Genomic and metagenomic analysis of the ongoing speciation between *Priestia megaterium* and *Priestia aryabhattai*

Sam Spence

Asian School of the Environment

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy, 2023.

Candidate Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

11/01/2023

Date

Sam Spence

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is of sufficient quality and grammatical clarity to be examined. To the best of my knowledge, it is free of plagiarism and the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. To the best of my knowledge, the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

11/01/2023

Date

KIM Hie Lim

Authorship Attribution Statement

Please select one of the following; *delete as appropriate:

*(B) This thesis contains material from 1 paper published in the following peer-reviewed journal in which I am listed as an author.

Data used in Chapter 3 is published in Drautz-Moses DI, Luhung I, Gusareva ES, Kee C, Gaultier NE, Premkrishnan BNV, Lee C, Leong S, Park C, Yap Z, Heinle CE, Lau KJX, Purbojati RW, Lim SBY, Lim YH, Kutmutia SK, Aung NW, Oliveira EL, Ng SG, Dacanay J, Ang PN, Spence S, Phung WJ, Wong A, Kennedy RJ, Kalsi N, Sasi SP, Chandrasekaran L, Uchida A, Junqueira ACM, Kim HL, Hankers R, Feuerle T, Corsmeier U, and Schuster SC. Vertical stratification of the air microbiome in the lower troposphere. PNAS. 2022 Feb 15;119(7):e2117293119.

My contribution as a co-author:

- Discussion of project direction and discussion of data interpretation

Chapters 1 and 2 are in preparation as a manuscript to be published after thesis submission, with myself as first author.

Signed:

Sam Spence
Student
Asian School of the Environment
Nanyang Technological University
Date: 11/01/2023

KIM Hie Lim
Supervisor
Asian School of the Environment
Nanyang Technological University
Date: 11/01/2023

Assoc Prof Benoit Taisne
Associate Chair (Research)
Asian School of the Environment
Nanyang Technological University
Date: _____

Acknowledgements

My PhD supervisor, Professor Hie Lim Kim.

My thesis advisory committee, Professors Adriana Lopes dos Santos and Stephan C Schuster.

The past and present members of Professor Kim's research lab: Amit Gourav Ghosh, Elena Gusareva, Namrata Kalsi, Regine Tiong, Hung Nguyen, and Misato Ogasahara.

The past and present members of Professor Schuster's air microbiome research team at SCELSE: Akira Uchida, Ngu War Aung, Elaine Oliveira, Wen Jia Phung, Irvan Luhung, Serene Lim, Anthony Wong, Cassie Heinle, Justine Dacanay, Carmon Kee, Kenny Lau, Rikky Purbojati, Ryan Kennedy, Lakshmi Chandrasekaran, Daniela Moses, Shruti Kutmutia, Yee Hui Lim, Balakrishnan Premkrishnan, Prabu Sekar, and all others.

Professors Janelle Thompson and Yann Boucher for their comments and suggestions.

Table of Contents

Acknowledgements i

Table of Contents ii

List of Figures iv

List of Tables vi

Abstract 1

Chapter 1 — Introduction on the definition of bacterial species 2

Chapter 2 — Phylogenetic study of *Priestia megaterium* and *Priestia aryabhattai* 15

Introduction 15

Methods 18

Priestia strain isolation 18

Whole genome sequencing 18

Genome assembly 19

Phylogenetic analysis 20

Protein-coding marker gene distance 20

16S rRNA gene phylogeny 21

Whole-genome genetic distance phylogeny 21

Clonal genome phylogeny 22

Phenotype testing 22

Results 23

Newly generated whole genome sequencing datasets 23

Phylogenetic analysis 24

Individual marker gene trees 27

Clonal genome phylogeny 31

Whole-genome genetic distances 32

Phenotype testing 33

Discussion 36

Chapter 3 — Gene content comparison of *Priestia megaterium* and *Priestia aryabhattai* 45

Introduction 45

Methods 51

Orthologous gene clustering 51

Pan-genome accumulation curve 53

Calculation of dN and dS 54

Whole genome synteny plots 56

Clustering of genomes by gene content 56

PCR 57

Bacterial culture and DNA template preparation 58

PCR conditions 58

PCR product purification 58

Results 59

Pan-genome accumulation curve 59

Core genome analysis 60

Orthologs exclusive to *P. megaterium* and to *P. aryabhattai* 62

PCR test for discriminating *P. megaterium* and *P. aryabhattai* 64

Orthologs with sequence divergence between *P. megaterium* and *P. aryabhattai* 66

Alignment of whole genomes to reference genomes 70

Nucleotide identity of orthologs to reference genomes....70
Clustering of genomes by gene content....71
Discussion....75
Ortholog clustering by the reciprocal best hits approach....75
Pangenome curve modelling....76
Clustering of genomes by gene content....78
Whole-genome alignment....79
Orthologs with high dN score between species....81
Cobalamin synthesis pathway genes....81
PTS sugar transport genes....84
Prophage proteins....85
Sporulation-related genes....86
Flagellar assembly genes....87
Implications for bacterial life cycle....89
P. aryabhattai as a plant growth promoting bacterium....91
Chapter 4 — Metagenomics study of the global distributions of *Priestia megaterium* and *Priestia aryabhattai*....94
Introduction....94
Methods....99
Metagenomics sampling and sequencing....100
Environmental data access....101
Species reclassification of reads....104
Calculation of species' relative abundance....105
Modelling of species majorities....105
Results....106
Changes in relative abundances after reclassification of reads....106
Vertical distribution of *P. megaterium* and *P. aryabhattai*....108
Global distribution of *P. megaterium* and *P. aryabhattai*....111
Onsite vs weather station metadata....111
Distributions and correlations of environmental variables....112
Ordination of global metadata and the global distribution of *Priestia* species....117
Discussion....130
Conclusions....134
Appendix....140
References....154

List of Figures

Figure 1-1: Schools of thought regarding bacterial species 3

Figure 2-1: Phylogenetic trees of 190 *Priestia* genomes, showing the distinct *aryabhattai* clade 28

Figure 2-2: Neighbour-joining tree of 74 housekeeping genes from 190 *Priestia* genomes 29

Figure 2-3: Maximum likelihood trees of single marker genes from 190 *Priestia* genomes 31

Figure 2-4: Clonal genome phylogeny of *P. aryabhattai* and *P. megaterium* genomes ... 34

Figure 2-5: Pairwise dDDH and ANI values between *P. megaterium* and *P. aryabhattai* complete genomes 35

Figure 3-1: Phylogenetic tree of the *aryabhattai* clade, showing countries of origin for each strain 46

Figure 3-2: Illustration of paralogous and orthologous genes 50

Figure 3-3: Pan-genome accumulation curves for 30 *Priestia* complete genomes 61

Figure 3-4: Core genome and pangenome sizes of 30 genomes from *P. megaterium* and *P. aryabhattai* 62

Figure 3-5: PCR test to identify *P. megaterium* and *P. aryabhattai*, targeting species-specific gene blocks 67

Figure 3-6: Amino acid sequence alignment of the cobalamin synthesis gene *cobU* from 30 *Priestia* genomes 68

Figure 3-7: Amino acid sequence alignment of the ascorbate-specific PTS system EIIB subunit from 30 *Priestia* genomes 68

Figure 3-8: Amino acid sequence alignment of *paiA* acetyltransferase and sporulation negative regulatory gene from 13 *Priestia* genomes 69

