Swansea University
Prifysgol Abertawe

Facultative bacterial symbionts from European Orius species: Evidence for an ancestral symbiotic association.

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

SWANSEA UNIVERSITY
Medical School

XIAORUI CHEN ([redacted])

Year of Submission: 2022

Supervisors:

Dr. Ricardo Del Sol and Prof. Paul Dyson

ABSTRACT

Pest control in agriculture employs diverse strategies, among which the use of predatory insects has steadily increased. The use of several species within the genus Orius in pest control is widely spread, particularly in Mediterranean Europe. The use of predatory insects in pest control in agriculture has spread worldwide and increased significantly, especially in the use of various *Orius* species. Currently, most studies about *Orius* species have been focused on the diet manipulation or selective breeding methods to reduce the rearing costs and improve the efficiency, only a few studies were associated to their *Wolbachia* symbionts. The characterisation and contribution of microbial symbionts to Orius sp. fitness, behaviour, and potential impact on human health has been neglected. Therefore, there is a lack of knowledge regarding *Orius*' symbionts such as their taxonomic characterisation, the functions of the symbionts and potential influences on human health. This project was focused on the first comparative genomics report of genome sequences level description of the predominant culturable facultative bacterial symbionts associated with the analyses of draft genomes of facultative symbionts using Next Generation Sequencing (NGS) technique related to five *Orius* species (*Orius laevigatus*, *Orius niger*, *Orius pallidicornis*, *Orius majusculus* and *Orius albidipennis*) and collected from various European countries (Greece, Italy, and Spain). Initially, *cox*l (COI) based taxonomic classification of the *Orius* species used was performed, followed by the isolation of culturable bacteria from live insects. The whole genome sequences of the bacterial isolates were generated and assembled into draft genomes using NGS. The isolates of two predominant bacteria belong to *Serratia* and *Leucobacter* genera, the third predominant bacteria are most likely to be a new genus within the Erwiniaceae. *Orius* sp. *Serratia* isolates genomes are more similar to *Serratia* sp. SCBI. Pan-genome analysis of *Serratia* sp. *Orius* isolates evidenced an open pan-genome, and 279 accessory genes were related to the insect symbiosis trait. Additionally, pan-genome analyses of the Serratia sp. isolates offered clues linking Type VI secretion system effector–immunity proteins from the Tai4 sub-family to the symbiotic lifestyle. These symbionts were found to colonise all the insect specimens tested, which evidenced an ancestral symbiotic association between these bacteria and the genus *Orius*. Additionally, plasmid sequence analyses suggest sequence exchanges between *Serratia* sp. *Orius* isolates and pathogenic *Serratia* species, which may have implications for food safety and human health.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .....[redacted]..... (candidate)

Date .....30/03/2022.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... [redacted]..... (candidate)

Date ..... 30/03/2022.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... [redacted]..... (candidate)

Date ..... 30/03/2022.....

THESIS SUMMARY

This summary sheet should be completed after you have read the accompanying notes. The completed sheet should be submitted by you to your Head of College at the time of submission of your work and the supporting documentation.

Candidate's Surname / Family Name: Chen

Candidate's Forenames: Xiaorui

Candidate for the Degree of PhD

Full title of thesis: Facultative bacterial symbionts from European *Orius* species: Evidence for an ancestral symbiotic association

Summary:

The use of predatory insects in pest control in agriculture has spread worldwide and increased significantly, especially in the use of various *Orius* species. Currently, most studies about *Orius* species have been focused on the diet manipulation or selective breeding methods to reduce the rearing costs and improve the efficiency, only a few studies were associated to their Wolbachia symbionts. There is a lack of knowledge regarding *Orius*' symbionts such as their taxonomic characterisation, the functions of the symbionts and potential influences on human health. This project was focused on the analyses of draft genomes of facultative symbionts related to five Orius species (*Orius laevigatus, Orius niger, Orius pallidicornis, Orius majusculus and Orius albidipennis*) and collected from various European countries (Greece, Italy, and Spain). Initially, coxl (COI) based taxonomic classification of the *Orius* species used was performed, followed by the isolation of culturable bacteria from live insects. The genome sequences of the bacterial isolates were generated and assembled into draft genomes. The isolates of two predominant bacteria belong to *Serratia* and *Leucobacter* genera, the third predominant bacteria are most likely to be a new genus within the Erwiniaceae. *Serratia* sp. *Orius* genomes are more similar to *Serratia* sp. SCBI. Pan-genome analysis of *Serratia* sp. *Orius* isolates evidenced an open pan-genome, and 279 accessory genes were related to the insect symbiosis trait. Additionally, plasmid sequence analyses suggest sequence exchanges between *Serratia* sp. *Orius* isolates and pathogenic *Serratia* species, which may have implications for food safety and human health.

ACKNOWLEDGEMENTS

I would like to express special appreciation to my supervisors Dr. Ricardo Del Sol and Prof. Paul Dyson for supervising this project, and for their continuous support, care of my wellbeing and guidance throughout the work. They were always helpful in guiding me on path of becoming an independent researcher. Thanks Dr. Ricardo Del Sol for your encouragement and engaging me in various academic activities. This research would not have been possible without the help and support of many people. Special thanks go to Professor Paul Dyson's group for their valuable comments and support during my PhD. I would like to thank Swansea University and the Medical School for the many ways in which they have helped me and made me feel at home.

Finally, I would like to thank my lovely family. Words cannot express how grateful I am to my parents for all the support, encouragement they provided me with during this research. I would also give special thanks to my friends, especially Rachel Locklin, Helen Han, Xiao Han, and Guowen zhang for their faith and support whilst I completed this thesis.

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Pest control management: general description

Today's agricultural challenges are alarming: the world's population is growing rapidly but the amount of arable land per capita is declining, urban development is depleting available arable land at an unprecedented rate, the climate is changing, and crop yields are stagnating.

The traditional agrochemicals that once helped increase agricultural productivity are no longer effective, but the overuse of agrochemicals is also causing irreversible damage to the environment. The use of agrochemicals such as pesticides, fertilizers and plant growth promoters has been critical to humanity over the last century. They have allowed agricultural productivity to keep pace with the most dramatic population growth in our history and have saved billions of people from starvation. However, their environmental impact has become so profound that it cannot be ignored, and they are increasingly seen as 20th century tools that cannot meet the challenges of the 21st century. Many countries are introducing policies to limit the use of agrochemicals, and there is an urgent need to find new solutions to increase crop yields without damaging the environment. Consequently, the search is on for alternative approaches to devise new strategies of pest control. As a result, Integrated Pest Management (IPM) has been created to suppress pest populations below a certain level of economic damage to crops and labour costs, without any environmental hazards (van Lenteren, 2012). Initially, IPM was used in pollination of greenhouse crops by honey and bumble bees. Since this method of pollination was successful in reducing labour costs and increasing crop production, more growers were encouraged to use biological control not only for reduction of pest populations, but also for prevention of diseases (Albajes et al., 1999).

IPM approaches have been developed as a systematic method that utilizes the natural enemies of insect pests to monitor the numbers of insect pests and to protect crops without causing environmental problems (Abrol and Shankar, 2016). The application of insect predators as biological control agents has a long history in the field of agriculture and is popular because it is more environmentally safe and economically viable than other pest control approaches in the context of IPM programmes. Types of biological control “tactics” can be divided into natural, conservation, inoculative (=classical) and augmentative

strategies. Natural biological control is the natural reduction of pest populations by their natural enemies without human intervention and originally developed during the evolution of the first terrestrial ecosystems about 500 million years ago (van Lenteren, 2012).

In contrast, conservation biological controls (CBC) are mainly managing the agroecosystems from human actions to allow the protection and stimulation of naturally residential enemies and the functions they offer (Romeis et al., 2019). There are two common approaches used in CBC. One is increasing the number and activity of natural enemies by manipulating the habitat, because the increased complexity of landscape composition can increase the abundance of natural enemies (Veres et al, 2013). According to their study, the landscape complexity includes the level of biodiversity, the intensity of agriculture in the landscape, and the abundance of predation or parasitism in the landscape. The semi-natural habitats, such as woodland and grassland, have more complexity of landscape composition than fully cultivated area. In the result, these semi-natural habitats or lower intensity of agriculture in the farm have higher proportions of CBCs effectiveness than fully cultivated area (Veres et al, 2013). Another approach is focus on reducing the use of control strategies which could be harmful to the natural enemies, such as the selective use of chemical pesticide (Romeis et al., 2019). For example, the soybean IPM in Brazil, the utilization of spiders (*Geocoris spp*. and *Nabis capsiformis*) in soybean fields can consume the higher proportion of the eggs and larval level of velvetbean caterpillar (*Anticarsia gemmatalis*). Since soyabean IPM adopted with the use of selective pesticides which are harmless to the natural enemies in the Brazilian soybean agroecosystem, the profit of soybean dramatically increased than before and balanced agricultural system in the Brazil (Torres and Bueno, 2018).

CBC is a sustainable method in part of IPM which aim to supress pest growth by conservation of natural enemies. It includes reduction of insecticide use and prevention of the natural habitat loss by decreasing the environmental disturbance related to intensive crop production. In the principle, the increase of natural enemies' population should correspond to a variety of conservation strategies which including the increased landscape complexity, reduction of cropping intensity and increasing plant diversity. However, the response of natural enemy population usually is not consistent to the conservation strategies. Furthermore, it could be resulted in failure of pest control or improved crop yield and reduction of utilisation in commercial crop production environments (Begg et

al., 2017). Therefore, the future work of CBC could focus on the accuracy of effective CBC in suppression of the target pests (Begg et al., 2017).

Inoculative biological control (IBC, also called classical biological control) is long-term manage pest populations by introduction of new natural enemies to the targeting pest areas and it commonly used in invasive species such as management of *Sirex noctilio* using the parasitoid *Ibalia leucospoides* (Hymenoptera: Ibaliidae)(van Lenteren, 2012; Fischbein and Corley, 2015). The main approach of IBC includes collection of the exotic natural predators of invasive pest from original area of its prey and construction of a new population of its natural enemies in new areas where the pests have been accidentally introduced and causes heavy damage. (Fischbein and Corley, 2015). For instance, *Lantana camara*, (Verbenaceae) an important invasive species in the Palaeotropics over a century have been invaded a majority of global agricultural areas such as Africa, Asia, and Australasia due to the global warming (Thomas et al., 2021). Due to the high genetic diversity of this species, it is difficult to suppress. Recently, a pathotype of the microcyclic rust *Puccinia lantanae* fungus collected from the Amazonian rainforest, can attack most biotypes of the *L. camara* species (Thomas et al., 2021). According to the results from greenhouse screening, this pathotype is highly specific to *L. camara*, it causes seedling death of *L. camara* and wide range of disease symptoms in *L. camara*, especially in forest. Therefore, it has been considered as a potential IBC to against invasions by *L. camara* in forest ecosystem (Thomas et al., 2021). Although IBC has been become the most cost-effective and environmentally safe management tool for invasive species internationally, it still has restricted by political, regulatory, and institutional issues and these issues related to new ICB discovery, pre-releasing, and post-releasing regulatory stages, such as shipping issues and political instability of source countries (DiTomaso et al., 2017). In the future, if the IBC project program could be transparent criteria and simplified shipment process of the IBC, IBC could be improved more effectively to suppress the invade species (DiTomaso et al., 2017).

Augmentative biological controls (ABC) are commercial biological control which are followed by massive production of indigenous or exotic predators on a semi-industrial scale (van Lenteren, 2012). Augmentative biological controls have proven very successful as global commercial activities, especially in Europe. Worldwide, this market generates more than 200 million euros in total at end-user level with more than 230 species of natural

predators (mostly Arthropoda, 219 out of 230 species = 95.2%) which are used in pest management worldwide (van Lenteren, 2012). However, this kind of biological controls has similar limitations of CBC and IBC. Therefore, the selections of ABC are essential factor to the farmers and also improving the effectiveness of ABC could apply similar ways of IBC and CBC.

Microbial control agents (MCA) are the most common approach for applications of entomopathogens to manage the population of pest arthropods and MCA are also a subset of ABC (Lacey et al., 2015). Entomopathogens are insect-specific viruses, bacteria, fungi, and nematodes to control the pest growths in the agricultural fields and over half percentage of them have been commercialized (Lacey et al., 2015). Among entomopathogenic bacteria, the best known is *Bacillus thuringiensis*. It has been known since 1901 and is used to manage several major insect pests in agriculture, forestry, and medicine (Sharma et al., 2019). It often used to kill the pest families and species of *Lepidoptera* (Lacey et al., 2015). Additionally, Serratia entomophila is a typical *Serratia* species which is widely used in control of numerous insect pests in pasture, and it can control the pest from the family of Scarabaeidae such as *Costelytra zealandica* (Lacey et al., 2015).

1.2 *Orius* as a natural biological control agent

Among the natural biological agents used, the order Hymenoptera is well represented by several species from the genus *Orius* (minute pirate bug) as augmentative biological controls in IPM, the majority of which constitute a substantial share of the market in the late 20th century. All the *Orius* species are Anthocorid bugs, members of the order Hemiptera, family Anthocoridae. They are also referred to as flower bugs or minute pirate bugs and comprise around 400 to 600 species distributed worldwide. More than 70 species of the genus *Orius* are commonly distributed in the Palearctic, with others located in the Nearctic and Neotropic realms (Horton et al., 2016). *O. sauteri* (Poppius) and *O. strigicollis* (Poppius), among others, have been used in Japan as biological control agents (Yano, 2004). Most *Orius* species in Europe are used in pest control and these species include *Orius albidipennis* (Hemiptera: Anthocoridae), *O. laevigatus, O. strigicollis, O. niger*, and *O. insidiosus* (Kim et al. 2008); Table 1-1 presents information on all currently

known *Orius* genus biological control agents. *Orius laevigatus* is one of the augmentative bio-control agents used against *Frankliniella occidentalis,* the essential pest of sweet pepper (Bouagga et al., 2017).

Table 1-1: Commercially used *Orius* genus biological control agents with region of use and references

SPECIES NAME	LOCATIONS	TARGET	REFERENCE
Orius albidipennis	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012
Orius armatus	Australia	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012
Orius horvathi	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	Gomez-Polo et al., 2013
Orius insidiosus	Europe, North and Latin America	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012
Orius laevigatus	Europe, Africa North, Asia	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012
Orius laticollis	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	Gomez-Polo et al., 2013
Orius majusculus	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012
Orius minutus	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012
Orius niger	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	Gomez-Polo et al., 2013
Orius strigicollis	Asia, Japan	*Frankliniella occidentalis* (Thrips)	Yano, 2004
Orius sauteri	Asia, Japan	*Frankliniella occidentalis* (Thrips)	Yano, 2004
Orius tristicolor	Europe, Mediterranean	*Frankliniella occidentalis* (Thrips)	van Lenteren, 2012

Several *Orius* species have been successfully used for pest control in many crops, including sweet pepper, cucumber, and some ornamentals, in greenhouses and open fields (Lorenzana et al., 2010). Especially, these are widely used to control *Frankliniella occidentalis*, *Thrips palmi* and *Thrips tabaci* (Yano, 2004). However, some *Orius* species also display various levels of facultative phytophagous habits, with *O. pallidicornis*, the only currently known exception, being reared primarily on pollen of *Ecballium elaterium* (Lattin, 1999; Pericart, 1972).

Orius insects are mainly facultatively phytophagous, *O. insidious* feeds on xylem and mesophyll contents, foods which provide different essential amino acids and sugars, during the survival time with lack of prey. *Orius laevigatus* also feeds on certain plant such as fresh pepper pollen, but nowadays *Orius laevigatus* is more frequently fed with *Ephestia kuehniella* Zeller eggs and *Artemia franciscana Kellogg* cysts as artificial diets to increase their production (De Clercq and Bonte, 2008).

Currently, the majority of published works concerning *Orius* genus insects are concentrated on the molecular classifications of *Orius* species for controlling field

sustainability and effectiveness, while some are focused on artificial diet amelioration. In fact, animal prey makes up an important dietary component for ideal development of *O. albidipennis* and *O. laevigatus* (Vacante et al., 1997). Paticularly, this relates to the high cost of rearing large numbers of *O. laevigatus* species due to the need for *Ephestia kuehniella* as the primary food resource (Bonte and DeClercq, 2008). An obvious gap in the knowledge of *Orius* biology relates to the symbiotic associations present within the genus, in particular the role played by bacterial symbionts hosted by *Orius* species.

Traditionally, the identification and classification of *Orius* by morphological characteristics is, however, time consuming, and in some groups also very difficult such as *Orius niger*. The identification of nymphal stages or eggs is even more critical or even impossible, due to high levels of morphological variation in diagnostic characteristics as result of putative hybridizations it is not surprising that in some species the taxonomic status is subject of discussion. Therefore, molecular methods are seen as promising complementary tool to morphological based methods (Raupach, et al., 2014).

Furthermore, the molecular identification of *Orius* species is important to understand their behaviour biology. DNA barcoding has become an effective molecular method for species identification regardless of the development stage of the analysed specimen representing an efficient approach for valid species identification for large-scale biodiversity studies. the classical barcode fragment consists of the fragment of the mitochondrial cytochrome c oxidase subunit 1 (CO1) gene. The idea of DNA barcoding relies on the concept that each species will most likely have unique DNA barcodes and that intraspecific CO1 variation is typically lower than the interspecific variability. As consequence, a so-called barcoding gap is given which allows an undoubted molecular species identification. Currently, DNA barcodes have become an important and increasingly used tool as part of an integrative taxonomy in modern species descriptions as well as various other biological disciplines, such as forensics, pest biology, and conservation biology (Raupach et al., 2014). However, only a few studies of *Orius* species were related to the identification and classification of *Orius*. Therefore, Chapter 3 will illustrate the classification of all available *coxl* sequences of *Orius* species in NCBI (National Centre for Biotechnology) database accessed on Jan 2017 and also the *Orius* specimens identification and classification by *coxl* sanger sequencing and its phylogenic study. Despite their extensive

use in IPM, there is a conspicuous absence of studies exploring the role of microbial symbionts on development, speciation, fitness, and behaviour of *Orius* species.

1.3 General symbiosis descriptions

The term symbiosis (from Greek) was first coined by Heinrich Anton de Bary in 1879 to represent "The living together of unlike organisms". Currently, symbiosis is defined as two or more organisms maintaining long-term interactions, especially in close physical contact. The catalogues of these interactions range from outright parasitism, which is detrimental to the host, or neutral as in commensalism (host neither suffers from symbiotic fitness nor gains benefits from its symbiont) through ecologically contingent mutualism (both host and symbiont gain advantages from this relationship) to obligate co-dependence (Figure 1-1) (Ewald, 1987; Brownlie and Johnson, 2009). Furthermore, most organisms on Earth enter some forms of 'symbiosis', as either host or symbiont, in order to adapt to different living environments and this is typically to the advantage of both host and symbionts. Mutualism is reciprocal exploitation that is also beneficial to each partner in this association (Herre et al., 1999). However, If an insect host loses its mycetomic symbionts which can provide nutrients necessary for host development, the host will develop more slowly than usual and will be unable to produce offspring (Sasaki et al., 1991; Douglas, 1992; Baumann et al., 1995). For example, slow evolutionary development of cockroaches is caused by disruption of nitrogen recycling when they lose their hindgut symbionts (Cochran, 1985; Cruden & Markoverz, 1987). Tsetse flies (*Glossina morsitans*) will reduce their oviposition rate because of elimination of a gut symbiont (Nogge & Gerresheim, 1982). Therefore, the elimination of a symbiont may potentially to be an essential strategy of pest controls. Obligate mutualists that are in highly interdependent host-symbiont relationships, such as the aphid–*Buchnera* endosymbiosis, have also been detected in other insects, such as *Wigglesworthia* in tsetse flies, *Baumannia* in sharpshooters, *Carsonella* in psyllids, *Tremblaya* in mealybugs, *Blochmannia* in carpenter ants, and *Nardonella* in weevils (Kikuchi, 2009).

Figure 1-1: The scale of all symbiosis relationship in different level. The lighter colour of blue, the less severe the severity the host will receive from its symbionts.

One typical characteristic of symbiotic interactions is that they extend to include very distinct parties, often with the members coming from different biological kingdoms. These associations can also cover all levels of biological complexity. For example, ancient symbiotic relationships have resulted in the evolution of eukaryotic cell structure (e.g., symbiotically derived organelles such as plastids, chloroplasts, and mitochondria). Some symbiotic relationships also form between different eukaryotes, or between a eukaryotic host and prokaryotic symbionts (Brownlie and Johnson, 2009). Symbiotic interactions also can range from transient to residential symbioses (Rosenberg and Zilber-Rosenberg, 2011). Transient interactions usually occur when there is a high frequency of different host contacts and is only present in part of one host generation. Residential or permanent symbioses are similar to chronic 'infections' in that symbionts are carried in multiple generations of their host, effectively becoming a heritable component between a host and its symbionts (Rosenberg and Zilber-Rosenberg, 2011).

The ubiquity of symbiosis in nature could be reasonably considered to be one of the major driving forces of evolutionary innovation on Earth, since a symbiont could confer either single or multiple phenotypic traits on its host (Sudakaran et al., 2017). Such traits may allow the expansion of the host into previously inaccessible ecological niches and subsequent lineage diversification, resulting in increased diversity of organisms on Earth (Sudakaran et al., 2017).

1.4 Microbial symbiosis in insects

Microbial symbioses between hosts and microorganisms occur when a microorganism lives as a permanent and non-invasive partner within the host. Most insects are always involved in symbiotic relationships, especially between herbivorous insect hosts and microbial symbionts such as parasitic or mutualistic microbial bacteria, mycorrhizal root fungi or social parasites (Jurkevitch, 2011). The importance of these microbial symbionts is that they can increase the ability of an insect to adapt to its environment.

As an example, an insect may symbiotically interact with a plant. Plants and insects have a vast variety of ways to interact with each other, although they belong to two of the most different multicellular kingdoms in the terrestrial environment. During the Cretaceous period especially, insects lived mainly on angiosperms, and this was one of the milestones that strongly supported the development of the rich diversity of insect species now extant. However, plant tissues consist of multiple plant polymers such as lignin, pectin, cellulose, and hemicellulose, which are difficult for herbivorous insects to digest. In addition, plants can produce plenty of toxic chemicals to defend themselves against attack by their natural enemies such as herbivores. In order to overcome these issues with plant defences and dietary challenges, insects interacted with microbial symbionts to improve their own physiological and behavioural adaptations. These symbionts can supply essential nutrients to the hosts and also detoxify toxic compounds produced by plants. Over 2 billion years of eukaryotic evolution, microbial symbiosis has played an essential role in helping eukaryote biological development such as their reproduction, immunity, defence against natural enemies and nutritional provision (Henry et al., 2015).

Insect microbial symbionts (Figure 1-2) commonly localise intracellularly in bacteriocytes and bacteriomes (Baumann, 2005), and extracellularly in insects' gut and gastric caeca (Kikuchi et al., 2011). Bacteriocytes are specialised host cells containing intracellular symbionts (endosymbionts). Bacteriomes are constructed from multiple bacteriocytes, and form specialised organs to contain endosymbionts (Baumann, 2005). Extracellular symbionts are usually found in insects specialised gastric caeca or crypts located in the digestive tract (Kikuchi et al., 2011). Additionally, the term of endosymbionts is frequently used in the study of symbionts, as they are commonly localised in internal organs and also within animal cells (Douglas, 2016).

These symbionts are acquired from the environment, and they also can be transmitted via egg-surface contamination by coprophagy (consumption of faeces) in species such as termites and stinkbugs (Kikuchi et al., 2011). Microorganisms in the insect digestive tract include protists, fungi, archaea and bacteria (Engel and Morgan, 2013). Protists were discovered in the lower termites and wood roaches, in which their persistence relies on social transmission (Hongoh, 2010). Fungi are frequently found in the insect gut, they are originated from wood or detritus, and these fungal symbionts usually take part in the process of host digestion (Engel and Morgan, 2013). Methanogenic archaea are most

studied in insects that eat wood or detritus such as beetles and termites (Egert et al., 2003; Lemke et al., 2003; Brune, 2010). Bacteria are found in the digestive tract of most insect species. However, many studies have been relied on bacterial 16S rRNA gene primers, probably biasing the structure of insect gut communities because of the variation of function and positions of these bacterial communities within the insect gut (Engel and Morgan, 2013). The groups of bacterial phyla often found in insect digestive tracts include Gammaproteobacteria, Alphaproteobacteria, Betaproteobacteria, Bacteroidetes, Firmicutes including *Lactobacillus* and *Bacillus* species, *Clostridia, Actinomycetes, Spirochetes, Verruomicrobia, Actinobacteria*, and others (Colman et al., 2012).

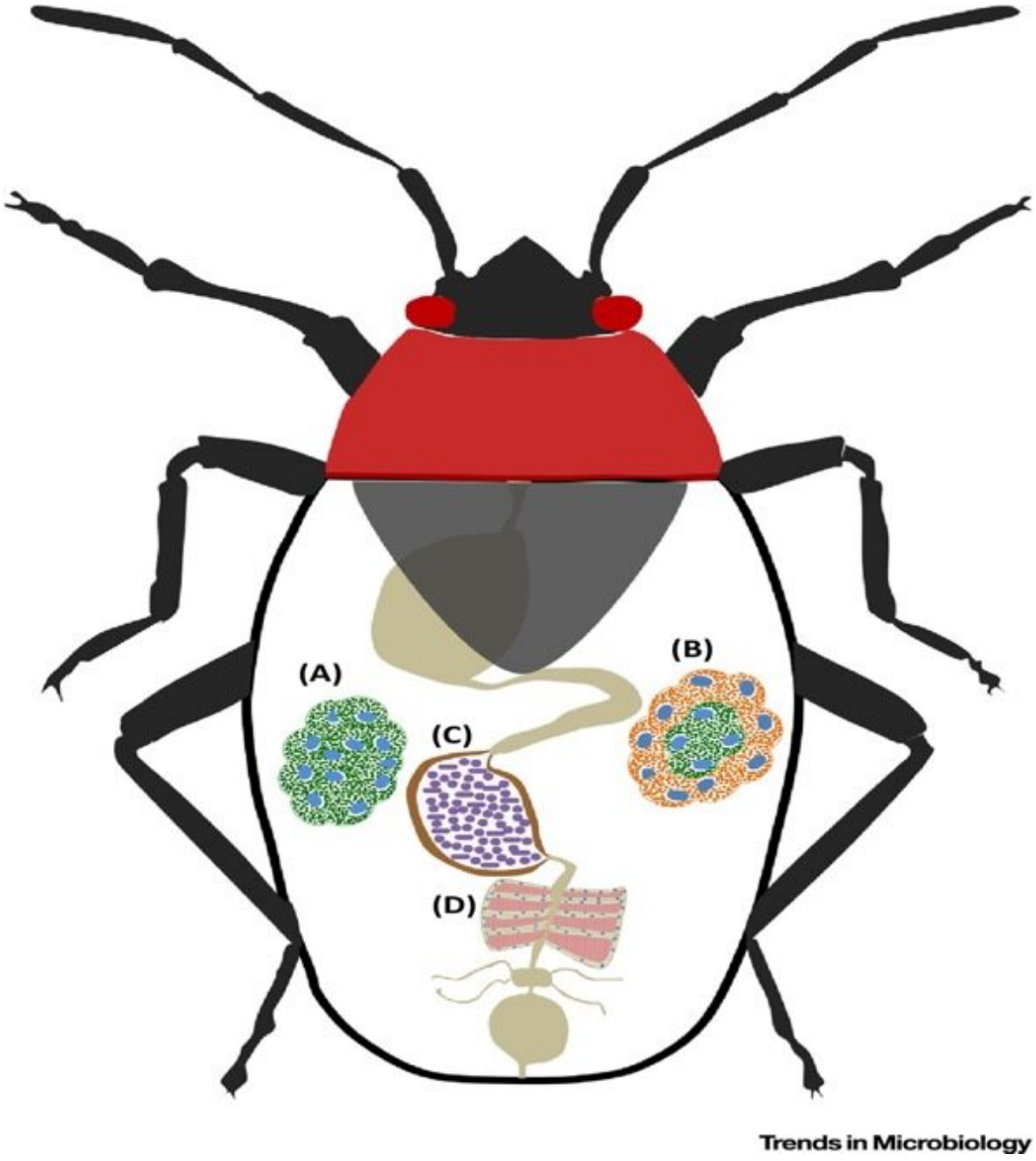


Figure 1-2: Different locations of symbionts associated with insects:

(A: Intracellular symbionts localised in a bacteriome (symbiont cells in green) B: Dual intracellular symbionts living in independent bacteriocytes within a bacteriome (green and orange fillings are symbionts). C: Extracellular symbionts in purple-filled circles located in the midgut. D: Extracellular symbionts (red filling) hidden in specialised gastric caeca or crypts (blue-filled small circles are host cell nuclei), (Figure reference from Sudakaran et al., 2017).)

The two categories of symbiotic associations in insect hosts are obligate (or primary) and facultative (or secondary) symbionts. Obligate symbionts are essential for their hosts'

successful development and reproduction, but the host will also suffer from the absence of its obligate symbionts. The features of obligate symbionts include that they usually share long term co-evolutionary history with the hosts, and they are always vertically transmitted (symbionts are transmitted from parents to offspring of the host and continue the relationship with the host's offspring), particularly in maternal transmission (transmission from mother site to next generation). Additionally, these obligate symbionts are localised intracellularly in bacteriomes. Commonly, such obligate symbionts can supply nutrients to their insect host and these nutrients are lacking in the host's plant or blood diet.

Obligate symbionts are commonly associated with the Hemiptera (Figure 1-2) and the orders of Hemiptera are composed of 82,000 known species in four main suborders which are Sternorrhyncha, Auchenorrhyncha, Coleorrhyncha, and Heteroptera (Cryan and Urban, 2011). In the suborder of Sternorrhyncha, most of the insects are primarily phloem sap-feeding, but this type of food resource is lacking in amino acids and cofactors, and also rich in sugars (Sandstrom and Moran, 1999). In order to balance the host's diet, microbial symbionts associated with the suborder of Sternorrhyncha can provide essential amino acids, vitamins and other beneficial microelements, allowing insect hosts to expand the range of host plants used as a primary food resource. Genome analysis of the bacterial endosymbiont *Carsonella* associated with the insect hosts psyllids (Psylloidea) is one example of symbionts providing essential nutrients to their host (Nakabachi et al., 2006).

Throughout the evolutionary history of Sternorrhynchan symbiosis, the symbionts have been replaced and transited numerous times from mono- to dual-symbiosis (Figure 1-3). For example, *Portiera* symbionts in the whitefly (*Bemisia tabaci*) can aid cofactor biosynthesis to produce several specific essential amino acids (Santos-Garcia et al., 2012). Recently, a putative co-obligate symbiont, *Hamiltonella defensa*, was found to complement the biosynthetic capabilities of Portiera, namely both *Hamiltonella defensa* and *Portiera* are obligate symbionts, residing within the same bacteriome in cells, where they help each other by producing certain amino acids essential for host survival (Luan et al., 2015). In the last Sternorrhynchan superfamily, Coccoidea, several co-obligate symbionts colonising the insects belonging to this order have been observed recently. For example, mealybugs (*Pseudococcidae*) are hosts to a unique nested symbiotic relationship consisting of Candidatus *Tremblaya* princeps from Betaproteobacteria and an intracellular gamma-proteobacterial symbiont named *Candidatus Moranella endobia* (McCutcheon

and von Dohlen., 2011; von Dohlen et al., 2001). These symbionts are also functional in providing essential amino acids to their host (McCutcheon et al., 2011). The order of Heteroptera comprises 40,000 species and is the largest hemipteran suborder. Nutritional symbiotic associations always occur in the infraorder Cimicorpha. For example, of two species in Cimicorpha, blood-sucking species of the *Rhodnius* genus (kissing bugs) have a relationship with symbiotic *Rhodococcus* (Actinobacteria) in their gut cavity while *Cimicidae* (bedbugs) harbour *Wolbachia* (Alphaproteobacteria) in specialised bacteriomes. Although these relationships involve two different symbiotic bacterial species in different locations, both symbionts supplement the diet of their host with vitamin B, which is lacking in vertebrate blood (Hosokawa, et al., 2010; Ben-Yakir, 1987).

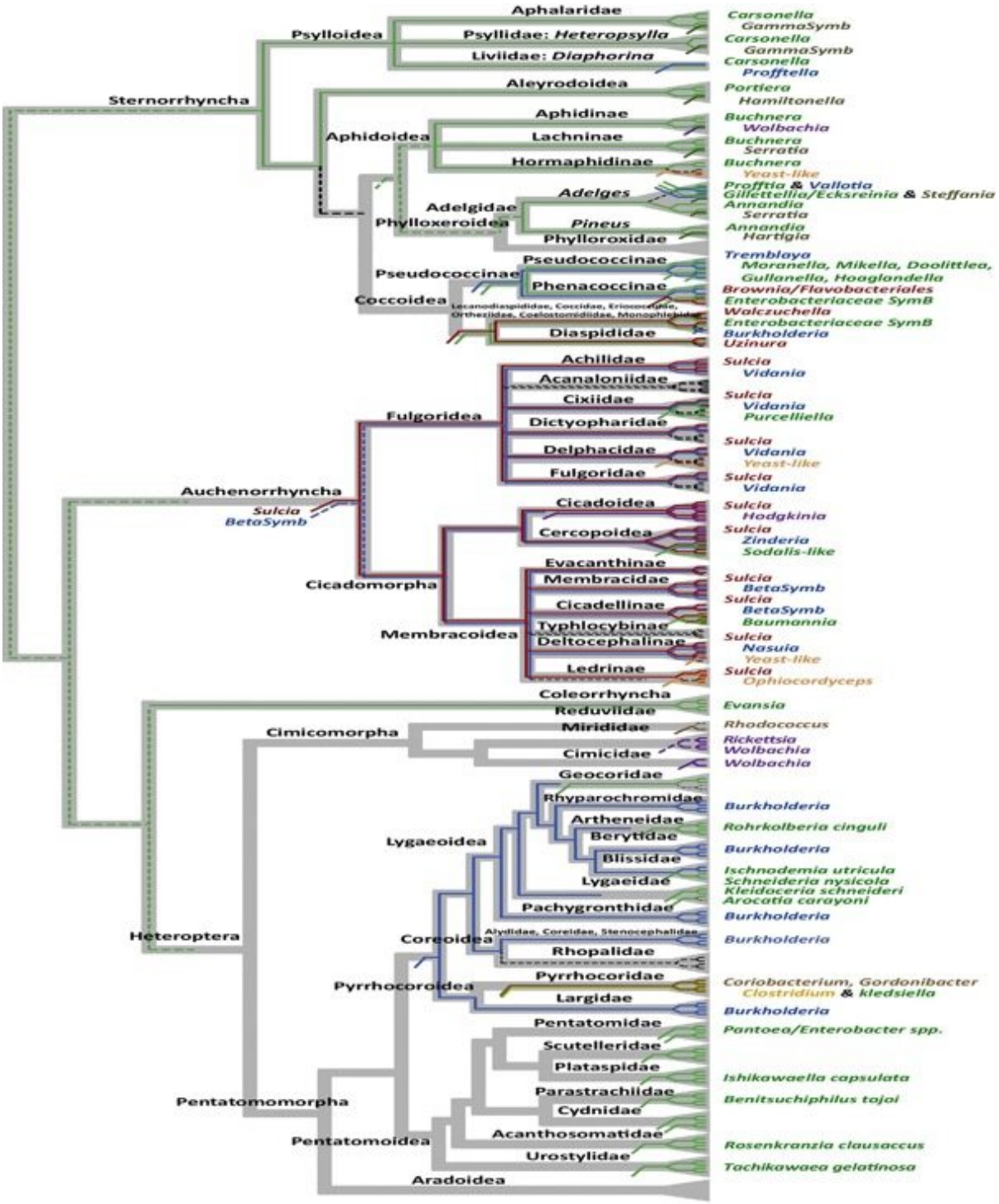


Figure 1-3: Evolutionary shifts in symbiotic associations across Hemiptera.

(A schematic host insect phylogeny is indicated in grey. Symbionts are colour-coded based on their taxonomic identity at phylum or class level: Gammaproteobacteria: green (light and dark green were used for dual gammaproteobacterial symbioses); Betaproteobacteria: blue; Alphaproteobacteria: purple; Bacteroidetes: red; Firmicutes: yellow; Actinobacteria: brown; Yeast-like fungal symbionts: orange. Black dashed lines: putative losses of symbionts, and coloured dashed lines indicate possible relatedness between symbionts that requires further experimental support. (Figure reference from Sudakaran et al., 2017).)

Facultative symbionts have been associated with the host for a shorter period, usually being transmitted via various routes both vertically and horizontally (horizontal transmission: symbionts are transmitted between hosts individually and establish the symbiotic relationship in each host generation anew) between different unrelated conspecific hosts. They also can be located intra- or extracellularly. Their presence in the host is not essential for the host's survival or development, although their absence may cause some light host fitness (Jurkevitch, 2011 and Sudakaran et al., 2017). The impact of facultative symbionts on populations of their eukaryotic host is unquestionable. They drive the evolution of numerous key traits such as disease resistance, predator defence and sex determination (Oliver et al., 2003; Hedges et al., 2008), as well as influencing reproductive behaviour (Rousset et al., 1992; Charlat et al., 2007). They facilitate niche expansion (Joy, 2013) and promote diversity and reproductive isolation (Bordenstein et al., 2001).

1.5 The functions of insect symbionts

The functions of endosymbionts can be defined as traits that are beneficial to the host. These functions include promotion of the host's evolutionary development, pathogen resistance or protection from environmental stress, and nutrient provision. Obligate endosymbionts are usually functional in supplying necessary nutrients to hemipteran hosts. Indeed, these endosymbionts are localised in bacteriomes or bacteriocytes to supplement essential amino acids (EAAs). EAAs comprise 10 of the total 20 amino acids–arginine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine–that contribute to eukaryotic protein production but cannot be synthesised by eukaryotes themselves. In particular, the hemipterans lack the urea cycle that can synthesise arginine, and endosymbionts can provide extra energy cost to synthesise this amino acid (Douglas, 2016).

Microbial species entering via the gut can be easily digested and used per se as nutrients (nutritional bacteria). For example, lysozymes expressed in the midgut of *Drosophila* have been shown to be beneficial to nutrition rather than immunity (Daffre et al., 1994). Drosophila produces over 10 distinct lysozymes in the midgut and contains a transporter displaying a high affinity for D-amino acids, which were discovered in peptidoglycan (PGN) (Miller et al., 2008).

The stinkbugs (Heteroptera), comprise more than 40,000 species, and plant-feeding stinkbugs in the family Plataspidae and Acanthosomatidae contain the gammaproteobacterial gut symbionts *Ishikawaella capsulata* and *Rosenkranzia clausaccus* localised in the inside of the midgut caecae or crypts, especially *I. capsulata* which always lives in the specialised caeca of the gut of stinkbugs species *Megacopta punctatissima* (family: Plataspidae) (Engel and Moran, 2013; Fukatsu & Hosokawa, 2002). *Ishikawaella capsulata* have a reduced genome size and only maintain the gene that provides nutrients to the host on a restricted diet of plant sap. Therefore, they lack numerous genes such as the genes responsible for synthesis of the cell wall and lipid metabolism for specialisation to conditions supplied by the host. Additionally, they are obligate symbionts, and their host species suffers a high rate of retarded growth and nymphal mortality in their absence, probably because these symbionts are extracellularly associated with the host. One evolutionary pattern of all host stinkbugs is the development of postnatal mechanisms for vertical symbiont transmission, while the intracellular symbiont undergoes prenatal transmission such as trans-ovarial transmission. Therefore, these symbionts could be easily invaded or replaced by foreign microbes (Kikuchi, 2009).

The gut microbiota of the fruit fly (*Drosophila melanogaster*) is an example of an open system, where both larvae and adults acquire the symbiotic microorganisms from the environment and previous life stages. The gut bacteria of *Drosophila melanogaster* include Lactobacillus*,* Acetobacteraceae, and Orbaceae genus bacteria. The roles played by gut bacteria of *Drosophila melanogaster* include promoting development of the host's immune system, as well as affecting metabolism and mating preferences.

The symbiont of the pea aphid (*Acyrthosiphon pisum*) is an example of bacteriocyte symbiosis, where vertical transmission of the microorganisms takes place through the female ovaries and symbionts are inserted into the developing embryo (Douglas, 2014). This symbiosis involves one type of heritable symbiont that is an obligate endosymbiont, located in the cytosol of specific host cells and providing limited necessary nutrients to the host. Bacteria of the pea aphid include *Staphylococcus, Pseudomonas, Acinetobacter,* and *Pantoea*. Their transmission route is from the environment such as phloem sap. They act as attractants and stimulate oviposition to increase the egg laying rate in the aphid (Leroy et al., 2011).

Genetically, *Ishikawaella capsulata* and *Rosenkranzia clausaccus* cluster together with the aphid intracellular symbiont *Buchnera.* Both the plataspid and acanthosomatid symbionts display some special genetic features including an AT-biased nucleotide structure, advanced molecular evolution, and significantly reduced genome size (Leroy et al., 2011). These features are usually retained in the obligate intracellular symbionts in many different insects such as *Buchnera* in the aphid and *Wigglesworthia* in tsetse flies, and it has been considered that the intracellular conditions inhabited by the symbiont may be relevant to the special features. However, the stinkbug gut symbionts illustrated that the intracellular location is not relevant to these molecular features. Instead, attenuated purified selection owing to a small population size and strong obstruction is a candidate mechanism that better explains the mode of evolution found in genes of the obligate mutualists.

Another functional trait indicates that a symbiont can protect the host in a specific ecological environment, such as protection against abiotic stress and natural enemies or pathogens (viruses & fungi). These traits are easily found in facultative endosymbionts, especially in *Sternorrhynchan hemipterans* (Douglas, 2016; Brownlie and Johnson, 2009). Hosts are protected from virus infection by *Wolbachia pipientis*. Commonly, the endosymbionts *Wolbachia* can manipulate reproductive behaviours of the hosts, but the insect *Drosophila melanogaster* is naturally infected by *Wolbachia* without strong reproductive parasitism and these bacteria are found in multiple different tissues within *D. melanogaster* and protect against viral infection in *D. melanogaster* due to the role of *Wolbachia*–modulated host factors towards RNA virus resistance in arthropods, alongside establishing methyltransferase gene Mt2's novel antiviral function against especially in Sindbis virus (Bhattacharya et al., 2017). *Hamiltonella defensa* in aphids increase host resistance against parasitisation by the parasitoid wasp (*Aphidius ervi*) and increase the mortality of the wasp larvae during their development (Brandt et al., 2017). The presence of *Regiella insecticola* and Streptomyces species prevent infection of pea aphids by fungal pathogens (mainly the Entomorphthorales fungus *Pandora neoaphidis*) by strong suppression of fungal sporulation (Ferrari and Vavre, 2011).

1.6 Co-evolution of symbiont and host

Most insects have an ancient symbiotic relationship (over one million years) with several vertically transmitted microbial endosymbionts. The host usually forces the symbiont to adapt to their obligate intracellular lifestyle though long-term genomic evolutionary changes, such as genomic reduction (Wilson and Duncan, 2015). Three significant features of genomic co-evolution between host and symbionts (termed holobiont genome evolution) are collaboration, acquisition, and constraint (Wilson and Duncan, 2015).

Collaboration of holobiont genome evolution is determined by symbiotic genomic involvement of the host metabolic pathway. For instance, the symbiosis helps the metabolism of the host at the molecular level. These obligate symbionts show adenine and thymine-biased nucleotide composition and high-base substitution rates in their genomes, as well as drastic genome reduction (Kikuchi, 2009). The genomes of these obligate symbionts (ranging from 0.9 Mb to below 0.1 Mb) are significantly smaller than cultivable relatives such as *Escherichia coli* K12 (4.6 Mb) and *Salmonella typhi* (4.8 Mb) (Wernegreen, 2002). However, these small genomes contain biosynthetic pathways that supply nutritional requirements of the host insects such as amino acid biosynthesis in *Planococcus citri* and vitamin B5 (pantothenate) biosynthesis in *Acyrthosiphon pisum*, while lacking several essential genes such as dnaA and/or recA that are important to microbial replication, at least in cultivable bacteria (Akman, 2002; McCutcheon and Dohlen, 2011; Husnik, 2013).

Acquisition is a second essential feature of holobiont genomic evolution. This feature alters genomic pathways associated with nutritional collaboration between insect and symbiont over millions of years by three key genomic mechanisms: gene duplication, lateral gene transfer and partial or full symbiont replacement (Wilson and Duncan, 2015). Gene duplication events facilitate refinement of existing function, the evolution of new spatial and temporal gene expression patterns, and even the evolution of new gene functions. Recent work in insect nutritional endosymbionts has focused on the evolution of nutrient amino acid transporters in the genomes of insects that feed on plant sap. Those studies find that the evolutionary history of amino acid transporter genes in plant sap-feeding insects is dynamic with respect to both duplication events and the recruitment of duplicated genes to the host/symbiont interface. The dynamic evolution of amino acid transporters in these insects, including some very recent duplications in aphids, demonstrates that despite millions of years of host/ endosymbiont coevolution, host

genomes are in flux. Lateral transfer of bacterial genes to host genomes has the potential to relieve coevolutionary constraints in a marvellous way. Such lateral gene transfer in the context of a nutritional endosymbiosis was first reported in *Acyrthosiphon pisum*. Whole-genome analysis of *A. pisum* identified the transfer of 12 genes or gene fragments from bacteria to the insect genome. Of those were highly expressed in bacteriocyte cells but none were implicated in amino acid or vitamin biosynthesis (Nikoh N, et al., 2010). Partial or full endosymbiont replacement by previously facultative insect symbionts is a third mode by which holobionts gain new genetic material. Even though symbiont genome evolution is characterized by genome degradation, many endosymbionts have coevolved with their insect hosts since ancient times. When bacteria become obligate symbionts, important population genetic parameters immediately change. Bacterial symbionts, unlike their free-living relatives, experience relaxed selection and greatly reduced population size, resulting in elevated genetic drift. in the Auchenorrhyncha, which typically have two primary symbionts, one symbiont (Sulcia) has been maintained through evolution whereas its coprimary symbiont has been replaced many times such that different Auchenorrhynchan lineages have alphaproteobacterial, betaproteobacterial, gammaproteobacterial, or yeast-like symbionts. Symbiont replacement can even arise when an ancestral symbiont diverges into two interdependent lineages (McCutcheon et al., 2009). The replacement or complementation of a primary endosymbiont by a more recently acquired (or derived) symbiont has the potential to functionally "reset" genes and pathways that have been eroded by mutation accumulation over evolutionary time (Wilson and Duncan, 2015).

In the third feature of holobiont genomic is constraints which can limit acquisition and collaboration features in the co-evolution of the holobiont to maintain the balance of holobiont co-evolution and thus prevent strong effects on host health fitness such as mortality or retarded growth (Wilson and Duncan, 2015). The same insect genes in different holobionts are engaged in host/symbiont metabolic collaboration, complementing the same symbiont gene losses. Opportunities to evolve collaborative biosynthesis are constrained by the gene content of host genomes. Furthermore, the coevolutionary potential of an endosymbiosis is constrained by the cell type that gave rise to the bacteriome in each taxon. Although all nucleated cell types contain the complete host genome, all cell types do not express all genes. Therefore, differences in basal expression of the cell lineage that gives rise to the bacteriome in each taxon will constrain

patterns of holobiont co- evolution. For example, the nutritional symbiosis between *A. pisum* and *Buchnera aphidicola*. This symbiosis generated numerous signatures of co-evolution between the partners, where the symbionts adapted to promote host-level fitness and vice-versa. For example, the biosynthesis of amino acids appears to have evolved in concert between the two partners. On one hand, the symbiont overproduces and regulates the biosynthesis of essential amino acids for the host. Interestingly, this biosynthesis pathway is itself shared between the partners: a reduced number of steps missing in the symbiont are performed by the host (Douglas, 2014). On the other hand, amino acids that are non-essential to the host are provided to the symbionts, which, in return, lose the biosynthesis pathways for these amino acids (Baumann, 2005). The host genome has also evolved multiple structures to ‘‘farm’’ and regulate endosymbiont activity. For example, many genes underlying responses to Gram-negative bacteria have been eliminated, including the immune-deficiency (IMD) signalling pathway (Douglas et al., 2011). The strict vertical inheritance of symbionts causes a drastic reduction of symbiont population sizes at each host generation, decreasing recombination rates within symbiont genomes (Groussin, Mazel and Alm, 2020). Metabolic collaboration, the acquisition of novel genomic material, and host genomic constraints are emergent features of host/symbiont genome coevolution. As more holobiont genomes are sequenced, the more studies will anticipate that these signatures will continue to be supported and that other as yet unidentified signatures will likely emerge.

1.7 *Serratia* species as symbionts and pathogens

Serratia species were first described as a new genus of bacteria by Merlino in 1924 (Merlino, 1924). Later studies of Serratia species focused on their evolution from the free form to a symbiotic lifestyle. Since *Serratia marcescens* was found to be related to bacterial endocarditis (Wheat et al., 1951), *S. marcescens* has been identified as a pathogen which can cause various infections, such as bacteraemia, pneumonia, keratitis, endocarditis, urinary tract infection, meningitis, and necrotizing fasciitis (Petersen and Tisa, 2013). The common virulence factors of *Serratia marcescens* mainly consist of protease activity, haemolysis, and adhesion (Petersen and Tisa, 2013). Especially, protease activity is essential in causing infection; it produces a protease which targets epithelial cells and immune resistance proteins to cause protein degradation and ultimately, cell lysis (Petersen and Tisa, 2013).

However, when *Serratia* species live in a symbiotic relationship with insects, they can protect the host from various conditions, an important example being the symbiotic relationship between *Serratia symbiotica* and different aphid species. *Serratia symbiotica* have been isolated from a variety of aphid species and they are divided into two phylogenetic groups (cluster A and B) based on the 16 rRNA gene (Lamelas et al., 2008). One member of *Serratia symbiotica* cluster B has become one of the primary symbionts in the aphid *Cinara cedri Mineur* (Lamelas et al., 2011). Since *Serratia symbiotica* transferred from a facultative symbiont to a co-obligate symbiont, its genome size has reduced, but it has high retention of transcription- and translation-related genes, having retained rpoD and rpoH genes coding for sigma 70 and sigma 32, respectively. Sigma 32 is important to the normal expression of heat-shock genes and the regulation of heat-shock proteins, so it can protect the host from heat shock (Manzano-Marin and Latorre, 2016). Furthermore, *Serratia symbiotica* also protects the host from parasitoid wasps by killing parasitoid larvae after oviposition (Oliver et al., 2005). Additionally, *Serratia symbiotica* and *Hamiltonella defensa* work together to defend aphids against predators by reducing the fitness and reproduction of ladybird beetles (Coleoptera: Coccinellidae) (Costopoulos et al., 2014).

Serratia sp. SCBI (termed South African *Caenorhabditis Briggsae* Isolate) is one species complex associated with the nematode *Caenorhabditis briggsae* KT0001 and is found in South Africa. The virulence factors of this species are similar with *Serratia marcescens* Db11, but virulence factors of these two species work differently, and *Serratia* sp. SCBI is not pathogenic to its host. *Serratia* sp. SCBI might be a strong pathogen, since its virulence factors are similar with *Serratia marcescens* Db11, but it does not negatively affect *Caenorhabditis briggsae* survival or reproduction (Lancaster et al., 2012).

1.8 *Orius*' symbionts

The study of association between *Orius* species and their symbionts has been neglected, with interest focused mainly on *Wolbachia* endosymbionts. *Wolbachia* belong to the order Rickettsiales in the family of Anaplasmataceae. This order of bacteria is associated with the genera *Anaplasma, Ehrlichia* and *Rickettsia*. It has attracted considerable attention during the past decade, primarily because it can alter the reproductive behaviour of its host

and could therefore be applied as a potential pest and disease control (Werren et al., 2008). *Wolbachia pipientis* is one of the species in the *Wolbachia* genus, first discovered in the mosquito *Culex pipiens*. According to 16S ribosomal sequence information, so far eight distinct supergroups of *Wolabachia* spp. have been discovered; the C and D supergroups usually appear in nematodes, while the other six supergroups are commonly associated with the arthropods, especially A and B supergroup which are the most related to insects (Werren et al., 2008).

One study on *Orius strigicollis* described that this insect is superinfected by two strains of *Wolbachia* sp. (wOus1 and wOus2), both inducing varying degrees of cytoplasmic incompatibility (CI). The occurrence of CI permits colonising bacteria to induce infertility between males carrying other *Wolbachia* strains and uninfected females, therefore ensuring transmission of the pre-existing endosymbionts to the progeny. The results from this study illustrated that wOus1 suppresses the ability of wOu2 to colonise by interfering with wOus2 densities in *Orius strigicolli*. Therefore, infection by CI-causing *Wolbachia* sp. in *Orius strigicollis* prevents additive infection (Watanabe et al., 2012). The presence of *Wolbachia* endosymbionts have been documented in several *Orius* species (*Orius sauteri, Orius nagaii, Orius minutus, Orius strigicollis*, and *Orius tantillus*), in all cases both wOus1 and wOus2 strains were identified (Watanabe et al., 2012).

There is a significant lack of studies concerning the functions of microbial symbionts on the deployment, speciation, fitness, and behaviour of *Orius* species, although *Orius* genus as an ABC are extensively used in IPM.

1.9 The bioinformatic techniques related to this study

1.9.1 genome assembly method

In order to achieve complete genome sequences of prokaryotes, the process of assembly are necessary, but always leave unordered assemblies and gaps due to short read length. The approach of Velvet assembly mainly manipulates de Bruijn graph through the simplification and compression without any loss of information by merging non-intersecting paths into single nodes. It also recognises and removes three main types of errors: tips causing by the errors at the edges of reads, “bubbles” because of internal read errors or to nearby tips connecting, and erroneous connections owing to cloning errors or

to distant merging tips. Finally, it combined short reads and read pairs to generate the contigs of reasonable length (Figure 4-5) (Zerbino et al., 2008).

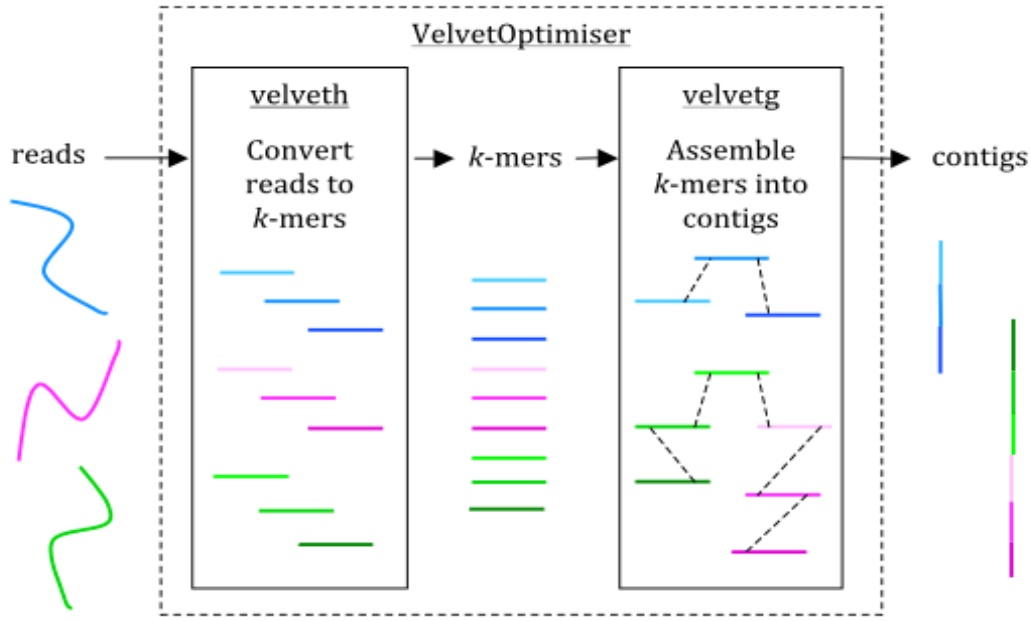


Figure 1-4: Genome assembly with Velvet. Reads are assembled into contigs using Velvet and Velvet Optimiser in two steps, (1) velveth converts reads to k-mers using a hash table, and (2) velvetg assembles overlapping k-mers into contigs via a de Bruijn graph. Velvet Optimiser can be used to automate the optimisation of k-mer length for velveth and velvetg and generate an optimal assembly (Edwards and Holt, 2013).

The SPAdes assembly method initially constructed an assembly graph by using multi-sized de Bruijn graph, and carried out new algorithms for removing tips, bubble, and chimeric reads. Then, pair of k-mers (k-bimer) were adjusted to derive accurate distance estimates between k-mers in the genome and paths in the assembly graph. After adjustment of k-bimer, paired assembly graph was constructed. Finally, contigs were produced. (Bankevich et al., 2012). The biggest difference between velvet and SPAdes assembly is the improvement of de Bruijn graph algorithms construction especially in iteration over values of kmer sizes, and incorporation of k-bimers, which allows information from paired end reads to be introduced into the computation at an earlier stage.

1.9.2 Genomic islands

The term horizontal gene transfer is used when bacteria acquire genetic material from species other than the immediate ancestor. The spread and exchange of antimicrobial resistance genes (ARGs) among different bacteria is mediated via Mobile Genetic

Elements (MGEs), which can spread to the same bacteria species, to species that are not closely related, and even across different phyla (Frost et al., 2005). MGEs include conjugative plasmids, gene cassettes within integrons, transposons and inserted sequence (IS) elements (Frost et al., 2005, Summers, 2006). MGEs, and especially plasmid conjugation, is the most prevalent HGT mechanism involved in the spread of ARGs (Nikaido, 2009). There are many reservoirs for the horizontal transfer of genes. Some limitations that control the extent of HGT include bacterial competence for transformation and the similarity of DNA to be taken and integrated, as well as the controlling role played by the recipient (Aminov, 2011). Additionally, HGT is the most widely known method in nature to induce diversity among the bacterial population (Thomas and Nielsen, 2005). HGT in *E.coli* may also contribute to the virulence of bacteria (Juhas et al., 2009).

1.9.3 Pangenome analysis

The pangenome is composed of a core genome, which includes all genes present in all the strains studied, encoding functions related to the basic biology and phenotypes of the species. A second component is an "accessory or dispensable genome," including genes present in some but not all strains studied as well as strain-specific genes. The dispensable genome is generally associated with nonessential functions, in addition to conferring selective advantages such as adaptability to ecological niches, the ability to colonize new environments, or antibiotic resistance (Rivera-Ramírez et al., 2022). To gain insight into the process of reductive evolution undergone by the adaptation from a free-living state to an endosymbiotic-lifestyle, the construction of a pan-genome for several representative genomes from the genus *Serratia* and *Serratia* sp. *Orius* isolates is necessary in this study.

1.9.4 Type 6 Secretion system (T6SS) detection

Serratia species as antagonistic bacteria exploit several different strategies to outperform their competitors. The most characterized antimicrobial compound in *Serratia* species is the red pigmented prodiginine, of which five types have been identified so far (prodigiosin, undecylprodigiosin, cycloprodigiosin, cyclo-nonylprodigiosin, and butyl-meta-cyclo-

heptylprodiginine) (Li et al., 2015). Prodigiosin is commonly produced by environmental isolates of *S. marcescens*, but not the clinical isolates (Li et al., 2015). In addition to its anti-bacterial, anti-fungal and anti-protozoal properties, prodiginine was recently reported to exhibit immunosuppressive and anticancer traits (Li et al., 2015).

Another central strategy employed is manifested via the protein secretion systems, through which Gram-negative bacteria secrete effectors to their exterior (Li et al., 2015). In particular, the Type VI secretion system (T6SS), the most recently described of the six, has highly versatile functions, which include eukaryotic and bacterial cell targeting, gene regulation, conjugation, and cellular adhesion (Li et al., 2015). Recent studies demonstrated that T6SSs in *S. marcescens* and other bacteria, such as *Vibrio cholerae*, *Burkholderia thailandensis* (Li et al., 2015), can target other bacterial competitors resulting in either growth inhibition or death (Li et al., 2015).

1.10 The aims of this project:

Most *Orius* species are natural pest control agents, but the boundaries between pathogenic and mutualistic microorganisms are unclear, especially in the case of facultative symbionts which can transfer horizontally between prey and predator, so these symbionts could cause human disease as well as plant pathogenic infections. To address the knowledge gap in understanding of cultivable microorganisms associated to *Orius* species, this project aimed to fill this knowledge gap of multiple different Orius species. Investigations consisted of:

- Taxonomic classification of the insect specimens by coxI amplification and coxI sequence phylogeny to confirm the identity of multiple *Orius* specimens and to indicate the evolutionary association between different *Orius* sp. insect hosts.
- Collection of isolated culturable microorganisms from *O. laevigatus* and various other *Orius* species at various geographical locations and initial identification of these isolates by morphological differences. Initial classification of these isolates to amplify and sequence the 16s rRNA. Due to the limitation of 16s rRNA classification, we went on to sequence the whole bacterial genomes and assemble these genome sequences to obtain draft genome sequences.
- Phylogenomic study of three predominant facultative symbionts by performing whole genome sequence comparisons to define the diversity or stability of the

genomes of the main symbionts isolated from these *Orius* species through Multi-locus Sequence Analysis (MLSA) and Genome-to-Genome Distance Calculation (GGDC) phylogenomic analysis.

- The genome plasticity of these symbionts was estimated by Genomic Island (GI) prediction to differentiate between *Serratia* sp. *Orius* isolates.
- Generation of a pan-genome sequence of *Serratia* sp. *Orius* isolates to define genetic traits and genes associated with the host–symbiont interaction.
- Detection of Type VI secretion system in *Serratia* sp. *Orius* isolates.

CHAPTER 2: GENERAL MATERIALS AND METHODS

2.1 *Orius* sp. insect sampling and maintenance

Orius sp. insect specimens were sampled at 15 locations in four different countries (Italy, Spain, Switzerland, and Greece) in the European continent (Table 2-1), prior to 2013 and used to establish both laboratory and agricultural field populations by our Spanish collaborator (Dr. Pablo Bielza Lino, Departamento de Producción Vegetal Escuela Técnica Superior de Ingeniería Agronómica Universidad Politécnica de Cartagena, Paseo Alfonso XIII, 4830203 CartagenaÉ 968325541). The commercial strains of *Orius laevigatus* were obtained from Syngenta Bioline®, Koppert (THRIPOR-L), BioBest (*Orius*-system) and AgroBio (Oricontrol). *O. pallidicornis* samples were collected from the field in southeast Spain and processed immediately or preserved in 70% ethanol.

The lab-rearing procedures for these insect populations were as follows: They were caged in one litre plastic containers covered by filter paper with air vents. Sterilised buckwheat husk (Fagopyrum) used as refuge was placed in the container and the insects were reared on a diet of UV-sterilized *Ephestia kuehniella* (Lepidoptera: Pyralidae) eggs and sterile water with cotton. The insect eggs were laid on surface-sterilized pods of flat green beans (Phaseolus vulgaris) with the ends sealed by paraffin wax. The laboratory was held at 25 to 26°C, with humidity of 70–80% RH and a light:dark ratio of 16L:8D daylight cycle. Every two or three days, *Orius* sp. eggs were collected from these insect populations to build another new age cohort of insects in case of cannibalism.

Live insects were shipped to Swansea and upon arrival a portion of adults were stored at -20°C for total insect DNA extraction, while others were used to isolate and culture insect symbiotic bacteria.

2.2 Isolation of bacteria from insects

Around 10 insects from each population were surface sterilised with 70% ethanol, followed by repeated rinses in sterile water. These samples were mechanically homogenised in 50 µL of liquid NB (nutrient broth) using micro-pestles. Two samples of

the same serial dilutions of the homogenate were plated onto NB agar using aseptic techniques, solid versions of this medium were prepared by adding 2% agar. All the plates were incubated at 28°C until colonies were visible (around 2–3 days), but one set of plates was placed in an AnaeroPack™ 2.5 L rectangular jar (Thermo Fisher Scientific) for bacterial culture in anaerobic conditions, and liquid cultures were shaken at 250 rpm overnight at 28°C. The procedures for isolation of bacteria from *Ephestia kuehniella* eggs, flat green bean pods and buckwheat husks were the same as for determination of the insects' microorganism content.

Initial classification of all isolated bacteria was carried out based on colony morphology, then colonies were sub-cultured to purity and single colonies were used to prepare 40% glycerol stocks to be stored at -20°C.

2.3 DNA Extraction

2.3.1 From *Orius* sp. insects

DNeasy Blood & Tissue kit (Cat No./ID: 51104) from Qiagen was used to extract *Orius* sp. DNA. First, every insect sample was surface sterilised as described above and crushed with a sterilised pestle in an autoclaved micro-centrifuge tube. Then 180 µL of ATL buffer from the kit and 20 µiL of Proteinase K (20 mg/mL) were added and incubated at 56°C overnight in a rotating thermo-incubator for increasing DNA yield. Total insect DNA extraction then proceeded following the insect protocol provided with the kit.

2.3.2 From bacterial cells

Bacteria isolated from all insect populations were grown in NB liquid at 28°C with shaking (250 rpm) until mid-logarithmic growth phase. Then genomic DNA was extracted from liquid cultures using the QIAmp mini kit (Qiagen), following the protocol provided for extraction from bacteria.

2.4 PCR (Polymerase Chain Reaction)

All PCR reactions were performed using MangoMix™, Bioline. All reactions were carried out in a 20 µL reaction mix, with the addition of 2.5 µL of 4 µM forward and reverse primers and completed in a Bio-Rad C1000TM thermo-cycler (Bio-Rad, Hercules, CA, USA).

