	*Solanum lycopersicum* (tomato), Jersey (Channel Is., UK)	6.082	5619	Draft (304)	Cai et al. (2011)
tomato	T1	tomT	*Solanum lycopersicum* (tomato), Canada	6.146	5702	Draft (122)	Almeida et al. (2009)

*4-letter identifier used in this study.

†Based on the number of published protein sequences available at NCBI.

‡The pathovar *maculicola* has been recently found to have been originally misidentified, and should in fact be renamed *Pseudomonas cannabina* pv. *alisalensis* (reported in Baltrus et al. 2011); thus the tag '*Pcan*' has been used in this study.

Literature Cited

Almeida NF, et al. 2009. A draft genome sequence of *Pseudomonas syringae* pv. *tomato* T1 reveals a type III effector repertoire significantly divergent from that of *Pseudomonas syringae* pv. *tomato* DC3000. Mol Plant Microbe Interact. 22:52–62.

Baltrus DA, et al. 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. PLoS Pathog. 7:e1002132.

Buell CR, et al. 2003. The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. Proc Natl Acad Sci USA. 100:10181–10186.

Cai R, et al. 2011. The plant pathogen *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection to evade tomato immunity. PLoS Pathog. 7:e1002130.

Clarke CR, Cai R, Studholme DJ, Guttman DS, Vinatzer BA. 2010. *Pseudomonas syringae* strains naturally lacking the classical *P. syringae* hrp/hrc locus are common leaf colonizers equipped with an atypical type III secretion system. Mol Plant Microbe Interact. 23:198–210.

Feil H, et al. 2005. Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. Proc Natl Acad Sci USA. 102:11064–11069.

Green S, et al. 2010. Comparative genome analysis provides insights into the evolution and adaptation of *Pseudomonas syringae* pv. *aesculi* on *Aesculus hippocastanum*. PLoS ONE. 5:e10224.

Joardar V, et al. 2005. Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol. 187:6488.

Qi M, Wang D, Bradley CA, Zhao Y. 2011. Genome sequence analyses of *Pseudomonas savastanoi* pv. *glycinea* and subtractive hybridization-based comparative genomics with nine Pseudomonads. PLoS ONE. 6:e16451.

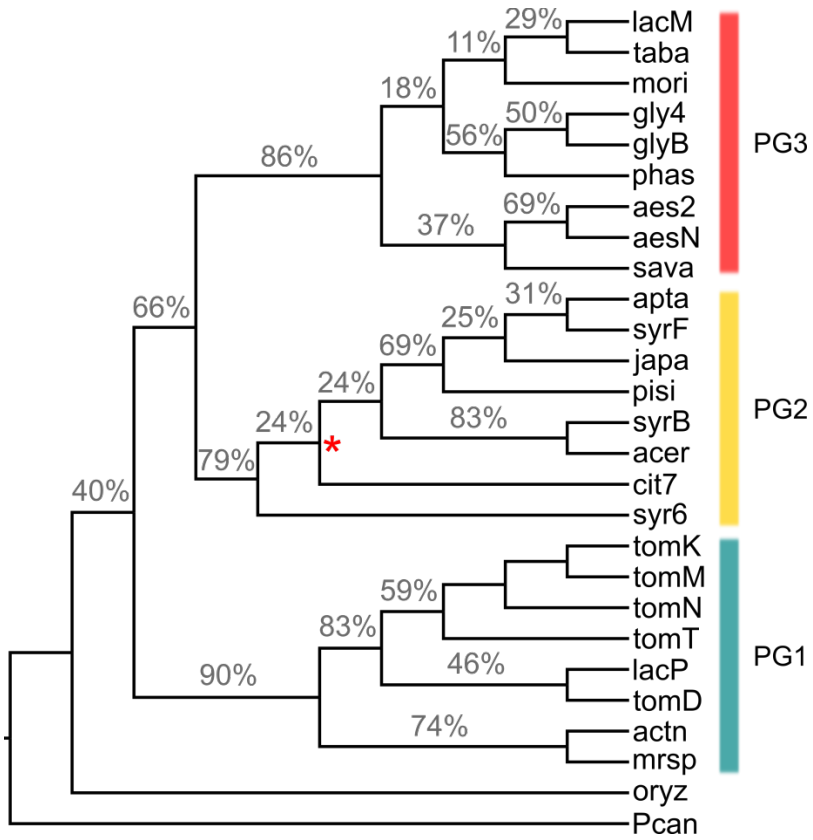
Reinhardt JA, et al. 2008. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. Genome Res. 19:294–305.

