Master’s thesis

Genetic methods for detection of plant-associated aerobic anoxygenic phototrophic bacteria

Martin Chilman

University of Jyväskylä
Department of Biological and Environmental Science
Nanoscience, Cell and Molecular Biology
01.07.2024

UNIVERSITY OF JYVÄSKYLÄ, Faculty of Mathematics and Science
Department of Biological and Environmental Science
Nanoscience, Cell and Molecular Biology

Martin Chilman:	Genetic methods for detection of plant-associated aerobic anoxygenic phototrophic bacteria
M.Sc. thesis:	66 p., 4 appendices (5 p.)
Supervisors:	Professor Janne Ihalainen (University of Jyväskylä) and University Lecturer Riitta Nissinen (University of Turku)
Reviewers:	Academy Research Fellow Nerea Abrego Antia (University of Jyväskylä) and Doctoral Researcher Benjami Laine (University of Jyväskylä)

May 2024

Keywords: Next generation sequencing, massive parallel sequencing, phyllosphere, endophyte, epiphyte, phototrophic

Aerobic anoxygenic phototrophic bacteria (AAPB) are a polyphyletic class of heterotrophic prokaryotes capable of anoxygenic photosynthesis in oxygenated environments. Since their discovery, they have been found in a great diversity of terrestrial niches and are considered ubiquitous in aerated bodies of water. The few metagenomic and culture-based studies that have investigated their presence and role in plants show that AAPB are also common in the phyllosphere. This study aims to augment research of plant-associated AAPB through the development of amplicon-based genetic methods. The specificity and sensitivity of frequently used AAPB-primers were evaluated using both a screen of AAPB isolated from plants, and DNA from the surface and tissue of lingonberry (*Vaccinium vitis-idaea*) leaves. pufM_uniF/pufM_uniR were shown to be the most suitable primers. Multiple variations in their PCR protocol were investigated to test if their efficacy with environmental samples could be optimized. Extensibility of these methods to community-analyses of plant-associated AAPB was evaluated with next-generation sequencing (IonTorrent PGM). In concert, the eubacterial population was analysed through *16S* ribosomal gene fragments. Results from the sequence data indicate that *Methylobacteria* are the most dominant AAPB genus found in the samples, and that epiphytic AAPB present in greater abundance and species diversity than endophytic AAPB. Only a single species of the genus *Sphingomonas* was found in the AAPB sequences, despite the primers being proven to be particularly sensitive to them and it being the most abundant genus in the eubacterial sequence data. Our results show that the techniques developed offer a good framework for future studies. More specifically, the sequence data show that primers pufM_uniF and pufM_uniR have good sensitivity and species coverage for plant-associated AAPB, and that they are well suited for analysis of epiphytic AAPB. The low number of detected endophytes may be due to their natural low abundance, and/or the inherent difficulty amplicon-based methods have with tissue extracts. Future work should validate the full range of

AAPB taxa that can be detected and the potential effects that low abundance may have on results.

JYVÄSKYLÄN YLIOPISTO, Matemaattis-luonnontieteellinen tiedekunta
Bio- ja ympäristötieteiden laitos
Nanotiede, Solu- ja molekyylibiologia

Martin Chilman:	Kasviassosioituneiden aerobisten happeatuottamattomien yhteyttävien bakteerien tunnistaminen geneettisillä menetelmillä
Pro gradu -tutkielma:	66 s., 4 liitettä (5 s.)
Työn ohjaajat:	Professori, Janne Ihalainen (Jyväskylän yliopisto), ja Yliopistolehtori, Riitta Nissinen (Turun yliopisto)
Tarkastajat:	Akatemiatutkija, Nerea Abrego Antia (Jyväskylän yliopisto), ja Väitöskirjatutkija, Benjami Laine (Jyväskylän yliopisto)

Toukokuu 2024

Hakusanat: Rinnakkaissekvensointia, fyllosfääri, endofyytti, epifyytti, fototrofi, kasviassosioituneet bakteerit

Aerobiset happeatuottamattomat yhteyttävät bakteerit (*eng.* aerobic anoxygenic phototrophic bacteria – AAPB) ovat polyfyleettisia ja toisenvaraisia esitumallisia, jotka kykenevät yhteyttämään happipitoisissa elinympäristöissä. AAPB löytyvät monenlaisista terrestrisista ympäristöistä ja niitä löydetään myös kaikissa happipitoisissa vesistöissä. Metagenomiset ja bakteeriviljelyyn perustavat tutkimukset ovat osoittaneet, että niitä löytyvät myös kasvien päältä ja sisältä. Tämä tutkielma pyrkii kehittämään menetelmän amplikonien avulla AAPB:n yhteisön kasvin kartoitukseen. Usein käytettyjen alukkeiden tarkuutta ja herkkyyttään tarkasteltiin kayttäen kahdeksaa AAPB kantaa ja DNA:sta sekä puolukan (*Vaccinium vitis-idaea*) lehtien pinnoilta että kudoksista. Alukkeet pufM_uniF ja pufM_uniR osoitettiin olevan sopivimpia tähän tehtävään, ja niiden tehokkuutta optimoitiin ympäristöperäisten näytteiden kanssa. Kehitettyä menetelmää verrattiin rinnakkaissekvensointimenetelmään. Samalla analysoittiin koko bakteeriyhteisörakennetta ja havaittiin, että *Methylobacteria* ovat yleisin tutkittujen kasvien AAPB-suku. Epifyyttiset AAPB:t esiintyvät endofyyttisiä runsaammin sekä monipuolisemmin. Ainoastaan yksi AAPB laji *Sphingomonas*-suvusta löytyi ja vain yhdeltä näytteeltä, vaikka tämä suku todistettiin olevaksi yleisin bakteeri-sekvenssidatassa. Lisäksi, käytetyt alukkeet olivat erityisen herkkiä tätä sukua kohta. Esitetty menetelmä toimii hyvänä pohjana pufM_uniF/pufM_uniR perustuviin AAPB tunnistamiseen kasveista. Ne ovat erityisen sopivia epifyyttejä analysoinnissa. Vähäinen AAPB määrä endosfääristä saattaa johtua luonnollisen vähäisestä määrästä tai toisaalta amplikoni-menetelmän haasteista kasvikudosnäytteissä. Tulevaisuudessa endofyyttisten näytteiden analyysimentelmää tulee kehittää, jolloin AAPB:n esiintyvyys kasvien sisällä voidaan kartoittaa tarkemmin.

TABLE OF CONTENTS

TERMS AND ABBREVIATIONS

VOCABULARY

Phototrophic (*Adj.*) a physiology determined by the capacity to perform photosynthesis (*n.*–phototroph)

Phyllosphere The above-ground/aerial portion of plants (such as stems and leaves)

Endophyte Micro-organisms residing within plant tissue.

Epiphyte Micro-organisms residing upon the surface of a plant.

Phyllophyte Microbes of the phyllosphere.

ABBREVIATIONS

AP	Anoxygenic photosynthesis
AAP	Aerobic anoxygenic photosynthesis
AAP$^+$	Aerobic anoxygenic phototrophic
AAPB	Aerobic anoxygenic phototrophic bacteria
AnAPB	Anaerobic anoxygenic phototrophic bacteria
AGE	Agarose gel electrophoresis
BChl a	Bacteriochlorophyll a
HGT	Horizontal gene transfer
NGS	Next generation sequencing
RC	Reaction Center
OTU	Operational taxonomic unit
VOC	Volatile organic compounds

1 INTRODUCTION

Photosynthesis—the conversion of solar radiation into chemical energy—is a manifold process found in all three kingdoms of life. Organisms capable of photosynthesis are classified physiologically as phototrophic. In the most abundant form of photosynthesis utilised by primary producers—algaea, cyanobacteria, and plants—light-induced oxidation (photolysis) of two water molecules ultimately produces diatomic oxygen and four protons. While the latter is used along with captured electrons in the synthesis of ATP and NADPH, the former is a waste product exuded to the atmosphere; photosynthesis which uses water as a reductant is thus an oxygenic process. In the Siderian period approximately 2.5 billion years ago, this oxygenic photosynthesis evolved in the ancestors of cyanobacteria, resulting in the rapid oxygenation of the Earth's atmosphere. Phototrophs, however, existed up to one billion years prior to this event, using various other reductants, and were therefore *anoxygenic*.

These ancient, anoxygenic phototrophs remain extant in diverse forms in the prokaryote and archaean kingdoms. Reflecting their evolutionary roots prior to the Great Oxidation Event, organisms capable of anoxygenic photosynthesis (AP) had long been observed to be exclusively anaerobic. However, in 1979, a strain of anoxygenic phototrophic prokaryote was discovered associated with green seaweed on the coast of Japan which, unusually, was capable of synthesising and utilising bacteriochlorophyll (BChl) *a* in the presence of oxygen (Shiba *et al.* 1979). These aerobic anoxygenic phototrophic bacteria (AAPB) were ostensibly considered at the time to be an oddity, yet as research has increasingly focused on AAPB, it appears that their phylogenetic diversity exceeds that of anaerobic anoxygenic phototrophs (AnAPB) (Imhoff *et al.* 2018). AAPB evolved from AnAPB at some time after the Great Oxidation Event, and have ancient phylogenetic origins (Imhoff *et al.* 2019). They appear not only to be ubiquitous in the aerated plane of water environments, but form a noteable portion of the marine microbiome (Koblížek 2015) and also are found in great diversity and abundance of fresh-water environments (Ferrara *et al.* 2017). Given their abundance and extent, they may even be a keystone in the overall carbon turnover of ocean- and sea-waters (Koblížek 2015). Recently, AAPB have also been detected in terrestrial environments such as soils (Csotonyi *et al.* 2010; Tang *et al.* 2018; Yang and Hu 2022).

Despite the phyllosphere—*viz.* the above-ground portion of plants—being in terms of surface area alone the largest potential terrestrial substrate for microbial colonisation (Vorholt 2012), research into AAPB's presence and role here has been relatively scarce. Yet the few studies performed in the past decade using metagenomic (Atamna-Ismaeel *et al.* 2012b; Florez-Núñez *et al.* 2020), culture-based (Nissinen *et al.* 2023; Zervas *et al.* 2019), and spectroscopic (Nissinen *et al.* 2023; Steifel *et al.* 2013) methods to detect AAPB in the phyllosphere all indicate that plant-associated AAPB are common, and may be as abundant as in aquatic environments (Atamna-Ismaeel *et al.* 2012b). In particular, the latest and most comprehensive study investigating 23 host-plant species across a large latitudal range indicates that they may indeed also be *ubiquitous* in this environmental niche (Nissinen *et al.* 2023). Considering both the noted importance of aquatic AAPB and of plant-microbe interactions and their impact on our atmosphere (reviewed in *i.a.* Bringel and Couée 2015; Liu *et al.* 2020; Vorholt 2012) further investigation into plant-associated AAPB is certainly warrented.

1.1 Aerobic anoxygenic phototrophs

One of the characteristics of AAPB which differentiate them from AnAPB is that they may synthesise their photosynthetic pigment, BChl *a*, in both aerated environments and the presence of light (Yurkov and Hughes 2013). Doing so results in accumulation of singlet oxygen ($^1O_2$) and concomitant cellular oxidative stress (Yurkov and Csotonyi 2009). While AnAPB avoid this by *i.a.* exclusively synthesising BChl in anaerobic conditions, how AAPB have adapted to these consequences is not fully understood. AAPB have been shown to differentially translate reaction centers (RC) and BChl *a* in light and dark conditions (Fecskeová *et al.* 2019; Koblížek *et al.*

2003; Selyanin *et al.* 2016). Therefore, mitigation may arise from BChl *a* biosynthesis (Selyanin *et al.* 2016) and other photosynthetic apparatus expression (Fecskeová *et al.* 2019) being performed preferentially in dark conditions, whereas remediation may arise from AAPB's high abundance of carotenoids; their carotenoids are typically evenly distributed throughout their cells (Yurkov and Csotonyi 2009) suggesting a general role for quenching radical oxygen species and harmful wavelengths of light (Hamilton 2019). They also act as accessory pigments to the light-harvesting (LH) complex by donating excitation energy to BChl *a* (Yurkov and Hughes 2013). AAPB produce a notable amount and variety of carotenoids, affording them their intense colorations; accordingly genes in the *crt* operon are found in great diversity (Zheng *et al.* 2011).

The second and most salient characteristic of AAPB is that their photosynthesis is *obligately* aerobic. Anoxygenic phototrophs utilise pheophytin-quinone in a Q-type (*a.k.a.* type-II-like) RC and AAPB use type-1 LH (LH1) complexes with the potential for simultaneous (and possibly modulatory) expression of LH2 (Yurkov and Hughes 2013) The primary electron acceptor (Q_A) species of some AAPB RC have been shown to have a positive redox midpoint potential in comparison to the much lower potential of AnAPB's Q_A (Rathgeber *et al.* 2012; Rathgeber *et al.* 2004). This implies that in the absence of oxygen, key components in the AAPB electron-transport apparatus become electronically saturated (dihydroquinol) thus stalling AAPB's cyclic electron transport.

The genetic components of AAP are structured within large, operonic (groups of adjacent genes under the control of a single promoter) scaffolds of variable content and orientation, often called the photosynthetic gene cluster (PGC) (Bauer *et al.* 1991; Igarashi *et al.* 2001; Waidner and Kirchman 2005; Young *et al.* 1989; Yutin *et al.* 2007). This roughly 45 kilobase contiguous "superoperon" contains some 40 (Waidner and Kirchman 2005) or 27 (Zheng *et al.* 2011) genes. It consists of five different operons; *bch* (BChl *a* biosynthesis enzymes), *puf* and *puh* (RC and LH components), *crt* (carotenoid biosynthesis enzymes), and regulatory components (Yurkov and Hughes 2013; Zheng *et al.* 2011). These operons are not always clearly or structurally deliminated, nor functionally independent (*ibid.*) hence the terms "superoperon" or PGC are salient in this instance. Evidence suggests that operons within the superoperon are co-expressed (Liotenberg *et al.* 2008).

AAPB are diverse and heterogenic in many terms—metabolically, taxonomically, phenotypically—so their one common characteristic being determined by operonic structures hints at a tantalising potential for (mass) horizontal gene transfer in AAPB speciation (Béjà *et al.* 2002; Igarashi *et al.* 2001; Nagashima *et al.* 1997; Yurkov and Beatty 1998). There exist subsets of bacterial species which both are, and are not, AAP, such as members of the core phyllosphere genus *Methylobacteria* (Steifel *et al.* 2013). In other words, AAP can be an interspecific trait. However, the evolution of AAPB through these manners remains contentious (Imhoff *et al.* 2018). Proteorhodopsin phototrophy (Atamna-Ismaeel *et al.* 2012a), for example, is readily transferred horizontally, though requires only two genes in contrast to the 27 to 45 of the PGC. Some evidence suggest that the PGC, though under low selective pressure (Zervas *et al.* 2019), is instead a very stable component of the AAPB genome, whose structure is a critical part of their adaptation (Zheng *et al.* 2011). Therefore, the probability of a "meaningful" instance of HGT with AAP is low, and accordingly AAPB speciation is more likely to happen via other manners.

Perhaps the most intensely studied component of the PGC is the *puf* operon: a core component of AP, critically encoding RC apoproteins. In APB, the general structure of the *puf* operon includes a BChl *a*-biosynthesis regulation component, *pufQ* (Bauer and Marrs 1988; Bauer *et al.* 1988) LH1 genes (*pufB* and *pufA*; β and α subunits) two of the three RC genes (*pufL* and *pufM*; light and medium subunits), and optionally *pufC* and *pufX*, which respectively encode cyctochrome *c* subunits and a structural protein which breaks symmetry of the LH1 (Koblížek *et al.* 2014). Transcription of these genes is a highly regulated process coordinated by the upstream BChl *a* biosynthesis operon, *bchCA*, the action of which produces a long, polycistronic mRNA transcript spanning several operons which can then be differentially modified (Bauer *et al.* 1991; Liotenberg *et al.* 2008). This operonic structure—coordinating synthesis of BChl *a* with other photosynthetic apparatus structures—ensures that neither can inordinately accumulate then cause oxidative damage (Koblížek *et al.* 2005), and that AAPB can respond quickly to changes in nutrients and oxygen tension (Liotenberg *et al.* 2008).

The structure of the *puf* operon itself in AAPB is variable (Béjà *et al.* 2002; Waidner and Kirchman 2005; Yutin *et al.* 2007; Zervas *et al.* 2019), allowing for classification of AAPB into 12 different phylotypes (Yutin *et al.* 2007). Ostensibly, various (supra)operonal structures facilitate different post-transcriptional modifications and differential translation of components of the photosynthetic apparatus, permitting for adaptation to the present environmental conditions that AAPB find themselves in, such as changes in oxygen tension (Liotenberg *et al.* 2008) and the aforementioned varied BChl *a*/RC ratios in light/dark conditions. This is particularly salient in regards to plant-associated AAPB—abiotic changes in the environment may come fast, unpredictably, and oftimes rarely; adaptation to this extreme environment requires rapid acclimation to these variables.

Nevertheless, photosynthesis ultimately accounts for only a fraction of AAPB energy acquisition (Yurkov and Csotonyi 2009) with the remainder derived from heterotrophic activity. Within the group of AAPB, the three clades (alpha-, beta-, and gamma-proteobacteria) may have preferential environments or niche association (Waidner and Kirchman 2008), suggesting a phylogenic/ecological specialisation. The abundance of AAPB in oligotrophic and/or extreme environments indicate that AAP may be an additional determinant of fitness in already ecologically robust taxa; that is to say, it appears likely that AAP is a trait in bacteria already specialised in metabolization of recalcitrant organic sources; photosynthesis ameliorates their metabolic needs, permitting extension into even more abiotically inhospitable niches and—on a community level—potentially outcompeting neighbouring heterotrophs. However, this idea—though reasonable and probable—remains hypothetical (see Cottrell *et al.* 2010). Nevertheless, evidence suggests that in oligotrophic conditions AAPB may thrive better than AnAPB (Ferrera *et al.* 2011). Conversely, some AAP taxa may contain the genetic capacity to perform AAP, yet are rarely observed doing so (Fecskeová *et al.* 2019). Some research indiates that AAPB are able to modify synthesis of BChl *a* in response to nutrient availability, and not only levels of darkness or oxygen tension (Cottrell *et al.* 2010). Likewise, it may be that in reducing their phototrophy in environments rich in labile organic carbon sources, some AAPB have simply jettisoned their phototrophic genes in favour of a purely heterotrophic physiology (discussed in Jiao *et al.* 2007 and Koblížek *et al.* 2013).