2.4.1 Colony PCR

The 16S rRNA gene U1 primers (Table 2-1) were used to amplify the 16S rRNA gene and the gene product size should be 800bp from each morphological type of colony for initial taxonomic classification with isolates. The success of the PCR reaction was assessed by agarose gel electrophoresis. The reactions were completed in a Bio-Rad C1000™ thermo-cycler using the program: initial denaturation of 3 minutes at 95°C, 35 cycles of 30 seconds at 95°C, 30 seconds at 55°C annealing temperature, and one-minute extension at 72°C. As a final step, PCR products were held at 12°C and cleaning up following the steps of 2.4.4 section. Finally, all the PCR products were stored at -20°C until sanger sequencing, using the primers used for PCR.

2.4.2 *Orius* sp. specimens *cox*I (mitochondrial cytochrome c oxidase subunit I-MTCOI) PCR

*Cox*I primers (Table 2-1) were used to amplify the *coxI* gene from total DNA of *Orius* sp. specimens (Folmer et al., 1994), for taxonomic classification. Reactions were amplified through 35 cycles as follows: 1 minute at 95°C, 1 minute at 49°C (annealing temperature), and 90 seconds at 72°C, followed by a final extension step at 72°C for 7 minutes and incubation at 12°C and cleaning up following the steps of 2.4.4 section. Finally, the PCR products were stored at -20°C until sanger sequencing using the primers used for PCR.

Table 2-1: *Cox*I and 16S rRNA primers information

PRIMER NAME	PRIMER SEQUENCE 5'-3'	SOURCE
COXILCO1490F (INITIAL COXI)	GGTCAACAAATCATAAAGATATTGG	Folmer et al., (1994)
COXIHCO2198R (INITIAL COXI)	TTAACTTCAGGGTGACCAAAAAATCA	Folmer et al., (1994)
16 rRNA U1F	ACGCGTCGACAGAGTTTGATCCTGGCT	James, G. (2010)

16 rRNA U1R	GGACTACCAGGGTATCTAAT	James, G. (2010)

2.4.3 Genome-specific PCR for detection of *Orius* sp. symbionts

Once the draft genome sequences of putative bacterial symbionts were obtained (see section 2.5 below), genome-specific genes were identified by searching existing databases with BLASTN, using as query the all the ORFs identified in the sequenced genomes from each species group identified (*Serratia*, *Leucobacter* and *Erwiniaceae*). ORF without homologous sequences in databases were selected as genome-specific targets. The primer sets (Table 2-2) were designed using the primer Basic Local Alignment Search Tool (Primer-Blast; Ye et al., 2012) with the selected genome-specific ORFs. One micro-litre of total insect DNA was used as DNA template in the reactions. *F. occidentalis* insect DNA and *Ephestia* egg DNA were used as negative controls in the PCR reactions. The reactions were performed using the following program: denaturation for 3 minutes at 95°C, 35 cycles of 30 seconds at 95°C, 30 seconds at 55°C (annealing temperature) and 1 minute at 72°C, and a final extension for 5 minutes at 72°C. A final holding step at 12°C was used until samples were collected. The primer sets are shown in Table 2-2.

Table 2-2 List of *Orius* symbiotic-specific primer sets

Primer	Sequence	locus_tag	Application
OLBL1620F	GCAACGTTTCGGCATTGAGT	BMF92_08790	*Serratia* sp *Orius* isolate specific PCR
OLBL1620R	CATGCGTGGCTTCCTCAGTA	BMF92_08790	*Serratia* sp *Orius* isolate specific PCR
OLAL1106F	CCAGTCATGCTGGTTCCTGT	BVU99_18640	*Serratia* sp *Orius* isolate specific PCR
OLAL1106R	ATGCCTCGCTAGATTCAGGC	BVU99_18640	*Serratia* sp *Orius* isolate specific PCR
OLFS546F	GCCGGAGATTTTTGGGGAGA	BL249_13365	*Erwinia* sp. *Orius* isolate specific PCR
OLFS546R	CACCGGGGTGAAAGTAACGA	BL249_13365	*Erwinia* sp. *Orius* isolate specific PCR
OLAS2480F	AAGGCATCCACTTCTACGGC	BMH27_09045	*Leucobacter* sp. *Orius* isolate specific PCR
OLAS2480R	GAACGGGTCGTCGTTCTTCT	BMH27_09045	*Leucobacter* sp. *Orius* isolate specific PCR

2.4.4 PCR product precipitation for Sanger sequencing

PCR amplicons were precipitated by adding one tenth volume of 3 M sodium acetate and three volumes of absolute ethanol and stored at -20°C overnight. Then the precipitated PCR products were centrifuged for 30 minutes at 13,000 rpm in a cold room, and the supernatant was carefully discarded. The pellets were washed with 70% ethanol and

centrifuged again as described, then the supernatant was discarded, and the pellet was air dried. All amplicons were ready to sequence in both forward and reverse directions using the original PCR primers. All PCR products were sequenced at a commercial sequencing facility (LGC Genomics, Berlin, Germany).

2.5 Genome sequencing

Bacterial genomic DNA was quality controlled by spectrophotometry and electrophoresis. Genome sequencing was performed on an Illumina MiSeq platform. Dr. Matt Hitchings of Swansea University created genomic DNA libraries in preparation for sequencing using Illumina Nextera XT sample preparation technology. The paired-end sequencing raw reads were trimmed for Illumina and Nextera adapters strings using the Trim Galore wrapper tool (Martin, 2011). Additionally, these reads were qualified by QC report in the Galaxy server (https://orione.crs4.it). Low-quality base calls were removed prior to assembly of the reads into contigs using the SPAdes assembler v3.5.0 (Bankevich et al., 2012) on the Galaxy server. In addition, contigs below 500 bp in length were filtered before the assemblies were evaluated by Quality Assessment Tool for Genome Assemblies (QUAST) (Gurevich et al., 2013) and initially annotated using the Rapid Annotation using Subsystem Technology server (RAST) (Aziz et al., 2008).

2.6 Genomic data analyses

2.6.1 *Cox*I phylogeny methods

All *cox*I sequence alignments were performed using maximum-likelihood (ML) phylogenetic trees for these sequences, implemented in Molecular Evolutionary Genetic Analysis- (MEGA)-7 (Kumar et al., 2016) using the Tamura-Nei model (Tamura and Nei, 1993) and uniform rates among sites.

The phylogenetic tree support was obtained using 1,000 bootstrap pseudoreplicates (Felsenstein, 1985). In parallel, BLAST (Basic Local Alignment Search Tool) (Altschul et

al., 1990) was used to retrieve all *Orius* sp. *cox*I sequences available at the National Centre for Biotechnology Information (NCBI) genetic sequence database.

2.6.2 Bacterial symbionts phylogeny methods

The initial taxonomic classification of the draft genomes of all large and small isolates was attempted on AmphoraNet (Kerepesi et al., 2014). Annotation of all draft genomes was undertaken using RAST. The evolutionary relationship between all available the Enterobacteriaceae, Actinobacteria as reference genomes (accessed on Jan 2017) and bacterial symbiotic genomes of *Orius* sp. specimens was determined using a multi-locus sequence alignment tool using encoded amino acid sequences by the genomes under analysis. The combination and alignment of the 400 marker protein sequences were implemented by PhyloPhlAn (Segata et al., 2013). This analysis was undertaken by Dr. Paul Facey, Swansea University, due to the computational capacity needed. The principle is that translated CDS files (.faa) from reference genomes were searched by the GenBank FTP site. An unrooted ML phylogeny was constructed from these alignments in FastTree MP (Price et al., 2010) and carried out using the model of JTT+CAT with 20 discrete categories (-cat 20) in the server of Cipres Science Gateway (Miller et al., 2010). The phylogeny support was obtained using 1,000 bootstrap pseudoreplicates. Additionally, due to the whole phylogentic tree is too large to add in later chapters, so the whole tree won't add in following chapters, if you want to see it, please email me or my supervisor.

2.6.3 Species delimitation of *Orius'* symbionts by genome-to-genome distance calculation (GGDC)

The reference genomes used in GGDC were chosen by PhyloPhlan phylogeny based on their phylogenetic clade distributions. Then similarity within strains of isolates was calculated using *in silico* DNA–DNA hybridization (DDH) implemented with the genome-to-genome distance calculator (Meier-Kolthoff et al, 2013) and a distance threshold of 70% as recommended by Formula 2 (identities / HSP -high-scoring segment pairs length) for draft genomes. Additionally, a data matrix was created according to the value of the

distance from Formula 2 between different isolates and reference genomes, in order to generate a phylogeny of GGDC.

2.6.4 Genomic Island (GI) detection

GI predictions were used for analysis of the horizontal gene transfer of *Serratia* sp. isolates draft genomes using *Serratia* sp. SCBI as a reference. GIs of nine bacterial genomes were obtained from the IslandViewer 3 database (Dhillon et al 2015) (http://www.pathogenomics.sfu.ca/islandviewer). Predicted GI sequences were grouped by CD-Hit (http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-est) using an 80% identity threshold to identify representative GIs. A hierarchical clustering analysis was carried out using the presence of clustered GIs to assess the relatedness of the isolates based on genomic GI content. The dendrogram and binary matrix were visualized using iTOL (Letunic and Bork, 2016).

2.6.5 Pangenome analysis

2.6.6 Detection of Type VI secretion system (T6SS) in *Serratia* sp. *Orius* isolates

All draft genomes from *Serratia* sp. *Orius* isolates were combined into pseudogenomes, using combining contigs website (https://www.bioinformatics.org/sms2/combine_fasta.html), which combined all the contigs from a genome to a single contig and annotated by Prokka with GenBank formats. SecReT6 (https://bioinfo-mml.sjtu.edu.cn/SecReT6/) was used for the detection of T6SS component genes and effectors, and classification of T6SS subtypes in these genomes. BLASTP was used to identify genes with unknown function in SecReT6 detections for analysis of T6SS adaptors and effectors.

CHAPTER 3: Molecular taxonomic classification of *Orius* sp. specimens

3.1 Abstract in this chapter

- The taxonomic classification of all *Orius* insect specimens used in the study was confirmed by the cytochrome c oxidase (*cox*I) sequences phylogeny.

3.2 Introduction

This chapter is aiming to provide conformation of the taxonomic classification of *Orius* specimens by molecular phylogeny using *cox*I gene marker. The fact that many insect species are difficult to be discriminated at the morphological level, as well as the huge number of cryptic species, makes the traditional classification uncertain (Scheffers et al., 2012). The adoption of DNA-based molecular markers represents a satisfactory alternative. Since the proposal of DNA barcoding in 2003, subunit I (658 bp) of the mitochondrial cytochrome C oxidase (COX) gene (namely COI) became the most universal marker for species identification in the animal kingdom (Hebert et al., 2003).

Commonly, phylogeny is a branch of genomics that studies the evolution of genomes. Phylogeny analysis is the process of reconstructing evolutionary relationships among taxa, using a phylogenetic tree. In the phylogeny tree, all the branches on the tree are called "branches" or "taxa". Commonly, the results of phylogeny analysis illustrate how closely related two or more species are to each other (the topology of the tree) (McLennan, 2010). Phylogenies can be used for many purposes including: identifying related organisms based on their common ancestry; reconstructing evolutionary relationships among taxa using genetic data (e.g., inferring ancestral states); determining whether similar traits evolve independently within lineages; detecting patterns of selection on gene regions across populations/species; determining whether similar phenotypes evolve under similar selective pressures during speciation events; searching for conserved regulatory elements across genomes (McLennan, 2010). It uses phylogenetics to identify and reconstruct evolutionary trees, which are then used as a framework for understanding how genes evolved within those lineages (Gregory, 2008). Phylogenetics is also used for classifying

species into higher taxonomic categories such as genus or family. The analysis can be applied to all organisms, from prokaryotes to humans. Molecular phylogenetic trees have been constructed using both protein-coding genes (DNA) and noncoding DNA (e.g., rRNA and tRNA) (Gregory, 2008). Furthermore, the phylogeny of insects is the science about the evolutionary history of insects (McLennan, 2010). The result is an evolutionary tree or cladogram that shows how different species are related to one another based on their shared characteristics (Gregory, 2008).

3.3 Method

The approaches used for *Orius* insect sampling has been described in the Methods section (Chapter 2). All *Orius* samples were collected in Spain mainly, and randomly across other European countries such as Italy, Greece, and Switzerland (Table 3-1 and Figure 3-1). The specimens from different commercial lines were chosen as controls for populations not reared in the lab, and field collected specimens. Furthermore, the samples of *O. pallidicornis* cannot be reared in the lab because this species had to feed with the fresh pollen from *Ecballium elaterium*, so all the samples from this species used in this study were collected in the field. Total DNA from multiple insects was extracted in the yield of DNA was calculated by the DNA concentration x total sample volume. The minimum yield of DNA and the maximum yield of DNA in these samples were 25 ng and 180.55 ng respectively. The amount of template DNA used for PCR amplification of *cox*I was at least 5ng.

Table 3-1 Sampling information for all *Orius* species in the project

Sample name	Insect collection	Country and city, description	Reared condition
OLA	*O. laevigatus*	Samaria (Crete, Greece)	Lab reared
OLB	*O. laevigatus*	Cazorla (Jaen, Southeast Spain)	Lab reared
OLC	*O. laevigatus*	Hellín (Albacete, Southeast Spain)	Lab reared
OLD	*O. laevigatus*	Policoro (Matera, South Italy)	Lab reared
OLE	*O. laevigatus*	Acate (Sicilia, Italy)	Lab reared
OLF	*O. laevigatus*	Portonovo (Pontevedra, Northwest Spain)	Lab reared
OLH	*O. laevigatus*	Carmona (Sevilla, Southwest Spain)	Lab reared
OLI	*O. laevigatus*	Cabo de Gata (Almería, Southeast Spain)	Lab reared
OLJ	*O. laevigatus*	Ruidera (Ciudad Real, Central Spain)	Lab reared
OLS12	*O. laevigatus*	Syngenta Bioline	Lab reared
OLCA19	*O. laevigatus*	Cartagena (Murcia, Southeast Spain)	Lab reared

OLT20	*O. laevigatus*	Teruel (North Spain)	Lab reared
OLMT26	*O. laevigatus*	Méntrida (Toledo, Central Spain)	Lab reared
OLLO30	*O. laevigatus*	Logroño (La Rioja, North Spain)	Lab reared
OLMD33	*O. laevigatus*	Mérida (Extremadura, West Spain)	Lab reared
OAM11	*O. albidipennis*	Different origins (mixed population)	Lab reared
OA2	*O. albidipennis*	Different origins (mixed population)	Lab reared
OPN1	*O. pallidicornis*	Cartagena (Murcia, Southeast Spain)	Lab reared
OPM	*O. pallidicornis*	Miranda (Cartagena, Southeast Spain)	Lab reared
OPT	*O. pallidicornis*	Torre Pacheco (Murcia, Southeast Spain)	Lab reared
OP2	*O. pallidicornis*	Cartagena (Murcia, Southeast Spain)	Lab reared
OSP9	*O. niger*	Basel (Swizerland)	Lab reared
OM2	*O. mujusculus*	Cabrils (Barcelona, Northeast Spain)	Lab reared
OLSgn	*O. laevigatus*	Cartagena (Murcia, Southeast Spain), Syngenta	commercial line
OLBio	*O. laevigatus*	Cartagena (Murcia, Southeast Spain), Biobest	commercial line
Olwild	*O. laevigatus*	Cartagena (Murcia, Southeast Spain), wild	agriculture field collected
OLkopp	*O. laevigatus*	Cartagena (Murcia, Southeast Spain), Koppert	commercial line
OLAgr	*O. laevigatus*	Cartagena (Murcia, Southeast Spain), Agrobio	commercial line
OA18156	*O. albidipennis*	Cartagena (Murcia, Southeast Spain), wild	agriculture field collected
ON8516	*O. niger*	La Manga (Murcia, Southeast Spain)	agriculture field collected
OM11516	*O. mujusculus*	Cabrils (Barcelona, Northeast Spain)	agriculture field collected
OP18516	*O. pallidicornis*	Cartagena (Murcia, Southeast Spain)	agriculture field collected

Figure 3-1: The geographic distributions of all *Orius* populations from 5 different *Orius* species in European countries.

(*O. majusculus* (OM2 & OM11516), *O. pallidicornis* (OP2, OPN1, OP18516, OPM & OPT), *O. niger* (OSP9), *O. laevigatus* (the rest of the populations, commercial lines of *O. laevigatus* were purchased

from commercial suppliers and presumably the details are provided in the materials and methods chapter) and *O. albidipennis* were collected from mix populations in all these European countries.)

Initial taxonomical classification of all field collected insect specimens were based on their morphological differences and preserved in 70% ethanol or reared in the lab to establish the populations. In this study, the taxonomic classification of these insect specimens was confirmed by *cox*I sequence phylogeny, using representative *cox*I sequences from several different *Orius* species.

3.4 Results

3.4.1 Initial genetic taxonomic classification of *Orius* specimens

Initial taxonomic classification of field collected *Orius* specimens was performed by a collaborator (Dr Bielza) based on their morphological characteristics (Ferragut and Gonzalez-Zamora, 1994). For genetic taxonomy, *cox*I gene was amplified for all specimens of *Orius* species in table 3-1 and subjected to sanger sequencing. All the *cox*I alignments of *Orius* specimens were identified on the NCBI database using BLAST (Alschul et al., 1990), and the sequences from closely related species retrieved (Table 3-1).

Query coverage in BLAST searching is the percentage of *cox*I sequences aligned. At least 89% of each specimen's *cox*I sequence was aligned against the database, and the *cox*I alignments of *Orius* specimens showed high percentage of identity on NCBI database, except *O. albidipennis* and *O. pallidicornis* samples due to unavailable sequences from these species in the database (Table 3-2).

Table 3-2: *Orius* coxI BLAST alignments information

Sample name	BLAST alignment name	Query cover	Identity	Accession
OLA	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H4	90%	100%	FM210186.1
OLB	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H4	90%	99%	FM210186.1
OLC	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H5	91%	99%	FM210187.1
OLD	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H5	90%	99%	FM210187.1
OLE	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H4	91%	99%	FM210186.1
OLF	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H4	91%	99%	FM210186.1
OLH	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H2	89%	99%	FM210184.1
OLI	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H2	90%	99%	FM210184.1
OLJ	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H5	90%	99%	FM210187.1
OLS12	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H3	90%	99%	FM210185.1
OLCA19	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H1	90%	100%	FM210183.1
OLT20	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H5	90%	99%	FM210187.1
OLMT26	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H5	90%	99%	FM210187.1
OLLO30	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H3	90%	99%	FM210185.1
OLMD33	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H4	90%	100%	FM210186.1
OA2	Orius niger voucher EUBUG_965_m_Oriunige10 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	90%	KM022197.1
OAM11	Orius niger voucher EUBUG_965_m_Oriunige10 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	90%	KM022197.1
OPN1	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	94%	KM023138.1
OPM	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	94%	KM023138.1
OPT	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	94%	KM023138.1
OP2	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	94%	KM023138.1
OM2	Orius majusculus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H3	89%	99%	KJ467501.1
OLSgn	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H4	90%	100%	FM210186.1
OLBio	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H5	90%	100%	FM210186.1
Olwild	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H6	90%	100%	FM210186.1
OLkopp	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H7	90%	100%	FM210186.1
OLAgr	Orius laevigatus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H8	90%	100%	FM210186.1
OA18156	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	90%	KM022197.1
ON8516	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	90%	KM022197.1
OM11516	Orius majusculus mitochondrial partial coi gene for cytochrome oxidase sub-unit 1, exon 1, allele H3	89%	99%	KJ467501.1
OP18516	Orius niger voucher EUBUG_605_f_Oriunige4 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	94%	94%	KM023138.1

3.4.2 *Cox*I sequence Phylogeny

To depict the evolutionary relationship of the five *Orius* species used in the study an ML phylogenetic tree was constructed from the alignment of all *cox*I sequences in Table 3-1 and three *coxl* sequences of *Orius laevigatus, Orius niger* and *Orius majusculus* randomly choosen from NCBI database. One non-*Orius* species (*Accompocoris alpinus*, NCBI accession number JSTR001340204) *cox*I sequence was used as a root (Figure 3-2).

Ultimately, all five species share a single common ancestor at the deepest internal node, also known as the "root" of the tree. Overall, the shape of the tree and therefore the pattern of branching that it hypothesizes are known as its "topology" (McLennan, 2010).

By the definition, the more common ancestors that the species share to the exclusion of other species, the more closely related they are. In both figures, from the terminal nodes to the root, all the species of *Orius laevigatus* share 7 common ancestors, species *Orius laevigatus* and *Orius niger* share 6 common ancestors, and species *Acompocoris alpinus* shares only one ancestor (the root itself) with any of the other *Orius* species. specimens of *Orius laevigatus* and reference *Orius laevigatus* species linked through a recent common ancestor that not shared by any other taxa on the tree and therefore known as "sister taxa". The next closest relative of species *Orius laevigatus* is *Orius niger* mainly, with whom they share an ancestor to the exclusion of *Orius majusculus, Orius albidipennis* and other reference *Orius* species in these phylogenic trees (Figure 3-2 and 3-3). Additionally, the other reference *Orius* species (*Orius tristicolor, Orius minutus, Orius vicinus, Orius laticollis, Orius minutus,* and *Orius sauteri*) formed an independent clade, these species tend to be a same species of *Orius*, it could be reviewed in the future for further classification of *Orius* species. *Acompocoris alpinus*, by contrast, it does not link to any of other species beyond a single distant ancestor and it known as the "outgroup". An outgroup is necessary to root a tree (unrooted trees also can be drawn, but these are less informative and are not covered here).

Since the genetic information of *O. albidipennis* and *O. pallidicornis* are unavailable in databases, there are no *cox*I reference sequences from these two species in the phylogeny. Additionally, both species represent two independent monophyletic lineages (Figure 3-2). It is essential to indicate that this is the first study to provide any sequences information for these two *Orius* species and to confirm their status as potentially unique species.

Interestingly, both results of BLAST searching, and ML phylogeny indicate a close evolutionary relationship between the *O. niger* populations clade and *O. pallidicornis* clade. To confirm the evolutionary relationship between these two species, additional reference *cox*I sequences from *O. niger* are required in the phylogeny. Furthermore, the lineage of *O. albidipennis* clade is completely independent to other *Orius* species; Therefore, it is difficult to understand its evolutionary history. Therefore, the *Orius* species *cox*I phylogeny were extended to all available distinct *Orius* species *cox*I sequences in the NCBI database (Figure 3-3).

The *cox*I derived phylogeny grouped all specimens used in this study, as well as all available other *Orius* species in NCBI database, according to their predicted genealogy. In addition, all the available *cox*I sequences of *O. niger* from NCBI are included in this phylogeny and the *O. niger* clade is represented by two clearly defined groups, with the clade of *O. pallidicornis* as a sister-taxa, next to one of *O. niger* clades. Therefore, this dichotomous grouping in *O. niger* clades suggests that there are at least two sub-species of *O. niger* presented in current available NCBI database. It also confirmed a close evolutionary relationship between *O. pallidicornis* and *O. niger*. They seem to share a last common ancestor in a recent speciation event that contributed to split their lineage to two different descents. In the clade of *O. albidipennis*, it is confirmed the uniqueness of this species because it forms a completely independent group in the phylogeny, without any closer evolutionary relationship with other available *Orius cox*I sequence.

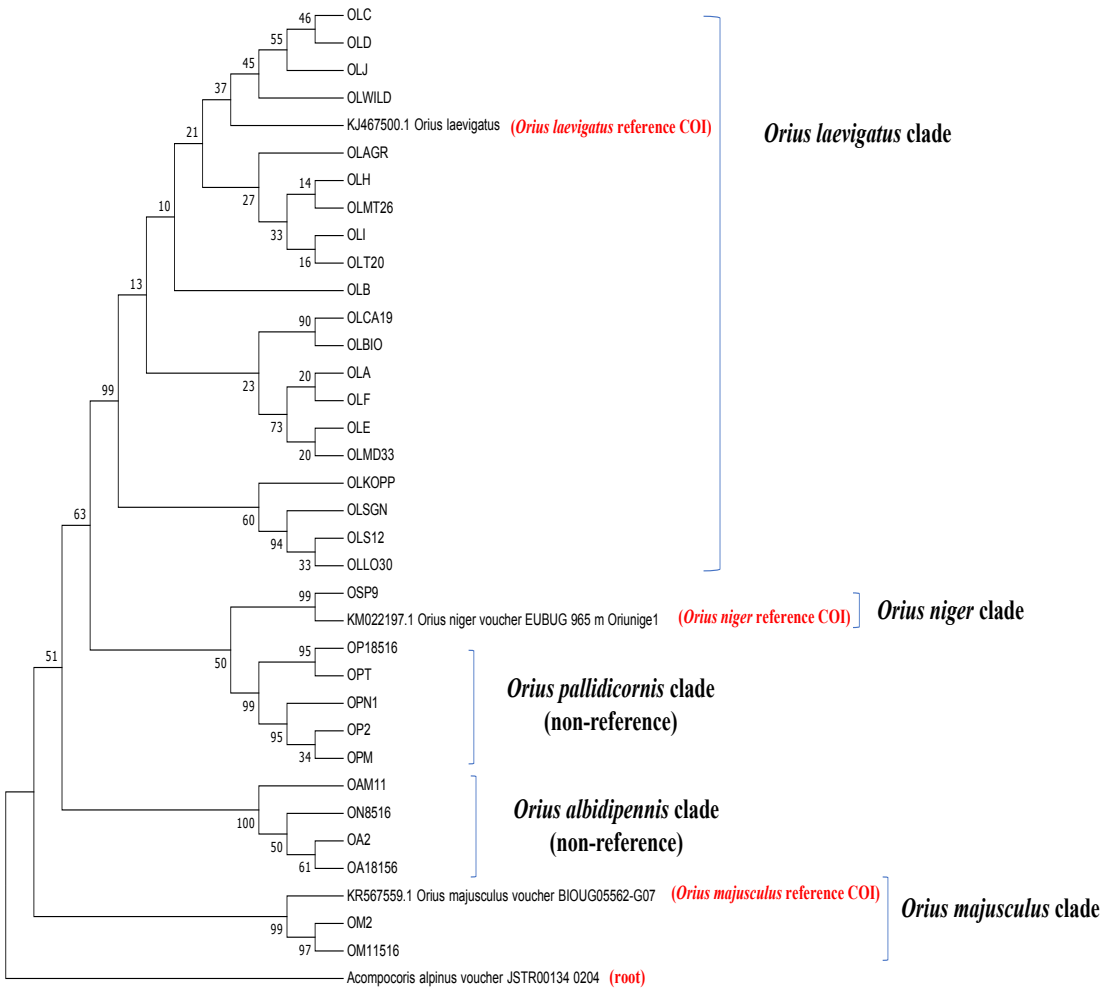


Figure 3-2: Project used *Orius* specimen's evolutionary history was inferred by using ML method based on the Tamura-Nei model.

(The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analysed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed)

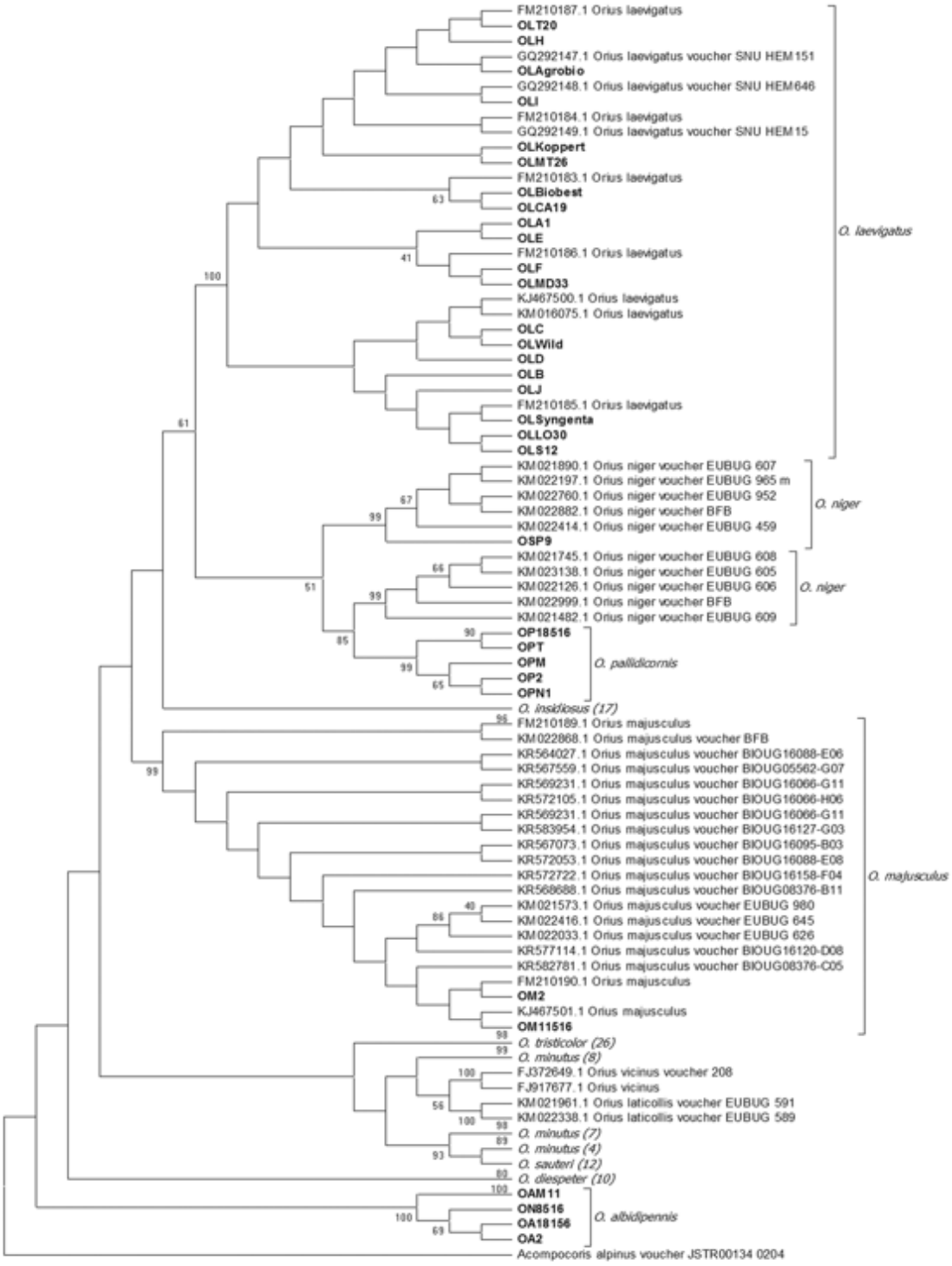


Figure 3-3: Evolutionary history of *Orius* samples inferred using all available CoxI sequences alignments and ML method (100 replicates, topology tree).

(Evolutionary distances were computed using Kimura 2-parameter method. Specimens from this study are shown in bold. *Acompocoris alpinus coxI* sequences was used as root. Only bootstrap values higher than 40% are shown in MEGA 7.)

3.5 Discussion and conclusion

This project used insect specimens from the five *Orius* species (*O. laevigatus, O. majusculus, O. niger, O. pallidicornis* and *O. albidipennis*), although most of the study focused on *O. laevigatus* specimens. These specimens were collected from different European countries, namely Spain, Italy, and Greece (Figure 3-1). All specimens were subject to amplification by universal *coxI* PCR. All *coxI* sequences were submitted for processing using BLAST to identify the most closely related species.

As expected, the sequences all exhibited a high degree of similarity to various *Orius* species. There are no available COI sequences in the NCBI database from *O. pallidicornis* and *O. albidipennis* at the time of writing. Taxonomic classification of all insect specimens used in the project was confirmed by *coxI* sequence phylogeny, using representative *coxI* sequences from each of the different *Orius* species available in the NCBI database.

Interestingly, the *O. niger* clade was separated into two sub-groups on the tree, suggesting that at least two sub-species of *O. niger* exist according to the current classification of *Orius* species. The *O. pallidicornis* clade distributed next to one of the two *O. niger* clades as a sister grouping. This suggests that *O. niger* and *O. pallidicornis* have a close evolutionary relationship, and that they may derive from a recent common ancestor. Evidence to show that *O. pallidicornis* could be a cryptic species within the *O. niger* species group would be important. One study in Germany used DNA barcode fragments of COI genes to identify 457 species of Heteroptera insects. The result of this study provided evidence for the putative existence of a cryptic species within *Orius niger.* Furthermore, it uses statistical parsimony to identify haplotype sharing between *O. niger* and other related species and reveals on-going hybridisation or recent speciation. Additionally, it can detect different lineages, so the putative existence of cryptic species of *O. niger* is supported (Raupach et al., 2014). However, a recent study of *O. niger* mentioned that *O. niger* and *O.sauteri* are same species, and *Ttraphleps aterrimus* is sister to *O. niger* and *O.sauteri* in their whole mitochondrial genome phylogenetic analysis. It indicates that the sequence of whole insect mitochondrial genomes of *Orius* species have more genetic information to identify their species rather than an amplicon sequence amplified by partial mitochondrial genes (Zhang et al., 2019).

The strict food requirements of *O. pallidicornis* can only be supported by pollen from *Ecballium elaterium* and mean that it cannot be reared in captivity, while *O. niger* is an omnivorous species. As a result, these two different species would not occur together, ruling out any laboratory-propagated contamination, and supporting the interpretation of two distinct, albeit closely related species.

Additionally, the clade of *O. albidipennis* distributes as an independent distant lineage from other *Orius* species, suggesting that in phylogenomic terms *O. albidipennis* is distantly related to the rest of the *Orius* species studied in this project. Initial molecular identification of Egyptian *O. albidipennis* was conducted using internal transcribed spacer 1 (ITS1) of ribosomal DNA and the result revealed that it was closely related to *O. sp*-Taif strain (Accessory number: HQ699724), both taxa belonging to the same species (Sayed et al., 2013). In the future, additional genetic markers could be used in further phylogenomic study of *Orius* species to illustrate the evolutionary relationship between the species and the timescale over which segregation of these species occurred, the discussion chapter will detailly discuss the future work about this chapter.

CHAPTER 4: Isolation of culturable bacteria from *Orius* specimen and genome assembly

4.1 Abstract in this chapter

• The isolation of microorganisms from *Orius* insect homogenates revealed three predominant bacterial colony morphologies across the whole range of insect specimens tested. According to the 16S rRNA classification, three of them are closely related to *Serratia*, *Leucobacter* and *Erwiniaceae*.

• The whole genome sequences of the isolates were assembled and annotated by SPAdes assembler.

• Representative genome sequences were used to design the primers for genome-specific PCR. Genome-specific PCR confirmed the presence of these bacteria which are true symbionts of their insect hosts.

4.2 Introduction

This chapter is aiming to explore the microbial community associated to several different species of *Orius* and create an initial view of the symbiotic community related to *Orius* genus. The first step is preparation of homogenates from *Orius* insects, culture bacteria from these homogenates, obtain the genome sequences and then perform whole genome sequence comparisons between different genomic assembly methods to choose the proper approach to assembly the genome sequences of these isolates, and further confirm the main symbionts isolated from *Orius* species mentioned in Chapter 3.

4.3 Method

All insect specimens were processed as described in Chapter 3, to isolate putative bacterial symbionts. Nutrient broth (NB) and Nutrient agar (NA) were used to culture bacteria from insect homogenates. The insect homogenates were subjected to serial dilutions (1/10,

1/100 and 1/1000 in NB) and plated on NA plates. From the initial sample set (October 2014), two predominant colony morphologies were detected on Nutrient agar plates after 2 days of incubation, a large white, regularly shaped colony, and a very small slow growing white colony. Several colonies of each type (classified as L=large and S=small) from each insect population, were re-streaked to ensure purity of the culture and 20% glycerol stocks were prepared for each isolate selected. The cultured abundance level of these bacterial colonies was not similar from any homogenate. In parallel, colony PCR was performed using 16S rRNA gene primers, and the PCR products were sequenced.

Additional insect specimens (June 2015) were sampled , and cultured, and additional colony morphologies were detected when grown on the NA plates (classified as LW= large and white, SP=small and pale, SW=small and white, LP=large and pale, I=irregular shape, NO=anaerobic condition, Y=Yellow, LY=light yellow, and DY=deep yellow), those samples were re-streaked twice to confirmed the purity of each population (Table 4-1), and used for total DNA extraction and genome sequencing (Illumina MiSeq) later.

Table 4-1: List of all isolate's colony morphology and host population.

Isolate name	Host population	Host species name	Colony morphology	Isolation collection date
OLAL2	A	*O. laevigatus*	large	2014 October
OLBL1	B	*O. laevigatus*	large	2014 October
OLCL1	C	*O. laevigatus*	large	2014 October
OLDL1	D	*O. laevigatus*	large	2014 October
OLEL1	E	*O. laevigatus*	large	2014 October
OLFL2	F	*O. laevigatus*	large	2014 October
OLHL2	H	*O. laevigatus*	large	2014 October
OLIL1	I	*O. laevigatus*	large	2014 October
OLJL1	J	*O. laevigatus*	large	2014 October
OLMTLW26	OLMT26	*O. laevigatus*	large and white	2015 June
OLLOLW30	OLLO30	*O. laevigatus*	large and white	2015 June
OPNLW1	OPN1	*O. pallidicornis*	large and white	2015 June
OPWLW2	OP2	*O. pallidicornis*	large and white	2014 October
OMLWL3	OM2	*O. mujusculus*	large and white	2014 October
OLFS4	F	*O. laevigatus*	small	2014 October
OLSSP12	OLS12	*O. laevigatus*	small and pale	2015 June
OLTSP20	OLT20	*O. laevigatus*	small and pale	2015 June
OLCASP19	OLCA19	*O. laevigatus*	small and pale	2015 June
OLMTSP26	OLMT26	*O. laevigatus*	small and pale	2015 June
OLMDSP33	OLMD33	*O. laevigatus*	small and pale	2015 June
OLMDLW33	OLMD33	*O. laevigatus*	large and white	2015 June
OAMSP11	OAM11	*O. albidipennis*	small and pale	2015 June
OPLPL6	OP2	*O. pallidicornis*	large and pale	2014 October
OLAS13	A	*O. laevigatus*	small	2014 October
OLCS4	C	*O. laevigatus*	small	2014 October
OLDS2	D	*O. laevigatus*	small	2014 October
OLES1	E	*O. laevigatus*	small	2014 October
OLIS6	I	*O. laevigatus*	small	2014 October
OLJS4	J	*O. laevigatus*	small	2014 October
OLTLW20	OLT20	*O. laevigatus*	large and white	2015 June
OAMLP11	OAM11	*O. albidipennis*	large and pale	2015 June
OAMSW11	OAM11	*O. albidipennis*	small and white	2015 June
OPYLY2-2014	OP2	*O. pallidicornis*	light yellow	2014 October
OPLPL6	OP2	*O. pallidicornis*	large and pale	2014 October
OPSWS1	OP2	*O. pallidicornis*	small and white	2014 October
OPLGL1-2014	OP2	*O. pallidicornis*	large and green	2014 October
OPNSW1	OPN1	*O. pallidicornis*	small and white	2015 June

OPNY1	OPN2	*O. pallidicornis*	yellow	2015 June
OALPL1	OA2	*O. albidipennis*	large and pale	2015 June
OAMY11	OAM11	*O. albidipennis*	yellow	2015 June
OAMSP11	OAM11	*O. albidipennis*	small and pale	2015 June
OAMLP11NO	OAM11	*O. albidipennis*	large and pale in anaerobic condition	2015 June
OAMI11	OAM11	*O. albidipennis*	Irregular	2015 June
OSPY9	OSP9	*O. niger*	yellow	2015 June
OSPSP9	OSP9	*O. niger*	small and pale	2015 June
OSPI9	OSP9	*O. niger*	Irregular	2015 June
OMLPL6	OM2	*O. mujusculus*	large and pale	2014 October
OMSS2-2014	OM2	*O. mujusculus*	small	2014 October
OLSI12	OLS12	*O. laevigatus*	Irregular	2015 June
OLSLY12	OLS12	*O. laevigatus*	light yellow	2015 June
OLSDY12	OLS12	*O. laevigatus*	deep yellow	2015 June
OLCALW19	OLCA19	*O. laevigatus*	large and white	2015 June
OLCAY19	OLCA19	*O. laevigatus*	yellow	2015 June
OLCAI19	OLCA19	*O. laevigatus*	Irregular	2015 June
OLTI20	OLT20	*O. laevigatus*	Irregular	2015 June
OLTLY20	OLT20	*O. laevigatus*	light yellow	2015 June
OLTDY20	OLT20	*O. laevigatus*	deep yellow	2015 June
OLMTSW26	OLMT26	*O. laevigatus*	small and white	2015 June
OLMTDY26	OLMT26	*O. laevigatus*	deep yellow	2015 June
OLMTLY26	OLMT26	*O. laevigatus*	light yellow	2015 June
OLMTLP26	OLMT26	*O. laevigatus*	large and pale	2015 June
OLMDLW33	OLMD33	*O. laevigatus*	large and white	2015 June
OLMDLY33	OLMD33	*O. laevigatus*	light yellow	2015 June

Initial taxonomic classification of these isolates was performed using a combinatorial approach of 16s rRNA colony amplification. Due to the limitation of 16s rRNA phylogeny classification, whole genome sequences of all the isolates were obtained and assembled (Figure 4-1). Initial and filtered reads quality control was carried out using FastQC (Andrews et al., 2015). *De novo* assembly of high- quality reads was performed with SPAdes (V3.5.0 assembler) using K-mers of 21, 33, 55, 77, 99, 127, and Velvet with a range of K-mers scale in careful mode utilizing reads error correction. After corrections of contigs, the contigs in the length longer than 1000bp were considered for downstream analysis. QUAST was used to assess assembly parameters such as number and size of contigs, total genome size N50 and overall GC content (Gurevich et al., 2013). Representative genome sequences from three predominant isolates were used to design genome specific PCR primers. It targeted to detect the presence of these isolates in insect specimens to confirm that these isolates are true symbionts of *Orius* specimens.

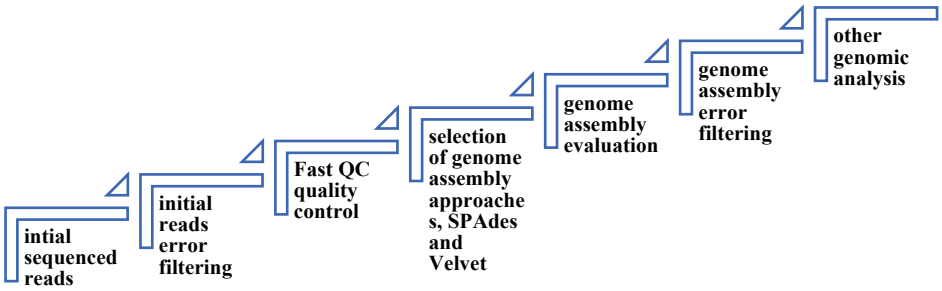


Figure 4-1: Workflow of the genome assembly of all isolates in this study.

4.4 Results

4.4.1 Initial classification of *Orius'* isolates

Initial taxonomic classification of the partial microbial isolates collected from October 2014 was performed using a combinatorial approach of 16S rRNA PCR amplification, clean up and Sanger sequencing. The isolates derived from insects collected in 2015 June were directly subjected to whole genome sequencing, the sequencing method has described in Chapter 2. Additionally, the information of all whole genome sequences of bacterial isolates were same as Table 4-1. Only the bacterial isolates cultured from A-J and OP2 populations of *Orius* specimens in the 2014 October insect collection were subjected to 16S rRNA gene PCR amplification and sequencing, followed by homology searches using BLAST and ML phylogeny (Figure 4-2). There are multiple clades presented in the ML phylogenic tree. The large colony and small colony isolates grouped together in *Serratia*-like clade and *Leucobacter*-like clade. Apart from OLJS4, OLEL1 and OLDL1 isolates, these three isolates were contaminated by bacteria plating errors, and they were re-cultured again for bacterial culture purity.

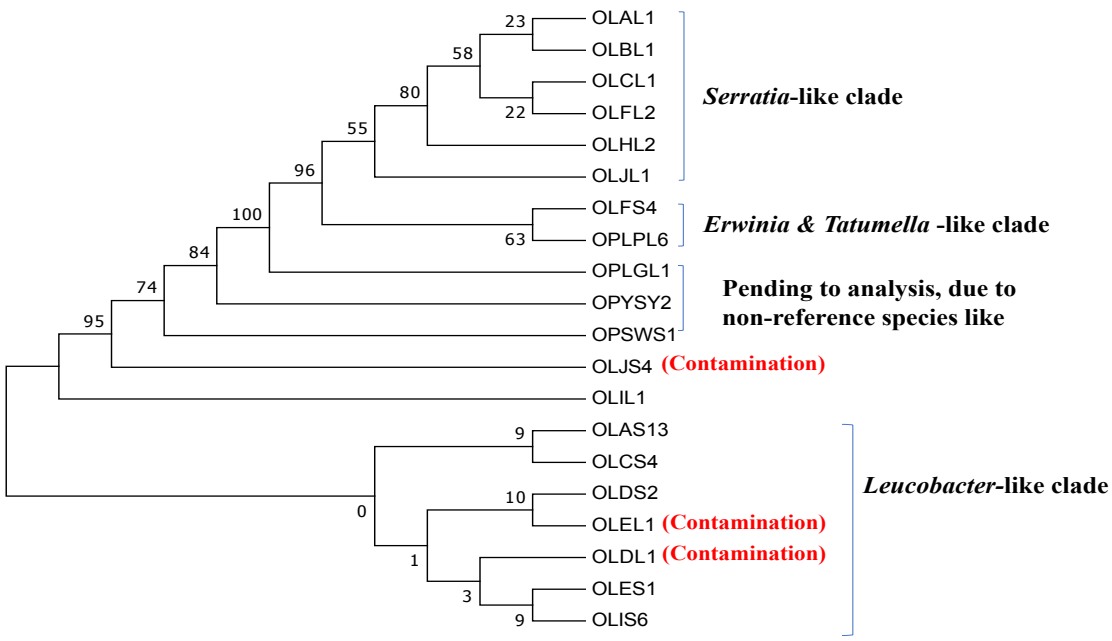


Figure 4-2: The phylogenetic analysis of 16s rRNA Sanger sequences of bacteria isolated from A to J and OP2 populations of *Orius* specimens.

(The evolutionary history was inferred by using ML method based on the Tamura-Nei model. The tree with the highest log likelihood (-989.79) is shown. Evolutionary analyses were conducted in MEGA7.)

Although there were a variety of isolates cultured from sample's homogenates, only three different morphological types of isolates can be cultured across the range of the insect samples. Therefore, only three predominant isolates were chosen to analyse by the following procedures. Sequence homology searches using BLASTn in the option of Nucleotide collection and 16S rRNA gene sequences from the predominant isolates revealed that the three predominant representative colonies were closely related to Actinobacteria (*Leucobacter* sp. *Orius*, 10 isolates) and Enterobacteriales (15 *Serratia* sp. *Orius* isolates, 8 *Erwinia* sp. *Orius* isolates and 1 related to Tatumella).

After initial identification of bacterial isolates derived from *Orius* species, the total bacterial DNA was isolated and whole genome sequenced by Illumina MiSeq platform and detailed information mentioned in Chapter 2.Initially, all the paired-end raw reads were processed by QC report and the raw reads filtered by Trim Galore wrapper tool. SPAdes assembler was later used to assembly all the genomes across the range of sample tested.

4.4.2 Quality control analyses of sequence raw reads for whole genome assembly

According to the FastQC report on the reads, the qualities of these sequenced reads were mainly determined on the parts of per base sequence quality and per sequence quality scores in FastQC report. The graph of per base sequence quality in FastQC report is an overview of the range of quality values across all bases at each position in the FastQ file. The graph of per sequence quality scores plots the average quality score over the full length of all reads on the x-axis and gives the total number of reads with this score on the y-axis.

As an example of this analysis, Figures 4-3 show the results of the per base sequence quality of sample OLAL2 initial and filtered sequence reads, respectively. The graphs indicated the range and average of sequence quality across the raw and after filtered reads of OLAL2. The Y-axis of the graphs is horizontally divided to three portions where the green colour section shows the highest quality of base calls, the orange part represents reasonably good quality of the calls, and the red part shows poorness of reads. For Illumina sequence reads, a Phred score of 30 means an error rate of 1 base in 1000, or an accuracy of 99.9%, while a Phred score of 40 indicates an error rate of 1 base in 10,000, or an accuracy of 99.99%. The top level of quality score typically averages 33 to 35, if the average quality of all the reads is over 30, the quality of sequence is good. On all the per base sequence quality graphs, the blue line on the plot means the mean value of overall base quality score, yellow box defined the inter-quartile range (25- 75%), and the central red lines in yellow boxes are median value. The upper and lower error bars represent the 10% and 90% points on the graphs (Andrews, 2010).

The percentage of poor reads of OLAL2 initial reads are higher than the good quality of reads (Figure 4-3). After filtering the poor quality of bases from 3’ end of reads and length filtering, also trimming the sequence adaptor and overall reads quality, the per base sequence score of OLAL2 is improved which indicates average quality of reads is over 30 quality score after filtering the initial reads.

Per sequence quality scores of OLAL2 initial and quality-filtered sequence reads are shown in Figures 4-4. The Y-axis on both figures indicates the number of sequences, and the X-axis show Phred scores according to a logarithmic scale. A poor quality of sequences will represent an error rate of 0.2 or higher which shows the Phred score lower than 27. If

Phred score is lower than 20, the sequence quality will be the poorest and fail to do further analysis. In the initial OLAL2 sequence reads, the Phred score is higher than 30 on average quality per reads. After quality filtered, OLAL2 sequence quality improve to 36 on Phred scores, which is very good quality of overall sequences.

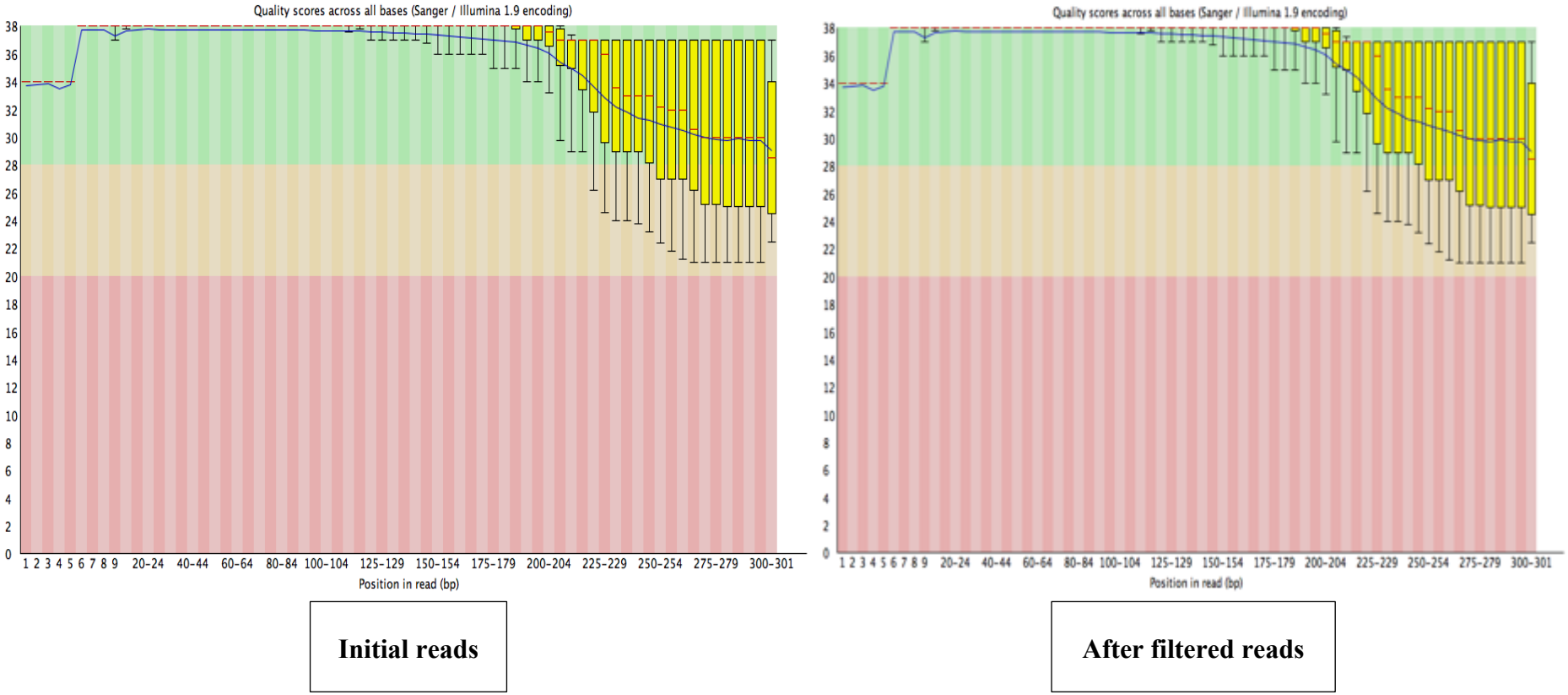
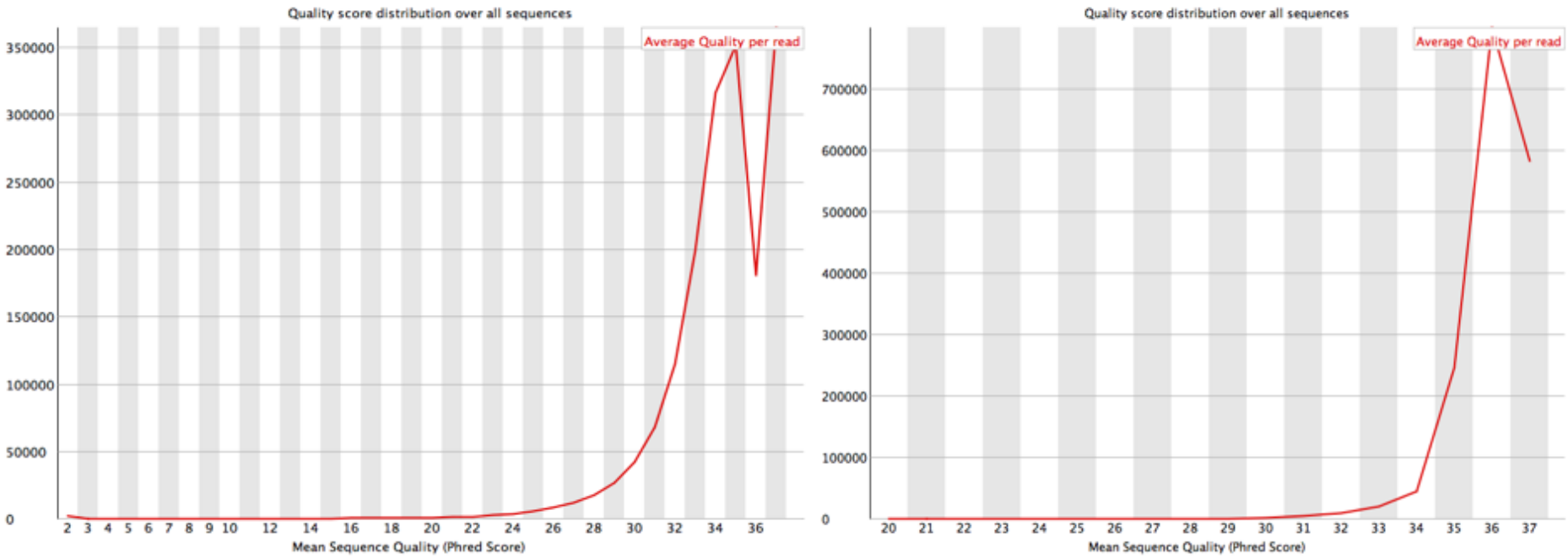


Figure 4-3: The comparisons of per base sequence quality of OLAL2 raw sequence reads and filtered sequence reads



Initial reads

After filtered reads

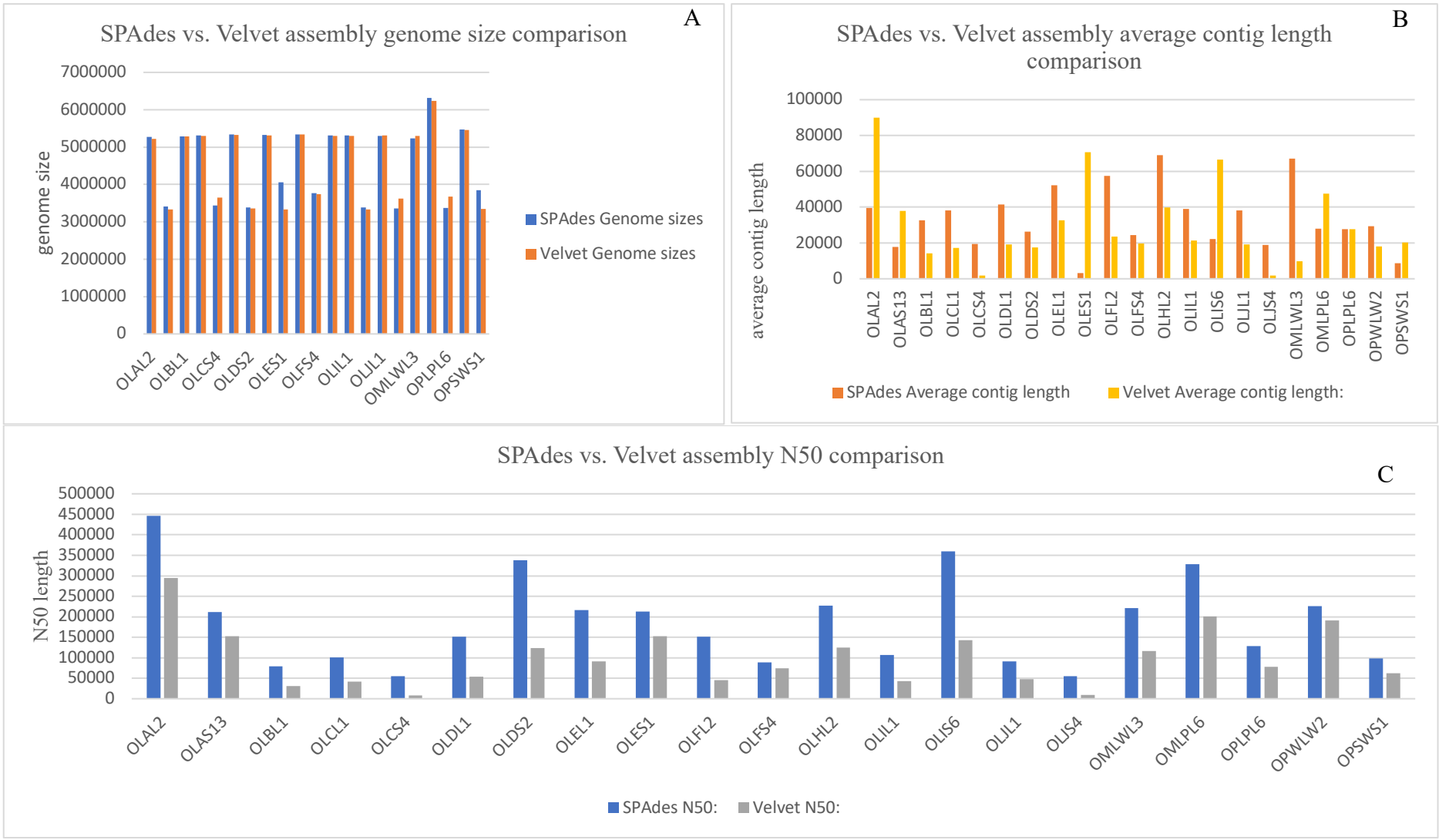
Figure 4-4: The comparison of per sequence quality scores of OLAL2 raw sequence reads and filtered sequence reads.

4.4.3 Draft genome assembly comparisons of all *Orius* bacterial isolates

A comparison of *O. laevigatus* potential symbiotic bacteria from A to J and OP2 populations, based on the basic statistic information of the assemblies between SPAdes and Velvet assembly approaches (Table 4-2 Galaxy server initially and Figure 4-5), shows that all the genome sizes in SPAdes assembly method are slightly larger than those genomes assembled with Velvet (Figure 4-5A). In addition, OLES1 was sequenced with poor quality, so the quality of assembly was very poor with large number of contigs and it was neglect to use due to the poor assembly quality. Additionally, the average contig length of the most SPAdes assembled genomes is significantly higher than Velvet (Figure 4-5B). Moreover, the proportion of SPAdes assembled contigs larger than 200 bp is higher than for Velvet assembled contigs (Figure 4-5D). SPAdes N50 values of these genomes were frequently larger than velvet (Figure 4-5C). N50 is the length for which the collection of all contigs of that length or longer covers at least half the assembly. Therefore, filtered SPAdes assembled genomes were chosen to do further analysis because these assemblies contain fewer contigs and larger genome size than Velvet assembly. It basically can give more genetic information of these isolates rather than using velvet assembly. Furthermore, it is important to optimise assembly quality because the quality of genome assemblies are most likely to affect further analysis of each genome such as the quality of genome annotation, GI predictions, and pangenome construction. In the future, more various assemblers will used to assembly the genome sequences and choose better assemblier to assemble the sequences. Alternatively, choosing different sequencing method to improve the sequencing quality before the assembly.

Table 4-2: The statistic information of bacterial genome assemblies isolated from populations of Orius specimens using by SPAdes assembly and Velvet assembly

Samples Name	SPAdes Genome Sizes	Velvet Genome Sizes	SPAdes N. Contigs	Velvet N. Contigs	SPAdes Average Contig Length	Velvet Average Contig Length:	SPAdes N. Contigs >= 200 bp:	Velvet N. Contigs >= 200 bp:	SPAdes N. Contigs >= 2,000 bp:	Velvet N. Contigs >= 2,000 bp:	SPAdes N50:	Velvet N50:
OLAL2	5274530 bp	5225339 bp	133	58	39658	90092	132 (99.2 %)	53 (91.4 %)	23 (17.3 %)	30 (51.7 %)	445886	294332
OLAS13	3409448 bp	3336722 bp	193	88	17666	37917	192 (99.5 %)	80 (90.9 %)	21 (10.9 %)	35 (39.8 %)	212055	152607
OLBL1	5290882 bp	5285382 bp	162	374	32660	14132	160 (98.8 %)	341 (91.2 %)	114 (70.4 %)	259 (69.3 %)	79489	30477
OLCL1	5313488 bp	5296422 bp	139	309	38227	17141	139 (100.0 %)	284 (91.9 %)	84 (60.4 %)	202 (65.4 %)	100736	41538
OLCS4	3433663 bp	3651271 bp	176	2061	19509	1772	176 (100.0 %)	1243 (60.3 %)	91 (51.7 %)	386 (18.7 %)	54529	8358
OLDL1	5336217 bp	5319822 bp	129	279	41366	19067	129 (100.0 %)	236 (84.6 %)	57 (44.2 %)	150 (53.8 %)	151517	54122
OLDS2	3379326 bp	3353712 bp	129	193	26196	17377	129 (100.0 %)	145 (75.1 %)	21 (16.3 %)	68 (35.2 %)	337459	123827
OLEL1	5325900 bp	5306985 bp	102	163	52215	32558	102 (100.0 %)	158 (96.9 %)	46 (45.1 %)	102 (62.6 %)	216784	91529
OLES1	4057764 bp	3327574 bp	1314	47	3088	70799	1313 (99.9 %)	45 (95.7 %)	19 (1.4 %)	35 (74.5 %)	212160	153099
OLFL2	5338242 bp	5333270 bp	93	226	57400	23599	93 (100.0 %)	218 (96.5 %)	67 (72.0 %)	172 (76.1 %)	151517	45112
OLFS4	3766039 bp	3742501 bp	155	191	24297	19594	155 (100.0 %)	152 (79.6 %)	74 (47.7 %)	91 (47.6 %)	88274	74850
OLHL2	5310903 bp	5302459 bp	77	133	68973	39868	77 (100.0 %)	128 (96.2 %)	44 (57.1 %)	79 (59.4 %)	226792	124478
OLIL1	5310395 bp	5296250 bp	136	249	39047	21270	135 (99.3 %)	247 (99.2 %)	92 (67.6 %)	180 (72.3 %)	106186	43003
OLIS6	3389765 bp	3326502 bp	153	50	22155	66530	153 (100.0 %)	50 (100.0 %)	19 (12.4 %)	41 (82.0 %)	359559	142436
OLJL1	5302898 bp	5307540 bp	139	278	38150	19092	138 (99.3 %)	248 (89.2 %)	91 (65.5 %)	170 (61.2 %)	91313	47244
OLJS4	3356222 bp	3626027 bp	177	1979	18962	1832	177 (100.0 %)	1174 (59.3 %)	110 (62.1 %)	389 (19.7 %)	54520	9093
OMLWL3	5231497 bp	5297745 bp	78	534	67070	9921	77 (98.7 %)	350 (65.5 %)	41 (52.6 %)	83 (15.5 %)	221016	116780
OMLPL6	6313714 bp	6241671 bp	226	131	27937	47646	225 (99.6 %)	120 (91.6 %)	45 (19.9 %)	63 (48.1 %)	328095	200548
OPLPL6	3372564 bp	3674795 bp	122	132	27644	27839	120 (98.4 %)	130 (98.5 %)	42 (34.4 %)	89 (67.4 %)	128515	77907
OPWLW2	5474171 bp	5452261 bp	187	303	29274	17994	185 (98.9 %)	269 (88.8 %)	53 (28.3 %)	76 (25.1 %)	226107	191331
OPSWS1	3852139 bp	3346247 bp	445	165	8656	20280	442 (99.3 %)	158 (95.8 %)	72 (16.2 %)	85 (51.5 %)	98947	62634


A
SPAdes vs. Velvet assembly genome size comparison
genome size
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
OLAL2
OLBL1
OLCS4
OLDS2
OLES1
OLFS4
OLIL1
OLJL1
OMLWL3
OPLPL6
OPSWS1
SPAdes Genome sizes
Velvet Genome sizes
B
SPAdes vs. Velvet assembly average contig length comparison
average contig length
100000
80000
60000
40000
20000
0
OLAL2
OLAS13
OLBL1
OLCL1
OLCS4
OLDL1
OLDS2
OLEL1
OLES1
OLFL2
OLFS4
OLHL2
OLIL1
OLIS6
OLJL1
OLJS4
OMLWL3
OMLPL6
OPLPL6
OPWLW2
OPSWS1
SPAdes Average contig length
Velvet Average contig length:
SPAdes vs. Velvet assembly N50 comparison
C
N50 length
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
OLAL2
OLAS13
OLBL1
OLCL1
OLCS4
OLDL1
OLDS2
OLEL1
OLES1
OLFL2
OLFS4
OLHL2
OLIL1
OLIS6
OLJL1
OLJS4
OMLWL3
OMLPL6
OPLPL6
OPWLW2
OPSWS1
SPAdes N50:
Velvet N50:

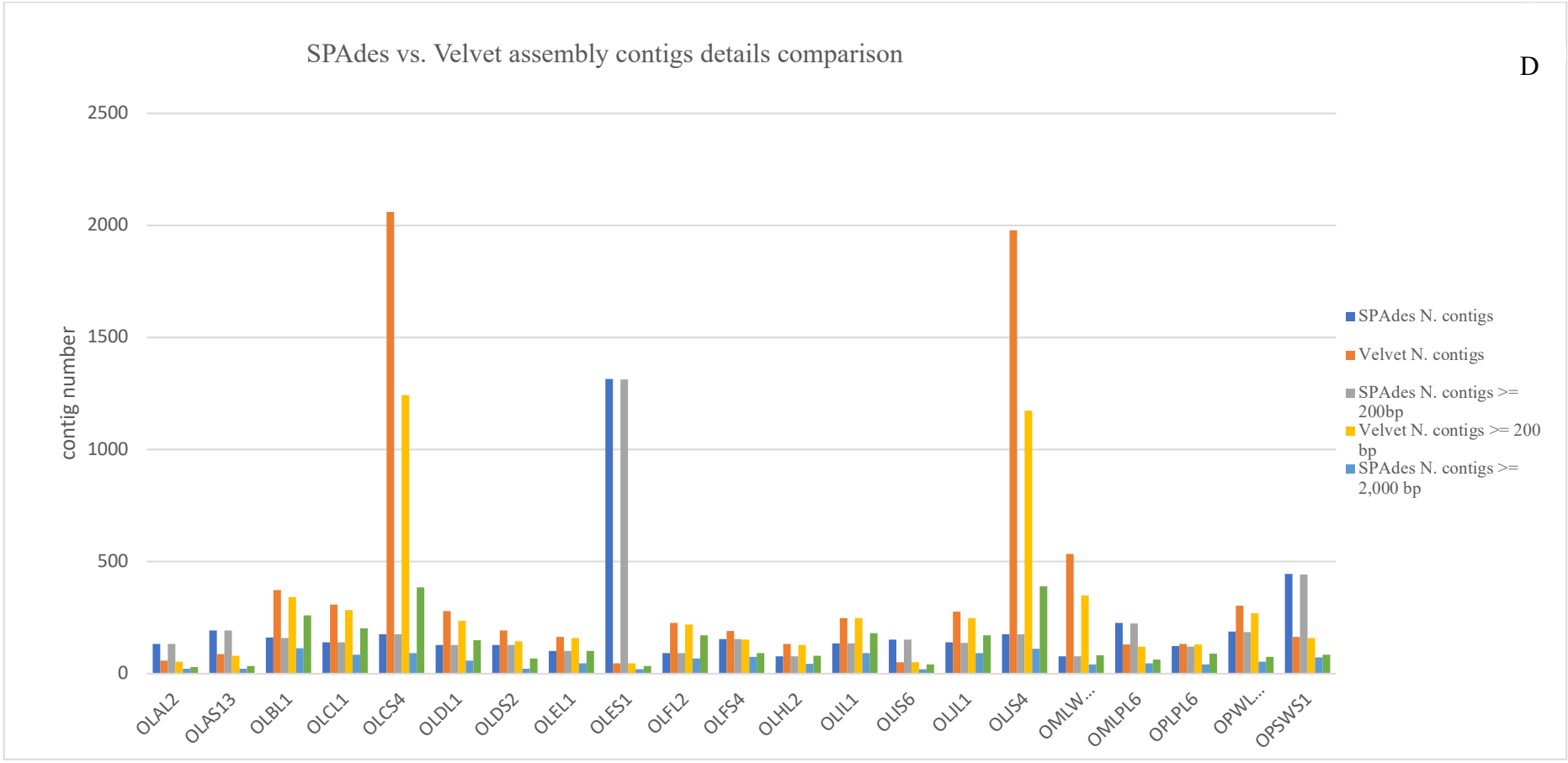


Figure 4-5: The comparisons of SPAdes and Velvet assembly.