Figure 3-9: Amino acid sequence alignment of flagellar biosynthesis gene *flhF* from 30 *Priestia* genomes 69

Figure 3-10: Nucleotide identity of orthologs from each of the four clades to reference genomes of *P. megaterium* and *P. aryabhattai* 71

Figure 3-11: Genome trees of produced by clustering of gene content data from *P. megaterium/aryabhattai* 73

Figure 3-12: Hierarchical clustering on the presence/absence of 9,338 orthologs in 30 *P. megaterium/aryabhattai* genomes 74

Figure 3-13: Adenosylcobalamin synthesis pathway 81

Figure 4-1: Changes in relative abundance of *Priestia* species after reclassifying reads 107

Figure 4-2: Relative abundances of *P. megaterium* and *P. aryabhattai* in two vertically stratified metagenomics experiments 109

Figure 4-3: Ratios of *P. aryabhattai* and *P. megaterium* relative abundances in two vertically stratified metagenomics experiments 110

Figure 4-4: Temperature and humidity readings taken during global air sampling 113

Figure 4-5: Distributions of environmental metadata for 1,180 metagenomics air samples 114

Figure 4-6: Distributions of continuous environmental variables across levels of categorical environmental variables for 1,180 metagenomics air samples 115

Figure 4-7: Correlations between the continuous variables in the global metagenomics dataset 116

Figure 4-8: Distributions of air quality index and monsoon season for 106 air samples taken in Singapore 116

Figure 4-9: Global map of *Priestia* species ratio in air samples for metagenomic sequencing 119

Figure 4-10: Ordination of environmental metadata of 1,180 air samples from around the world. 120

Figure 4-11: Effects of variables in the ordination of 1,180 global metagenome samples from air 121

Figure 4-12: Ordination of environmental metadata of 274 air samples from Southeast Asia 123

Figure 4-13: Ordination of environmental metadata of 106 air samples from Singapore 125

Figure 4-14: Effects of variables in the ordination of 106 air samples from Singapore.. 126

Figure 4-15: Boxplots of *Priestia* relative abundances across levels of geographical region, season, temperature, and humidity 128

List of Tables

Table 1-1: Prokaryote species concepts 5
Table 2-2: Average pairwise ANI similarities and marker gene p-distances for 189 *Priestia* genomes 36
Table 2-3: List of *Priestia* genomes in GenBank which are recommended to be renamed 41
Table 3-1: ANI scores between pairs of *Bacillus cereus* sensu lato strains 48
Table 3-2: PCR primers used to detect strains of the species *P. megaterium* and *P. aryabhattai* 57
Table 4-1: Typical dates of the monsoon seasons of Singapore 102
Table 4-2: Scale of air quality index values and their health implications 103
Table 4-3: Confusion matrices of Random Forest and Support Vector Machine classifiers on global samples 122
Table 4-4: Confusion matrices of Random Forest and Support Vector Machine classifiers on Southeast Asian samples 124
Table 4-5: Confusion matrices of Random Forest and Support Vector Machine classifiers on Singaporean samples 127
Table A-1: Genome assembly statistics of new *P. aryabhattai* genomes 140
Table A-2: Phenotype tests on 12 *P. aryabhattai* isolates and one *P. megaterium* isolate 141
Table A-3: Housekeeping genes used in the multilocus sequence analysis of *P. aryabhattai* and *P. megaterium* 143
Table A-4: Species-specific core genes for *P. megaterium* and *P. aryabhattai* 145
Table A-5: Top 100 orthologs by average dN between *P. aryabhattai* and *P. megaterium* sequences 147

Abstract

Priestia megaterium and *Priestia aryabhattai* are two closely related groups of bacteria with important industrial uses, but it has been debated whether they are the same or different species. In this thesis, I aimed to resolve the identity of these two groups and reveal the functional and ecological differences between them. My phylogenetic study of 190 *Priestia* genome assemblies showed that the two groups are in separate evolutionary clades. The whole-genome genetic distance between them was sufficient to classify them as separate species; the presence of recombinant strains, isolated from environmental samples, suggests a still ongoing speciation. To examine the functional difference between the two species, I identified the orthologous genes in 14 *P. aryabhattai* and 13 *P. megaterium* complete genomes and compared the gene content of the two species. The results showed substantial amino acid sequence divergence between the two species in genes related to cobalamin synthesis, sugar transport, sporulation, and flagellar assembly. Genomes of *P. aryabhattai* also had increased copy numbers of several iron transport genes which are important for plant growth promotion. Species-specific genes were used to design PCR primers for the easily replicated identification of *P. megaterium* and *P. aryabhattai* isolates without whole-genome sequencing. Finally, a global metagenomics analysis showed that *P. megaterium* was the generally more abundant species, but that *P. aryabhattai* may have an ecological niche in Southeast Asia.

Chapter 1 — Introduction on the definition of bacterial species

This thesis concerns the relationship between two groups of bacteria, *Priestia megaterium* and *Priestia aryabhattai*, and whether they are part of the same species. However, there is a long-standing debate over whether bacteria actually form groups that should be called species (1). This question actually consists of two separate problems. The first is whether bacterial lineages form self-similar genomic groups — the species phenomenon, or 'genomic clustering' — and the second is if there is any generalisable category, with properties or cut-offs that are common between different groups, that we can define as a species — the 'species concept' (2).

The first question can be conceptualised as a map of bacterial genomes in sequence space (Figure 1-1). If bacterial genomes naturally form clusters, in which the genomes are more similar to those in the same group than to those in other groups, and clusters are separated by some genetic gap, then it can be said that a type of bacterial species exists, regardless of whether different species are consistent in their properties (3). For animals such as mammals, it is obvious that the process of sexual reproduction enforces such clustering because it creates a strong reproductive barrier between species. For the clonally reproducing bacteria however, it was previously unclear if genetic clusters exist in nature, or if instead the rate of horizontal gene transfer (HGT) between distant genomes is high enough to result in a genetic continuum across sequence space.

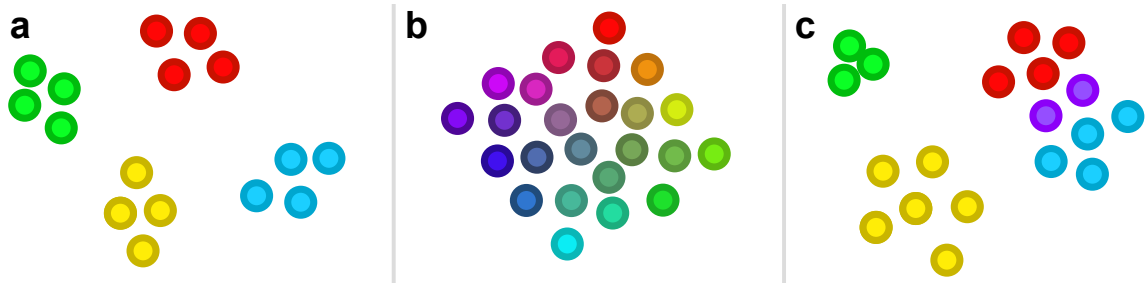


Figure 1-1: Schools of thought regarding bacterial species, conceptualised as clusters in genetic space. The distance between strains represents genetic distance. Colours represent strains with distinct phenotypes. **a:** Species monism; bacteria form uniformly diverse species, with distinct phenotypes, and with a gap in genetic distance between them. **b:** Genetic continuum; different phenotypes are observed but rampant recombination between groups prevents distinct species clusters from forming. **c:** Species pluralism; clusters are often formed but may be inconsistent in size and diversity. Horizontal gene transfer between closely related species may lead to mosaic genomes which blur the line between these 'fuzzy species'.