1.2 The plant as a microbial niche

1.2.1 Definitions

The phyllosphere is defined as the aerial or above-ground portion of vascular plants, and is most often implied to mean the surface of *i.a.* leaves. However, when discussing only the surface of the phyllosphere, One could argue that a more precise term would be the phylloplane. As in Newton *et al.* 2010, the term phyllosphere is used to indicate both the interior and exterior of the aerial portion of plants, but limit discussion to leaves only. Microbes of the phyllosphere are phyllophytes, and of those, endophytes reside within the plant tissue, and epiphytes on the surface.

1.2.2 Challenges and qualities

Although not the most extreme environment imaginable, the phyllosphere—much like a tidal pool—experiences rapid fluctuations in stressors and resources: diurnal, seasonal, and stochastic changes in light, heat, water availability, wind, and predation (Vorholt 2012). Furthermore, oxygen both present in the atmosphere and exuded from the plant, along with UV radiation exposure, demand phyllophytes to cope with intense oxidative stress. As a living matrix, the phyllosphere is itself subject to growth, senescence, predation, abscission, dehiscence, and a plant's metabolism can also vary. Although commensal and mutualistic relationships between plants and their phyllophytes do exist, leaves are nevertheless a critical component of a plant's wellbeing, and act as portals to the apoplast; they are accordingly well-defended by antimicrobial agents, and leaching of valuable resources such as sugars is actively minimized (Vorholt 2012). Therefore, the phyllosphere is an oligotrophic, transient, hostile, and abiotically challenging environment, with strict demands for residents.

1.2.3 Adaptations and colonization

Given the harshness of the leaf as an environment, at one time the phyllosphere was considered to be somewhat barren, hosting a small and static population of microbes. The modern view of the phyllosphere however, is that of a highly dynamic and heterogenous environment; not only are there interspecific differences, but there are also intraspecies variation (Kinkel 1997), down to the level of leaves of an individual (Hirano and Upper 1991). In addition, different ages of plant can harbour different populations (Lindow and Brandl 2003) with evidence suggesting that a species can “cultivate” a preferential microbiome (Leveau 2019). Phyllophyte abundance and diversity are determined by four main factors—growth, death, immigration, and emigration. Epiphyte population structure is principally determined by the latter two (Kinkel 1997), where mechanical effects (such as precipitation, aerial distribution, and abscission) govern the redistribution of microorganisms between leaves, plants, leaf litter, and soil (Hirano and Upper 1991; Kinkel 1997); epiphyte communities are accordingly very labile. Within the apoplast, endophytes are afforded more stability and protection from the elements and have a more ready access to nutrients (Lindow and Brandl 2003). However, access is more challenging, either by foliar invasion, transfer from the soil into the rhizosphere followed by vascular migration (Bringel and Couée 2015), or vertically *i.e.* through seeds *etc.* (Koskella 2020). Otherwise, migration is a relatively minor factor in endophyte population, with growth and decline being the principle shaper of their community. Endophytes thus observed to have a more steady community (Kinkel 1997) and present generally in lower diversity and abundance than on the phylloplane or rhizosphere (Given *et al.* 2020).

In general, phyllophyte wellbeing is closely tied to their plant hosts', where warm, shaded, and humid environments are particularly amenable to endophytes, with drought and intense solar radiation being most detrimental (Kinkel 1997). However, to survive at all, phyllophytes must be well adapted. For example, as is common with other bacteria—such as AAPB—which are routinely exposed to sunlight, phyllophytes are characteristically highly pigmented with carotenoids which remediate oxidative damage (Hamilton 2019). A particular challenge is accessing nutrients, which are scarce and distributed non-homogenously (Leveau and Lindow 2001; Remis-Emsermann and Schlecther 2018). The leaf has a complex micro-scale architecture—from a bacterium's perspective even the grooves between epidermal cells would seem significant. Generally, nutrients are located among these microniches such as leaf-veins and trichomes, where water tends to collect along with "leakage" of plant metabolites (Leveau and Lindow 2001), and bacteria tend to aggregate around them. That is to say, such microniches may be oases of water and nutrients saturated to carrying capacity by aggregates of persistent bacteria (Newton *et al.* 2010; Remis-Emsermann and Schlecther 2018) with outliers representing transient incomers which are not likely to thrive (Knief *et al.* 2010). Koh *et al.* (2011) hypothesised that AAPB would be found in the liminal zone between ice sheets and water, where they would be afforded protection form sunlight and planktonic predation; a similar observation has been made for epiphytes: greater numbers of epiphytes are located on the abaxial plane (*viz.* underside) of the leaf, where there is a thinner leaf-cuticle, greater density of hospitable microniches, and fewer chances of being struck by direct sunlight or precipitation (Newton *et al.* 2010). As the absorbance spectra of plant chlorophyll and BChl *a* do not overlap, neither endophytic nor abaxial epiphytic AAPB's phototrophy would suffer greatly for their locations.

In a population characterised by constant flux, established aggregate populations have a distinct advantage (Rastogi *et al.* 2013). An interaction of all these variables indicate that those microorganisms specialised to tolerate the challenges the environment poses can outcompete other residents/imigrants, grow to a sufficient population size (such that death and emigration do not result in a net population loss), and can form aggregates (reducing dispersal/removal) will thus become dominant.

This is interesting for AAPB, as their common presence in the phyllospher implies either supreme adaptability and/or a very large presence in immigrant populations (*e.g.* through rain or wind) potentially from distant origins (Hirano and Upper 1991). For AAPB to establish themselves as persistent residents, they must be well adapted to grow in population numbers, and not perish due to the nature of the niche. Clearly, phototrophy is an advantage over strict heterotrophy, but AAPB's close association with oligotrophic environments and diverse metabolism also indicate that they are well suited to utilizing recalcitrant resources. For example, a highly pigmented phyllophyte consistently found as a part of the core phyllosphere microbiome, *Methylobacteria*, is able to metabolise plant-derived VOC (Knief *et al.* 2010) and was later found to frequently be AAP$^+$ (Atamna-Ismaeel *et al.* 2012b). The surface of the leaf is also likely to be coated with pollen and other dispersed detritus—such factors AAPB have been noted to preferentially thrive alongside (Cottrell *et al.* 2010; Waidner and Kirchman 2007; Waidner and Kirchman 2008) which may supply them with organic matter. However, in the context of an aquaeous environment, detritus may aid in predation avoidance and reduction of direct sunlight, which would be less applicable in the phyllosphere (*ibid.*).

To which picture AAPB belong (diverse, abundant, resident, transient) is not yet clear. Because life in the phyllosphere can be so demanding, there are two methods for adaptation: tolerance or avoidance (Beatty and Lindow 1995), with avoidance in this context meaning migration or dormancy. All the data available from AAPB in aquatic bodies suggest that they are particularly amenable to tolerance, though their ability to survive drought is questionable. However, some evidence suggests that epiphytes do not experience quite as severe water-shortages as would be assumed (Lindow and Brandl 2003) and AAPB have also been found in arid soils (Csotonyi *et al.* 2010; Tang *et al.* 2018; Yang and Hu 2022).

Given this idea, one would assume that AAPB would form a more stable component of the inner-leaf and a relatively more dynamic one on the phylloplane. The general trend of diverse or abundant implies that epiphytic AAPB are more likely to be diverse and transient, though well-tolerant of conditions.

1.3 Methods for detecting AAPB

Uncovering the true diversity of plant-associated AAPB through metagenomic and culturomic methods alone, however, presents some issues. First of all, AAPB are metabolically and physiologically diverse, meaning that for a given growth medium, a subset of the AAPB population may be non-amenable to culturing (Yurkov and Beatty 1998). (For a more general critique of culturing methods in the phyllosphere, see *i.a.* Hirano and Upper (1991), Müller and Ruppel (2013) and Rastogi *et al.* (2013)). Furthermore, some AAPB require long incubation periods prior to observable colony formation, and—being that there is no selective growth medium for AAPB—they must be cultured alongside non-AAPB, raising the risk of antibiosis effects which distorts true diversity. Although generally AAPB colonies are detectable with the naked eye due to the richness of their pigments, confirmation requires spectroscopic detection of BChl *a* (Yurkov and Hughes 2017)) or sequencing (*e.g.* small-subunit rRNA and/or AP-related genes (Müller and Ruppel 2013)). One limitation of spectroscopic detection is that it relies on sufficient quantities of BChl *a*, which vary between species and under the given growth conditions. However, culture-dependent methods are still viable and important, particularly for physiological analysis; it is important to not overstate the limitations of culturomics: regarding the phyllosphere, one study found that approximately two thirds of the bacteria isolated were culture-amenable (Steifel *et al.* 2013), though this proportion is also dependent on plant species and many other factors (Knief *et al.* 2010).

Likewise, though shotgun metagenomic studies are free from the aforementioned complications and are not subject to primer selection-biases (Mao *et al.* 2012), (*i.e.* they are sequence agnostic), the richness of the detected populations is impacted by the numerical abundance of species and genera in the sample. That is to say, microbes of a certain class—though phylogenetically diverse—may not be so well represented in the metagenomic data should their genetic markers be dwarfed by more populous genera; metaphorically speaking, they may be so small that they fall through the gaps in the net. In addition, such methods can be costly in monetery, resource, space, time, and computational terms.

Amplicon-based genetic methods—where selected primer pairs target then amplify specific sequences of DNA from a pool—are in many terms more desirable. Polymerase chain reaction (PCR) is a fast, mature, and inexpensive technique which is readily available to most research institues, well documented, and is amenable to many downstream analyses. Although primer design and PCR-protocol optimisation is a complex and lengthy process in itself, well-formulated protocols tend to be robust, transferable and—in principle—amenable to context-specific optimisation. As PCR exponentially increases the concentration of the target amplicon, it can be used to detect trace amounts of the target: the investigated DNA sample itself may consist of a very low mass of DNA (like to "magnifying"), or the target within a sample may be present in relatively tiny concentrations (akin to "finding a needle in a haystack").

In its most primitive manifestation, analysis of PCR products (*e.g.* with agarose gel electrophoresis (AGE)) can offer a binary answer to the presence of a target; this has immediate value in *e.g.* epidemiology, where a microscopic virion can be genetically detected from a sample readily and with confidence. Downstream PCR-based analyses, such as amplicon-based sequencing, has long been a cornerstone for sequencing in general, be it Sanger sequencing or the more recently developed high-throughput *a.k.a.* next-generation sequencing (NGS). In contrast to Sanger sequencing, which requires the sequenced target to be homogenous, NGS is particularly applicable to studies investigating diversity, such as phylogenetic, population, and community analyses. On a cost-per-sequence analysis, NGS is vastly cheaper than so-called "first generation" sequencing. Unlike metagenomic shotgun methods, amplicon-based NGS can be highly selective.

The variable expression of both RC and BChl *a* means that RT-PCR and spectroscopic detection methods rely on sufficient transcription/translation of these components. With the exception of RT-PCR, amplicon-based methods can generally overcome this issue.

The design of primers is a key component of successful sequencing, where non-specificity—such as nucleotide mismatches—can cause erroneous amplification and poor results (Mao *et al.* 2012). This can be caused by sequence similarity between the target and other genes present in the sample, which can occur, for example, when using degenerate primers. Degenerate bases in a primer attempt to account for variations in the target sequence, which can be common with primers which are intended to be "universal". Good design of primers, knowledge of potential sequence similarities in the sample, and selecting primers which are specific to the target generally overcome such problems.

Unfortunately, there are no genes strictly exclusive to AAPB, making creation of selective primers a challenge (Yurkov and Hughes 2017). Instead, almost all primers used in AAPB research target the *puf* operon genes *pufM* and *pufL*, which encode RC2 apoproteins. These are shared with other AnAPB, notably with purple non-sulphur bacteria (PNSB). An exception are the primers bchY_fwd and bchY_rev, whose target gene functions in bacteriochlorophyll synthesis, and is thus universal to all anoxygenic phototrophs (Yutin *et al.* 2009). Such amplicon-based methods are able to differentiate between AAPB and AnAPB *implicitly*: when the sample is retrieved from an oxic environment, positive matches are classified as aerobic and, by extension, AAPB.

A long fragment spanning *pufL* to *pufM* was the first amplicon in AAPB research (Nagashima *et al.* 1997), and the forward primer, pufL, has seen consistent usage since its discovery. In contrast, the reverse primer, pufM, has been replaced several times. Achenbach *et al.* (2001) designed primers to amplify a fragment of the *pufM* gene, pufM.557 and pufM.750. While the former is original, the latter is very similar to Nagashima *et al.*'s pufM. Their target was all phototrophic bacteria utilizing RC2. These were shortly followed by primers pufMF and pufMr; the reverse primer is explicitly modified from primer pufM, and the forward primer is also unique, lying upstream of pufM.557 (Béjà *et al.* 2002). Based on vast metagenomic data, two further primers were created (pufM_uniF and pufM_uniR) to account for the growing knowledge of AAPB diversity, intended to be "universal" to all AAPB, as previous primers had been shown to be biased to certain taxa (Yutin *et al.* 2005). To account for diversity, these primers were designed to be highly degenerate. In addition, a new conserved region was discovered downstream from previous reverse primers, creating pufM_WAW (Yutin *et al.* 2005). Primers pufL67F and pufM781R were first used for purple non-sulfur bacteria (Tank *et al.* 2009) though have also been used for AAPB-research (Steifel *et al.* 2013). In the estimation of Gazulla *et al.*, universal primers pufM_uniF and pufM_uniR have often been problematic in certain environments (*e.g.* see Koh *et al.* 2011), and instead the primer pair pufMf paired with pufM_WAW has been the most successful (Gazulla *et al.* 2023). They additionally designed a new primer, pufMF_Y, intended to be paired with pufM_WAW, and assessed its performance in relation to other primer pairs. They found that the common pufMf/pufM_WAW pair suffered similar selectivity bias in contrast to pufM_uniF/pufM_uniR and the new pufMF_Y/pufM_WAW, which were reported to work optimally (Gazulla *et al.* 2023). A visual representation of most frequently used primers, and how they overlap and relate to one another is provided in Appendix A.

For this study, we chose to assess primers pufL (Nagashima *et al.* 1997), pufMr (Béjà *et al.* 2002), pufM_uniF, pufM_uniR, and pufM_WAW (Yutin *et al.* 2005) for plant-associated AAPB sensitivity. Previous work in the group with pufMf (Béjà *et al.* 2002) and a screen of AAPB strains selected from the phyllosphere determined that it did not respond to any of the strains tested (data not shown), and with pufM.750 bearing such similarilty to pufMr, I considered it redundant for this instance. With the taxonomic structure of plant-associated AAPB being in a comparatively early stage, I hypothesised that universal primers pufM_uniF/pufM_uniR would provide a more detailed community structure. Primer pufM_WAW is commonly used, and provides a slightly longer amplicon, which could allow for taxonomic analysis at a greater depth. Amplifying both the *pufL* and *pufM* gene with primer pufL also bears the same rationale; however, this amplicon would be too long for IonTorrent NGS.

1.4 Aim of the study

This study investigated the use of amplicon-based genetic methods for detection of plant-associated AAPB using 8 AAPB strains and 36 environmental samples; AAPB strains were derived from plants, and environmental samples were DNA extracted from the leaves of lingonberry (*Vaccinium vitis-idaea*), a plant found to reliably harbour AAPB (Nissinen *et al.* 2023). Several AAPB-specific primers were evaluated for specificity and efficiency, the most superior of which was then used in high-throughput sequencing alongside 16S primers designed to quantify the samples' eubacterial population.

The aim of the sequencing component of the study was a "proof-of-concept" potentially indicating a direction for future studies. The method of sampling was aimed to be minimally sufficient for a rudimentary phylogenetic and community analysis of detected AAPB, and the addition of the eubacterial population in the methods and results was hoped to both corroborate and bolster the validity of the analysis.

Though amplicon-based methods should not be used in place of culturomics or metagenomics, it would be a welcome addition to the field, potentially broadening our understanding of these fascinating bacteria's role in plants. Every scientific method is burdened with biases and confounding variables, thus various techniques or -omics can be used to "triangulate" results, hopefully illuminating the actual state of affairs in the subject under study: in two parallel studies, genetic and microscopic methods showed differing levels of epiphytic AAPB (Atamna-Ismaeel *et al.* 2012a; Atamna-Ismaeel *et al.* 2012b) (see also Steifel *et al.* 2013.)

Given the transient and unstable nature of plants as a microbial substrate, microbial population and commuity dynamics in the phyllosphere (and by extension—though to a lesser extent—plant endosphere) are remarkably fluid. It has been observed that marine AAPB exist in high diversity but low abundance in oligotrophic waters, with the inverse true for eutrophic waters (Jiao *et al.* 2007). Given that the phyllosphere is an oligotrophic environment with a fluid community structure, it may be that plant-associated AAPB also exist in a similar low-abundance, high-diversity manner. If this is so, understanding the plant AAPB microbiome may indeed require methods similar to those outlined in this study,

A well designed NGS protocol offers rapid and reliable analysis of microbiome, and with optimisation (such as incresed sample flowthrough and choice of superior primers) studies in plant-associated AAPB may be expanded to incorporate a broader swathe of samples and time-frames. This the present state of the majority of AAPB studies, which predominantly focus on marine and freshwater environments, and have helped diversify the field so much in the past four decades.

2 METHODS

2.1 Sample collection

From three different sites, three small branches of lingonberry (*Vaccinium vitis-idaea*) were collected, representing biological replicates. In turn, leaves from each of the biological replicates were portioned into two batches to make technical replicates. From each technical replicate, DNA from both the surface and interior of the leaves were extracted, thus giving a total of 36 DNA samples.

2.1.1 Collection site and times

I collected the specimens of lingonberry from the north-west slopes of Ylistönrinne, Jyväskylä on 26.08.2022. There were three different sites, each at least 100 m from one another: Site A (62.228 103° N, 25.742 185° E) Site B (62.227 015° N, 25.739 390° E) Site C (62.226 023° N, 25.738 890° E)

The following criteria were used when selecting specimens: Foliage not evidently diseased, soiled, dessicated, or in contact with soil or other plants; stems at least 150 mm long and unbranched, and being generally representative of the plant (*i.e.* not having too few/many leaves, leaves not being unusually large or small.)

Using a sterile technique, suitable specimens were collected by cutting the stem approximately 150 mm from the apical tip and then storing in individual premarked boxes. Specimens were immediately transferred to a 7 °C refridgerator whilst awaiting further processing.