(A part is the comparison of after assembly genome size between SPAdes and Velvet. B part is the comparison of average contig length between SPAde and Velvet assembly. C part is the comparison of N50 between SPAdes and Velvet. D part is the comparisons between SPAdes and Velvet in the total numbers of contigs, and the number of contigs in single contig length larger than 200bp and 2000bp.)

Table 4-3: A genome assembly final results in genome size and number of contigs in QUAST.

Assembly	Number of contigs (>= 0 bp)	Number of contigs (>= 1000 bp)	Total length of base in contigs (>= 1000 bp)	Total length(bp)	GC (%)	N50
OLAL2	31	26	5222816	5226333	59.18	445886
OLAS13	26	22	3327383	3330152	70.49	216619
OLBL1	143	128	5271561	5282384	59.53	79489
OLCL1	105	97	5291285	5296967	59.48	106650
OLCS4	97	94	3393143	3395490	70.49	54529
OLDL1	72	64	5302568	5308254	59.47	151517
OLDS2	24	21	3326717	3328964	70.5	337459
OLEL1	59	53	5300104	5304362	59.42	216784
OLES1	376	49	3361481	3597532	69.65	381168
OLFL2	82	78	5329974	5332885	59.37	151517
OLFS4	92	82	3728094	3736053	51.86	88274
OLHL2	54	50	5296848	5299433	59.45	226792
OLIL1	111	104	5293434	5298474	59.47	109387
OLIS6	22	19	3326135	3328333	70.49	380684
OLJL1	109	102	5284683	5290386	59.48	91313
OLJS4	117	112	3323196	3327083	70.47	54520
OMLWL3	51	46	5215646	5219380	59.2	221016
OMLPL6	68	51	6225404	6237447	60.11	328095
OPWLW2	85	65	5409512	5424191	59.29	226107
OPSWS1	115	86	3674220	3695425	64.68	100555
OPLPL6	63	61	3343345	3345108	47.55	127622
OLGL4-2014	238	35	2704879	2763280	33.47	247306
OLCALW19	162	25	3329425	3356810	70.43	264133
OAMLP11	410	27	3343666	3461247	70.1	212278
OPNLW1	132	36	5306864	5324074	59.38	277893
OLMTSW26	197	36	2920863	2960431	33.12	327436
OAMSW11	553	37	3340030	3492626	70.08	212093
OSPLW9	148	38	5318231	5341629	59.36	278043
OLTLW20	271	39	3343269	3390491	70.24	237733
OPYLY2-2014	180	40	2858065	2884756	32.56	136329
OLLOSP30	188	43	5314283	5343885	59.36	243240
OLMTLW26	205	55	5312985	5349589	59.38	222720
OLSLY12	135	59	2755426	2768009	33.04	90006
OLTI20	275	60	2698871	2736756	33.22	93337
OAMY11	120	78	6515909	6520875	39.81	140440
OLLOLW30	202	81	5309859	5331920	59.39	131128
OSPSP9	220	82	3738733	3766343	51.87	88223
OLCAY19	396	82	2701021	2764000	33.46	64307
OLMDSP33	196	83	3693919	3721966	51.91	97347
OLCASP19	149	84	3727627	3737139	51.87	81144
OAMSP11	298	86	3737623	3788585	51.91	97347
OLMTSP26	311	87	3739578	3797925	51.92	88223
OLSSP12	164	89	3701221	3714401	51.9	83938
OLMDLY33	236	92	2692335	2720460	33.15	52591
OSPY9	486	92	6922049	6970275	63.38	167637
OMSS2-2014	1035	93	6253397	6567656	59.25	158043
OLTSP20	186	106	3678473	3694453	51.92	68465
OLMTLY26	263	122	2748868	2778961	33.09	38510
OLCAI19	395	157	4739211	4794162	70.73	51537
OLMTLP26	381	165	6544504	6590142	66.57	104806
OAMI11	381	178	6705115	6739281	66.53	83798
OPNSW1	448	198	3660389	3710045	64.61	32910
OPLGL1-2014	345	206	3728279	3759317	38.72	35598
OLMDLW33	304	209	5036971	5073004	56.55	41620
OLTLY20	479	223	2741843	2795903	33.27	24269
OLMTDY26	372	233	4834884	4878045	36.85	46729
OSPI9	418	253	5918560	5954035	35.05	76258
OLTDY20	453	320	4812495	4859447	36.91	32044
OAMLP11NO	1701	516	6702003	6876424	66.45	27043
OLSDY12	715	567	4776656	4841289	36.99	15026
OPNY1	1803	841	1744769	2277425	32.6	1706
OLSI12	1945	1196	4374262	4692303	70.11	4391
OALPL2	56	51	6215869	6219609	60.13	228649

After using SPAdes to assemble all the bacterial genomes from *Orius* specimens, these genomes were filtered again for sequence adaptors removal. All the genomic statistic information was performed in QUAST (Table 4-3). The parameters of QUAST are slightly different than Galaxy online server, but the statistic status of these genomes is better to present in QUAST, as it can provide more details about the genomes, such as GC content (Table 4-3).

All the draft genomes were submitted to RAST annotation online server. The SPAdes assembled and RAST annotated information of the whole genome sequences from the three predominant isolate types (Table 4-4) that the average genome sizes for *Serratia* sp. *Orius*, *Leucobacter* sp. *Orius* and *Erwini*a sp. *Orius* strains were approximate 5.3, 3.7 and 3.6 Mb, respectively. The average GC content was 59.4%, 51.9% and 70.0% respectively.

Table 4-4: Summary of draft genome sequence features

Genome name	Average genome size (bp)	Genome sizes range	Average GC content	GC content range	Average number of CDs	Range of number of CDs
Serratia* sp. *Orius	5315187	5.2Mb-5.4Mb	59.4	59.3-59.5	4959	4937-5062
Erwinia* sp. *Orius	3730646	3.6Mb-3.7Mb	51.9	51.9	3832	3775-3895
Leucobacter* sp. *Orius	3654077	2.7Mb-3.3Mb	70.0	66.5-70.5	2999	2445-3105

Additionally, those genomes have been submitted to NCBI for annotation with accession numbers (Table 4-5). Since NCBI required genome annotation using Prokaryote Genome Annotation Pipeline (PGAP; Tatusova et al, 2016), all the genomes were annotated again by PGAP and added genome annotation to NCBI database in 2017.

Table 4-5: List of bacterial isolates and initial taxonomic classification. Genebank accession numbers are provided.

Isolate name	Source insect specimen	16s RNA gene-based classification (genus)	Genebank accession number
OLFS4	F	*Erwinia*	NZ_MOMB00000000.1
OAMSP11	OAM11	*Erwinia*	NZ_MTCI00000000.1
OLCASP19	OLCA19	*Erwinia*	NZ_MNKW00000000.1
OLMDSP33	OLMD33	*Erwinia*	NZ_MNKY00000000.1
OLMDLW33	OLMD33	*Erwinia*	NZ_MTCH00000000.1
OLMTSP26	OLMT26	*Erwinia*	NZ_MNKX00000000.1
OLSSP12	OLS12	*Erwinia*	NZ_MNCH00000000.1
OLTSP20	OLT20	*Erwinia*	NZ_MOMA00000000.1
OLAS13	A	*Leucobacter*	NZ_MRAS00000000.1
OLCS4	C	*Leucobacter*	NZ_MRAS00000000.1
OLDS2	D	*Leucobacter*	NZ_MRAT00000000.1
OLES1	E	*Leucobacter*	NZ_MRAU00000000.1
OLIS6	I	*Leucobacter*	NZ_MRAV00000000.1
OLJS4	J	*Leucobacter*	NZ_MRAW00000000.1
OAMSW11	OAM11	*Leucobacter*	NZ_MTCK00000000.1
OAMLP11	OAM11	*Leucobacter*	NZ_MTCJ00000000.1
OLCALW19	OLCA19	*Leucobacter*	NZ_MPIM00000000.1
OLTLW20	OLT20	*Leucobacter*	NZ_MRAQ00000000.1
OLAL2	A	*Serratia*	NZ_MSTL00000000.1
OLBL1	B	*Serratia*	NZ_MORD00000000.1
OLCL1	C	*Serratia*	NZ_MORE00000000.1
OLDL1	D	*Serratia*	NZ_MORF00000000.1
OLEL1	E	*Serratia*	NZ_MORG00000000.1
OLFL2	F	*Serratia*	NZ_MORH00000000.1
OLHL2	H	*Serratia*	NZ_MORI00000000.1
OLIL2	I	*Serratia*	NZ_MOWN00000000.1
OLJL1	J	*Serratia*	NZ_MOWO00000000.1
OLLOLW30	OLLO30	*Serratia*	NZ_MKYT00000000.1
OLMTLW26	OLMT26	*Serratia*	NZ_MNBD00000000.1
OMLWL3	OM2	*Serratia*	NZ_MSTK00000000.1
OPWLW2	OP2	*Serratia*	NZ_MTCF00000000.1
OPNLW1	OPN1	*Serratia*	NZ_MTCE00000000.1
OSPLW9	OSP9	*Serratia*	NZ_MSTM00000000.1
OPLPL6	OP2	*Tatumella*	NZ_MTCG00000000.1

4.4.4 Genome specific PCR for amplification of *Orius* sp. facultative symbionts in total DNA from the host.

The selected genome-specific ORFs from above representative genome sequences (Table 4-6) were used as template to design genome-specific PCR primers. These primers were used to detect the presence of the corresponding target sequences in total DNA extracted from all insect specimens used as source of isolates. Since all insect species tested were lab-reared, total insect DNA was also extracted from specimens collected in the field or acquired from commercial sources, but not propagated in the lab, except the specimens from *O. pallidicornis*. Genome-specific PCR amplification detected the presence of *Serratia* sp. *Orius* isolates in insect specimens of all available lab-reared, agriculture field collected and non-lab reared insect specimens in 2017 Swansea university laboratory of Institute of Life Science 1 Building, Singleton park campus.

Additionally, *Serratia* sp. *Orius* isolates were not detected by culturing techniques in some *Orius* specimens, and it may cause by the lower abundance level of *Serratia* sp. *Orius* isolates in some insect homogenates. *F. occidentalis* total DNA was used as negative control in the PCR and no PCR product was amplified as expected. Another type of *Serratia* sp. *Orius* isolates (OLAL2 and OMLW3) was not detected in any of the insect DNA samples tested and therefore it was concluded that there was insufficient evidence to catalogue these isolates as true symbionts and were not further analysed (Figure 4-7A). Furthermore, OLAL2 and OMLW3 are similar with *Serratia marcescens* Db11 in Chapter 5 phylogenetic study, but genome specific PCR did not detect the amplicons of OLAL2 in all the *Orius* specimens' total DNA, these two strains are most likely to be contaminant to *Orius* specimens. Therefore, OLAL2 and OMLW3 will not analysis further for this study.

Similarly, the presence of *Erwinia* sp. *Orius* isolates was confirmed across several *Orius* species in most of the specimens tested, including none of lab-reared hosts like *O. pallidicornis* (OPN1, OP18516). Additionally, the *Erwinia* sp. *Orius* isolates was only detected in lab-reared *O. albidipennis* and not detected in some commercial specimens (Figure 4-7B).

Attempts to detect the presence of the *Leucobacter* sp. *Orius* strains using a similar approach were successful in lab-reared specimens, but not from field collected or commercially purchased ones (results not shown), questioning the symbiotic association of these isolates with *Orius* sp. This finding, together with the limited number of genome sequences from the same species, led to the decision of not performing further analyses on these strains until their symbiotic nature is confirmed by future studies.

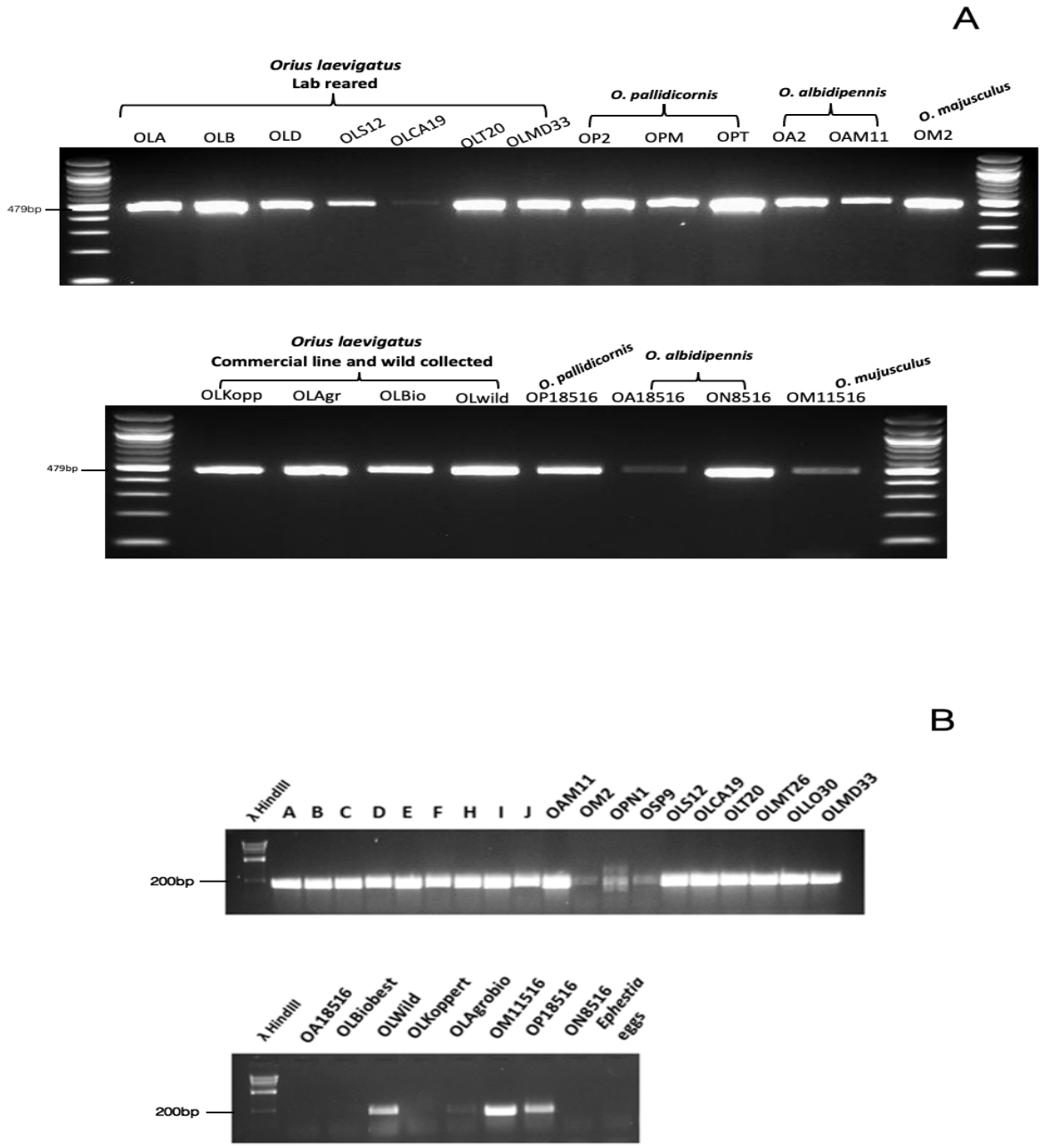


Figure 4-7: Genome-specific PCR for amplification of *Serratia* sp. *Orius* and *Erwinia* sp. *Orius* isolates DNA in total DNA of all insect specimens revealed that the (A) *Serratia* and (B)*Erwinia* are associated to all *Orius* specimens under study, *F. occidentalis* total DNA was used as negative control in the PCR of *Serratia* and no PCR product was amplified as expected. Negative control of (B)*Erwinia* is *Ephestia* eggs for both PCR, the DNA ladders for both PCR product were λHindIII DNA ladder.

4.5 Discussion and conclusion

The isolation of culturable bacteria from homogenates of *Orius* species insects revealed three different predominant bacterial colony morphologies across the whole range of insect specimens tested. Representative colony types from each insect population were initially used to amplify and classify the bacteria by 16S rRNA genetic analysis. The results of 16S rRNA phylogenomic analysis illustrate that two main types of bacteria, closely related to *Erwiniaceae* and *Microbacteriaceae*, respectively, are most likely to be two new species. However, no more information could be obtained from the 16S rRNA genes of these predominant bacterial species, which could be used to classify them and identify how different they are more accurately. Therefore, the entire genome sequences of these isolates were assembled and annotated using different approaches.

Based on the comparison between SPAde and Velvet genome assembly results, SPAde was preferable to Velvet assembly. This is mainly because SPAde assembly can automatically correct errors caused by k-mer-ed fastqs sequences, while Velvet can only assemble the genomes according to k-mer size without repairing any errors due to k-mer reduction. The approach of Velvet assembly mainly manipulates de Bruijn graphs through simplification and compression, without any loss of information, by merging non-intersecting paths into single nodes. It also recognises and removes three main types of errors: tips caused by errors at the edges of reads, “bubbles” due to internal read errors or to nearby tips connecting, and erroneous connections owing to cloning errors or to distant merging tips. Finally, it combines short reads and read pairs to generate contigs of reasonable length (Zerbino et al., 2008).

The SPAde assembly method initially constructed an assembly graph using multi-sized de Bruijn graphs, then performed new algorithms to remove tips, bubbles and chimeric reads. Next, pairs of k-mers (k-bimers) were adjusted to derive accurate distance estimates between k-mers in the genome and paths in the assembly graph. After adjustment of k-bimers, a paired assembly graph was constructed. Finally, contigs were produced (Bankevich et al., 2012). The biggest difference between Velvet and SPAde assembly is the improved construction of de Bruijn graph algorithms,

especially in iteration over values of k-mer sizes, and incorporation of k-bimers, which allows information from paired-end reads to be introduced into the computation at an earlier stage. Hence, SPAde assemblies were chosen to assemble all the bacterial genomes. However, genomes assembled by SPAde sometimes produce short and inaccurate contigs as well as combined sequence adaptors that can result in issues when using bioinformatic tools such as GGDC. Consequently, the SPAde contigs were also filtered to remove sequence adaptors and short base-pair reads. This part of the study allowed me to determine the best pipeline to process raw genome sequence data and apply it to all ongoing and future genome analyses.

After genome assembly, the representative genomes sequences were used to design the primers for genome-specific PCR. Since the predominant colony morphologies were not recovered in each insect population by culturing techniques, genome-specific PCR can confirm whether these isolates are true symbionts that are present in the DNA isolated from *Orius* sp. specimens. According to the PCR results, the presence of the *Serratia* sp. *Orius* isolates was confirmed in total DNA from representative insect specimens, so it confirmed the presence of *Serratia* sp. *Orius* and *Erwinia* sp. *Orius* isolates which are true symbionts in the insect hosts.

Due to this chapter only focus on the assembly of all bacterial isolates from all *Orius* specimens and the confirmation of presence of predominant isolates in *Orius* specimens, so the genome sizes of each predominant isolate were not expected to be a certain genome size yet, because they haven't classified accurately. Furthermore, according to Professor Ian Goodhead theory about genome reduction of endosymbiont states that when bacteria transfer from a free-living environment to a more specific (intracellular) niche, the initial state of bacteria having large genomes, some reduction in both size and coding capacity is seen as the organism adapts to an intracellular environment as the selective constraints on some genes are relaxed due to protection or nutrients offered by the host cell. Such genes are prone to 'pseudogenisation' by mutation. Further specialisation results in deletion of redundant sequences and a much-reduced genome size. The long-term maintenance of reduced coding capacities in recent symbionts may be as a result of some residual function (Goodhead and Darby, 2015). However, these isolates still can be cultured or growing on outside of host. It potentially indicates that these bacteria have not lose essential gene functions such as DNA replications. Therefore, they haven't tend to be 'pseudogenisation' by mutation yet.

CHAPTER 5: phylogenomic analysis of facultative symbionts from *Orius* sp. identified three putative new species

5.1 Abstract in this chapter

• In multilocus sequence analysis (MLSA) phylogenomic analysis, all the *Serratia* sp. *Orius* isolates were close to *Serratia* sp. SCBI, except OLAL2 & OMLWL3 which belong to another clade of *S. marcescens*, close to *S. marcescens* Db11.

• *Erwinia* sp. *Orius* and *Leucobacter* sp. *Orius* isolates were distributed into two novel monophyletic groups, which appear to be two new genera.

• The results of genome–genome distance calculation (GGDC) comparisons were similar with the results of MLSA phylogeny, confirming the accuracy of MLSA phylogenomic analysis.

5.2 Introduction

This chapter is aiming to identify and explain the taxonomic classification of previous assembled genomes by MLSA phylogenomic method. MLSA is currently a widely used method to obtain a higher resolution of the phylogenetic relationships of species within a genus or genera within a family and partial sequences of genes coding for proteins with conserved functions ('housekeeping genes') are used to generate phylogenetic trees and subsequently deduce phylogenies (Glaeser and Kämpfer, 2015). PhyloPhlAn is an integrated pipeline for large-scale phylogenetic profiling of genomes and metagenomes (Segata et al., 2013). PhyloPhlAn is an accurate, rapid, and easy-to-use method for large-scale microbial genome characterization and phylogenetic analysis at multiple levels of resolution. PhyloPhlAn can assign both genomes and metagenome-assembled genomes (MAGs) to species-level genome bins (SGBs). PhyloPhlAn can reconstruct strain-level phylogenies using clade-specific maximally informative phylogenetic markers and can also scale to very-large phylogenies comprising >17,000 microbial species (Segata et al., 2013). For the last 25 years species delimitation in prokaryotes (Archaea and Bacteria) was to a large extent based on

DNA-DNA hybridization (DDH), a tedious lab procedure designed in the early 1970s that served its purpose astonishingly well in the absence of deciphered genome sequences. With the rapid progress in genome sequencing time has come to directly use the now available and easy to generate genome sequences for delimitation of species. GBDP (Genome Blast Distance Phylogeny) infers genome-to-genome distances between pairs of entirely or partially sequenced genomes, a digital, highly reliable estimator for the relatedness of genomes. Its application as an in-silico replacement for DDH was recently introduced. Despite the high accuracy of GBDP-based DDH prediction, inferences from limited empirical data are always associated with a certain degree of uncertainty. It is thus crucial to enrich in-silico DDH replacements with confidence-interval estimation, enabling the user to statistically evaluate the outcomes. Such methodological advancements, easily accessible through the GGDC web service at http://ggdc.dsmz.de, are crucial steps towards a consistent and truly genome sequence-based classification of microorganisms. In earlier study of *Frankliniella occidentalis* symbionts (BFo1 and BFo2), concatenated MLSA phylogenies indicated that it may have shared a common ancestor to the Erwinia and Pantoea genera, and based on the clustering of rMLST genes, it was most closely related to *Pantoea ananatis* but represented a divergent lineage (Facey et al., 2015). In this study, MLSA phylogeny analysis first time used in all available genomes (access in 2017) from *Enterobacteriales* and *Actinobacteria* genus to identify the bacterial isolates from *Orius* specimen.

5.3 Method

The whole genome sequences of these isolates were obtained, assembled, and annotated as described on Chapter 4. A multi-locus Sequence Analysis (MLSA) phylogeny was used to accurately classify the cultured symbionts. GGDC analysis with the genome sequences from *Serratia*-like isolates using digital DNA: DNA hybridization confirmed that all genomes in this clade belong to the same species and will be referred as *Serratia* sp. *Orius* isolates. Since the value of similarity < 70% threshold in in Formula 2 shows the probability that reference genomes are same species as query genome, there were different values representing in the different *Serratia* species (Table 5-2).

GGDC comparison was also used to confirm the taxonomic classification of provided by the MLSA approach. Furthermore, using GGDC distance values of *Serratia* sp. *Orius* isolates and several typical *Serratia* species form a distance matrix made by DendroUPGMA (Garcia-Vallve et al, 1999) and represented in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Four hundred orthologous protein sequences from all Genebank FTP available at databases from *Enterobacteriales* and *Actinobacteria* genome sequences were retrieved, aligned, and used to create a concatenated sequence alignment using PhyloPhlAn. These alignments were used to construct a ML phylogenetic tree. It will improve the accuracy of taxonomic classification of different symbionts related to *Orius* species.

5.4 Results

5.4.1 MLSA Phylogenomic analysis of identified three putative *Orius sp.* facultative symbiotic new species

In the MLSA phylogeny (Figure 5-1), the 15 *Serratia* sp. *Orius* isolates distributed into two strongly supported clades, predominantly populated by *Serratia marcescens*. Several species within these clades are not classified as *S. marcescens*, suggesting that a phylogenomic-based revision of the taxonomic classification of this genus is due. We refer to these clades as "*S. marcescens* clade" and "*Serratia* SCBI clade" to simplify the discussion. 13 *Serratia* sp. *Orius* isolates distributed to a tight small clade together with 10 *Serratia* species within the "*Serratia* SCBI complex," while two isolates (OMLWL3 and OLAL2) grouped with the reference strain *S. marcescens* DB11 in a smaller clade (DB11 complex) within the major "*S. marcescens* complex." Since some of the organisms grouping within the SCBI complex are classified as *S. marcescens* sub-species, it suggested the revision of the *S. marcescens* taxonomic classification in some genomes. Since the above *Orius* sp. symbionts are close to bacteria from the *Serratia* genus, the importance of several representative *Serratia* species were summarized (Table 5-1), to support the observation of our isolates being true symbionts. There are at least 5 *Serratia* species closely associated with insect. Particularly, *Serratia* sp. SCBI from nematode is most similar with the *Orius Serratia*-like isolates (Figure 5-1). Additionally, *Serratia symbiotica str. 'Cinara cedri'* is one of Aphid's facultative endosymbiont to help the host gain nutrients and some essential

metabolism (Monnin et al., 2020). More than 10 *Serratia* species are shown to be associated to plants and the rest of them are either human pathogens or environmental free-living bacteria.

The genome sequences from the *Erwinia* sp. *Orius* and *Leucobacter* sp. *Orius* isolates were also analysed using PhyloPhlAn. According to the phylogeny (Figure 5-2A), only OLMDLW33 isolate grouped within the *Erwinia*/*Pantoea* to form a small clade with the *F. occidentalis* symbiont BFo1 (Facey et al., 2016). Additionally, OPLPL6 isolate from *O. pallidicornis* grouped with the other main *F. occidentalis* symbiont Bfo2. Further confirmation for the value of these distributions were analysed by GGDC comparison. Surprisingly, the remaining seven "*Erwinia* sp. *Orius*" isolates distributed to a novel monophyletic group within the *Erwiniaceae* (Figure 5-2A). It is most likely constituting a new genus pending to be properly classified. All *Leucobacter* sp. *Orius* isolates formed novel monophyletic clade within the *Microbacteriaceae*, as part of the *Leucobacter* clade (Figure 5-2B). Furthermore, GGDC comparisons also confirmed the *Leucobacter* sp. *Orius* isolates are not similar with any other existed *Leucobacter* species in NCBI database. It seems to be new genus of *Leucobacter*.

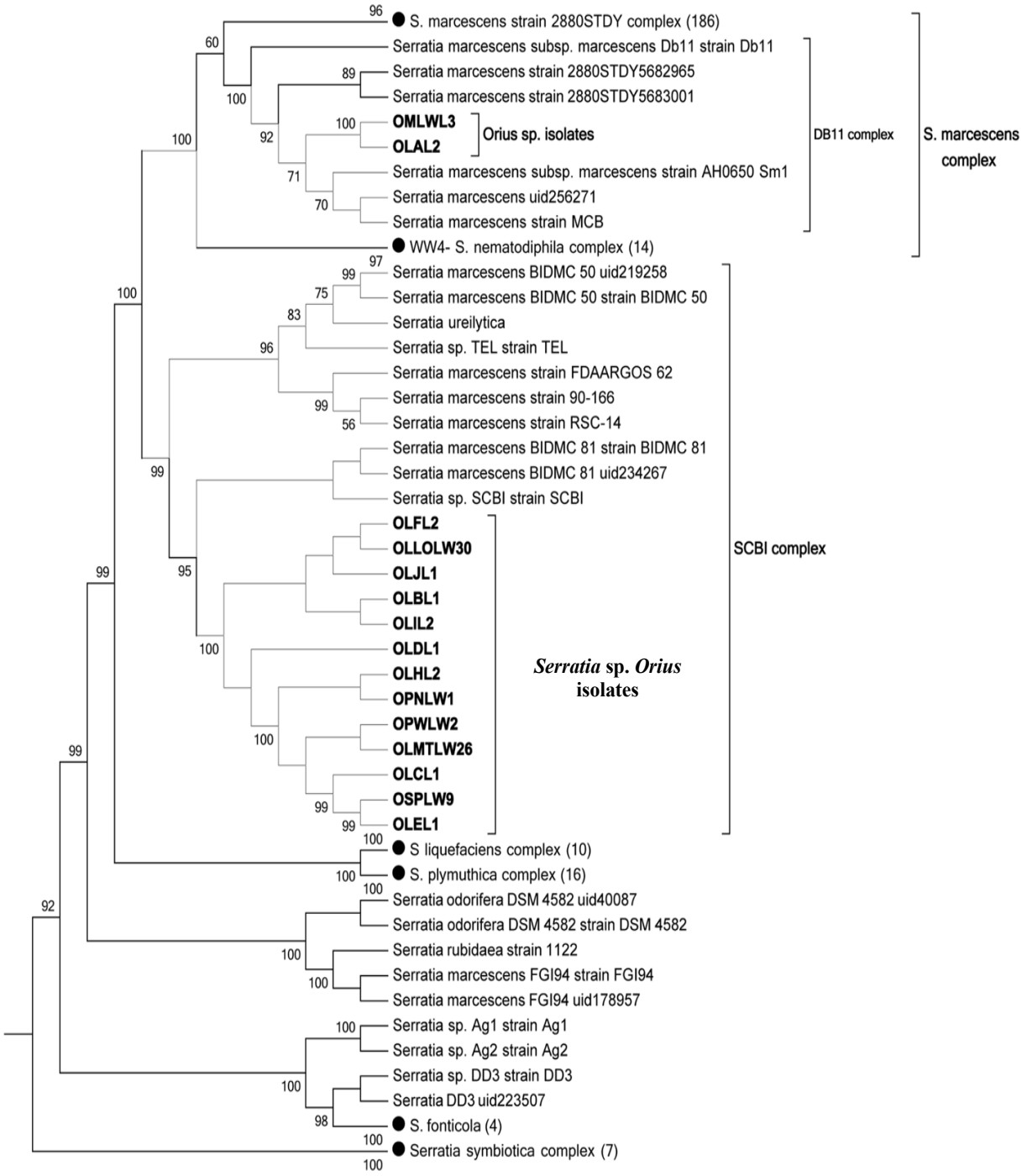


Figure 5-1: Multilocus phylogeny of *Serratia* sp. *Orius* isolated from *Orius* species.

(Topology bootstrap consensus tree (PhyloPhlAn) is shown, using collapsed nodes (black dots) for simplicity, with number of genomes per node indicated within brackets.)

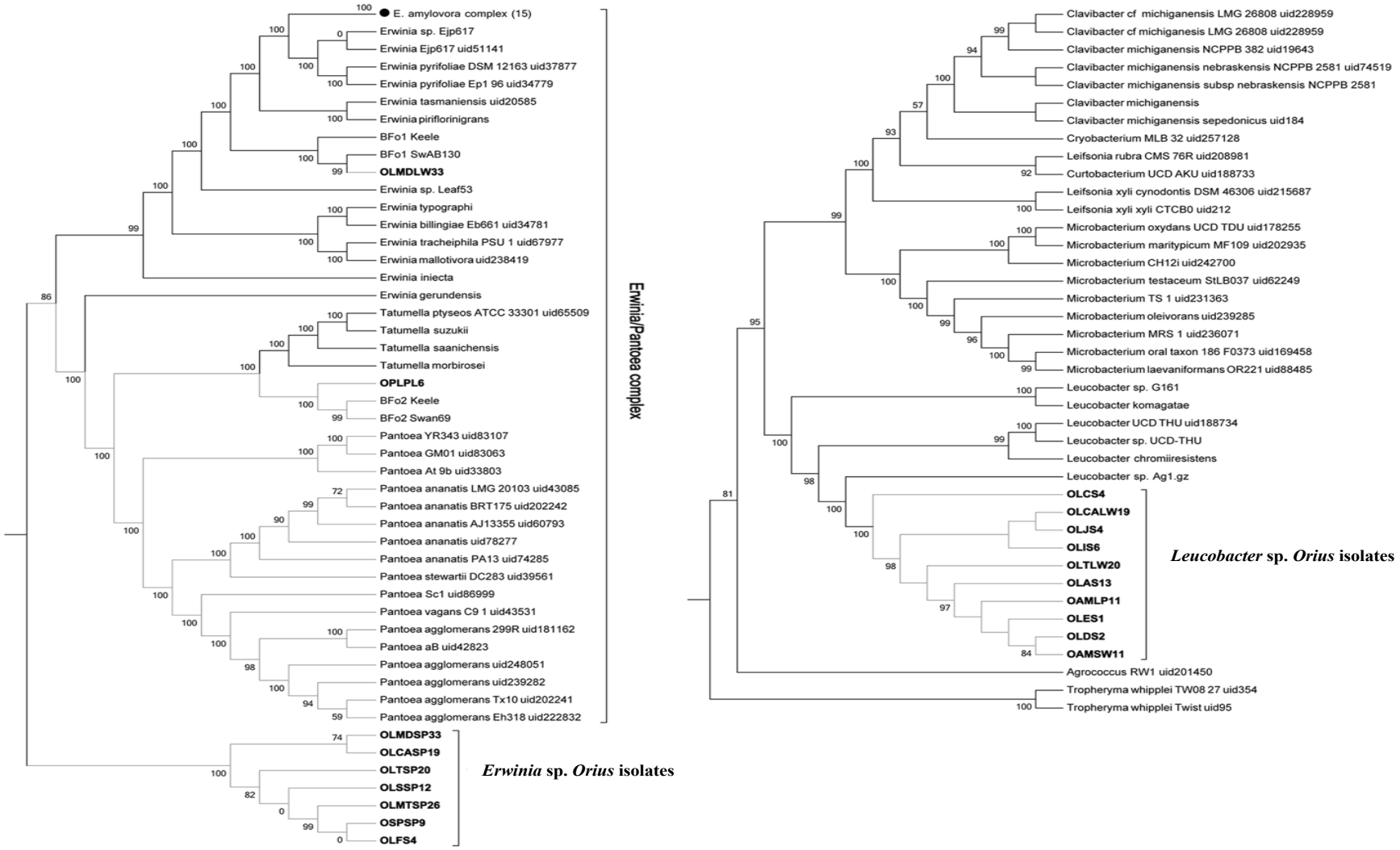


Figure 5-2: Multi-locus phylogenetic distribution of *Erwinia* sp. *Orius* (A) and *Leucobacter* sp. *Orius* (B) isolates.

(Obtained using PhyloPhlan. Collapsed nodes were used for simplicity, with number of genomes per node indicated within brackets.)

Table 5-1: The hosts and importance of representative *Serratia* species obtained from NCBI database.

Genome name	Host or collected resources	Importance of isolates	Biosample No.
***Serratia* sp. SCBI**	Host: Caenorhabditis briggsae, **(nematode)**	Agricultural, symbiotic study	SAMN02990345
***Serratia marcescens* FGI94**	Host: leaf-cutter ant, *Atta colombica* (**Insect**)	Inhibition of its symbiotic fungul garden	SAMN02604288
Serratia symbiotica str. 'Cinara cedri'	Host: giant conifer aphids, *Cinara cedri* **(Insect)**	Comparative genome analysis, symbiotic study	SAMN13545458
***Serratia* sp. Tel**	Host: entomopathogenic nematode, *Oscheius* species TEL-2014 **(nematode)**	Biological control agent against insect *galleria mellonella*	SAMN03701014
***Serratia marcescens* Db11**	Host: moribund fly, *Drosophila melanogaster* (**Insect**)	Pathogen, symbiotic study	SAMEA3138834
***Serratia plymuthica* 3rp8**	Host: Rapeseed, *Brassica napus* (**Plant**)	Biological control agent for seed coatings	SAMN03841799
***Serratia plymuthica* 3re4-18**	Host: Rapeseed, *Brassica napus* (**Plant**)	Biological control agent for seed coatings	SAMN03841798
***Serratia plymuthica* S13**	Host: Styrian pumpkin anthrosphere (**Plant**)	Biological control agent for seed coatings	SAMN02603297
***Serratia marcescens* strain 90-166**	Host: field-grown plant, **(Plant)**	Biological control agent against Rhizoctonia solani on cotton	SAMN03610504
***Serratia plymuthica* AS9**	Host: rape **(Plant)**	Comparative genome analysis	SAMN00713621
***Serratia sp.* AS12**	Host: rape **(Plant)**	Comparative genome analysis	SAMN00713623
***Serratia sp.* AS13**	Host: Rapeseed (**Plant)**	Comparative genome analysis	SAMN00713631
***Serratia sp*. FS14**	Host: Cāng zhú (Chinese herb), *Atractylodes macrocephala Koidz* (**Plant**)	Comparative genome analysis	SAMN03081466
***Serratia marcescens* RSC-14**	Host: European black nightshade, *Solanum nigrum* (**Plant**)	Agricultural effective microbe	SAMN04029108
***Serratia marcescens* B3R3**	Host: Corn, *Zea mays* (**Plant**)	Agricultural	SAMN04214975
***Serratia fonticola* GS2**	Host: Sesame (**Plant**)	Agricultural	SAMN04390144
***Serratia plymuthica* 4RX13**	Host: roots of a potato **(Plant)**	Agricultural	SAMN02603255
***Serratia marcescens* SM39**	Host: Patient	Pathogen	SAMD00061009
***Serratia marcescens* CAV1492**	Host: Patient	Pathogen	SAMN03733805
***Serratia marcescens* SmUNAM836**	Host: patient	Pathogen	SAMN03733572
***Serratia marcescens* U36365**	Host: Patient	Pathogen	SAMN04621337
***Serratia rubidaea* 1122**	Host: Patient	Pathogen	SAMN04481172
***Serratia liquefaciens* HUMV-21**	Host: Patient	Pathogen	SAMN03481691
***Serratia liquefaciens* FDAARGOS_125**	Host: Patient	Pathogen	SAMN03996271
***Serratia marcescens* BIDMC 50**	Host: Patient	Pathogen	SAMN02356590
***Serratia marcescens* BIDMC 81**	Host: Patient	Pathogen	SAMN02581400
***Serratia marcescens strain* FDAARGOS_62**	Host: Patient	Pathogen	SAMN02934511
***Serratia plymuthica* PRI-2C**	Collected resource: maize rhizosphere soil	Environmental	SAMN02470676
***Serratia* sp. YD25**	Collected resource: rhizosphere soil	Environmental	SAMN04226521
Serratia ureilytica	Collected resource: geothermal spring water	Environmental	SAMN18104335
***Serratia liquefaciens* ATCC 27592**	Collected resource: Environmental	Astrobiology	SAMN02604177
***Serratia fonticola* DSM 4576**	Collected resource: water	Environmental	SAMN03450772
***Serratia marcescens* WW4**	Collected resource: paper machine	Biofilm	SAMN02602965

5.4.2 GGDC analysis confirmed the similarity of symbiotic species in MLSA classification

Genome distance comparisons using as query genome of *Serratia* sp. SCBI to all available fully assembled genomes of bacteria classified as *S. marcescens* in the NCBI nucleotide database were conducted, revealing in some cases (e.g., *S. marcescens* DB11, *S. marcescens* SM39, *S. marcescens* FGI94) genome sequence distances that disprove their taxonomic classification as the same species. Additionally, the genome distances of OLAL2 and OMLWL3 were distinct to other *Serratia* sp. *Orius* isolates and other "SCBI complex" species, it indicates the species within 'SCBI complex' are different with the 'Db11 complex', there are two main different types of species closely associated with *S. marcescens*.

Based on the distribution derived from the GGDC distance values (Figure 5-3), only two isolates (OLAL2 & OMLWL3) associate with the clade of *Serratia marcescens* Db11, while the *Serratia* sp. *Orius* isolates distributed with the "*Serratia* SCBI complex," and it is similar with the result of PhyloPhlan phylogeny (Figure 5-1). Therefore, GGDC comparison of *Serratia* sp. *Orius* isolates confirmed the result of PhloPhlan phylogeny.

Table 5-2: One example of *Serratia* sp. *Orius* isolates in GGDC comparison in Formula 2, *Serratia* sp. SCBI compared with representative *Serratia* species and *Serratia* sp. *Orius* isolates.

Query genome	Reference genome	DDH	Distance	Prob. DDH >= 70%
***Serratia sp.* SCBI**	*Serratia marcescens* FGI94	26.7	0.162	0.02
***Serratia sp.* SCBI**	*Serratia marcescens* WW4	62.9	0.0468	61.39
***Serratia sp.* SCBI**	*Serratia marcescens* SM39	57.8	0.0556	44.79
***Serratia sp.* SCBI**	*Serratia marcescens* CAV1492	57.6	0.0559	44.17
***Serratia sp.* SCBI**	*Serratia marcescens* RSC-14	88.5	0.0138	95.26
***Serratia sp.* SCBI**	*Serratia marcescens* SmUNAM836	57.5	0.056	43.89
***Serratia sp.* SCBI**	*Serratia marcescens* B3R3	62.3	0.0478	59.56
***Serratia sp.* SCBI**	*Serratia marcescens* U36365	62.5	0.0474	60.37
***Serratia sp.* SCBI**	*Serratia plymuthica* PRI-2C	28.4	0.1515	0.05
***Serratia sp.* SCBI**	*Serratia rubidaea* 1122	26.7	0.162	0.02
***Serratia sp.* SCBI**	*Serratia liquefaciens* ATCC 27592	27.1	0.1592	0.03
***Serratia sp.* SCBI**	*Serratia fonticola* DSM 4576	24.9	0.1748	0.01
***Serratia sp.* SCBI**	*Serratia fonticola* GS2	24.7	0.1761	0.01
***Serratia sp.* SCBI**	*Serratia liquefaciens* HUMV-21	27.3	0.1584	0.03
***Serratia sp.* SCBI**	*Serratia plymuthica* 4Rx13	28.3	0.1518	0.05
***Serratia sp.* SCBI**	*Serratia symbiotica str. 'Cinara cedri'*	22.2	0.1973	0
***Serratia sp.* SCBI**	*Serratia plymuthica* AS9	28.2	0.1527	0.05
***Serratia sp.* SCBI**	*Serratia* sp. AS12	28.2	0.1527	0.05
***Serratia sp.* SCBI**	*Serratia liquefaciens* FDAARGOS_125	27.2	0.1587	0.03
***Serratia sp.* SCBI**	*Serratia* sp. AS13	28.2	0.1527	0.05
***Serratia sp.* SCBI**	*Serratia plymuthica* S13	28.4	0.1515	0.05
***Serratia sp.* SCBI**	*Serratia symbiotica* STs	34.3	0.121	0.53
***Serratia sp.* SCBI**	*Serratia plymuthica* 3Rp8	28.4	0.1514	0.05
***Serratia sp.* SCBI**	*Serratia* sp. FS14	64	0.045	64.65
***Serratia sp.* SCBI**	*Serratia plymuthica* 3Re4-18	28.4	0.1514	0.05
***Serratia sp.* SCBI**	*Serratia* sp. YD25	55.8	0.0593	37.8
***Serratia sp.* SCBI**	*Serratia* sp. SCBI	100	0	98.3
***Serratia sp.* SCBI**	*Serratia marcescens* BIDMC_50	15	0.2912	0
***Serratia sp.* SCBI**	*Serratia marcescens* BIDMC_81	14.5	0.301	0
***Serratia sp.* SCBI**	*Serratia ureilytica*	88.7	0.0136	95.32
***Serratia sp.* SCBI**	*Serratia marcescens strain* FDAARGOS_62	89.2	0.013	95.51
***Serratia sp.* SCBI**	*Serratia marcescens strain* 90-166	83.2	0.0196	92.76
***Serratia sp.* SCBI**	*Serratia sp.* TEL	90.6	0.0115	95.99
***Serratia sp.* SCBI**	filtered_OLCL1	88.9	0.0133	95.41
***Serratia sp.* SCBI**	filtered_OLDL1	88.9	0.0133	95.41
***Serratia sp.* SCBI**	filtered_OLEL1	88.9	0.0133	95.41
***Serratia sp.* SCBI**	filtered_OLFL2	89	0.0133	95.43
***Serratia sp.* SCBI**	filtered_OLHL2	88.9	0.0133	95.41
***Serratia sp.* SCBI**	filtered_OLIL2	89	0.0133	95.43
***Serratia sp.* SCBI**	filtered_OLJL1	89.1	0.0131	95.47
***Serratia sp.* SCBI**	filtered_OLLOLW30	88.9	0.0133	95.43
***Serratia sp.* SCBI**	filtered_OLMTLW26	88.9	0.0133	95.42
***Serratia sp.* SCBI**	filtered_OLBL1	89.1	0.0131	95.47
***Serratia sp.* SCBI**	filtered_OSPLW9	88.9	0.0133	95.42
***Serratia sp.* SCBI**	filtered_OPNLW1	88.9	0.0133	95.41
***Serratia sp.* SCBI**	filtered_OPWLW2	88.8	0.0134	95.38
***Serratia sp.* SCBI**	filtered_OLAL2	62.9	0.0468	61.33
***Serratia sp.* SCBI**	filtered_OMLWL3	62.9	0.0468	61.4
***Serratia sp.* SCBI**	*Serratia marcescens* Db11	63	0.0466	61.8

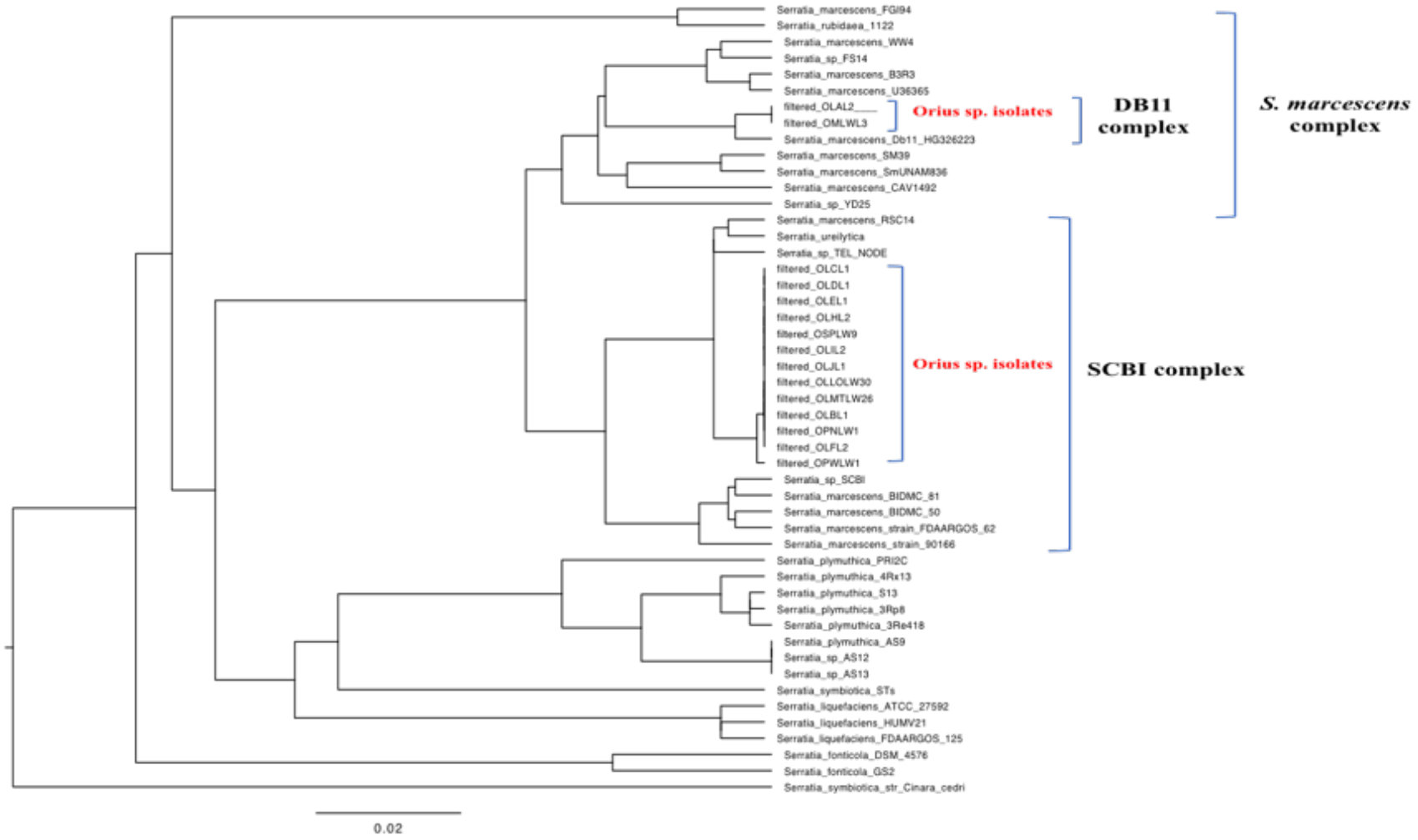


Figure 5-3: GGDC-based distribution of *Serratia* sp. *Orius* isolates and several representative *Serratia* species made by GGDC distance matrix using DendroUPGMA drawing the tree.

Since GGDC distance values of *Erwinia* sp. *Orius* isolates were too distinct from any of the available *Erwinia, Pantoea*, or *Tatumella* species genomes which it cannot form an accurate GGDC distance phylogeny, so the phylogeny of *Erwinia*-like isolates was not further explored. Further attempts with GGDC analyses failed to identify a close relative in publicly available genomes. The relationship of OLMDLW33 and Bfo1 (Figure 5-2A) was confirmed by GGDC comparisons, which showed that OLMDLW33 and BFo1 belong to the same species. The isolate OPLPL6 from *O. pallidicornis* is related to the other main *F. occidentalis* symbiont BFo2, although it displayed 68% probability by GGDC comparison of being the same species as BFo2 (i.e., similarity <70% threshold).

Genome distance comparisons estimated by GGDC revealed that none of the isolates have higher percentage of similarity to be considered the same species as the *Leucobacter* sp. genomes available at NCBI databases. Further exploration identified one additional *Leucobacter* sp. genome sequence, namely *Leucobacter* sp. AEAR, which was assembled from the raw genome sequences of the nematodes *Caenorhabditis angaria* and *Caenorhabditis remanei*, although never isolated and cultured (Percudani, 2013). When compared by GGDC, *Leucobacter* sp. AEAR genome was shown to be similar enough to be considered the same species to the *Leucobacter* sp. Orius isolates, and together they should be considered a new species within the *Leucobacter* genus. Despite this similarity, comparison of genome statistics revealed differences, mostly in terms of number of coding sequences (*Leucobacter sp*. AEAR: 2778, *Leucobacter* sp. *Orius*: ~3010 ± 189) that suggest different sub-species. Indeed, while GGDC produced a high probability score (82.34%, higher than the 70% threshold) confirming that these genomes belong to the same species, the probability of them being the same sub-species (33.31%) falls well below the defined threshold (>79%).

5.5 Discussion and conclusion

Assembled genomes from three types of predominant *Orius* sp. isolates and all available *Enterobacteriales* and *Actinobacteria* genome sequences were retrieved and concatenated sequence alignments using PhyloPhlAn to create MLSA phylogenies. All the *Serratia* sp. *Orius* isolates were closely related to *Serratia* sp. SCBI, except OLAL2 & OMLWL3 belong to another clade of *S. marcescens,* which were associated with *S. marcescens* Db11. *Erwinia* sp. *Orius* and *Leucobacter* sp. *Orius* isolates were distributed to two novel monophyletic groups,

it seems to be two new genera of bacteria species. Further confirmation of the three kinds of predominant *Orius sp.* isolates taxonomic classification was analysed using GGDC comparisons. The results of GGDC comparisons matched the results of MLSA phylogeny. Therefore, it confirmed the high accuracy of MLSA phylogenic analysis.

Erwinia sp. *Orius* and *Leucobacter* sp. *Orius* isolates were distributed to two novel monophyletic groups, likely to be two new genera of bacteria. However, GGDC comparisons failed to identify similar reference genomes that were sufficiently similar with consider the genomes as the same species, based on the relevant genomes in the clades of *Leucobacter* and *Erwinia* species in the MLSA phylogenomic study. The lack of available reference genomes prevented further genomic comparative studies of these two species.

As mentioned previously, *Leucobacter* species are likely to be the true symbionts of *Orius* species because related *Leucobacter* species have been isolated from various insect hosts, although it has not been possible to amplify their genes from the total DNA of *Orius* species. For instance, *Leucobacter* sp. Ag1 (BioSample: SAMN03481186) has been isolated from adult mosquito gut (*Anopheles gambiae*) as the facultative symbiont. *L. chironomi* strain MM2LBT (Halpern et al., 2009) was isolated from the eggs of biting midges (*Chironomus* sp.) and functions to protect the insect hosts from toxic metals and confer toleration of metals (Senderovich et al., 2013). *L. holotrichiae* sp. nov was isolated from the gut of scarab beetle (*Holotrichia oblita*) larvae (Zhu et al., 2016). These symbiotic associations of *Leucobacter* species with their insect hosts could be the important evidence supporting the feasibility of a facultative symbiotic relationship between *Leucobacter*-like isolates and *Orius* species. In the future, if more genomes of this species become available, pangenome analysis will identify the genes and gene functions necessary for the facultative symbiotic lifestyle.

CHAPTER 6: Genomic Islands (GIs) predictions differentiates lineages within *Serratia* sp. *Orius* facultative symbiont strains.

6.1 Abstract in this Chapter

- Since *Serratia* sp. *Orius* isolates were closely related to *Serratia sp.* SCBI as described in Chapter 5 (MLSA phylogeny), the differences of *Serratia* sp. *Orius* isolates were analysed by genomic island (GI) predictions using as reference genome *Serratia* sp. SCBI. Additionally, *Serratia* sp. *Orius* isolates closely related to *Serratia marcescens* Db11 (OLAL2 and OMLWL3) were confirmed that they are not true symbionts of *Orius* species by genomic specific PCR amplification in chapter 4 section 4.4.4, so both OLAL2 and OMLWL3 are not further analysis in following chapters.
- Since *Erwinia* sp. *Orius* and *Leucobacter* sp. *Orius* isolates are most likely to be new species without any close related species available on the NCBI database, these genomes were not analysed further due to the lack of suitable reference genomes.
- In the results, total 87 GIs with sizes ranging from 50 Kb to 262Kb, were predicted in all *Serratia* sp. *Orius* genomes from Island Viewer 3 using *Serratia* sp. SCBI as reference.
- There is a small correlation between GI numbers and length, and poor correlation between GI number and genome size.
- All 32 genome-specific GIs were found out by clustering in CD-Hit, and these representative GIs provide an indication of different lineages among strains.
- Multiple genome alignment revealed that genomic regions not shared between these isolates and the reference genome are most likely to be the locations of GIs in each genome.

6.2 Introduction

This chapter is targeted to predict GIs of the *Serratia* sp. *Orius* isolates to determine diversity and stability of these strains. Genetic variation of bacteria can be achieved through mutations, rearrangements, and horizontal gene transfers and recombination (Moon et al., 2016). Recently,

significant increasing numbers of bacterial genome sequences are being determined, and their comparisons have demonstrated that bacterial genomes are composed of a conserved “core gene pool” and strain-specific “flexible gene pools” (Ogura et al., 2008). The “core genes pool” encoding house-keeping functions such as essential metabolic activities, information processing, and bacterial structural and regulatory components. In contrast, the “flexible gene pools” are often carried on so-called “genomic islands (GIs)”, which have been acquired by Horizontal Gene Transfer (HGT). Most of them are also related to mobile genomic elements (MGEs) which are a heterogeneous group of genes, such as bacteriophages, plasmids, chromosomal cassettes, integrative conjugative elements, and transposons (Ogura et al., 2008). They usually contribute to adaptation and survival under certain environmental conditions by acquisition of genetic traits are beneficial to colonise at diverse ecological niches (Moon et al., 2016). GIs also carry multiple genes encoding a variety of biologically functional proteins that are closely associated with niche colonization, catabolism of diverse substrates, symbiotic relationships, resistance to antimicrobial agents, or enhanced virulence, antimicrobial resistance, and metabolic pathways (Ogier et al., 2010; Farrugia et al., 2015). Therefore, GIs are essential factor to drive bacterial evolution.

GIs are horizontally transferred DNA segments integrated into bacterial chromosomes (Boyd, Almagro-Moreno, and Parent, 2009). They are characterized by a G + C content, codon usage bias and dinucleotide frequencies, among other sequence signatures, which usually differs from those of the genome（Che, Hasan, and Chen, 2014）. Most of them are integrated at the 3′-end of tRNA and tmRNA genes, although distinct families of GIs can show preference for other genes as integration sites (Ambroset et al.,2016). GIs range in size from approximately 10 to 500 kbp (Piña-Iturbe et al., 2018). GIs or the genes on the GIs that are mobile element on a bacterial genome confer various benefits to the host bacterium, such as increased fitness in a specific host environment and increased pathogenicity. Therefore, the determination and functional analysis of GIs are essential steps toward a proper understanding of pivotal strain-specific features.

Many GIs contain an excision/ integration module that includes an integrase/recombinase gene which facilitate chromosome integration and excision (Hsiao et al., 2005). They usually come from the two classical protein superfamilies of the tyrosine recombinase and serine recombinase. These proteins can facilitate recombination reactions between Direct Repeated Sequences (DRS), also known as Left and Right attachment sites (attL and attR) at both ends

of GIs. Recombination is resulted in GIs excision from the chromosome and the consequent formation of a circular epitomal element that carries one copy of the DRS (attP), while another DRS (attB) remains in the host DNA. After excision the attB and attP sites can play as substrates for integrase-mediated recombination, resulting in the re-integration of the GI into the bacterial chromosome (Piña-Iturbe et al., 2018). In addition to this, excised islands can also be transferred to other hosts by utilization of co-resident prophages for high-frequency transduction inside their capsids or transferred by conjugation. It explains that that some GIs can replicate in their circular form, but that others cannot be able to replicate again in the circular form (Piña-Iturbe et al., 2018).

Importantly, GIs are vulnerable to the genes loss or gain during their transit from one bacterium to another. However, genes encoding the key functions of excision/integration, mobilisation and their regulation remain as a conserved core, as reported for different families of GIs such as the Mobilizable Genomic Islands and the SXT/R391 family of integrative and conjugative GIs present in different Gram-negative bacterial families, or the conjugative and MGEs recently found in streptococci, and the Phage-Inducible Chromosomal Islands (PICIs) of *Staphylococcus aureus* and other Gram-positive strains (Piña-Iturbe et al., 2018).

GI prediction is an ideal tool to assess genome plasticity and to segregate closely related strains. IslandViewer 3 (Dhillon et al., 2015) is a freely online server that incorporates three of the most accurate GI prediction methods: IslandPick, IslandPath-DIMOB and SIGI-HMM and they have different features. IslandPick uses a comparative genomics-based method to identify unique regions by comparing a user-specified genome against closely related genomes. Comparative genomic GI prediction methods may disclose genomic regions that are not present within related strains, suggesting that the region was horizontally transferred. Other tools (SIGI-HMM and IslandPath-DIMOB) detect regions with abnormal sequence composition and use sequence composition approaches to detect GIs. Sequence composition GI prediction methods are based on the fact that most bacterial genomes usually have differences in sequence structure, such as GC% and codon bias. If there is a region within a genome that has abnormal sequence composition, it could indicate that this genome is differentiated from another genome because of this region. Particularly, IslandPath-DIMOB identifies islands with dinucleotide bias and the presence of an associated mobility gene (integrases, transposases, etc.), but SIGI-HMM identifies codon usage bias with a hidden Markov model approach.

6.3 Method

In this study, IslandViewer 3 were used to analysis the GI content of the *Serratia* sp. *Orius* isolates genomes when compared to *Serratia* sp. SCBI as reference genome. After one *Serratia* sp. *Orius* isolates genome Genebank file is uploaded to IslandViewer 3, a basic circular genome visualizer allows users to view GIs, also with a table providing annotations of the genes in the predicted GI regions (supplementary data) and sequences from each GI. On the table, the positions and length of GIs were presented as island start, island end, and length respectively. It also represents Genebank locus names with annotated proteins with different gene positions. Additionally, Microsoft Office Excel was used to calculate the correlations between genome size, number of GI and GI length of each genome from the data of IslandViewer 3. Furthermore, chapter 2 section 2.6.4 mentioned method of GI clustering analysis. Additionally, PowerPoint were used for forming gene map of all GIs positions and genes coding in GIs of all 13 *Serratia* sp. *Orius* isolates genome and predominant GIs content in each *Serratia* sp. *Orius* isolates genome. Finally, Mauve alignments were performed to each *Serratia* sp. *Orius* isolates for differentiation of the lineages of these genomes, because GI prediction is a suitable tool to assess genome plasticity and to segregate closely related strains. Furthermore, Mauve 2.4.1 (Darling et al., 2004) was used to perform multiple genome comparisons to each *Serratia* sp. *Orius* isolates and *Serratia* sp. SCBI as reference. Initially, the contigs from each draft genome were reordered against the reference *Serratia* sp. SCBI using the 'move contigs' option in Mauve. After these draft genomes were reordered against the reference genome, these genomes were aligned. The results display as one horizonal panel per input genome sequence, a scale representing the sequence coordinates for that genome, and a single black horizontal centre line. Each coloured block area corresponds to a Locally Collinear Block (LCB), a region of the genome sequence that aligned to a part of another genome, and probably homologous and internally free from genomic rearrangement. The coloured area is higher where the similarity is high. Conversely, areas of low similarity are identified by larger white portions. Areas that are completely white within a Locally Collinear Block (LCB) are not aligned and probably contain sequence elements specific to a particular genome, like GIs. Additionally, the LCBs shown above the centre line of the aligned regions are the forward oriented direction against

reference genome. If several LCBs lie below the centre line, the regions are in a reverse complement orientation. The contigs from the rest of the draft genomes were also reordered by 'move contigs' in Mauve.