Bacteria can gain horizontally transferred genes from environmental DNA, from viral infection, and from cell-to-cell DNA transfer (4), at rates which vary across the tree of life (5). Horizontally acquired sequences which are not immediately beneficial for the cell may still become useful later if environmental conditions change, so long as they are neutral or only mildly deleterious and therefore not removed by selection (6,7). It is typical for bacteria to gain horizontally transferred sequences at several times the rate at which mutations are caused by nucleotide substitutions (8), which has led to uncertainty over whether discrete genetic clusters could be maintained in nature (9). However, like other genetic changes, these new sequences are often deleterious and quickly removed (10). Further, the rate of HGT between bacterial genomes decreases as the genetic distance between genomes increases (11). The fact that HGT is more common between closely related genomes that are in frequent physical contact means that it is simplistic to understand HGT as a process that removes species clusters in favour of a genetic continuum; theoretically, it may also increase the genetic clustering that we observe by making genomes within the same cluster more similar to each other than to those outside (12).

With the increasing availability of whole-genome and metagenomics data, it has become clear that bacteria do in fact form genetic clusters despite the horizontal gene transfer between them. For example, recombination between two *Synechococcus* lineages did not prevent them from separating into ecologically distinct sequence clusters (13). Large-scale metagenomics studies have shown bacterial populations that cluster into self-similar groups of over 90% nucleotide identity, with less than 80% identity to the nearest other clusters (12). This gap between sequence clusters in the 80–90% range, where few reads are found, can be interpreted as a natural species boundary, but one that is not rigid nor consistent. The sequence identity values that define this boundary vary between species, as does the prevalence of intermediate sequences. Similarly, the development of the Average Nucleotide Identity (ANI) test for similarity between whole genomes showed a genetic discontinuity between 90,000 genomes: almost all genome pairs were over 96% similar or less than 83% similar to each other, but pairs of genomes at 83–96% similarity were remarkably rare (14).

The discontinuity in genetic distance between species is visible with ANI as well as other distance methods (15), and discontinuity has also been observed at the subspecies level (16). The species-level discontinuity does not seem to be a result of any sampling bias, but may instead be caused by the drop in recombination frequency that occurs around 90–95% ANI, or also by selective sweeps that remove diversity (14). These theories correspond to the recombination species concept and the ecological species concept (Table 1-1). Regardless, ANI discontinuity is now recognised as a useful marker for species delineation in bacteria (17) as well as fungi (18,19) and bacteriophages (20), when used with the understanding that the cut-offs are not exact and that borderline

cases may require a more holistic examination of the organisms involved (19). The existence of bacterial genome clusters is therefore now commonly accepted, even if the rules for species demarcation may differ from those used for eukaryotes (21).

Table 1-1: Prokaryote species concepts that have been proposed to describe the causes of genomic clustering and how to delineate species. It has been debated whether there can be one species concept that can fit all species, or whether multiple concepts should be used to describe different groups.

Species concept	**Description**	**Further reading**
Biological species concept	Species are groups with isolated gene pools due to barriers to gene exchange.	Bobay and Ochman 2017 (22)
Ecological species concept	Species are evolutionary lineages adapted to specific environments. These ecotypes are held together by periodic selection for survival in their niche, reducing diversity by removing less well-adapted individuals.	Cohan 2002 (3)
Recombination species concept	Species are groups whose genomes can recombine. Individual gene trees should agree for genomes from different species, but can disagree for genomes from the same species.	Dykuizen and Green 1991 (23), Fraser et al. 2007 (9)
Phylo-phenetic species concept	Species are monophyletic groups with similar characteristics, and can be discerned by some discriminative phenotype. DNA relatedness metrics may be used to draw species boundaries in a flexible manner.	Rosselló-Mora and Amann 2001 (24)
General lineage concept	Species are separately evolving metapopulation lineages, a general concept that aims to include and unify other species concepts. Other properties such as unique niches or phenotypes are secondary.	de Queiroz 2005, 2007 (25,26), Achtman and Wagner 2008 (27)
Pangenome species concept	Species are groups with internally similar gene repertoires. Requires sequencing of many genomes to cluster genomes using low frequency accessory genes.	Moldovan and Gelfand 2018 (28)

Although many bacteria do indeed form genetic clusters, the second question has been long debated: what is the definition of a prokaryote species? Some authors have taken the strict view that any species definition should be universally generalisable; it should be able to classify all prokaryote cells into identically-defined species using the

same rules, and the criteria that are used – such as genome similarity metrics – should correspond to some natural biological property by which species are distinct from other taxonomic levels (29). Since species groups could not be universally delineated using any rules based on ecology, selection, or recombination, then any species definition could be at best a nominal or functional category without any rigid theoretical backing. The reticulate nature of bacterial phylogenetics that results from lateral gene transfer has also been problematic for any species definition based on evolutionary descent (30).

As a proponent of the view that there are no true bacterial species, Ereshefsky (31) described the state of the field in 2010 as being divided between several schools of thought: 'optimists', who hoped that more research would uncover a definitive species definition (32); 'pessimists', who accepted that there was no consistent category of species for bacteria despite the existence of genetic clusters (24); and others who argued that prokaryote biology could be better conceptualised in terms of other evolutionary units (33). Doolittle (1) similarly described the opposing views as those of species monists, who assert that a strict species definition can be formulated that is generalisable to all organisms; versus species pluralists, who hold that multiple species concepts can each be valid in the case of different organisms.

The species pluralism framework embraces the 'problem' that there may be no reason why all bacteria must form species clusters by the same mechanism — or even at all — but asserts that some do, and that the causes of speciation and diversification may vary between species (1). Whereas Ereshefsky's eliminative pluralism (34) rejects the species concept as anything more than nominal classification because it cannot be strictly and universally defined, other pluralists accept a broader concept of the species

that allows for variations in evolutionary history (24). As an answer to the debate over whether speciation is caused by selective sweeps on clonal lineages (35), or instead by higher levels of recombination within groups than between them (23), pluralism accepts that different groups may be affected by both to various degrees (36), and that these groups can still all be referred to as 'species' despite their differences (37). On a philosophical level, pluralists attempt to find working species concepts that accurately describe the messy reality of prokaryote evolution, rather than trying to force an inappropriately rigid model onto nature (1).

In practice, decisions on where to draw the line between species begin with looking for the gap between clusters which occurs at around 70% DNA hybridisation (38), or with modern methods, at 95–96% ANI (14). Having accepted the pluralist caveat that this cut-off may not be universal, these values are considered as starting points that may be relaxed if phylogenetic or phenotype evidence indicates a more diverse species cluster that is more coherent with lower values (24). This practice has been criticised for relying on arbitrary and flexible cut-offs that make the identified species no more 'correct' than those described by other theories (31). It is true that the value of 70% hybridisation was originally calibrated to match the species groupings which had already been described (39), and this may appear to be a use of circular logic. However, those pre-existing species were not arbitrary groupings, but were decided based on multiple phenotypic and rRNA similarities. The various methods used by taxonomists are in approximate agreement precisely because they are identifying real clusters of organisms which are more similar to each other than to other clusters, and this fact makes these species categories 'correct'.