2.2 Sample processing

Prior to chemical extraction of DNA from the samples, the epiphytic and endophytic bacteria had to be physically isolated from one another. To prevent cross-contamination, only intact leaves could be selected from the stems and—after surface material collection—the exterior of these leaves had to be sterilized.

In order to chemically extract DNA from the leaf tissue, the entire leaves first had to be mechanically homogenized. Although it is technically possible to proceed directly to chemical extraction of leaf-surface DNA, these samples were also subjected to the same mechanical homogenization process as the leaf tissue. This was to prevent any potential confounding variables in the process from arising, thus simplifying analysis of any differences between epiphyte and endophyte results.

The homogenization vessels mentioned below were 2 mL, reinforced, plastic, screw-cap tubes into which I had added two chromed-steel ball bearings (ø = 3.2 mm) and approximately 50 µg of glass beads (ø = 0.1 mm.)

2.2.1 Leaf-surface material isolation

I removed 300 mg of intact leaves from each biological replicate proceeding from the apical tip downwards. These leaves were placed into preweighed, sterile 15 mL conical tubes.

Isolating the surface matieral from the leaves was achieved by sonication in an isolation buffer—a potassium phosphate solution containing a surfactant (20 mmol L^{-1} KPi, pH 6.5, 0.005 % (v/v) Silwet). 4.5 mL of the isolation buffer was added to each vessel, which sufficed to cover all the leaves within. These were then sonicated for 3 min in a sonication bath (FB 15 046, Fisherbrand®) at RT.

After sonication, 4 mL of extraction buffer from each of the replicates was extracted and aliquoted equally into two, 2 mL Eppendorf microcentrifuge tubes. Each tube was centrifuged on a tabletop centrifuge for 3 min at 13 000 × g. From each of the microcentrifuge tubes, 1.8 mL of the supernatant was discarded, then 400 µL of KPi buffer (20 mmol L^{-1}, pH 6.5) was added to resuspend the pellets. These were centrifuged once more at 13 000 × g, and 400 µL of supernatant was removed, leaving behind a pellet in 200 µL of solution. These technical replicants of the leaf-surface material were resuspended then transferred to homogenization vessels and stored at −80 °C until homogenization.

2.2.2 Leaf-tissue DNA isolation

The leaves which remained in the conical flasks after sonication required further processing to ensure that no residual epiphytes would pass through to the endophyte isolation. The leaves from each biological replicate were gathered into labelled, stainless steel tea-balls and then immersed in a 3 % sodium hypochlorite solution for 3 min. They were then washed thrice in sterile, double-distilled water for 3 min. The leaves were then portioned into three, 100 mg batches. One batch was reserved, unprocessed, for storage. The other two batches were dissected—each leave was cut once medially, and thrice transversely. The two dissected batches of leaves were transferred to homogenization tubes and stored at −80 °C until homogenization.

2.2.3 Mechanical homogenization

The samples were transferred from −80 °C directly to a −20 °C ice block, which was stored at 7 °C for half an hour to allow the samples to slowly thaw. The samples were then placed in a bead mill homogenizer (Bead Rupter Elite, Omni International) which had been precooled to 0 °C with liquid nitrogen according to the manufacturer's instructions. Homogenization was performed at 5 m s^{-1} with 3 × 30s homogenization and 10 s dwells.

2.2.4 DNA extraction

After homogenization, lysis and DNA extraction was performed using Spin Plant Mini Kit (Invisorb®) according to the manufacturer's directions, though with a few modifications; lysis was performed for 45 min instead of the recommended 30 min—we had found that the waxy cuticle and starchy content of the lingonberry leaves demanded a longer incubation time in order to ensure a thorough lysis. Also, the lysate was briefly centrifuged to pellet the densest, unlysated tissue mass, and only the supernatant was transferred to the pre-filter—placing the whole lysate on the pre-filter as recommended caused it to block and tear in downstream steps. The extracted DNA samples were transferred to 2 mL cryogenic tubes and stored at −80 °C.

2.3 Measuring DNA concentration

DNA concentration was measured using QuBit®dsDNA HS Assay kit (Invitrogen Life Technologies) according to the manufacturor's instructions.

2.4 AAPB strains

The Jaettu Valo project has cultured a variety of plant-associated bacteria which have been determined to be AAP^{+} via NIRis – a 3D-printed imaging device for detection of BChl *a* directly from bacterial colonies on Petri dishes (Franz *et al.* 2023; Nissi-

Table 1: AAPB strains used to determine the response of various APB primers used in this study. Strains were collected as part of the Jaettu Valo project (Nissinen *et al.* 2023). Names given in parenthesis indicate the probable genus.

Code name	**Taxonomy/genus**	**Source plant**	**DNA conc.** ng µL^{-1}
Methyl	*Methylobacterium*	*Betula pubescens czerepanovii*	0.290
SphingA	*Sphingomonas*	*Saxifraga oppositifolia*	26.6
LichenA	*Lichenibacter*	*Huperzia selago*	3.11
Aureim	*Aureimonas*	*Vaccinium vitis-idaea*	6.50
SphingB	*Sphingomonas*	*Vaccinium vitis-idaea*	25.0
LichenB	*Lichenibacter*	*Pinus sylvestris*	6.62
Caball	*(Caballeronia)*	*Betula pubescens*	48.5
SphingC	*Sphingomonas faeni*	*Vaccinium vitis-idaea*	45.5

nen *et al.* 2023). Many of the cultures have been taxonomically identified using sequencing methods. In order to test the responsiveness of the APB-specific primers to various species of AAPB, my supervisor, Riitta Nissinen, supplied me with 8 such bacterial strains which she had determined to be phylogenetically diverse. Table 1 lists these bacterial strains, along with their DNA concentration.

2.5 PCR

All PCR reactions were carried out in a C1000 Touch™Thermal Cycler (Bio-Rad) controlled with C1000 Manager and CTX Maestro software (Bio-Rad). The PCR buffer used was DreamTaq Green PCR Master Mix (2X) which contained all the required buffers (2X DreamTaq Green buffer), dNTPs (25 µmol L^{-1}), ions (0.4 mmol L^{-1} $MgCl_2$), polymerase (DreamTaq DNA polymerase), and density reagent & tracking dyes allowing for direct loading onto gels. Primers and mastermixes were diluted to appropriate concentration using the nuclease-free water supplied with the Master Mix kit.

Table 2 lists all of the primers used in this study. All primers were diluted from stock to a concentration of $100\,\mu\text{mol}\,\text{L}^{-1}$ with nuclease-free water and stored at $-20\,°\text{C}$. Four dfferent PCR protocols (A-D) were used (Table 3). Unless otherwise stated, the primer concentration for all four of the PCR protocols was $0.4\,\mu\text{mol}\,\text{L}^{-1}$ and reaction volumes 25 µL. In the following sections, the amount of template added to each reaction is expressed in mass (ng). Based on the concentration of the template, the volume equivalent to the target mass was added using formula 1.

$$V = m/C \tag{1}$$

Equation for determining volume (V) of template to add, given a target mass (m) and known concentration (C) of template.

2.5.1 AGE

All PCR reactions were evaluated with AGE using a separation chamber (Owl separation System, Model B2) with a 1X TA running buffer at 100 V (controlled by a VWR Power Source 250 V) for 1 h. Agarose gels were made on site prior to AGE. Unless otherwise stated, each gel was a 1 % w/v powdered agarose/TA (1X) solution with a volume of 100 mL. Once the solution had reached a temperature of approximately 70 °C, 2 µL of SYBR™Safe DNA gel stain (10 000× in DMSO, ThermoFisher Scientific®) was added, mixed well, poured into a casette, then allowed to cool and polymerise to solidity at RT. Volume of samples loaded into gels was always 5 µL. The DNA ladder used was GeneRuler DNA Ladder Mix (ThermoScientific®) with a volume of 1 µL.

Completed AGE gels were visualised using a Gel Documentation System (Axygen®) controlled with Axygen®Capture software.

2.6 Testing of APB-specific primers

In order to test the responsiveness of three APB-specific primer pairs (pufM_uniF/-pufM_uniR, pufM_uniF/pufM_uniR_WAW, and pufLf/pufMr), two different sets

Table 2: Primers used in the study. Lower case letters in the sequences indicate degenerate nucleotides.

Name	Sequence (5′–3′)	Reference
pufM_uniF	GGnAAyyTnTwyTAyAAyCCnTTyCA	(1)
M13_pufM_uniF	TGTAAAACGACGGCCAGTGGnAAyyT–nTwyTAyAAyCCnTTyCA	
pufM_uniR	yCCATnGTCCAnCkCCArAA	(1)
P1_pufM_uniR	CCTCTCTATGGGCAGTCGGTGAyCCA–nGTCCAnCkCCArAA	
pufM_WAW	AYnGCrAACCACCAnGCCCA	(1)
pufLf	CTsTTCGACTTCTGGGsGG	(2)
pufMr	CCCATsGTCCAGCGCCAGAA	(3)
799F	AACmGGATTAGATACCCkG	(4)
1062F	GTCAGCTCGTGyyGTGA	(5)
M13_1062F	TGTAAAACGACGGCCAGTGTCAGCTC–GTGyyGTGA	
1390R	ACGGGCGGTGTGTrCAA	(6)
P1_1390R	CCTCTCTATGGGCAGTCGGTGATACG–GGCGGTGTGTrCAA	
1492R	GGyTACCTTGTTACGACTT	(4)
SA429f	TAAAGCTCAAGCTCTTTTACCCG	(7)
SA933r	AAACCACATGCTCCACC	(7)

Degenerate nucleotide key: n=A/T/G/C; y=C/T; w=A/T; k=G/T; r=A/G; s=G/C; m=A/C.
References: (1)=Yutin *et al.* 2005; (2)=Nagashima *et al.* 1997; (3)=Béjà *et al.* 2002; (4)=Chelius and Triplett 2001; (5)=Ghyselinck *et al.* 2013; (6)=Zheng *et al.* 1996; (7)=Zhou *et al.* 2012

Table 3: PCR protocols used in this study. After initial denaturation, the denaturation, annealing, and extension stages were repeated sequentially as described in the column *Cycles* before a final extension phase. The PCR protocols were taken from the referred sources.

Name	**I.D.**		**Cycles**	**D.**		**A.**		**E.**		**F.E.**		**Ref.**
	°C	T.		°C	T.	°C	T.	°C	T.	°C	T.	
A	94	3 min	30	94	45 s	54	45 s	72	60 s	72	5 min	(1)
B	94	5 min	35	94	60 s	55	45 s	72	90 s	72	8 min	(2)
C	94	3 min	35	94	30 s	50	45 s	72	30 s	72	10 min	(3)
D	94	3 min	35	94	30 s	54	45 s	72	30 s	72	10 min	(4)

Key. I.D.=Initial denaturation; D.=Denaturation; A.=Annealing; E.=Extension; F.E.=Final extension.
References. (1)=Kumar *et al.* 2017; (2)=Zhou *et al.* 2012; (3)=Yutin *et al.* 2005; (4)=Achenbach *et al.* 2001

of templates were used. In the first, 1 ng of the 8 AAP$^+$ strains (section 2.4) were used as templates. In the second, 30 ng of leaf-tissue DNA and 3 ng of leaf surface DNA were used as templates. Previous preliminary tests had demonstrated that leaf-surface DNA responded better to these primers than leaf-tissue DNA (data not shown), which is why more leaf-tissue samples were chosen. PCR protocol C was used for primer pairs pufM_uniF/pufM_uniR and pufM_uniF/pufM_uniR_WAW, and protocol D for primers pufLf/pufMr (Table 3). The primer pair SA429f/SA933r (Table 2) also targets a variable region of the 16S rDNA. This target sequence is, however, specific to the genus *Sphingomonas* (Zhou *et al.* 2012), which has been determined to be common plant-associated AAP bacteria (Nissinen *et al.* 2023). In the latter test, *Sphingomonas*-specific primer pair (SA429f/SA933r) were also used with protocol B (Table 3).

2.7 PCR optimization for primer pair pufM_uniF/pufM_uniR

The primer pair pufM_uniF/pufM_uniR was determined to be the most suitable primer pair for this study (see section 3.2), though required optimization for leaf-tissue DNA samples. In all tests, primer protocol C (Table 3) was used with these primer pairs.

2.7.1 Template concentration

I tested the effect of varying template masses on reaction efficiency using a single leaf-surface sample (B2pi) and three leaf-tissue samples (A2si, B2si, and C2si). DNA masses added were 3 ng, 1 ng, 0.5 ng and 0.1 ng for leaf-surface samples, and 30 ng, 10 ng, 5 ng and 1 ng for leaf-tissue samples. In a second test, I investigated the effect of increasing template mass of leaf-tissue DNA using samples A2si, B2si, C2si, and C3si. Masses added were 30 ng, 60 ng and 90 ng. In each test 2 ng of sample SphingC (*Sphingomonas*, see table 1) was used as a positive control.

2.7.2 Annealing temperatures

The annealing temperatures of the reaction was tested on a gradient of 45 °C to 64 °C. Templates B2si, B2pi, and SphingC (*Sphingomonas*) were used with masses of 30 ng, 3 ng and 2 ng respectively.

2.7.3 Annealing cycle, annealing temperature, and primer concentration

Because a single PCR variable may not improve reaction efficiency on its own, I tested the effect of three variables in concert: annealing temperatures 50 °C and 47 °C, primer concentrations 0.4 $\mu\text{mol}\,\text{L}^{-1}$ and 0.2 $\mu\text{mol}\,\text{L}^{-1}$, and number of reaction cycles from 35 to 42. Templates tested were A1si, A2sii (30 ng), A1pi, and A3pii (3 ng). The lower primer concentration (0.2 $\mu\text{mol}\,\text{L}^{-1}$) was only tested with the leaf-tissue DNA samples.

2.8 Library preparation

In this study, library preparation for IonTorrent utilized the "barcoding" method as detailed in Mäki *et al.* (2016). This involves a minimum of two stages. In the first stage, an M13-tagged version of the forward primer is used along with the standard reverse primer. DNA "barcode" polynucleotides linked to a cognate M13 sequence are used as forward primers in the second stage, along with P1-tagged versions of the standard reverse primer. The final amplicons are thus genetically labelled according to their sample source and are equipped with the P1 sequence required for IonTorrent's emulsion PCR. In the former case, any samples without barcodes can be computationally filtered from the sequence data, and in the latter instance any amplicons without the P1 tag are not sequenced at all.

The leaf-tissue homogenization process liberates into solution host-cell organelle DNA which can quantitatively dominate the target bacterial DNA. A process which can either exclude or physically differentiate the host organelles' DNA/amplicons from the bacterial targets not only facilitates the analysis of sequence DNA, but also allows for a PCR which is less skewed by unwanted template genes thus creating a eubacterial sequence library with higher fidelity.

Detection and sequencing of the leaf-associated eubacterial community was achieved using a similar method as described in Kumar *et al.* (2017). In PCR-1, primers designed to differentially amplify bacterial and mitochondrial 16S rDNA sequences to the exclusion of plastid rDNA (799f and 1492r (Chelius and Triplett 2001)) create amplicons 735 and 1090 bp long for eubacteria and mitochondria respectively. PCR-2 uses the amplicons from the PCR-1 as template and primers 1062f (Ghyselinck *et al.* 2013) & 1390r (Zheng *et al.* 1996), producing eubacterial amplicons covering V6-V8 of the 16S gene which are both short enough to be processed by IonTorrent PGM, and can still be size fractionated from the larger, mitochondrial, 18S-derived amplicons.

Using the methods described above, two sequencing libraries were prepared. The first library was composed of *pufM* gene fragments from the primer pair pufM-_uniF/pufM_uniR, and the second was composed of *16S* gene fragments. Technical and logistical restraints meant that only a limited number of barcodes were available, and that sequencing would have to be performed in two different runs (batches). This necessitated the processing of the libraries in four different pools: **Pool 1**: *pufM* (batch 1); **Pool 2**: *16S* (batch 1); **Pool 3**: *pufM* (batch 2); **Pool 4**: *16S* (batch 2).

The APB-specific (*pufM*) library was prepared in two PCR stages.

PCR 1: *Primers*–M13_pufM_uniF, pufM_uniR; *Protocol*–C; *Templates*–see table 5, appendix B.

PCR 2: *Primers*–M13-barcode (1 µL), P1_pufM_uniR; *Protocol*–C, with only 8 cycles; *Templates*–1 µL from PCR-1.

Eubacterial (*16S*) library preparation was achieved in three PCR stages:

PCR 1: *Primers*–799F, 1492R; *Protocol*–A; *Templates*–see table 5, appendix B.

PCR 2: *Primers*–M13_1062F, 1390R; *Protocol*–A; *Templates*–2 µL from PCR-1.

PCR 3: *Primers*–M13-barcode (1 µL), P1_1390R; *Protocol*–A, with only 8 cycles; *Templates*–Leaf-surface samples, 0.5 µL of PCR-2; leaf-tissue samples, 1 µL of PCR-2.

2.8.1 Evaluating amplicon concentration

All samples from the final PCRs were analysed with a 1.5 % (w/v) AGE. The apparent brightness of the amplicons in the AGE images were considered to correspond to their relative concentrations. Thus, samples could be grouped according to similar target amplicon concentration. In this manner, samples in the *pufM* library were divided into 5 and 6 groups of similar target concentration for leaf-tissue and -surface DNA respectively; amplicons from the leaf-surface samples in the *16S* library were all of similar concentration, and leaf-tissue samples could be divided into two different groups. Representative samples from each of the groups were analysed with TapeStation (section 2.8.3) to determine the concentration of target amplicons. This data was used to extrapolate the target amplicon concentration of all the samples, and from this a volume equivalent to 20 ng was calculated using formula 1. For each pool, the appropriate volume of each sample was added to a 2 mL Eppendorf microcentrifuge tube to form an approximately equimolar mixture of target amplicons, which were subsequently stored at −80 °C.

2.8.2 PippinPrep

Target amplicons were isolated and purified from the PCR product using PippinPrep (Sage Science®) according to the manufacturor's instructions with a 2 % Ethidium-free Agarose Gel Cassette (Sage Science®). Amplicons in the range of 220 to 320 nt were isolated from pools 1 and 3 (*pufM* amplicons), and amplicons between 350 to 550 nt long from pools 2 and 4 (*16S* amplicons). After isolates were collected into 1.5 mL Eppendorf microcentrifuge tubes, samples were analysed for quality and concentration using TapeStation as described in section 2.8.3.