After reordering the contigs of draft genomes, multiple reordered draft genomes from *Serratia* sp. *Orius* isolates were aligned by 'Progressive Mauve' against the reference genome- *Serratia sp.* SCBI (Figure 6-4). This alignment method could improve the accuracy of aligning regions conserved in some genomes for better visualisation of multiple draft genomes alignments. The multiple genome alignments evidenced the presence of large homologous regions shared across the genomes compared. This similarity is represented by LCB length (>100bp) and number.

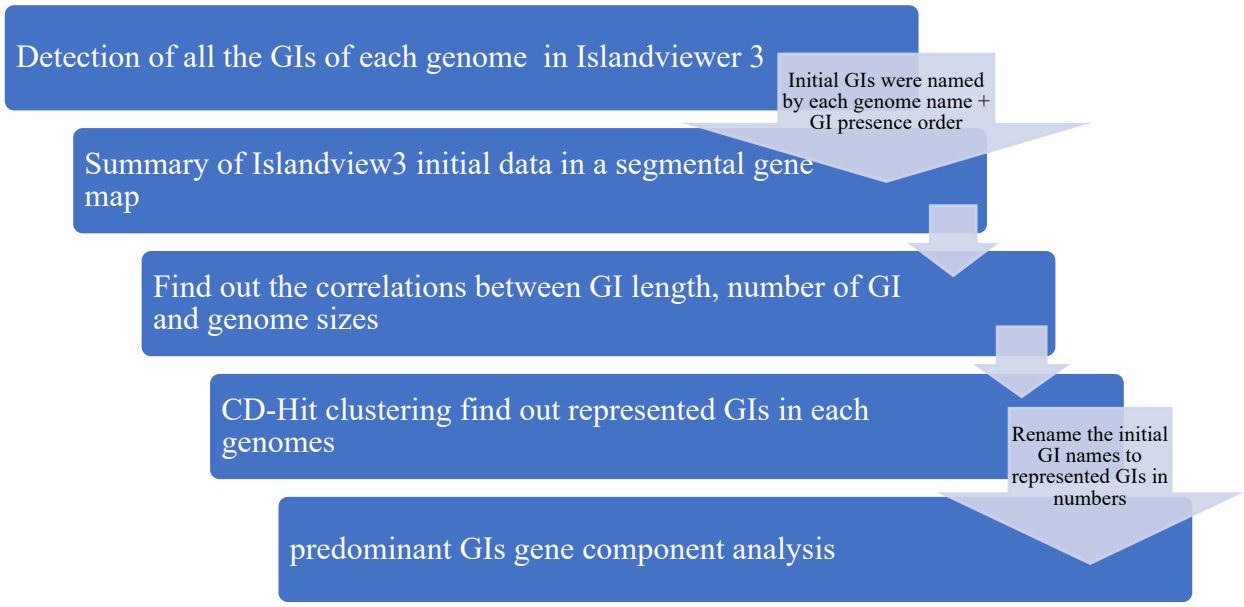


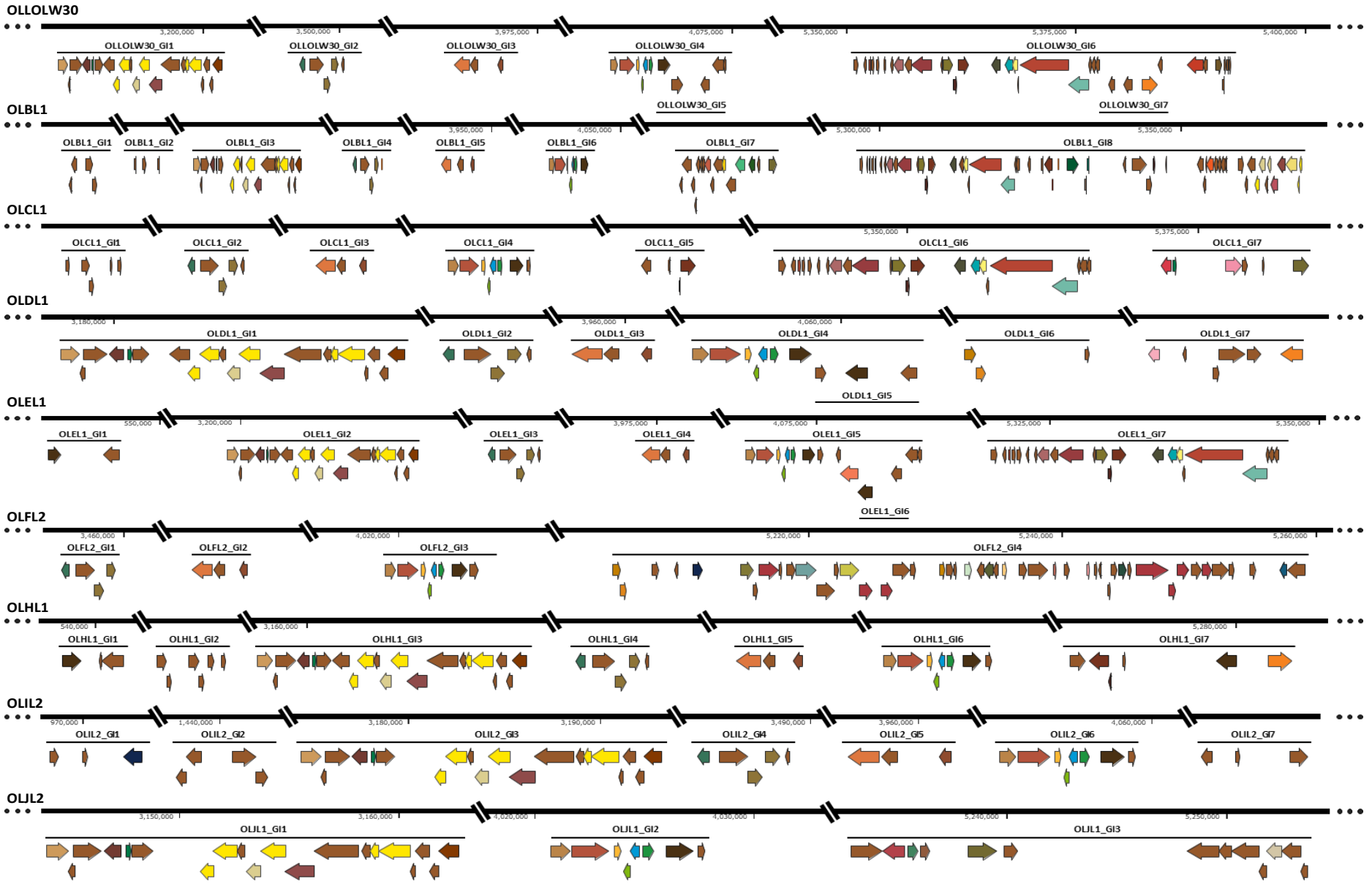
Figure 6-1: The process of GIs detections and analyses in each isolate.

6.4 Results

6.4.1 Prediction of GIs

Due to the redundancy of GIs initial data downloaded from IslandViewer 3, the segmental gene map of initial GIs gene components and their locations in each isolate's genome were illustrated in Figure 6-1. Interestingly, the number and length of GIs per genome are variable,

the numbers of GI per genome ranges from 3 GIs in OLJL1 to 11 GIs in OSPLW9 (Table 6-1). The segmental map of the 13 genomes is shown in Figure 6-1, with alien segments (color-coded, each representing a distinct cluster) (Figure 6-1). Each type of genomic island is polymorphic in gene content but several of them are conserved within strain lineages. Furthermore, there are over 150 different genes presented in these GI, they mainly annotated as conjugal transfer proteins, integrase related proteins, phage related proteins, multiple types of regulator proteins and type 4 and 5 secretion system related proteins. These proteins are important to GI formation and help to distinguish which type of GI they are belong to, based on their component proteins or genes in these GIs. The GI sequence annotation revealed a variety of different regulatory protein families and most of these proteins promote the lifestyles of bacteria under optimal conditions, such as TetR, LacI, AsnC, and LuxR family transcriptional regulator proteins. For instance, TetR family transcriptional regulator proteins are present in most GIs; members of this protein family usually control genes involved in multidrug resistance, enzymes regulated in various catabolic pathways, biosynthesis of antibiotics, osmotic stress, and pathogenicity of both gram-negative and gram-positive bacteria (Ramos et al., 2004). In some *Serratia* species such as *Serratia* sp. ATCC 39006, TetR family proteins modulate two secondary metabolic pathways of this species: synthesis of the pigment prodigiosin and carbapenem antibiotics (Gristwood et al., 2008). Several phage related genes present in GIs of *Serratia* sp. *Orius* and they might be prophages but they need further annotations using different annotation tool in the future and also to find out their function to *Orius* symbiosis in future work.


OLLOLW30
OLLOLW30_GI1
OLLOLW30_GI2
OLLOLW30_GI3
OLLOLW30_GI4
OLLOLW30_GI6
OLBL1
OLLOLW30_GI5
OLLOLW30_GI7
OLBL1_GI1
OLBL1_GI2
OLBL1_GI3
OLBL1_GI4
OLBL1_GI5
OLBL1_GI6
OLBL1_GI7
OLBL1_GI8
OLCL1
OLCL1_GI1
OLCL1_GI2
OLCL1_GI3
OLCL1_GI4
OLCL1_GI5
OLCL1_GI6
OLCL1_GI7
OLDL1
OLDL1_GI1
OLDL1_GI2
OLDL1_GI3
OLDL1_GI4
OLDL1_GI6
OLDL1_GI7
OLDL1_GI5
OLEL1
OLEL1_GI1
OLEL1_GI2
OLEL1_GI3
OLEL1_GI4
OLEL1_GI5
OLEL1_GI7
OLEL1_GI6
OLFL2
OLFL2_GI1
OLFL2_GI2
OLFL2_GI3
OLFL2_GI4
OLHL1
OLHL1_GI1
OLHL1_GI2
OLHL1_GI3
OLHL1_GI4
OLHL1_GI5
OLHL1_GI6
OLHL1_GI7
OLIL2
OLIL2_GI1
OLIL2_GI2
OLIL2_GI3
OLIL2_GI4
OLIL2_GI5
OLIL2_GI6
OLIL2_GI7
OLJL2
OLJL1_GI1
OLJL1_GI2
OLJL1_GI3

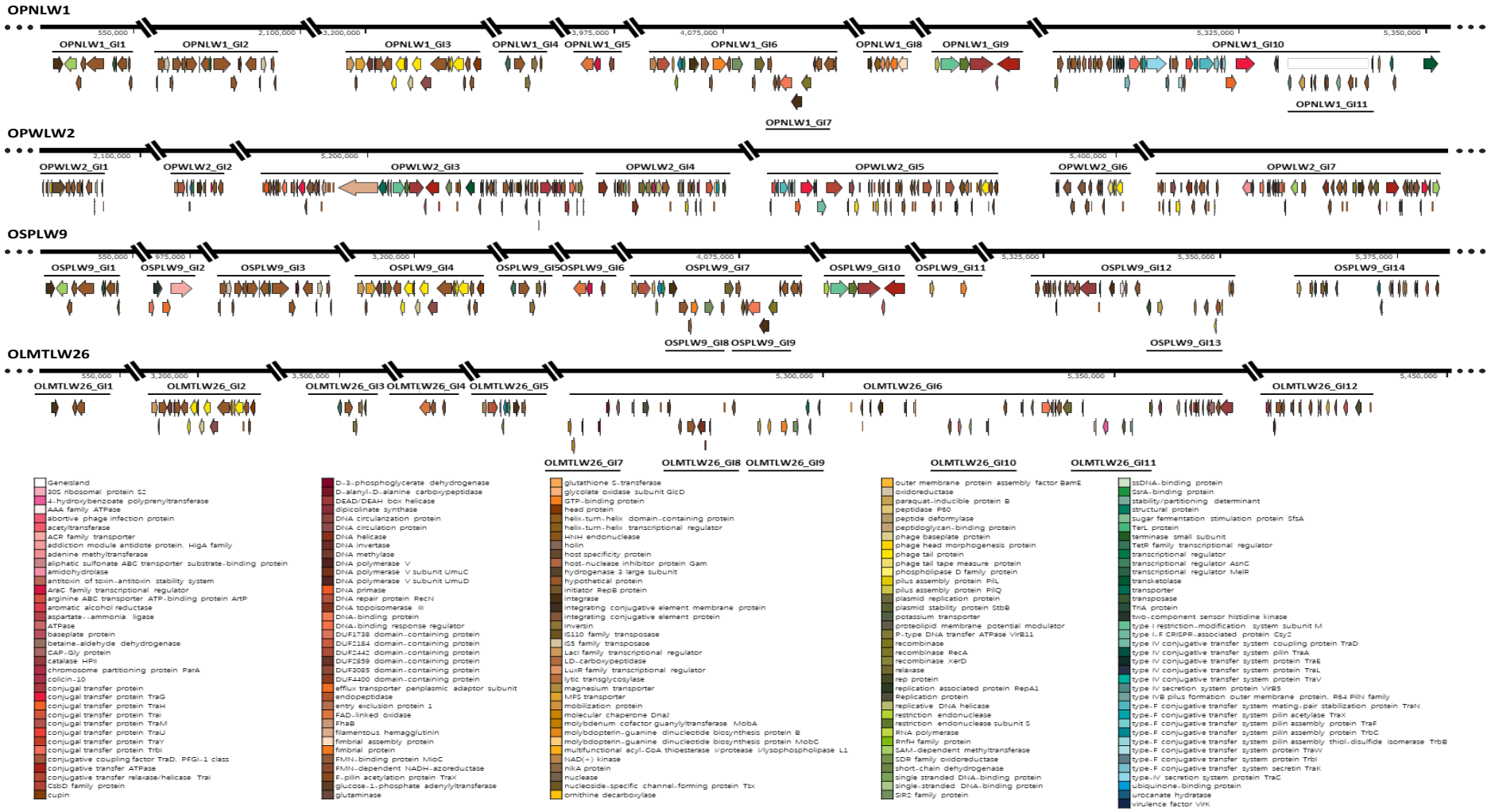


Figure 6-2: Segmental gene map of GIs and genes encoding in GIs of all 13 *Serratia* sp. *Orius* isolates genomes. The initial names of GIs in each genome by their presence order in Island viewer 3 initial data sheets (supplementary data).

A total of 87 GIs were predicted, with sizes ranging from 50 to 262 Kb by IslandViewer 3. According to initial data downloaded from IslandViewer 3 for GIs predictions from each *Serratia* sp. *Orius* isolates, Genomic Islands content (total number of GIs, total length of GIs in each genome) were summarised to analyse the association of GIs and each genome size (Table 6-1). Most of the *Serratia* sp. *Orius* isolates displayed a comparable number of GIs (ranging from 5 to 9), except for OLFL2 (4) and OLJL1 (3). Overall, there is poor correlation between genome size and GI number (Figure 6-3A). Unexpectedly, a small correlation was observed between GI number and length (Figure 6-3B), triggered by outliers like OPWLW2 (7 GIs, 262 Kb) and OSPLW9 (11 GIs, 145 Kb).

Table 6-1 Genomic Islands content in *Serratia* sp. *Orius* isolates. *Serratia* sp. SCBI was used as reference for GI prediction, and its GI content is included for comparison.

Isolate name	Number of GIs	GI length (bp)	Genome size (bp)
OLBL1	8	133655	5282384
OLCL1	7	68811	5289356
OLDL1	6	53944	5308254
OLEL1	6	78233	5302439
OLFL2	4	72266	5332885
OLHL2	7	59015	5297056
OLIL2	7	52002	5288552
OLJL1	3	50760	5290134
OLMTLW26	7	180039	5340605
OLLOLW30	5	81754	5308385
OPNLW1	9	142344	5314870
OSPLW9	11	145165	5328579
OPWLW2	7	262028	5413936
***Serratia sp.* SCBI (REF.GENOME)**	12	99913	5034688

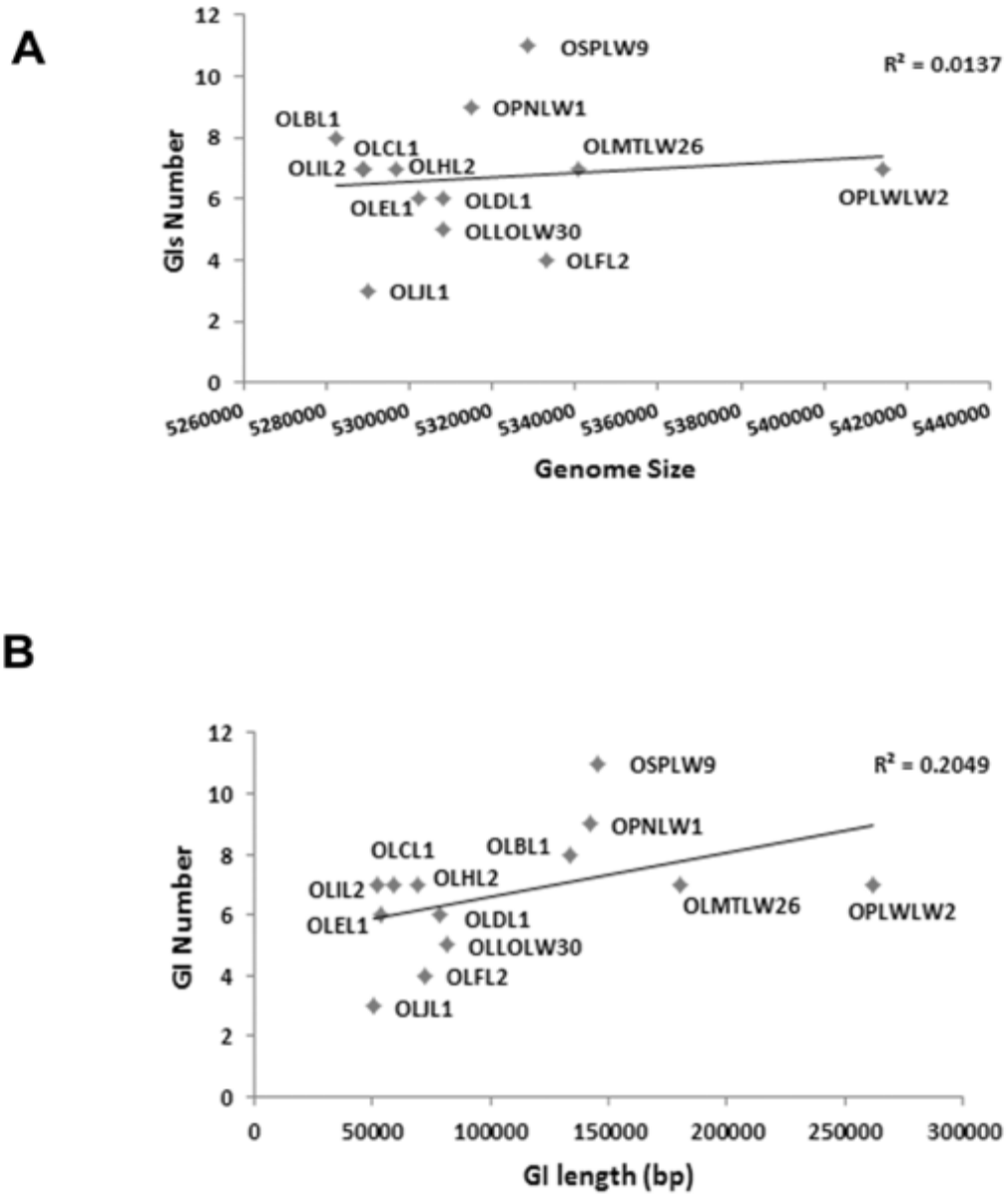


Figure 6-3: Linear correlation plots displaying correlation of GI number per genome and corresponding genome size (A) and GI length (B) by Microsoft Excel drawing the plot.

(Correlation coefficient values are shown.)

6.4.2 Further analysis of GI sequences from *Serratia* sp. *Orius* isolates

The GI sequences were clustered into representative GIs (32 unique GIs, ≥80% identity, ≥30% query coverage), using CD-Hit as described in the Methods Chapter, to identify those shared across genomes from genome-specific GIs. Supplementary Table 6-1 displays the occurrence of each representative GI and annotated features per genome analysed. Many representative GIs were genome specific, indicating that the isolates constitute independent strains. *Serratia* sp. *Orius* isolates shared numerous GIs, while *O. niger*-(OSP9) and *O. pallidicornis* (OPN1, OPWLW2)-derived isolates contained, respectively, many genome-specific GIs (Figure 6-3). Interestingly, the distribution of these two insect hosts based on presence/absence of representative GIs in genomes of corresponding symbionts is reminiscent of the COI-based closely phylogenetic evolutionary relationship of *O. niger* and *O. pallidicornis*, as proposed earlier in Chapter 3 Figure 3-2 and 3-3.

Based on the CD-Hit clustering results of GI prediction (Figure 6-3 and Supplementary Table 6-1), representative GI12 encodes a protease which is a virulence factor associated with invasion and destruction of various mammalian cell lines (Petersen and Tisa, 2014). Only OLMTLW26 presents this GI, it is likely to show OLMTLW26 is the strain carrying the virulence factor. Additionally, almost every *Serratia* sp. *Orius* isolate contains GI28, GI2, GI3 and GI4, except the strains OLJL1 and OPWLW2. GI2 encodes the TetR protein which functional as a repressor of the tetracycline efflux pump encoded by the *tetA* gene. GI3 contains FAD-linked oxidases proteins which regulate carbohydrate metabolic pathways such as carbon source utilization or sugar metabolism (Ravcheev et al., 2014). However, GI3 is absent from OLCL1, OLFL2 and OPWLW2 strains. Similarly, GI4 contains various binding proteins and DNA repair proteins, and is present in all strains except OPWLW2. In addition, OPWLW2 presents six unique GIs (Figure 6-2), while OLJL1 contains GI7, and these unique GIs contain over 50% of hypothetical proteins. Furthermore, GI distribution of OPNLW1 (from *O. pallidicornis*) is similar with that of OSPLW9 (from *O. niger*), but different from OPWLW2. OPWLW2 and OPNLW9 were isolated at different time. This species may undergo horizontal gene transfer events from *O. niger* because of the phylogenic similarities between *O. niger* and *O. pallidicornis*, it is similar to close evolutionary relationship mentioned on COI-based phylogenetic tree of *Orius* specimens (Figure 3-2 and 3-3) in Chapter 3. Alternatively, OPWLW2 and OPNLW9 belong to distinct lineages, so they are more likely contain different GI distributions.

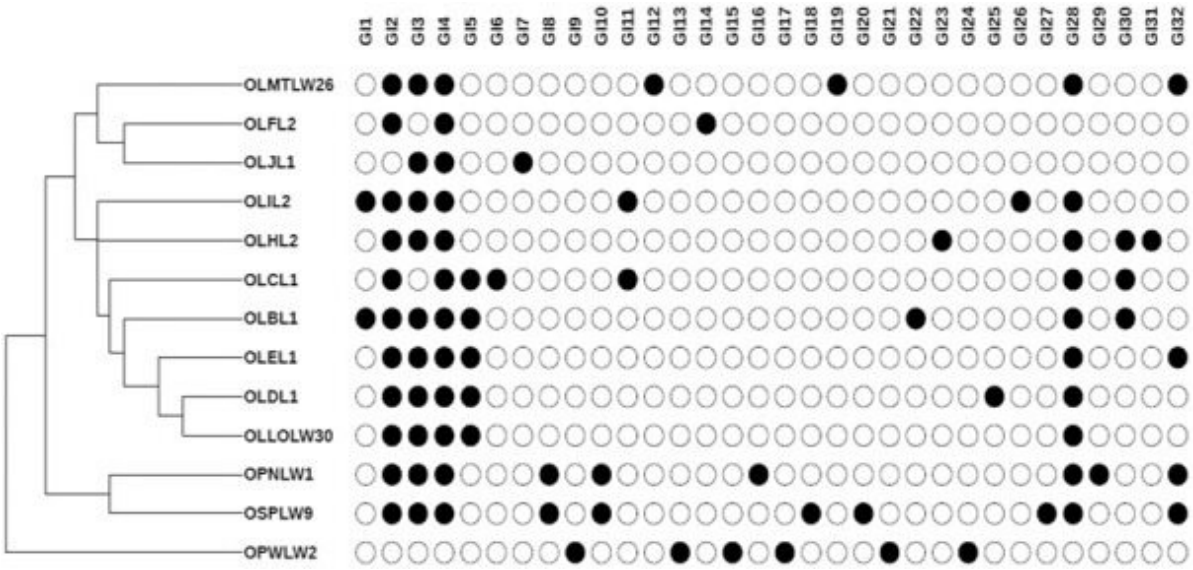


Figure 6-4: Genomic Islands profiles in *Serratia* sp. *Orius* isolates.

(Presence–absence map showing the distribution of representative GIs per genome. GI number corresponds to those described in Chapter 10 Supplementary Table 6-1).

6.4.3 Predominant Genomic Islands general description

4 representative GIs (GI28, GI2, GI3 and GI4) were predominant in most of *Serratia* sp. *Orius* isolates (Figure 6-5). All four GIs incorporate into chromosomes at multiple distinct locations and carry diverse functions of genes, indicating they may have the different origin and have evolved into different lineage branches.

PREDOMINANT GENOMIC ISLAND GI28

Based on the annotation results of genomes, the predicted GIs mostly contain phage and hypothetical protein encoding genes. Especially, GI28 harbours 21 genes, many with homologies to known genes and some with unknown functions. Most of genes in GI28 are phage related genes. The three common mechanisms mainly meant for HGT are through Phage integrase (Transduction), Transposon – Transposase (Insertion Sequences) and tRNA (Rao et al., 2020). Mostly, phage-related integrase genes are present on these predicted GIs, suggesting that they are integrated and excised in a method similar with prophages such as *Staphylococcus aureus* Pathogenicity Island (SaPI) (Rao et al., 2020). Furthermore, phages can mediate GIs transfer which confers virulence and resistance in some *Staphylococcus aureus* strains (Rao et

al., 2020). Additionally, phage related genes are also closely associated with symbiosis island acquiring by HGT (Sullivan et al., 2002). Apart from phage-related genes, some genes encoded in GI28 are associated with metabolism functions as well such as LacI family transcriptional regulator. Most LacI family transcription factors are regulators of linked carbohydrate metabolism genes, but in some *Serratia* species regulate a diverse set of metabolic pathways in bacteria such as virulence, motility, and antibiotic production, such as carbapenem antibiotic production (Lee et al., 2017). Therefore, this GI28 could be potentially act as a pathogenicity island or symbiosis island in these *Serratia* sp. *Orius* isolates.

PREDOMINANT GENOMIC ISLAND GI2

Inversin genes are the main genes in this GI, and one TetR family transcriptional regulator gene is located at the beginning of this GI (Figure 6-5). The rest of genes are hypothetical protein which are the proteins with unknown functions. For survival of diverse living environment, bacteria always prepare a wide range of rapid and adaptive responses. These responses are generally mediated by regulatory proteins, which modulate transcription, translation, or other events in gene expression so that the physiological responses are appropriate to the environmental changes (Chattoraj et al., 2011). According to sequence similarity as well as on structural and functional characteristics, these regulatory proteins are classified into multiple protein families. Of these, the tetracycline repressor (TetR) family transcriptional regulators compose the third most common transcriptional regulator family found in bacteria (Chattoraj et al., 2011). It is ubiquitous in bacteria, where it plays an important role in bacterial gene expression. The TetR family is named after the transcriptional regulators that control the expression of the tet genes, whose product confers resistance to tetracycline. However, TetR family proteins are also involved in various other important biological processes, such as biofilm formation, biosynthesis of antibiotics, catabolic pathways, multidrug resistance, nitrogen fixation, stress responses, and the pathogenicity of Gram-negative and Gram-positive bacteria (Ramos et al., 2005). Therefore, this gene in GI2 is almost likely to be functional to bacterial metabolism in these isolates. Another known gene is inversin, which usually come from eukaryote such as human, or other mammalian animals. It usually helps renal functions of mammalian animals and related to calmodulin binding (Bergmann et al., 2008). This gene may be horizontally transferred from host, but there no evidence to prove *Orius* species have

inversin gene currently, it still needs to do further analysis for host genes. According to these two genes with known functions in GI2, this GI may act as a metabolism island for facilitating isolates adapted from free living to a symbiosis lifestyle.

PREDOMINANT GENOMIC ISLAND GI3

This GI only contains three genes. Two of them were annotated as FAD-linked oxidase and FMN-dependent NADH azoreductase. The FAD-linked oxidases are known to catalyse vital redox reactions, especially for basic metabolism in many microbes (Gao et al. 2015; Heikal et al. 2014). Currently, FAD-linked oxidases could promote the expression of type III secretion system in *Ralstonia solanacearum* (Chen et al., 2021). FMN-dependent NADH azoreductase as an enzyme catalyses the reductive cleavage of azo groups (- N=N-) in aromatic azo compounds and reduction of indigo compounds as substrates (Yoneda et al., 2020). In some *Bacillus* species, it even retained complete activity even after incubation at 100 °C for 10 min (Yoneda et al., 2020). Therefore, this enzyme has very high thermostability in the bacteria. Thus, this gene could potentially to facilitate isolates to resistant to high temperatures in living environment. Therefore, this GI could be a symbiosis island helping isolates to survival with host in different living environment.

PREDOMINANT GENOMIC ISLAND GI4

Six annotated genes in GI4 possess a known function and they include NAD (+) kinase, DNA repair protein RecN, outer membrane protein assembly factor BamE, RnfH family protein, SsrA-binding protein, and integrase. The rest of genes were annotated as hypothetical proteins. NAD (+) kinase known as Nicotinamide adenine dinucleotide (NAD) kinases (NADK) which are ubiquitous enzymes. It usually catalyses the phosphorylation of NAD to nicotinamide adenine dinucleotide phosphate (NADP), which is subsequently reduced to NADPH. Since it is the only known enzyme producing NADP *de novo*, NAD kinase plays a crucial role in controlling the intracellular balance of NAD(H) and NADP(H) in many cellular metabolic pathways of bacteria (Clément et al., 2020). The RecN protein is a member of the structural maintenance of chromosomes (SMC) family of proteins. SMC proteins have important functions in a variety of housekeeping DNA processes including chromosomal condensation,

sister chromatid cohesion and recombinational DNA repair (Uranga et al., 2017). BamE belong to assembly complex of the outer membrane protein (Bam), which is involved in assembly and insertion of beta-barrel proteins into the outer membrane (Iadanza et al., 2016). RnfH proteins are located at a membrane related complex involved in transporting electrons for various reductive reactions such as nitrogen fixation (Iyer, Burroughs and Aravind, 2006). In bacteria, SsrA RNA (also known as tmRNA or 10Sa RNA) acts both as a tRNA and an mRNA in a process that clears stalled ribosomes and tags the nascent polypeptides associated with such ribosomes with a C-terminal peptide that results in their degradation (Karzai and Sauer, 2000).

Integrase is key element in this GI, it closely related to integration. There are two main requirements for the integration: functional element-encoded recombination enzymes and presence of short attachment sites, phage attachment site (*attP*) and bacterial attachment site (*attB*), recognized in site- specific recombination (Antonenka et al., 2005). Integration is catalysed by an integrase protein (Int) mediating site- specific recombination between *attP* and *attB*, and generating direct repeats called *attL* and *attR* as products of the reaction (Singh et al., 2013). The integration of GIs into the bacterial chromosome and their excision can occur naturally (Ubeda et al., 2007). There are two modes of excision, excisionase-dependent, and excisionase-independent, but both require interaction with the integrase (Bergemann et al., 1995; Schubeler et al., 1997; Semsey et al., 1999; Mir-Sanchis et al., 2012).

Integrase play a core role in GIs and are similar with a bacteriophage. The mechanism of site-specific recombination is similar with bacteriophage integration. Integrases interact with *attL* and *attR* direct repeats of the GIs, mediating excision from the host chromosomes, and formation of the cyclization intermediates. The Cre-loxP site-specific recombination system was encoded by the *E. coli* λ phage P1. The process of integration and excision only requires the Cre integrase. The Lambda site- specific recombination system was encoded by the *E. coli* λ phage. The process of integration and excision requires lambda integrase as well as Xis and integration host factors. (Wozniak and Waldor, 2010; Bellanger et al., 2014).

Furthermore, the gene components of symbiosis islands are similar with pathogenicity islands, they all contain phage and integrase. Pathogenicity islands are defined regions of chromosomal DNA containing clusters of genes required for virulence which are absent from benign isolates of the same or related species, and the available evidence suggests that pathogenicity islands were obtained by lateral transfer, but it has proved difficult to demonstrate their transfer under laboratory conditions. In contrast, the symbiosis island is ready to transfer in any situations

(Sullivan et al., 2002). Therefore, GI28 contain multiple phages related genes and GI4 contains integrase genes. It is difficult to determine which GI is pathogenicity islands or symbiosis islands by current stage, it will further analysis in the lab condition to identify their specific feature of these two GIs. However, the other components of GI4 are related to bacteria metabolism or DNA repair, there are not any phage related element or virulence factors. This GI more tends to be a symbiosis island.

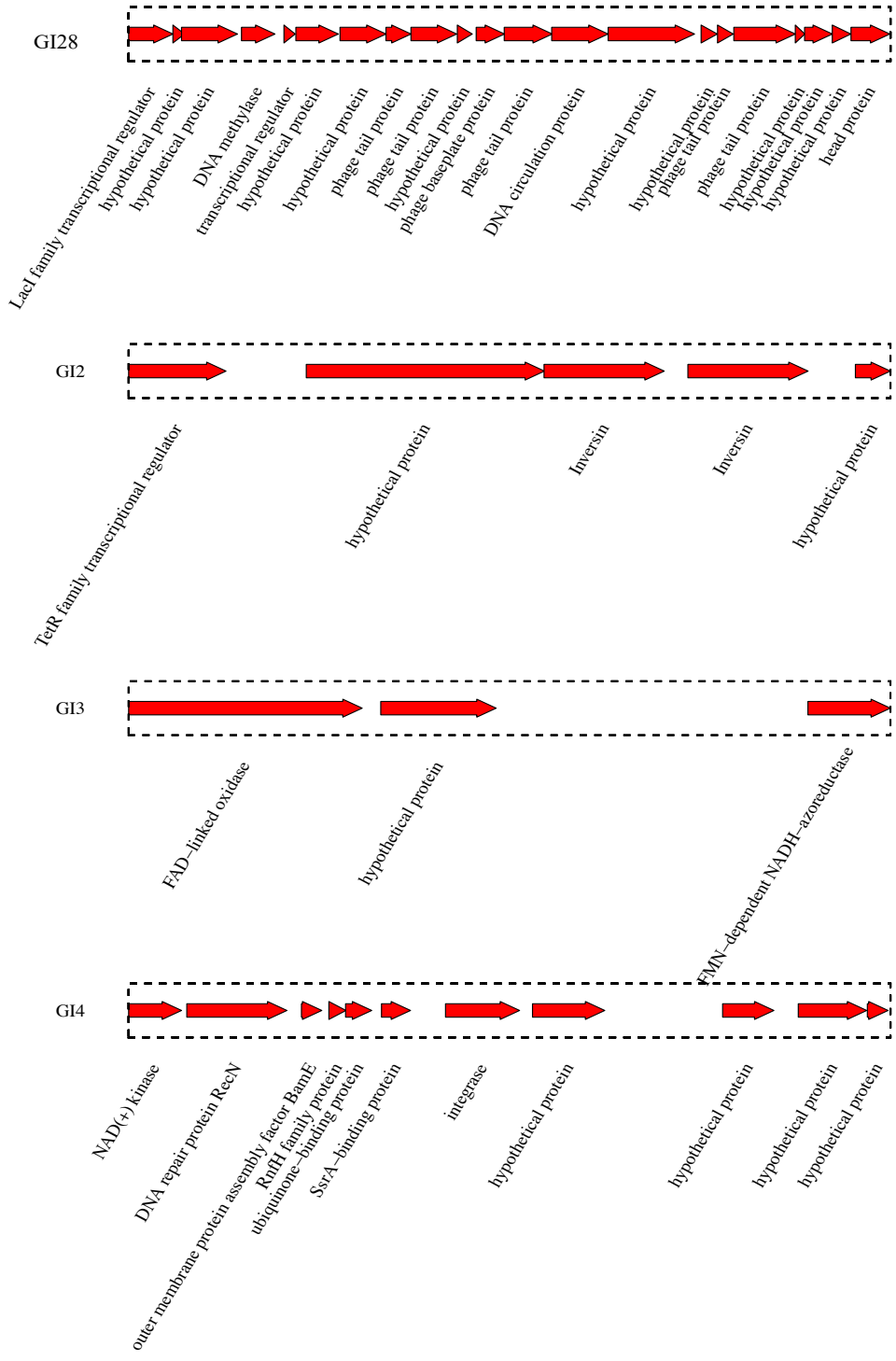


Figure 6-5: Gene map of 4 predominant GIs and genes encoding in GIs presented in all 13 *Serratia* sp. *Orius* genomes.

6.4.3 Mauve alignments for *Serratia* sp. *Orius* isolates

Figure 6-6B shows the presence of 6 blocks of homologous regions shared between OLBL1 and the *Serratia* sp. SCBI genomes. There are several gaps present within LCBs that presumably correspond to GIs.

However, there are several regions within the genomes from the isolates that displayed no synteny to the reference genome (Figure 6-4), indicative of DNA gain or loss within the isolates, as well as genome rearrangements. Figure 6-4B shows the additional LCBs (inserted/additional DNA segment) that is conserved among all 13 draft genomes but are not present in *Serratia* sp. SCBI. These regions may contain sequences originated by horizontal gene transfer events in each isolate. According to these alignments, most of the GIs corresponded to LCBs located in the end of genome that are absent from the reference *Serratia* sp. SCBI. Therefore, these regions present on the multiple alignments of genomes (Figure 6-4) but not shared with the reference genome are most likely to be the positions of GIs in each genome.

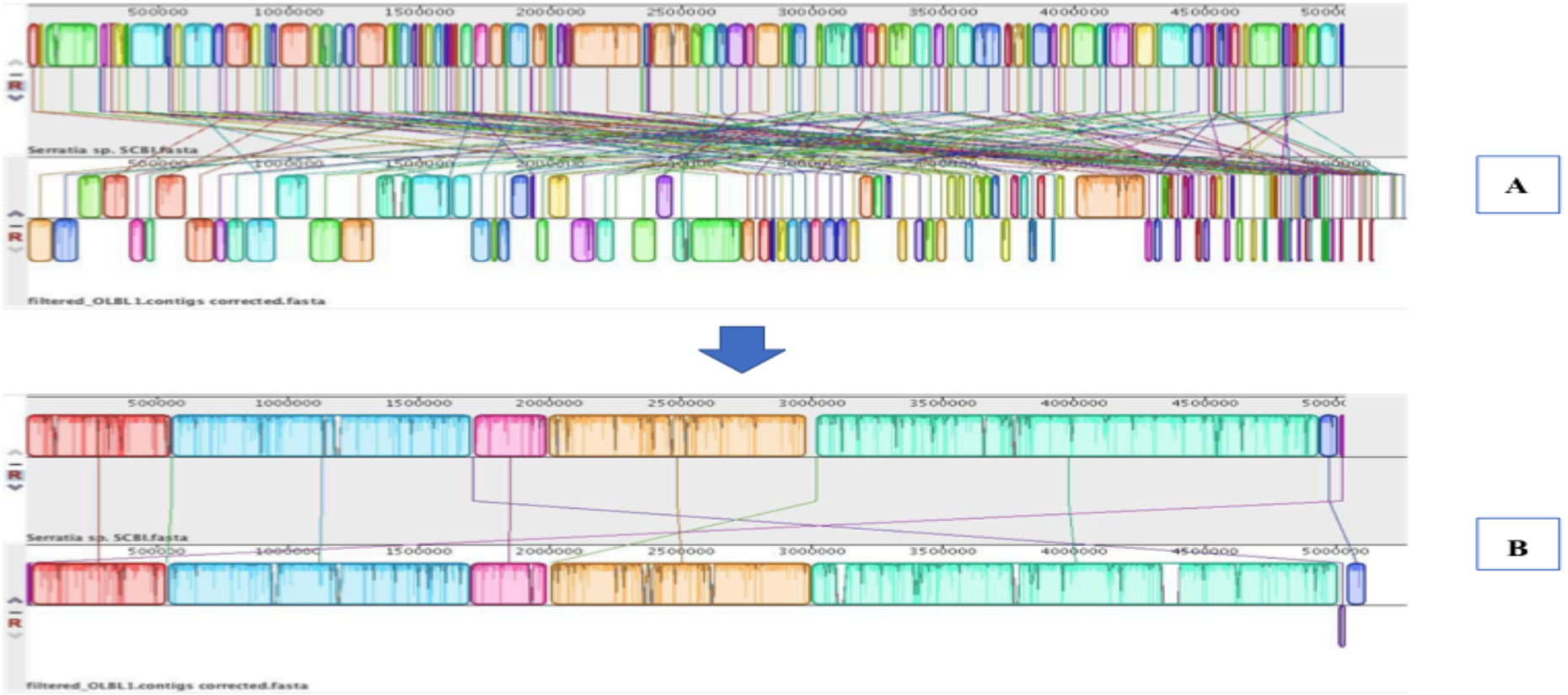


Figure 6-6: Reordering of contigs from genome OLBL1 using as reference *Serratia* sp. SCBI and MAUVE 2.4.1 software.

(Part A is original contigs order of the draft genome against reference genome. Part B shows the reordered contigs of OLBL1 by 'move contigs' in Mauve. LCB are represented by blocks of different colours. The degree of similarity is indicated using coloured areas.)

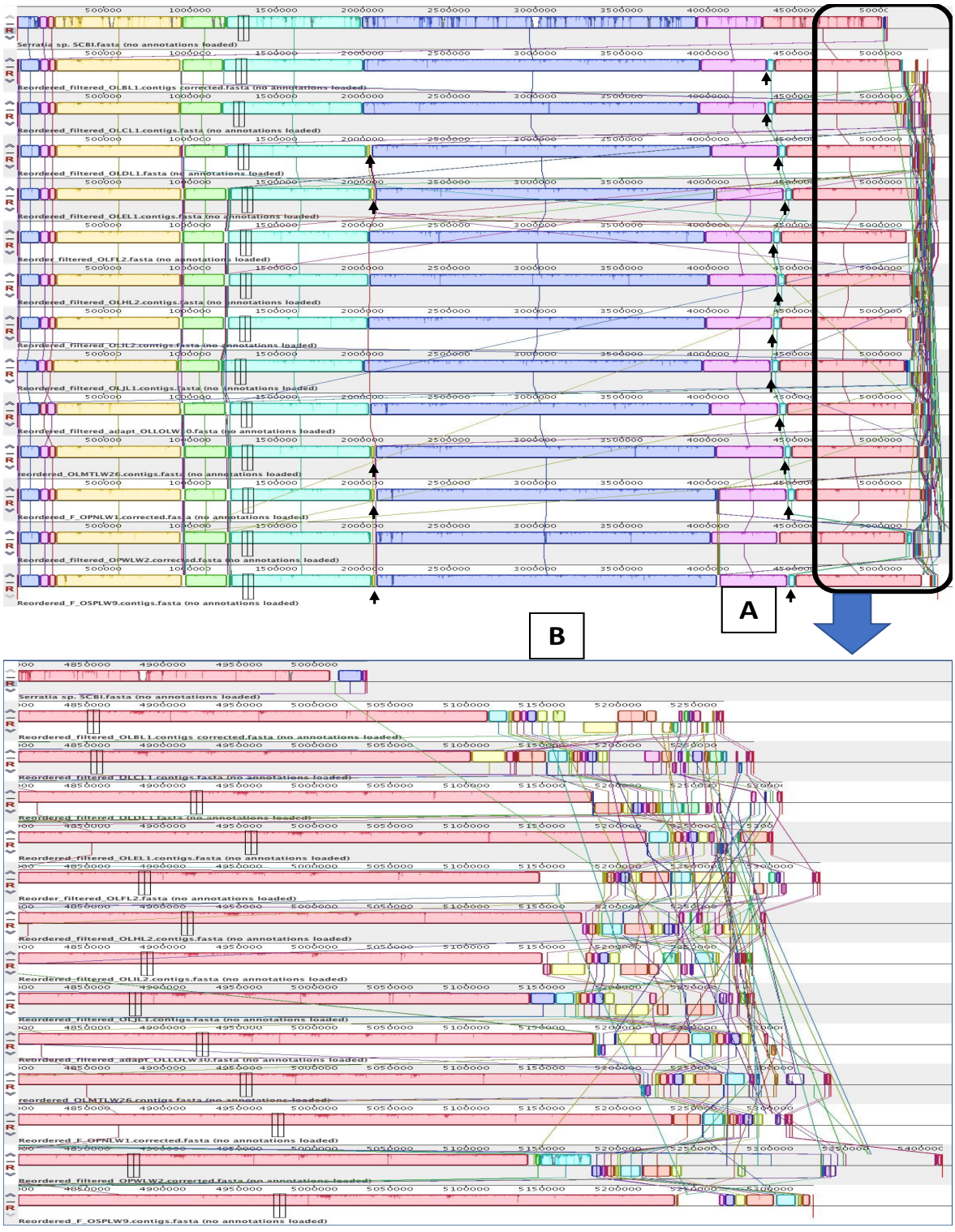


Figure 6-7: Multiple alignments of all reordered genomes of *Serratia* sp. *Orius* isolates aligned by 'Progressive Mauve' using the MAUVE aligner version 2.4.1.

(Part A: The original version of all progressive mauve alignments, the black arrows indicate the regions (LCBs) within the draft genomes that are not present in the reference genome. Part B: enlarged sections of genome ends in all progressive mauve alignments, these LCBs are not present in the reference genome.)

6.5 Discussion and conclusion

GI prediction is important in the investigation of horizontal gene transfer between microbial genomes, particularly for the functions of bacterial niche colonization, catabolism of diverse substrates, symbiotic relationships to their hosts, resistance to antimicrobial agents or enhanced virulence (Dobrindt et al., 2004). GI prediction indicates the important mobile genetic factors that contribute to potential rapid changes of virulence in many bacterial pathogens.

The analysis of GI prediction among all the *Serratia* sp. *Orius* isolates reveals that all these isolates are different from each other since they all present different GI features from each genome and most GIs are genome specific. This is mainly because these insects are living in different geographic locations, their symbionts need to adapt to different living environments and have undergone distinct evolutionary events or horizontal gene transfer, in turn to become putative independent strains due to their distinct GI characteristics. The analysis also confirmed that these isolates are symbiotically related to their hosts, rather than the results being due to contamination of the *Orius* species under lab-rearing conditions. Since the distributions of GI-based hierarchical clustering in these isolate genomes closely reflect their host phylogenetic associations, the GIs of these isolates even differ between the same insect host species, especially in *O. pallidicornis.* The availability of same species genome sequences allowed the extensive analysis of the *Serratia* sp. *Orius* isolates. Despite the close genome sequence similarity across the isolates, the characterisation of genomic islands permitted their segregation into putative independent strains, suggesting that despite their common origin independent lineages are emerging within their respective host populations. Indeed, GI-based hierarchical clustering of these genomes closely resembles the host phylogenetic relationship.

When these genomes exhibited large variability in total GI length, they still contained similar genome sizes. This indicates that the symbiotic associations between these isolates and various *Orius* species were established in the ancient ancestor to recent European *Orius* species, although they were not affected by genome size reduction. It also indicates that this symbiotic relationship is the result of the routine acquisition of *Serratia* species from environmental resources such as soil and water. Furthermore, previous bacterial isolations from *Orius* specimens failed to isolate other *Serratia* species, which also supports the opinion that this is a long-established symbiotic relationship.

CHAPTER 7: Pangenome analysis of *Serratia* sp. *Orius* isolates

7.1 Abstract of this chapter

- A pangenome constructed using Roary revealed an open pangenome for the species within the *Serratia* sp. SCBI complex.
- 279 accessory genes were identified as related to *Orius* facultative symbionts, within the 13 *Orius* associated *Serratia* isolates present in the SCBI complex.
- Accessory genes of *Orius* associated *Serratia* isolates reveal many mobile elements and several plasmids associated genes. BLASTN sequence homology search indicated plasmid exchange across the *Serratia* genus.

7.2 Introduction

This chapter aims to generate a pangenome of the *Serratia* isolates for identification of genetic traits and related genes related to insect symbiotic association. Over decades of genomic sequencing development, experimental and mathematical modelling predictions have shown that new genes can be detected even after sequencing more than hundreds of genomes per species. Thus, distinct strains of the same species had their genome sequenced, it becomes obvious that there are numerous intraspecific variations in prokaryotic genome content. Furthermore, the term of 'pangenome' was created for the set of orthologous and unique genes of a specific group of organisms to provide a better understanding of genotype-phenotype associations of the genes in these organisms (Tettelin et al., 2005; Kim, Gu, Kim, and Lee, 2020). The terms of 'core' and 'accessory' genomes represent the genomic variability and stability of different strains from same species, respectively (McInerney, McNally and O'Connell, 2017). The pangenome consists of all the gene families that have been found in the species, the core genome refers to 'essential' gene that are found in all members sequenced thus far, and the accessory genome refers to 'dispensable' genes that are not in each genome (Costa et al., 2020).

The core and accessory genomes also represent the stability and diversity of the species, respectively. Most of core genes are involved in vital roles for bacterial survival. However,

genes of the core genome may also be involved in pathogenicity and virulence in some bacterial species (Koonin, Makarova and Wolf, 2021). Accessory and unique genes are acquired by horizontal gene transfer (HGT) or evolved due to mutations in pre-existing genes. They are commonly related to a specific metabolism, virulence, antibiotic resistance mechanism, or other environmental adaptation for their specific lifestyles and evolutionary trajectories (Kim, et al., 2020).

Furthermore, the current crucial corollary of the discovery of pangenomes is that the essential evolutionary process in prokaryotes is not point mutations but rather gene replacement via HGT and gene loss (Koonin, Makarova and Wolf, 2021). In this chapter, it is important to distinguish the terms genomic plasticity and accessory genome. Genomic plasticity is used to describe the mobile genetic elements (MGEs) and hypervariable regions that transform the genome into a dynamic molecule. Therefore, it is a concept used to discuss the genetic variability of a single or multiple genomes without necessarily making use of a pan-genomic approach. But in some cases, MGEs and hypervariable regions comprise most of the accessory genes in pangenome (Costa et al., 2020). For instance, in strains of *Bacillus amyloliquefaciens*, most gene clusters to produce secondary metabolites are present in the accessory genome of the species (Belbahri et al., 2017).

Most comparative genomic analyses begin by identifying the homologous characteristics between 2 or more genomes. These homologies range from large chromosomal segments to genes or even point mutations. In a pan-genome analysis, the genes are the main characteristics evaluated. From an evolutionary perspective, a gene is classified as homologous or analogous. Homologous genes are those originated from a common ancestor, whereas the analogous genes evolved independently through convergent evolution. In both cases, they will present the same function but in different organisms. About 15% of the genes of a bacterium are acquired through HGT (Paquola et al., 2018). A pan-genomic analysis searches for homologous genes within the set of analysed genomes. These homologous genes are divided into orthologous and paralogous genes (Altenhoff et al., 2012). Orthologous genes diverged via evolutionary speciation, whereas paralogous genes diverged via gene duplication. Therefore, orthologous genes are those shared by 2 or more bacteria and have equivalent biological function. It is worth noting that orthologous genes tend to be more conserved than paralogous genes (Chen and Zhang, 2012). In contrast, paralogous genes commonly having experienced several mutations

after their duplication, and it is resulted in an alteration in their biological function (Gabaldón, et al., 2012).

Theoretically, a bacterial species whose population is highly clonal (representing a closed pan-genome) is more successful in colonizing stable environments such as the human or animal tissues. In contrast, free-living (representing an open pangenome) microorganisms have greater gene variability to adapt to different environmental conditions. For example, the coagulasea negative staphylococci *Staphylococcus lugdunensis* is a commensal bacterium with closed pan-genome (Argemi et al., 2018). However, several other research studies illustrated that this theory is not always be right (Kawai et al., 2011). The genome of *Helicobacter pylori* displays significant divergence depending on the geographical location of the isolate, especially in East Asia lineage and European lineage (Kawai et al., 2011). It is worth to point out that Lapierre and Gogarten illustrated that the whole bacteria domain appears to have an open pan-genome (Lapierre and Gogarten, 2009). Therefore, it is difficult to define whether closed pan-genomes are true evidence of species with limited gene frequency or if they are only artifacts from analysis with a limited number of genomes (Costa, et al., 2020). The maintenance of gene frequency in a pan-genome has been subject of several studies. Rodriguez-Valera et al raised the hypothesis that the pan-genome of a bacterial population is maintained and equalized by phage predation (Rodriguez-Valera et al., 2009). Many works analyse the relationship between microbial communities and abiotic factors. However, bacteria also need to adapt to biotic factors such as phage predation. A bacterial population under constant phage predation is also under constant modulation of its gene content. This process is called pan-selectome, and the pan-genome is a snapshot of the gene frequency in each population under constant phage predation (Rodriguez-Valera et al., 2009). Subsequently, Rodriguez-Valera et al postulated that this pan-selectome is the evolutionary unit of selection in the microbial world. Therefore, at the genomic level the unit of selection is the pan-genome (Rodriguez-Valera et al., 2009).

Pangenome analysis typically involves a collection of genome data, homology clustering based on multiple sequence alignment, and profiling of core and accessory genomes. These three main steps are considered a core part of the pan-genome analysis (Koonin, Makarova and Wolf, 2021). Representative biological information that can be reaped from the pan-genome analyses includes phylogenetic distances, presence, or absence of genes across a target clade, and functional distribution of proteins (Koonin, Makarova and Wolf, 2021). Because of the immense volume and diversity of genome data that need to be processed, development of new

computational tools is an active area of studies in the field of pan-genome analysis (Figure 7-1). The homogenization of the annotations avoids the wrong identification of core genes into the shared subset and shared genes assigned to singletons. This step should be applied by genome annotation computational tools, such as RAST and Prokka. The clustering analysis is normally achieved by first performing an all-vs-all bidirectional BLAST analysis followed using an orthology identification software, such as OrthoMCL. The complete table with all the orthologous genes may then be used for pan-genome development analyses, which will fit the specific curve generated from permutations of all genomes in all positions. Normally, the software performs curve fitting of the pan-genome using Heaps law or Power Law, whereas the curve fitting of shared genome and core genome are performed by means of exponential regression decay (Costa et al., 2020).

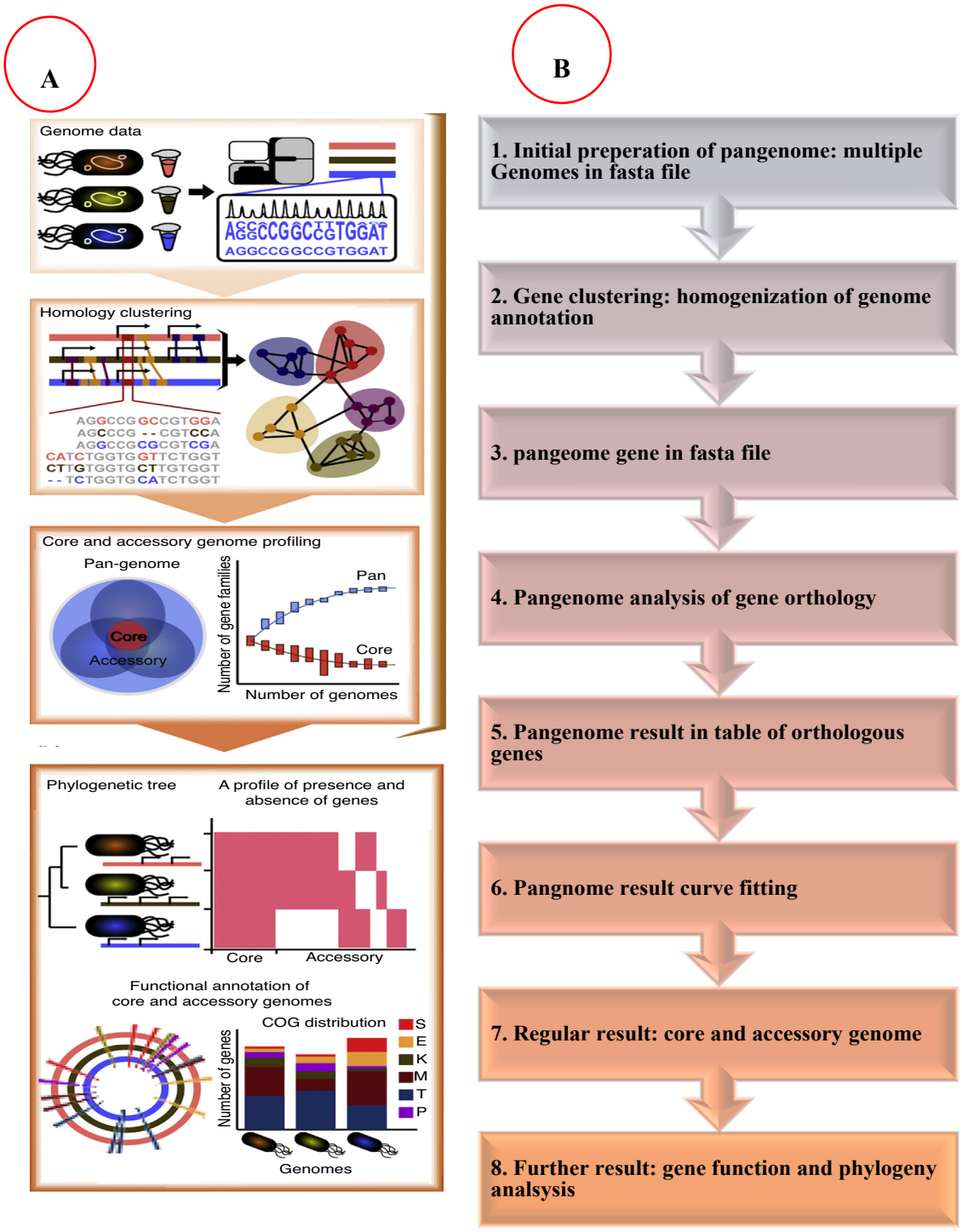


Figure 7-1: Flow diagram represents main steps working for essential part of pan-genome analysis in most softwires.

((A) it detailly comprises preparation of genome data, homology clustering, and profiling of core and accessory genomes. Anticipated biological information obtainable from the pan- genome analysis, including a phylogenetic tree, presence, or absence of genes in a microbial clade, and functional annotation of core and accessory genes (Koonin, Makarova and Wolf, 2021). (B) explained and summarised the main steps in a pan-genomic analysis. Each process (represented by blocks) can be performed by different methods.)

Roary (Page et al. 2015) can construct a large pangenomes even on a typical desktop machine, yielding fairly accurate results. Roary also uses CD-HIT, BLAST and MCL for the orthology analyses. For instance, it can digest up to 1000 strains (13 GB of RAM) building the pangenome in ~4 h. The accuracy of Roary is attributable to utilization of the context of conserved gene neighborhood information (Figure 7-2). A suite of command line tools is provided to interrogate the dataset providing union, intersection, and complement (Costa et al., 2020).

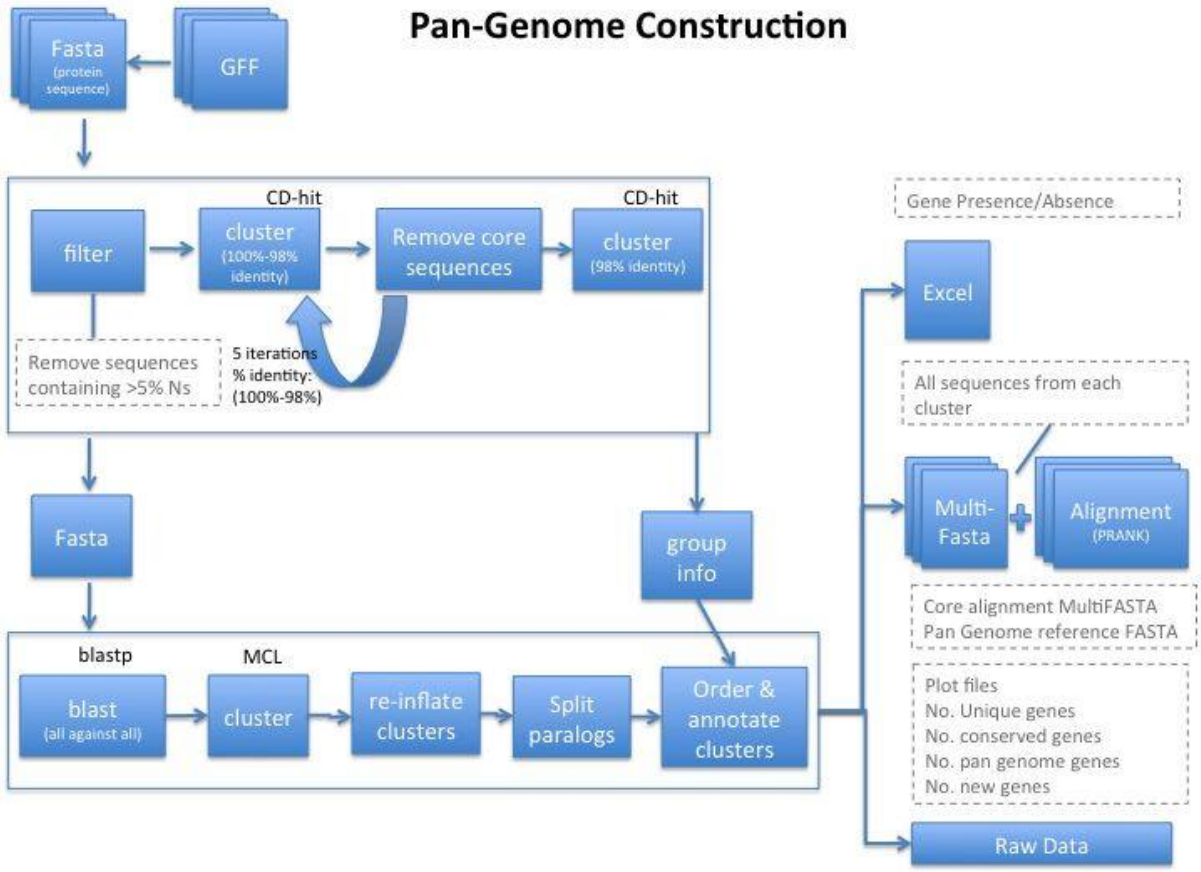


Figure 7-2: The flowchart of the steps in Roary application (Page et al. 2015).

(The pipeline takes as input GFF3 files created by Prokka and clusters the predicted proteins to allow for the full genomic variation of the input set to be explored. The basic method is to filter and precluster the proteins, perform an all against all comparison using BLASTP, and cluster with MCL.)

SCOARY is used for studying the association between pangenome genes presence or absence and observed phenotypes and usually apply with Roary pangenome construction. It is termed the method "pan-GWAS" to distinguish it from traditional SNP-based Genome-wide

association studies (GWAS). Each candidate gene in the accessory genome is sequentially scored the components of the pan-genome for associations to observed phenotypic traits while accounting for population stratification, with minimal assumptions about evolutionary processes. Scoary is implemented in Python and is available under an open source GPLv3 license (Brynildsrud et al., 2016).

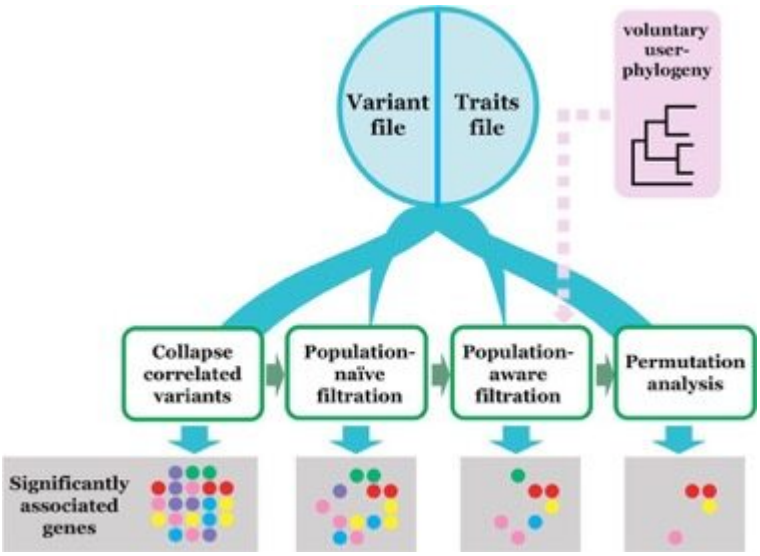


Figure 7-3: Flow diagram shows the main steps of Scoary analysis.

(The input files are a genotype and a phenotype matrix, and optionally a phylogenetic tree which will define the sample pedigree. If the latter is not provided it is calculated internally through the isolate Hamming distances of the input genotype file. Each candidate variant goes through a set of filtration steps, the thresholds for every set by the user. Fewer and fewer candidate variants will be left to analyse as the computational complexity of operations increase. Variants that pass all filters are returned as results (Brynildsrud et al., 2016).)