In the end, the facts are clear and agreed upon: that bacteria commonly form clusters rather than a genetic continuum, with a gap between clusters that can be measured as a discontinuity in ANI values. Species concepts have been criticised for having no property which distinguishes them from other taxonomic levels (31), but this ANI discontinuity seems to be one such property. Given these facts, the debate over whether these clusters should be spoken of as bacterial 'species' seems to be a matter of semantics rather than science. As Mallet wrote in 2005 on species in general, "the debate about species reality boils down, sadly, to different interpretations of the word 'real'" (40).

Previously, several species concepts that described different routes of speciation were seen as competing to be the one correct species definition (31), but it was also known that bacterial species vary greatly in their rates of homologous recombination (41). Given the modern understanding that multiple species concepts can describe real causes of species divergence and cohesion under different conditions (27), recent research has focused on understanding speciation rather than defining species (37). Speciation processes have been modelled under simple drift (42) and under combinations of high or low recombination and selection (36) in order to understand the formation of species under an implicitly pluralist paradigm.

New terminology, such as the concept of 'fuzzy species', has allowed researchers to describe the nuanced realities of certain closely related species. This term comes from a study on 770 strains of 11 *Neisseria* species (43), in which phylogenetic trees made from concatenated housekeeping genes showed that most of the strains were clustered into large clades that matched the previous species descriptions. However, some strains formed smaller clades on the branches between the larger clades. These intermediate

strains had mosaic genomes which had formed by frequent interspecies recombination that had also affected the housekeeping gene sequences, leading to their position on the trees between the two major clades. It was unknown if the high rates of recombination in these intermediate clades were due to some genomic feature that made them more prone to recombination, or if particular environmental conditions had caused more frequent contact and genetic exchange between the ancestors of these strains. Intermediate strains in phylogenetic trees could also be caused by mixed cultures being mistaken as clonal isolates, but this should result in individual hybrid strains placed on separate branches rather than whole clades showing lineages of mixed genomes (44).

The authors used the term 'fuzzy species' to describe such cases where the boundary between species is unclear due to lateral gene transfer, and later described more examples of fuzzy species in the *Streptococcus* (45) and *Pneumococcus* (46) genera. They theorised that having a high enough frequency of recombination to cause this state would require the two species to be capable of growing in the same environment, but that there must also be some physical or genetic barrier to recombination in order for the two species to have partially separated (43).

On the other hand, some fuzzy species could be the result of the convergence of two species that were previously separated. This was suggested by Sheppard *et al*. (47) as the cause of gene flow between *Campylobacter jejuni* and *Campylobacter coli*. The authors found that the two species formed two clades, but with some hybrid strains placed in an intermediate position on the trees. They argued that the two species had previously been genetically similar enough for recombination to occur but that this had

been prevented by physical separation, and that a recent reintroduction of the two species to each other in an agricultural environment had allowed the rate of recombination to increase. This situation was referred to by the authors as a 'despeciation' rather than a fuzzy species.

Regardless of the cause of the indistinct clade boundaries that have been observed among the strains in these fuzzy species, the authors who described them emphasised the importance of large-scale phylogenetic studies to determine whether distinct clusters can be resolved among groups of organisms, rather than relying on strict DDH cut-offs to make taxonomic decisions (45). Their goal was to describe real clusters of organisms, even if these clusters do not correspond to any current species concept (44). In effect, their idea was to make species descriptions that describe reality, which may be fuzzy and imprecise, rather than forcing organisms into boxes which do not fit them.

The genomics era may have settled the debate on whether bacteria form species clusters, but the technical challenges of how to identify and delineate them are still being addressed (48). Since it was previously more challenging to sequence whole genomes, older techniques for species identification relied on sequencing small portions of the genome that were known to show enough variation to be able to identify species. These methods include 16S rRNA gene sequencing (49), which targets one of the bacterial ribosome genes which has both conserved and variable regions, and multi-locus sequence typing (MLST) (50), which uses loci from multiple genes which have known variations between species. The main drawback of both of these methods is that using only short sequences from specific genes will miss any variation that may be

present in other regions, and the variety of differences that can be identified is constrained by the length of the sequences and the number of variable sites within.

Sequencing the 16S rRNA gene has historically been a useful approach to identifying bacterial species because this essential gene is found in every bacterial cell, but this method is also complicated by the fact that bacterial genomes often contain multiple non-identical copies of the 16S rRNA gene, which becomes a problem when the variation between 16S copies within a genome is too high and the variation between 16S copies from different species is too low (51). A 2015 study (52) identified the species of 617 clinical isolates using three methods: with the 16S rRNA gene, using species-specific marker gene probes, and using culture-based biochemical phenotype tests. The study found that species assignments made using the 16S rRNA gene and using the culture phenotype methods agreed only 81% of the time, and that the identifications made using the 16S rRNA gene agreed with the identifications made using other RNA sequences in only 77.5% of cases. They also found several important species groups in which the 16S rRNA gene sequences from different species were over 97% identity, the conventional same-species threshold (53), including species from *Streptococcus*, *Staphylococcus*, *Pseudomonas*, *Klebsiella*, and *Escherichia*. Conversely, their results frequently found same-species 16S sequences that were less than 97% identical. Similarly, 16S sequences from different genera often showed over 95% similarity, which has been used as a same-genus threshold (54). These results suggest that, like other biomarkers, the resolution of the 16S rRNA gene in discriminating species is not universally equal across taxa, and that the groups with highly similar 16S sequences should be identified carefully using additional methods.

A disadvantage of sequencing a specific few marker genes for species identification by MLST is that it requires prior knowledge of which genes have the appropriate level of variation for the group being studied — the substitution rate of each gene varies across species (55). The MLST scheme for a particular group of organisms needs to be designed using genes which are present in only one copy per genome and which are evolving under purifying selection; ideally, 12–18 such loci should be identified that are distributed equally across the genome (56). Because of these requirements, MLST is better suited for the identification and population studies of clinically relevant and well-studied bacteria than for identifying novel species (57). Once a species group has been thoroughly characterised, MLST data has also been used to clarify the relationships between species (58–65), but the method is becoming increasingly redundant as more whole genome sequences have become available (57).

The increase in availability of annotated whole genomes allows for a greater number of genes to be used for species identification, rather than using only one or a few genes which are expected to show sufficient variation between species. A typical approach, called multilocus sequence analysis (MLSA), is to concatenate the whole sequences of multiple conserved genes that are common to all genomes of interest (66). Further, modern whole-genome sequence-based methods such as digital DNA-DNA hybridisation (dDDH) (67) and Average Nucleotide Identity (ANI) (68) are now available instead of or as a complement to the previously dominant methods. The two methods of dDDH and ANI are conceptually similar, and work by aligning fragments of two whole genomes together in order to count the differences between every possible aligned nucleotide.

These methods are digital equivalents of the older method of DNA-DNA hybridisation (DDH) for taxonomic identification, for which it was found that most pairs between strains of the same species produced results higher than 70% (69), a number which became the standard for species delineation (38). However, DDH involves wet lab experiments that are difficult to perform, and the results cannot be compared between studies (70). The rise of whole-genome sequencing data permitted the development of replacement *in silico* methods of genome comparison such as ANI (71). This method was shown to give values that are closely correlated with DDH results, with the 70% DDH cut-off corresponding to 95–96% ANI (68,72). The ANI method is simple; fragments of one genome are aligned against another genome and the successful alignments which have at least 70% alignment length are used to calculate the mean percentage of identical nucleotides (71).