2.8.3 TapeStation

Profiling of DNA concentration and size was performed using TapeStation 2200 (Agilent Technologies) according to the manufacturer's instructions with a ScreenTape gel (Agilent Technologies) and analyzed using Tapestation Analysis Software (version A.02.02 (SR1)). For evaluation of the final amplicon concentration (section 2.8.1) all samples were diluted by a factor of 1:10, except *pufM* samples B1sii, C1sii, and C2sii. For analysing the DNA concentration of the pooled samples, pools 2 and 4 were diluted by a factor of 1:5 with nuclease-free PCR-grade H_2O whereas pools 1 and 3 were undiluted.

2.8.4 NGS and sequence analysis

Sequencing was performed at the University of Jyväskylä using IonTorrent PGM with an Ion PGM Hi-Q sequencing kit, following manufacturer's instructions. This was performed in two batches as described above, each containing 400 ng of pooled, equimolar samples. The sequenced reads were binned into samples according to their cognate barcodes on an IonTorrent server. Quality control, read filtering, reference sequence assignment, and OTU clustering were performed on CLC genomics workbench software with the Microbial Genomics module (https://digitalinsights-.qiagen.com).

More precisely, sequences with low quality and which did not contain both forward and reverse primer sequences were excluded. After binning to sample and type (eubacterial/AAPB), barcodes and primer sequences were removed. Any *pufM* sequences less than 120 and greater than 220 nt long were also excluded. OTU clustering was performed *de novo* at a 97 % identity level for *pufM*, and with SILVA at 99 % resolution for the *16S* fragments. All mitochondrial and chloroplast sequences were removed. Finally, OTU tables were transferred to PRIMER Permanovo software for further analysis.

3 RESULTS

3.1 DNA concentrations from extractions

Concentrations of DNA extracted from leaf-surface and -tissue are listed in appendix B. Despite the mass of leaf tissue being constant, the concentrations of DNA extracted from them varied between 18.0 ng μL^{-1} to 72.2 ng μL^{-1}. There is a much larger variation in DNA concentration between the samples from the leaf-surface (0.066 ng μL^{-1} to 7.50 ng μL^{-1}). This is to be expected, as—unlike the leaf-tissue—the mass of surface material could not be controlled. Variations in DNA concentration within or between sites, specimens, replicates, or samples is randomly distributed, though in the case of the leaf-surface samples the variations in concentration between technical replicates is most pronounced.

3.2 Response of primer pairs to various templates

The three APB-specific primer pairs tested responded in vastly different ways to the screen of 8 plant-associated AAPB strains (Figure 1). PufL/PufM produced amplicons in only one of the strains. pufM_UniF/pufM_UniR_WAW produced inconsistent results, responding appropriately to 5 strains, and of these, three (Sphingomonas_A, Lichenibacter, Caballeronia) effected multiple and inordinately large amplicons. In contrast, pufM_UniF/pufM_UniR responded to 6 strains, and all am-

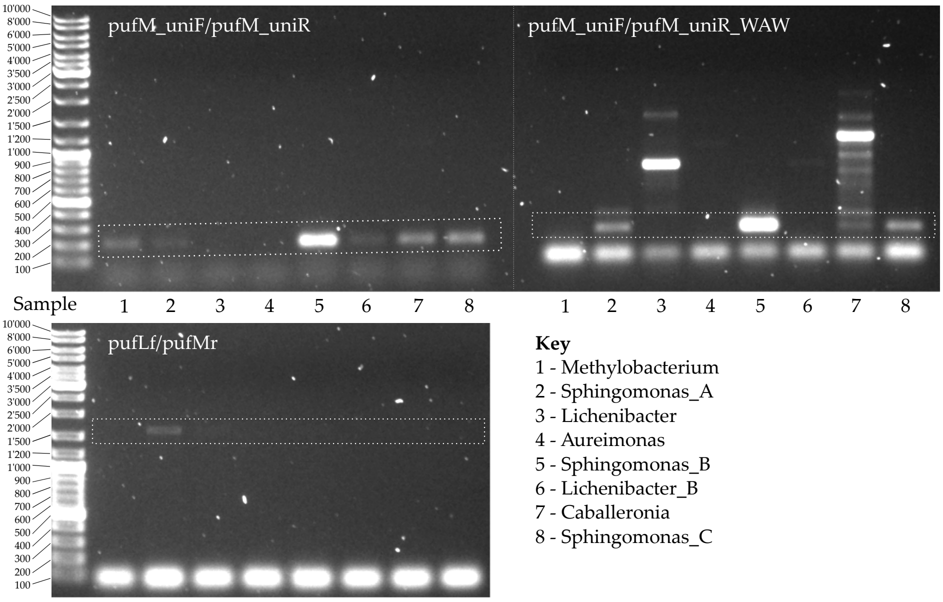

Figure 1: Response of 8 AAPB strains to APB-specific primer pairs (pufM_uniF-/pufM_uniR, pufM_uniF/pufM_WAW, pufLf/pufMr). Target amplicon region is shown with a white, dotted box within the AGE. Template mass for all samples was 1 ng.

plicons were of the target size. None of the tested primer pairs responded to the *Aureimonas* strain.

In a similar manner to how they reacted to plant-associated AAPB the primer pairs pufM_uniF/pufM_WAW and pufLf/pufMr produced poor results with leaf-surface and leaf-tissue DNA samples (Figure 2). With pufM_uniF/pufM_WAW large quantities of non-specific amplification were observed in all samples tested, with amplicons being longer than the target. Although it appears that target amplicons were found in the leaf-surface templates, there were so many non-specific amplicons that they manifested as “streaks” instead of bars. There was also a slight difference in the size of the nominal target amplicons between leaf-surface and leaf-tissue DNA templates. Likewise, for primer pair pufLf/pufMr we observe little to no amplification whatsoever with plant-tissue samples, though some very weak target amplicons in two of the three leaf-surface samples tested.

Primers pufM_uniF/pufM_uniR had optimal response with leaf-surface DNA samples, though did not produce any visible target amplicons from the leaf-tissue DNA

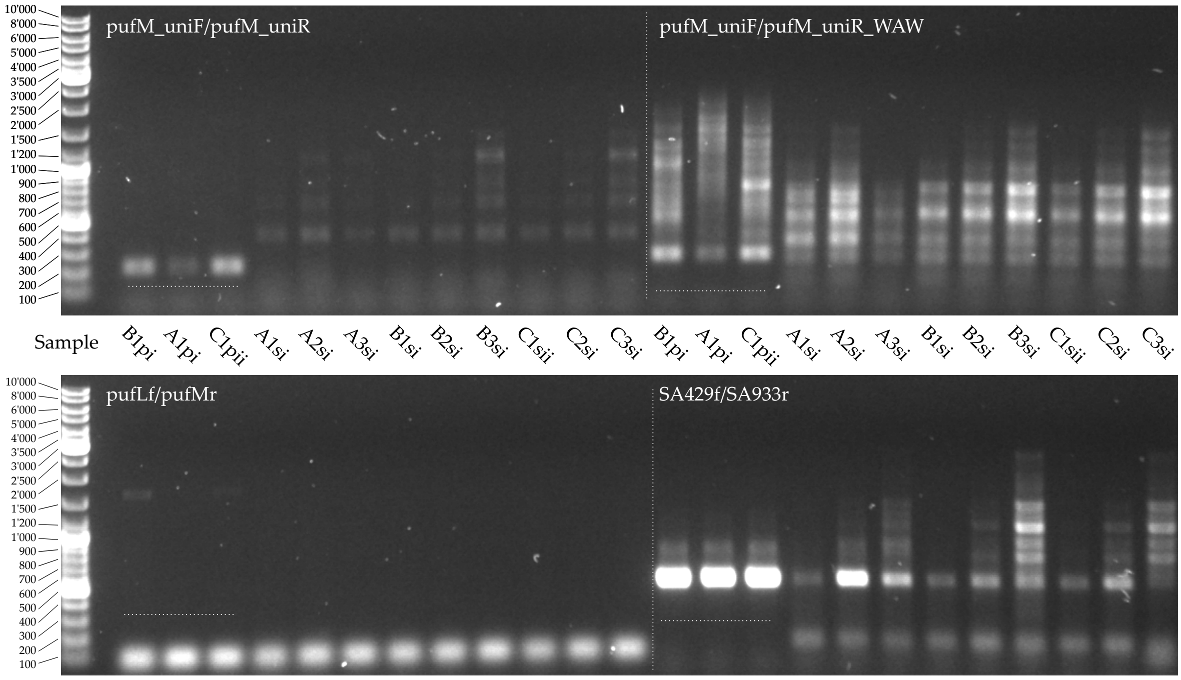

Figure 2: Primer response to a screen of three leaf-surface (dotted horizontal lines) and 9 leaf-tissue DNA extracts. In addition to the three APB-specific primer pairs (pufM_uniF/pufM_uniR, pufM_uniF/pufM_WAW, pufLf/pufMr) the response of an additional *Sphingomonas*-specific primer pair (SA429f/SA933r) was also investigated.

tested. Instead, it produced stereotypical amplicons of sizes approximately 400, 700, 900 and 1200 bp long. When comparing the AGE of pufM_uniF/pufM_uniR and pufM_uniF/pufM_WAW one notices that all of the tested samples produced similar amounts of amplicons regardless of which primer was tested—*e.g.* samples B1pi, A3si and C1sii had the lowest overall PCR efficiency, C1pii, A2si, B3si, and C3si had the highest efficiency.

The *Sphingomonas*-specific primer pair SA429f/SA933r produced target amplicons in all of the samples, though, similar to pufM_uniF/pufM_uniR, also produced non-specific amplicons of stereotypical lengths in quantities that varied between samples. The variation in intensities between the leaf-tissue DNA is somewhat similar to those observed in pufM_uniF/pufM_uniR and pufM_uniF/pufM_uniR_WAW, but not identical. The pattern of target amplicon intensity does not match the pattern of the non-specific amplicons/general intensity: A1si and C3si have similar target amplicon intensity, though the former is nearly "optimal" whereas the latter has very intense non-specific amplicons.

3.3 Optimisation of pufM_uniF/pufM_uniR

3.3.1 Template mass

Alterations in the mass of template added to the reaction showed that the pufM-_uniF/pufM_uniR primers could detect AAPB in leaf surface samples of DNA masses low as 100 pg (Figure 3). However, changing the leaf-tissue DNA masses from 1 ng to 90 ng could not ameliorate target amplification. Instead, the highest (60 ng to 90 ng) template masses reduced the concentration of the 1200, 900 and 700 bp amplicons and strengthened the 400 bp amplicon. Lowering the leaf-tissue template mass below 10 ng appears to have simply lowered the general PCR efficiency.

3.3.2 Annealing temperature

The temperature gradient, which was not linearly distributed amongst the reaction vessels, resulted in the following 8 different annealing temperatures: 45.0 °C, 46.2 °C, 48.9 °C, 52.7 °C, 57.6 °C, 61.6 °C, 63.8 °C and 65.0 °C. The PCR reaction ceased functioning when the annealing temperature was raised beyond 52.7 °C for the leaf-tissue sample and 57.6 °C for the leaf-surface sample and positive control (Figure 4). Annealing temperatures lower than the standard 50 °C (Table 3, protocol C) did not appear to impede the efficiency of the positive control or leaf-surface DNA sample, though for the leaf-tissue DNA sample the amplicon pattern changed in a manner similar to when template mass was increased (section 3.3.1), namely the 400 bp amplicon dominated as other, larger ones faded.

3.3.3 Other PCR parameters in conjunction

The lower primer concentration of 0.2 μmol L^{-1} in combination with the other parameters tested did cause a visible improvement in target amplification from leaf-

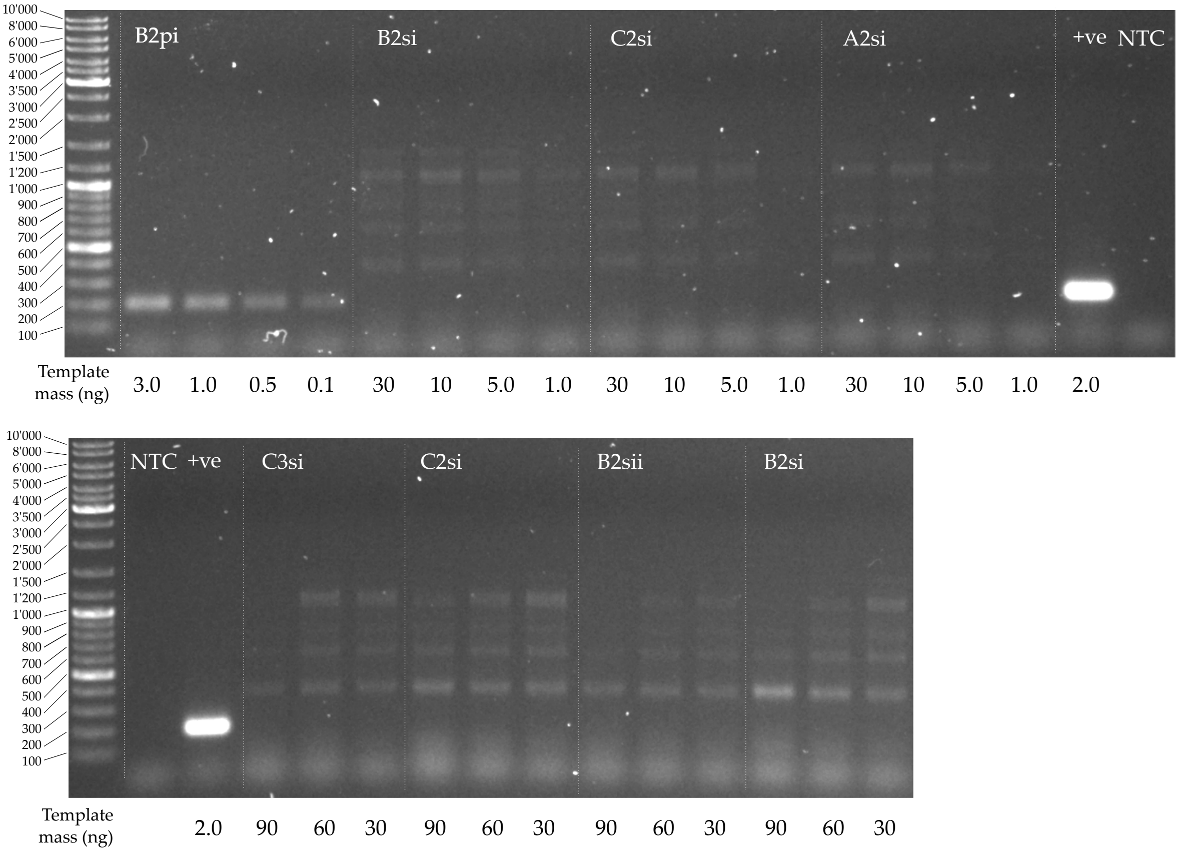

Figure 3: Effect of varying masses of added template masses from 3 ng to 0.1 ng for leaf-surface extracts, and 90 ng to 1 ng for leaf-tissue extracts. Primer pair used was pufM_uniF/pufM_uniR. +ve control used was SphingomonasC (*Sphingomonas*), 2 ng.

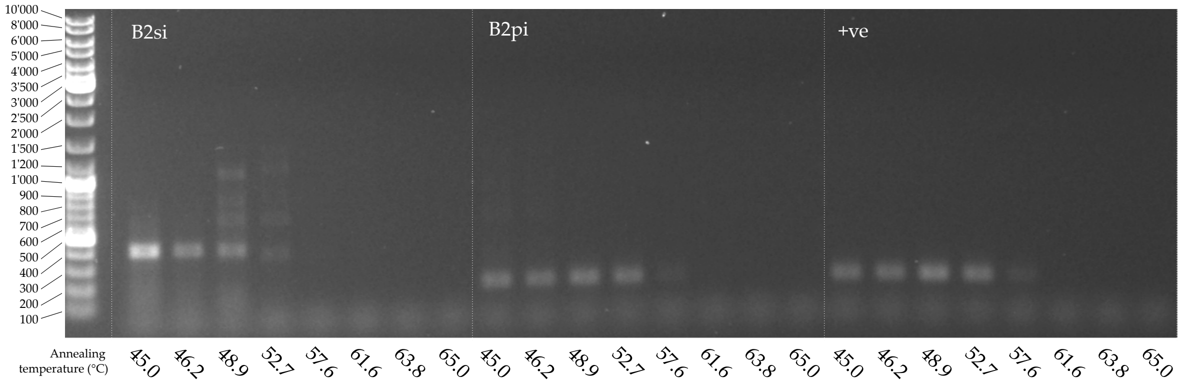

Figure 4: Effect of various annealing temperatures on the PCR efficiency of primer pair pufM_uniF/pufM_uniR with templates Bs2i (30 ng), Bp2i (3 ng), and positive control SphingomonasC (*Sphingomonas*, 2 ng).

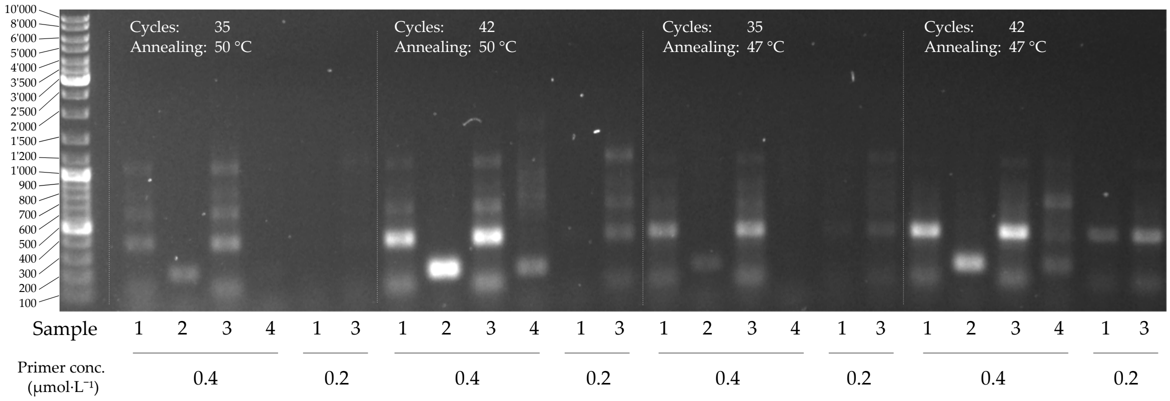

Figure 5: Effect of various parameters (cycle number, annealing temperature, and primer concentration) in different permutations on the PCR efficiency of primer pair pufM_uniF/pufM_uniR.

Sample legend: 1=A1si, 2=A1pi, 3=A2si, 4=A3pii.

tissue samples (Figure 5). Likewise, no improvements were made for leaf-tissue samples when annealing temperature was lowered, reaction cycles were increased, or when both variables were combined.