Bacterial Pangenome Analysis (BPGA) (Chaudhari et al. 2016), comes with several new functions, the most significant optimised execution speed. Additionally, it further provides phylogeny of various entities (core-, pangenome and MLST), subset analysis, atypical sequence component analysis, orthologous, and functional annotation for all gene datasets, user-selection of gene clustering algorithm, command line interface, and nice graphics. It runs both in Windows and in Linux as executables files (source code in Perl). BPGA has dependencies with other tools that needs to be installed. In terms of input files, BPGA allow to accept following file formats: GenBank (.gbk) files, protein sequence file (faa or .fsa or fasta format), binary (0,1) matrix (tab-delimited) file as output of other tools. The seven functional

modules of BPGA algorithm involve: Pangenome profile analysis, pangenome sequence extraction, exclusive gene family analysis, atypical GC content analysis, pangenome functional analysis, species phylogenetic analysis, and subset analysis (optional) (Tettelin and Medini, 2020). BPGA software uses USEARCH, or CD-HIT, or OrthoMCL software for the orthology analyses and power- law regression and exponential curve fit for the pan-genome and core genome developments, respectively. It also implements other relevant analyses such as core/pan/MLST (Multi Locus Sequence Typing) phylogeny, subset analysis, and Kyoto Encyclopaedia of Genes and Genomes (KEGG) & Clusters of Orthologous Genes (COG) mapping (Costa et al., 2020).

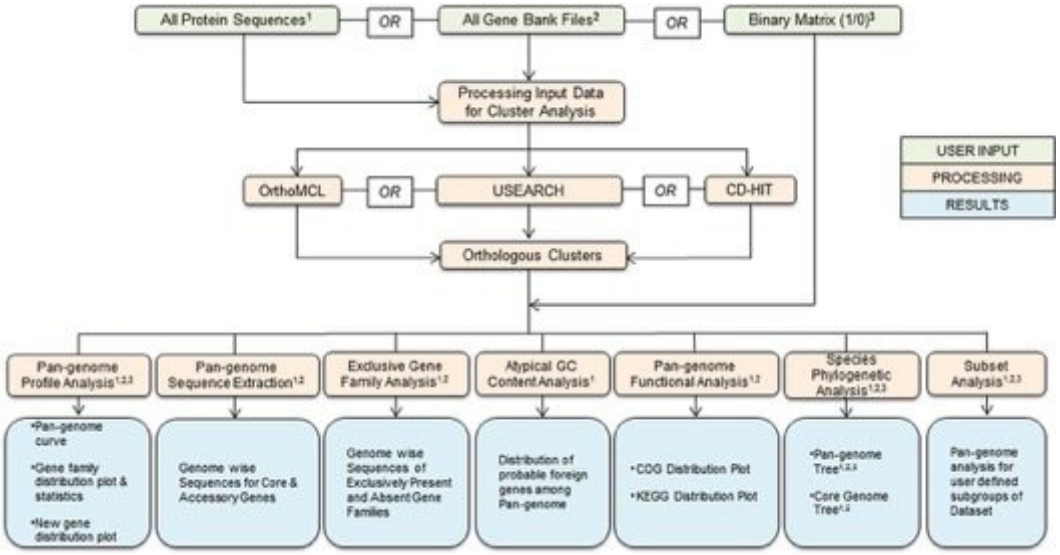


Figure 7-4: BPGA workflow. Initially, BPGA prepare sequence data for clustering.

(BPGA then runs USEARCH for fastest clustering (using 50% sequence identity cut-off by default; user may change this cut-off value). The clustered output is processed to generate tab delimited gene presence absence binary matrix (pan-matrix) which is then used for pan-genome profile calculations with iterations (default 20 or user defined) as well as pan genome-based phylogeny. MUSCLE is used to align concatenated core genes to generate phylogeny tree based on core genome and MLST based on user selected housekeeping genes. For assigning COG and KEGG IDs, best hits with respective reference databases obtained from ublast function of USEARCH are used. Gnuplot is used for plotting all the results as high-quality pdf images (Chaudhari et al. 2016).)

7.3 Methods

In this study, Prokka was used to annotate the genome sequences from all *Serratia* sp. SCBI-like isolates. The resulting gff files were used as input in Roary to generate a *Serratia* SCBI

pangenome. Basically, the method of pangenome construction is to filter and pre-cluster the predicted protein sequences by CD-HIT iteratively from 100% down to a default to 98% sequence identity, thus reducing computation of subsequent steps. An all against all comparison using BLASTP was then implemented on remaining sequences and finally clustering with MCL (Markov Cluster Algorithm) was performed. Scoary was subsequently used for pan-GWAS to identify genes considered related to the trait of insect symbiosis. The detailed method to construct and analyse the pangenome using Roary and Scoary has been described in the Methods Chapter. BPGA was later used for functional annotation of all genes in the pangenome and calculated the curve to confirm the classification of *Serratia* sp. *Orius* isolates pangenome whether the pangenome is open or closed. Furthermore, sequence homology searches using these plasmid- related genes identified single contig assemblies corresponding to putative plasmid sequences, well conserved across all the *Serratia* sp. draft genomes reported. A typical sequence (NODE_21 in OSP9LW9, Accession number: NZ_MSTM01000051.1) was used as query in BLASTN (Standard Nucleotide BLAST) to search available *Serratia* sequences.

7.4 Results

7.4.1 Statistics of *Serratia* sp. *Orius* isolates Pangenome

After the pangenome was constructed by Roary, it produced multiple output files. This involves a spreadsheet with information of the presence and absence of each gene in each isolate, number of isolates a gene is found in, frequency of gene per isolates, functional annotation, QC information and sorting information, but due to the excessive data capacity in Excel worksheet, it cannot show in supplementary data. The summary of pan-genome statistics represented in Table 7-1. The cloud genes, soft core genes, and shell genes are involved in accessory genes of pangenomes (Figure 7-5). Shell genes are shared by more than 15% to 95% of the genomes in the pangenome and soft-core genes shared by more than 95% to 99% of genes in the pangenome (Figure 7-6). The total gene number within the pangenome was 8120, with 3517 core genes shared across isolates yielding 43% of the pangenome. A further 515 soft core genes (95% <= strains < 99%), 1165 shell genes (15% <= strains < 95%) and 2923 cloud genes (0% <= strains < 15%) were identified (Figure 7-6). In this pangenome pie chart (Figure 7-6), the size of accessory genes is larger than core gene, it tends to be an open pangenome and confirmed in following curving calculations (Table 7-2).

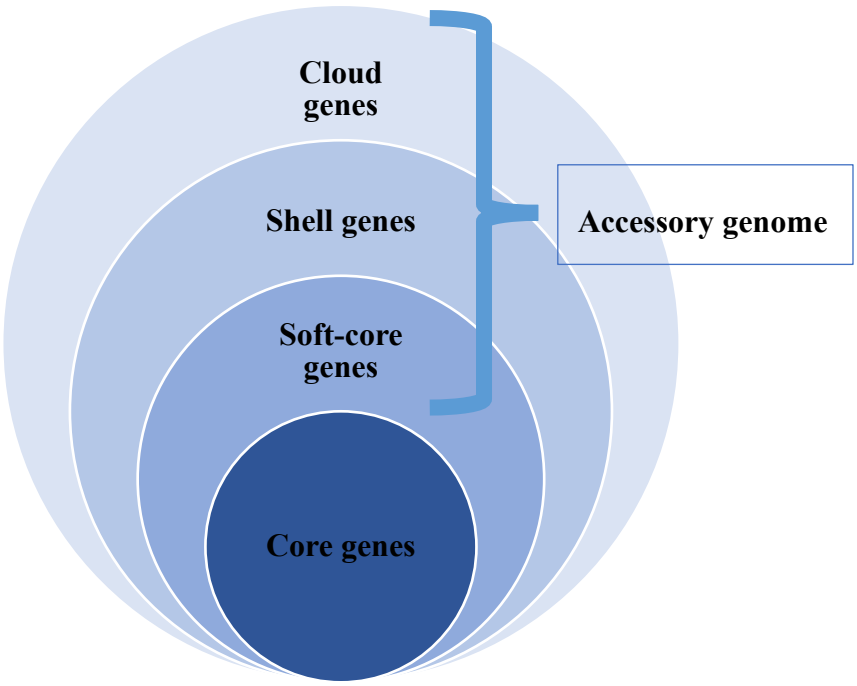


Figure 7-5: The structure of a pangenome.

Table 7-1: Summary of pan-genome genes statistics

Gene name	Percentage range of strains	Gene number
Core genes	(99% <= strains <= 100%)	3517
Soft core genes	(95% <= strains < 99%)	515
Shell genes	(15% <= strains < 95%)	1165
Cloud genes	(0% <= strains < 15%)	2923
Total genes	(0% <= strains <= 100%)	8120

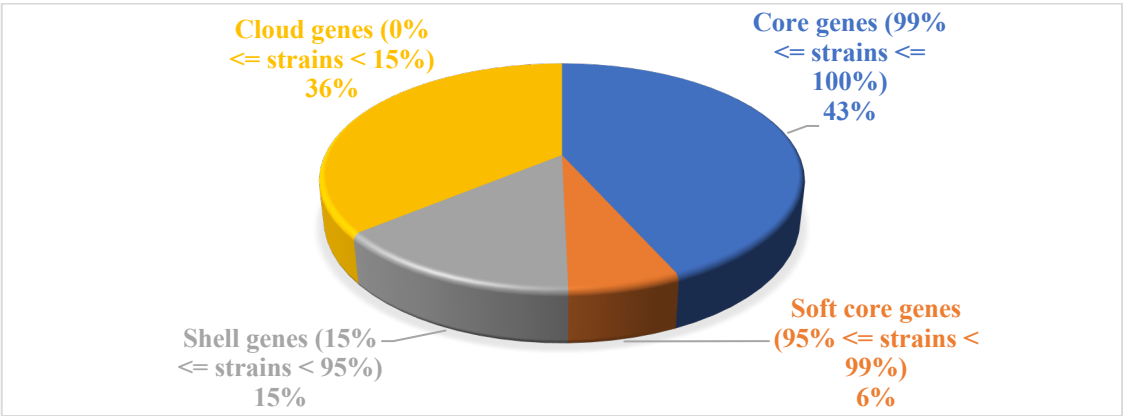


Figure 7-6: The pie chart of *Serratia* sp. *Orius* isolates pangenome statistics.

(This chart includes multiple percentage range of each component in the pangenome. Yellow part is the percentage of cloud genes. The section of soft-core genes represented as red colour. Shell genes is prey part. The rest part is blue colour.)

Based on the tab delimited files (the gene presence and absence file) of Roary pangenome outputs, two graphs were generated displaying the number of conserved genes and total genes when the number of genomes increases ((Figure 7-6), and the number of new genes and unique genes when the number of genomes increase (Figure 7-7). The number of conserved genes represents the size of the core genome, while the total number of genes including core and accessory genes corresponds to pangenome size.

Figure 7-7 shows that the total number of genes in the pangenome is increased whenever a new genome is added, while the number of conserved genes remains stable or is slightly reduced. Figure 7-8 shows that the number of unique genes increased when adding genomes to the pangenome, while the overall number of new genes remains stable (Figure 7-7). These observations, where the addition of a new genome to the pangenome results in an increase of the total number of genes, confirms that the *Serratia* sp. SCBI complex (Figure 5-1 in chapter 5) possess an open pangenome. A pan-genome is classified as open or closed depending on the probability of detecting new gene families as new genomes are added into the analysis. In an open pan- genome, the number of gene families will continuously increase with the addition of new genomes to the analysis. In contrast, in a closed pan-genome, the number of gene families will not increase considerably (Costa et al., 2020).

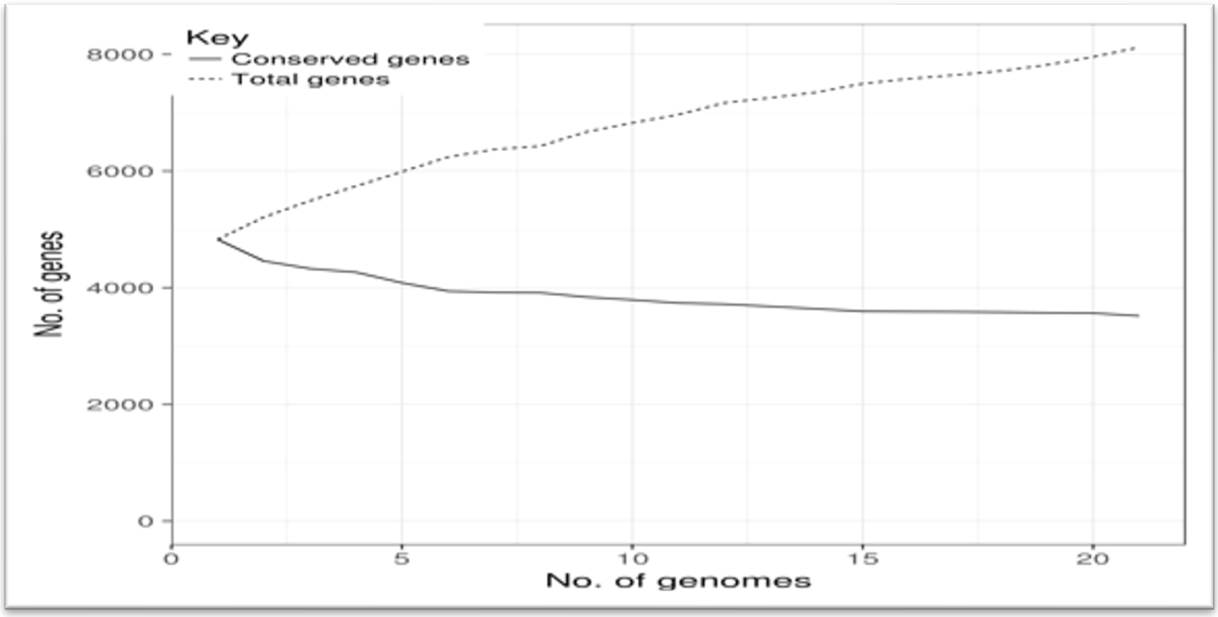


Figure 7-7: Pangenome analysis of *Serratia* sp. *Orius* isolates.

(Overview of the complete pangenome, displaying how the addition of genomes does not lead to an increase of conserved gene content, while total gene content continues to augment, suggestive of an open pangenome for *Serratia* sp. *Orius* isolates.)

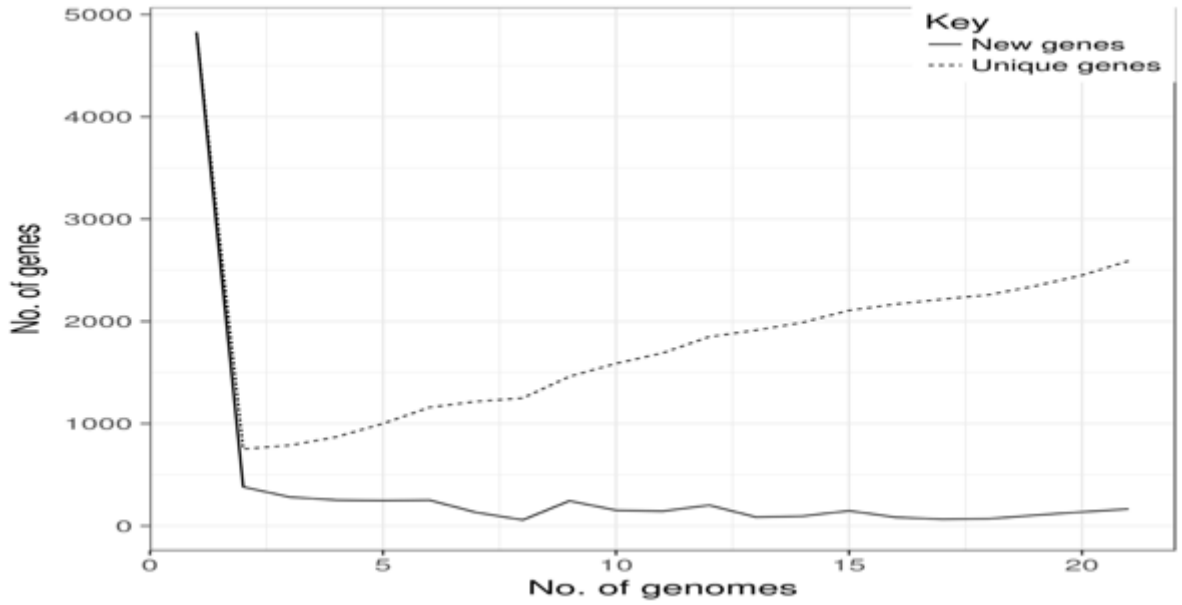


Figure 7-8: Pangenome analysis of *Serratia* sp. *Orius* isolates.

(Overview of the complete pangenome revealed that as new genomes are added, the number of new genes does not increase while unique gene content continues to augment, suggestive of an open pangenome for the SCBI complex.)

Table 7-2 The curve calculation of BPGA Pangenome classification.

	PAN GENOME	CORE GENOME
FIT LAW	POWER	EXPONENTIAL
EQUATION	f(x)=a.x^b	f1(x)=c.e^(d.x)
PARAMETERS	a= 4637.3	c= 4644.22
_____	b= 0.0258825	d= -0.00253827
EXPECTED SIZE	4983	0
ESTIMATED SIZE	4955.61	4493.47

In Table 7-2, the power-law regression model for the pan-genome data and an exponential curve fit model in case of the core genome data. Where, a, b, c and d are the fitting parameters. F(x) and fl(x) are calculated pan-genome and core genome sizes respectively. If parameter 'b'≤1, then the pangenome is considered open, according to power- law regression and exponential curve fit for the pan-genome and core genome developments (Table 7-2). The calculated parameter 'b' = 0.0258825 indicates that the pan genome is open. (Costa, et al., 2020). BPGA Pangenome calculations confirmed the classification of Roary pangenome, both pangenome represents the pangenome of *Serratia* sp. *Orius* isolates is an open pangenome.

Furthermore, a Newick tree, based on presence and absence of genes in the accessory genome using FastTree (Price et al., 2010), was combined with a matrix of presence and absence of core and accessory genes created from the information contained in the gene_presence_absence matrix and accessory binary genes in Newick format (Figure 7-8A). The grey section shown in the heat map shared across all genomes without any spaces corresponds mostly to genes present in the core genome, while some grey sections with spaces shared among a few genomes only are accessory genes. The black section corresponds to the genes only presented in *Serratia* sp. *Orius* isolates, which were analysed using Scoary to define the association of these genes and the insect symbiosis trait.

7.4.2 Scoary output analysis for pan-GWA (pan-genome-wide association) study of symbiotic trait related genes of *Serratia* sp. *Orius* isolates

Scoary applies the pairwise comparisons algorithm to identify the maximum number of non-intersecting pairs of isolates that contrast in the state of both gene and trait. The inputs of Scoary required the gene_presence_absence.csv file from Roary and as a trait presence only in *Orius* derived isolates, to test association of genes to insect symbiosis. The Scoary output is a single csv file per trait in the traits file. The results consist of genes that were found to be associated with the trait, sorted according to significance (Supplementary data Table 7-1). In Scoary output, 'Number_pos_present_in' shows the number of trait-positive isolates a gene was found in. In this case it refers to genes found in 13 *Serratia* sp. *Orius* isolates, but not found in the rest (8) of the SCBI complex genomes (Figure 5-1 in chapter 5), using a naïve p-value of 4.91E-06. This analysis identified a total of 279 accessory genes only found in *Orius Serratia sp.* SCBI-like facultative symbionts, but not in any of the other *Serratia* species within the SCBI complex. Most of the gene annotations show hypothetical proteins with unknown functions.

Only 79 known genes have annotated by Scoary which related to their symbiotic lifestyle (Figure 7-8B and Table 7-3). The section B of Figure 7-8 also mentioned their gene annotation as well. Most of these genes belong to the category of cellular metabolic pathways, virulence factor and Phages. For example, *udh* (uronate dehydrogenase) gene relates to pathway D-galacturonate degradation and in Carbohydrate acid metabolism, and it is closely associated with d-Galacturonate, the primary constituent of pectin, an abundant polymer found in plant cell walls. Following hydrolysis by pectinases, d-galacturonate can be utilized as a sole carbon

source by many soil bacteria (Bouvier et al., 2014). This gene may relate to insect host diet habit, and horizontally transfer from the host for optimise their symbiotic lifestyles.

Furthermore, katE, azoR, dnaB and group_2404 genes (unnamed genes) are also presented in genomic islands of some isolates. KatE is hydroperoxidase II (HPII) and bacteria always possesses multiple distinct catalases to defend itself against oxidative stress, including hydroperoxidase I (HPI), katG and hydroperoxidase II (HPII), KatE (Baoshan et al., 2019). AzoR is mediated azoreductase activity in cellular metabolic pathway (Ryan et al., 2014). Potentially it horizontally transfers from other bacteria and helps the expression of these reductases during bacterial growth (Mercier et al.,2013). DnaB participates in initiation and elongation during chromosome replication (Poehlein, Freese, Daniel and Simeonova, 2014). Group_2404 gene is Acetyltransferase. It is responsible to the regulation of cell shape and peptidoglycan biosynthetic process (Lamelas et al., 2011). The rest of genes were annotated as hypothetical proteins, which are the proteins of unknown functions. Therefore, BPGA software used KEGG and Cluster of Orthologous Groups (COGs) databases to re-annotate all the genes in the pangenome to classify their functions.

Interestingly, within the Roary *Orius* accessory genome a Type VI secretion system (T6SS) associated protein was identified and annotated as an immunity protein belonging to the Tai4 protein family (Supplementary data 7-2). Considering the implication of T6SS in bacterial virulence and inter-bacterial competition, the T6SS-related genes in the genomes analysed were scrutinised (Chapter 8).

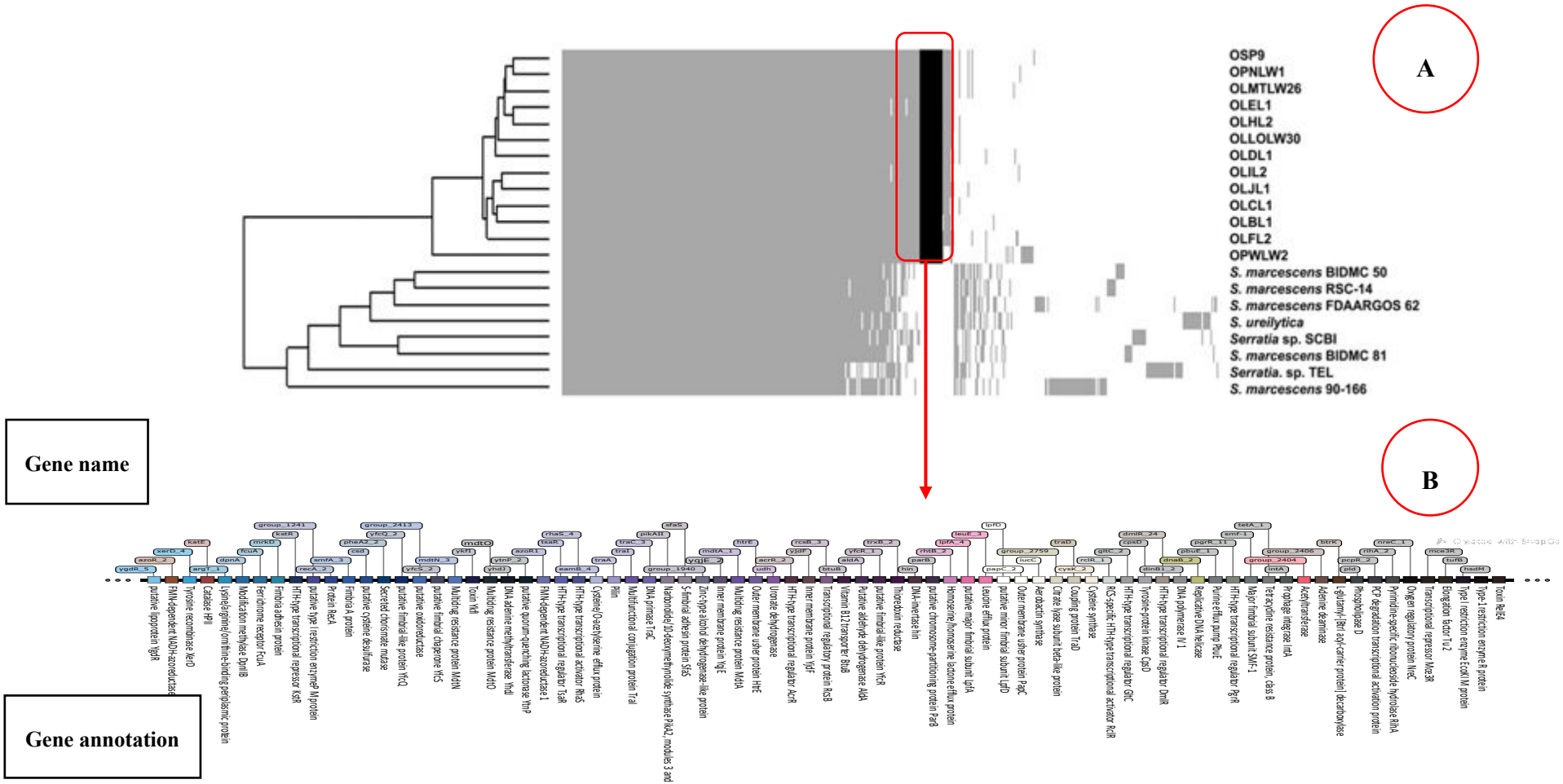


Figure 7-8: A representation of the pangenome gene presence/absence metrics displayed as a heatmap highlights the diversity within the SCBI complex.

(Grey shading confirms gene presence, white space confirms gene absence, and black shading highlights all *Orius* symbiotic-associated genes absent from all other members of the SCBI complex (Figure 5-1 in chapter 5). B illustrates the part of A black part of all known gene names and annotations, and the detail of these gene annotations were shown on Table 7-3 below.)

Table 7-3: Scoary output as list of significant genes per symbiotic trait of *Serratia* sp. *Orius* isolates

gene	Gene Function
ygdr_5	Regulation of antibiotic sensitivity (Yu et al., 2020)
azor_2	Exhibits azoreductase activity (Nishiya and Yamamoto, 2007)
xerd_4	Component of dif site and processing recombinatiom (Midonet and Barre, 2016)
kate	Protect cells from the toxic effects of hydrogen peroxide (Grzechowiak, Sekula, Jaskolski and Ruszkowski, 2021)
argt_1	Part of an ABC transporter involved in lysine, arginine, and ornithine transport (Hayashi et al., 2006)
dpna	Protects the DNA from cleavage by the DpnII endonuclease (Lacks, Ayalew, de la Campa and Greenberg, 2000)
fcua	Signaling receptor activity: Receptor for the hydroxamate siderophore, ferrichrome (Grim et al., 2012)
mrkd	Cell adhesion (Rêgo et al., 2012)
kstr	Controls a number of genes involved in cholesterol and fatty acid catabolism (Young et al., 2021)
group_1241	Site-specific DNA-methyltransferase (adenine-specific) activity (Quintieri et al., 2020)
reca_2	Cellular response to DNA damage stimulus (Shinohara et al., 2015)
smfa_3	Cell adhesion (Shinohara et al., 2015)
csd	Cysteine desulfurase activity to help the biosynthesis of selenoproteins (Esaki and Mihara, 2002)
phea2_2	Prephenate biosynthesis (Khanapur et al., 2017)
yfcq_2	Cell adhesion (Garbeva, van Elsas and de Boer, 2012)
group_2413	Lipid metabolic process: oxidoreductase activity (Marques-Pereira, Proença and Morais, 2020)
yfcs_2	Cell wall organization and chaperone-mediated protein folding (Garbeva, van Elsas and de Boer, 2012)
mdtn_3	Xenobiotic transmembrane transporter activity (Nordstedt and Jones, 2021)
ykfi	Toxic component of a type IV toxin-antitoxin (TA) system (Wen et al., 2017)
mdto	Involved in resistance to puromycin, acriflavine and tetraphenylarsonium chloride (Sulavik et al., 2001)
yhdj	DNA binding and N-methyltransferase activity (Rajagopala et al., 2014)
ytnp_2	The hydrolase activity as defence strategy to compete other bacteria (Schneider et al., 2011)
azor1	Azoreductase activity (Ryan et al., 2014)
tsar	Degradation of para-toluenesulfonate (TSA) as sole source of carbon and energy (Monferrer et al., 2010)
rhas_4	DNA-binding transcription factor activity (Fineran et al., 2013)
eamb_4	Amino acid transmembrane transporter activity: exporter of O-acetylserine (OAS) and cysteine (Hayashi et al., 2006)
traa	Conjugation related to antibiotic resistance (Cabezón et al., 2014)
trai	Conjugation related to antibiotic resistance (Ilangovan et al., 2017)
trac_3	Related to plasmid transfer during conjugation (Parker and Meyer, 2005)
pikaii	Involved in the biosynthesis of 12- and 14-membered ring macrolactone antibiotics (Zheng and Keatinge-Clay, 2011)
sfas	Cell adhesion: enable bacteria to colonize the epithelium of specific host organs (Babai, Stern, Hacker, and Ron, 2000)
group_1940	Zinc ion binding and oxidoreductase activity (Youn et al., 2006)
yqje_2	Integral component of membrane (Manzano-Marín and Latorre, 2014)

gene	Gene Function
mdta_1	Resistance to antibiotics, antimicrobial peptides, metals, detergents, and bile salts (Abi Khattar et al., 2019)
htre	Contribute to adhesion to various surfaces in specific environmental niches (Korea et al., 2010)
udh	Involved in the pathway D-galacturonate degradation and in Carbohydrate acid metabolism (Bouvier et al., 2014)
acrr_2	Potential regulator protein for the acrab genes. (Rajagopala et al., 2014)
yjdf	Inner membrane protein with six predicted transmembrane domains (Goodall et al., 2018)
rcsb_3	Component of the Rcs signaling system and binds to regulatory DNA regions (Castanié-Cornet et al., 2010)
btub	Derives its energy for transport by interacting with the trans-periplasmic membrane protein tonb (Pieńko and Trylska, 2020)
alda	Aldehyde dehydrogenase (NAD+) activity (Takeuchi et al., 2005)
yfcr_1	Cell adhesion (Garbeva, van Elsas and de Boer, 2012)
trxb_2	Essential thiol-reducing enzyme that protects the cell from thiol-specific oxidizing stress (Lin et al., 2016)
hin	DNA binding: the inversion of the flagellin controlling region (Koskiniemi et al., 2013)
parb	Involving in chromosome partition (Ogata et al., 2005)
rhtb_2	Amino acid transmembrane transporter activity (Tsu and Saier, 2015)
lpfa_4	Cell adhesion (Ferdous et al., 2016)
leue_3	Amino acid transport (Garbeva, van Elsas and de Boer, 2012)
lpfd	Cell adhesion (Garbeva, van Elsas and de Boer, 2012)
papc_2	Fimbrial usher porin activity and identical protein binding (Omattage et al., 2018)
iucc	Siderophore biosynthetic process and ligase activity (Sibanda and Ramganesh, 2021)
group_2759	Lyase activity and metal ion binding (Dimroth, 2004)
trad	Conjugative DNA transfer and related to type IV secretion system (Lu et al., 2008)
cysk_2	Cysteine synthase activity and hydrolase activity (Burke and Moran, 2011)
rclr_1	DNA-binding transcription factor activity and sequence-specific DNA binding (Garbeva, van Elsas and de Boer, 2012)
gltc_2	DNA-binding transcription factor activity (Garbeva, van Elsas and de Boer, 2012)
cpsd	Non-membrane spanning protein tyrosine kinase activity (Poehlein, Freese, Daniel and Simeonova, 2014)
dmlr_24	DNA-binding transcription factor activity (Garbeva, van Elsas and de Boer, 2012)
dinb1_2	DNA repair, DNA-directed DNA polymerase activity and magnesium ion binding (Poehlein, Freese, Daniel and Simeonova, 2014)
dnab_2	Chromosome replication (Poehlein, Freese, Daniel and Simeonova, 2014)
pbue_1	Transmembrane transporter activity (Garbeva, van Elsas and de Boer, 2012)
pgrr_11	DNA-binding transcription factor activity (Garbeva, van Elsas and de Boer, 2012)
smf-1	Involved in adherence to eukaryotic epithelial cells and abiotic surfaces. (de oliveira-garcia et al., 2003)
teta_1	Transmembrane transporter activity (Poehlein, Freese, Daniel and Simeonova, 2014)
inta	Prophage integrase activity and DNA binding (Fan et al., 2020)
group_2404	Regulation of cell shape and peptidoglycan biosynthetic process (Lamelas et al., 2011)
group_2406	Plays an important role in the purine salvage pathway and in nitrogen catabolism (Kamat et al., 2011)
btrk	Involved in the pathway butirosin biosynthesis, which is part of Antibiotic biosynthesis (Li et al., 2005)
pld	Phospholipase D activity and functions in the lipid metabolism (Jenkins and Frohman, 2005)

gene	Gene Function
pcpr_2	DNA-binding transcription factor activity (Garbeva, van Elsas and de Boer, 2012)
riha_2	Hydrolase activity, hydrolysing N-glycosyl compounds (Petersen and Møller, 2001)
nrec_1	DNA-binding (Garbeva, van Elsas and de Boer, 2012)
mce3r	Represses the transcription of mce3 operon in lipid metabolism (Santangelo et al., 2009)
tufb	Translation elongation factor activity (Burke and Moran, 2011)
hsdm	DNA-binding (Kennaway et al., 2008)
hsdr	Type I site-specific deoxyribonuclease activity (Loenen et al., 2013)
rele4	Toxic component of a type II toxin-antitoxin (TA) system (Fiebig, et al., 2010)

7.4.3 KEGG and COGs distribution of Core, Accessory and Unique genes by BPGA

A search for core, accessory and unique gene families were conducted to compare the distribution of functional categories COGs database through BPGA. Table 7-4 and Figure 7-9 shows the differential distribution of COGs functional categories in core, accessory, and unique gene families. The most common functions (44%) in the core genomes are associated with metabolism (Figure 7-9). Category E (Amino acid transport and metabolism) was the most enriched (15%) metabolic function. Meanwhile, category J (Translation, ribosomal structure, and biogenesis) belonging to cellular processing and signalling functions showed different degree of enrichment (5%) with category E in the core genomes. According to the result of the COGs distribution, many genes belonging to the core group were related to housekeeping functions. Additionally focused on the accessory genome in category E and category J, category J genes were more conserved in *Serratia* sp. *Orius* isolates (2.00% of that accessory genome), while category E genes comprised 5% of that group. It was suggested that class E genes might suggest the different abilities depending on the species of *Serratia*. Likewise, when comparing the COGs groups for metabolism, the percentage of category P (Inorganic ion transport and metabolism) genes relatively conserved and were found in higher fractions (8%) of core genome versus the accessory genome (4%). On the contrary, category C (Energy production and conversion) and category H (Coenzyme transport and metabolism) genes were variable in *Serratia* species, which were 3.0% and 4% in the core genome versus 3% and 1% in the accessory genome.

About 20% of the core genome content was grouped under category R (General function prediction only) and class S (Function unknown) having poorly characterized function. Likewise, among the genes from the accessory genome and unique genes, approximately 26.7–27.8% of the total gene content was grouped under the COGs same classes, with no specific function assigned to these genes (Table 7-4). The *Serratia* species have potential pathways or abilities not yet estimated by the present COGs categories. Pan-genome and COG analyses showed that the majority of the conserved core genes are involved in basic cellular functions, while genomic factors such as prophages contribute considerably to genome diversity.

Table 7-4 COGs detail distribution of core, accessory, and unique genes.

COGs FUNCTION CATEGORIES	CORE	ACCESSORY	UNIQUE
[S] function unknown	8.517924963	21.27659574	9.626955475
[J] translation, ribosomal structure, and biogenesis	3.863801014	12.95097132	1.203369434
[R] general function prediction only	13.34767623	10.17576318	14.44043321
[L] replication, recombination, and repair	3.183245154	9.250693802	19.25391095
[K] transcription	9.549735461	7.400555042	8.423586041
[E] amino acid transport and metabolism	11.13038133	6.475485661	6.016847172
[G] carbohydrate transport and metabolism	7.793462273	6.475485661	4.813477738
[U] intracellular trafficking, secretion, and vesicular transport	2.634409782	4.625346901	8.423586041
[P] inorganic ion transport and metabolism	7.266580317	4.625346901	3.610108303
[I] lipid transport and metabolism	2.700270027	3.700277521	1.203369434
[Q] secondary metabolites biosynthesis, transport, and catabolism	2.656363197	2.775208141	1.203369434
[M] cell wall/membrane/envelope biogenesis	5.268819565	2.775208141	0
[T] signal transduction mechanisms	3.929661259	1.85013876	4.813477738
[V] defense mechanisms	1.163530987	1.85013876	4.813477738
[N] cell motility	2.173388071	1.85013876	3.610108303
[C] energy production and conversion	5.115145661	0.92506938	2.406738869
[F] nucleotide transport and metabolism	2.019714167	0.92506938	1.203369434
[O] post-translational modification, protein turnover, and chaperones	3.336919058	0	2.406738869
[H] coenzyme transport and metabolism	3.578406621	0	1.203369434
[D] cell cycle control, cell division, chromosome partitioning	0.76836952	0	1.203369434

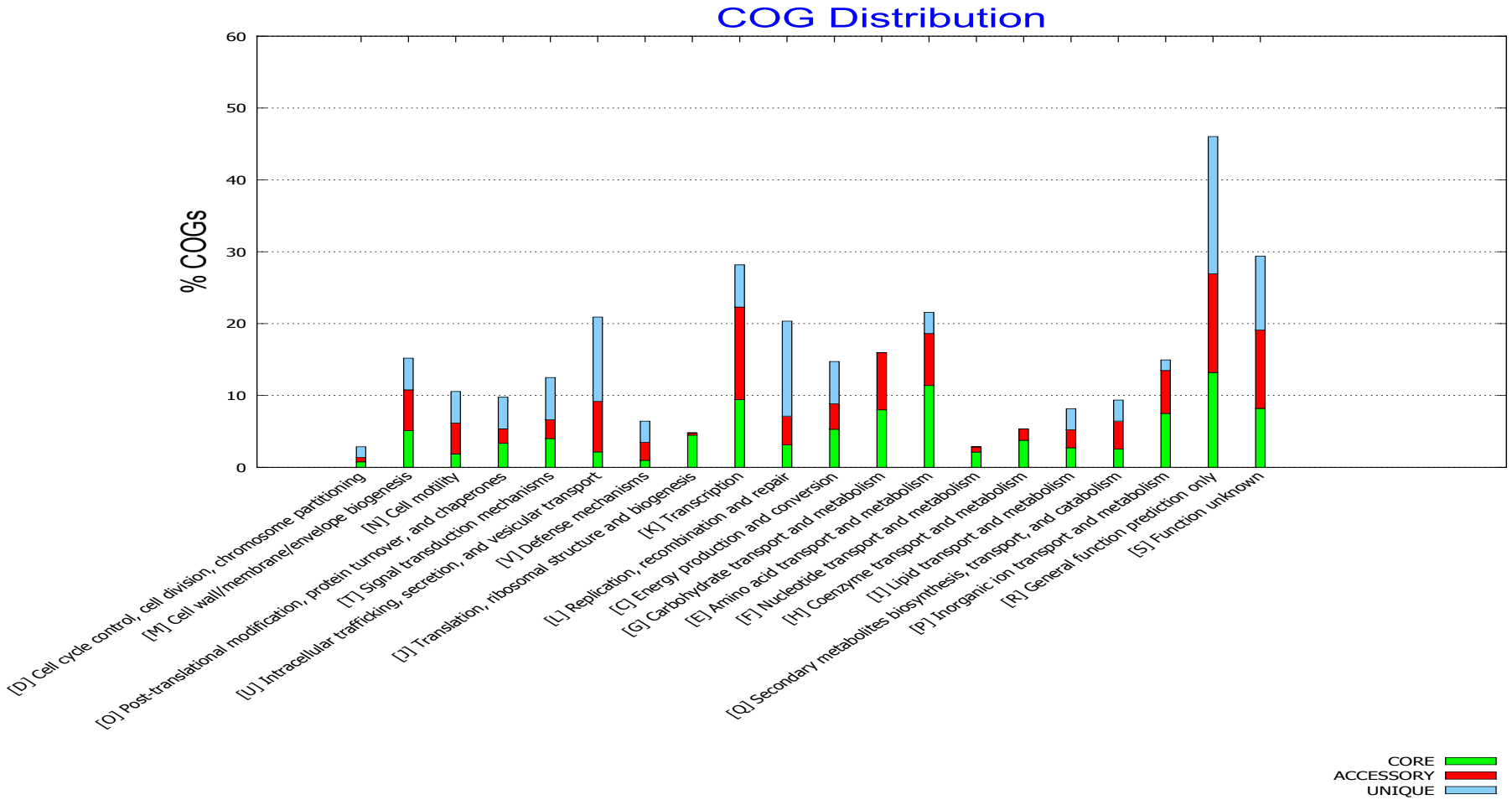


Figure 7-9: COGs distribution of core, accessory, and unique genes.

(Classes D, M, N, O, T, U and V belonging to the category of Cellular processing and signalling. Classes J, K, and L belonging to the category of Information storage and processing. Classes C, G, E, F, H, I, Q and P belonging to the category of Metabolism. Classes R and S belonging to the category of Poorly characterize.)

The pan-genome functional analysis module of BPGA was also used for KEGG pathway mapping of representative protein sequences of core, accessory genomes, and unique genes of the pangenome in this study. (Supplementary data Table 7-3 listed all countable KEGG pathways with the KEGG major and sub-categories in the pangenome where at least one gene was detected). According to the Supplementary data information, it was suggested that various pathways might be conserved in the *Serratia* sp. *Orius* isolates to adapt to the natural environment. In addition, these pathways might vary by accessory and unique genes. KEGG assignments from BPGA revealed overall higher representation of metabolism related pathways (Figure 7-10). This result also strongly supported the result of the COGs distribution regarding metabolic function. amino acids, carbohydrate metabolism and membrane transport were specifically enriched (more than 10% with enrichment significance) suggesting that these three KEGG pathways were more conserved in the *Serratia* species. However, due to the high percentage of hypothetical proteins in the pangenome, so the extra potential pathways still need to analysis further, after find out the role of hypothetical proteins in the future work.

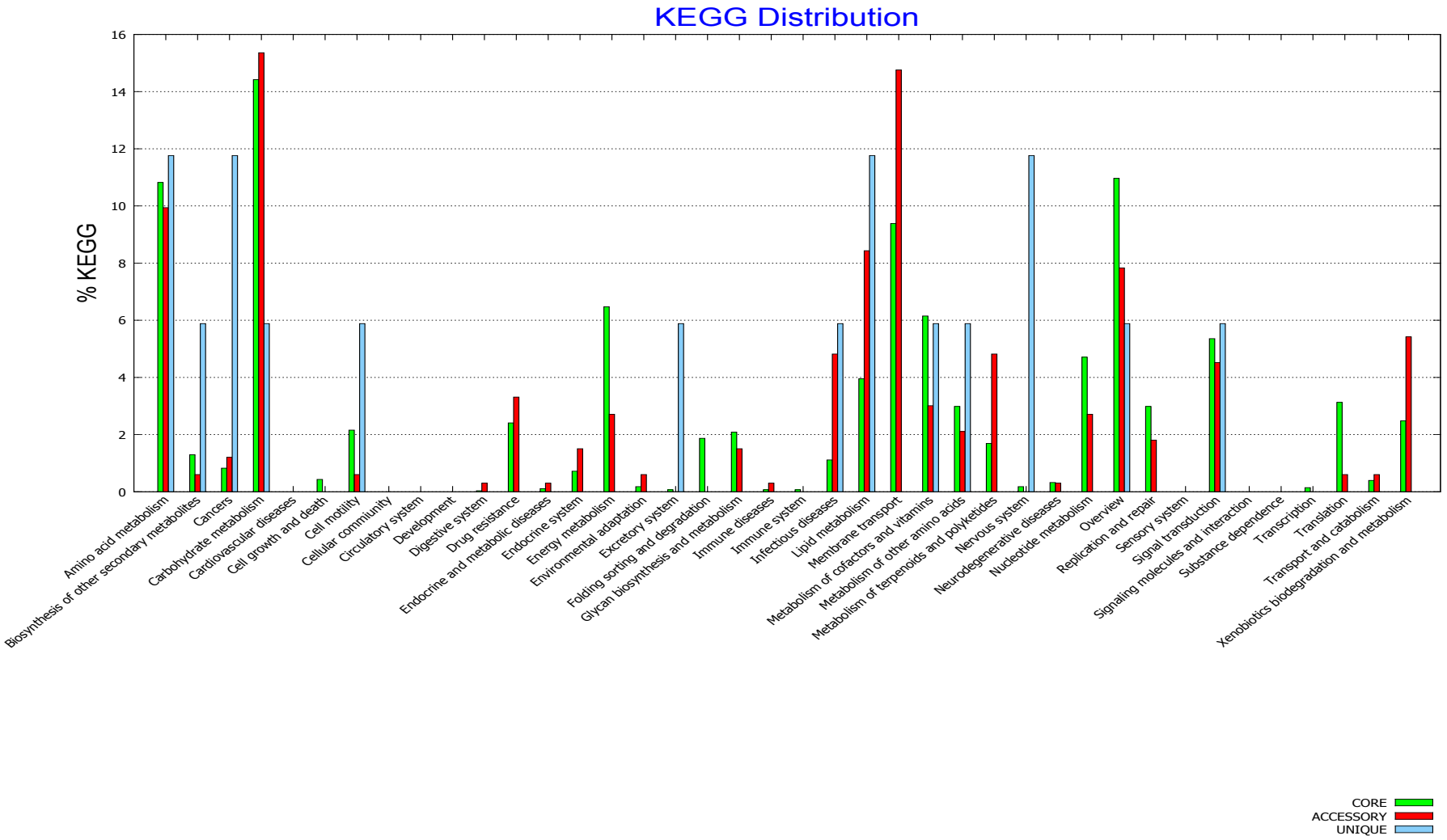


Figure 7-10: KEGG distribution of the representative proteins in the core, accessory, and unique genome.

7.4.4 Annotated *Serratia* sp. *Orius* isolates accessory genome reveals the presence of plasmid associated features.

All 21 genomes of '*Serratia* sp. SCBI complex' (Figure 5-1 in chapter 5) described here most likely contained putative plasmid sequences as part of the assemblies, since they are similar to *Serratia* sp. SCBI and this organism contains one plasmid called plasmid SCBI_PI (Accession number: CP003425.1), but other reference genomes within the '*Serratia* sp. SCBI complex' and these draft genomes were not segregated into replicons. Therefore, the construction of the pangenome included plasmid sequences as part of the draft genome sequence.

The results suggested that these plasmid sequences contained high sequence homology to the plasmids SCBI_P1 from *Serratia* sp. SCBI, plasmid PSM22 from the strain *S. marcescens* B-6493, and PWN146p1 from the strain *S. marcescens* PWN146 in Table 7-5. A Mauve alignment of the available plasmids in NCBI and OSPLW9 plasmid sequence is shown in Figure 7-11. This explains the presence of some plasmid genes in the *Orius* accessory genes. *S. marcescens strain* B-6493 is a human pathogen and *S. marcescens* PWN146 was isolated from the nematode *Bursaphelenchus xylophilus* and both are *Serratia* species distinct from those in the SCBI complex as confirmed by GGDC. Since there is a lack of replicon separation within available *Serratia* species genomes from the NCBI database, it is difficult to expand on plasmid sequence comparisons at the current stage to define a plasmid genealogy for the genus. However, these findings suggest that plasmid exchange take place across the *Serratia* genus and is worthy of future studies.

Table 7-5: BLAST alignments of *Serratia* sp. *Orius* plasmid related contigs

Reference plasmid sequence name	Query Orius plasmid sequences	Total score	Query cover	Identity
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OSPLW9_Plasmidslcl|Query_127032 NODE_21_length_59884_cov_29.7327_ID_41	71492	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OPWLW2_Plasmids_lcl|Query_48248 NODE_22_length_59884_cov_36.9523_ID_43	71678	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OPNLW1_Plasmids_lcl|Query_29577 NODE_24_length_59773_cov_27.5704_ID_47	71492	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLMTLW26_Plasmids_lcl|Query_65718 NODE_26_length_59884_cov_27.5358_ID_51	71919	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLLOLW30_Plasmids_lcl|Query_42268 NODE_32_length_59774_cov_48.962_ID_63	71492	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLJL1_Plasmids_lcl|Query_57525 NODE_45_length_39346_cov_15.489_ID_89	56341	53%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLHL1_Plasmids_lcl|Query_98442 NODE_25_length_59739_cov_41.1314_ID_49	71492	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLFL2_Plasmid_lcl|Query_197525 NODE_31_length_59884_cov_23.6257_ID_61	71722	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLEL1_Plasmidlcl|Query_243724 NODE_32_length_41021_cov_19.0231_ID_63	56341	53%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLDL1_Plasmid_lcl|Query_145855 NODE_32_length_59766_cov_30.9586_ID_63	71492	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLCL1_Plasmid_lcl|Query_40266 NODE_35_length_59770_cov_18.8429_ID_69	71492	74%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLBL1_Plasmid_lcl|Query_161550 NODE_46_length_39040_cov_11.9416_ID_91	56335	53%	99%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLJL1_Plasmids_lcl|Query_57550 NODE_66_length_19819_cov_16.7846_ID_131	15151	21%	94%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLIL1_Plasmids_lcl|Query_218301 NODE_67_length_18742_cov_15.8063_ID_133	15151	21%	94%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLEL1_Plasmidlcl|Query_243745 NODE_39_length_18762_cov_21.8539_ID_77	15151	21%	94%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLBL1_Plasmid_lcl|Query_161585 NODE_80_length_18729_cov_14.1741_ID_159	15151	21%	94%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLIL1_Plasmids_lcl|Query_218300 NODE_60_length_22805_cov_14.5215_ID_119	29498	30%	98%
CP003425.1 Serratia sp. SCBI plasmid SCBI_Pl	OLIL1_Plasmids_lcl|Query_218327 NODE_72_length_16792_cov_14.0056_ID_143	26745	24%	97%

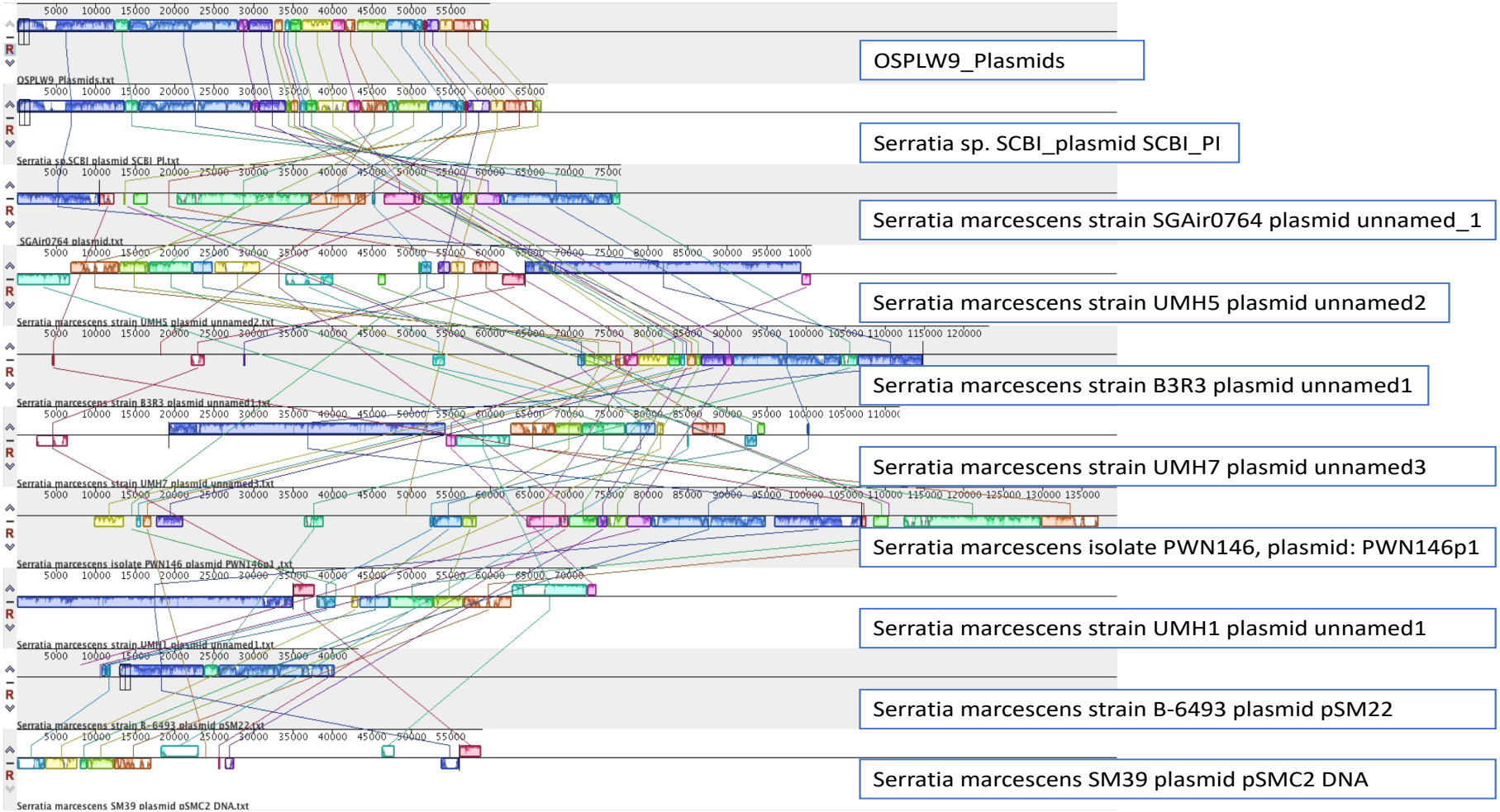


Figure 7-11: A Mauve progressive alignment of OPSLW9 plasmids related single contigs and various *Serratia* species plasmids sequences

7.5 Discussion and Conclusion

For understanding the evolutionary processes undergone by *Serratia* sp. *Orius* isolates during their adaptation from a free-living state to a facultative endosymbiotic lifestyle, an 'open' pangenome was constructed from the genomes within the *Serratia* sp. SCBI complex (Figure 5-1) in chapter 5. General features of each genome, together with the two strains from insect and nematode pathogens (*Serratia* sp. TEL and SCBI), three strains associated with plants (*Serratia marcescens* 90-166, *Serratia ureilytica*, and *Serratia* sp. RSC-14), and three human pathogens (*Serratia marcescens* BIDMC 50, *Serratia marcescens* BIDMC 81 and *Serratia marcescens strain* FDAARGOS_62) were recovered from their respective sources.

This 'open' pangenome of *Serratia* sp. *Orius* isolates indicated that the symbiotic lifestyles of these strains would require a vast available gene reservoir to help them adapt to the niches within different environments by expanding their accessory and pangenome through different means of lateral gene transfer. However, annotation of the accessory genomes failed to reach a definite conclusion regarding its role in supporting the symbiotic lifestyle, due to the presence of many hypothetical proteins which are the proteins unknown functions. A previous Scoary pan-GWA study identified 279 accessory genes in accessory genomes which are associated with the insect symbiosis trait. In future, further gene annotations will be processed using different tools to increase understanding of the functional roles of these genes in the symbiotic lifestyle and also to identify the roles of these hypothetical proteins. Additionally, plasmids were found in the *Serratia* sp. *Orius* isolates and were confirmed by mauve alignments of the plasmids of various *Serratia* species and one representative isolate plasmid. This result indicates that plasmids can be exchanged between various *Serratia* species, due to the high similarity of plasmid alignments, in particular human pathogens, so that virulence factors may be transferred between different *Serratia* species, an area which will be worthy of future study.

CHAPTER 8: Prediction of type 6 secretion systems (T6SS) encoded by *Serratia* sp. *Orius* isolates

8.1 Abstract in this chapter

- Due to the discovery of T6SS in reference genome of *Serratia* sp. SCBI which is available on SecReT6 database (https://bioinfo-mml.sjtu.edu.cn/SecReT6/), T6SS of *Serratia* sp. *Orius* isolates are identified and analysed in this chapter.
- Two different T6SS loci found out in all the isolates.
- The classification of T6SS subtype is i3 in these strains and it is associated with interbacterial competition.

8.2 Introduction

In the previous chapter, within the accessory genome of *Serratia* sp. isolates pangenome, a T6SS-associated protein was identified and annotated as an immunity protein belonging to the Tai4 protein family. Considering the implication of T6SS in bacterial virulence and inter-bacterial competition, the T6SS-related gene in the genomes of *Serratia* sp. isolates analysed were scrutinised (Supplementary data Table 7-2). In recent studies of bacterial secretion systems, T6SS has been shown to be used by bacteria to attack bacterial competitors, or to defeat the host defence mechanisms. in order to colonize a host niche and/or survive in competition. Particularly, some studies demonstrated that T6SSs in several *Serratia* species can target other bacterial competitors resulting in either growth inhibition or death. (Li et al., 2015). This chapter aims at identifying T6SS encoding loci in all the *Serratia* sp. *Orius* isolates to predict their ability to establish interbacterial competition.

The microbiome of insects is composed of complex communities comprising bacteria, fungi, and viruses. Within these communities, the microorganisms commonly compete for limited resources and space. These drivers have forced the co-development of specialized collaboration and competitive mechanisms. Especially, pathogens and symbiotic bacteria employ several strategies to attack competitors. One key strategy is using specialised secretion systems to delivery functional proteins termed effectors (Allsopp et al., 2019). Type VI secretion systems

(T6SS) are highly conserved mechanisms that directly penetrates effector proteins into the target cell and are widely found in 25% of Gram-negative bacteria (Bingle, Bailey, & Pallen, 2008). T6SS is a complete secretion apparatus within the membrane that delivers toxic effectors to eukaryotic and prokaryotic cells in a contact-dependent manner, with effector cell wall degrading enzymes, cell membrane targeting proteins, and nucleases ((Alcoforado Diniz and Coulthurst, 2015).

The T6SS apparatus is assembled into a double-membrane-spanning structure in three major complexes (the membrane complex, the tail complex, and the baseplate complex) with a minimal set of 13 core subunits named *tss*A-M (several units have alternative names) (Navarro-Garcia et al., 2019). It also contains several 'accessory' proteins which might be essential to T6SS assembly or regulation in different system, and these proteins are important to facilitate the diversity of the T6SSs (Cianfanelli et al., 2016). The minimal membrane complex of T6SS is assembled by three discrete multiprotein subunits (*tss*J, *tss*L and *tss*M). *tss*B and *tss*C form the tail complex of T6SS. *Tss*AFGK are the key constituent of baseplate complex (Navarro-Garcia et al., 2019). The list of all core and accessory T6SS gene components found in the T6SS gene clusters, including their putative functions and COGs (Clusters of Orthologous Groups of proteins) classification is presented in Table 8-1.

Table 8-1 : The conserved T6SS components (Cianfanelli et al., 2016 ; Shyntum et al., 2014)

Tss/Tag	Alternative	COG	Location/Role	Related Proteins
*tss*A		COG3515	Baseplate	
*tss*B	*Vip*A	COG3516	Contractile sheath/tail	gp18 (TssBC)
*tss*C	*Vip*B	COG3517	Contractile sheath/tail	gp18 (TssBC)
*tss*D	*Hcp*	COG3157	Expelled tube	gp19
*tss*E	*HsiF*	COG3518	Baseplate	gp25
*tss*F	*VasA*	COG3519	Baseplate	
*tss*G		COG3520	Baseplate	
*tss*H	*ClpV*	COG0542	Sheath recycling	AAA+ ATPases
*tss*I	*VgrG*	COG3501	Expelled spike	gp27/gp5
*tss*J	*Lip, SciN*	COG3521	Membrane complex	
*tss*K	*impJ, vasE*	COG3522	Baseplate	
*tss*L	*IcmH/DotU, VasF*	COG3455	Membrane complex	IcmH (T4bSS)
*tss*M	*IcmF, VasK*	COG3523	Membrane complex	IcmF (T4bSS)
*tag*D	PAAR (Proline-alanine-alanine-arginine repeat protein)	COG4104	Tip of expelled spike	gp5.4
*tag*E	*pknA/ppkA*	COG0515	Serine/threonine kinase, post-translational regulation	
*tag*F		COG3913	Accessory protein: post-translational regulation	
*tag*J	*HsiE*	COG4455	Accessory protein: sheath recycling	
*tag*L	*SciZ*		Accessory protein: cell wall anchoring	
*ppk*A		COG0515	Accessory protein: post-translational regulation	

pppA	*tag*G	COG0631	Accessory protein: post-translational regulation	
Fha	*tag*H	COG3456	Accessory protein: post-translational regulation	

The 'gp' proteins are from bacteriophage T4; IcmF and IcmH are components of the Type IVb secretion system (T4bSS).

Toxic effectors secreted by the type VI secretion system (T6SS) facilitate interbacterial warfare, as well as pathogenesis toward humans, animals, and plants. Adaptor proteins are mediators that help to load their cognate effectors onto the T6SS spike complex (Navarro-Garcia et al., 2019). The contextual genes of the known adaptor proteins (DUF1795, DUF2169 or DUF4123) all exhibited a high proportion of encoding T6SS spike complex protein (*tss*I or PAAR) and effector proteins. Due to Serine/threonine protein kinase (STPK) in *Pseudomonas aeruginosa* provide virulence, PRK06147 might be a novel adaptor protein which was discovered in a recent study (Liu et al., 2020). In the T6SS, the *tss*I protein (in a trimer form) and the conical PAAR protein form the spike complex, which is located at the top of the T6SS secretion structure and responsible for creating an opening in the target cell envelope (Brackmann et al., 2013). Toxic effectors can be fused to *tss*I or PAAR as an extension domain or as an independent protein loaded onto the secretory component through protein–protein interactions (Silverman et al., 2012). Therefore, the *tss*I and PAAR proteins, which have an extension domain at the C-terminus, are promising effector candidates. In addition, the N-terminal domain of the multidomain effector proteins may have specific motifs, such as rearrangement hotspots (RHS), YD repeats, MIX motifs and FIX motifs, which can serve as markers for unknown T6SS effectors. However, many T6SS effectors do not contain the above-mentioned identification characteristics, especially single-domain proteins. These effectors require additional assistance to load onto the T6SS, and this assistance is provided by proteins known as adaptors or chaperones (Unterweger et al., 2017). the adaptors assist with loading effectors onto the T6SS spike complex, and they are not secreted by the T6SS. In species such as *Pseudomonas aeruginosa* and *Serratia marcescens*, the DUF1795 protein is used as an adaptor to bind to the effector protein, and it is delivered with the effector to the binding site of *tss*I during the assembly process of the T6SS. In *Agrobacterium tumefaciens*, adaptors characterized by the conserved DUF2169 domain or DUF4123 domain have been identified, and they are necessary for T6SS effector secretion. Moreover, DUF4123 has also been identified as an adaptor in species such as *Vibrio cholerae*, *P. aeruginosa* and *Aeromonas hydrophila* (Liu et al., 2020).

The classification of T6SS gene clusters were recently divided in four sub-types. (i) the majority of T6SSs belong to subtype T6SS^{i} and are present in Proteobacteria, T6SS^{i} was also classified into another six distinct subtypes T6SS-i1, i2, i3, i4a, i4b, and i5 based on the organization and phylogenetic relationship of its core components; (ii) the *Francisella* pathogenicity island-like systems were classified as T6SS^{ii} (Broms et al., 2010); (iii) Bacteroidetes T6SSs are distinct from the first two and were classified as T6SS^{iii} (Russell et al., 2014); and (iv) a contractile system from *Amoebophilus asiaticus* was classified T6SS^{iv} (Bock et al., 2017). The functions of these T6SS subtypes are all related to interbacterial antagonism, interbacterial competition and metal ion acquisition (Lennings, West and Schwarz, 2019).

Therefore, the identification and classification of T6SS components and effectors in these sequences is required to support future experimental work to confirm their contribution to interbacterial antagonism within the host.

8.3 Method

In this chapter, each draft genomes from *Serratia* sp. *Orius* Isolates were combined into pseudogenomes which is a virtual, artificial concatemer of all assembled contigs from sequencing a particular airtificial genome, using combining contigs website (https://www.bioinformatics.org/sms2/combine_fasta.html), which combined all the contigs from a genome to a single contig and then annotated by Prokka with GenBank formats. SecReT6 (https://bioinfo-mml.sjtu.edu.cn/SecReT6/) was used for the detection of T6SS component genes and effectors, and classification of T6SS subtypes in these genomes. It is a high reliable web-based resource that currently provides 11167 core T6SS components mapping to 906 T6SSs detected in 498 bacteria strains representing 240 species, it also covered over 600 directly related references. The SecReT6 database also contains 1340 T6SS candidate effectors and 196 immunity proteins (Li, et al., 2015). Furthermore, BLASTP was used to identify genes with unknown function in SecReT6 detections for analysis of T6SS adaptors and effectors. Standard T6SS gene nomenclature (Shalom et al., 2007) were used in this study. According to this nomenclature, the conserved T6SS genes were designated *tss*A-M/ (Type Six

Secretion A-M), while the accessory or non-conserved T6SS genes were designated *tag*A-P (Type Six Associated Genes A-P) (Shalom et al., 2007).

8.4 Results

8.4.1 In silico identification of T6SS gene clusters in draft genomes from *Serratia* sp. *Orius* Isolates

Homologs of the T6SS genes were clustered in 4 distinct genomic regions in all sequenced strains from *Serratia sp. Orius* Isolates. These regions were designated *Serratia* sp. *Orius* isolates T6SS locus 1, 2 (SS T6SS-1, SS T6SS-2). All the T6SS regions contained all 13 core gene components. Both SS T6SS-1 and SS T6SS-2 gene clusters were located on the genome of all 13 sequenced strains of *Serratia* sp. *Orius* isolates. The overall genetic organisation of each T6SS of *Serratia* sp. *Orius* isolate is presented in Figure 8-1.