Around the same time, a different group developed a similar method called digital DDH (dDDH), which aimed to produce values which were consistent with those from classical DDH experiments so that the results could be compared to those from DDH with the same 70% species cut-off (73). The dDDH similarity $d(X, Y)$ between two genomes X and Y is given by the formula (74):

$$d(X,Y) = 1 - \frac{2 \cdot I_{XY}}{H_{XY} + H_{YX}}$$

Where

XY := a BLAST search with genome X as the query and genome Y as the subject

I_{XY} := the sum of identical base pairs over all BLAST matches

H_{XY} := the total length of all BLAST matches

ANI and dDDH were both developed in comparison to and validated by comparison to wet lab DDH values, and so the same-species thresholds of both whole-genome methods (95–96% and 70%) stem from the 70% cut-off of DDH. In the original DDH method, this 70% value was found to be the value which best separates the bacterial species which had been identified in the past by older methods based on morphology, phenotyping, and 16S rRNA gene sequencing (69). Both the ANI and dDDH methods have the limitation that they do not consider genomic regions that cannot be aligned between the two genomes, therefore species-specific genes and other information such as genome rearrangements are not considered in the species classification decision (75). Additionally, whole-genome sequencing can be a barrier to entry because of its cost and is not always used when describing new species, e.g. (76). Overall, a lack of standardisation in the use of the various species identification methods can lead to disagreements over the species status of newly discovered bacteria, such as in the case of *Priestia megaterium* and *Priestia aryabhattai.*

Chapter 2 — Phylogenetic study of *Priestia megaterium* and *Priestia aryabhattai*

Introduction

Priestia megaterium (basonym: *Bacillus megaterium*) is a gram-positive, ubiquitous environmental bacterium in the phylum Bacillota (previously Firmicutes) which has been well-studied since its discovery in the 19th century (77,78), in part due to its large cell size which facilitates physiological study of cell division, sporulation, and biofilm formation (79,80). A spore-forming generalist, it grows in many types of environments, with genome assemblies sequenced from diverse environments including soil (81), water (82), air (83), and plant and human tissue (84,85). *P. megaterium* is an important species in research and biotechnology (86). First, as a natural producer of cobalamin (vitamin B12), *P. megaterium* has been used both for industrial production of cobalamin (87) and for research into the natural B12 synthesis pathway (88,89). Second, recent research has focused on the role of *P. megaterium* as a plant growth promoter, whether by suppressing plant pathogens (90), by increasing nutrient availability (91), or by producing plant growth hormones (92). Finally, some *P. megaterium* strains have been investigated for their potential in bioremediation (93).

The closely-related species *Priestia aryabhattai* (basonym: *Bacillus aryabhattai*) was first named in 2009 after being identified in air samples taken at a height of 41 km (76). Strains in this new species have attracted industrial interest for their capabilities for plant growth promotion (94), bioremediation (95), and synthesis of biodegradable plastics (96). Both *P. aryabhattai* and *P. megaterium* were previously in the genus

Bacillus, a diverse group in the family Bacillaceae that had long been long known as a heterogeneous group in need of revision (97). The genus was in this state because historically, *Bacillus* had been the default group for aerobic rod-shaped spore-formers, without better understanding of their phylogenetic relationships (98). A 2020 study eventually categorised these disparate groups into 17 new genera, moving *megaterium*, *aryabhattai,* and *flexa* to the new genus *Priestia*, along with a few species that lacked whole genome sequences (99). However, the phylogeny within this new genus has not been adequately resolved due to a dispute over the species status of *P. aryabhattai*.

When describing the first *P. aryabhattai* strain, Shivaji *et al.* (76) found disagreement between the species identification methods that were used. They compared the new *aryabhattai* strain with a representative of *P. megaterium* and found that the sequence similarity between the 16S rRNA genes was 99.7%, which is above the 97% threshold that is used to delineate species (53). However, they argued that *P. aryabhattai* should be considered a new species because the whole-genome similarity to *P. megaterium* was only 35%, and because the two strains differed in their responses to several biochemical tests. A later study (100) instead argued that *P. aryabhattai* strains should be renamed to *P. megaterium*. They found ANI values of 95.3 and 95.8% between the type strains of the two bacteria, higher than the same-species level of 95–96% (72), and considered the physiological differences between the two strains to be too few to name *P. aryabhattai* as a new species.

The approaches used by both previous studies (76,100) provide an incomplete picture of the issue because of the lack of data. Only two genomes were compared, with one from each species, therefore it is not known if *P. aryabhattai* forms a separate,

monophyletic clade from *P. megaterium*, which would be important evidence as to the species status of *P. aryabhattai.* The number of gene sequences used for comparison have also been limited, and dDDH results have not been reported between the groups. The two studies also used biochemical test results to reach opposite conclusions because there was no standardised method for delineating the species using these tests.

As of February 2022, there are 42 *P. aryabhattai* whole genome assemblies in the GenBank database, as recent authors who have identified new strains have named them based on their closest database match without further genomic analysis. These 42 genomes, which may be important in biotechnology, are all potentially misclassified until the question of *P. aryabhattai*'s species status is resolved. For these reasons, it is important that the taxonomic dispute be resolved and that the *Priestia* genomes involved be definitively classified.

Whilst previous studies of these two bacteria have only compared a single genome from each species for species identification, a larger dataset would allow for a more detailed view of the diversity and structure within the *Priestia* genus. In this study, 12 whole genome de novo assemblies were generated and analysed together with 178 genome assemblies that were retrieved from the GenBank database to examine the phylogenetic relationships between *P. megaterium* and *P. aryabhattai* genomes, in order to clarify the species identification of the two groups.

Methods

The sample collection, whole-genome sequencing, genome assembly, and phenotype testing for the unpublished genomes presented here were performed in collaboration with a research team in SCELSE, Nanyang Technological University. The methods can be found in a previous study (83).

Priestia strain isolation

Air samples were collected at multiple indoor and outdoor locations (including residential areas, offices, parks, and beaches) in Singapore in 2015 using Andersen single-stage impactors (SK, USA). The air was impacted onto multiple agar types (M9 minimal salts, R2A, potato dextrose, malt extract, brain heart infusion, and marine agar) which were then incubated overnight. The resulting colonies were sub-cultured on Tryptic Soy Agar (Sigma-Aldrich, USA) and the strains were individually inoculated in lysogeny broth (LB, Becton–Dickinson, USA) at 30 °C overnight to obtain axenic cultures.

Whole genome sequencing

The DNA was purified using the Wizard genomic DNA purification kit (Promega, USA). A genomic DNA library was then prepared with the SMRTbell template prep kit 1.0 (Pacific Biosciences, USA). The DNA was sheared using the g-Tube shearing method and size selected using BluePippin size selection (cutoff = 15 kb). Single-molecule real-time (SMRT) sequencing was done on a PacBio RS II sequencer (DNA sequencing kit 4.0 v2).

Whole-genome shotgun libraries were constructed using the TruSeq Nano DNA library preparation kit (Illumina, USA), and short-read data were generated via a paired-end Illumina MiSeq run with a 300 bp read length. All software was run with default settings unless otherwise stated.

Quality control of the PacBio reads was done using PreAssembler Filter v1 from the Hierarchical Genome Assembly Process v3 (HGAP3) (101) protocol, as implemented in the PacBio SMRT Analysis 2.3.0 package. The MiSeq reads quality control was done using Cutadapt v1.8.1 (102).