For leaf-surface samples, the deviations from the standard protocol brought no improvements. Interestingly for sample A3pii, when exposed to the lower annealing temperature and higher cycle number, amplicons of 400 and 700 bp stereotypical of the leaf-tissue DNA began to appear.

3.4 Library preparation

3.4.1 AGE of barcoded and labelled samples

After barcoding and P1-tagging, the target *pufM* amplicons were approximately 300 nt long (Appendix D, Figure 9). These amplicons were found in all of the leaf-surface samples, though in varying quantities. Interestingly, unlike it previous tests with the primers pufM_uniF/pufM_uniR, after barcoding and labelling, the target amplicon was visible in many of the leaf-surface samples: particularly notable are samples A2si and A3si. Due to a suspected pipetting error, the reaction for sample B1si failed completely.

The eubacterial *16S* amplicons were approximately 450 nt long and present in all of the samples (Appendix D, Figure 10). In the leaf-surface samples all amplicons were of equal intensity, though there was some variation in the leaf-tissue samples. The latter also contained evident 800 nt long mitochondrial-derived amplicons.

3.4.2 Amplicon analysis

TapeStation analysis of the size-fractionated pools showed that the *pufM* amplicons had a peak size of 300 nt. Prepared library *pufM* amplicons from sample A1si was analysed with TapeStation, and compared to size-filtered pool 3 (isolated target *pufM* amplicons) in Appendix C. In sample A1si, no target amplicons are detected in the AGE, and are hardly detectable at all in the TapeStation data. In contrast, the approximately 300 bp target fragment is clearly visible in the AGE of sample A2si. Despite PippinPrep filtering between 220 and 320 bp fragments, the TapeStation data indicate that fragments between approximately 280 and 380 were extracted, with peak at 306 bp.

3.5 Sequencing results

After removal of low-quality reads from the eubacterial (16S) sequence data, a total of 105 656 reads were binned into 310 OTUs, which could subsequently be grouped into 104 genera. The 30 most common genera made up over 85 % of the dataset, with the remaining 74 genera occurring in less than 0.5 % relative abundance.

Figure 6 displays these most common genera, with the data grouped into site of origin and separated into “epiphytes” and “endophytes.” The most abundant genus common to the entire phyllosphere is *Sphingomonas*. Where epiphyte population structure is quite homogenous between the three sites—with the four principle genera being *Sphingomonas, Hymenobacter, Methylobacteria* and *Beijerinckiaceae*—there are marked differences between the endophytes: site A’s endophytes have a somewhat equal distribution of *Sphingomonas, Ralstonia,* Candidatus *Portiera, Methylobacterium,* and *Hymenobacter*. In contrast, the three genera *Sphingomonas, Ralstonia,* and

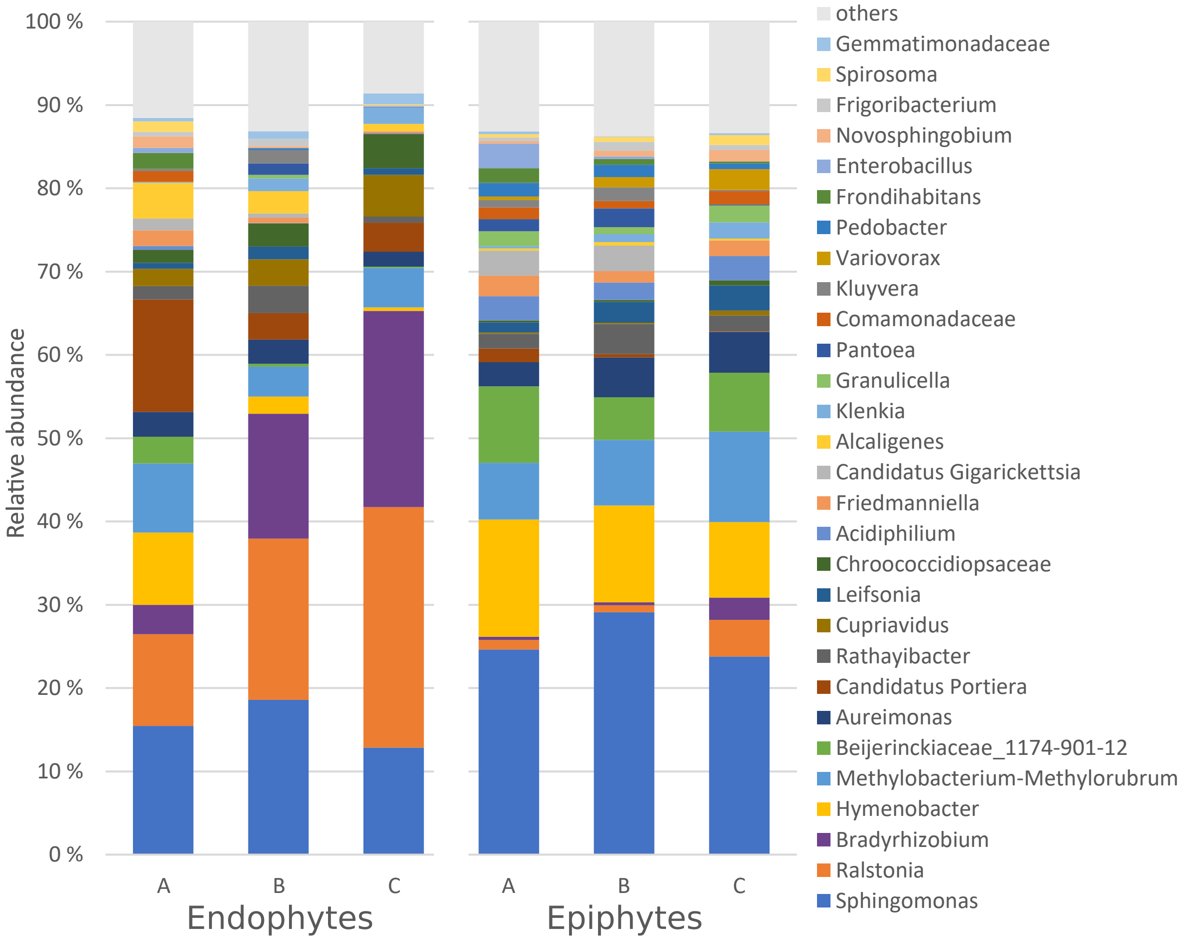

Figure 6: Relative abundance of the 30 most abundant genera of the sequenced lingonberry phyllosphere microbiome, grouped according to source site and niche (epiphyte/endophyte). Note that the after the class "others," the key is structured in ascending order of relative abundance in the total sequenced microbiome.

Bradyrhizobium alone account for more than 50 % and 65 % of the endophytes from sites B and C respectively.

Of the 128 490 *pufM* reads which remained after quality control, 654 OTUs could be formed; however 179 of these OTUs were discarded due to there being no corresponding matches in the database. The remaining 475 OTUs were then clustered based on reference sequences, forming a final total of 281 OTUs. Only 12 of these occurred in relative abundances greater than 0.5 %. Three of the endosphere samples (A1si, A2sii, and B1si) produced less than 100 reads, and so were discarded from the dataset. Figure 7 displays the 41 most common OTUs in each of the samples. Notable is the preponderance of genus *Methylobacterium* (indented in the fig-

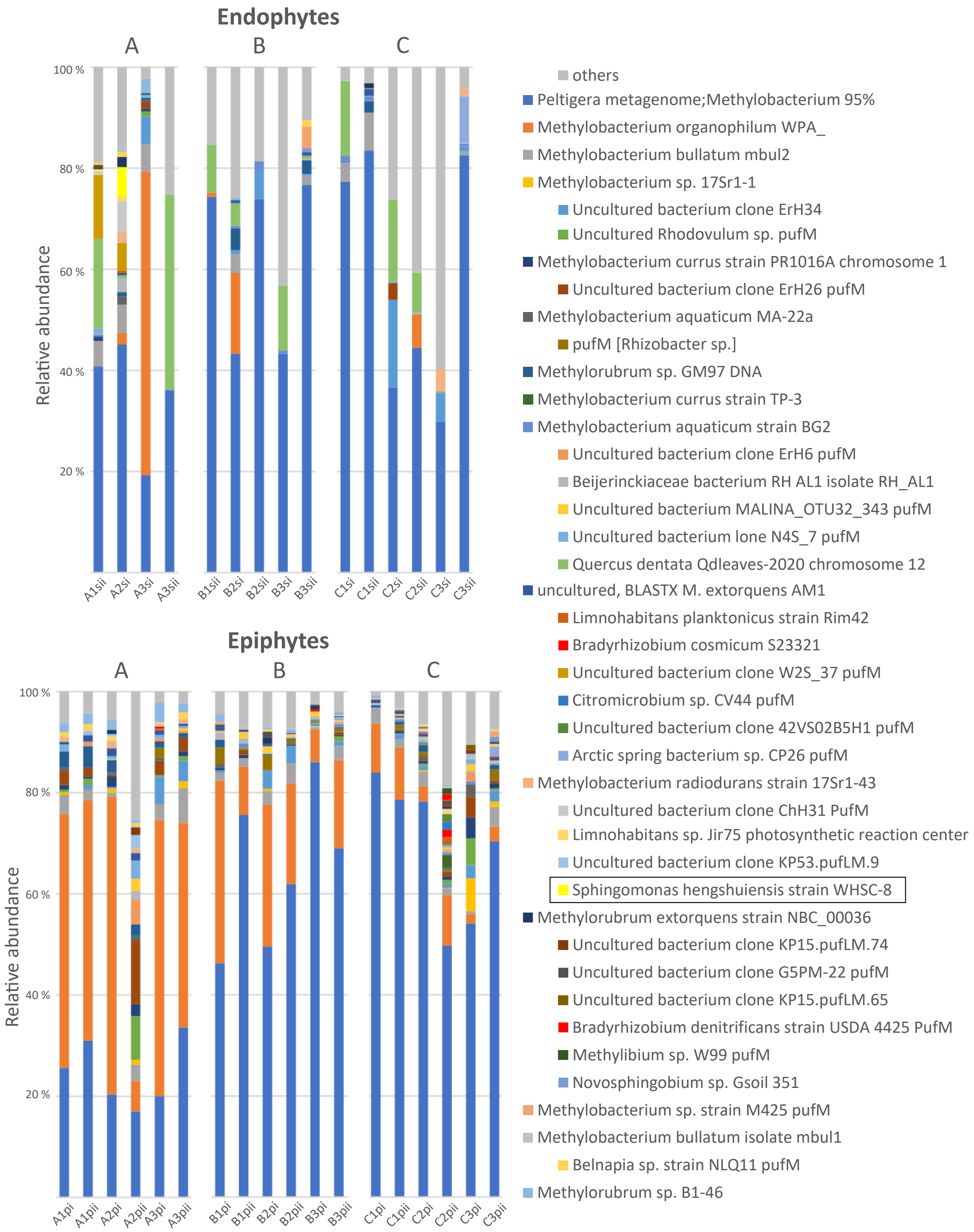

Figure 7: The relative abundances of AAPB in the lingonberry leaf phyllosphere as determined by presence of *pufM* gene. With the exception of the class "others," the key is structured in descending order of total relative abundance in the whole data set. *Methylobacterium* genera in the key are offset for visual clarity, and the single *Sphingomonas* species (bright yellow) found in sample A2si is outlined.

ure legend) which accounts for more than a third of the most common reference sequence OTUs, and at least 79 % of the OTUs. The *Methylobacterium* reference sequence from "*Peltigera* metagenome" was definitively the most abundant match in the entire *pufM* dataset at 64 % of all binned reads. The second species of note is *Methylobacterium organophilum* which has an evident niche preference for the phylloplane over the leaf endosphere. The epiphytes have a marked higher diversity than endophytes, and also had in general more reads. Lastly, despite the eubacterial data set showing that *Sphingomonas* was a notable genus in the whole phyllosphere (Figure 6), only a single AAP$^+$ strain was sequenced—*Sphingomonas hengshuiensis*—and furthermore in only one instance (endosphere sample A2si) though with a notable relative abundance of 6 %.

4 DISCUSSION

As methods for analysing plant-based AAPB have until this point been either metagenomic or culturomic, exploration of alternative methods is warrented. The amplicon based technique presented in this thesis nominally combines some of the virtues of metagenomics and culturomics, namely *en mass* batch processing and specificity. In some ways, amplicon-based methods are also easier; they are faster than the growth, detection, isolation, regrowth steps of culturing, and require considerably less computation, analytical power, and specialised equipment than metagenomics demands. Perhaps as a result of this, amplicon-based methods are ostensibly the most commonly used ones in modern AAPB research, thus its very absence in plant-based AAPB studies may be telling—when performing genetic work with environmental samples, primers which work "*in vitro*" may suddenly behave in unanticipated manners due to the considerable presence of non-target DNA and other factors which may interfere with PCR processes. In other words, amplicon-based studies of plant-associated AAPB may not exist simply due to the fact that inherent limitations in the techniques render such analyses either suboptimal or non-viable.

To further complicate the discussion, there remains a distinct possibility that the plants investigated in this study had a low abundance and/or diversity of AAPB, the extraction methods were too distruptive, the primers and their PCR protocols were not suitable, or some combination of these factors. In order to fully address these possibilities, a much more elaborate study would have had to have been devised involving a considerably larger sample size from many different locations, different host species, and simultaneous evaluation of different analysis techniques to "triangulate" a reliable and meaningful result. This is not to imply a defeatest attitude—instead, the technique and methods presented here are valuable as they indicate some of the challenges that may arise during development of such methods, and explore the ways that they may be resolved.

For example, by making a clear distinction between microbes of the surface and interior of leaves in this study, the methods resulted in quite a different image of epiphytic and endophytic AAPB's presence in lingonberry; though all evidence suggest that these two niches contain significantly different bacterial populations, the nature of the niches also mean that a common method to investigate them may either exaggerate or minimise these differences. Thus, in this analysis we must try to disentangle the possibilities of stochastic variations, choice of primers, the very nature of the samples, *etc.* to determine how closely the results conform to reality, and thus *ultimately evaluate how reliable the presented techniques were,* such that later research is more prepared to investigate the actual case with plant-associated AAPB. With this in mind, this thesis can be viewed as simply one instance of a broader question; there will doubtless exist other cases where an analogous set of plant-associated micro-organisms must be analysed in a similar fashion, and will encounter corresponding limitations and challenges. It is thus hoped that this thesis and the questions it raises have a role outside the world of AAPB too. Finally, the concurrent technique used in this study for querying the entire plant-associated eubacterial population is a good example of how an apparent impasse can be overcome; despite chloroplasts also possessing a 16S gene and thus dominating a leaf-tissue DNA sample, specific primers in a nested PCR allow for their exclusion in an analysis of bacterial endophytes — by describing and delineating problems, we are more capable of finding solutions to them.

4.1 Sampling

Variations in the concentration of DNA between all of the samples can be accounted for by both the nature of the samples and the methods used to extract DNA. Leaf surfaces will have had varying amounts of detritus on them, which after sonication will have been suspended in the extraction buffer; it is likely that varying amounts of debris will have been transferred to the homogenization vessels, resulting in such large variations. The homogenization of the leaf tissue resulted in a very viscous mass, and even after an extended lysis stage the mixture was turbid and uneven. This mass was so dense, that in previous runs the lysate could not flow through it to the prefilter during centrifugation, and as a result I had to resort to using only the supernatant in the extraction phase; this somewhat imprecise solution is likely the main cause for variation in DNA concentration of the leaf-tissue extracts. I had collected leaves in the morning in order to minimize the amount of starch in the leaves, but this did not appear to help. This may be due to the thick, waxy cuticle of the lingonberry leaf; future work may require an extended homogenizaion phase, however this could comprimise the integrity of the extracted DNA. It could also be that 100 mg of lingonberry leaves was too much, and a smaller amount would have resulted in a more efficacious extraction.

4.2 Primers

Using 1 ng of DNA from each AAPB strain in the screen resulted in variable response from all primers (Figure 1). It is notable that even with primers pufM_uniF-/pufM_uniR—which did not cause non-specific amplification—the quantity of amplicons varied between strains. Previous research attempting to quantify AAPB via genetic means did so on the assumption that each AAP bacterium contains a single *pufM* gene (Schwalbach and Fuhrman 2005). Chromosome size will of course vary considerably between genera. Thus, one must take into consideration that interspecies discrepencies between amplicon concentration—in spite of equal quantities of DNA being tested—do not necessarily signify a cognate primer's variance in efficacy, but may instead be caused by differences in target gene prevalence relative to chromosome size. However, even with universal primers, amplicon specificity may vary for a given taxon.

With the exception of Achenbach *et al.* (2001), I have never seen any published AGE images for any of the *pufM*-directed primers, nor have I read any comment on nonspecific amplification arising from them. This made evaluation of PCR difficult: for example, with primers pufM_unif/pufM_WAW in sample 7 (*Caballeronia*, Figure 1) we do observe one amplicon which is of the expected size, though it is accompanied by numerous other, larger fragments, the most intense of which is some 1300 nt long. Similar observations from these primer were also made with environmental samples, where one finds the target amplicon in all samples, even from leaf-tissue; in comparison to pufM_uniF/pufM_uniR, the total mass of amplicons is much greater. Indeed, for leaf-tissue DNA, which did not produce detectable target amplicons, the total amount of amplicons was comparitively low. *Sphingomonas* is a part of the core phyllosphere (Vorholt 2012) and its primers—SA429f/SA933r—also produced variable amounts of amplicons, sometimes quite low. This may indicate that the ratio of endophyte DNA in the leaf-tissue extracts is also low and/or variable. Considering—for example—that 1 ng of the *Lichenibacter* strain produced amplicons at the threshold of detection with pufM_uniF/pufM_uniR, it may be the case that the amount of AAP$^+$ endophytes in 30 ng is simply far too low for detection with AGE and pufM_uniF/pufM_uniR. Attempts to increase the mass of leaf-tissue DNA template up to 90 ng (Figure 3), however, did not ameliorate target amplification.

Though not quantifiable, there is a distinct pattern to the intensity of the amplicons in plant-tissue extracts independent of the primer used (Figure 2). As amount of template added was equivalent for each sample (30 ng) and each of the samples represent individual biological replicates, this could indicate different leaves contain varying abundances of bacteria. Considering the ratio of endophyte to host DNA in leaf-tissue extracts, and putatively tiny amounts of AAPB, one may assume that slight primer preference for non-AAPB DNA results in a case where exponential increase of non-specific amplicons "masks" target amplicons and DNA.