8.4.2 Operon structure of the T6SS

The T6SS encoded by most bacteria is organized in discreet transcriptional units or operons, suggesting coordinated expression (Bingle et al., 2008; Boyer et al., 2009; Dandekar et al., 1998; Williams et al., 1992). Therefore, the organizations of conserved genes in *Serratia* sp. T6SS have been investigated. The core genes of SS T6SS-1 were clustered in three highly conserved operons; group 1 (*tss*J-*tss*K-*tss*L-*tss*M) group 2 (*tss*A-*tss*B-*tss*C) and group 3 (*tss*E-*tss*F-*tss*G-*tss*H). SS T6SS-2 showed a considerable level of potential gene shuffling compared to SS T6SS-1, with gene order being highly variable between each of the different groupings. The consensus grouping in SS T6SS-2 included *tss*A-*tss*B-*tss*C-*tss*D-*tss*E-*tss*F-*tss*G and *tss*J-*tss*K-*tss*L-*tss*M, while t*ss*H-*tss*I was stand-alone operons linked to non- conserved T6SS genes. These different operon structures suggested the independent acquisition of the T6SS clusters, each of which may play a different function in the biology of *Serratia* sp. *Orius* isolates in this study.

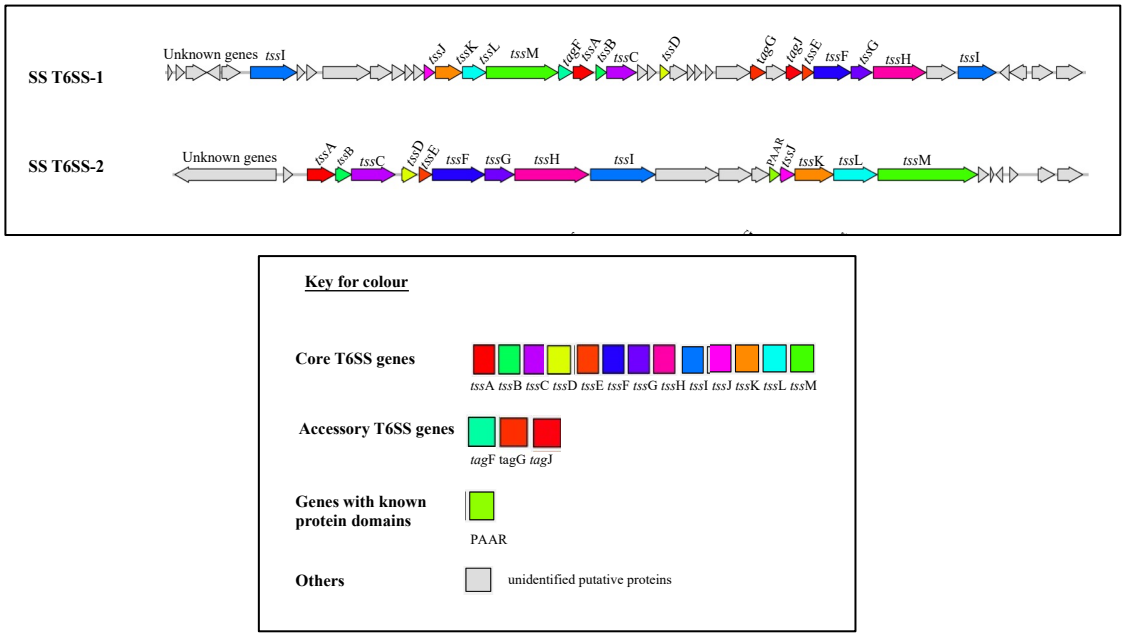


Figure 8-1: Genetic organization of the different T6SS gene clusters in *Serratia* sp. *Orius* isolates.

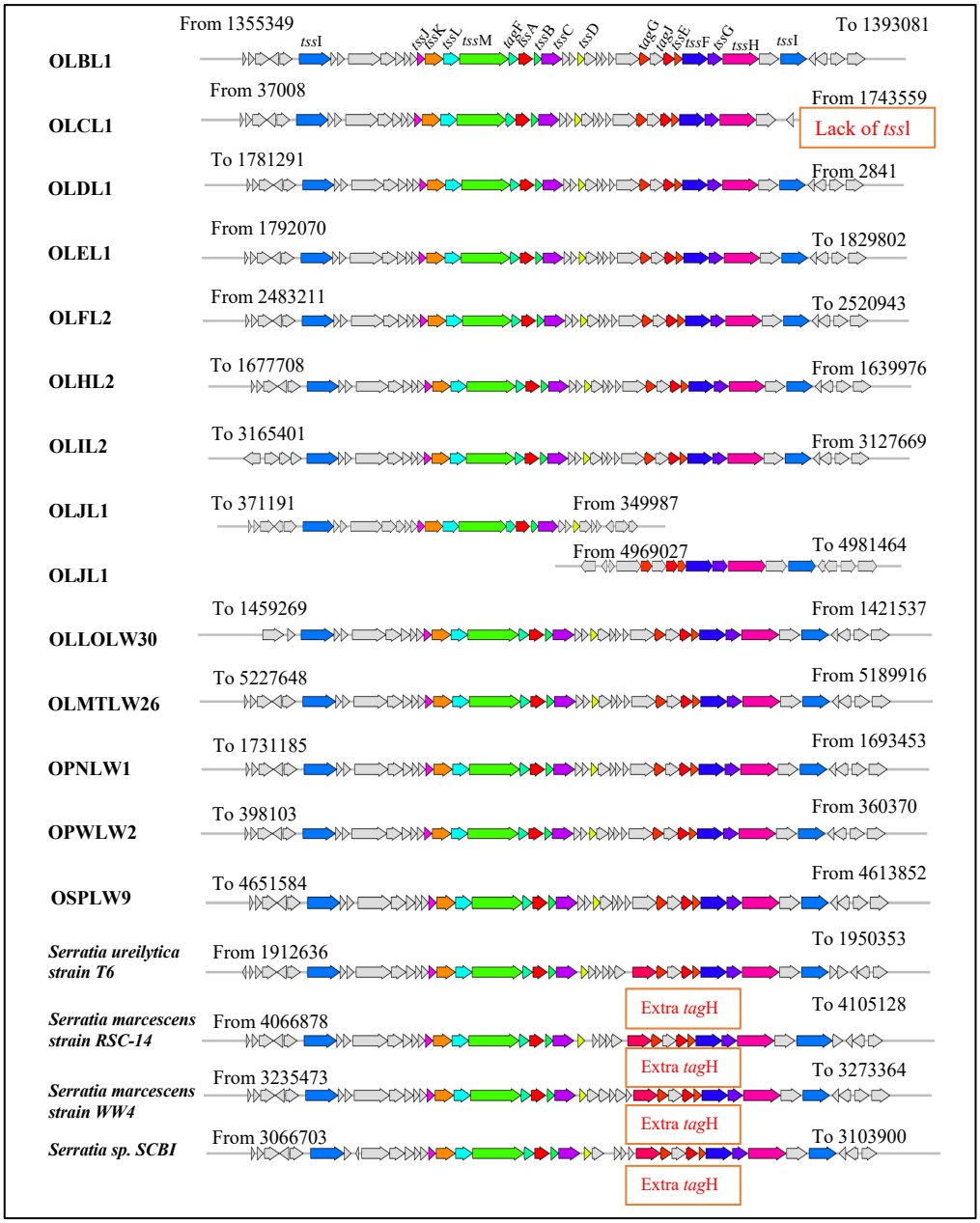
(Genes are indicated by arrows and the direction of the arrows represented the direction of transcription of the gene related to the rest of the genome. Conserved gene components of the T6SS (*tss*A-M) are indicated in different colour which represented in first line of key for colour. Non-conserved genes associated with T6SS of limited bacteria (*tag*A-P) are also indicated in distinct colours which showed in second line of key for colour. Grey colour means unidentified putative proteins which could not identified by SecReT6.)

8.4.3 Comparative analysis of T6SS gene clusters from all *Serratia* sp. *Orius* isolates

Homologous SS T6SSs encoded by different strains of *Serratia* sp. *Orius* isolates were highly conserved in terms of sequence similarity, gene content and operon structure (Figure 8-2). A detailed description of the genes found in individual T6SSs encoded by all sequenced *Serratia* sp. *Orius* isolates analysed in this study are provided in supplementary data.

The genetic architecture of SS T6SS-1 was shown to be conserved amongst all *Serratia* sp. *Orius*. The *tss*D and *tss*I genes found in this cluster encoded *Tss*D and *Tss*I proteins that do not have C-terminal extensions as found in "evolved" *tss*D and *tss*I proteins (Blondel et al., 2009; Pukatzki et al., 2007; Suarez et al., 2010). The C-terminal extension of some evolved *tss*I

proteins, such as *tss*I of *V. cholerae* and *Aeromonas hydrophila*, have been associated with actin cross-linking and actin ADP ribosylation activity in mammalian host cells, respectively (Pukatzki et al., 2007; Suarez et al., 2010). Only strain OLCL1 possess a single *tss*I gene, while other strains had an additional copy of *tss*I. It is possible, therefore, that the different *tss*I proteins encoded by each *tss*I gene are mobilized to the T6SS baseplate under different physiological conditions or play different roles either as effectors, structural elements, or both. Regions associated with *tss*D and *tss*I contain genes that encode a variable number of accessory and hypothetical proteins that account for strain specific differences. The first interesting region in SS T6SS-1 is located between the tssJ and *tss*I. Genes found in this region encode mostly hypothetical proteins and proteins with a PAAR repeat or pentapeptide_4 domains. PAAR repeat proteins of several bacteria have effector domains on the N or C-terminal terminus. Some of these effector domains include transthyretin, lipase, nuclease, deaminase, and ADP-ribosyl transferase (Vocadlo and Withers, 2005). A recent study showed that the PAAR repeat proteins of *E. coli* and *V. cholerae* bind to the Gp5- *tss*I complex by means of non-covalent interactions. In addition, PAAR repeat proteins of *V. cholerae* and *Acinetobacter baylyi* were shown to be bactericidal effectors associated with T6SS-mediated killing of *E. coli*. These findings have led to the speculation that PAAR repeat proteins carrying different effector domain located on either their N or C-terminal extensions may also bind to the *tss*I spike and mediate secretion of these effectors by the T6SS (Shneider et al., 2013). It is also speculated that PAAR repeat proteins may form non-covalent interactions with different effectors, thereby recruiting them to the T6SS spike complex. Therefore, the PAAR repeat proteins encoded by genes located in the *tss*I island of SS T6SS-1 gene cluster may either be T6SS effectors associated with inter-bacterial competition or may mediate secretion of other effectors. OLJL2 have two separate gene clusters of T6SS locus in SS T6SS-1, due to the gene components of SS T6SS-1 presented in two separate contigs of OLJL2. Four reference genomes (*Serratia ureilytica strain* T6*, Serratia marcescens strain* RSC-14, *Serratia marcescens strain* WW4 and *Serratia* sp. SCBI) contain an extra *tag*H gene in the locus.



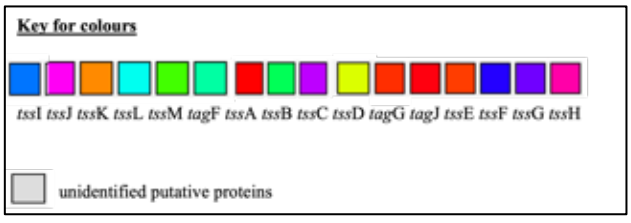
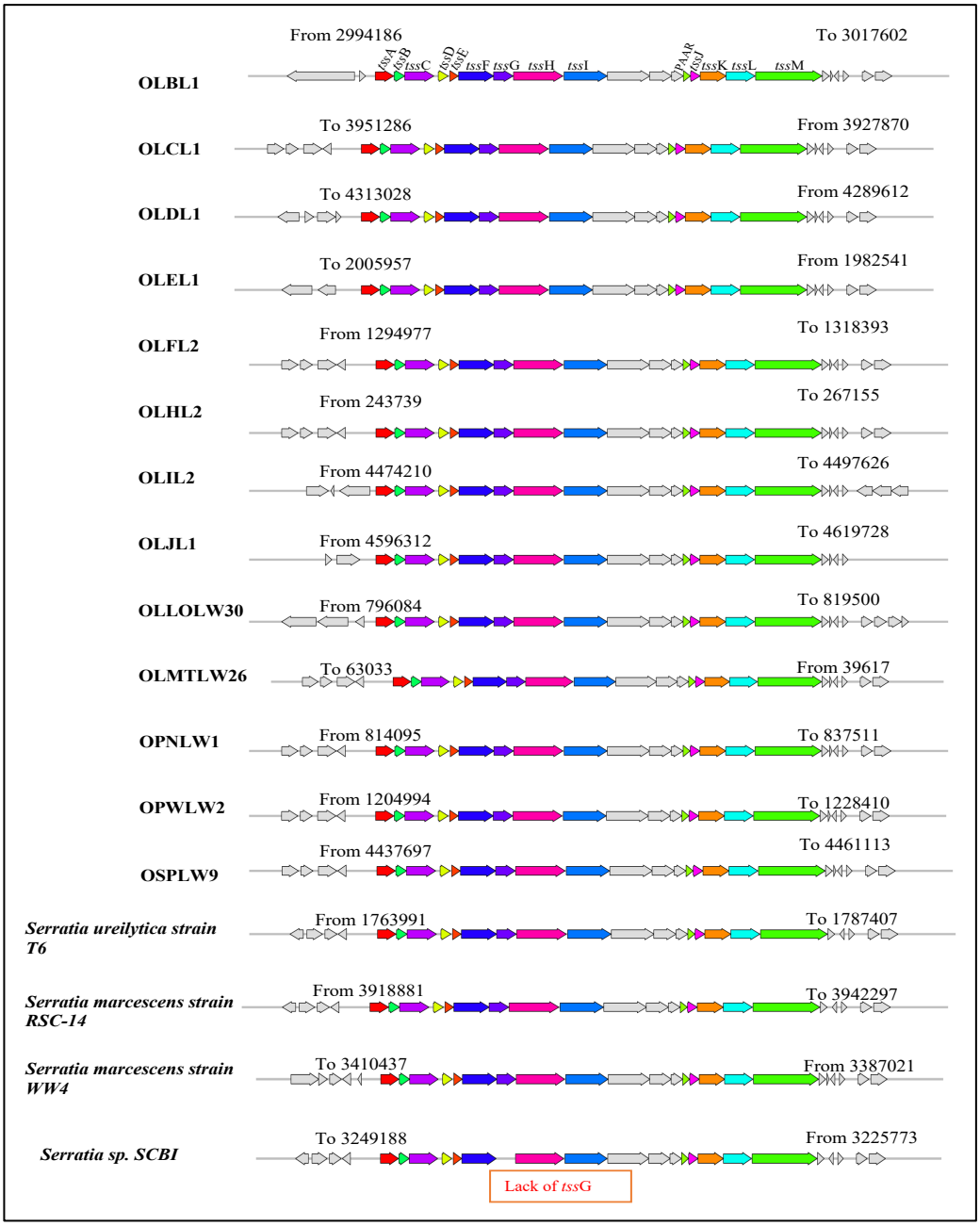


Figure 8-2: Comparison of all the sequenced strains from all *Serratia* sp. *Orius* isolates T6SS in the regions of SS T6SS-1.

(All core genes and non-conserved genes components of the T6SS are indicated in different colours showing on above key for colours. Grey colour key is unidentified putative proteins by SecReT6. SS T6SS-1 was found in *Serratia* sp. *Orius* isolates analysed, while four reference genomes have an extra *tag*H gene and OLCL1 lack of *tss*I gene in SS T6SS-1.)



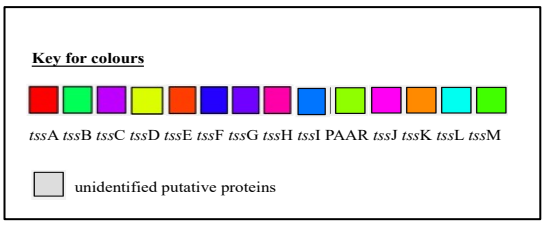


Figure 8-3: Comparison of all *Serratia* sp. *Orius* isolates T6SS in the regions of SS T6SS-2.

(All core genes and non-conserved genes components of the T6SS are indicated in different colours showing on above key for colours. Grey colour key is unidentified putative proteins by SecReT6. SS T6SS-2 was found in all *Serratia* sp. *Orius* isolates analysed, while *Serratia* sp. SCBI reference genome lack of *tss*G gene in SS T6SS-2.)

The genetic architecture of SS T6SS-2 (Figure 8-3) is highly conserved in almost every strain from *Serratia* sp. *Orius* isolates that harbour the cluster, except reference genome *Serratia* sp.SCBI lack of *tssG* gene in this cluster. SS T6SS-2 was found to contain a single *tss*I gene that encodes a TssI protein binding with additional DUF2169 and DUF 3540 of unknown function proteins. Only *Serratia* sp. SCBI lacks a *tss*G gene in the locus, the reason for that still unknown. Comparative analysis of SS T6SS-2 showed that there was no variability of this cluster between the isolates. The genetic architecture, gene order and gene content of SS T6SS-2 was conserved in *Serratia* sp. *Orius* isolates. The high conservation of this cluster suggests a strong selective pressure to maintain the gene content and order, although its specific role is unknown.

8.4.4 Identifying T6SS adaptors and effectors in all *Serratia* sp. *Orius* isolates

In SecReT6 website, the effector genes information of all *Serratia* sp. *Orius* isolates were detected in their database and the results were multiple excel worksheets downloaded from database, but due to the excessive data capacity, it cannot show in supplementary data. Additionally, the identified putative proteins, that could not be identified by SecReT6 website, were re-identified by BLASTP. *Tae4 – Tai4* effector-immunity pairs are only conserved on all *Serratia* sp. *Orius* isolates (Figure 8-4). Four broadly distributed and phylogenetically distinct families of T6SS peptidoglycan (PG) amidase effectors-immunity (EI) pairs have been recently identified based on overall primary sequence homology and different substrate specificities (Zhang et al., 2013). *Tae4* (type VI amidase effector 4) and *Tai4* (type VI amidase immunity 4) are T6SS effector-immunity pairs from the fourth family. Most of T6SS EI pairs are discovered in pathogens that colonize polymicrobial sites in the host and natural environment, such as the gastrointestinal tract (GI tract), and the soil (Zhang et al., 2013). This suggests they are closely associated with interbacterial interactions in the formation of environmentally and clinically relevant microbial communities. Under these conditions, the cross-immunity against multiple effectors may promote the cooperation of some species and play an important role in interbacterial competition (Zhang et al., 2013). Therefore, the *Tae4 – Tai4* EI pair in these strains might be functional to the interbacterial competition in polymicrobial environment of *Orius* insects.

Furthermore, Russell et al found 27% of the immunity proteins they identified were not encoded adjacent to intact effector genes, while the effector genes always co-occur with immunity genes (Russell et al., 2012). This indicates there is a selective pressure to retain immunity even in the absence of cognate effectors for the antagonistic interspecies competition. It may explain the reason for Tae4 absence in sometime, but *Tai4* still exist in T6SS locus to support antagonistic interbacterial competition. Based on the results of SecReT6 effectors and immunity proteins detections, *Tse, Tme, Tde* and *Tle* are the well-known antibacterial effectors of *Serratia marcescens* (Russell et al., 2012) with cognate immunity proteins detected in all *Serratia* sp. *Orius* isolates (Russell et al., 2012), supported information shown on the supplementary data. It further confirmed that the role of T6SS in *Serratia* sp. *Orius* isolates is supporting the interbacterial competition in a polymicrobial environment.

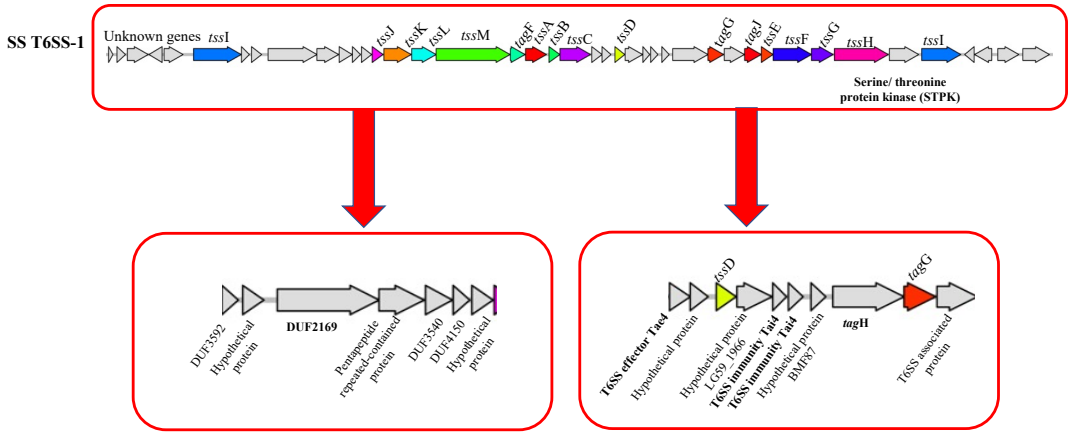


Figure 8-4: Genetic organisation of the different loci in grey colour associated with SS T6SS-1 gene cluster.

(The grey colour sections are unidentified putative proteins identified by SecReT6 and these proteins were re-identified by BLASTP. Most of these proteins related to T6SS formations and detected in all sequenced strains of Serratia sp. Orius isolates.)

8.4.5 Classification of T6SS subtypes

The phylogenetic analysis of 825 proteins in SecReT6 online database assigned as the *tss*B component matches the classification scheme of T6SS subtypes, suggesting that TssB alone may be a suitable classification marker once appropriate quantitative phylogenetic criteria are

established. All the strains from both different *Serratia* clades belong to T6SS subtype i3 (T6SS^{i3}) and the function of i3 subtype is directly related to antibacterial in SecReT6 online database search.

8.5 Conclusion

Comparative analyses of the T6SS in the genomes of sequenced strains of *Serratia* sp. SCBI clade identified two putative gene clusters SS T6SS-1 and SS T6SS-2. Two of these were potentially functional as they contained the 13 core genes necessary for synthesis of T6SS. SS T6SS-1 and SS T6SS-2 were widespread in the genomes of all sequenced strains including environmental isolates. However, the finding that SS T6SS-1 and SS T6SS-2 were present in both pathogenic and non- pathogenic strains of *Serratia* species supports the concept that the T6SS may evolve to play different roles unrelated to pathogenicity, such as competition against other microbes, fitness and/or niche adaptation, and its need to study in the future.

CHAPTER 9: DISCUSSION

9.1 Summary and interpretation of findings

9.1.1 Classification of *Orius* specimens

Molecular identification of five different *Orius* species using *cox*I gene marker was conducted for the first time. The phylogenetic relationships were determined using the maximum-likelihood method supported with strong bootstrap probabilities clustering of all taxa. Two phylogenetic trees were generated utilizing different selective reference *Orius* species, and consistently the reference *Orius* species segregated as a clade with *Orius tristicolor, Orius minutus, Orius vicinus, Orius laticollis, Orius minutus,* and *Orius sauteri* grouped closely together. The other major clade observed in the tree included the five different *Orius* species from European countries. Interestingly, the specimens of *Orius albidipennis* seem to be an independent clade with an independent ancestor. Furthermore, molecular phylogeny based on COI sequences revealed a close evolutionary relationship between *O. niger* and *O. pallidicornis*, indicating that *O. pallidicornis* can be considered a cryptic species within *O. niger*. However, a recent study of *O. niger* mentioned that *O. niger* and *O. sauteri* are the same species, and *Ttraphleps aterrimus* is sister to *O. niger* and *O.sauteri* in their whole mitochondrial genome phylogenetic analysis (Zhang et al., 2019). In future work, if there are additional *Orius* population specimens present, when the sample range increases, the more complete taxonomic classification of *Orius* species will be complete. Additionally, if there are any chance to sequence the whole mitochondrial genome of *Orius* species, the classification of *Orius* species will be more accurate than partial sequences of *Orius* mitochondrial genes used in the molecular classifications.

9.1.2 Isolation of facultative symbionts of *Orius* sp.

The isolation of culturable microorganisms from *Orius* sp. specimen homogenates revealed three predominant bacterial colony morphologies across the whole range of insect specimens tested. The representative colony types from each insect population were initially used to amplify and classify by 16S rRNA gene. Due to the limitation of 16S rRNA classification, the

whole genome sequences of the above isolates were assembled and annotated by different approaches. Based on the comparison between SPAde and Velvet assembly results, SPAde was preferred than Velvet assembly because of automatic error corrections of SPAde function. Therefore, SPAde assemblies were chosen for assembling all the bacterial genomes.

9.1.3 Confirmation of the presence of *Orius'* facultative symbionts by genome-specific PCR

After genome assembly, unique, species-specific sequences were used to design primers for genome-specific PCR. This PCR aimed to confirm the conserved relationship between three predominant facultative symbionts and the *Orius* specimens. Since the predominant colony morphologies were not recovered from each insect population by culture techniques, genome-specific PCR was performed to detect the presence of these isolates in total insect DNA isolated from all the specimens collected from wide ranges of *Orius* species. Samples were taken from field collections and from *Orius* sp. specimens reared in the laboratory and covered the range of *Orius* sp. specimens as well as commercially used *Orius* species. This section of the study provided conclusive evidence for the existence of a symbiotic association between the predominant isolates and their *Orius* hosts.

The PCR results confirmed the presence of the *Erwinia* sp. *Orius* and *Serratia* sp. *Orius* predominant isolates in total DNA from representative insect specimens, apart from *Leucobacter* species. This may be due to the low abundance of *Leucobacter* sp. in these specimens, meaning that insufficient DNA was recovered to allow successful PCR. The artificial diets and environmental elements of lab-reared *Orius* species may have resulted in the enrichment of this species in lab-reared *Orius* species homogenates. Additionally, the *Leucobacter* sp. *Orius* strains identified within *Orius* specimens were similar with the nematode related *Leucobacter* sp. AEAR. However, *cox*I PCR primers used sufficient homology of *Orius* specimens to amplify homologous target sequences from *Nematoda*, none of the *cox*I sequences retrieved from COI PCR shared any similarity to available *Nematoda* sequences in the NCBI database. Additionally, on-going next-generation sequencing of the *Orius*-derived COI amplicons described here have not resulted in the detection of COI sequences from Nematoda (unpublished), further supporting the observation that *Leucobacter*

sp. *Orius* are hosted by the insects tested. Several *Leucobacter* species have been isolated from several insects hosts such as the non-biting midges *Chironomid* sp. (Laviad et al., 2015), the scarab beetle *Holotrichia oblita* (Zhu et al., 2016), and *Anopheles gambiae* (*Leucobacter* sp. Ag1, BioSample: SAMN03481186). Therefore, it is most likely that *Orius* species have a facultative symbiotic association with *Leucobacter* sp. *Orius.* In the future, once more genomes from this species become available, pangenome analyses will lead to the identification of genes and gene networks required for the facultative symbiotic lifestyle.

9.1.4 Identification of two putative new species of facultative symbionts from *Orius* sp. by phylogenomic analysis.

Due to the limitations of 16s rRNA taxonomic classification, assembled genomes from three types of predominant isolates and all available *Enterobacteriales* and *Actinobacteria* genome sequences from the NCBI database were retrieved and concatenated to create 400 protein alignments using PhyloPhlAn to create MLSA phylogenies. Two of these putative symbiotic bacteria, belonging to Erwiniaceae and Microbacteriaceae, are likely to be the first representatives of new species in MLSA phylogeny. Despite the sequence similarity shown by their 16S rRNA gene sequences, classifying them as members of the *Erwinia* and *Leucobacter* genera, GGDC failed to identify genomes similar enough to be classified as the same species, which further supports the proposal of them being new species in need of taxonomic classification. Furthermore, All the *Serratia* sp. *Orius* isolates were closely related to *Serratia* sp. SCBI and were found within the *Serratia* sp. SCBI complex. The exceptions were OLAL2 and OMLWL3 which belong to another clade of *S. marcescens* and were close to *S. marcescens* Db11. To confirm the results of MLSA phylogenies in various *Serratia* genus bacteria, the GGDC tool was used to calculate genome-to-genome distances between different *Serratia* species and *Serratia* sp. *Orius* isolates. Using *in silico* DDH, the GGDC results confirmed the high accuracy of the MLSA phylogenies. Consequently, *Serratia* sp. *Orius* isolates are considered to be the same species as *Orius Serratia* sp. SCBI. However, none of the available genome sequences within the *Serratia* sp. SCBI complex is similar with the rest of the *S. marcescens* genomes available, evidenced by long genome-to-genome distances, above the threshold set for same-species classification. For example, another study of *Serratia ureilytica*

revealed the low level of DNA–DNA hybridisation (43.7%) with the species of *S. marcescens* by using a dot-blot hybridisation method with a DIG DNA labelling and detection kit (Roche Diagnostics) (Bhadra et al., 2005). Additionally, other species of *S. marcescens* such as the human pathogens *S. marcescens* strains SM39 and SmUNAM836 are unlikely to be the same *S. marcescens* species as *S. marcescens* WWW4 or Db11, because of their low level of similarity in GGDC comparison and the fact that they belong to different clades of MLSA phylogeny. This may indicate that previous studies reporting the identification of *S. marcescens* derived from 16S rRNA sequences may need to be revised, due to the inaccuracy of 16S rRNA classification in some cases. Furthermore, it suggests that the classification of *S. marcescens* could be revised by both GGDC comparisons and MLSA phylogenies, in order to successfully separate different species and establish reference genomes for comparative genomic studies such as GI predictions.

9.1.5 GI prediction of differentiated lineages within the *Serratia* sp. *Orius* symbiont strains.

The presence of several genes encoding putative virulence factors in horizontally acquired genomic islands suggest the symbiotic associations can be established by dissimilar sets of mechanisms for killing, bioconversion, sanitization, and colonization. The presence of these sets of genes in the isolates further suggests that the major hurdle in symbiotic complex formation may be the initial development of co-tolerance between potential partners. The complete genome sequence of one of the partners in a nascent or ancient symbiotic association should enable future analysis of the *Serratia* complex. Additionally, most of Genome specific GIs contain over 50% of hypothetical proteins. This suggests that these typical GIs need to be further annotated in future study for understanding more about the functions and horizontal gene transfer events of this symbiont. In a future study, more distinct *Orius* species collected from different geographic locations can determine the range of these symbiotic associations.

9.1.6 Pangenome analysis

The genus *Serratia* is an important constituent of the insect symbiotic microbiome. This work represents the first characterization of predominant 13 *Serratia* sp. *Orius* isolates using pan-genomic analysis. The gene accumulation curve showed that the numbers of the core genome size decreased continually with addition of new strains, while the pangenome size showed an increasing trend in Chapter 7 (figure 7-7). The change of both curves (figure 7-7 and 7-8) slowed down because the pangenome of *Serratia* sp. *Orius* isolates was in an open state, indicating that unique genes would be added along with the addition of new strains. Furthermore, the unique genes represented only a small number of strains, suggesting that the evolution of *Serratia* sp. *Orius* isolates was relatively conservative. After obtaining these pangenome sequences statistics, the distribution of their functional categories was compared using the COGs and KEGG database. Because Orthologs are genes in different species that have evolved from a common ancestral gene via speciation and often retain the same function during evolution. Comparing orthologs is essential to identify events of gene gain or loss. Especially, COGs provide genome-scale analysis of protein function prediction. However, it was still many genes did not exist in both databases, the characteristics of relatively new species are scarce, thus suggesting the need to strengthen in-depth studies of *Serratia* species. This lack of in-depth knowledge may explain why functional categories could not be determined for many genes. The most abundant functions in the core genome of *Serratia* sp. *Orius* isolates were associated with metabolism in both databases. The overall proportion of pangenome in COG database related to metabolic functions was 42.2%, 25.9% and 21.6% in the core genome, accessory genome, and unique genes of strain-specific, respectively. More specifically, amino acid transport and metabolism (E), and carbohydrate transport and metabolism (G) were abundant in the core genes, suggesting that these genes were relatively conserved in *Serratia* sp. *Orius* isolates. Furthermore, information storage and processing genes are more abundant in accessory genome, such as the high proportion of (J) translation, ribosomal structure and biogenesis, (K) transcription, and (L) replication, recombination and repair functions. It suggested that these accessory genes are less likely transferred horizontally from other species or even from another genus because there is no mobilome-related prophage and transposase genes detected by both databases and low abundance of phage elements shown in Roary or Scoary results as well. However, several GIs gene elements were found in the pangenome of *Serratia* sp. *Orius* isolates (supplementary data), and accessory genome contains seemingly plasmid-related genes (Supplementary data). Sequence homology searches results in chapter 7

explains why some plasmid genes are included within the *Orius* accessory genes. *S. marcescens* strain B-6493 is a human pathogen and *S. marcescens* PWN146 was isolated from the nematode *Bursaphelenchus xylophilus*, and both are different species from those in the SCBI complex as confirmed by GGDC. With a lack of replicon segregation within available *Serratia* species genome assemblies, it is difficult to expand on plasmid sequence comparisons at this stage to define a plasmid genealogy for the genus, but our findings are certainly indicative of plasmid exchange and sequence plasticity across the *Serratia* genus and worthy of future studies. It also suggested that some horizontal gene transfer events may occur in plasmid exchange between their genera. Furthermore, the pangenome analyses described here clearly associate gene from accessory genome of *Serratia* sp. *Orius* isolates with the insect symbiosis trait and segregates them from similar species from the SCBI complex, including a known nematodes symbiont like Serratia sp. SCBI. These functions of accessory genes might be relatively important to *Serratia* sp. *Orius* isolates. Its need to further study in the future. Specifically, the discovery of the conserved pathway of metabolism in isolates increases understanding of the metabolic network within the *Serratia* isolates. Moreover, these results may impact understanding of the symbiosis of *Orius* as well in the future. Finally, the results of this study increase our understanding of the characteristics of *Serratia* species as an insect symbiont and will facilitate future studies of this genus.

9.1.7 Prediction of T6SS encoded by *Serratia* sp. *Orius* isolates strains isolated from multiple *Orius* species

Two different T6SS loci were identified in all *Serratia* sp. *Orius* isolates. The genetic organization of SS T6SS-1 and -2 loci further suggests that these clusters were independently acquired to play differing roles in the different strains of *Serratia* sp. *Orius* isolates. Furthermore, the variable regions associated with *tss*D and *tss*I genes could account for specialization of each T6SS based on the needs of the specific strain. The classification of T6SS subtype is i3 in these *Serratia* sp. *Orius* isolates, and it is associated with interbacterial competition (Amaya et al., 2022). *Salmonella enterica* serotype Dublin (*S. Dublin*) is a cattle-adapted pathogen that harbours both T6SS_SPI-6 and T6SS_SPI-19 and both systems have been linked to virulence and host colonization in *S. Dublin*. T6SSSPI-6 belongs to subtype i3

and encoding three candidate antibacterial effectors located within SPI-6 (Amaya et al., 2022). Each antibacterial effector gene is located upstream of a gene encoding a hypothetic immunity protein, thus conforming an effector/immunity (E/I) module. Of note, the genes encoding these effectors and immunity proteins are widely distributed in *Salmonella* genomes, it was suggesting a relevant role in interbacterial competition and virulence (Amaya et al., 2022). In *Serratia* sp. *Orius* isolates, *Tae4–Tai4* effector–immunity cognate partners presented in T6SS loci of these isolates suggests that the establishment of a symbiotic lifestyle requires molecular mechanisms ensuring successful interspecies competition which is quite similar to *S. Dublin*.

Additionally, the T6SS of *Xanthomonas citri* is the only example experimentally characterized so far within the Xanthomonadales order and displays anti-eukaryotic function by providing resistance to predation by amoeba. This T6SS is regulated at the transcriptional level by a signalling cascade involving a *Ser*/Thr kinase and an extracytoplasmic function (ECF) sigma factor. In *silico* predictions identified a series of proteins with known toxic domains as putative T6SS effectors, suggesting that the T6SSs of Xanthomonadales display both anti-prokaryotic and anti-eukaryotic properties depending on the phylogenetic group and bacterial species (Bayer-Santos, Ceseti, Farah and Alvarez-Martinez, 2019). This study suggested in future study of our isolates' T6SS could involve in different phylogenetic groups of *Serratia* sp. groups to predict T6SS in both in *silico* predictions.

Furthermore, the accessory elements of T6SS clusters are highly variable; for example, T6SS-a of *Serratia* contains multiple accessory proteins, such as *tag*G and *tag*H, Since the bacteria carrying T6SS gene clusters are found in diverse environments and the function of T6SS is highly versatile, these accessory proteins might be involved in regulation or might confer additional functions to the system. For instance, homologs of *tag*G and *tag*H in T6SS was characterized to play important roles in activation of T6SS at transcriptional or post-translational levels in *Serratia marcescens* FS14 (Li et al., 2015). The function of T6SS in *Serratia marcescens* FS14 demonstrated high antagonistic activities against both bacterial and fungal phytopathogens (Li et al., 2015). Some of genes found within the T6SS cluster were reported as secretory effector proteins or self-immunity proteins, such as the ssp proteins which are novel toxins recently identified in *S. marcescens* Db10 (Li et al., 2015). Therefore, it is speculated that other genes assigned with unknown functions in T6SS clusters are likely novel effectors or immunity proteins, whose roles remain to be experimentally verified.

In addition, a recent study about honeybee gut symbionts working together protects bees from invasion by a bacterial pathogen *Serratia marcescens.* In honeybees, perturbing or depleting the gut microbiota increases host mortality rates upon challenge with the opportunistic pathogen *Serratia marcescens*, suggesting antagonism between *Serratia marcescens* and one or more members of the bee gut microbiota. In laboratory culture, *Serratia marcescens* uses a T6SS to kill bacterial competitors, but the role of this T6SS within hosts is unknown. They found that *S. marcescens* is rapidly eliminated in the presence of the microbiota but persists in microbiota-free guts. Protection is reduced in noncolonised and antibiotic-treated bees, possibly because different symbionts occupy distinct niches. *Serratia marcescens* uses a T6SS to antagonize *Escherichia coli* and other *S. marcescens* strains but shows limited ability to kill bee symbionts. Furthermore, wild-type and T6SS-deficient *S. marcescens* strains achieved similar abundance and persistence in bee guts (Steele et al., 2021). Thus, an intact gut microbiota offers robust protection against this common pathogen, whose T6SSs do not confer the ability to compete with commensal species. In this study, bacteria native to the honeybee gut work together to exclude the opportunistic pathogen *Serratia marcescens*. Although *S. marcescens* has a T6SS that can kill bacteria, bee gut bacteria seem resistant to its effects. This limitation may partially explain why ingestion of *S. marcescens* is rarely lethal to insects with healthy gut communities, but other species of *Serratia* could present in insect symbiotic bacteria community.

Additionally, Bacteria employ diverse competitive strategies to enhance fitness and promote their own propagation. Because these mechanisms can be costly to use, their expression and function are often restricted to specific environments where the benefits outweigh the costs. However, little is known about how symbiotic bacteria modulate competitive mechanisms as they compete for a host niche. In recent study, some researchers used the bioluminescent squid and fish symbiont *Vibrio fischeri* to probe for host and environmental conditions that control interbacterial competition via T6SS. their findings identified a new host-specific cue that promotes competition among many but not all *V. fischeri* isolates, underscoring the utility of studying multiple strains to reveal how competitive mechanisms may be differentially regulated among closely related populations as they evolve to fill distinct niches (Speare et al., 2021). Furthermore, products secreted by T6SS systems encoded by other *Serratia* species have been confirmed the antimicrobial properties targeting microbial competitors to ensure survival in polymicrobial environments (Murdoch et al., 2011; English et al., 2012;

Srikannathasan et al., 2013). It suggested that the host-symbiont association between *Orius* sp. and the *Serratia* sp. *Orius* isolates described may have driven the acquisition and specialisation of strain- specific T6SS effector–immunity partnerships to preserve niche colonisation from closely related, invading environmental or pathogenic *Serratia* species.

In the future, key questions that need to be asked to determine: 1) whether the T6SSs of *Serratia* species are functionally active and what roles they play in host-pathogen interactions and fitness; 2) which *in vitro* and *in vivo* conditions activate the T6SSs; 3) the presence of different potential effectors secreted by the T6SSs of *Serratia* species and their physiological relevance to fitness and host-symbiont interactions; and 4) how T6SSs are regulated in these strains.

9.2 The limitations of current study

In the *Orius* insect species classification, because the samples analysed in this study were collected from limited areas only in several European countries, the collection of samples and species from wider areas could better indicate potential differences in morphological characteristics and genetic markers. Also, the selection of one sample of any *Orius* species in each location for molecular investigation, as a limiting factor in the present study, should be considered.

In the part of symbiotic bacteria extractions, amplification and sequencing biases also made it unfeasible to determine the relative abundances of different species, the 16s rRNA sequencing in the isolates of this study is unable to determine the relative abundances of different species from insect total DNA. Furthermore, it could not detect and classify obligate or unculturable symbiotic bacteria of *Orius* specimens. Therefore, it is hard to know which bacterial species exactly colonise or associate to *Orius* species. In the future, the deep sequencing of total DNA recovered from insect specimens will permit the detection and classification of most of the microorganisms associated to *Orius* species.

9.3 General Conclusion

This study filled some knowledge gaps in the understanding of facultative symbionts associated with various European *Orius* species hosts across a wide range of geographic locations. Firstly, the taxonomic classifications of *Orius* specimens were confirmed by the phylogeny constructed by *Orius' cox*I sequences. Two *Orius* species (*O. pallidicornis* and *O. albidipennis*) are absent from the NCBI database. In the phylogeny, *O. pallidicornis* is closely related to *O. niger*, while *O. albidipennis* has an independent linkage, so these two species could be further classified in the future. After isolation and assembly of the *Orius* isolates, a new *Serratia* species was identified and found to be a facultative symbiont of European *Orius* species, closely related to *Serratia* sp. SCBI. However, the GI predictions of this *Serratia* species illustrate that some strains of this species contain virulence factors, and most European *Orius* species used as pest control agents in IPM applications. Some *Orius* species will be massively produced, so these virulence factors are most likely to transfer horizontally between *Orius* species and environmental bacteria on crops. Additionally, there is one report mentioning that *O. majusculus* could bite humans and therefore might transfer bacteria to humans (Kampen and Werner, 2011). Although earlier identifications of *Serratia* sp. *Orius* isolates showed they are different to human pathogens, concerns remain regarding the safety of these *Orius* bio-control agents in IPM applications. Furthermore, this type of situation could also arise with the use of other bio-control agents. Therefore, monitoring of the use of this biological control by characterising its microbiome should be considered as part of standard quality control. Additionally, these results are intended to serve as a guide for future functional studies on the symbiosis. Hopefully they represent a leap forward in a system that holds great potential for future research.

CHAPTER 10: Supplementary Data

Supplementary Table 6- 1: Representative genomic islands (GI) identified in *Serratia* sp. *Orius* genomes.

(GI coordinates, length, locus name and annotation in corresponding genome Genebank entry is provided. The occurrence of each GI per genome is indicated by Y, and query coverage lower than 100% is shown between brackets.)

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
GI1	OLIL_1437758..1441926	4168	BMH23_21805	hypothetical protein	Y (97)						Y						
			BMH23_21800	hypothetical protein													
			BMH23_06825	hypothetical protein													
			BMH23_06820	hypothetical protein													
GI2	OLBL_3467782..3472585	4803	BMF92_01405	TetR family transcriptional regulator	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	
			BMF92_01410	hypothetical protein													
			BMF92_01415	Inversin													
			BMF92_01420	Inversin													
			BMF92_01425	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
GI3	OLJL_3143931..3162739	18808	BMH24_23345	LacI family transcriptional regulator	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	
			BMH24_23350	hypothetical protein													
			BMH24_23355	hypothetical protein													
			BMH24_23360	DNA methylase													
			BMH24_23365	transcriptional regulator													
			BMH24_23370	hypothetical protein													
			BMH24_00530	phage tail protein													
			BMH24_00525	phage tail protein													
			BMH24_00520	hypothetical protein													
			BMH24_00515	phage baseplate protein													
			BMH24_00510	phage tail protein													
			BMH24_00505	DNA circulation protein													
			BMH24_00500	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMH24_0 0495	hypothetical protein													
			BMH24_0 0490	phage tail protein													
			BMH24_0 0485	phage tail protein													
			BMH24_0 0480	hypothetical protein													
			BMH24_0 0475	hypothetical protein													
			BMH24_0 0470	hypothetical protein													
			BMH24_0 0465	head protein													
GI4	OPNLW1_4065274..4 090072	24798	BOM26_1 3715	NAD(+) kinase	Y	Y	Y	Y (61)	Y	Y	Y	Y	Y (48)	Y (46)	Y	Y (92)	
			BOM26_1 3710	DNA repair protein RecN													
			BOM26_1 3705	outer membrane protein assembly factor BamE													
			BOM26_1 3700	RnfH family protein													
			BOM26_1 3695	ubiquinone-binding protein													
			BOM26_1 3690	SsrA-binding protein													
			BOM26_1 3680	integrase													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_13675	hypothetical protein													
			BOM26_13670	hypothetical protein													
			BOM26_13665	GTP-binding protein													
			BOM26_13660	hypothetical protein													
			BOM26_13655	SIR2 family protein													
			BOM26_13650	hypothetical protein													
			BOM26_13645	relaxase													
			BOM26_13640	hypothetical protein													
			BOM26_13635	hypothetical protein													
			BOM26_13630	hypothetical protein													
			BOM26_13625	DNA-binding protein													
			BOM26_13620	integrase													
			BOM26_13615	recombinase													
			BOM26_13610	dipicolinate synthase													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_13605	hypothetical protein													
			BOM26_13600	hypothetical protein													
			BOM26_13595	hypothetical protein													
GI5	OLBL_5296773..5369914	73141	BMF92_18545	hypothetical protein	Y	Y	Y	Y						Y (42)			
			BMF92_18550	hypothetical protein													
			BMF92_18555	CsbD family protein													
			BMF92_18560	hypothetical protein													
			BMF92_18565	hypothetical protein													
			BMF92_18570	hypothetical protein													
			BMF92_18575	hypothetical protein													
			BMF92_18580	CAP-Gly protein													
			BMF92_18585	hypothetical protein													
			BMF92_18590	catalase HPII													
			BMF92_18595	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF92_18600	recombinase RecA													
			BMF92_18605	DNA polymerase V subunit UmuD													
			BMF92_18610	DNA polymerase V subunit UmuC													
			BMF92_18615	Replication protein													
			BMF92_18620	F-pilin acetylation protein TraX													
			BMF92_18625	phospholipase D family protein													
			BMF92_18630	hypothetical protein													
			BMF92_18635	conjugative transfer relaxase/helicase TraI													
			BMF92_18640	type IV conjugative transfer system coupling protein TraD													
			BMF92_18645	hypothetical protein													
			BMF92_18650	hypothetical protein													
			BMF92_24745	hypothetical protein													
			BMF92_24640	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF92_24645	DNA polymerase V subunit UmuC													
			BMF92_24650	DNA polymerase V subunit UmuD													
			BMF92_24655	hypothetical protein													
			BMF92_24730	transketolase													
			BMF92_24760	holin													
			BMF92_24765	structural protein													
			BMF92_24685	hypothetical protein													
			BMF92_24690	hypothetical protein													
			BMF92_24695	hypothetical protein													
			BMF92_24700	DNA polymerase V													
			BMF92_24735	hypothetical protein													
			BMF92_24475	hypothetical protein													
			BMF92_24480	hypothetical protein													
			BMF92_24485	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF92_24490	DNA primase													
			BMF92_24495	hypothetical protein													
			BMF92_24500	hypothetical protein													
			BMF92_24505	hypothetical protein													
			BMF92_24510	hypothetical protein													
			BMF92_21740	hypothetical protein													
			BMF92_21745	hypothetical protein													
			BMF92_21750	hypothetical protein													
			BMF92_21755	phage tail protein													
			BMF92_21760	phage baseplate protein													
			BMF92_21765	hypothetical protein													
			BMF92_21770	phage baseplate protein													
			BMF92_21775	baseplate protein													
			BMF92_21780	DNA circularization protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF92_2 1785	phage tail tape measure protein													
			BMF92_2 1790	phage tail protein													
			BMF92_2 1795	phage tail protein													
GI6	OLCL_5264245..5268890	4645	BMF85_2 4875	hypothetical protein		Y											
			BMF85_2 4435	hypothetical protein													
			BMF85_2 4440	DNA polymerase V subunit UmuD													
			BMF85_2 4445	DNA polymerase V subunit UmuC													
GI7	OLJL_5232873..5253687	20814	BMH24_2 4520	hypothetical protein								Y					
			BMH24_2 4525	aspartate--ammonia ligase													
			BMH24_2 4530	transcriptional regulator AsnC													
			BMH24_2 4535	FMN-binding protein MioC													
			BMH24_2 4700	relaxase													
			BMH24_2 4705	hypothetical protein													
			BMH24_2 3150	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMH24_23155	hypothetical protein													
			BMH24_23160	hypothetical protein													
			BMH24_23165	hypothetical protein													
			BMH24_23170	oxidoreductase													
			BMH24_23175	hypothetical protein													
			BMH24_23180	hypothetical protein													
GI8	OSPLW9_2078225..2094262	16037	BVV03_01360	hypothetical protein											Y	Y	
			BVV03_01355	hypothetical protein													
			BVV03_01350	oxidoreductase													
			BVV03_01345	hypothetical protein													
			BVV03_01340	hypothetical protein													
			BVV03_01335	hypothetical protein													
			BVV03_01330	hypothetical protein													
			BVV03_01325	phage tail protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BVV03_01320	hypothetical protein													
			BVV03_01315	transcriptional regulator													
			BVV03_01310	hypothetical protein													
			BVV03_01305	hypothetical protein													
			BVV03_01300	DNA polymerase V													
			BVV03_01295	hypothetical protein													
			BVV03_01290	hypothetical protein													
			BVV03_01285	IS5 family transposase													
			BVV03_01280	hypothetical protein													
			BVV03_01275	hypothetical protein													
GI9	OPWLW2_5175957..5249411	73454	BOM25_12150	hypothetical protein													Y
			BOM25_12155	DNA primase													
			BOM25_12160	hypothetical protein													
			BOM25_12165	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_12170	hypothetical protein													
			BOM25_12175	hypothetical protein													
			BOM25_12180	hypothetical protein													
			BOM25_12185	DNA-binding protein													
			BOM25_12190	hypothetical protein													
			BOM25_12195	hypothetical protein													
			BOM25_12200	conjugal transfer protein TraG													
			BOM25_12205	hypothetical protein													
			BOM25_12210	integrating conjugative element protein													
			BOM25_12215	integrating conjugative element protein													
			BOM25_12220	integrating conjugative element protein													
			BOM25_12225	hypothetical protein													
			BOM25_12230	FhaB													
			BOM25_12235	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_12240	hypothetical protein													
			BOM25_12245	hypothetical protein													
			BOM25_12250	filamentous hemagglutinin													
			BOM25_12255	transporter													
			BOM25_12260	hypothetical protein													
			BOM25_12265	restriction endonuclease													
			BOM25_12270	type I restriction-modification system subunit M													
			BOM25_12275	restriction endonuclease subunit S													
			BOM25_12280	DEAD/DEAH box helicase													
			BOM25_12285	acetyltransferase													
			BOM25_12290	conjugative transfer ATPase													
			BOM25_12295	conjugal transfer protein													
			BOM25_15780	SIR2 family protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_16000	hypothetical protein													
			BOM25_16005	hypothetical protein													
			BOM25_16010	hypothetical protein													
			BOM25_15825	transketolase													
			BOM25_13350	hypothetical protein													
			BOM25_13355	hypothetical protein													
			BOM25_13360	hypothetical protein													
			BOM25_13365	hypothetical protein													
			BOM25_13370	hypothetical protein													
			BOM25_13375	hypothetical protein													
			BOM25_13380	molybdopterin-guanine dinucleotide biosynthesis protein MobC													
			BOM25_13385	hypothetical protein													
			BOM25_13390	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_13395	single stranded DNA-binding protein													
			BOM25_13400	hypothetical protein													
			BOM25_13405	hypothetical protein													
			BOM25_13410	hypothetical protein													
			BOM25_13415	replication associated protein RepA1													
			BOM25_13420	hypothetical protein													
			BOM25_13425	proteolipid membrane potential modulator													
			BOM25_13430	hypothetical protein													
			BOM25_13435	hypothetical protein													
			BOM25_13440	hypothetical protein													
			BOM25_13445	hypothetical protein													
			BOM25_13450	hypothetical protein													
			BOM25_13455	hypothetical protein													
			BOM25_13460	conjugal transfer protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_13465	conjugal transfer protein													
			BOM25_13470	hypothetical protein													
			BOM25_13475	conjugal transfer protein													
			BOM25_13480	hypothetical protein													
			BOM25_13485	conjugal transfer protein													
			BOM25_13490	conjugal transfer protein TrbI													
			BOM25_13495	hypothetical protein													
			BOM25_13500	hypothetical protein													
			BOM25_13505	P-type DNA transfer ATPase VirB11													
			BOM25_13510	hypothetical protein													
GI10	OPNLW1_4508651..4517077	8426	BOM26_25005	restriction endonuclease											Y	Y	
			BOM26_25000	type I restriction-modification system subunit M													
			BOM26_24995	restriction endonuclease subunit S													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_24990	DEAD/DEAH box helicase													
			BOM26_24985	acetyltransferase													
			BOM26_24980	conjugative transfer ATPase													
GI11	OLCL_5377229..5384534	7305	BMF85_23215	amidohydrolase		Y					Y						
			BMF85_23220	hypothetical protein													
			BMF85_23225	hypothetical protein													
			BMF85_24475	relaxase													
GI12	OLMTLW26_5256304..5371494	115190	BK415_16520	molybdenum cofactor guanylyltransferase MobA									Y				
			BK415_16525	molybdopterin-guanine dinucleotide biosynthesis protein B													
			BK415_18620	transcriptional regulator MelR													
			BK415_20650	D-3-phosphoglycerate dehydrogenase													
			BK415_18155	D-alanyl-D-alanine carboxypeptidase													
			BK415_20615	betaine-aldehyde dehydrogenase													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BK415_18120	short-chain dehydrogenase													
			BK415_16565	Replication protein													
			BK415_20300	hypothetical protein													
			BK415_18665	glucose-1-phosphate adenylyltransferase													
			BK415_18085	hypothetical protein													
			BK415_16865	hypothetical protein													
			BK415_16870	DNA polymerase V subunit UmuC													
			BK415_16875	DNA polymerase V subunit UmuD													
			BK415_16880	hypothetical protein													
			BK415_20605	hypothetical protein													
			BK415_18630	hypothetical protein													
			BK415_18100	paraquat-inducible protein B													
			BK415_18615	multifunctional acyl-CoA thioesterase I/protease I/lysophospholipase L1													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BK415_16885	GTP-binding protein													
			BK415_18070	SIR2 family protein													
			BK415_18135	structural protein													
			BK415_18140	holin													
			BK415_18170	hypothetical protein													
			BK415_18090	mobilization protein													
			BK415_20290	hypothetical protein													
			BK415_20295	hypothetical protein													
			BK415_20275	glutaminase													
			BK415_16890	integrase													
			BK415_20625	hydrogenase 3 large subunit													
			BK415_20460	glycolate oxidase subunit GlcD													
			BK415_18125	hypothetical protein													
			BK415_16855	abortive phage infection protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BK415_16860	RNA polymerase													
			BK415_20455	potassium transporter													
			BK415_18175	hypothetical protein													
			BK415_18105	transposase													
			BK415_18660	magnesium transporter													
			BK415_15690	DNA-binding protein													
			BK415_15695	hypothetical protein													
			BK415_15700	hypothetical protein													
			BK415_15705	hypothetical protein													
			BK415_15710	relaxase													
			BK415_17615	nikA protein													
			BK415_18165	IS5 family transposase													
			BK415_18150	4-hydroxybenzoate polyprenyltransferase													
			BK415_18145	plasmid replication protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BK415_18600	type I-F CRISPR-associated protein Csy2													
			BK415_20620	nucleoside-specific channel-forming protein Tsx													
			BK415_18130	colicin-10													
			BK415_20645	peptidoglycan-binding protein													
			BK415_16570	hypothetical protein													
			BK415_16575	hypothetical protein													
			BK415_16580	CsbD family protein													
			BK415_16585	hypothetical protein													
			BK415_16590	hypothetical protein													
			BK415_16595	hypothetical protein													
			BK415_16600	hypothetical protein													
			BK415_16605	CAP-Gly protein													
			BK415_16610	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BK415_16615	catalase HPII													
GI13	OPWLW2_5412131..5476832	64701	BOM25_15975	hypothetical protein													Y
			BOM25_15980	hypothetical protein													
			BOM25_15985	hypothetical protein													
			BOM25_15990	hypothetical protein													
			BOM25_15995	hypothetical protein													
			BOM25_15865	hypothetical protein													
			BOM25_15870	DNA circulation protein													
			BOM25_15875	phage tail protein													
			BOM25_15880	phage baseplate protein													
			BOM25_15745	DNA topoisomerase III													
			BOM25_15750	hypothetical protein													
			BOM25_15755	integrating conjugative element protein													
			BOM25_15760	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_15765	hypothetical protein													
			BOM25_15770	hypothetical protein													
			BOM25_15775	hypothetical protein													
			BOM25_15850	restriction endonuclease													
			BOM25_15855	hypothetical protein													
			BOM25_13865	adenine methyltransferase													
			BOM25_13870	transposase													
			BOM25_13875	DNA-binding protein													
			BOM25_13880	hypothetical protein													
			BOM25_13885	hypothetical protein													
			BOM25_13890	hypothetical protein													
			BOM25_13895	hypothetical protein													
			BOM25_13900	DUF3085 domain-containing protein													
			BOM25_13905	DUF1738 domain-containing protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_13910	hypothetical protein													
			BOM25_13915	hypothetical protein													
			BOM25_15815	SAM-dependent methyltransferase													
			BOM25_16020	hypothetical protein													
			BOM25_14090	DNA topoisomerase III													
			BOM25_14095	hypothetical protein													
			BOM25_14100	integrating conjugative element protein													
			BOM25_14105	hypothetical protein													
			BOM25_14110	hypothetical protein													
			BOM25_14115	hypothetical protein													
			BOM25_14120	hypothetical protein													
			BOM25_15245	recombinase													
			BOM25_15250	integrase													
			BOM25_14970	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_14975	relaxase													
			BOM25_14980	hypothetical protein													
			BOM25_13535	conjugal transfer protein													
			BOM25_13540	conjugative transfer ATPase													
			BOM25_13545	acetyltransferase													
			BOM25_13550	hypothetical protein													
			BOM25_13555	hypothetical protein													
			BOM25_13560	integrating conjugative element protein													
			BOM25_13565	integrating conjugative element protein													
			BOM25_13570	integrating conjugative element protein													
			BOM25_13575	hypothetical protein													
			BOM25_13580	conjugal transfer protein TraG													
			BOM25_13585	hypothetical protein													
			BOM25_13590	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_13595	SAM-dependent methyltransferase													
GI14	OLFL_5204426..5259208	54782	BMF91_13890	molybdenum cofactor guanylyltransferase MobA					Y								
			BMF91_13895	molybdopterin-guanine dinucleotide biosynthesis protein B													
			BMF91_25065	hypothetical protein													
			BMF91_25100	hypothetical protein													
			BMF91_25075	virulence factor VirK													
			BMF91_23835	P-type DNA transfer ATPase VirB11													
			BMF91_23840	hypothetical protein													
			BMF91_23845	conjugal transfer protein													
			BMF91_23850	hypothetical protein													
			BMF91_23855	hypothetical protein													
			BMF91_23860	type IVB pilus formation outer membrane protein, R64 PilN family													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF91_23865	hypothetical protein													
			BMF91_23870	hypothetical protein													
			BMF91_23875	pilus assembly protein PilQ													
			BMF91_23880	conjugal transfer protein													
			BMF91_23885	conjugal transfer protein													
			BMF91_23890	hypothetical protein													
			BMF91_23895	hypothetical protein													
			BMF91_23900	molecular chaperone DnaJ													
			BMF91_23905	hypothetical protein													
			BMF91_23910	hypothetical protein													
			BMF91_23915	ssDNA-binding protein													
			BMF91_23920	hypothetical protein													
			BMF91_23925	plasmid stability protein StbB													
			BMF91_23930	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF91_23935	molybdopterin-guanine dinucleotide biosynthesis protein MobC													
			BMF91_23940	hypothetical protein													
			BMF91_23945	hypothetical protein													
			BMF91_23950	antitoxin of toxin-antitoxin stability system													
			BMF91_23955	hypothetical protein													
			BMF91_23960	addiction module antidote protein, HigA family													
			BMF91_23965	hypothetical protein													
			BMF91_23970	hypothetical protein													
			BMF91_23975	hypothetical protein													
			BMF91_23980	hypothetical protein													
			BMF91_23985	TriA protein													
			BMF91_23990	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF91_2 3995	conjugal transfer protein													
			BMF91_2 4000	conjugal transfer protein													
			BMF91_2 4005	conjugal transfer protein													
			BMF91_2 4010	hypothetical protein													
			BMF91_2 4015	conjugal transfer protein													
			BMF91_2 4020	hypothetical protein													
			BMF91_2 4025	hypothetical protein													
			BMF91_2 5060	hypothetical protein													
			BMF91_2 4945	urocanate hydratase													
			BMF91_2 4950	hypothetical protein													
GI15	OPWLW2_5320836..5373095	52259	BOM25_1 3010	type-F conjugative transfer system pilin assembly protein TrbC													Y
			BOM25_1 3015	type-F conjugative transfer system mating-pair stabilization protein TraN													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_13020	type-F conjugative transfer system pilin assembly protein TraF													
			BOM25_13025	hypothetical protein													
			BOM25_13030	type-F conjugative transfer system pilin assembly thiol-disulfide isomerase TrbB													
			BOM25_13035	conjugal transfer protein TraH													
			BOM25_13040	conjugal transfer protein TraG													
			BOM25_13045	hypothetical protein													
			BOM25_13050	hypothetical protein													
			BOM25_13055	type IV conjugative transfer system coupling protein TraD													
			BOM25_13060	conjugative transfer relaxase/helicase TraI													
			BOM25_15785	hypothetical protein													
			BOM25_15790	DNA circulation protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_15795	phage tail protein													
			BOM25_15800	phage baseplate protein													
			BOM25_14010	hypothetical protein													
			BOM25_14015	single-stranded DNA-binding protein													
			BOM25_14020	hypothetical protein													
			BOM25_14025	DUF2442 domain-containing protein													
			BOM25_14030	pilus assembly protein PilL													
			BOM25_14035	hypothetical protein													
			BOM25_14040	integrating conjugative element protein													
			BOM25_14045	lytic transglycosylase													
			BOM25_14050	DUF2859 domain-containing protein													
			BOM25_14055	restriction endonuclease													
			BOM25_15230	hypothetical protein													
			BOM25_15235	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_15240	hypothetical protein													
			BOM25_16045	hypothetical protein													
			BOM25_16050	hypothetical protein													
			BOM25_16055	hypothetical protein													
			BOM25_16060	hypothetical protein													
			BOM25_16065	host-nuclease inhibitor protein Gam													
			BOM25_15720	conjugative coupling factor TraD, PFGI-1 class													
			BOM25_15725	DUF4400 domain-containing protein													
			BOM25_15730	hypothetical protein													
			BOM25_15735	hypothetical protein													
			BOM25_15740	integrating conjugative element protein													
			BOM25_14060	conjugative coupling factor TraD, PFGI-1 class													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_14065	integrating conjugative element membrane protein													
			BOM25_14070	IS110 family transposase													
			BOM25_14075	LD-carboxypeptidase													
			BOM25_14080	hypothetical protein													
			BOM25_14085	integrating conjugative element protein													
			BOM25_16080	hypothetical protein													
			BOM25_16085	hypothetical protein													
			BOM25_16090	phage tail protein													
			BOM25_16095	phage tail protein													
			BOM25_16100	hypothetical protein													
			BOM25_16105	hypothetical protein													
			BOM25_16110	hypothetical protein													
			BOM25_16115	head protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
GI16	OPNLW1_5300888..5351644	50756	BOM26_17215	recombinase											Y		
			BOM26_17220	hypothetical protein													
			BOM26_17225	hypothetical protein													
			BOM26_17230	hypothetical protein													
			BOM26_17235	hypothetical protein													
			BOM26_17240	lytic transglycosylase													
			BOM26_17245	conjugal transfer protein TraM													
			BOM26_17250	hypothetical protein													
			BOM26_17255	hypothetical protein													
			BOM26_17260	hypothetical protein													
			BOM26_17265	hypothetical protein													
			BOM26_17270	type IV conjugative transfer system pilin TraA													
			BOM26_17275	type IV conjugative transfer system protein TraL													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_17280	type IV conjugative transfer system protein TraE													
			BOM26_17285	type-F conjugative transfer system secretin TraK													
			BOM26_17290	conjugal transfer protein TrbI													
			BOM26_17295	hypothetical protein													
			BOM26_17300	type IV conjugative transfer system protein TraV													
			BOM26_17305	type-IV secretion system protein TraC													
			BOM26_17310	type-F conjugative transfer system protein TrbI													
			BOM26_17315	hypothetical protein													
			BOM26_17320	hypothetical protein													
			BOM26_17325	type-F conjugative transfer system protein TraW													
			BOM26_17330	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_17335	conjugal transfer protein TraU													
			BOM26_17340	type-F conjugative transfer system pilin assembly protein TrbC													
			BOM26_17345	type-F conjugative transfer system mating-pair stabilization protein TraN													
			BOM26_17350	type-F conjugative transfer system pilin assembly protein TraF													
			BOM26_17355	hypothetical protein													
			BOM26_17360	type-F conjugative transfer system pilin assembly thiol-disulfide isomerase TrbB													
			BOM26_17365	conjugal transfer protein TraH													
			BOM26_17370	conjugal transfer protein TraG													
			BOM26_13310	hypothetical protein													
			BOM26_13315	transposase													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_15130	type IV secretion system protein VirB5													
			BOM26_13300	paraquat-inducible protein B													
			BOM26_12600	hypothetical protein													
			BOM26_12605	hypothetical protein													
			BOM26_15165	transposase													
			BOM26_15170	hypothetical protein													
			BOM26_15135	sugar fermentation stimulation protein SfsA													
			BOM26_12580	hypothetical protein													
			BOM26_12585	rep protein													
			BOM26_10965	hypothetical protein													
			BOM26_10970	rep protein													
			BOM26_10975	mobilization protein													
			BOM26_13290	structural protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_13295	holin													
			BOM26_12590	transketolase													
GI17	OPWLW2_5253017..5282376	29359	BOM25_13530	DNA topoisomerase III													Y
			BOM25_13605	hypothetical protein													
			BOM25_13610	integrase													
			BOM25_13615	hypothetical protein													
			BOM25_13620	hypothetical protein													
			BOM25_13625	SDR family oxidoreductase													
			BOM25_13630	LuxR family transcriptional regulator													
			BOM25_13635	DNA helicase													
			BOM25_13640	relaxase													
			BOM25_13645	recombinase XerD													
			BOM25_13650	chromosome partitioning protein ParA													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_13655	replicative DNA helicase													
			BOM25_13660	hypothetical protein													
			BOM25_13665	hypothetical protein													
			BOM25_13965	hypothetical protein													
			BOM25_13970	single-stranded DNA-binding protein													
			BOM25_13975	hypothetical protein													
			BOM25_13980	pilus assembly protein PilL													
			BOM25_13985	hypothetical protein													
			BOM25_13990	integrating conjugative element protein													
			BOM25_13995	lytic transglycosylase													
			BOM25_14000	conjugal transfer protein													
			BOM25_14005	restriction endonuclease													
			BOM25_12735	conjugal transfer protein TraI													
			BOM25_12740	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_12745	phospholipase D family protein													
			BOM25_12750	type-F conjugative transfer system pilin acetylase TraX													
			BOM25_12755	Replication protein													
			BOM25_13000	hypothetical protein													
			BOM25_13005	conjugal transfer protein TraU													
GI18	OSPLW9_5323865..5352855	28990	BVV03_12535	hypothetical protein												Y	
			BVV03_12540	hypothetical protein													
			BVV03_12545	CsbD family protein													
			BVV03_12550	hypothetical protein													
			BVV03_12555	hypothetical protein													
			BVV03_12560	hypothetical protein													
			BVV03_12565	hypothetical protein													
			BVV03_12570	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BVV03_12575	CAP-Gly protein													
			BVV03_12580	hypothetical protein													
			BVV03_12585	catalase HPII													
			BVV03_12590	hypothetical protein													
			BVV03_12595	hypothetical protein													
			BVV03_11340	integrase													
			BVV03_11255	AAA family ATPase													
			BVV03_11260	entry exclusion protein 1													
			BVV03_11275	hypothetical protein													
			BVV03_11395	transposase													
			BVV03_11310	paraquat-inducible protein B													
			BVV03_09590	mobilization protein													
			BVV03_11370	nuclease													
			BVV03_11240	phage tail protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BVV03_11245	endopeptidase													
			BVV03_11250	phage tail protein													
			BVV03_11345	MFS transporter													
			BVV03_11350	hypothetical protein													
GI19	OLMTLW26_5418865..5442200	23335	BK415_16830	chromosome partitioning protein ParA									Y				
			BK415_16835	stability/partitioning determinant													
			BK415_16840	DNA invertase													
			BK415_16845	hypothetical protein													
			BK415_16850	hypothetical protein													
			BK415_20630	glutathione S-transferase													
			BK415_20635	hypothetical protein													
			BK415_18160	hypothetical protein													
			BK415_18180	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BK415_18605	peptide deformylase													
			BK415_18610	hypothetical protein													
			BK415_18650	ornithine decarboxylase													
			BK415_20285	hypothetical protein													
			BK415_18625	aliphatic sulfonate ABC transporter substrate-binding protein													
			BK415_16895	Replication protein													
			BK415_20280	hypothetical protein													
GI20	OSPLW9_5360831..5380823	19992	BVV03_11315	paraquat-inducible protein B												Y	
			BVV03_09085	hypothetical protein													
			BVV03_09090	rep protein													
			BVV03_09095	helix-turn-helix domain-containing protein													
			BVV03_09100	rep protein													
			BVV03_11385	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BVV03_11290	structural protein													
			BVV03_11295	holin													
			BVV03_09105	mobilization protein													
			BVV03_09110	hypothetical protein													
			BVV03_09115	hypothetical protein													
			BVV03_09120	mobilization protein													
			BVV03_09125	mobilization protein													
			BVV03_11320	entry exclusion protein 1													
			BVV03_11325	transposase													
			BVV03_11390	arginine ABC transporter ATP-binding protein ArtP													
			BVV03_11330	cupin													
GI21	OPWLW2_5385241..5401997	16756	BOM25_16070	transposase													Y
			BOM25_16075	ATPase													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_15805	integrating conjugative element protein													
			BOM25_15810	integrating conjugative element protein													
			BOM25_14985	integrating conjugative element protein													
			BOM25_14990	integrating conjugative element protein													
			BOM25_14995	integrating conjugative element protein													
			BOM25_15000	hypothetical protein													
			BOM25_16025	hypothetical protein													
			BOM25_16030	hypothetical protein													
			BOM25_16035	hypothetical protein													
			BOM25_16040	hypothetical protein													
			BOM25_15885	phage tail tape measure protein													
			BOM25_15890	hypothetical protein													
			BOM25_15895	phage tail protein													
			BOM25_15900	phage tail protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_15905	hypothetical protein													
GI22	OLBL_5258066..5274131	16065	BMF92_22125	hypothetical protein	Y												
			BMF92_22130	hypothetical protein													
			BMF92_22135	hypothetical protein													
			BMF92_22140	hypothetical protein													
			BMF92_22145	hypothetical protein													
			BMF92_22150	hypothetical protein													
			BMF92_22155	hypothetical protein													
			BMF92_22160	DUF2184 domain-containing protein													
			BMF92_22165	hypothetical protein													
			BMF92_22170	hypothetical protein													
			BMF92_22175	phage head morphogenesis protein													
			BMF92_22180	hypothetical protein													
			BMF92_22185	TerL protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BMF92_2 2190	terminase small subunit													
			BMF92_2 2195	hypothetical protein													
			BMF92_2 4675	relaxase													
GI23	OLHL_5269072..5283 744	14672	BMF90_1 7910	hypothetical protein						Y							
			BMF90_1 7915	DNA polymerase V subunit UmuC													
			BMF90_1 7920	DNA polymerase V subunit UmuD													
			BMF90_1 7925	hypothetical protein													
			BMF90_1 7975	integrase													
			BMF90_1 7980	GTP-binding protein													
GI24	OPWLW2_2077456.. 2091785	14329	BOM25_1 6545	peptidase P60													Y
			BOM25_1 6540	hypothetical protein													
			BOM25_1 6535	phage tail protein													
			BOM25_1 6530	host specificity protein													
			BOM25_1 6525	hypothetical protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM25_16520	hypothetical protein													
			BOM25_16515	hypothetical protein													
			BOM25_16510	DNA polymerase V													
			BOM25_16505	hypothetical protein													
			BOM25_16500	hypothetical protein													
			BOM25_16495	IS5 family transposase													
			BOM25_16490	hypothetical protein													
			BOM25_16485	hypothetical protein													
GI25	OLDL_5323768..5332189	8421	BMF88_24990	30S ribosomal protein S2			Y										
			BMF88_12930	hypothetical protein													
			BMF88_24950	hypothetical protein													
			BMF88_24955	hypothetical protein													
			BMF88_24960	hypothetical protein													
			BMF88_24975	GTP-binding protein													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
GI26	OLIL_4450954..4456528	5574	BMH23_10905	hypothetical protein							Y						
			BMH23_13600	hypothetical protein													
			BMH23_13605	hypothetical protein													
GI27	OSPLW9_969850..975288	5438	BVV03_09155	DNA-binding response regulator												Y	
			BVV03_09150	two-component sensor histidine kinase													
			BVV03_09145	efflux transporter periplasmic adaptor subunit													
			BVV03_09140	ACR family transporter													
GI28	OLBL_3941812..3947183	5371	BMF92_24530	FAD-linked oxidase	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	
			BMF92_08785	hypothetical protein													
			BMF92_08790	FMN-dependent NADH-azoreductase													
GI29	OPNLW1_4260592..4265795	5203	BOM26_10355	integrase											Y		
			BOM26_10350	helix-turn-helix transcriptional regulator													