Genome assembly

De novo assembly of the PacBio subreads was performed using HGAP3 and polished with Quiver (101). The quality of the draft assembly was further improved by using MiSeq paired-end reads with Pilon v1.16 (103) (tracks –changes –vcf –fix all –mindepth 0.1 –mingap 10 –minmq 30 –minqual 20 –K 47), to create platinum-grade genome assemblies. The completed assemblies consist of 4–14 contigs each (Table 2-1), with GC content ranging from 37.18 to 38.14%. Contig lengths and GC content were obtained with the Quality Assessment Tool for Genome Assemblies (QUAST) (104). The completeness and circularity of the chromosomes and plasmids were evaluated using BUSCO (105) and Circlator v1.5.6 (106). The isolation, sequencing, and assembly of the 13 *Priestia* strains was performed by the authors of the genome announcement (83).

Phylogenetic analysis

Protein-coding marker gene distance

Two phylogenetic trees were constructed using marker gene sequences, using the 'bac120' set of 120 marker genes (107), of which 74 were consistently found in the *Priestia* genomes. These 74 genes are listed in Table A-3. To obtain the marker gene sequences for each genome, a reference sequence for each gene was taken from the GenBank-annotated reference genome *P. megaterium* 22-2 (accession number GCA_009935415.1). These reference gene sequences were used as BLASTN queries to find the homologous sequence with the lowest e-value from each other genome. Five genome assemblies which were missing more than one of the 74 genes were excluded from the analysis, leaving 190 genomes including a *Priestia flexa* outgroup. *P. flexa* was selected as the outgroup because it is the next closest species to *P. megaterium/aryabhattai* by ANI.

The DNA nucleotide sequences of the genes from each genome were aligned using MUSCLE v3.8.31 with default parameters (108), and the 74 marker gene alignments were concatenated into one large alignment for phylogenetic inference. The alignment was trimmed using trimAl v1.4 (109), removing 109 poorly aligned nucleotide sites and leaving a 84,180 bp alignment. A maximum likelihood tree with bootstrap support was constructed from the alignments using RAxML version 8.2.12 with the option –GTRCAT (General Time Reversible model of nucleotide substitution under the Gamma model of rate heterogeneity) (110). Trees were also constructed for each of the 74 individual gene alignments using the same method. A neighbour-joining tree was constructed from the

concatenated gene alignment using ape (111) with the option 'TN93' for the distance matrix (Tamura and Nei, 1993). Both trees were extremely similar as shown in Figure 2-1.

16S rRNA gene phylogeny

To construct a phylogenetic tree from 16S rRNA gene sequences, the 16S sequence of the representative genome *Priestia megaterium* 22-2 (accession number GCA_009935415.1) was used as a BLASTN query against the other 189 genomes to find the closest matching sequence from each genome. The sequences from each genome were aligned together with MUSCLE (108). Eight poorly aligned sites were removed using trimAl, leaving a 1,488 bp alignment. RAxML was used for maximum likelihood tree building and bootstrapping.

Whole-genome genetic distance phylogeny

Trees were also constructed using whole-genome ANI and dDDH. ANI was calculated between all pairs of genomes using FastANI (14) and used to build a Neighbour-Joining tree using the R package 'ape' (111). The tree was plotted using the R package 'ggtree' (112). The pairwise ANI values were also used to construct a network tree using the Neighbour-net algorithm in SplitsTree5 (113). Pairwise dDDH values were calculated between the 30 genomes which were listed as 'complete' on GenBank, as well as the 12 newly sequenced genomes, using the GGDC web server (114).

Clonal genome phylogeny

To reduce the confounding effect of recombination between genomes on the phylogenetic analysis, a tree was also created using only the genomic sections which appear to have been inherited clonally. Due to the high computational load, this analysis was done using a reduced set of 18 high-quality genomes: 6 from the *aryabhattai* clade, 6 from the *megaterium* clade, 4 from recombinant clade 1, and 2 from recombinant clade 2 (Figure 2-1). These genomes included 11 assemblies from the GenBank database and seven of the newly sequenced isolates (Table 2-1).

A core genome alignment was created from the genome assemblies using progressiveMauve v2.4.0 (115). A phylogenetic tree was made from the core genome alignment with RAxML v8.2.12 (110). ClonalFrameML v1.12 (116) was run using the whole-genome alignment, the phylogenetic tree, and the transition/transversion ratios calculated by RAxML. The ClonalFrameML method compares the aligned genomes and marks sites that have unusually high substitution rates as putative recombination events from external genomes. It then constructs a tree with these sites excluded, in order to estimate the phylogeny of the sections of the genomes which have evolved clonally.

Phenotype testing

Isolates were freshly revived from –80 °C storage and cultured on trypticase soy agar at 30 °C overnight. After 24 hr, a colony was sub-cultured in trypticase soy broth at 30 °C for 48 hr. Biochemical phenotype tests were then performed using the API 20 E and 50 CH kits (bioMerieux, USA) following the manufacturer's standard protocol. After

incubating for 48 hr, the results were manually observed according to the manufacturer's standard instruction. API 50 CHB (version 4.1) in APIWEB was used as the reference for species interpretation by matching the isolates' biochemical profiles. The phenotype testing was performed by SCELSE.

Results

Newly generated whole genome sequencing datasets

Twelve new platinum-grade whole-genome assemblies were generated of isolates from outdoor air samples collected in various locations in Singapore (Table 2-1). The assemblies have 4–14 contigs, with an average total size of 5,762,982 bp and an average N50 of 5,095,237 bp, indicating thorough assembly of the main chromosomes. The genomes were determined to have low contamination and heterogeneity by CheckM 1.0.7 (117) (Table A-1), and completeness was found to be over 98% by BUSCO 5.0.0 (118).

Species identification of the 12 genomes was performed using FastANI version 1.32 (14) to search for the closest matching genome in the NCBI RefSeq database (downloaded April 2021). Nine out of the 12 assemblies had ANI above 98% to database genomes named *P. megaterium*, and one to *P. aryabhattai*, whilst two genomes showed ANI matches of just over 96% to *P. megaterium*. However, the 16S rRNA genes of all the genomes had the greatest similarity to *P. aryabhattai* (Table 2-1). These mismatches between the results of the species identification by the two methods of ANI and 16S

rRNA indicate an issue with species identity in the *Priestia* genus. The 12 genome assemblies in Table 2-1 were all named *Priestia aryabhattai* due to the findings of the phylogenetic study below: the two genomes which were the closest GenBank matches to the 12 new genomes — *P. megaterium* Q3 and YC4-R4 —are part of a separate *P. aryabhattai* clade, and are recommended to be renamed as such (Table 2-3).

Phylogenetic analysis

To determine the phylogenetic relationship between the two species *P. aryabhattai* and *P. megaterium*, I conducted a thorough phylogenetic study of all available genomes of the two species, including the 12 genomes generated in this study (Table 2-1) as well as 178 whole-genome assemblies that were downloaded from GenBank (including one outgroup). I constructed five phylogenetic trees based on protein-coding gene sequences (Figure 2-1a, Figure 2-2), whole-genome ANI (Figure 2-1b, Figure 2-1c), and 16S rRNA gene sequences (Figure 2-1d).

Table 2-1: Summary statistics of new *Priestia* genomes sequenced and used in this study. More assembly details in Table A-1. The species identity of these genomes varies depending on the identification method used. The genome assemblies *P. megaterium* YC4-R4 and *P. megaterium* Q3, which are the closest ANI match to most of these new genomes, are recommended to be renamed as *P. aryabhattai* following the results of this Chapter (Table 2-3).