The rationale for proceeding with the universal primers instead of pufM_uniF/-pufM_WAW—despite the latter producing targets in all environmental samples—was a measured one; although it is possible that the target amplicons could be extracted with PippinPrep and erroneous results removed from the sequenced data through size-filtering alone, this seemed to me to be excess data manipulation. Furthermore, it seemed that in producing so many abberant amplicons in the screen of AAPB strains, the reliability of their results in sequencing would be questionable and not reflect the full diversity to be found in the endosphere. Last of all, these universal primers appeared to work optimally for leaf-surface extracts, whereas pufM_uniF/pufM_WAW caused large "streaks" instead of clearly defined amplicons. Gazulla *et al.* (2023) report also that this primer pair did not function with their samples and was not amenable to optimization. It is clear that pufLf/pufMr did not work well in comparison with the other primers tested. Replacing the reverse primer with pufM_uniR and pufM_WAW resulted in no detectable amplification at all (data not shown). Should any of these primer pairs have worked in either the screen or with environmental samples, it would have provided the opportunity for a nested PCR.

The length of the amplicon from primers pufM_uniF/pufM_uniR is frequently stated to be approximately 150 bp long. In this experiment, the target amplicons in AGE images are closer to 220 bp (Figure 1), and after library preparation were closer to 270 bp (Appendix D, Figure 9). The peak length of the amplicon as determined by TapeStation was approximately 300 bp. In retrospect, an AGE performed with a higher agarose w/v along with a DNA ladder of a smaller and finer size range would have resulted in a better analysis, with easier resolution of target fragments. Discrepancies in fragment sizes through the aforementioned AGE setup may describe some of the observed inaccuracies. Gazulla *et al.* (2023) makes clear that their stated 145 bp pufM_uniR/pufM_uniR amplicon length was the smallest amplicon that they found; in this study, size filtration was performed in the range of 220 to 320, meaning that for a minimum of 145 bp with an additional 40 nt in M13 and P1 labels (total 185 bp) there is a possibility that some amplicons were excluded prior to sequencing. Because the peak size was measured accurately to be 300 bp, these sequences would be at the tail end of the distribution curve, and indeed the distribution in sizes is not Gaussian instead tending toward the larger side of the peak; therefore if any reads were indeed disposed, then they are likely to be very few. This could be taken as an example to not trust excessively in such a subjective method such as AGE for analysis which should be accurate.

4.3 PCR-optimization

Various attempts at optimization of pufM_uniF/pufM_uniR did not result in any visible improvement in reaction efficiency (Section 3.3). Typically, increasing the annealing temperature reduces non-specific amplification, though in this instance it caused a complete cessation in amplification. Lowering the temperature instead reduced the number of larger amplicons, and resulted in a single, roughly 450 bp long amplicon (Figure 4) suggesting a order or preference for primer binding sites, with the latter being particularly potent. In hindsight, a better judge of protocol optimization would have been through more quantitative means, such as with TapeStation. In a recent comparison of *puf* primers, pufM_uniF/pufM_uniR were found to be somewhat inefficient in comparison to the study's newly designed primers and sequencing was apparently only "possible after a cleaning step", though were evaluated to represent taxonomic diversity well (Gazulla *et al.* 2023). What bearing the cleaning process has is unclear; size-fractionation with PippinPrep in this study constitutes a cleaning step, which is assumed to be a compulsory component of sequencing.

Interestingly, target amplicons were detected in a few samples in the prepared library—sample A3si worked almost optimally (Appendix D, Figure 9). Library preparation was performed in winter; the air was dry and caused electrostatic charges to build up on my clothes, pipette tips, and PCR plate. The structure and layout of the method at this stage precluded any possibility of contamination from epiphytes. This made pipetting difficult, and clearly with sample B1si it did not work at all. It is possible that a smaller volume was added for sample A3si, which may have effected the result somehow, though previous optimizations using template masses from 30 ng to 1 ng did not result in any improvement in reaction efficiency at all. Another reason may be due to the fact that in this two-stage PCR, the primers used had additional sequences on their 5′ end. The cumulative 43 PCR cycles may also have contributed, though earlier this did not result in observable improvements (Figure 5).

Ultimately there were sufficient amounts of target of *pufM* amplicons for sequencing. The emulsion phase PCR in IonTorrent PGM requires 400 ng of template; in essence, each individual P1-labelled amplicon in library is sequenced and read up to a certain molar threshhold. Accordingly, part of the workflow requires diluting the amplicon libraries to this necessary concentration, as was done with the *16S* library in this study. The *pufM* library was sufficiently concentrate for sequencing without dilution. A charitable meterological view would be that low amplification would be equivalent to the process of dilution, and that we could consider the methods described in this thesis to be sufficient for sequencing endophytic AAPB. The nature of PCR is such, however, that preferential amplification of non-specific targets causes their molar amount to increase exponentially causing consecutively smaller probabilities that the target will be amplified. As a result, the more abundant target species in the sample will be overrepresented in the sequence data, skewing the true diversity in their favour. Due to the sensitivity of IonTorrent, it is likely that the overwhelming majority of targets will have at least one read, but compounded with unavoidable technical limitations in the methods—such as incomplete amplification, failure for P1 or M13 labels to be added *etc.*—the reliability of the results can be thoroughly degraded.

4.4 Sources of non-specific amplicons

Unlike the other primers tested, the universal primers produced clean amplicons in both the screen and the leaf-surface extracts. The source of the universal primer's non-specific amplicons are therefore likely to be exogenous, *i.e.* from the plant host, whereas pufM_uniF/pufM_WAW may be a combination of endogenous and exogenous. TapeStation analysis of one of the *puf* library samples (A1si) prior to pooling showed that these non-specific amplicons from the universal primers with leaf-tissue DNA had peak lengths of 207, 438, 479, 722 and 1200 bp with the target being approximately 300 bp (Appendix C). It is highly likely that sequences in the lingonberry (or perhaps even plant) genome are cognate with the primer sequences, causing so much non-specific amplification. As there are very few available primer positions in the *pufM* gene, should there actually be some overlap with a plant's genome then amplicon-based research of endophytic AAPB will be very limited, if not virtually impossible.

The universal forward primer has a remarkable degree of degeneracy (pufM_uniF: 8192) a very low GC content (36.5 %) and is rather long. The melting temperatures for the universal primer pair are also low (54.2 °C to 56.9 °C), which may explain the difficulty with optimization to remove ostensibly exogenous amplification; a ready method to increase primer specificity is to increase the annealing temperature, but this did not appear to work in this case (Figure 4). A new forward primer (pufMF_Y) intended to be paired with pufM_WAW, seems to be promising and reveal as much—if not more—taxonomic diversity than the universal primers, and furthermore appears to be more efficient (Gazulla *et al.* 2023). It has superior properties compared to pufM_uniF (higher GC content, shorter, less degenerate nucleotides) but its melting temperature is even lower (47.5 °C). This primer was designed for marine environments, and its lower melting temperature implies that for investigation of endophytic AAPB it may not perform much better than the universal primers. In this context one must also note that the main differential response in amplicon profiles arose not from pufM_uniF, but from the effects of the reverse primers (pufM_uniR & pufM_WAW). Thus, a pufM_Y/pufM_WAW pairing may not be any more useful for investigating endophytic AAPB.

Previous tests using the primers presented here and 1 ng of DNA template from whole lingonberry leaves (*i.e.* no separation of epiphytes from epiphytes or surface sterilization) resulted in a amplicon profile similar to what was observed with the leaf-surface extracts in Figure 2, *viz.* pufM_uniF/pufM_uniR produced very few non-specific amplicons, pufM_uniF/pufM_WAW produced many, and pufLf/pufMr did not work at all [1] (data not shown). It seems certain that the source of the non-specific amplicons for pufM_uniF/pufM_uniR is derived from the host. It can be assumed therefore that the small amount of epiphytic AAPB DNA in a mixture composed overwhelmingly of host DNA was sufficient to "save" the PCR; the universal primers have a strong preference to *pufM* genes over those other sequences in the host DNA, indicating that there had been very little endophytic AAPB in the leaf-tissue DNA extracts. One could test this hypothesis by titrating amounts of DNA from leaf-surface or AAP$^+$ strains to a set quantity of leaf-tissue DNA template in PCR, then observing the amplification profile. If one observed a decrease in non-specific amplification after a certain amount of AAPB template DNA was added to the leaf-tissue DNA, it would show that there is indeed very little endophytic AAPB present in the sampled phyllosphere.

The endogenous non-specific amplicons with pufM_WAW which are longer than expected (Figure 1) may be due to its target being further downstream than anticipated, but to my knowledge the size of the *pufM* gene is quite conserved; using 197 *pufLM* sequences given as a supplementary table in Imhoff *et al.* (2018), the length of *pufM* genes can be calculated to be between 690 to 1038 bp. The forward primer pufM_uniF targets a site at the latter end of the gene (nominal position 639) with pufM_uniF at base 784 and pufM_WAW roughly 60 bp downstream at base 842—the tail end of the *pufM* gene (as reported relative to the shortest amplicon lengths in Gazulla *et al.* 2023). Therefore it is highly unlikely that there are variable intervening/intragenic regions 700 to 1200 bp long between the target sequences of

[1] This experiment was performed at a different time of year, from a different site, and using slightly different extraction methods, such as whole lysate transferred to prefilter, and homogenizer not cooled to 0 °C

pufM_uniF and pufM_WAW. However, if this is the case, then it certainly would be of considerable phylogenetic interest. Given that the *Caballeronia* strain in Figure 1 produces multiple amplicons, it is far more likely that there are cognate sequences far downstream than was intended. One should also note the alarming mismatch in melting temperatures between pufM_uniF and pufM_WAW, which are respectively 54.2 °C and 73 °C; a more pragmatic analysis of non-specific amplification could be that it is almost impossible to reconcile these two values in a PCR protocol, leading to abberant amplification.

Based on the name and sequence of pufM_WAW (5′-Ay nGC rAA CCA **CCA nGC CCA**-3′ → **WAW**WFA; where the 3′ sequence is cognate with amino acids tryptophan, alanine, tryptophan) it is likely that this domain is common in other genes, hence the amount of non-specific amplicons. This codon sequence is present in 94 of the 197 *pufM* sequences mentioned above, indicating that it is not a universal motif in AAPB *pufM* (however, a more thorough study with a considerably larger data set (approximately 1300 *vs.* 197 (Gazulla *et al.* 2023)) suggests otherwise). A gene reported to be commonly located downstream of *pufM* is *puf*C (Zheng *et al.* 2011)—it may be that this codon pattern is located there, with variable intragenic regions between the genes accounting for the variations in amplicon length.

Providing that pufM_uniF has a single target in the genome this would not cause problems for sequencing *per se*, as the intended primary 250 bp could still be usefully included in the sequence data. Yet, in this technique with IonTorrent, the P1-label is included in the reverse primer, meaning that if the PippinPrep size filtration step had not been employed, many reads from pufM_WAW would have included useless sequences. Sequencing size limitations of the NGS technology must also be considered. Frequently, in order to remove incomplete or erroneous reads in NGS, an inclusion criteria for a sequence library is that a datum contains both forward and reverse primer sequences. As we employed the M13 "barcoding" system with the forward primers, sequences which were longer than the limitation of the system would have either been excluded or untraceable to their origin. Ultimately, the stated target amplicon size should always be used to ensure rigor in results; size-exclusion methods to isolate target amplicons (such as PippinPrep in this instance) is one way to ensure of this and has the added benefit of cleaning up the PCR reaction to ensure good sequencing with NGS systems. Primers which are not amenable to this workflow are likely to cause results which do not reflect the true community structure or abundance in an environmental sample.

4.5 Community structure

As determined by *16S* sequencing, the abundance and diversity of the eubacterial populations appears to be somewhat homogenous for epiphytes, though there is substantial variation in endophyte population structure between the sites (Figure 6). Of note also is that no mitochondrial sequences were found in the eubacterial epiphyte data-set, indicating that the epiphyte/endophyte isolation phase succeeded.

For many reasons, the utility of the eubacterial population data is not solely to corroborate any findings with the results of the AAPB population, but is instead supposed to serve as a mediating factor when evaluating the techniques used to detect plant-associated AAPB. The eubacterial portion of the workflow functions here as a standard whose primers and their corresponding protocols have been designed specifically for a task such as this; the AAPB workflow instead emulates the eubacterial portion. This "emulation" functionally ends at the NGS portion, which is due to the nature of the sequence analysis: 16S rDNA sequences have long been used as a genetic standard in taxonomic identification and correspondingly there is a richer data set that the read sequences could be referred to. As such, the methods for structuring the NGS data into OTUs differs between eubacteria and AAPB, where the latter was performed *de novo* and the former through the SILVA database. As the performance of the sequencing worked nominally well with the *16S* primers and each sample had a good number of reads with an even species accumulation curve, the eubacterial data could be represented confidently in terms of genera and sites. This is in contrast with AAPB: three endophyte samples (A1si, A2sii, B1si) resulted in so few functional reads that they were discarded completely, meaning that grouping results into sites would have been misrepresentative.

According to our results *Methylobacteria* are the most dominant in the AAPB dataset, whilst simultaenously comprising less than 10 % of the eubacterial population. Metagenomic reseach into plant-associated AAPB have found that *Methylobacteria* are found in multiple phyllospheres and can account for more than a third of all genera detected (Atamna-Ismaeel *et al.* 2012b), and similarly high amounts were found by Nissinen *et al.* (2023). Therefore, the preponderance of this genera in the data may be a reliable result. However, the latter study also revealed an equal—if not greater—abundance of plant-associated AAP$^+$ *Sphingomonas.*

This leads us to one particular observation which may have major significance in the analysis: curiously, the most dominant genus in the eubacterial population—*Sphingomonas*—which has both been previously demonstrated to be a common plant-associated AAPB and one which AAPB-primers respond to well (Section 3.2) was almost completely unrepresented in the AAPB data set (Figure 7), with only a single species being detected in only one sample. This discord between observed primer-specificity, expectations based on previous studies, and final results may be accounted for in three ways;

1. There are little-to-no AAP$^+$ *Sphingomonas* in the sampled *V. vitis-idaea*.
2. AAP$^+$ *Sphingomonas* are not well represented in the reference sequence database.
3. The primers and/or methods used here were not suitable for representing the true diversity of plant-associated AAPB.

As mentioned above, the structure of this experiment is unable to unequivocally address point 1, though reframing the idea as a disjunctive syllogism can demonstrate that it is a possibility:

Sphingomonas may or may not be AAPB — most AAPB are not *Sphingomonas* — therefore a set of AAPB may not contain *Sphingomonas* (and similarly, a set of *Sphingomonas* may not contain AAPB).

However, the second proposition above may be misleading in its generality, because Nissinen *et al.* (2023) have shown that many *plant-associated* AAPB *are Sphingomonas*. So we may conclude that point 1, though entirely possible, is very unlikely.

This allows us to consider the second point; that a probable result has somehow "fallen through the net," *i.e.* they are either discarded during reference-sequence–OTU creation or are "disguised" in the dataset through being categorized as "other" or "uncultured". One could address this through a phylogenetic analysis, though yet again one is confronted with the limitation of sequencing *pufM* fragments, which are so small (in this instance, 120 to 220 nt long) that such a method initially seem troublesome. However, in this context it may be sufficient to do so at a low level, such as binning the *pufM* reads into orders and observing the amount of *Sphingomonodales*.

In these circumstances, we must err on the side of caution and explore how point 3—that is, technical errors—may have impacted the results. It first must be noted that all major problems in the workflow were associated with the endophytic AAPB, and that both the materials/methods, and majority of results for epiphytic AAPB were analogous to the eubacterial portion. The fact that three of the 18 endophytic *pufM* samples had to be removed in the final AAPB dataset due to low quality reflects this. Due to this, the NGS results for endophytic AAPB must be approached with some degree of scepticism; though the endophytic AAPB dataset is not likely to contain false-positives, it is clear that the techniques used here may not represent the full diversity of endophytic AAPB.

Why is it so, that despite all appearances of working well in earlier stages, many epiphytic samples in the final data have similarly low read numbers and are also missing an expected, core component of their population? The clear differences between the epiphytic and endophytic AAPB data indicate that on some level the techniques are capable of differentiating between different population structures.

It thus seems that there is sufficient *specificity* in the technique to generate a difference in epiphytic and endophytic AAPB populations, i.e. we can be confident in the lack of false positives. However, a noteable portion of the data is either not detected or is functionally discarded (due to primer–sequence mismatch, overabundance of DNA to which primers have (partial) affinity, low read numbers, a portion of the most common reads lacking a meaningful taxonomy in the reference database). More precisely, the lack of AAP$^+$ *Sphingomonas* can be presumed to be a *false negative*, and one so significant that in conjunction with other, lesser difficulties may indicate a general lack of *sensitivity*. Where precisely in the workflow this presumed lack of sensitivity arises is not evident, however, a mechanism whereby low target amplification can cause results skewed toward the most abundant taxa (section 4.3), and also how the *pufM* gene and its associated primers require further investigation in this field (section 4.4).

However, the *pufM* technique produced equivalent numbers of total reads and OTUs in comparison to the *16S* workflow (128 490 reads & 281 OTUs *vs.* 105 656 reads & 310 OTUs). Though concern regarding the probable lack of sensitivity is warrented, another interpretation of the results could be that it *does* provide a good representation of plant-associated AAPB—that we can conclude with some degree of confidence that AAP$^+$ phyllophytes exist in low abundance but moderate diversity, though the vast majority are represented by a very few genera. This account would still have to be tempered by all the other valid objections and criticisms mentioned above. Despite the shortcomings and difficulties encountered this would be the most fruitful outcome, as it would serve as a good foundation for future work; for example, does a similar pattern of abundance and diversity still manifest when optimisations result in greater and/or more evenly distributed reads among the samples?

4.6 Potential improvements and future work

As PCR is neither expensive nor particularly time-consuming, a more comprehensive test of primers with a larger screen of AAP$^+$ strains is warrented. As shown in Appendix A, there are many more potential primer pairs that can be tested. We have in our department the ability to perform a single AGE with 20 samples including DNA ladder, negative control, and no-template control. Should a screen of diverse AAPB—such as representatives of each of the 12 phyllogroups or as in this instance, cultured phyllophytes—be assembled, one could observe the response and amplicon profile for even more primer pairs used in AAPB research, including the recently developed pufMF_Y (Gazulla *et al.* 2023). Though sensitive techniques such as NGS may detect even small amounts of amplicons and produce noteworthy results, it is essential that primers function well and that readers of publications understand potential limitations of the materials and techniques presented therein. The above proposal would be a simple yet effective way of demonstrating the efficacy and behaviour of AAPB-specific primers, and allow future groups to make informed choices in their methodology and potentially prevent confusion such as I experienced in working with these primers for the first time. Some primers such as pufM_WAW produce endogenous non-specific amplicons; isolation and sequencing of these may shed some light on the genetic structure of certain AAPB and the appropriateness of certain primers for target groups, as it may be the case that the cause of sequence failures, low count numbers, or biases for certain primers is not to be accounted for by target sequence variability alone, but also due to operonic structure.