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_10345	LuxR family transcriptional regulator													
			BOM26_10340	fimbrial protein													
			BOM26_10335	fimbrial protein													
			BOM26_10330	fimbrial assembly protein													
GI30	OLCL_2434158..2438977	4819	BMF85_21550	hypothetical protein	Y (80)	Y				Y							
			BMF85_24910	hypothetical protein													
			BMF85_24915	hypothetical protein													
			BMF85_11105	hypothetical protein													
			BMF85_11100	hypothetical protein													
GI31	OLHL_537833..541885	4052	BMF90_05005	integrase						Y							
			BMF90_17100	hypothetical protein													
			BMF90_17095	hypothetical protein													
GI32	OPNLW1_539334..549540	10206	BOM26_06645	integrase				Y (55)					Y		Y	Y	

GI No.	Representative GI	Length (bp)	Locus ID	Annotated features	OLB L1	OLC L1	OLD L1	OLE L1	OLF L2	OLH L2	OLI L2	OLJ L1	OLMTL W26	OLLOL W30	OPNL W1	OSPL W9	OPWL W2
			BOM26_06640	SAM-dependent methyltransferase													
			BOM26_06635	hypothetical protein													
			BOM26_06630	hypothetical protein													
			BOM26_06625	hypothetical protein													
			BOM26_06620	transposase													
			BOM26_06615	transposase													
			BOM26_06610	HNH endonuclease													
			BOM26_06605	hypothetical protein													

Swansea University
Prifysgol Abertawe

Supplementary data 7-1: Scoary output_genes that were found to be associated with the trait.

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2026		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2027		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2733		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2666		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2564		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2036		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2643		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2626		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1097		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1095		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1093		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1091		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1099		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2047		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1984		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1242		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1240		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1245		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1244		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2251		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2253		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2340		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_573		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1354		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2559		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2417		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1458		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2552		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2553		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2419		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2551		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2556		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2557		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2554		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2043		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2118		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2707		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2706		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2705		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2704		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2703		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2702		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2701		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2709		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2708		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1236		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2675		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1616		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1613		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1619		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2057		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2056		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2336		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_462		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2458		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2456		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2455		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2453		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2451		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2450		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2677		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2562		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2069		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1968		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2429		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2422		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2423		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2420		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2421		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2426		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2427		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2424		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2425		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2748		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2637		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2711		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2713		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1941		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1100		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2497		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1101		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2083		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2435		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2125		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2124		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1237		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1234		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2341		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1230		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1231		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1239		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2062		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2064		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1456		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2549		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2548		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2466		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2467		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2460		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1539		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2543		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2542		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2545		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2544		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2546		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2013		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2547		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2261		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2262		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2239		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2718		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2164		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2162		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2174		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2160		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2760		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1567		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2767		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2768		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1568		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1569		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2024		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2022		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2021		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2663		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_575		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_577		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2135		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_579		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2501		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2337		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2719		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2007		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2439		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2438		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2431		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2430		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2433		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2344		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1944		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2434		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2436		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2665		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2437		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_3045		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2661		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2756		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1133		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2720		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2722		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2153		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1964		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2361		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2363		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2428		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2652		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1229		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2071		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2070		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2073		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2072		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2075		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2468		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2474		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2470		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2473		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2472		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2029		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_1566		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
rclR_3		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2558		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2432		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2550		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2555		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2443		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2750		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2035		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2031		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2033		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2039		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2038		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2538		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2444		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1313		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2245		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2241		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2407		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_1953		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2403		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2243		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2710		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2712		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2714		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2715		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2716		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2717		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1943		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_1955		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2045		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1
group_2040		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2042		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2664		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2448		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2445		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2446		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2447		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2563		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1
group_2561		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226202 8	6.88E-05	1	1	0	1	1

Gene	Non-unique Gene name	Annotation	Number_pos_present_in	Number_neg_present_in	Number_pos_not_present_in	Number_neg_not_present_in	Sensitivity	Specificity	Odds_ratio	Naive_p	Bonferroni_p	Benjamini_H_p	Max_Pairwise_comparisons	Max_supporting_pairs	Max_opposing_pairs	Best_pairwise_comp_p	Worst_pairwise_comp_p
group_2560		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226208	6.88E-05	1	1	0	1	1
group_2566		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226208	6.88E-05	1	1	0	1	1
group_2565		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226208	6.88E-05	1	1	0	1	1
group_2765		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226208	6.88E-05	1	1	0	1	1
group_107		hypothetical protein	13	0	0	8	100	100	inf	4.91E-06	0.0226208	6.88E-05	1	1	0	1	1

Supplementary data Table 7-2: List of ROARY predicted *Serratia* sp. *Orius* isolates accessory genome associated to GI number.

Query Name	Top Hit Identifier	Description
00692-acrR_2.fa.aln_1	TetR_N	Bacterial regulatory proteins, tetR family
04575-aldA.fa.aln_1	Aldedh	Aldehyde dehydrogenase family
03694-argT_1.fa.aln_1	SBP_bac_3	Bacterial extracellular solute-binding proteins, family 3
02771-azoR1.fa.aln_1	Flavodoxin_2	Flavodoxin-like fold
02772-azoR_2.fa.aln_1	Flavodoxin_2	Flavodoxin-like fold
00165-btuB.fa.aln_1	TonB_dep_Rec	TonB dependent receptor
04863-cpsD.fa.aln_1	AAA_31	AAA domain
00570-csd.fa.aln_1	Aminotran_5	Aminotransferase class V
00574-cysK_2.fa.aln_1	PALP	Pyridoxal-phosphate dependent enzyme
03339-dinB1_2.fa.aln_1	IMS	impB/mucB/samB family
04581-dmlR_24.fa.aln_1	LysR_substrate	LysR substrate binding domain
04865-dnaB_2.fa.aln_1	DnaB_C	DnaB-like helicase C terminal domain
01624-dpnA.fa.aln_1	N6_N4_Mtase	DNA methylase
04582-eamB_4.fa.aln_1	LysE	LysE type translocator
00695-fabG_1.fa.aln_1	adh_short	short chain dehydrogenase
04576-fabG_9.fa.aln_1	adh_short_C2	Enoyl-(Acyl carrier protein) reductase
00576-fcuA.fa.aln_1	TonB_dep_Rec	TonB dependent receptor
02625-gltC_2.fa.aln_1	LysR_substrate	LysR substrate binding domain
00700-group_105.fa.aln_1	Ank	Ankyrin repeat
01619-group_1093.fa.aln_1	GP46	Phage protein GP46
01612-group_1095.fa.aln_1	Phage_sheath_1	Phage tail sheath protein subtilisin-like domain
01609-group_1097.fa.aln_1	DUF1320	Protein of unknown function (DUF1320)
01604-group_1099.fa.aln_1	Phage_Mu_F	Phage Mu protein F like protein
01596-group_1101.fa.aln_1		
01584-group_1102.fa.aln_1	AAA_22	AAA domain
01583-group_1103.fa.aln_1	rve	Integrase core domain
00232-group_1135.fa.aln_1	Phage_GPA	Bacteriophage replication gene A protein (GPA)
04864-group_1231.fa.aln_1		
04868-group_1232.fa.aln_1		
04869-group_1233.fa.aln_1	DUF2857	Protein of unknown function (DUF2857)
04870-group_1236.fa.aln_1		
04873-group_1238.fa.aln_1		
04884-group_1239.fa.aln_1	TraG-D_C	TraM recognition site of TraD and TraG
04890-group_1241.fa.aln_1	DUF2895	Protein of unknown function (DUF2895)
04896-group_1242.fa.aln_1		
04899-group_1243.fa.aln_1	N6_Mtase	N-6 DNA Methylase
04904-group_1244.fa.aln_1	TraU	TraU protein

Query Name	Top Hit Identifier	Description
04905-group_1246.fa.aln_1		
04907-group_1247.fa.aln_1	TraG_N	TraG-like protein, N-terminal region
01491-group_1315.fa.aln_1	Trans_reg_C	Transcriptional regulatory protein, C terminal
01910-group_1356.fa.aln_1		
00036-group_1464.fa.aln_1	TolA	TolA C-terminal
00894-group_1545.fa.aln_1		
00489-group_1572.fa.aln_1	DinI	DinI-like family
00490-group_1573.fa.aln_1		
00494-group_1574.fa.aln_1		
00497-group_1575.fa.aln_1		
00502-group_1576.fa.aln_1		
04919-group_1623.fa.aln_1		
04916-group_1626.fa.aln_1	DUF3085	Protein of unknown function (DUF3085)
04915-group_1629.fa.aln_1		
00759-group_1938.fa.aln_1	LysE	LysE type translocator
00697-group_1945.fa.aln_1	ADH_zinc_N_2	Zinc-binding dehydrogenase
00696-group_1946.fa.aln_1		
00694-group_1948.fa.aln_1	adh_short	short chain dehydrogenase
00693-group_1949.fa.aln_1		
00940-group_1955.fa.aln_1	LysR_substrate	LysR substrate binding domain
00460-group_1958.fa.aln_1	Oxygenase-NA	Oxygenase, catalysing oxidative methylation of damaged DNA
00445-group_1960.fa.aln_1		
04013-group_1969.fa.aln_1	TolA	TolA C-terminal
03627-group_1973.fa.aln_1		
03695-group_1983.fa.aln_1	Peptidase_M20	Peptidase family M20/M25/M40
03078-group_1989.fa.aln_1		
04540-group_2012.fa.aln_1		
01641-group_2018.fa.aln_1	Tai4	Type VI secretion system (T6SS), amidase immunity protein
01617-group_2026.fa.aln_1		
01616-group_2027.fa.aln_1	DNA_circ_N	DNA circularisation protein N-terminus
01614-group_2029.fa.aln_1		
01613-group_2031.fa.aln_1		
01611-group_2032.fa.aln_1	DUF2635	Protein of unknown function (DUF2635)
01610-group_2034.fa.aln_1	DUF1834	Domain of unknown function (DUF1834)
01608-group_2036.fa.aln_1	Mu-like_gpT	Mu-like prophage major head subunit gpT
01607-group_2038.fa.aln_1		
01605-group_2040.fa.aln_1	Phage_tail_S	Phage virion morphogenesis family
01603-group_2041.fa.aln_1	DUF935	Protein of unknown function (DUF935)
01602-group_2043.fa.aln_1		

Query Name	Top Hit Identifier	Description
01601-group_2044.fa.aln_1		
01600-group_2045.fa.aln_1		
01599-group_2047.fa.aln_1	DUF1804	Protein of unknown function (DUF1804)
01598-group_2048.fa.aln_1		
01595-group_2050.fa.aln_1		
01593-group_2052.fa.aln_1	Mor	Mor transcription activator family
01582-group_2061.fa.aln_1	HTH_35	Winged helix-turn-helix DNA-binding
01581-group_2062.fa.aln_1	HTH_3	Helix-turn-helix
01346-group_2067.fa.aln_1		
01344-group_2069.fa.aln_1		
00233-group_2074.fa.aln_1		
00234-group_2075.fa.aln_1	DUF5405	Domain of unknown function (DUF5405)
00236-group_2076.fa.aln_1	DUF5405	Domain of unknown function (DUF5405)
00237-group_2077.fa.aln_1	DUF2732	Protein of unknown function (DUF2732)
00238-group_2078.fa.aln_1		
00240-group_2080.fa.aln_1	HTH_31	Helix-turn-helix domain
00628-group_2088.fa.aln_1		
02770-group_2120.fa.aln_1	ADH_zinc_N_2	Zinc-binding dehydrogenase
02769-group_2121.fa.aln_1	LysR_substrate	LysR substrate binding domain
00231-group_2123.fa.aln_1		
00219-group_2129.fa.aln_1	Phage_tail_S	Phage virion morphogenesis family
00218-group_2130.fa.aln_1	AIPR	AIPR protein
04972-group_2140.fa.aln_1		
04875-group_2158.fa.aln_1	DUF3577	Protein of unknown function (DUF3577)
04885-group_2165.fa.aln_1	DUF4400	Domain of unknown function (DUF4400)
04886-group_2167.fa.aln_1	Plasmid_RAQPRD	Plasmid protein of unknown function (Plasmid_RAQPRD)
04887-group_2169.fa.aln_1	DUF3262	Protein of unknown function (DUF3262)
04903-group_2179.fa.aln_1	DUF1525	Protein of unknown function (DUF1525)
02295-group_2241.fa.aln_1	adh_short_C2	Enoyl-(Acyl carrier protein) reductase
02296-group_2242.fa.aln_1	LysR_substrate	LysR substrate binding domain
02299-group_2244.fa.aln_1	DUF962	Protein of unknown function (DUF962)
00544-group_2246.fa.aln_1	DUF4180	Domain of unknown function (DUF4180)
00571-group_2248.fa.aln_1	Peripla_BP_2	Periplasmic binding protein
00573-group_2250.fa.aln_1	Octopine_DH	NAD/NADP octopine/nopaline dehydrogenase, alpha-helical domain
01478-group_2256.fa.aln_1		
01480-group_2258.fa.aln_1	FimH_man-bind	FimH, mannose binding
01487-group_2264.fa.aln_1		
01488-group_2265.fa.aln_1		
01489-group_2266.fa.aln_1	Trans_reg_C	Transcriptional regulatory protein, C terminal

Query Name	Top Hit Identifier	Description
01490-group_2267.fa.aln_1		
00825-group_2342.fa.aln_1		
03758-group_2345.fa.aln_1		
03755-group_2346.fa.aln_1		
03752-group_2349.fa.aln_1	Arm-DNA-bind_3	Arm DNA-binding domain
01515-group_2366.fa.aln_1		
01513-group_2368.fa.aln_1		
01485-group_238.fa.aln_1	PapD_N	Pili and flagellar-assembly chaperone, PapD N-terminal domain
02043-group_2403.fa.aln_1	Phage_integrase	Phage integrase family
02042-group_2404.fa.aln_1	GerE	Bacterial regulatory proteins, luxR family
04559-group_2408.fa.aln_1	Barstar	Barstar (barnase inhibitor)
04570-group_2409.fa.aln_1	Acetyltransf_10	Acetyltransferase (GNAT) domain
04572-group_2411.fa.aln_1	A_deaminase	Adenosine/AMP deaminase
04573-group_2412.fa.aln_1	MFS_1	Major Facilitator Superfamily
04580-group_2418.fa.aln_1	adh_short	short chain dehydrogenase
01864-group_2422.fa.aln_1	Phage_antiter_Q	Phage antitermination protein Q
01867-group_2424.fa.aln_1		
01868-group_2425.fa.aln_1	Phg_2220_C	Conserved phage C-terminus (Phg_2220_C)
01869-group_2426.fa.aln_1	DUF4222	Domain of unknown function (DUF4222)
01870-group_2427.fa.aln_1	DUF4222	Domain of unknown function (DUF4222)
01871-group_2428.fa.aln_1		
01872-group_2429.fa.aln_1		
01873-group_2430.fa.aln_1		
01874-group_2431.fa.aln_1		
01875-group_2432.fa.aln_1		
01876-group_2433.fa.aln_1		
01877-group_2434.fa.aln_1	DUF2303	Uncharacterized conserved protein (DUF2303)
01878-group_2435.fa.aln_1		
01879-group_2436.fa.aln_1	DUF1482	Protein of unknown function (DUF1482)
01880-group_2437.fa.aln_1	Methyltransf_11	Methyltransferase domain
01881-group_2438.fa.aln_1		
01882-group_2439.fa.aln_1		
01883-group_2440.fa.aln_1		
01884-group_2441.fa.aln_1		
01885-group_2442.fa.aln_1	DUF5051	3' exoribonuclease, RNase T-like
01886-group_2443.fa.aln_1	Exc	Excisionase-like protein
01887-group_2444.fa.aln_1	Phage_integrase	Phage integrase family
03362-group_2448.fa.aln_1	Vut_1	Putative vitamin uptake transporter
03361-group_2449.fa.aln_1		

Query Name	Top Hit Identifier	Description
03358-group_2451.fa.aln_1	PsiA	PsiA protein
03348-group_2457.fa.aln_1		
03338-group_2466.fa.aln_1		
03337-group_2467.fa.aln_1		
03334-group_2470.fa.aln_1		
03331-group_2472.fa.aln_1		
03330-group_2473.fa.aln_1		
00035-group_2474.fa.aln_1	Acetyltransf_3	Acetyltransferase (GNAT) domain
02794-group_2496.fa.aln_1	Ytca	Uncharacterised protein family
02785-group_2500.fa.aln_1	Abhydrolase_6	Alpha/beta hydrolase family
00880-group_2537.fa.aln_1	Tautomerase_2	Tautomerase enzyme
04100-group_2541.fa.aln_1		
04101-group_2542.fa.aln_1	DUF2313	Uncharacterised protein conserved in bacteria (DUF2313)
04102-group_2543.fa.aln_1	Baseplate_J	Baseplate J-like protein
04103-group_2544.fa.aln_1	GP46	Phage protein GP46
04104-group_2545.fa.aln_1	Phage_Mu_Gp45	Bacteriophage Mu Gp45 protein
04105-group_2546.fa.aln_1		
04106-group_2547.fa.aln_1	DNA_circ_N	DNA circularisation protein N-terminus
04107-group_2548.fa.aln_1	PhageMin_Tail	Phage-related minor tail protein
04108-group_2549.fa.aln_1	Phage_TAC_7	Phage tail assembly chaperone proteins, E, or 41 or 14
04109-group_2550.fa.aln_1	Tail_tube	Phage tail tube protein
04110-group_2551.fa.aln_1	Phage_sheath_1	Phage tail sheath protein subtilisin-like domain
04111-group_2552.fa.aln_1		
04112-group_2553.fa.aln_1		
04113-group_2554.fa.aln_1	Phage_H_T_join	Phage head-tail joining protein
04114-group_2555.fa.aln_1	Phage_connect_1	Phage gp6-like head-tail connector protein
04115-group_2556.fa.aln_1	Phage_capsid	Phage capsid family
04117-group_2557.fa.aln_1	Phage_portal	Phage portal protein
04118-group_2558.fa.aln_1		
04119-group_2559.fa.aln_1	Terminase_1	Phage Terminase
04120-group_2560.fa.aln_1	Terminase_4	Phage terminase, small subunit
04121-group_2561.fa.aln_1		
04122-group_2562.fa.aln_1		
04123-group_2563.fa.aln_1		
04124-group_2564.fa.aln_1		
04125-group_2565.fa.aln_1		
00132-group_2625.fa.aln_1	DDE_Tnp_1	Transposase DDE domain
03049-group_2636.fa.aln_1		
02478-group_2642.fa.aln_1	Trans_reg_C	Transcriptional regulatory protein, C terminal

Query Name	Top Hit Identifier	Description
02480-group_2644.fa.aln_1	PapD_N	Pili and flagellar-assembly chaperone, PapD N-terminal domain
00488-group_2651.fa.aln_1	DUF1367	Protein of unknown function (DUF1367)
00503-group_2660.fa.aln_1		
00505-group_2662.fa.aln_1	RecT	RecT family
00506-group_2663.fa.aln_1		
00507-group_2664.fa.aln_1	Exc	Excisionase-like protein
00508-group_2665.fa.aln_1	Arm-DNA-bind_1	Bacteriophage lambda integrase, Arm DNA-binding domain
01860-group_2674.fa.aln_1	PP-binding	Phosphopantetheine attachment site
00900-group_2676.fa.aln_1	Poly_export	Polysaccharide biosynthesis/export protein
03385-group_2699.fa.aln_1	TraH	Conjugative relaxosome accessory transposon protein
03384-group_2700.fa.aln_1	TraF	F plasmid transfer operon protein
03383-group_2701.fa.aln_1	TraQ	Type-F conjugative transfer system pilin chaperone (TraQ)
03382-group_2702.fa.aln_1	TraF	F plasmid transfer operon protein
03377-group_2706.fa.aln_1		
03375-group_2708.fa.aln_1	DSBA	DSBA-like thioredoxin domain
03374-group_2709.fa.aln_1		
03373-group_2710.fa.aln_1	TrbI_Ftype	Type-F conjugative transfer system protein (TrbI_Ftype)
03372-group_2711.fa.aln_1	TraC_F_IV	F pilus assembly Type-IV secretion system for plasmid transfer
03371-group_2712.fa.aln_1	TraV	Type IV conjugative transfer system lipoprotein (TraV)
03370-group_2713.fa.aln_1		
03369-group_2714.fa.aln_1	TrbI	Bacterial conjugation TrbI-like protein
03368-group_2715.fa.aln_1	TraK	TraK protein
03367-group_2716.fa.aln_1	TraE	TraE protein
03366-group_2717.fa.aln_1	TraL	TraL protein
04542-group_2719.fa.aln_1		
02669-group_2730.fa.aln_1		
01310-group_2745.fa.aln_1	HNH	HNH endonuclease
01307-group_2747.fa.aln_1		
04912-group_2753.fa.aln_1		
00578-group_2756.fa.aln_1	HpcH_HpaI	HpcH/HpaI aldolase/citrate lyase family
02676-group_2757.fa.aln_1	Relaxase	Relaxase/Mobilisation nuclease domain
00134-group_2762.fa.aln_1		
02609-group_2764.fa.aln_1	Phage_holin_3_3	LydA holin phage, holin superfamily III
02608-group_2765.fa.aln_1		
01865-group_3042.fa.aln_1	DUF968	Protein of unknown function (DUF968)
04830-group_459.fa.aln_1	Radical_SAM_N	Radical SAM N-terminal
04878-group_569.fa.aln_1		
04880-group_571.fa.aln_1		

Query Name	Top Hit Identifier	Description
04881-group_573.fa.aln_1	SLT	Transglycosylase SLT domain
04891-group_575.fa.aln_1	DUF3438	Protein of unknown function (DUF3438)
01345-hsdM.fa.aln_1	N6_Mtase	N-6 DNA Methylase
04897-hsdR.fa.aln_1	EcoR124_C	Type I restriction and modification enzyme - subunit R C terminal
02481-htrE.fa.aln_1	Usher	Outer membrane usher protein
02681-intA.fa.aln_1	Phage_integrase	Phage integrase family
00577-iucC.fa.aln_1	IucA_IucC	IucA / IucC family
02795-kstR.fa.aln_1	TetR_N	Bacterial regulatory proteins, tetR family
01484-lpfA_4.fa.aln_1		
02039-lpfD.fa.aln_1	Fimbrial	Fimbrial protein
00575-lysA_1.fa.aln_1	Orn_Arg_deC_N	Pyridoxal-dependent decarboxylase, pyridoxal binding domain
00704-mdtA_1.fa.aln_1	HlyD_D23	Barrel-sandwich domain of CusB or HlyD membrane-fusion
02793-mdtN_3.fa.aln_1	HlyD_3	HlyD family secretion protein
02792-mdtO.fa.aln_1	FUSC	Fusaric acid resistance protein family
02482-mrkD.fa.aln_1	Fimbrial	Fimbrial protein
01479-nreC_1.fa.aln_1	GerE	Bacterial regulatory proteins, luxR family
01486-papC_2.fa.aln_1	Usher	Outer membrane usher protein
03356-parB.fa.aln_1	ParBc	ParB-like nuclease domain
00941-pbuE_1.fa.aln_1	MFS_1	Major Facilitator Superfamily
04492-pcpR_2.fa.aln_1	LysR_substrate	LysR substrate binding domain
01514-pgrR_11.fa.aln_1		
03335-pld.fa.aln_1	PLDc_2	PLD-like domain
02784-rclR_1.fa.aln_1	Cupin_6	Cupin
04574-rclR_3.fa.aln_1	Cupin_6	Cupin
02041-rcsB_3.fa.aln_1	GerE	Bacterial regulatory proteins, luxR family
04973-relE4.fa.aln_1	ParE_toxin	ParE toxin of type II toxin-antitoxin system, parDE
04578-rhaS_4.fa.aln_1	HTH_18	Helix-turn-helix domain
04577-rhtB_2.fa.aln_1	LysE	LysE type translocator
04571-rihA_2.fa.aln_1	IU_nuc_hydro	Inosine-uridine preferring nucleoside hydrolase
02040-sfaS.fa.aln_1		
01483-smfA_3.fa.aln_1	Fimbrial	Fimbrial protein
00572-tetA_1.fa.aln_1	MFS_1	Major Facilitator Superfamily
03753-tetC.fa.aln_1	TetR_N	Bacterial regulatory proteins, tetR family
03365-traA.fa.aln_1	TraA	TraA
04917-traC_3.fa.aln_1	DUF1738	Domain of unknown function (DUF1738)
03332-traD.fa.aln_1	TrwB_AAD_bind	Type IV secretion-system coupling protein DNA-binding domain
03333-traI.fa.aln_1	TrwC	TrwC relaxase
03754-trxB_2.fa.aln_1	Pyr_redox_2	Pyridine nucleotide-disulphide oxidoreductase
03696-tsaR.fa.aln_1	LysR_substrate	LysR substrate binding domain

Query Name	Top Hit Identifier	Description
02476-tufB.fa.aln_1	GTP_EFTU_D3	Elongation factor Tu C-terminal domain
02297-udh.fa.aln_1	Epimerase	NAD dependent epimerase/dehydratase family
01481-yfcQ_2.fa.aln_1	Fimbrial	Fimbrial protein
01482-yfcR_1.fa.aln_1	Fimbrial	Fimbrial protein
01866-yhdJ.fa.aln_1	N6_N4_Mtase	DNA methylase
01492-yjdF.fa.aln_1	DUF2238	Predicted membrane protein (DUF2238)
04918-ykfI.fa.aln_1	CbtA_toxin	CbtA_toxin of type IV toxin-antitoxin system
02633-ytnP_1.fa.aln_1	Lactamase_B	Metallo-beta-lactamase superfamily
04579-ytnP_2.fa.aln_1	Lactamase_B	Metallo-beta-lactamase superfamily

Supplementary data 7-3: KEGG pathways with the KEGG major and sub-categories in the pangenome.

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
Cellular_Processes	Cell_growth_and_death	04110 Cell cycle [PATH:ko04110]	0	0	0
		04111 Cell cycle - yeast [PATH:ko04111]	0	0	0
		04112 Cell cycle - Caulobacter [PATH:ko04112]	12	0	1
		04113 Meiosis - yeast [PATH:ko04113]	0	0	0
		04114 Oocyte meiosis [PATH:ko04114]	0	0	0
		04115 p53 signaling pathway [PATH:ko04115]	0	0	0
		04210 Apoptosis [PATH:ko04210]	0	0	0
	Cell_motility	02030 Bacterial chemotaxis [PATH:ko02030]	20	5	1
		02040 Flagellar assembly [PATH:ko02040]	35	2	1
		04810 Regulation of actin cytoskeleton [PATH:ko04810]	0	0	0
	Cellular community	04510 Focal adhesion [PATH:ko04510]	0	0	0
		04520 Adherens junction [PATH:ko04520]	0	0	0
		04530 Tight junction [PATH:ko04530]	0	0	0
		04540 Gap junction [PATH:ko04540]	0	0	0
		04550 Signaling pathways regulating pluripotency of stem cells [PATH:ko04550]	0	0	0
	Transport_and_catabolism	04140 Regulation of autophagy [PATH:ko04140]	0	0	0
		04142 Lysosome [PATH:ko04142]	2	0	0
		04144 Endocytosis [PATH:ko04144]	0	0	0
		04145 Phagosome [PATH:ko04145]	0	0	0
		04146 Peroxisome [PATH:ko04146]	9	2	0
Environmental_Information_Processing	Membrane_transport	02010 ABC transporters [PATH:ko02010]	207	40	1
		02060 Phosphotransferase system (PTS) [PATH:ko02060]	30	7	2
		03070 Bacterial secretion system [PATH:ko03070]	17	13	0
	Signal_transduction	02020 Two-component system [PATH:ko02020]	122	22	5
		04010 MAPK signaling pathway [PATH:ko04010]	0	0	0
		04011 MAPK signaling pathway - yeast [PATH:ko04011]	3	0	0
		04012 ErbB signaling pathway [PATH:ko04012]	0	0	0
		04013 MAPK signaling pathway - fly [PATH:ko04013]	0	0	0
		04014 Ras signaling pathway [PATH:ko04014]	0	0	0
		04015 Rap1 signaling pathway [PATH:ko04015]	0	0	0
		04020 Calcium signaling pathway [PATH:ko04020]	0	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		04022 cGMP - PKG signaling pathway [PATH:ko04022]	0	0	0
		04024 cAMP signaling pathway [PATH:ko04024]	0	0	0
		04064 NF-kappa B signaling pathway [PATH:ko04064]	0	0	0
		04066 HIF-1 signaling pathway [PATH:ko04066]	3	0	0
		04068 FoxO signaling pathway [PATH:ko04068]	3	1	0
		04070 Phosphatidylinositol signaling system [PATH:ko04070]	5	0	0
		04071 Sphingolipid signaling pathway [PATH:ko04071]	0	0	0
		04075 Plant hormone signal transduction [PATH:ko04075]	0	0	0
		04150 mTOR signaling pathway [PATH:ko04150]	0	0	0
		04151 PI3K-Akt signaling pathway [PATH:ko04151]	1	0	0
		04152 AMPK signaling pathway [PATH:ko04152]	2	0	0
		04310 Wnt signaling pathway [PATH:ko04310]	0	1	0
		04330 Notch signaling pathway [PATH:ko04330]	0	0	0
		04340 Hedgehog signaling pathway [PATH:ko04340]	0	0	0
		04350 TGF-beta signaling pathway [PATH:ko04350]	0	0	0
		04370 VEGF signaling pathway [PATH:ko04370]	0	0	0
		04390 Hippo signaling pathway [PATH:ko04390]	0	0	0
		04391 Hippo signaling pathway -fly [PATH:ko04391]	0	0	0
		04630 Jak-STAT signaling pathway [PATH:ko04630]	0	0	0
		04668 TNF signaling pathway [PATH:ko04668]	0	0	0
	Signaling_molecules_and_interaction	04060 Cytokine-cytokine receptor interaction [PATH:ko04060]	0	0	0
		04080 Neuroactive ligand-receptor interaction [PATH:ko04080]	0	0	0
		04512 ECM-receptor interaction [PATH:ko04512]	0	0	0
		04514 Cell adhesion molecules (CAMs) [PATH:ko04514]	0	0	0
Genetic_Information_Processing	Folding,_sorting_and_degradation	03018 RNA degradation [PATH:ko03018]	16	0	0
		03050 Proteasome [PATH:ko03050]	0	0	0
		03060 Protein export [PATH:ko03060]	17	0	0
		04120 Ubiquitin mediated proteolysis [PATH:ko04120]	0	0	0
		04122 Sulfur relay system [PATH:ko04122]	15	2	0
		04130 SNARE interactions in vesicular transport [PATH:ko04130]	0	0	0
		04141 Protein processing in endoplasmic reticulum [PATH:ko04141]	1	1	1
	Replication_and_repair	03030 DNA replication [PATH:ko03030]	15	2	1

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		03410 Base excision repair [PATH:ko03410]	13	1	0
		03420 Nucleotide excision repair [PATH:ko03420]	7	1	0
		03430 Mismatch repair [PATH:ko03430]	20	3	1
		03440 Homologous recombination [PATH:ko03440]	24	3	1
		03450 Non-homologous end-joining [PATH:ko03450]	0	0	0
		03460 Fanconi anemia pathway [PATH:ko03460]	0	0	0
	Transcription	03020 RNA polymerase [PATH:ko03020]	4	0	0
		03022 Basal transcription factors [PATH:ko03022]	0	0	0
		03040 Spliceosome [PATH:ko03040]	0	0	0
	Translation	00970 Aminoacyl-tRNA biosynthesis [PATH:ko00970]	26	2	2
		03008 Ribosome biogenesis in eukaryotes [PATH:ko03008]	3	0	0
		03010 Ribosome [PATH:ko03010]	56	0	0
		03013 RNA transport [PATH:ko03013]	2	0	0
		03015 mRNA surveillance pathway [PATH:ko03015]	0	0	0
Human_Diseases	Cancers	05200 Pathways in cancer [PATH:ko05200]	2	0	0
		05202 Transcriptional misregulation in cancers [PATH:ko05202]	0	0	0
		05203 Viral carcinogenesis [PATH:ko05203]	2	1	0
		05204 Chemical carcinogenesis [PATH:ko05204]	5	2	0
		05205 Proteoglycans in cancer [PATH:ko05205]	1	0	0
		05206 MicroRNAs in cancer [PATH:ko05206]	1	1	3
		05210 Colorectal cancer [PATH:ko05210]	0	0	0
		05211 Renal cell carcinoma [PATH:ko05211]	1	0	0
		05212 Pancreatic cancer [PATH:ko05212]	0	0	0
		05213 Endometrial cancer [PATH:ko05213]	0	0	0
		05214 Glioma [PATH:ko05214]	0	0	0
		05215 Prostate cancer [PATH:ko05215]	1	0	0
		05216 Thyroid cancer [PATH:ko05216]	0	0	0
		05217 Basal cell carcinoma [PATH:ko05217]	0	0	0
		05218 Melanoma [PATH:ko05218]	0	0	0
		05219 Bladder cancer [PATH:ko05219]	1	0	0
		05220 Chronic myeloid leukemia [PATH:ko05220]	0	0	0
		05221 Acute myeloid leukemia [PATH:ko05221]	0	0	0
		05222 Small cell lung cancer [PATH:ko05222]	0	0	0
		05223 Non-small cell lung cancer [PATH:ko05223]	0	0	0
		05230 Central carbon metabolism in cancer [PATH:ko05230]	6	1	1
		05231 Choline metabolism in cancer [PATH:ko05231]	2	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
	Cardiovascular_diseases	05410 Hypertrophic cardiomyopathy (HCM) [PATH:ko05410]	0	0	0
		05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC) [PATH:ko05412]	0	0	0
		05414 Dilated cardiomyopathy (DCM) [PATH:ko05414]	0	0	0
		05416 Viral myocarditis [PATH:ko05416]	0	0	0
	Drug_resistance	01501 beta-Lactam resistance [PATH:ko01501]	23	4	0
		01502 Vancomycin resistance [PATH:ko01502]	6	1	0
		01503 Cationic antimicrobial peptide (CAMP) resistance [PATH:ko01503]	37	6	0
	Endocrine_and_metabolic_diseases	04930 Type II diabetes mellitus [PATH:ko04930]	2	1	0
		04932 Non-alcoholic fatty liver disease (NAFLD) [PATH:ko04932]	0	0	0
		04940 Type I diabetes mellitus [PATH:ko04940]	1	0	0
		04950 Maturity onset diabetes of the young [PATH:ko04950]	0	0	0
	Immune_diseases	05310 Asthma [PATH:ko05310]	0	0	0
		05320 Autoimmune thyroid disease [PATH:ko05320]	0	0	0
		05321 Inflammatiory bowel disease (IBD) [PATH:ko05321]	0	0	0
		05322 Systemic lupus erythematosus [PATH:ko05322]	0	0	0
		05323 Rheumatoid arthritis [PATH:ko05323]	0	0	0
		05330 Allograft rejection [PATH:ko05330]	0	0	0
		05332 Graft-versus-host disease [PATH:ko05332]	0	0	0
		05340 Primary immunodeficiency [PATH:ko05340]	2	2	0
	Infectious_diseases	05100 Bacterial invasion of epithelial cells [PATH:ko05100]	0	0	0
		05110 Vibrio cholerae infection [PATH:ko05110]	0	0	0
		05111 Vibrio cholerae pathogenic cycle [PATH:ko05111]	6	0	0
		05120 Epithelial cell signaling in Helicobacter pylori infection [PATH:ko05120]	2	6	1
		05130 Pathogenic Escherichia coli infection [PATH:ko05130]	0	0	0
		05131 Shigellosis [PATH:ko05131]	0	0	0
		05132 Salmonella infection [PATH:ko05132]	3	0	0
		05133 Pertussis [PATH:ko05133]	8	10	2
		05134 Legionellosis [PATH:ko05134]	4	0	1
		05140 Leishmaniasis [PATH:ko05140]	0	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		05142 Chagas disease (American trypanosomiasis) [PATH:ko05142]	1	0	0
		05143 African trypanosomiasis [PATH:ko05143]	1	0	0
		05144 Malaria [PATH:ko05144]	0	0	0
		05145 Toxoplasmosis [PATH:ko05145]	0	0	0
		05146 Amoebiasis [PATH:ko05146]	0	0	0
		05150 Staphylococcus aureus infection [PATH:ko05150]	0	0	0
		05152 Tuberculosis [PATH:ko05152]	4	0	0
		05160 Hepatitis C [PATH:ko05160]	0	0	0
		05161 Hepatitis B [PATH:ko05161]	0	0	0
		05162 Measles [PATH:ko05162]	0	0	0
		05164 Influenza A [PATH:ko05164]	0	0	0
		05166 HTLV-I infection [PATH:ko05166]	0	0	0
		05168 Herpes simplex infection [PATH:ko05168]	0	0	0
		05169 Epstein-Barr virus infection [PATH:ko05169]	0	0	0
	Neurodegenerative_diseases	05010 Alzheimer's disease [PATH:ko05010]	2	1	0
		05012 Parkinson's disease [PATH:ko05012]	0	0	0
		05014 Amyotrophic lateral sclerosis (ALS) [PATH:ko05014]	2	1	0
		05016 Huntington's disease [PATH:ko05016]	3	0	0
		05020 Prion diseases [PATH:ko05020]	1	0	0
	Substance_dependence	05030 Cocaine addiction [PATH:ko05030]	0	0	0
		05031 Amphetamine addiction [PATH:ko05031]	0	0	0
		05032 Morphine addiction [PATH:ko05032]	0	0	0
		05033 Nicotine addiction [PATH:ko05033]	0	0	0
		05034 Alcoholism [PATH:ko05034]	0	0	0
Metabolism	Amino_acid_metabolism	00250 Alanine, aspartate and glutamate metabolism [PATH:ko00250]	28	2	1
		00260 Glycine, serine and threonine metabolism [PATH:ko00260]	35	5	1
		00270 Cysteine and methionine metabolism [PATH:ko00270]	37	6	3
		00280 Valine, leucine and isoleucine degradation [PATH:ko00280]	15	6	0
		00290 Valine, leucine and isoleucine biosynthesis [PATH:ko00290]	20	1	1
		00300 Lysine biosynthesis [PATH:ko00300]	17	1	0
		00310 Lysine degradation [PATH:ko00310]	13	2	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		00330 Arginine and proline metabolism [PATH:ko00330]	43	8	2
		00340 Histidine metabolism [PATH:ko00340]	14	3	0
		00350 Tyrosine metabolism [PATH:ko00350]	15	2	0
		00360 Phenylalanine metabolism [PATH:ko00360]	23	4	2
		00380 Tryptophan metabolism [PATH:ko00380]	15	4	0
		00400 Phenylalanine, tyrosine and tryptophan biosynthesis [PATH:ko00400]	20	2	0
	Biosynthesis_of_other_secondary_metabolites	00231 Puromycin biosynthesis [PATH:ko00231]	0	0	0
		00232 Caffeine metabolism [PATH:ko00232]	1	0	0
		00254 Aflatoxin biosynthesis [PATH:ko00254]	0	0	0
		00261 Monobactam biosynthesis [PATH:ko00261]	10	0	0
		00311 Penicillin and cephalosporin biosynthesis [PATH:ko00311]	0	0	1
		00331 Clavulanic acid biosynthesis [PATH:ko00331]	0	0	0
		00332 Carbapenem biosynthesis [PATH:ko00332]	2	0	0
		00401 Novobiocin biosynthesis [PATH:ko00401]	4	0	0
		00402 Benzoxazinoid biosynthesis [PATH:ko00402]	0	0	0
		00403 Indole diterpene alkaloid biosynthesis [PATH:ko00403]	0	0	0
		00521 Streptomycin biosynthesis [PATH:ko00521]	7	3	0
		00524 Butirosin and neomycin biosynthesis [PATH:ko00524]	1	0	0
		00901 Indole alkaloid biosynthesis [PATH:ko00901]	0	0	0
		00940 Phenylpropanoid biosynthesis [PATH:ko00940]	2	0	0
		00941 Flavonoid biosynthesis [PATH:ko00941]	0	0	0
		00942 Anthocyanin biosynthesis [PATH:ko00942]	0	0	0
		00943 Isoflavonoid biosynthesis [PATH:ko00943]	0	0	0
		00944 Flavone and flavonol biosynthesis [PATH:ko00944]	0	0	0
		00945 Stilbenoid, diarylheptanoid and gingerol biosynthesis [PATH:ko00945]	1	0	0
		00950 Isoquinoline alkaloid biosynthesis [PATH:ko00950]	3	0	0
		00960 Tropane, piperidine and pyridine alkaloid biosynthesis [PATH:ko00960]	5	0	0
		00965 Betalain biosynthesis [PATH:ko00965]	0	0	0
		00966 Glucosinolate biosynthesis [PATH:ko00966]	0	0	0
		01058 Acridone alkaloid biosynthesis [PATH:ko01058]	0	0	0
	Carbohydrate_metabolism	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	40	11	1

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		00020 Citrate cycle (TCA cycle) [PATH:ko00020]	25	0	1
		00030 Pentose phosphate pathway [PATH:ko00030]	32	7	1
		00040 Pentose and glucuronate interconversions [PATH:ko00040]	14	8	0
		00051 Fructose and mannose metabolism [PATH:ko00051]	29	0	0
		00052 Galactose metabolism [PATH:ko00052]	17	3	0
		00053 Ascorbate and aldarate metabolism [PATH:ko00053]	6	10	0
		00500 Starch and sucrose metabolism [PATH:ko00500]	28	1	1
		00520 Amino sugar and nucleotide sugar metabolism [PATH:ko00520]	43	2	2
		00562 Inositol phosphate metabolism [PATH:ko00562]	8	2	0
		00620 Pyruvate metabolism [PATH:ko00620]	47	5	2
		00630 Glyoxylate and dicarboxylate metabolism [PATH:ko00630]	26	6	2
		00640 Propanoate metabolism [PATH:ko00640]	26	2	0
		00650 Butanoate metabolism [PATH:ko00650]	36	6	1
		00660 C5-Branched dibasic acid metabolism [PATH:ko00660]	14	0	1
	Energy metabolism	00190 Oxidative phosphorylation [PATH:ko00190]	42	2	0
		00195 Photosynthesis [PATH:ko00195]	7	2	0
		00196 Photosynthesis - antenna proteins [PATH:ko00196]	0	0	0
		00680 Methane metabolism [PATH:ko00680]	28	2	2
		00710 Carbon fixation in photosynthetic organisms [PATH:ko00710]	18	3	1
		00720 Carbon fixation pathways in prokaryotes [PATH:ko00720]	31	2	2
		00910 Nitrogen metabolism [PATH:ko00910]	16	2	0
		00920 Sulfur metabolism [PATH:ko00920]	32	6	0
	Glycan_biosynthesis_and_metabolism	00510 N-Glycan biosynthesis [PATH:ko00510]	0	1	0
		00511 Other glycan degradation [PATH:ko00511]	3	2	0
		00512 Mucin type O-glycan biosynthesis [PATH:ko00512]	0	0	0
		00513 Various types of N-glycan biosynthesis [PATH:ko00513]	0	0	0
		00514 Other types of O-glycan biosynthesis [PATH:ko00514]	0	0	0
		00531 Glycosaminoglycan degradation [PATH:ko00531]	4	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate [PATH:ko00532]	0	0	0
		00533 Glycosaminoglycan biosynthesis - keratan sulfate [PATH:ko00533]	0	0	0
		00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin [PATH:ko00534]	0	0	0
		00540 Lipopolysaccharide biosynthesis [PATH:ko00540]	22	3	1
		00550 Peptidoglycan biosynthesis [PATH:ko00550]	20	3	0
		00563 Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [PATH:ko00563]	1	0	0
		00601 Glycosphingolipid biosynthesis - lacto and neolacto series [PATH:ko00601]	0	1	0
		00603 Glycosphingolipid biosynthesis - globo series [PATH:ko00603]	2	0	0
		00604 Glycosphingolipid biosynthesis - ganglio series [PATH:ko00604]	2	0	0
	Lipid_metabolism	00061 Fatty acid biosynthesis [PATH:ko00061]	21	8	6
		00062 Fatty acid elongation [PATH:ko00062]	1	0	0
		00071 Fatty acid degradation [PATH:ko00071]	17	4	0
		00072 Synthesis and degradation of ketone bodies [PATH:ko00072]	4	3	0
		00073 Cutin, suberine and wax biosynthesis [PATH:ko00073]	0	0	0
		00100 Steroid biosynthesis [PATH:ko00100]	0	0	0
		00120 Primary bile acid biosynthesis [PATH:ko00120]	0	0	0
		00121 Secondary bile acid biosynthesis [PATH:ko00121]	0	0	0
		00140 Steroid hormone biosynthesis [PATH:ko00140]	0	0	1
		00561 Glycerolipid metabolism [PATH:ko00561]	15	4	0
		00564 Glycerophospholipid metabolism [PATH:ko00564]	28	4	2
		00565 Ether lipid metabolism [PATH:ko00565]	0	3	0
		00590 Arachidonic acid metabolism [PATH:ko00590]	2	1	0
		00591 Linoleic acid metabolism [PATH:ko00591]	0	1	0
		00592 alpha-Linolenic acid metabolism [PATH:ko00592]	2	1	0
		00600 Sphingolipid metabolism [PATH:ko00600]	1	0	0
		01040 Biosynthesis of unsaturated fatty acids [PATH:ko01040]	12	5	4
	Metabolism_of_cofactors_and_vitamins	00130 Ubiquinone and other terpenoid-quinone biosynthesis [PATH:ko00130]	19	2	1
		00670 One carbon pool by folate [PATH:ko00670]	14	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		00730 Thiamine metabolism [PATH:ko00730]	14	0	0
		00740 Riboflavin metabolism [PATH:ko00740]	8	0	0
		00750 Vitamin B6 metabolism [PATH:ko00750]	9	0	0
		00760 Nicotinate and nicotinamide metabolism [PATH:ko00760]	19	2	1
		00770 Pantothenate and CoA biosynthesis [PATH:ko00770]	21	1	1
		00780 Biotin metabolism [PATH:ko00780]	20	7	6
		00785 Lipoic acid metabolism [PATH:ko00785]	3	0	0
		00790 Folate biosynthesis [PATH:ko00790]	16	4	0
		00830 Retinol metabolism [PATH:ko00830]	1	1	0
		00860 Porphyrin and chlorophyll metabolism [PATH:ko00860]	18	2	0
	Metabolism_of_other_amino_acids	00410 beta-Alanine metabolism [PATH:ko00410]	14	2	0
		00430 Taurine and hypotaurine metabolism [PATH:ko00430]	6	1	0
		00440 Phosphonate and phosphinate metabolism [PATH:ko00440]	9	1	0
		00450 Selenocompound metabolism [PATH:ko00450]	18	1	2
		00460 Cyanoamino acid metabolism [PATH:ko00460]	7	1	0
		00471 D-Glutamine and D-glutamate metabolism [PATH:ko00471]	5	0	1
		00472 D-Arginine and D-ornithine metabolism [PATH:ko00472]	0	0	0
		00473 D-Alanine metabolism [PATH:ko00473]	4	0	0
		00480 Glutathione metabolism [PATH:ko00480]	20	2	0
	Metabolism_of_terpenoids_and_polyketides	00253 Tetracycline biosynthesis [PATH:ko00253]	3	1	0
		00281 Geraniol degradation [PATH:ko00281]	7	0	0
		00522 Biosynthesis of 12-, 14- and 16-membered macrolides [PATH:ko00522]	0	0	0
		00523 Polyketide sugar unit biosynthesis [PATH:ko00523]	2	3	0
		00900 Terpenoid backbone biosynthesis [PATH:ko00900]	11	1	0
		00902 Monoterpenoid biosynthesis [PATH:ko00902]	0	0	0
		00903 Limonene and pinene degradation [PATH:ko00903]	8	3	0
		00904 Diterpenoid biosynthesis [PATH:ko00904]	0	0	0
		00905 Brassinosteroid biosynthesis [PATH:ko00905]	0	0	0
		00906 Carotenoid biosynthesis [PATH:ko00906]	0	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		00908 Zeatin biosynthesis [PATH:ko00908]	0	1	0
		00909 Sesquiterpenoid and triterpenoid biosynthesis [PATH:ko00909]	0	0	0
		00981 Insect hormone biosynthesis [PATH:ko00981]	0	0	0
		01051 Biosynthesis of ansamycins [PATH:ko01051]	3	2	0
		01052 Type I polyketide structures [PATH:ko01052]	0	0	0
		01053 Biosynthesis of siderophore group nonribosomal peptides [PATH:ko01053]	8	6	2
		01054 Nonribosomal peptide structures [PATH:ko01054]	1	6	3
		01055 Biosynthesis of vancomycin group antibiotics [PATH:ko01055]	1	0	0
		01056 Biosynthesis of type II polyketide backbone [PATH:ko01056]	0	0	0
		01057 Biosynthesis of type II polyketide products [PATH:ko01057]	0	0	0
	Nucleotide_metabolism	00230 Purine metabolism [PATH:ko00230]	77	18	1
		00240 Pyrimidine metabolism [PATH:ko00240]	52	3	2
	Overview	01200 Carbon metabolism [PATH:ko01200]	101	13	3
		01210 2-Oxocarboxylic acid metabolism [PATH:ko01210]	30	0	1
		01212 Fatty acid metabolism [PATH:ko01212]	32	8	6
		01220 Degradation of aromatic compounds [PATH:ko01220]	9	4	2
		01230 Biosynthesis of amino acids [PATH:ko01230]	126	13	1
	Xenobiotics_biodegradation_and_metabolism	00351 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation [PATH:ko00351]	0	0	0
		00361 Chlorocyclohexane and chlorobenzene degradation [PATH:ko00361]	2	1	0
		00362 Benzoate degradation [PATH:ko00362]	9	3	1
		00363 Bisphenol degradation [PATH:ko00363]	1	0	0
		00364 Fluorobenzoate degradation [PATH:ko00364]	1	0	0
		00365 Furfural degradation [PATH:ko00365]	0	0	0
		00621 Dioxin degradation [PATH:ko00621]	0	0	1
		00622 Xylene degradation [PATH:ko00622]	1	1	1
		00623 Toluene degradation [PATH:ko00623]	1	0	0
		00624 Polycyclic aromatic hydrocarbon degradation [PATH:ko00624]	1	0	0
		00625 Chloroalkane and chloroalkene degradation [PATH:ko00625]	4	3	0
		00626 Naphthalene degradation [PATH:ko00626]	2	1	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		00627 Aminobenzoate degradation [PATH:ko00627]	6	1	0
		00633 Nitrotoluene degradation [PATH:ko00633]	5	1	0
		00642 Ethylbenzene degradation [PATH:ko00642]	2	0	0
		00643 Styrene degradation [PATH:ko00643]	2	1	0
		00791 Atrazine degradation [PATH:ko00791]	2	3	0
		00930 Caprolactam degradation [PATH:ko00930]	5	1	0
		00980 Metabolism of xenobiotics by cytochrome P450 [PATH:ko00980]	5	2	0
		00982 Drug metabolism - cytochrome P450 [PATH:ko00982]	5	2	0
		00983 Drug metabolism - other enzymes [PATH:ko00983]	10	0	0
		00984 Steroid degradation [PATH:ko00984]	0	1	0
Organismal_Systems	Circulatory_system	04260 Cardiac muscle contraction [PATH:ko04260]	0	0	0
		04261 Adrenergic signaling in cardiomyocytes [PATH:ko04261]	0	0	0
		04270 Vascular smooth muscle contraction [PATH:ko04270]	0	0	0
	Development	04320 Dorso-ventral axis formation [PATH:ko04320]	0	0	0
		04360 Axon guidance [PATH:ko04360]	0	0	0
		04380 Osteoclast differentiation [PATH:ko04380]	0	0	0
	Digestive_system	04970 Salivary secretion [PATH:ko04970]	0	0	0
		04971 Gastric acid secretion [PATH:ko04971]	0	0	0
		04972 Pancreatic secretion [PATH:ko04972]	0	0	0
		04973 Carbohydrate digestion and absorption [PATH:ko04973]	1	1	0
		04974 Protein digestion and absorption [PATH:ko04974]	0	0	0
		04975 Fat digestion and absorption [PATH:ko04975]	0	0	0
		04976 Bile secretion [PATH:ko04976]	0	0	0
		04977 Vitamin digestion and absorption [PATH:ko04977]	0	0	0
		04978 Mineral absorption [PATH:ko04978]	0	0	0
	Endocrine_system	03320 PPAR signaling pathway [PATH:ko03320]	3	3	0
		04614 Renin-angiotensin system [PATH:ko04614]	0	0	0
		04910 Insulin signaling pathway [PATH:ko04910]	3	0	0
		04911 Insulin secretion [PATH:ko04911]	0	1	0
		04912 GnRH signaling pathway [PATH:ko04912]	0	0	0
		04913 Ovarian Steroidogenesis [PATH:ko04913]	0	0	0
		04914 Progesterone-mediated oocyte maturation [PATH:ko04914]	1	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		04915 Estrogen signaling pathway [PATH:ko04915]	1	0	0
		04916 Melanogenesis [PATH:ko04916]	0	0	0
		04917 Prolactin signaling pathway [PATH:ko04917]	1	0	0
		04918 Thyroid hormone synthesis [PATH:ko04918]	3	0	0
		04919 Thyroid hormone signaling pathway [PATH:ko04919]	0	0	0
		04920 Adipocytokine signaling pathway [PATH:ko04920]	2	1	0
		04921 Oxytocin signaling pathway [PATH:ko04921]	0	0	0
		04922 Glucagon signaling pathway [PATH:ko04922]	5	1	0
	Environmental_adaptation	04626 Plant-pathogen interaction [PATH:ko04626]	5	2	0
		04710 Circadian rhythm [PATH:ko04710]	0	0	0
		04711 Circadian rhythm - fly [PATH:ko04711]	0	0	0
		04712 Circadian rhythm - plant [PATH:ko04712]	0	0	0
		04713 Circadian entrainment [PATH:ko04713]	0	0	0
	Excretory_system	04960 Aldosterone-regulated sodium reabsorption [PATH:ko04960]	0	0	0
		04961 Endocrine and other factor-regulated calcium reabsorption [PATH:ko04961]	0	0	0
		04962 Vasopressin-regulated water reabsorption [PATH:ko04962]	0	0	0
		04964 Proximal tubule bicarbonate reclamation [PATH:ko04964]	2	0	1
		04966 Collecting duct acid secretion [PATH:ko04966]	0	0	0
	Immune_system	04062 Chemokine signaling pathway [PATH:ko04062]	0	0	0
		04610 Complement and coagulation cascades [PATH:ko04610]	0	0	0
		04611 Platelet activation [PATH:ko04611]	0	0	0
		04612 Antigen processing and presentation [PATH:ko04612]	1	0	0
		04620 Toll-like receptor signaling pathway [PATH:ko04620]	0	0	0
		04621 NOD-like receptor signaling pathway [PATH:ko04621]	1	0	0
		04622 RIG-I-like receptor signaling pathway [PATH:ko04622]	0	0	0
		04623 Cytosolic DNA-sensing pathway [PATH:ko04623]	0	0	0
		04640 Hematopoietic cell lineage [PATH:ko04640]	0	0	0
		04650 Natural killer cell mediated cytotoxicity [PATH:ko04650]	0	0	0
		04660 T cell receptor signaling pathway [PATH:ko04660]	0	0	0

Major_KEGG_Category	KEGG Sub-Category	KEGG Pathway	Core	Accessory	Unique
		04662 B cell receptor signaling pathway [PATH:ko04662]	0	0	0
		04664 Fc epsilon RI signaling pathway [PATH:ko04664]	0	0	0
		04666 Fc gamma R-mediated phagocytosis [PATH:ko04666]	0	0	0
		04670 Leukocyte transendothelial migration [PATH:ko04670]	0	0	0
		04672 Intestinal immune network for IgA production [PATH:ko04672]	0	0	0
	Nervous_system	04720 Long-term potentiation [PATH:ko04720]	0	0	0
		04721 Synaptic vesicle cycle [PATH:ko04721]	0	0	0
		04722 Neurotrophin signaling pathway [PATH:ko04722]	0	0	0
		04723 Retrograde endocannabinoid signaling [PATH:ko04723]	1	0	0
		04724 Glutamatergic synapse [PATH:ko04724]	2	0	1
		04725 Cholinergic synapse [PATH:ko04725]	0	0	0
		04726 Serotonergic synapse [PATH:ko04726]	0	0	0
		04727 GABAergic synapse [PATH:ko04727]	2	0	1
		04728 Dopaminergic synapse [PATH:ko04728]	0	0	0
		04730 Long-term depression [PATH:ko04730]	0	0	0
	Sensory_system	04740 Olfactory transduction [PATH:ko04740]	0	0	0
		04742 Taste transduction [PATH:ko04742]	0	0	0
		04744 Phototransduction [PATH:ko04744]	0	0	0
		04745 Phototransduction - fly [PATH:ko04745]	0	0	0
		04750 Inflammatory mediator regulation of TRP channels [PATH:ko04750]	0	0	0

CHAPTER 11: Reference

Abebe-Akele F., Tisa L. S., Cooper V. S., Hatcher P. J., Abebe E., Thomas W. K. (2015). Genome sequence and comparative analysis of a putative entomopathogenic Serratia isolated from Caenorhabditis briggsae. BMC Genomics 16:531. 10.1186/s12864-015-1697-8.