Rodríguez-Palenzuela P, et al. 2010. Annotation and overview of the *Pseudomonas savastanoi* pv. *savastanoi* NCPPB 3335 draft genome

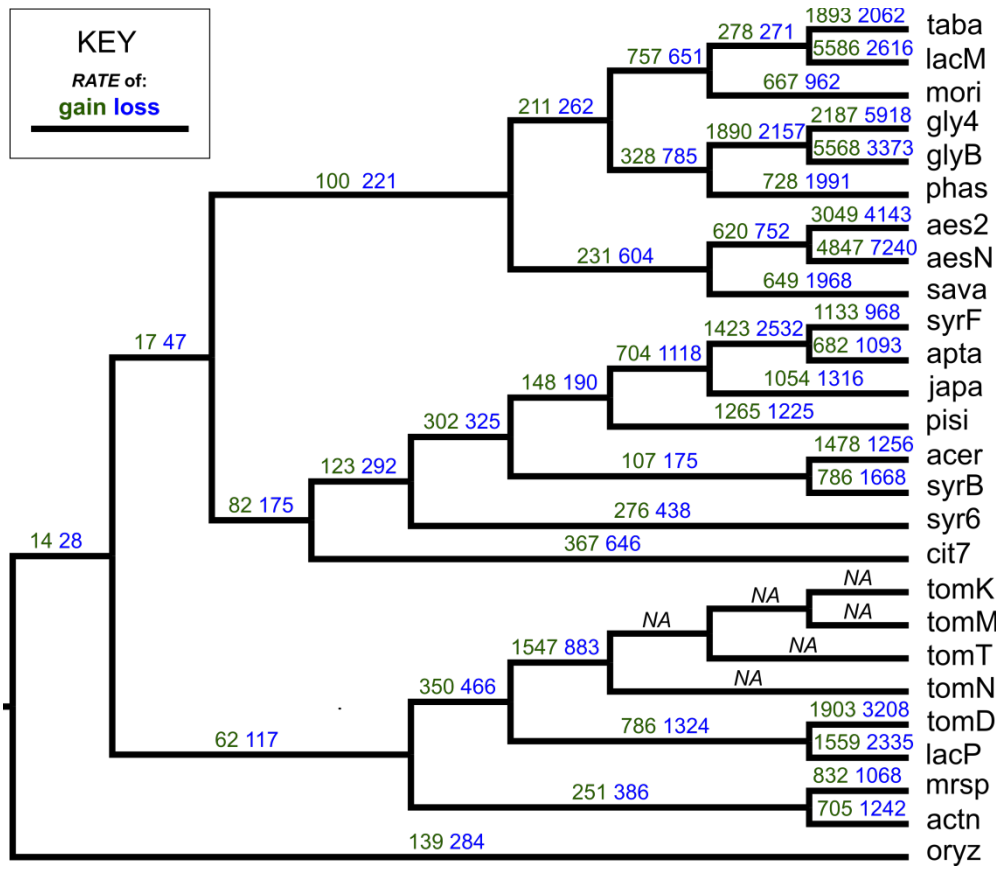
reveals the virulence gene complement of a tumour-inducing pathogen of woody hosts. Environ Microbiol. 12:1604–1620.

Sohn KH, Jones JDG, Studholme DJ. 2012. Draft genome sequence of *Pseudomonas syringae* pathovar *syringae* strain FF5, causal agent of stem tip dieback disease on ornamental pear. J Bacteriol. 194:3733–3734.