4.6.1 qPCR

The general specificity of the universal primers and their lower amounts of non-specific amplicons indicate of the primers tested here, this pair is the most suitable for qPCR. This would be applicable for studying quantities of epiphytic AAPB, but the presumably irremediable problems with leaf-tissue DNA preclude any such experiments with endophytic AAPB.

4.7 Summary

The use of amplicon-based methods for detection of plant-associated AAPB is a valid and useful technique. The primer pair used in this instance (pufM_uniF/-pufM_uniR) appear to be the most specific, and result in less non-specific amplification than the others tested. The converse is that it may result in *lower* target amplification, and this is most pronounced when attempting to detect endophytic AAPB. A second primer pair which reliably produces target amplicons (pufM_uniF/pufM_-WAW) in environmental samples, however, displayed lower specificity in the screen, and caused considerable amount of non-specific amplification. These non-specific amplicons may be both exogenous and endogenous, in contrast to pufM_uniF/-pufM_uniR, whose non-specific amplicons were exclusively endogenous and comparatively easy to differentiate from the target. None of the techniques used to optimize the PCR reaction for detection of endophytic AAPB resulted in noteable improvements in target amplification. Any future attempts at optimization should, however, rely on more objective measurements, as it remains possible that endophytic AAPB exist in low abundances, and thus even marginal improvements may impact the quality of downstream analyses.

When working with endophytic bacteria, techniques for isolating the target amplicon must be employed prior to batch sequencing. Size-fractionation was used in this instance, which also served as a PCR-cleanup stage, thus it dovetails well with the workflow and should not be considered a hindrance. Detection of endophytes though amplicon-based methods is a technical challenge irrespective of the studied taxa. However, AAPB present a particular challenge, as there are few genetic markers which can be employed to distinguish them. The *pufM* gene contains only three regions which can be used as a primer target, one of which is only moderately conserved, and as a result the researcher must consider a tradeoff; greater diversity can be detected with the highly degenerate universal primers (pufM_uniF/pufM_uniR) but at the price of considerable non-specific amplification with endophytes.

In this thesis, the use of these degenerate primers along with their recommended

PCR protocols resulted in sequence data suggesting that AAPB in the lingonberry phyllosphere are dominated by a small number of taxa (principally *Methylobacteria*), though the total number of OTUs generated was similar to that of the eubacterial dataset. Many samples, particularly those derived from the leaf interior, contained very few reads and had to be discarded. Furthermore, a key plant-associated genus, *Sphingomonas*, was virtually absent from the AAPB dataset, despite it being the most dominant in the eubacterial sequence data. These lacunae may be interpreted as either *true negatives* (*i.e.* AAPB are not at all abundant in the lingonberry leaf; none of the detected *Sphingomonas* were AAPB) or *false negatives* (*i.e.* the primers could not detect *Sphingomonas* and/or *Sphingomonas*–derived *pufM* amplicon reads could not be definitively matched to a taxon; the methods failed to generate a representative amount of reads from the actual AAPB population). Establishing the validity of this result is not possible with an experiment of this scale. This observation is so significant, however, that future studies should—at least partially—structure their experiments to prove it.

The results indicate that the techniques described here are provisionally valid, and indicate some paths that similar, future experiments could explore. One such experiment could observe the amplicon profile from common AAPB-primers with leaf-tissue DNA extract which has been titrated with AAPB DNA; a concomitant decrease in non-specific amplification would be a very strong signal that endophytic AAPB are found in low abundance. Likewise, a much larger screen of plant-associated AAPB with more AAPB-primers could indicate potential biases toward certain orders/genera and suggest how these biases may shape NGS data/analyses.

ACKNOWLEDGEMENTS

My supervisors Riitta Nissinen and Janne Ihalainen provided excellent feedback and guidance on writing this thesis. This thesis was part of the Jaettu Valo ("Shared Light") project. The entire Jaettu Valo team also gave good constructive criticism at

many points during the experimental phase, which helped me think more critically about the methods I used and the results that were gained from them. Ole Franz supplied me with the LaTeX template which was a significant help for me during the writing phase. Essentially all sequencing work was performed by Elina Virtanen; she was also a very supportive in assisting with the latter phases of the DNA work, *i.a.* using TapeStation and how to extract/analyse its data.

Finally, Riitta Nissinen performed the analysis of the sequenced data, and furthermore helped in all phases of the experimental work, giving invaluable direction and support from the beginning to end.

I acknowledge and give my warmest thanks to all mentioned here.

REFERENCES

Achenbach L.A., Carey J., & Madigan M.T. 2001. Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Appl. Environ. Mircobiol.*, 67: 2922–2926.

Atamna-Ismaeel N., Finkel O.M., Glaser F., Sharon I., Schneider R., Post A.F., Spudich J.L., von Mering C., Vorholt J.A., Iluz D., Béjà O., & Belkin S. 2012a. Microbial rhodopsins on leaf surfaces of terrestrial plants. *Environ. Microbiol.*, 14: (1), 140–146.

Atamna-Ismaeel N., Finkel O.M., Glaser F., von Mering C., Vorholt J.A., Koblížek M., Belkin S., & Béjà O. 2012b. Bacterial anoxygenic photosynthesis on plant leaf surfaces. *Environ. Mircrobiol. Rep.*, 14: (2), 209–216.

Bauer C.E., Buggy J.J., Yang Z.M., & Marrs B.L. 1991. The superoperonal organization of genes for pigment biosynthesis and reaction center proteins is a conserved feature in *Rhodobacter capsulatus* — analysis of overlapping *bchB* and *puhA* transcripts. *Mol. Gen. Genet.*, 228: 433–444.

Bauer C.E., & Marrs B.L. 1988. *Rhodobacter capsulatus puf* operon encodes a regulatory protein (PufQ) for bacteriochlorophyll biosynthesis. *Proc. Natl. Aca. Sci. USA*, 85: (19), 7074–7078.

Bauer C.E., Young D., & Marrs B. 1988. Analysis of the *Rhodobacter capsulatus puf* operon. *J. Biol. Chem.*, 263: (10), 4820–4827.

Beatty G.A., & Lindow S.E. 1995. The secret life of foliar bacterial pathogens on leaves. *Annu Rev Phytopathol*, 33: 145–172.

Béjà O., Suzuki M.T., Nelson W.C., Preston C.M., Hamada T., Eisen J.A., Fraser C.M., & DeLong E.F. 2002. Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature*, 415: 630–633.

Bringel F., & Couée I. 2015. Pivotal roles of phyllosphere microorganisms at the interface between plant functioning and atmospheric trace gas dynamics. *Front. Microbiol.*, 6: (486).

Chelius M.K., & Triplett E.W. 2001. The diversity of archaea and bacteria in association with the roots of *Zea mays* L. *Microb. Ecol.*, 41: 252–263.

Cottrell M.T., Ras J., & Kirchmann D.L. 2010. Bacteriochlorophyll and community structure of aerobic anoxygenic phototrophic bacteria in a particle-rich estuary. *The ISMA journal*, 4: 945–954.

Csotonyi J.T., Swiderski J., Stackebrandt E., & Yurkov V. 2010. A new environment for aerobic anoxygenic phototrophic bacteria: biological soil crusts. *Environ. Microiol.*, 2: 651–656.

Fecskeová L.K., Piwosz K., Hanusová M., Nedoma J., & Koblížek M. 2019. Diel changes and diversity of *pufM* expression in freshwater communities of anyoxygenic phototrophic bacteria. *Sci. Rep.*, 9: (18766).

Ferrara I., Sarmento H., Priscu J.C., Chiuchiolo A., González J.M., & Grossart H.-P. 2017. Diversity and distribution of freshwater aerobic anoxygenic phototrophic bacteria across a wide latitudinal gradient. *Front. Microbiol.*, 8: (175).

Ferrera I., Gasol J.M., Sebastián M., Hojerová E., & Koblízek M. 2011. Comparison of growth rates of aerobic anoxygenic pototrophic bacteria and other bacterioplankton groups in coastal Mediterranean waters. *Appl. Environ. Microbiol.*, 77: (21), 7451–7458.

Florez-Núñez V.M., Fonseca-García C., Desgarennes D., Eloe-Fadrosh E., Woyke T., & Partida-Martínez L.P. 2020. Functional signatures of the epiphytic prokaryotic microbiome of agaves and cacti. *Front. Microbiol.*, 10: (3044).

Franz O., Häkkänen H., Kovanen S., Heikkilä-Huhta K., Nissinen R., & Ihalainen J. 2023. NIRis: A low-cost, versatile imaging system for NIR fluorescence detection of phototrophic cell colonies used in science and education. *bioRxiv*.

Gazulla C.R., Cabella A.M., Sánchez P., Gasol J.P., Sánchez O., & Ferrera I. 2023. A metagenomic and amplicon sequencing based combined approach reveals the best primers to study marine aerobic anoxygenic phototrophs. *Microb. ecol.*, 86: 2161–2172.

Ghyselinck J., Pfieffer S., Heylen K., Sessitsch A., De Vos P., & Larsen P. 2013. The effect of primer choice and short read sequences on the outcome of 16s rRNA gene based diversity studies. *PLOS ONE*, 8: (8), e71360.

Given C., Häikiö E., Kumar M., & Nissinen R. 2020. Tissue-specific dynamics in the endophytic bacterial communities in Arctic pioneer plant *Oxyria digyna*. *Frontiers in Plant Sci.*, 11: (561).

Hamilton T.L. 2019. The trouble with oxygen: The ecophysiology of extant phototrophs and implications for the evolution of oxygenic photosynthetisis. *Free Radical Biology and Medicine*, 140: 233–249.

Hirano S.S., & Upper C.D. 1991. Bacterial community dynamics. In J. Andrews & S. Hirano (Eds.), *Microbial ecology of leaves* (pp. 271–294). Springer New York.

Igarashi N., Harada J., Nagashima S., Matsuura K., Shimada K., & Nagashima K. 2001. Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J Mol Evol.*, 52: 333–341.

Imhoff J.F., Rahn T., Künzel S., & Neuinger S.C. 2018. Photosynthesis is widely distributed among proteobacteria as demonstrated by the phylogeny of *puflm* reaction center proteins. *Front. Microbiol.*, 8: (2679).

Imhoff J.F., Rahn T., Künzel S., & Neulinger S.C. 2019. Phylogeny of anoxygenic photosynthesis based on sequences of photosynthetic reaction center proteins and a key enzyme in bacteriochlrophyll biosynthesis, the chlorophyllide reductase. *Microorganisms*, 7: (576).

Jiao N., Zhang Y., Zeng Y., Hong N., Liu R., Chen F., & Wang P. 2007. Distinct distribution pattern of abundance and diversity of aerobic anoxygenic phototrophic bacteria in the global ocean. *Environ. Microbiol.*, 9: (12), 3091–3099.

Kinkel L.L. 1997. Microbial population dynamics on leaves. *Annu. Rev. Phytopathol.*, 35: (1), 327–347.

Knief C., Ramette A., Frances K., Alonso-Blanco C., & Vorholt J.A. 2010. Site and plant species are important determinants of the Methyllobacterium community composition in the plant phyllosphere. *ISME J.*, 4: 719–728.

Koblížek M. 2015. Ecology of aerobic anoxygenic phototrophs in aquatic environments. *FEMS Microbiol. Re.*, 39: (6), 854–870.

Koblížek M., Béjà O., Bidigare R.R., Christensen S., Benitez-Nelson B., Vetriani C., Kolber M.K., Falkowski P.G., & Kolber Z.S. 2003. Isolation and characterization of *Erythrobacter* sp strains from the upper ocean. *Arch. Microbiol.*, 180: (8), 327–338.

Koblížek M., Mousilová V., Muroňová M., & Oberník M. 2014. Horizontal transers of two types of *puf* operons among phototrophic members of the Roserobacter clade. *Folia Microbiol.*, 60: 37–65.

Koblížek M., Shih J.D., Breitbart S.I., Ratcliffe E.C., Kolber Z.S., Hunter C.N., & Niederman R.A. 2005. Sequential assembly of photosynthetic units in *Rhodobacter sphaeroides* as revealed by fast repetition rate analysis of variable bacteriochlorophyll *a* fluorescence. *Biochem Biophys Acta*, 1706: (3), 220–231.

Koblížek M., Zeng Y., Horák A., & Oborník M. 2013. Chapter thirteen - regressive evolution of photosynthesis in the roseobacter clade. In J.T. Beatty (Ed.), *Genome evolution of photosynthetic bacteria* (pp. 385–405). Academic Press.

Koh E.Y., Phua W., & Ryan K.G. 2011. Aerobic anoxygenic phototropic bacteria in Antarctic sea ice and seawater. *Environ. Microbiol. Rep.*, 3: 710–716.

Koskella B. 2020. The phyllosphere. *Curr. Biol.*, 30: (9), 1143–1146.

Kumar M., Brader G., Sessitsch A., Mäki A., van Elsas J.D., & Nissinen R. 2017. Plants assemble species specific bacterial communities from common core taxa in three Arcto-Alpine climate zones. *Front. Microbiol.*, 8: (12).

Leveau J.H.J., & Lindow S.E. 2001. Appetite of an epiphyte: Quantitative monitoring of bacterial sugar consumption in the phyllosphere. *Proc Natl Acad Sci USA*, 98: 3446–3453.

Leveau J. 2019. A brief from the leaf: Latest research to inform our understanding of the phyllosphere microbiome. *Curr. Opin. Microbiol.*, 49: 41–49.

Lindow S.E., & Brandl M.T. 2003. Microbiology of the phyllosphere. *Applied and environmental microbiology*, 69: (4), 1875–1883.

Liotenberg S., Steunou A., Picaud M., Reiss-Husson F., Astier C., & Ouchane S. 2008. Organization and expression of photosynthesis genes and operons in anoxgenic photosynthetic proteobacteria. *Environ. Microbiol.*, 10: (9), 2267–2276.

Liu H., Brettell L.E., & Singh B. 2020. Linking the phyllosphere microbiome to plant health. *Trends Plant Sci.*, 25: (9), 841–844.

Mäki A., Rissanen J.A., & Tiirola M. 2016. A practical method for barcoding and size-trimming PCR templates for amplicon sequencing. *BioTechniques*, 60: 88–90.

Mao D.P., Zhou Q., Chen C.Y., & Quan Z.X. 2012. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC microbiol*, 12: (66).

Müller T., & Ruppel S. 2013. Progress in cultivation-independent phyllosphere microbiology. *FEMS Microbiol. Ecol.*, 87: 2–17.

Nagashima K., Hiraishi A., Shimada K., & Matsuura K. 1997. Horizontal transfer of genes coding for the photosynthetic reaction centers of purple bacteria. *J. Mol. Evol.*, 43: 131–136.

Newton A.C., Gravouil C., & Fountaine J.M. 2010. Managing the ecology of foliar pathogens: Ecological tolerance in crops. *Annals of Appl Biol.*, 157: (3), 343–359.

Nissinen R., Franz O., Kovanen S., Mäkelä M., Kraft V., Ketola K., Liukkonen A., Heikkilä-Huhta K., Häkkänen H., & Ihalainen J.A. 2023. Aerobic anoxygenic phototrophic bacteria are ubiquitous in phyllo- and endosphere microbiomes of boreal and subarctic plants. *bioRxiv*.

Oz A., Sabehi G., Koblížek M., Massana R., & Béjà O. 2005. *Roseobacter*-like bacteria in Red and Mediterranean Sea aerobic anoxygenic photosynthetic populations. *Appl. Environ. Microbiol.*, 71: (1), 344–353.

Rastogi G., Coaker G.L., & Leveau J. 2013. New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches. *FEMS microbiol. lett.*, 348: 1–10.

Rathgeber C., Alric J., Hughes E., Vermélgio A., & Yurkov V. 2012. The photosynthetic apparatus and photoinduced electron transfer in the aerobic phototrophic bacteria *Roseicyclus mahoneyensis* and *Porphyrobacter meromictius*. *Photosynth. res.*, 110: 193–203.

Rathgeber C., Beatty J.T., & Yurkov V. 2004. Aerobic phototrophic bacteria: New evidence for the diversity, ecological importance and applied potential of this previously overlooked group. *Photosynth. res.*, 81: 113–128.

Remis-Emsermann M., & Schlecther R.O. 2018. Phyllosphere microbiology: At the interface between microbial individuals and the plant host. *New Phytologist*, 218: 1327–1333.

Schwalbach M.S., & Fuhrman J.A. 2005. Wide-ranging abundancies of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. *Limnol. Oceanogr.*, 50: (2), 620–628.

Selyanin V., Hauruseu D., & Koblížek M. 2016. The variability of light-harvesting complexes in aerobic anoxygenic phototrophs. *Photosynth. Res.*, 128: 35–43.

Shiba T., Simidu U., & Taga N. 1979. Distribution of aerobic bacteria which contain bacteriochlorohyll *a*. *Appl. Environ. Microbiol.*, 38: 43–45.

Steifel P., Zambelli T., & Vorholt J.A. 2013. Isolation of optically targeted single bacteria by application of fluidic force microscopy to aerobic anoxygenic phototrophs from the phyllosphere. *Appl. Environ. Microbiol.*, 79: (16), 4895–5905.

Tang K., Jia L., Yuan B., Yang S., Li H., Meng J., Zeng Y., & Feng F. 2018. Aerobic anoxygenic phototrophic bacteria promote the development of biological soil crusts. *Front. Microbiol.*, 9: (2715).

Tank M., Thiel V., & Imhoff J.F. 2009. Phylogenetic relationship of phototrophic purple sulfur bacteria according to *pufl* and *pufm* genes. *Int. Microbiol.*, 12: (3), 175–185.

Vorholt J. 2012. Microbial life in the phyllosphere. *Nat. Rev. Microbiol.*, 10: (12), 828–840.

Waidner L.A., & Kirchman D.L. 2005. Aerobic anoxygenic photosynthesis genes and operons in uncultured bacteria in the Delaware river. *Env. Microbio.*, 7: (12), 1896–1908.

Waidner L.A., & Kirchman D.L. 2007. Aerobic anoxygenic phototrophic bacteria attached to particles in turbid waters of the Delaware and Chesapeake estuaries. *Appl. Environ. Microbiol.*, 12: (3), 3936–3944.