Abi Khattar, Z., Lanois, A., Hadchity, L., Gaudriault, S. and Givaudan, A. (2019) Spatiotemporal expression of the putative MdtABC efflux pump of Phtotorhabdus luminescens occurs in a protease-dependent manner during insect infection. PLOS ONE, 14(2), p.e0212077.

Abrol, D. P. and Shankar, U. (2016) Pesticides, Food Safety and Integrated Pest Management, pp. 167–199, In: Pimental, D., Peshin, R., (eds.) Integrated Pest management- Pesticide problems, Springer Science+Business Media.

Alcoforado Diniz, J. and Coulthurst, S., (2015) Intraspecies Competition in Serratia marcescens Is Mediated by Type VI-Secreted Rhs Effectors and a Conserved Effector-Associated Accessory Protein. Journal of Bacteriology, 197(14), pp.2350-2360.

Allsopp, L., Bernal, P., Nolan, L. and Filloux, A., (2019) Causalities of war: The connection between type VI secretion system and microbiota. Cellular Microbiology, 22(3).

Antonenka, U., Nolting, C., Heesemann, J., and Rakin, A. (2005). Horizontal transfer of Yersinia high-pathogenicity island by the conjugative RP4 attB target-presenting shuttle plasmid. Mol. Microbiol. 57, 727–734. doi: 10.1111/j. 1365- 2958.2005.04722.x

Amaya, F., Blondel, C., Barros-Infante, M., Rivera, D., Moreno-Switt, A., Santiviago, C. and

Pezoa, D., (2022) Identification of Type VI Secretion Systems Effector Proteins That Contribute to Interbacterial Competition in Salmonella Dublin. Frontiers in Microbiology, 13.

Ambroset, C., Coluzzi, C., Guédon, G., Devignes, M., Loux, V., Lacroix, T., Payot, S. and Leblond-Bourget, N., (2016) New Insights into the Classification and Integration Specificity of Streptococcus Integrative Conjugative Elements through Extensive Genome Exploration. Frontiers in Microbiology, 6.

Akman, K., Yamashita, A., Watanabe, K., Oshima, K., Shiba, T., Hattori M., and Aksoy, S. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flie, Wigglesworthia glossinidia. Nat. Genet. 32: 402-407.

Albajes R, Gullino ML, van Lenteren JC, Elad Y (eds) (1999) Integrated pest and disease management in greenhouse crops. Kluwer Publishers, Dordrecht

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8:e1002514.

Altschul, S.F. Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. J Mol Biol. 215(3):403-10

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available

online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Argemi, X., Matelska, D., Ginalski, K., Riegel, P., Hansmann, Y., Bloom, J., Pestel-Caron, M., Dahyot, S., Lebeurre, J. and Prévost, G., (2018) Comparative genomic analysis of Staphylococcus lugdunensis shows a closed pan-genome and multiple barriers to horizontal gene transfer. BMC Genomics, 19(1).

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., Zagnitko O. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 9:75. doi: 10.1186/1471-2164-9-75.

Babai, R., Stern, B., Hacker, J. and Ron, E., (2000) New Fimbrial Gene Cluster of S-Fimbrial Adhesin Family. Infection and Immunity, 68(10), pp.5901-5907.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. doi: 10.1089/cmb.2012. 0021

Baumann, P., Baumann, L., Lai, C-Y., Rouhbakhsh, D., Moran, N.A., and Clark M.A. (1995) Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids. Annu. Revs. Microbiol. 49, 55–94.

Baumann, P. (2005) Biology of bacteriocyte-associated endo- 44. symbionts of plant sap-sucking insects. Annu. Rev. Microbiol.

59, 155–189.

Bayer-Santos, E., Ceseti, L., Farah, C. and Alvarez-Martinez, C., 2019. Distribution, Function and Regulation of Type 6 Secretion Systems of Xanthomonadales. Frontiers in Microbiology, 10.

Begg, G., Cook, S., Dye, R., Ferrante, M., Franck, P., Lavigne, C., Lövei, G., Mansion-Vaquie, A., Pell, J., Petit, S., Quesada, N., Ricci, B., Wratten, S. and Birch, A., (2017) A functional overview of conservation biological control.

Bellanger, X., Payot, S., Leblond-Bourget, N., and Guedon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol. Rev. 38, 720–760. doi: 10.1111/1574-6976. 12058

Belbahri, L., Chenari Bouket, A., Rekik, I., Alenezi, F., Vallat, A., Luptakova, L., Petrovova, E., Oszako, T., Cherrad, S., Vacher, S. and Rateb, M., (2017) Comparative Genomics of Bacillus amyloliquefaciens Strains Reveals a Core Genome with Traits for Habitat Adaptation and a Secondary Metabolites Rich Accessory Genome. Frontiers in Microbiology, 8.

Bergmann, C., Fliegauf, M., Brüchle, N., Frank, V., Olbrich, H., Kirschner, J., Schermer, B., Schmedding, I., Kispert, A., Kränzlin, B., Nürnberg, G., Becker, C., Grimm, T., Girschick, G., Lynch, S., Kelehan, P., Senderek, J., Neuhaus, T., Stallmach, T., Zentgraf, H., Nürnberg, P., Gretz, N., Lo, C., Lienkamp, S., Schäfer, T., Walz, G., Benzing, T., Zerres, K. and Omran, H., 2008. Loss of Nephrocystin-3 Function Can Cause Embryonic Lethality, Meckel-Gruber-like Syndrome, Situs Inversus, and Renal-Hepatic-Pancreatic Dysplasia. The American Journal of Human Genetics, 82(4), pp.959-970.

Bergemann, J., Kuhlcke, K., Fehse, B., Ratz, I., Ostertag, W., and Lother, H. (1995). Excision of specific DNA-sequences from integrated retroviral vectors via site- specific recombination. Nucleic Acids Res. 23, 4451–4456. doi: 10.1093/nar/23. 21.4451

Ben-Yakir, D. (1987) Growth retardation of Rhodnius prolixus symbionts by immunizing host against Nocardia (Rhodococcus) rhodnii. J. Insect Physiol. 33, 379–383

Bhadra, B., Roy, P., and Chakraborty, R. (2005). Serratia ureilytica sp. nov., a novel urea-utilizing species. Int J. Syst. Evol. Microbiol. 55, 2155–2158. doi: 10.1099/ijs.0.63674-0

Bhattacharya T, Newton ILG, Hardy RW. (2017) Wolbachia elevates host methyltransferase expression to block an RNA virus early during infection. PLoS Pathog. Jun 15;13(6):e1006427. doi: 10.1371/journal.ppat.1006427. PMID: 28617844; PMCID: PMC5472326.

Bingle, L. E., Bailey, C. M., & Pallen, M. J. (2008). Type VI secretion: A beginner's guide. Current Opinion in Microbiology, 11(1), 3–8. https:// doi.org/10.1016/j.mib.2008.01.006

Blondel CJ, Jimenez JC, Contreras I, Santiviago CA (2009) Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems differentially distributed in Salmonella serotypes. BMC Genomics, 10:1–17.

Bock, D., Medeiros, J. M., Tsao, H. F., Penz, T., Weiss, G. L., Aistleitner, K. (2017). In situ architecture, function, and evolution of a contractile injection system. Science 357, 713–717. doi: 10.1126/science.aan7904

Bouagga, S., Urbaneja, A., Rambla, J., Granell, A. and Pérez-Hedo, M., (2017) Orius laevigatus strengthens its role as a biological control agent by inducing plant defenses. Journal of Pest Science, 91(1), pp.55-64.

Bonte M and De Clercq P. (2008) Developmental and reproductive fitness of Orius laevigatus (Hemiptera: Anthocoridae) reared on factitious and artificial diets. J Econ Entomol;101(4):1127-33.

Bordenstein S.R., O'Hara F.P., and Werren J.H. (2001) Wolbachia-induced incompatibility

precedes other hybrid incompatibilities in Nasonia. Nature. 409(6821):707-10.

Bouvier, J., Groninger-Poe, F., Vetting, M., Almo, S. and Gerlt, J. (2014) Galactaro δ-Lactone Isomerase: Lactone Isomerization by a Member of the Amidohydrolase Superfamily. Biochemistry, 53(4), pp.614-616.

Boyer FG, Berthod FJ, Vandenbrouck Y, Attree I. (2009) Dissecting the bacterial type VI secretion system by a genome wide in silico analysis: what can be learned from available microbial genomic resources? BMC Genomics, 10:104.

Boyd, E. F., Almagro-Moreno, S. & Parent, M. A. (2009) Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.* **17**, 47–53.

Burke, G. and Moran, N., (2011) Massive Genomic Decay in Serratia symbiotica, a Recently Evolved Symbiont of Aphids. Genome Biology and Evolution, 3, pp.195-208.

Brackmann M, Nazarov S, Wang J, Basler M. (2017) Using Force to Punch Holes: Mechanics of Contractile Nanomachines. Trends Cell Biol;27(9):623–32.

Brandt JW, Chevignon G, Oliver KM, Strand MR. (2017) Culture of an aphid heritable symbiont demonstrates its direct role in defence against parasitoids. Proc Biol Sci. 284(1866).

Broms, J. E., Sjostedt, A., and Lavander, M. (2010). The role of the Francisella tularensis pathogenicity island in Type VI secretion, intracellular survival, and modulation of host cell signaling. Front. Microbiol. 1:136.

Brownlie JC., and Johnson KN. (2009) Symbiont-mediated protection in insect hosts. Trends

Microbiology. 17(8):348-54. doi: 10.1016/j.tim.2009.05.005.

Brune A (2010) Methanogens in the digestive tract of termites. (Endo)symbiotic Methanogenic Archaea (Microbiology Monographs) (Hackstein JHP, ed.), pp. 81–100. Springer, Berlin.

Brynildsrud O., Bohlin J., Scheffer L., Eldholm V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 17 238. 10.1186/s13059-016-1108-8

Cabezón, E., Ripoll-Rozada, J., Peña, A., de la Cruz, F. and Arechaga, I., (2014) Towards an integrated model of bacterial conjugation. FEMS Microbiology Reviews, p.n/a-n/a.

Castanié-Cornet, M., Cam, K., Bastiat, B., Cros, A., Bordes, P. and Gutierrez, C. (2010) Acid stress response in Escherichia coli: mechanism of regulation of gadA transcription by RcsB and GadE. Nucleic Acids Research, 38(11), pp.3546-3554.

Chandler D., Bailey A. S., Tatchell G. M., Davidson G., Greaves J., Grant W. P. (2011). The development, regulation and use of biopesticides for integrated pest management. Philos. Trans. R. Soc. Lond. B Biol. Sci. 366 1987–1998. 10.1098/rstb.2010.0390

Charlat S., Hornett E.A., Fullard J.H., Davies N., Roderick G.K., and Wedell N. (2007). Extraordinary Flux in Sex Ratio. Science (New York, N.Y.) 317: 214–214.

Chaudhari NM, Gupta VK, Dutta C. (2016) BPGA—an ultra-fast pan-genome analysis pipeline. Sci Rep.; 6:24373.

Chattoraj, P., Mohapatra, S., Rao, J. and Biswas, I., (2011) Regulation of Transcription by SMU.1349, a TetR Family Regulator, in Streptococcus mutans. Journal of Bacteriology, 193(23), pp.6605-6613.

Che, D., Hasan, M. S. & Chen, B. (2014) Identifying pathogenicity islands in bacterial pathogenomics using computational approaches. *Pathogens* **3**, 36–56.

Chen, M., Chen, N., Wang, J., Zhou, Y., Han, L., Shi, X., Hikichi, Y., Ohnishi, K., Li, J. and Zhang, Y., (2021) Involvement of a FAD-Linked Oxidase RSc0454 for Expression of the Type III Secretion System and Pathogenicity in Ralstonia solanacearum. Molecular Plant-Microbe Interactions®, 34(11), pp.1228-1235.

Chen X, Zhang J. (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol. 8: e1002784.

Cianfanelli, F., Monlezun, L. and Coulthurst, S., (2016) Aim, Load, Fire: The Type VI Secretion System, a Bacterial Nanoweapon. Trends in Microbiology, 24(1)51-61.

Clément, D., Leseigneur, C., Gelin, M., Coelho, D., Huteau, V., Lionne, C., Labesse, G., Dussurget, O. and Pochet, S., (2020) New Chemical Probe Targeting Bacterial NAD Kinase. Molecules, 25(21), p.4893.

Cochran D. G. (1985) Nitrogen Excretion in Cockroaches. Annual Review of Entomology. Vol. 30:29-49. https://doi.org/10.1146/annurev.en.30.010185.000333

Colman D.R., Toolson EC & Takacs-Vesbach CD (2012) Do diet and taxonomy influence insect gut bacterial communities?Mol Ecol 21: 5124–5137.

Costa, S., Guimarães, L., Silva, A., Soares, S. and Baraúna, R., (2020) First Steps in the Analysis of Prokaryotic Pan-Genomes. Bioinformatics and Biology Insights, 14, p.117793222093806.

Costopoulos K, Kovacs J.L., Kamins A., Gerardo N.M. (2014) Aphid facultative symbionts reduce survival of the predatory lady beetle Hippodamia convergens. BMC Ecol. 14:5. doi: 10.1186/1472-6785-14-5.

Cruden D.L., and Markovetz A.J. (1987) Microbial ecology of the cockroach gut. Annu Rev

Microbiol. 1987; 41:617-43.

Cryan, J.R. and Urban, J.M. (2012) Higher-level phylogeny of the insect order Hemiptera: is Auchenorrhyncha really paraphyletic? Syst. Entomol. 37, 7–21

Daffre S., Kylsten P., Samakovlis C., and Hultmark D. (1994) The lysozyme locus in Drosophila melanogaster: an expanded gene family adapted for expression in the digestive tract. Mol Gen Genet. 242(2):152-62.

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci, 23:324–328.

de Oliveira-Garcia, D., Dall'Agnol, M., Rosales, M., Azzuz, A., Alcántara, N., Martinez, M. and Girón, J. (2003) Fimbriae and adherence ofStenotrophomonas maltophiliato epithelial cells and to abiotic surfaces. Cellular Microbiology, 5(9), pp.625-636.

Dharam P. A., Uma S. (2016) Breeding Oilseed Crops for Sustainable Production, Chapter 20 – Integrated Pest Management, Pages 523–549, Academic Press

Dhillon B. K., Laird M. R., Shay J. A., Winsor G. L., Lo R., Nizam F. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 43 W104–W108. 10.1093/nar/gkv401

Dimroth, P., (2004) Molecular Basis for Bacterial Growth on Citrate or Malonate. EcoSal Plus, 1(1).

DiTomaso, J., Van Steenwyk, R., Nowierski, R., Meyerson, L., Doering, O., Lane, E., Cowan, P., Zimmerman, K., Pitcairn, M. and Dionigi, C. (2017) Addressing the needs for improving classical biological control programs in the USA. Biological Control, 106, pp.35-39.

Dobrindt U., Hochhut B., Hentschel U., and Hacker J. (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2(5):414-24.

Douglas, A.E. (1992) Requirement of pea aphids (Acyrthosiphon pisum) for their symbiotic

bacteria. Ent. Exp. Appl. 65, 195–198.

Douglas, A.E. (2014) Symbiosis as a general principle in eukaryotic evolution. Cold Spring Harb Perspect Biol. 6(2). pii: a016113. doi: 10.1101/cshperspect. a016113.

Douglas, A.E. (2016) How multi-partner endosymbioses function. Nat.

Rev. Microbiol. 14, 731–743

Egert M., Wagner B., Lemke T., Brune A. & Friedrich MW (2003) Microbial community structure in midgut and hindgut of the humus-feeding larva of Pachnoda ephippiata (Coleoptera: Scarabaeidae). Appl Environ Microbiol 69: 6659–6668.

Engel P., and Moran N. A. (2013). The gut microbiota of insects - diversity in structure and function. FEMS Microbiol. Rev. 37 699–735. 10.1111/1574-6976.12025

English, G., Trunk, K., Rao, V. A., Srikannathasan, V., Hunter, W. N., and Coulthurst, S. J. (2012). New secreted toxins and immunity proteins encoded within the Type VI secretion system gene cluster of *Serratia marcescens*. Mol. Microbiol. 86, 921–936. doi: 10.1111/mmi.12028

Ewald PW. (1987) Transmission modes and evolution of the parasitism-mutualism continuum. Ann N Y Acad Sci. 503:295-306.

Facey P.D., Méric G., Hitchings M.D., Pachebat J.A., Hegarty M.J., Chen X., Morgan L.V., Hoeppner J.E., Whitten M.M., Kirk W.D., Dyson P.J., Sheppard S.K., Del Sol R. (2015) Draft Genomes, Phylogenetic Reconstruction, and Comparative Genomics of Two Novel Cohabiting Bacterial Symbionts Isolated from Frankliniella occidentalis. Genome Biol Evol.7(8):2188-202. doi: 10.1093/gbe/evv136.

Fan, H., Su, B., Ma, C., Rowley, P. and Jayaram, M., (2020) A bipartite thermodynamic-kinetic contribution by an activating mutation to RDF-independent excision by a phage serine integrase. Nucleic Acids Research, 48(12), pp.6413-6430.

Farrugia, D. N., Elbourne, L. D., Mabbutt, B. C., and Paulsen, I. T. (2015). A novel family of integrases associated with prophages and genomic islands integrated within the tRNA-dihydrouridine synthase A (dusA) gene. Nucleic Acids Res. 43, 4547–4557. doi: 10.1093/nar/gkv337

Ferdous, M., Friedrich, A., Grundmann, H., de Boer, R., Croughs, P., Islam, M., Kluytmans-van den Bergh, M., Kooistra-Smid, A. and Rossen, J., (2016) Molecular characterization and phylogeny of Shiga toxin–producing Escherichia coli isolates obtained from two Dutch regions using whole genome sequencing. Clinical Microbiology and Infection 22 (2016) 642.e1e642.e9

Ferrari J. and Vavre F. (2011) Bacterial symbionts in insects or the story of communities affecting communities. Philos Trans R Soc Lond B Biol Sci. 366(1569):1389-400. doi: 10.1098/rstb.2010.0226.

Fiebig, A., Castro Rojas, C., Siegal-Gaskins, D. and Crosson, S. (2010) Interaction specificity, toxicity and regulation of a paralogous set of ParE/RelE-family toxin-antitoxin systems. Molecular Microbiology, 77(1), pp.236-251.

Fineran, P., Iglesias Cans, M., Ramsay, J., Wilf, N., Cossyleon, D., McNeil, M., Williamson, N., Monson, R., Becher, S., Stanton, J., Brügger, K., Brown, S. and Salmond, G., (2013) Draft Genome Sequence of Serratia sp. Strain ATCC 39006, a Model Bacterium for Analysis of the Biosynthesis and Regulation of Prodigiosin, a Carbapenem, and Gas Vesicles. Genome Announcements, 1(6).

Finn R. D., Clements J., Eddy S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39 W29–W37. 10.1093/nar/gkr367

Fischbein, D. and Corley, J. (2015) Classical biological control of an invasive forest pest: a world perspective of the management of Sirex noctiliousing the parasitoidIbalia leucospoides (Hymenoptera: Ibaliidae). Bulletin of Entomological Research, 105(1), pp.1-12.

Folmer O., Black M., Hoeh W., Lutz R., Vrijenhoek R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. 3 294–299.

Fukatsu, T. and Hosokawa, T. (2002) Capsule-transmitted gut symbiotic bacterium of the Japanese common plataspid stink- bug, Megacopta punctatissima. Appl. Environ. Microbiol. 68, 389–396

Gabaldón, T., Martin, T., Marcet-Houben, M., Durrens, P., Bolotin-Fukuhara, M., Lespinet, O., Arnaise, S., Boisnard, S., Aguileta, G., Atanasova, R., Bouchier, C., Couloux, A., Creno, S., Almeida Cruz, J., Devillers, H., Enache-Angoulvant, A., Guitard, J., Jaouen, L., Ma, L., Marck, C., Neuvéglise, C., Pelletier, E., Pinard, A., Poulain, J., Recoquillay, J., Westhof, E., Wincker, P., Dujon, B., Hennequin, C. and Fairhead, C. (2013) Comparative genomics of emerging pathogens in the Candida glabrata clade. BMC Genomics, 14(1), p.623.

Gao, C., Wang, Y., Zhang, Y., Lv, M., Dou, P., Xu, P., and Ma, C. (2015) NAD-independent L-lactate dehydrogenase required for L-lactate utili- zation in Pseudomonas stutzeri A1501. J. Bacteriol. 197:2239-2247.

Garbeva, P., van Elsas, J. and de Boer, W., (2012) Draft Genome Sequence of the Antagonistic Rhizosphere Bacterium Serratia plymuthica Strain PRI-2C. Journal of Bacteriology, 194(15), pp.4119-4120.

Garcia-Vallve S., Palau J. and Romeu A. (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in Escherichia coli and Bacillus subtilis. Molecular Biology and Evololution 16(9):1125-1134**.**

Gilmour, M., Thomson, N., Sanders, M., Parkhill, J. and Taylor, D., (2004) The complete nucleotide sequence of the resistance plasmid R478: defining the backbone components of incompatibility group H conjugative plasmids through comparative genomics. Plasmid, 52(3), pp.182-202.

Gomez-Polo P., Alomar O., Castane C., Riudavets J., Agusti N. (2013). Identification of Orius

spp. (Hemiptera: Anthocoridae) in vegetable crops using molecular techniques. Biol. Control 67 440–445. 10.1016/j.biocontrol.2013.09.017

Goodall, E., Robinson, A., Johnston, I., Jabbari, S., Turner, K., Cunningham, A., Lund, P., Cole, J. and Henderson, I. (2018) The Essential Genome of Escherichia coli K-12. mBio, 9(1).

Glaeser, S.P. and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology*, 38(4), pp.237–245.

Gregory, T., (2008) Understanding Evolutionary Trees. Evolution: Education and Outreach, 1(2), pp.121-137.

Grim, C., Kothary, M., Gopinath, G., Jarvis, K., Beaubrun, J., McClelland, M., Tall, B. and Franco, A., (2012) Identification and Characterization of Cronobacter Iron Acquisition Systems. Applied and Environmental Microbiology, 78(17), pp.6035-6050.

Gristwood T., Fineran P.C., Everson L., Salmond G.P. (2008) PigZ, a TetR/AcrR family repressor, modulates secondary metabolism via the expression of a putative four-component resistance-nodulation-cell-division efflux pump, ZrpADBC, in Serratia sp. ATCC 39006. Mol Microbiol. 69(2):418-35. doi: 10.1111/j.1365-2958.2008.06291. x.

Groussin, M., Mazel, F. and Alm, E., (2020) Co-evolution and Co-speciation of Host-Gut Bacteria Systems. Cell Host & Microbe, 28(1), pp.12-22.

Grzechowiak, M., Sekula, B., Jaskolski, M. and Ruszkowski, M., (2021) Serendipitous crystallization of E. coli HPII catalase, a sequel to “the tale usually not told”. Acta Biochimica Polonica.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment

tool for genome assemblies. Bioinformatics 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Halpern M., Shaked T., Pukall R., and Schumann P. (2009) Leucobacter chironomi sp. nov., a chromate resistant bacterium isolated from a chironomid egg mass. Int J Syst Evol Microbiol. 59:665–670. doi: 10.1099/ijs.0.004663-0.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B., Mori, H. and Horiuchi, T. (2006) Highly accurate genome sequences of Escherichia coliK-12 strains MG1655 and W3110. Molecular Systems Biology, 2(1).

Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc. Biol. Sci. 2003; 270:313–32

Hedges L.M., Brownlie J.C., O'Neill S.L., and Johnson K.N. (2008) Wolbachia and virus protection in insects. Science. 322(5902):702. doi: 10.1126/science.1162418.

Heikal, A., Nakatani, Y., Dunn, E., Weimar, M. R., Day, C. L., Baker, E. N., Lott, J. S., Sazanov, L. A., and Cook, G. M. (2014) Structure of the bacterial type II NADH dehydrogenase: A monotopic membrane protein with an essential role in energy generation. Mol. Microbiol. 91:950-964.

Henry LM, Maiden MC, Ferrari J, Godfray HC. (2015) Insect life history and the evolution of bacterial mutualism. Ecol Lett.18(6):516-25. doi: 10.1111/ele.12425.

Herre E.A., Knowlton N., Mueller U.G., and Rehner S.A. (1999) The evolution of mutualisms: exploring the paths between conflict and cooperation. Trends Ecol Evol. 14(2):49-53.

Hongoh Y. (2010) Diversity and genomes of uncultured microbial symbionts in the termite gut. Biosci Biotechnol Biochem. 74(6):1145-51.

Hosokawa, T., Kikuchi, Y., Nikoh, N., Shimada, M., and Fukatsu, T. (2006) Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria. PLoS Biol,

4:e337.

Hosokawa T., Koga R., Kikuchi Y., Meng X. Y., and Fukatsu T. (2010) Wolbachia as a bacteriocyte-associated nutritional mutualist. Proc Natl Acad Sci U S A. 107(2):769-74. doi: 10.1073/pnas.0911476107.

H. Mihara; N. Esaki (2002). Bacterial cysteine desulfurases: their function and mechanisms. , 60(1-2), 12–23.

Husnik F. (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153(7):1567–1578.

Iadanza, M., Higgins, A., Schiffrin, B., Calabrese, A., Brockwell, D., Ashcroft, A., Radford, S. and Ranson, N., (2016) Lateral opening in the intact β-barrel assembly machinery captured by cryo-EM. Nature Communications, 7(1).

Ilangovan, A., Kay, C., Roier, S., El Mkami, H., Salvadori, E., Zechner, E., Zanetti, G. and Waksman, G., 2017. Cryo-EM Structure of a Relaxase Reveals the Molecular Basis of DNA Unwinding during Bacterial Conjugation. Cell, 169(4), pp.708-721.e12.

Iyer, L., Burroughs, A. and Aravind, L., (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like β-grasp domains. Genome Biology, 7(7), p.R60.

James G., Schuller M., Sloots T. P., James G. S., Halliday C. L., Carter I. W. J. (2010). “Universal bacterial identification by PCR and DNA sequencing of 16S rRNA gene,” in PCR for Clinical Microbiology: An Australian and International Perspective (Dordrecht: Springer Netherlands); 209–214.

Jenkins, G. and Frohman, M., (2005) Phospholipase D: a lipid centric review. Cellular and Molecular Life Sciences, 62(19-20), pp.2305-2316.

Joy J.B. (2013) Symbiosis catalyses niche expansion and diversification. Proc Biol Sci. 280(1756):20122820. doi: 10.1098/rspb.2012.2820.

Jurkevitch, E. (2011) Riding the Trojan horse: combating pest insects with their own symbionts.

Microbial Biotechnology, 4(5), 620–627.

Kamat, S., Bagaria, A., Kumaran, D., Holmes-Hampton, G., Fan, H., Sali, A., Sauder, J., Burley, S., Lindahl, P., Swaminathan, S. and Raushel, F. (2011) Catalytic Mechanism and Three-Dimensional Structure of Adenine Deaminase. Biochemistry, 50(11), pp.1917-1927.

Karzai, A. and Sauer, R., 2000. Protein factors associated with the SsrA·SmpB tagging and ribosome rescue complex. *Proceedings of the National Academy of Sciences*, 98(6), pp.3040-3044.

Kampen H., Werner D. (2011) Human-biting potential of the predatory flower bug Orius majusculus (Hemiptera: Anthocoridae). Parasitol. Res. 108 1579–1581. 10.1007/s00436-010-2231-1

Kawai, M., Furuta, Y., Yahara, K., Tsuru, T., Oshima, K., Handa, N., Takahashi, N., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I. and Kobayashi, I., (2011) Evolution in an oncogenic bacterial species with extreme genome plasticity: Helicobacter pyloriEast Asian genomes. BMC Microbiology, 11(1).

Kennaway, C., Obarska-Kosinska, A., White, J., Tuszynska, I., Cooper, L., Bujnicki, J., Trinick, J. and Dryden, D. (2008) The structure of M. EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein. Nucleic Acids Research, 37(3), pp.762-770.

Kerepesi C., Bánky D., Grolmusz V. (2014) AmphoraNet: The webserver implementation of the AMPHORA2 metagenomic workflow suite, Gene. 533(2):538-40. doi: 10.1016/j.gene.2013.10.015.

Khanapur, M., Alvala, M., Prabhakar, M., Shiva Kumar, K., Edwin, R., Sri Saranya, P., Patel, R., Bulusu, G., Misra, P. and Pal, M., (2017) Mycobacterium tuberculosis chorismate mutase: A potential target for TB. Bioorganic & Medicinal Chemistry, 25(6), pp.1725-1736.

Kim, Y., Gu, C., Kim, H. and Lee, S. (2020) Current status of pan-genome analysis for pathogenic bacteria. Current Opinion in Biotechnology, 63, pp.54-62.

Kikuchi, Y. (2009) Endosymbiotic bacteria in insects: their diversity and culturability.

Microbes Environ, 24:195–204.

Kikuchi Y, Hosokawa T, and Fukatsu T. (2011) An ancient but promiscuous host-symbiont association between Burkholderia gut symbionts and their heteropteran hosts. ISME J. 5(3):446-60. doi: 10.1038/ismej.2010.150.

Kim Y.H., Kim J.H., Kim H.W., and Byun Y.W. (2008) Biological Characteristics of Two Natural Enemies of Thrips, Orius strigicollis (Poppius) and O. Laevigatus (Fieber) (Hemiptera: Anthocoridae). Korean Journal of Applied Entomology 47: 421–429.

Kim, Y., Gu, C., Kim, H. and Lee, S., (2020) Current status of pan-genome analysis for pathogenic bacteria. Current Opinion in Biotechnology, 63, pp.54-62.

Koonin, E., Makarova, K. and Wolf, Y. (2021) Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century. Trends in Microbiology, 29(7), pp.582-592.

Koskiniemi, S., Gibbons, H., Sandegren, L., Anwar, N., Ouellette, G., Broomall, S., Karavis, M., McGregor, P., Liem, A., Fochler, E., McNew, L., Rosenzweig, C., Rhen, M., Skowronski, E. and Andersson, D. (2013) Pathoadaptive Mutations in Salmonella enterica Isolated after Serial Passage in Mice. PLoS ONE, 8(7), p.e70147.

Korea, C., Badouraly, R., Prevost, M., Ghigo, J. and Beloin, C., (2010) Escherichia coli K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. Environmental Microbiology, 12(7), pp.1957-1977.

Lacey, L., Grzywacz, D., Shapiro-Ilan, D., Frutos, R., Brownbridge, M. and Goettel, M., (2015) Insect pathogens as biological control agents: Back to the future.

Lacks, S., Ayalew, S., de la Campa, A. and Greenberg, B., (2000) Regulation of competence for genetic transformation in Streptococcus pneumoniae: expression of dpnA, a late competence gene encoding a DNA methyltransferase of the DpnII restriction system. Molecular Microbiology, 35(5), pp.1089-1098.

Lamelas, A., Gosalbes, M.J., Manzano-Marín, A., Peretó, J., Moya, A., and Latorre, A. (2011)

Serratia symbiotica from the aphid Cinara cedri: a missing link from facultative to obligate insect endosymbiont. PLoS Genet. 7: e1002357. doi: 10.1371/journal.pgen.1002357. PMID:22102823.

Lamelas, A., Pérez-Brocal, V., Gómez-Valero, L., Gosalbes, M.J., Moya, A., and Latorre, A. (2008) Evolution of the secondary symbiont "Candidatus Serratia symbiotica" in aphid species of the subfamily lachninae. Appl. Environ. Mi- crobiol. 74: 4236–4240. doi:10.1128/AEM.00022-08. PMID:18502932.

Lancaster J.D., Mohammad B., and Abebe E. (2012) Effect of the bacterium Serratia marcescens SCBI on the longevity and reproduction of the nematode Caenorhabditis briggsae KT0001. BMC Res Notes. 5:688. doi: 10.1186/1756-0500-5-688.

Lattin J. D. (1999). Bionomics of the Anthocoridae. Annu. Rev. Entomol. 44 207–231. 10.1146/annurev.ento.44.1.207

Lee, C., Monson, R., Adams, R. and Salmond, G., (2017) The LacI–Family Transcription Factor, RbsR, Is a Pleiotropic Regulator of Motility, Virulence, Siderophore and Antibiotic Production, Gas Vesicle Morphogenesis and Flotation in Serratia. Frontiers in Microbiology, 8.

Lemke T Stingl U Egert M Friedrich MW & Brune A (2003) Physicochemical conditions and microbial activities in the highly alkaline gut of the humus-feeding larva of Pachnoda ephippiata (Coleoptera: Scarabaeidae). Appl Environ Microbiol 69: 6650–6658.

Lennings, J., West, T. and Schwarz, S., (2019) The Burkholderia Type VI Secretion System 5:

Composition, Regulation and Role in Virulence. Frontiers in Microbiology, 9.

Leroy P.D., Sabri A., Heuskin S., Thonart P., Lognay G., Verheggen F.J., Francis F., Brostaux Y., Felton G.W., Haubruge E. (2011) Microorganisms from aphid honeydew attract and enhance the efficacy of natural enemies. Nat Commun. 2:348. doi: 10.1038/ncomms1347.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242–W245. doi: 10.1093/nar/gkw290

Lin K, O'Brien KM, Trujillo C, Wang R, Wallach JB, Schnappinger D, Ehrt S. (2016) Mycobacterium tuberculosis Thioredoxin Reductase Is Essential for Thiol Redox Homeostasis but Plays a Minor Role in Antioxidant Defense. PLoS Pathog 12 (6): e1005675.

Li, J., Yao, Y., Xu, H., Hao, L., Deng, Z., Rajakumar, K. and Ou, H., 2015. SecReT6: a web-based resource for type VI secretion systems found in bacteria. Environmental Microbiology, 17(7), pp.2196-2202.

Liu, Y., Zhang, Z., Wang, F., Li, D. and Li, Y. (2020) Identification of type VI secretion system toxic effectors using adaptors as markers. Computational and Structural Biotechnology Journal, 18, pp.3723-3733.

Li, Y., Llewellyn, N., Giri, R., Huang, F. and Spencer, J. (2005) Biosynthesis of the Unique Amino Acid Side Chain of Butirosin: Possible Protective-Group Chemistry in an Acyl Carrier

Protein-Mediated Pathway. Chemistry & Biology, 12(6), pp.665-675.

Loenen, W., Dryden, D., Raleigh, E., Wilson, G. and Murray, N. (2013) Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Research, 42(1), pp.3-19.

Lorenzana, A., de Mendoza, A. H., Seco, M. V., and Casquero, P. A. (2010) Population development of Phorodon humuli and predators (Orius spp.) within hop cones: Influence of aphid density on hop quality. Crop Protection. 29:832–837.

Luan J.B., Chen W., Hasegawa D.K., Simmons A.M., Wintermantel W.M., Ling K.S., Fei Z., Liu S.S., Douglas A.E. (2015) Metabolic Coevolution in the Bacterial Symbiosis of Whiteflies and Related Plant Sap-Feeding Insects. Genome Biol Evol. 7(9):2635-47. doi: 10.1093/gbe/evv170.

Lu, J., Wong, J., Edwards, R., Manchak, J., Frost, L. and Glover, J. (2008) Structural basis of specific TraD-TraM recognition during F plasmid-mediated bacterial conjugation. Molecular Microbiology, 70(1), pp.89-99.

Maddison WP. (2000) Testing character correlation using pairwise comparisons on a phylogeny. J Theor Biol. 202(3):195–204.

Manzano-Marín, A. and Latorre, A. (2014) Settling Down: The Genome of Serratia symbiotica from the Aphid Cinara tujafilina Zooms in on the Process of Accommodation to a Cooperative Intracellular Life. Genome Biology and Evolution, 6(7), pp.1683-1698.

Manzano-Marín, A., and Latorre, A. (2016). Snapshots of a shrinking partner: genome reduction in Serratia symbiotica. Sci. Rep. 6:32590. doi: 10.1038/ srep32590

Martin, M. (2011). Cut adapt removes adapter sequences from high-throughput sequencing

reads. EMBnet J. 17, 10–12. doi: 10.14806/ej.17.1.200

Marques-Pereira, C., Proença, D. and Morais, P., (2020) Genome Sequences of Serratia Strains Revealed Common Genes in Both Serratomolides Gene Clusters. Biology, 9(12), p.482.

McCutcheon, J.P. and von Dohlen, C.D. (2011) An interdependent metabolic patchwork in the nested symbiosis of mealy- bugs. Curr. Biol. 21, 1366–1372

McCutcheon JP, McDonald B.R., and Moran N.A. (2009) Convergent evolution of metabolicroles in bacterial co-symbionts of insects.Proc Natl Acad Sci USA106(36):15394–15399.

McLennan, D., (2010) How to Read a Phylogenetic Tree. *Evolution: Education and Outreach*, 3(4), pp.506-519.

McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol, 2:1-5.

Medini D., Donati C., Tettelin H., Masignani V., Rappuoli R. (2005) The microbial pan-genome. Curr Opin Genet Dev. 15(6):589-94.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC

Bioinformatics 14:60.

Mercier, C., Chalansonnet, V., Orenga, S. and Gilbert, C. (2013) Characteristics of majorEscherichia colireductases involved in aerobic nitro and azo reduction. Journal of Applied Microbiology, 115(4), pp.1012-1022.

Merlino, C.P. (1924). "Bartolomeo Bizio's Letter to the most Eminent Priest, Angelo Bellani, Concerning the Phenomenon of the Red Colored Polenta". Journal of Bacteriology. 9: 527–543. PMC 379088 Freely accessible. PMID 16559067.

Midonet, C. and Barre, F., (2016) How Xer-exploiting mobile elements overcome cellular control. Proceedings of the National Academy of Sciences, 113(30), pp.8343-8345.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). “Creating the CIPRES Science Gateway for inference of large phylogenetic trees,” in Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA. doi: 10.1109/ GCE.2010.5676129

Miller M.M., Popova L.B., Meleshkevitch E.A., Tran P.V., and Boudko D.Y. (2008) The invertebrate B(0) system transporter, D. melanogaster NAT1, has unique d-amino acid affinity and mediates gut and brain functions. Insect Biochem Mol Biol. 38(10):923-31. doi: 10.1016/j.ibmb.2008.07.005.

Mir-Sanchis, I., Martinez-Rubio, R., Marti, M., Chen, J., Lasa, I., Novick, R. P. (2012) Control of Staphylococcus aureus pathogenicity island excision. Mol. Microbiol. 85, 833–845. doi:

10.1111/j.1365-2958.2012. 08145.x

Monferrer, D., Tralau, T., Kertesz, M., Dix, I., Solà, M. and Usón, I., (2010) Structural studies on the full-length LysR-typeregulator TsaR fromComamonas testosteroniT-2 reveal a novel open conformation of the tetrameric LTTR fold. Molecular Microbiology, 75(5), pp.1199-1214.

Montllor, C.B., Maxmen, A., and Purcell, A.H. (2002) Facultative bacterial endo- symbionts benefit pea aphids Acyrthosiphon pisum under heat stress. Ecol. Entomol. 27: 189–195. doi:10.1046/j.1365-2311.2002.00393. x.

Moon, B., Park, J., Robinson, D., Thomas, J., Park, Y., Thornton, J. and Seo, K., (2016) Mobilization of Genomic Islands of Staphylococcus aureus by Temperate Bacteriophage. PLOS ONE, 11(3), p.e0151409.

Moran, N.A., Munson, M.A., Baumann, P., and Ishikawa, H. (1993). A molecu- lar clock in endosymbiotic bacteria is calibrated using the insect hosts. Proc. R. Soc. London. Ser. B Biol. Sci. 253, 167–171.

Moriniere J, de Araujo BC, Lam AW, Hausmann A, Balke M, Schmidt S (2016) Species identification in Malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. PLoS ONE; 11:e0155497.

Murdoch, S. L., Trunk, K., English, G., Fritsch, M. J., Pourkarimi, E., and Coulthurst, S. J. (2011). The opportunistic pathogen Serratia marcescens utilizes type VI secretion to target

bacterial competitors. J. Bacteriol. 193, 6057–6069. doi: 10.1128/JB.05671-11

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, and Hattori M. (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314, 267

Navarro-Garcia, F., Ruiz-Perez, F., Cataldi, Á. and Larzábal, M., (2019) Type VI Secretion System in Pathogenic Escherichia coli: Structure, Role in Virulence, and Acquisition. Frontiers in Microbiology, 10.

Nishiya, Y. and Yamamoto, Y., (2007) Characterization of a NADH:Dichloroindophenol Oxidoreductase fromBacillus subtilis. Bioscience, Biotechnology, and Biochemistry, 71(2), pp.611-614.

Nikoh, N., McCutcheon, J., Kudo, T., Miyagishima, S., Moran, N. and Nakabachi, A., (2010) Bacterial Genes in the Aphid Genome: Absence of Functional Gene Transfer from Buchnera to Its Host. PLoS Genetics, 6(2), p.e1000827.

Nogge, G., and Gerresheim, A. (1982). Experiments on the elimination of symbionts from the Tsetse-Fly, Glossina- Morsitans-Morsitans (Diptera, Glossinidae), by antibiotics and lysozyme. J. Invertebr. Pathol. 40, 166–179. doi: 10.1016/0022-2011 (82)90112-4

Nordstedt, N. and Jones, M., (2021) Genomic Analysis of Serratia plymuthica MBSA-MJ1: A Plant Growth Promoting Rhizobacteria That Improves Water Stress Tolerance in Greenhouse

Ornamentals. Frontiers in Microbiology, 12.

Ogata, H., Renesto, P., Audic, S., Robert, C., Blanc, G., Fournier, P., Parinello, H., Claverie, J. and Raoult, D. (2005) The Genome Sequence of Rickettsia felis Identifies the First Putative Conjugative Plasmid in an Obligate Intracellular Parasite. PLoS Biology, 3(8), p.e248.

Ogier, J. C., Calteau, A., Forst, S., Goodrich-Blair, H., Roche, D., Rouy, Z. (2010). Units of plasticity in bacterial genomes: new insight from the comparative genomics of two bacteria interacting with invertebrates, photorhabdus and xenorhabdus. BMC Genomics 11:568. doi: 10.1186/1471- 2164- 11- 568

Ogura, Y., Abe, H., Katsura, K., Kurokawa, K., Asadulghani, M., Iguchi, A., Ooka, T., Nakayama, K., Yamashita, A., Hattori, M., Tobe, T. and Hayashi, T., (2008) Systematic Identification and Sequence Analysis of the Genomic Islands of the Enteropathogenic Escherichia coli Strain B171-8 by the Combined Use of Whole-Genome PCR Scanning and Fosmid Mapping. Journal of Bacteriology, 190(21), pp.6948-6960.

Oliver K.M., Moran N.A., and Hunter M.S. (2005) Variation in resistance to parasitism in aphids is due to symbionts not host genotype.Proc Natl Acad Sci U S A. 102(36):12795-800.

Oliver K.M., Russell J.A., Moran N.A., and Hunter M.S. (2003) Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. Proc Natl Acad Sci U S A. 100(4):1803-7.

Omattage, N., Deng, Z., Pinkner, J., Dodson, K., Almqvist, F., Yuan, P. and Hultgren, S., (2018) Structural basis for usher activation and intramolecular subunit transfer in P pilus biogenesis

in Escherichia coli. Nature Microbiology, 3(12), pp.1362-1368.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Paquola, A., Asif, H., Pereira, C., Feltes, B., Bonatto, D., Lima, W. and Menck, C., (2018) Horizontal Gene Transfer Building Prokaryote Genomes: Genes Related to Exchange Between Cell and Environment are Frequently Transferred. Journal of Molecular Evolution, 86(3-4), pp.190-203.

Parker, C. and Meyer, R., (2005) Mechanisms of Strand Replacement Synthesis for Plasmid DNA Transferred by Conjugation. Journal of Bacteriology, 187(10), pp.3400-3406.

Percudani, R. (2013). A microbial metagenome (Leucobacter sp.) in Caenorhabditis whole genome sequences. Bioinform. Biol. Insights 7, 55–72. doi: 10.4137/BBI. S11064

Péricart J. (1972). Hémiptères. Anthocoridae, Cimicidae et Microphsidae de l'ouest-paléarctique. Paris: Masson.

Petersen, C. and Møller, L. (2001) The RihA, RihB, and RihC Ribonucleoside Hydrolases of Escherichia coli. Journal of Biological Chemistry, 276(2), pp.884-894.

Petersen, L. M., and Tisa, L. S. (2013). Friend or foe? A review of the mechanisms that drive Serratia towards diverse lifestyles. Can. J. Microbiol. 59, 627–640. doi: 10.1139/cjm-2013-

0343

Petersen L. M. and Tisa L. S. (2014) Molecular Characterization of Protease Activity in Serratia sp. Strain SCBI and Its Importance in Cytotoxicity and Virulence. J Bacteriol. 196(22): 3923–3936. doi: 10.1128/JB.01908-14

Piña-Iturbe, A., Ulloa-Allendes, D., Pardo-Roa, C., Coronado-Arrázola, I., Salazar-Echegarai, F., Sclavi, B., González, P. and Bueno, S., (2018) Comparative and phylogenetic analysis of a novel family of Enterobacteriaceae-associated genomic islands that share a conserved excision/integration module. Scientific Reports, 8(1).

Pieńko, T. and Trylska, J. (2020) Extracellular loops of BtuB facilitate transport of vitamin B12 through the outer membrane of E. coli. PLOS Computational Biology, 16(7), p.e1008024.

Poehlein, A., Freese, H., Daniel, R. and Simeonova, D. (2014) Draft Genome Sequence of Serratia sp. Strain DD3, Isolated from the Guts of Daphnia magna. Genome Announcements, 2(5).

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–maximum-likelihood trees for large alignments. PLOS ONE 5: e9490. doi: 10.1371/journal.pone.0009490

Pukatzki S, Ma AT, Revel AT, Sturtevant D, Mekalanos JJ (2007) Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. Proc Natl

Acad Sci U S A, 104:15508–15513.

Quintieri, L., Fanelli, F., Zühlke, D., Caputo, L., Logrieco, A., Albrecht, D. and Riedel, K., (2020) Biofilm and Pathogenesis-Related Proteins in the Foodborne P. fluorescens ITEM 17298 With Distinctive Phenotypes During Cold Storage. Frontiers in Microbiology, 11.

Rajagopala, S., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S., Phanse, S., Ceol, A., Häuser, R., Siszler, G., Wuchty, S., Emili, A., Babu, M., Aloy, P., Pieper, R. and Uetz, P., (2014) The binary protein-protein interaction landscape of Escherichia coli. Nature Biotechnology, 32(3), pp.285-290.

Ramos J. L., Martínez-Bueno M., Molina-Henares A. J., Terán W.,1 Watanabe K, Zhang X., Gallegos M.T.,1 Brennan R., and Tobes R. (2005) The TetR Family of Transcriptional Repressors. Microbiol Mol Biol Rev. 69(2): 326–356. doi: 10.1128/MMBR.69.2.326-356.2005.

Rao, R., Sharma, S., Sivakumar, N. and Jayakumar, K., (2020) Genomic islands and the evolution of livestock-associated Staphylococcus aureus genomes. Bioscience Reports, 40(11).

Raupach, M. J., Hendrich, L., Küchler, S. M., Deister, F., Morinière, J., and Gossner, M. M. (2014). Building-up of a DNA barcode library for true bugs (insecta: hemiptera: heteroptera) of Germany reveals taxonomic uncertainties and surprises. PLOS ONE 9:e106940. doi: 10.1371/journal.pone.0106940

Ravcheev D. A., Khoroshkin M. S., Laikova O. N., Tsoy O. V., Sernova N. V., and Petrova S. A. (2014). Comparative genomics and evolution of regulons of the LacI-family transcription

factors. Front. Microbiol. 5:294 10.3389/fmicb.2014.00294

Read A.F., and Nee S. (1995) Inference from binary comparative data. J Theor Biol.173(1):99–108.

Rêgo, A., Johnson, J., Geibel, S., Enguita, F., Clegg, S. and Waksman, G. (2012) Crystal structure of the MrkD1Preceptor binding domain ofKlebsiella pneumoniaeand identification of the human collagen V binding interface. Molecular Microbiology, 86(4), pp.882-893.

Rodriguez-Valera, F., Martín-Cuadrado, A., Rodriguez-Brito, B., Pasic, L., Thingstad, T., Rohwer, F. and Mira, A., (2009) Explaining microbial population genomics through phage predation. Nature Precedings.

Romeis, J., Naranjo, S., Meissle, M. and Shelton, A., (2019). Genetically engineered crops help support conservation biological control. Biological Control, 130, pp.136-154.

Rosenberg E and Zilber-Rosenberg (2011) Symbiosis and development: the hologenome concept. Birth Defects Res C Embryo Today. 93(1):56-66. doi: 10.1002/bdrc.20196.

Rousset F., Bouchon D., Pintureau B., Juchault P., and Solignac M. (1992) Wolbachia endosymbionts responsible for various alterations of sexuality in arthropods. Proc Biol Sci. 250(1328):91-8.

Russell AB, Singh P, Brittnacher M, Bui NK, Hood RD, et al. (2012) A wide spread type VI secretion effector superfamily identified using a heuristic approach. Cell Host Microbe 11:

538–549.

Russell, A., Wexler, A., Harding, B., Whitney, J., Bohn, A., Goo, Y., Tran, B., Barry, N., Zheng, H., Peterson, S., Chou, S., Gonen, T., Goodlett, D., Goodman, A. and Mougous, J., (2014) A Type VI Secretion-Related Pathway in Bacteroidetes Mediates Interbacterial Antagonism. Cell Host & Microbe 16, 227–236.

Ryan, A., Kaplan, E., Nebel, J., Polycarpou, E., Crescente, V., Lowe, E., Preston, G. and Sim, E. (2014) Identification of NAD(P)H Quinone Oxidoreductase Activity in Azoreductases from P. aeruginosa: Azoreductases and NAD(P)H Quinone Oxidoreductases Belong to the Same FMN-Dependent Superfamily of Enzymes. PLoS ONE, 9(6), p.e98551.

Sabir, J., Rabah, S., Yacoub, H., Hajrah, N., Atef, A., Al-Matary, M., Edris, S., Alharbi, M., Ganash, M., Mahyoub, J., Al-Hindi, R., Al-Ghamdi, K., Hall, N., Bahieldin, A., Kamli, M. and Rather, I., 2019. Molecular evolution of cytochrome C oxidase-I protein of insects living in Saudi Arabia. *PLOS ONE*, 14(11), p.e0224336.

Sandstrom, J. and Moran, N. (1999) How nutritionally imbal- 35. anced is phloem sap for aphids? Entomol. Exp. Appl. 91, 203–210

Santangelo, M., Klepp, L., Nuñez-García, J., Blanco, F., Soria, M., García-Pelayo, M., Bianco, M., Cataldi, A., Golby, P., Jackson, M., Gordon, S. and Bigi, F. (2009) Mce3R, a TetR-type transcriptional repressor, controls the expression of a regulon involved in lipid metabolism in Mycobacterium tuberculosis. Microbiology, 155(7), pp.2245-2255.

Santos-Garcia D., Farnier PA, Beitia F, Zchori-Fein E, Vavre F, Mouton L, Moya A, Latorre A, and Silva FJ. (2012) Complete genome sequence of “Candidatus Portiera aleyrodidarum” BT-QVLC, an obligate symbiont that supplies amino acids and carotenoids to Bemisia tabaci.

J. Bacteriol. 194, 6654–6655

Sasaki, T., Hayashi, H., and Ishikawa, H. (1991) Growth and reproduction of the symbiotic and aposymbiotic pea aphids, Acyrthosiphon pisum maintained on artificial diets. J. Insect Physiol. 37, 749–756.

Sayed, S. M., Montaser, M. M., Elsayed, G., and Amer, S. A. M. (2013) Preliminary Molecular Identification of a Predatory Bug, *Orius albidipennis*, Collected from Ornamental Plants. J Insect Sci., 13: 11.

Schaefer, C.W. and Panizzi, A.R. (2000) Economic Importance of Heteroptera: A General View, CRC Press-Taylor & Francis Group.

Schubeler, D., Mielke, C., and Bode, J. (1997). Excision of an integrated provirus by the action of FLP recombinase. Vitro Cell Dev. Biol. Anim. 33, 825–830. doi: 10.1007/s11626-997-0163-6.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF. What we know and don't know about Earth's missing biodiversity. Trends Ecol. Evol. 2012; 27:501–510.

Schloss, PD & Handelsman, J. (2006) Introducing SONS, a tool for OTU-based comparisons of membership and structure between microbial communities. Applied and Environmental Microbiology. 72:6773-9.

Schneider, J., Yepes, A., Garcia-Betancur, J., Westedt, I., Mielich, B. and López, D., (2011) Streptomycin-Induced Expression in Bacillus subtilis of YtnP, a Lactonase-Homologous

Protein That Inhibits Development and Streptomycin Production in Streptomyces griseus. Applied and Environmental Microbiology, 78(2), pp.599-603.

Schubeler, D., Mielke, C., and Bode, J. (1997). Excision of an integrated provirus by the action of FLP recombinase. Vitro Cell Dev. Biol. Anim. 33, 825–830. doi: 10.1007/s11626-997-0163-6.

Speare, L., Woo, M., Bultman, K., Mandel, M., Wollenberg, M. and Septer, A., 2021. Host-Like Conditions Are Required for T6SS-Mediated Competition among Vibrio fischeri Light Organ Symbionts. mSphere, 6(4).

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. 4:2304. doi: 10.1038/ncomms3304

Semsey, S., Papp, I., Buzas, Z., Patthy, A., Orosz, L., and Papp, P. P. (1999). Identification of site-specific recombination genes int and xis of the Rhizobium temperate phage 16-3. J. Bacteriol. 181, 4185–4192.

Senderovich Y., and Halpern M. (2013) The protective role of endogenous bacterial communities in chironomid egg masses and larvae. ISME J. 7(11):2147-58. doi: 10.1038/ismej.2013.100.

Sethupathy, S., Sathiyamoorthi, E., Kim, Y., Lee, J. and Lee, J. (2020) Antibiofilm and Antivirulence Properties of Indoles Against Serratia marcescens. Frontiers in Microbiology, 11.

Shalom G, Shaw JG, Thomas MS (2007) In vivo expression technology identifies a type VI secretion system locus in Burkholderia pseudomallei that is induced upon invasion of macrophages. Microbiol, 153:2689–2699.

Sharma, A., Sandhi, R. and Reddy, G., (2019) A Review of Interactions between Insect Biological Control Agents and Semiochemicals. Insects, 10(12), p.439.

Shinohara, T., Ikawa, S., Iwasaki, W., Hiraki, T., Hikima, T., Mikawa, T., Arai, N., Kamiya, N. and Shibata, T. (2015) Loop L1 governs the DNA-binding specificity and order for RecA-catalyzed reactions in homologous recombination and DNA repair. Nucleic Acids Research, 43(2), pp.973-986.

Shneider MM, Buth SA, Ho BT, Basler M, Mekalanos JJ, Leiman PG. (2013) PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. Nature, 500:350–353.

Shyntum, D., Venter, S., Moleleki, L., Toth, I. and Coutinho, T., (2014) Comparative genomics of type VI secretion systems in strains of Pantoea ananatis from different environments. BMC Genomics, 15:163

Sibanda, T. and Ramganesh, S. (2021) Taxonomic and functional analyses reveal existence of virulence and antibiotic resistance genes in beach sand bacterial populations. Archives of Microbiology, 203(4), pp.1753-1766.

Silverman JM, Brunet YR, Cascales E, Mougous JD. (2012) Structure and Regulation of the Type VI Secretion System. Annu Rev Microbiol;66(1):453–72.

Singh, S., Ghosh, P., and Hatfull, G. F. (2013). Attachment site selection and identity in Bxb1 serine integrase-mediated site-specific recombination. PLoS Genet. 9:e1003490. doi: 10.1371/journal.pgen.1003490.

Steele, M., Motta, E., Gattu, T., Martinez, D. and Moran, N., (2021) The Gut Microbiota Protects Bees from Invasion by a Bacterial Pathogen. Microbiology Spectrum, 9(2).

Suarez G, Sierra JC, Erova TE, Sha J, Horneman AJ, Chopra AK. (2010) A type VI secretion system effector protein, VgrG1, from Aeromonas hydrophila that induces host cell toxicity by

ADP ribosylation of actin. J Bacteriol, 192:155–169.

Sudakaran S, Kost C, and Kaltenpoth M. (2017) Symbiont Acquisition and Replacement as a Source of Ecological Innovation. Trends Microbiol. 25(5):375-390. doi: 10.1016/j.tim.2017.02.014.

Sulavik MC, Houseweart C, Cramer C, Jiwani N, Murgolo N, Greene J, DiDomenico B, Shaw KJ, Miller GH, Hare R, Shimer G. (2001) Antibiotic susceptibility profiles of Escherichia coli strains lacking multidrug efflux pump genes. Antimicrob Agents Chemother. 45(4):1126-36.

Sullivan, J., Trzebiatowski, J., Cruickshank, R., Gouzy, J., Brown, S., Elliot, R., Fleetwood, D., McCallum, N., Rossbach, U., Stuart, G., Weaver, J., Webby, R., de Bruijn, F. and Ronson, C., (2002) Comparative Sequence Analysis of the Symbiosis Island of Mesorhizobium loti Strain R7A. Journal of Bacteriology, 184(11), pp.3086-3095.

Srikannathasan, V., English, G., Bui, N. K., Trunk, K., O'Rourke, P. E., Rao, V. A., et al. (2013). Structural basis for type VI secreted peptidoglycan DL-endopeptidase function, specificity and neutralization in Serratia marcescens. Acta Crystallogr. D Biol. Crystallogr. 69, 2468–2482. doi: 10.1107/S0907444913022725

Takeuchi, F., Watanabe, S., Baba, T., Yuzawa, H., Ito, T., Morimoto, Y., Kuroda, M., Cui, L., Takahashi, M., Ankai, A., Baba, S., Fukui, S., Lee, J. and Hiramatsu, K., (2005) Whole-Genome Sequencing of Staphylococcus haemolyticus Uncovers the Extreme Plasticity of Its Genome and the Evolution of Human-Colonizing Staphylococcal Species. Journal of Bacteriology, 187(21), pp.7292-7308.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44, 6614–6624. doi: 10.1093/nar/gkw569

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al.: Genome analysis of multiple pathogenic isolates of

Streptococcus agalactiae: implications for the microbial "pan- genome." (2005) Proc Natl Acad Sci U S A, 102:13950-13955.

Thomas, S., Evans, H., Cortat, G., Koutsidou, C., Day, M. and Ellison, C., (2021) Assessment of the microcyclic rust Puccinia lantanae as a classical biological control agent of the pantropical weed Lantana camara. Biological Control, 160, p.104688.

Thermo Scientific, 2010. T042-TECHNICAL BULLETIN NANODROP SPECTROPHOTOMETERS, 260/280 AND 260/230 RATIOS. Available at:http://www.nanodrop.com/Library/T042-NanoDrop-Spectrophotometers-Nucleic-Acid-Purity-Ratios.pdf [Accessed 25 August 2010]

Tsu, B. and Saier, M., 2015. The LysE Superfamily of Transport Proteins Involved in Cell Physiology and Pathogenesis. PLOS ONE, 10(10), p.e0137184.

Torres, J. and Bueno, A., (2018) Conservation biological control using selective insecticides – A valuable tool for IPM. Biological Control, 126, pp.53-64.

Ubeda, C., Barry, P., Penades, J. R., and Novick, R. P. (2007). A pathogenicity island replicon in *Staphylococcus aureus* replicates as an unstable plasmid. *Proc. Natl. Acad Sci. U.S.A.* 104, 14182–14188. doi: 10.1073/pnas.0705994104

Unterweger D, Kostiuk B, Otjengerdes R, Wilton A, Diaz-Satizabal L, Pukatzki S. (2015) Chimeric adaptor proteins translocate diverse type VI secretion system effectors in Vibrio

cholerae. EMBO J;34(16):2198–210.

Uranga, L., Reyes, E., Patidar, P., Redman, L. and Lusetti, S., (2017) The cohesin-like RecN protein stimulates RecA-mediated recombinational repair of DNA double-strand breaks. *Nature Communications*, 8(1).

Vacante V., Cocuzza G. E., De Clercq P., Van de Veire M., Tirry L. (1997). Development and survival of Orius albidipennis and O. laevigatus (Het.: Anthocoridae) on various diets. Entomophaga 42 493–498. 10.1007/BF02769809

van Lenteren J. (2012). The state of commercial augmentative biological control: plenty of natural enemies, but a frustrating lack of uptake. Biocontrol 57 1–20. 10.1007/s10526-011-9395-1

Vacante V., Cocuzza G. E., De Clercq P., Van de Veire M., Tirry L. (1997). Development and survival of Orius albidipennis and O. laevigatus (Het.: Anthocoridae) on various diets. Entomophaga 42 493–498. 10.1007/BF02769809

Veres, A., Petit, S., Conord, C. and Lavigne, C., (2013) Does landscape composition affect pest abundance and their control by natural enemies? A review. Agriculture, Ecosystems & Environment, 166, pp.110-117.

Vocadlo DJ, Withers S (2005) Detailed comparative analysis of the catalytic mechanisms of beta-N-acetylglucosaminidases from families 3 and 20 of glycoside hydrolases.

Biogeosciences, 44:12809–12818.

von Dohlen C.D., Kohler S., Alsop S.T., McManus W.R. (2001) Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. Nature. 412(6845):433-6.

Wan B, Zhang Q, Ni J, Li S, Wen D, Li J, et al. (2017) Type VI secretion system contributes to Enterohemorrhagic Escherichia coli virulence by secreting catalase against host reactive oxygen species (ROS). PLoS Pathog 13(3): e1006246.

Watanabe, M., Tagami, Y., Miura, K., Kageyama, D. and Stouthamer, R. (2012) Distribution patterns of Wolbachia endosymbionts in the closely related flower bugs of the genus Orius: implications for coevolution and horizontal transfer. Microb Ecol., 64(2):537-45.

Wen, Z., Wang, P., Sun, C., Guo, Y. and Wang, X., (2017) Interaction of Type IV Toxin/Antitoxin Systems in Cryptic Prophages of Escherichia coli K-12. Toxins, 9(3), p.77.

Wernegreen, J. J. (2002) Genome evolution in bacterial endosymbionts of insects. Nat. Rev. Genet., 3:850-860

Werren, J. H. (1997) Biology of Wolbachia. Annu. Rev. Entomol.,42:587–609.

Werren J. H., Baldo L., and Clark M.E. (2008) Wolbachia: master manipulators of invertebrate

biology. Nature Review, Microbiology, 6:741-751

Wheat, R.P., Zuckerman, A., and Rantz, L.A. 1951. Infection due to Chromobacte- ria: report of eleven cases. Arch. Intern. Med. 88(4): 461–466. doi:10.1001/archinte.1951.03810100045004.

Williams SG, Greenwood JA, Jones CW (1992) Molecular analysis of the lac operon encoding the binding-protein-dependent lactose transport system and β-galactosidase in Agrobacterium radiobacter. Mol Microbiol, 6:1755–1768.

Wilson A. C. C. and Duncan R. P. (2015) Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. PNAS, 112(33):10255–10261.

Wozniak, R. A., and Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat. Rev. Microbiol. 8, 552–563. doi: 10.1038/nrmicro2382

Yano E. (2004). Recent development of biological control and IPM in greenhouses in Japan. J. Asia Pacific Entomol. 7 5–11. 10.1016/S1226-8615(08)60195-8

Youn, B., Camacho, R., Moinuddin, S., Lee, C., Davin, L., Lewis, N. and Kang, C., (2006) Crystal structures and catalytic mechanism of the Arabidopsis cinnamyl alcohol dehydrogenases AtCAD5 and AtCAD4. Organic & Biomolecular Chemistry, 4(9), p.1687.

Young, A., Carette, X., Helmel, M., Steen, H., Husson, R., Quackenbush, J. and Platig, J., (2021) Multiomic regulatory networks capture downstream effects of kinase inhibition in Mycobacterium tuberculosis. npj Systems Biology and Applications, 7(1).

Yoneda, K., Yoshioka, M., Sakuraba, H., Araki, T. and Ohshima, T., (2020) Structural and biochemical characterization of an extremely thermostable FMN-dependent NADH-indigo reductase from Bacillus smithii. International Journal of Biological Macromolecules, 164, pp.3259-3267.

Yu, L., Li, W., Li, Q., Chen, X., Ni, J., Shang, F. and Xue, T., (2020) Role of LsrR in the regulation of antibiotic sensitivity in avian pathogenic Escherichia coli. Poultry Science, 99(7), pp.3675-3687.

Zerbino, D. R. and Birney, E. (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". Genome Research. 18 (5): 821–829.

Zhang, D., Gao, J., Li, M., Yuan, J., Liang, J., Yang, H. and Bu, W., 2019. The complete mitochondrial genome of Tetraphleps aterrimus (Hemiptera: Anthocoridae): Genomic comparisons and phylogenetic analysis of Cimicomorpha. *International Journal of Biological Macromolecules*, 130, pp.369-377.

Zhang, H., Gao, Z., Wei, Y., Xu, J. and Dong, Y., (2013) Insights into the Cross-Immunity Mechanism within Effector Families of Bacteria Type VI Secretion System from the Structure of StTae4-EcTai4 Complex. PLoS ONE, 8(9), p.e73782.

Zheng, J. and Keatinge-Clay, A., (2011) Structural and Functional Analysis of C2-Type Ketoreductases from Modular Polyketide Synthases. Journal of Molecular Biology, 410(1), pp.105-117.

Zhu D., Zhang P., Li P., Wu J., Xie C., Sun J., and Niu L. (2016) Description of Leucobacter holotrichiae sp. nov., isolated from the gut of Holotrichia oblita larvae. Int J Syst Evol Microbiol. 66(4):1857-61. doi: 10.1099/ijsem.0.000957.