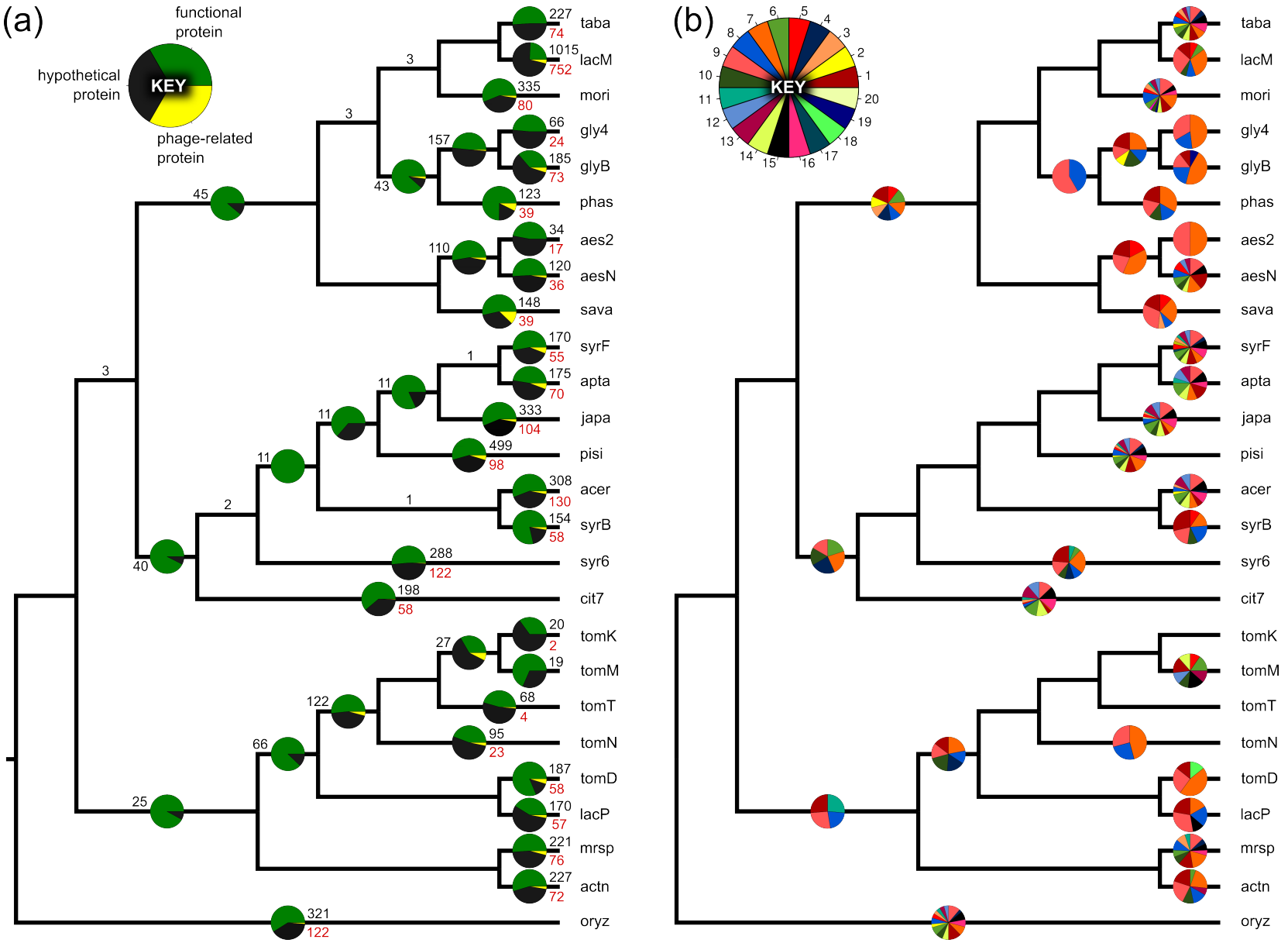
Studholme DJ, et al. 2009. A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of *Pseudomonas syringae* pathovar *tabaci* 11528. BMC Genomics. 10:395. doi: 10.1186/1471-2164-10-395.



SUPPLEMENTARY FIG. S1.— Extended majority consensus tree. Grey numbers indicate the consensus support for a given branch, given as the proportion of times (expressed as a percentage) a particular branch splits the set of taxa into two subsets separated by that branch. The asterisk indicates the only incongruence in branching pattern relative to the core-genome phylogeny.



SUPPLEMENTARY FIG. S2.— Rate of gain and loss. Numbers above each branch indicate the ratio of the number events along a given branch to the length of that branch, i.e. the number of gains (green) or losses (blue) per percentage amino-acid divergence of the core-genome phylogeny. These values are not given for some branches with zero branch length (*NA*). Lineage-specific genes are included in gains along terminal branches. Topology is based on the core-genome phylogeny and branch lengths are not to scale.


(a)
functional protein
hypothetical protein
KEY
phage-related protein
taba
lacM
mori
gly4
glyB
phas
aes2
aesN
sava
syrF
apta
japa
pisi
acer
syrB
syr6
cit7
tomK
tomM
tomT
tomN
tomD
lacP
mrsp
actn
oryz
(b)
KEY

SUPPLEMENTARY FIG. S3.—Functional categorisation of gained genes. (a) Simple categorisation. Numbers above each branch indicate the total number of gains (≥ 0.8) for that branch; red numbers on terminal branches indicate the number of terminal gains that are lineage-specific, *i.e.* unique to the given lineage. Pies along each branch represent the classification of genes into one of three basic categories: 'functional protein' (green), 'hypothetical protein' (black) or 'phage-related protein' (yellow). Only branches with ≥ 10 gains are delimited into pies. The topology of the cladogram is based on the core-genome phylogeny (branch lengths not to scale). (b) Gene ontology categorisation (molecular function, GO level 3). Genes from (a) in the 'functional protein' category are further decomposed into functional groups based on gene ontology (GO) information. Only GO terms with ≥ 3 sequences are shown. Functional descriptions corresponding to the numbers in the key are given in the table below (for further detail regarding GO terms see www.geneontology.org).

Table of GO categories:

#	Description	#	Description
1	Transferase activity	**11**	Transmembrane transporter activity
2	Lysase activity	**12**	Amine binding
3	Signal transducer activity	**13**	Carboxylic acid binding
4	Sequence-specific DNA binding transcription factor activity	**14**	Vitamin binding
5	Ion binding	**15**	Cofactor binding
6	Substrate-specific transporter activity	**16**	Ligase activity
7	Nucleic acid binding	**17**	Isomerase activity
8	Nucleotide binding	**18**	Recombinase activity
9	Hydrolase activity	**19**	Protein binding
10	Oxidoreductase activity	**20**	Metal cluster binding