Waidner L.A., & Kirchman D.L. 2008. Diversity and distribution of exoptypes of the aerobic anoxygenic phototrophy gene *pufM* in the Delaware estuary. *Appl. Environ. Microbiol.*, 74: (13), 4012–4021.

Yang H., & Hu C. 2022. Soil chemistry and nutrients influence the distribution of aerobic anoxygenic phototrophic bacteria and eukaryotic phototrophic microorganisms of physical soil crusts at different elevations on the Tibetan plateau. *Microb. Ecol.*, 83: 100–113.

Young D., Bauer C.E., Williams J.C., & Marrs B.L. 1989. Genetic evidence for superoperonal organization of genes for photosynthetic pigments and pigment-binding proteins in *Rhodeobacter capsulatus*. *Mol. Gen. Genet.*, 218: (1), 1–12.

Yurkov V., & Beatty J.T. 1998. Aerobic anoxygenic phototrophic bacteria. *Microbiol. Mol. Biol.*, 62: (3), 695–724.

Yurkov V., & Csotonyi J.T. 2009. New light on aerobic anoxygenic phototrophs. In C. Hunter, F. Daldal, M. Thurnauer & J. Beatty (Eds.), *The purple phototrophic bacteria* (pp. 31–55). Springer Netherlands.

Yurkov V., & Hughes E. 2013. Chapter eleven—genes associated with the peculiar phenotypes of the aerobic anoxygenic phototrophs. In J. Beatty (Ed.), *Genome evolution of photosynthetic bacteria* (pp. 327–358). Academic press.

Yurkov V., & Hughes E. 2017. Aerobic anoxygenic phototrophs: Four decades of mystery. In P. Hallenbeck (Ed.), *Modern topics in the phototrophic prokaryotes: Environmental and applied aspects* (pp. 193–214). Springer international publishing.

Yutin N., Suzuki M.T., & Béjà O. 2005. Novel primers reveal wider diversity among marine aerobic anoxygenic phototrophs. *Appl. Environ. Microbiol.*, 71: (12), 8958–8962.

Yutin N., Suzuki M.T., Rosenberg M., Rotem D., Madigan M.T., Süling J., Imhoff J.F., & Béjà O. 2009. *bchY*-based degenerate primers target all types of anoxygenic photosynthetic bacteria in a single PCR. *Appl. Environ. Microbiol.*, 75: (23), 7556–7559.

Yutin N., Suzuki M.T., Teeling H., Weber M., Venter J.C., Rusch D.B., & Béjà O. 2007. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling metagenomes. *Environ. Microbiol.*, 9: 1564–1475.

Zervas A., Zeng Y., Madsem A.M., & Hansen L.H. 2019. Genomics of aerobic photoheterotrophs in wheat phyllosphere reveals divergent evolutionary patterns of photosynthetic genes in *methyllobacterium* spp. *Genome Biol. Evol.*, 11: (10), 2895–2908.

Zheng D., Alm E.W., Stahl D.A., & Raskin L. 1996. Characterization of universal small-subunit rRNA hybridization probes for quantitative molecular microbial ecology studies. *Appl. Environ. Microbiol.*, 62: (12), 4504–4513.

Zheng Q., Zhang R., Koblížek M., Bolderave E.N., Yurkov V., Yan S., & Jiao N. 2011. Diverse arrangement of photosynthetic gene clusters in aerobic anoxygenic phototrophic bacteria. *PLoS ONE*, 6: (9).

Zhou L., Li H., Zhang Y., Wang Y., Han S., & Xu H. 2012. Abundance and diversity of *Sphingomonas* in Shenfu petroleum-wastewater irrigation zone, China. *Environ. Sci. Pollut. Res.*, 19: 282–294.

APPENDIX A. Primer alignment

Table 4: Primers frequently used in AAPB research. All sequences given in boldface are intended to represent alignment. Sequences are presented according to target region of the genes, divided by horizontal rules.

Name	**Reference**	**Sequence** (5′–3′)
pufL	Nagashima *et al.* 1997	**CTsTTCGACTTCTGGGsGG**
pufL67F	Tank *et al.* 2009	**TTCGACTTyTGGrTnGGnCC**
pufM.557	Achenbach *et al.* 2001	CGCACCTGGACTGGAC
pufM570f	Oz *et al.* 2005	CAGTTACTTTATTTTTCACAAC
pufMf	Béjà *et al.* 2002	**TACGGsAACCTGTwCTAC**
forward	Schwalbach and Fuhrman 2005	**TATAAyCCATTTCAyGC**
pufM_uniF	Yutin *et al.* 2009	**GGnAAyyTnTwyTAyAAyCCnTTyCA**
pufMF_Y	Gazulla *et al.* 2023	**GGsAAyCTsTwyTAyAAyC**
bchY_F	Yutin *et al.* 2009	CCnCArACnATGTGyCCnGCnTTyGG
pufM	Nagashima *et al.* 1997	**CCCATsGTCCAGCGCCAGAA**
pufM.750	Achenbach *et al.* 2001	**CCCATGGTCCAGCGCCAGAA**
pufMr	Béjà *et al.* 2002	**CCATsGTCCAGCGCCAGAA**
pufM_uniR	Yutin *et al.* 2009	**yCCATnGTCCAnCkCCArAA**
pufM781R	Tank *et al.* 2009	**CCAksGTCCAFCFCCAFAAnA**
pufM_WAW	Yutin *et al.* 2009	**AynGCrAACCACCAnGCCCA**
reverse	Schwalbach and Fuhrman 2005	**GCrAACCACCAAGCCCA**
bchY_R	Yutin *et al.* 2009	GGrtCnrCnGGrAAnATyTCnCC

APPENDIX B. DNA concentrations

Table 5: Concentrations of the extracted DNA samples, alongside the volumes used in the library preparation stage. Target DNA mass was was 30 ng from leaf-tissue samples and 1 ng from leaf-surface samples.

Leaf-tissue	**Conc.**	**Vol.**	**Mass**	**Leaf-surface**	**Conc.**	**Vol.**	**Mass**
Sample	ng µL^{-1}	µL	ng	Sample	ng µL^{-1}	µL	ng
A1si	55.0	0.5	27.5	A1pi	0.996	1.0	0.996
A1sii	53.3	0.5	26.65	A1pii	0.342	3.0	1.026
A2si	57.3	0.5	28.65	A2pi	4.98	0.5	2.49
A2sii	68.8	0.5	34.4	A2pii	0.389	3.0	1.167
A3si	26.3	1.0	26.3	A3pi	1.41	1.0	1.41
A3sii	71.8	0.5	35.9	A3pii	0.782	1.0	0.788
B1si	72.2	0.5	36.1	B1pi	0.590	2.0	1.18
B1sii	45.6	0.5	22.8	B1pii	0.200	3.0	0.600
B2si	53.9	0.5	26.95	B2pi	1.05	1.0	1.05
B2sii	55.4	0.5	27.7	B2pii	1.61	1.0	1.61
B3si	43.5	0.5	21.75	B3pi	0.412	2.0	0.824
B3sii	42.8	0.5	21.4	B3pii	1.31	1.0	1.31
C1si	41.8	0.5	20.9	C1pi	0.066	3.0	0.1977
C1sii	18.0	1.0	18.0	C1pii	1.93	0.5	0.965
C2si	36.5	1.0	36.5	C2pi	3.75	0.5	1.875
C2sii	40.7	0.5	20.35	C2pii	7.50	0.5	3.75
C3si	29.7	1.0	29.7	C3pi	0.11	3.0	0.33
C3sii	46.7	0.5	23.35	C3pii	1.70	0.5	0.85

APPENDIX C. TapeStation

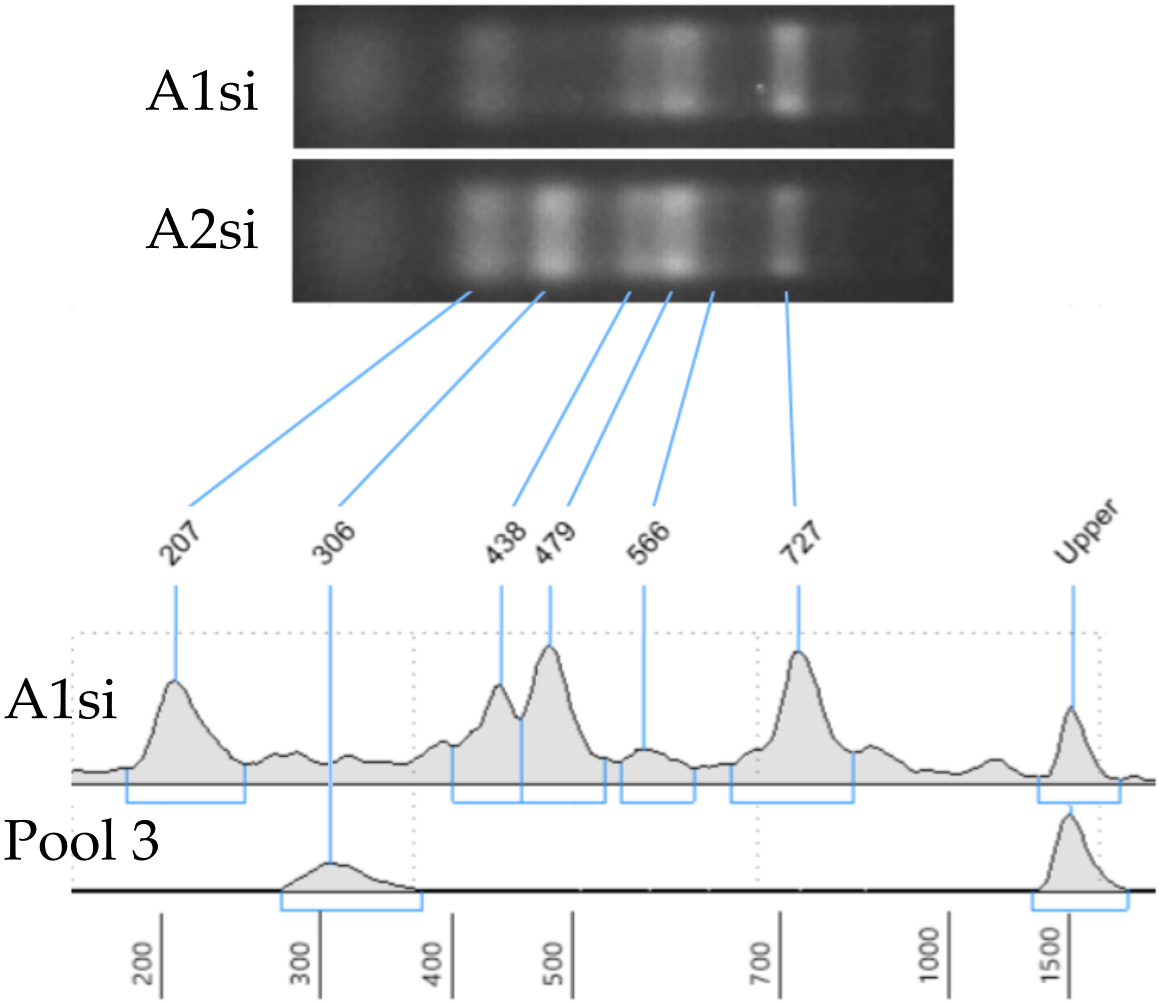

Figure 8: Upper AGE images from samples A1si and A2si (extracted from appendix figure 9) and TapeStation size analyses from A1si and pool 3 (isolated target *pufM* amplicons).

APPENDIX D. Final PCR

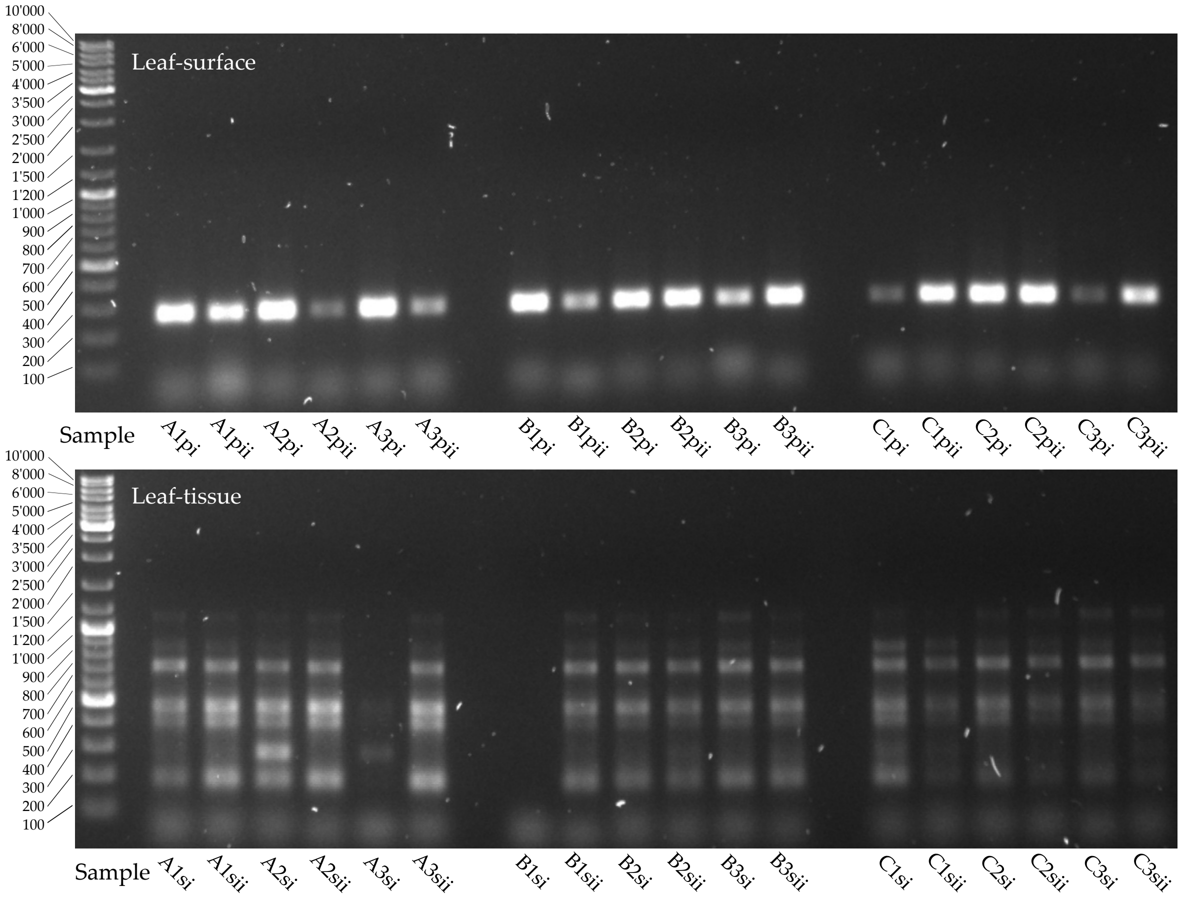

Figure 9: AGE for final PCR of the *pufM* amplicons.

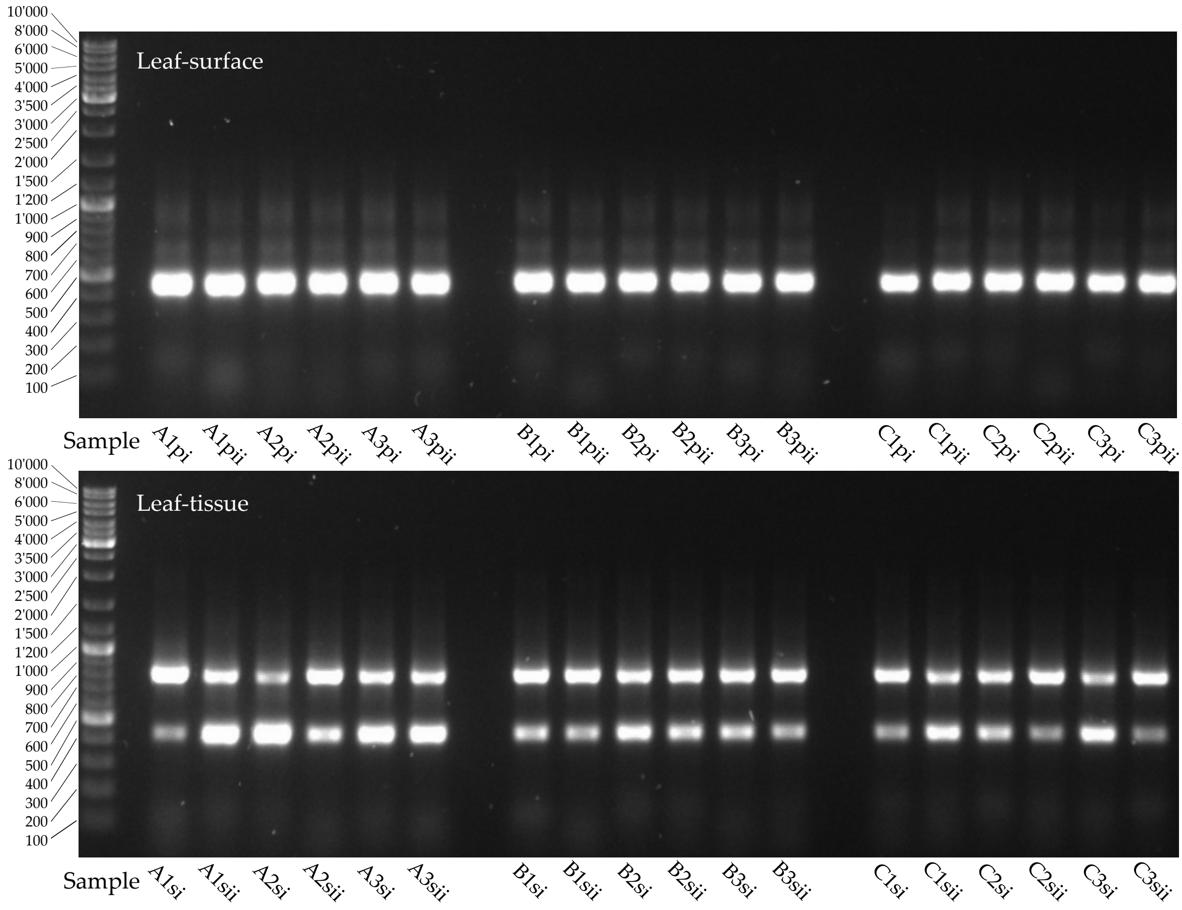

Figure 10: AGE for final PCR of the *16S* amplicons