GenBank accession no.	Strain name	Best match by ANI (%) to GenBank	Best match by 16S rRNA gene sequence to GenBank (% identity)	Number of contigs	Genome size (bp)	N50 (bp)	GC content (%)
CP028074-CP028080	*P. aryabhattai* SGAir0178	*P. megaterium* Q3 (98.5)	*P. aryabhattai* B8W22 (99.8)	7	5,473,470	5,001,814	38.03
CP025620-CP025623	*P. aryabhattai* SGAir0179	*P. megaterium* YC4-R4 (99.4)	*P. aryabhattai* B8W22 (99.9)	4	5,305,001	5,053,919	38.14
CP028043-CP028049	*P. aryabhattai* SGAir0202	*P. megaterium* Q3 (98.4)	*P. aryabhattai* B8W22 (99.7)	7	6,362,169	5,077,550	37.22
CP028019-CP028030	*P. aryabhattai* SGAir0257	*P. megaterium* YC4-R4 (99.4)	*P. aryabhattai* B8W22 (99.7)	13	5,576,818	5,017,599	37.93
CP027997-CP028008	*P. aryabhattai* SGAir0265	*P. megaterium* YC4-R4 (99.4)	*P. aryabhattai* B8W22 (99.8)	12	5,571,671	5,011,965	37.93
CP027989-CP027996	*P. aryabhattai* SGAir0269	*P. aryabhattai* K13 (99.1)	*P. aryabhattai* B8W22 (99.7)	8	5,516,307	5,028,224	37.94
CP027914-CP027919	*P. aryabhattai* SGAir0414	*P. megaterium* Q3 (98.5)	*P. aryabhattai* B8W22 (99.8)	7	5,390,002	5,159,113	38.01
CP027900-CP027906	*P. aryabhattai* SGAir0424	*P. megaterium* Q3 (96.3)	*P. aryabhattai* B8W22 (99.5)	8	5,659,875	5,262,388	37.95
CP027889-CP027899	*P. aryabhattai* SGAir0425	*P. megaterium* Q3 (98.5)	*P. aryabhattai* B8W22 (99.8)	14	6,481,834	5,183,207	37.18
CP027876-CP027885	*P. aryabhattai* SGAir0427	*P. megaterium* YC4-R4 (99.6)	*P. aryabhattai* B8W22 (99.6)	10	5,628,789	5,103,336	37.94
CP027870-CP027875	*P. aryabhattai* SGAir0428	*P. megaterium* Q3 (96.4)	*P. aryabhattai* B8W22 (99.9)	6	5,617,367	5,226,421	37.93
CP027931-CP027939	*P. aryabhattai* SGAir0563	*P. megaterium* YC4-R4 (99.4)	*P. aryabhattai* B8W22 (99.5)	10	6,572,477	5,017,311	37.19

The four trees (Figure 2-1a-c, Figure 2-2) separate most of the genomes (165 genomes) into two clear clades, which I refer to as the *megaterium* clade (130 genomes) and the *aryabhattai* clade (35 genomes). The classification of the two clades is supported by the protein-coding gene trees and the ANI trees, with the members of each clade being identical by each method. The phylogenetic tree constructed using the 16S rRNA gene sequence, a commonly used bacterial marker gene, is unable to distinguish the two clades (Figure 2-1d). This marker gene has very little variation among *P. megaterium* and *P. aryabhattai*, with many genomes from both clades having completely identical sequences. The lack of variation in the 16S rRNA gene caused its tree to show an unbranched structure in which strains from both clades were mixed together, rather than a tree structure with two clades as shown by the trees which used other methods.

The phylogenetic trees contained 24 *P. aryabhattai/megaterium* genomes which were not placed in the *megaterium* or *aryabhattai* clades, but instead formed two smaller clades with inconsistent positions across the trees. For example, one of the clades clustered with the *megaterium* clade in the protein-coding gene tree (Figure 2-1a) but with the *aryabhattai* clade in the whole-genome ANI tree (Figure 2-1b). The network tree (Figure 2-1c) showed clearly that the two small clades are placed in an intermediate position between the *megaterium* and *aryabhattai* clades. I therefore defined these two smaller groups as recombinant clades 1 and 2 (14 and 10 genomes each). Although recombinant clade 1 is not monophyletic in the protein-coding gene tree (Figure 2-1a), the clade is monophyletic in the ANI and network trees.

Most of the genome assemblies retrieved from GenBank have already been assigned the species name that corresponds to their clade in our trees. I found 17 genomes whose assigned species names do not match the new classifications based on this phylogenetic analysis; the suggested species names for these genomes are provided in Table 2-3.

I calculated the average distances within and between the four clades by whole-genome ANI and by nucleotide p-distance of the 74 genes (Table 2-2). The *megaterium* clade had greater diversity (2.67% +/- 1.23 average ANI distance) than the *aryabhattai* clade (1.37% +/- 0.39 average ANI distance), suggesting a more recent common ancestor for *P. aryabhattai* than for *P. megaterium*. The whole-genome ANI values were several times larger than the p-distances of the protein-coding genes, suggesting that ANI is more effective in showing the extent of genetic diversity at the strain level than the p-distances of the protein-coding genes, which are house-keeping genes with conserved sequences between genomes.

Individual marker gene trees

In addition to the tree constructed from the concatenated alignment of 74 protein-coding genes, I also made trees from each individual gene in order to assess the contribution of each gene to the consensus tree.

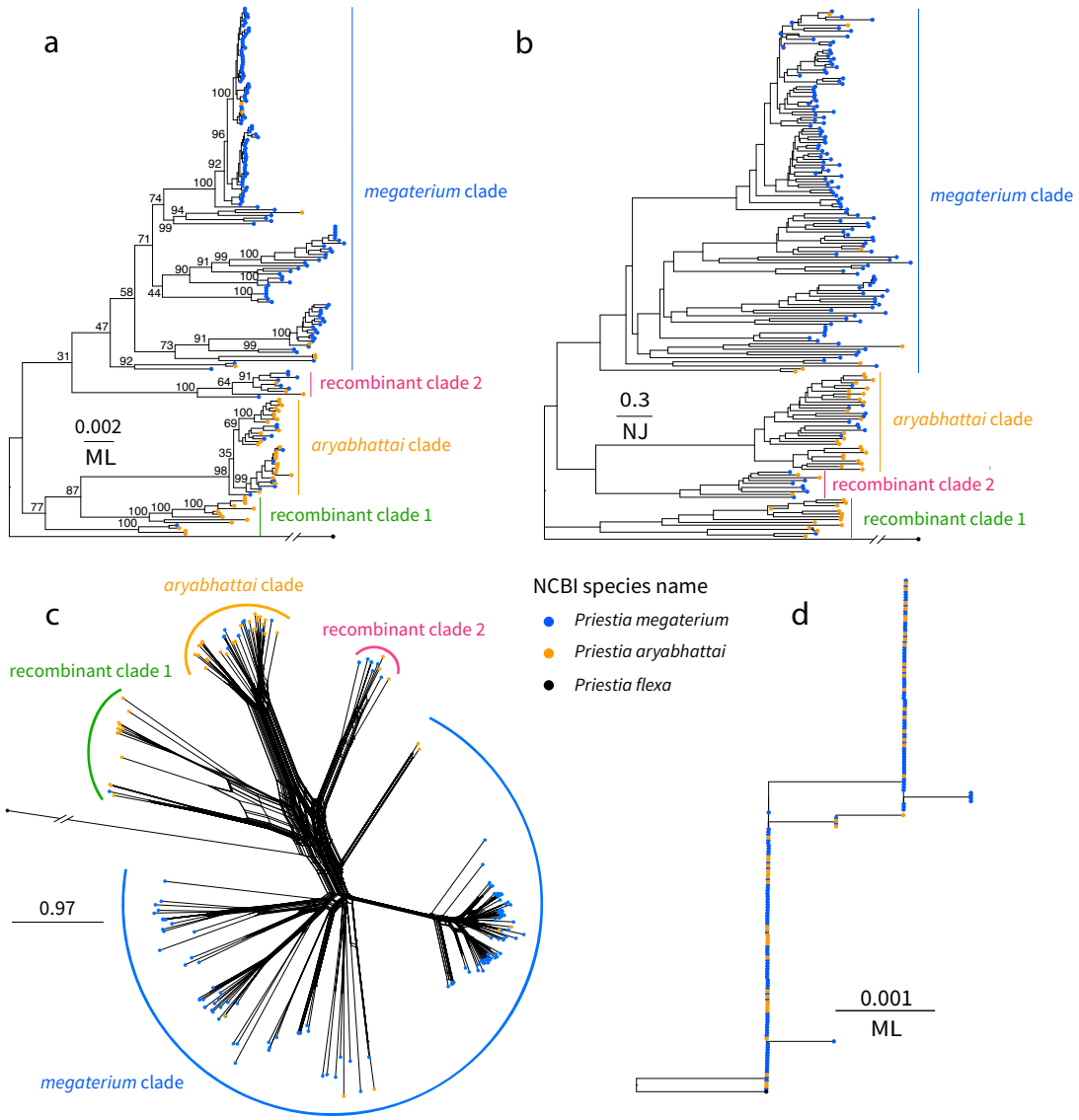


Figure 2-1: Phylogenetic trees of 190 *Priestia* genomes, showing the distinct *aryabhattai* clade (orange). Tree tip colours indicate the names currently given to the genomes in the GenBank database; the nearest species *Priestia flexa* is shown as an outgroup. The genomes form two clades that are mostly congruent with the assigned species names, separated by a long branch. **a:** Maximum likelihood tree, constructed from 74 marker gene sequences. Branch labels show bootstrap support out of 100 runs. Branches with low bootstrap support are due to a lack of phylogenetic signal in some marker genes, and conflicting signal for the recombinant clades. **b:** Neighbour-joining tree, constructed from pairwise ANI scores. **c:** Unrooted phylogenetic network of pairwise ANI scores, built using the Neighbour-net algorithm. The recombinant clades, which are inconsistently placed in **a** and **b**, are shown clearly as intermediate between the *megaterium* and *aryabhattai* clades. See details in Methods. **d:** Maximum likelihood tree of 16S rRNA gene sequences, which are not able to discriminate between most *P. megaterium* and *P. aryabhattai* genomes and thus do not separate the genomes into the clades shown in **a** and **b**.

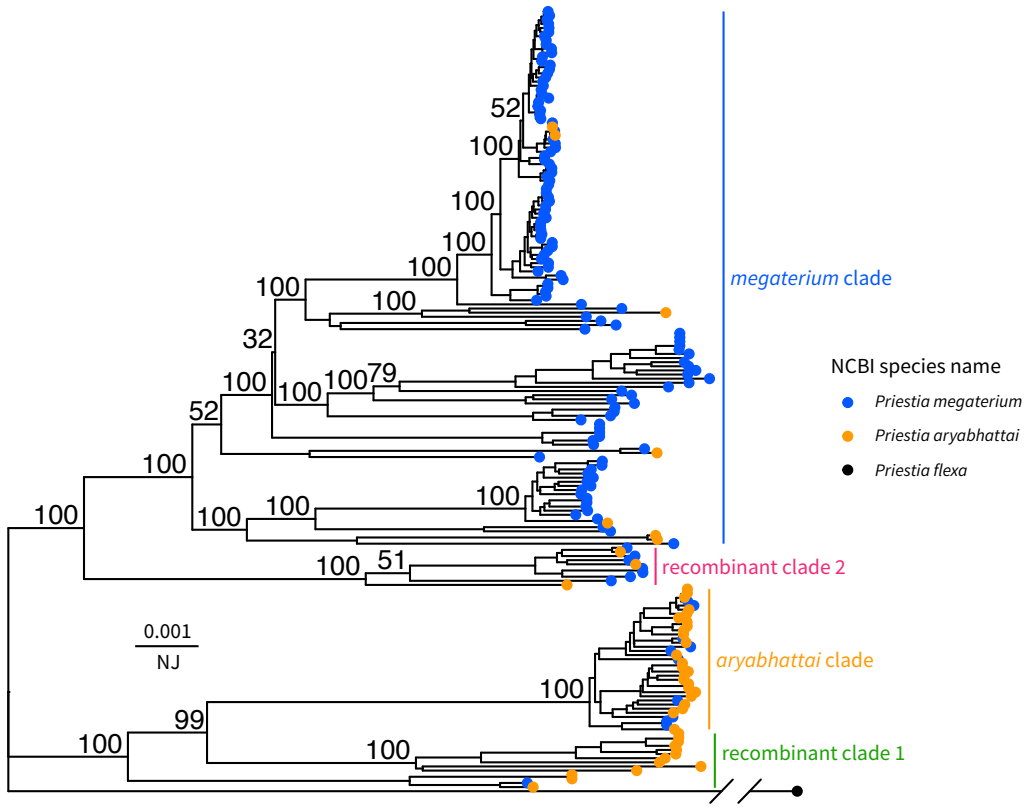


Figure 2-2: Neighbour-joining tree of an 84,180 bp alignment of 74 housekeeping genes from 190 *Priestia* genomes. The maximum-likelihood tree constructed from the same data is shown in Figure 2-1a and is identical save for minor variations in relative branch lengths.

Fifteen of the 74 gene trees placed the four clades in the same positions as the concatenated gene tree in Figure 2-1a (recombinant clade 1 grouped with *P. aryabhattai*, and recombinant clade 2 grouped with *P. megaterium*). One of these genes was *rpoB* (RNA polymerase beta subunit), which has previously been reported as a useful marker gene for phylogenetic analysis (119), particularly in Bacillales species for which it has previously been preferred over the 16S rRNA gene (120). The *rpoB* tree for the *Priestia* genomes (Figure 2-3a) provided a relatively high level of taxonomic resolution compared to other genes, shown by longer branches between clades, fewer groups of identical sequences, and fewer polytomies (where a branch splits into three or more subgroups, rather than being resolved into nested bifurcations).

The difficulty of determining the evolutionary history of the recombinant clades is illustrated by the remaining gene trees. Three gene trees (e.g. Figure 2-3b) were arranged similarly to the whole-genome ANI tree (Figure 2-1b), with recombinant clade 2 grouped with the *aryabhattai* clade and recombinant clade 1 placed outside of the *megaterium* and *aryabhattai* clades. Ten trees placed both recombinant clades in a group with the *aryabhattai* clade, and 11 trees placed both recombinant clades in a group with the *megaterium* clade. Eighteen more trees (e.g. Figure 2-3c) placed the recombinant clades in other positions, such as recombinant clade 1 with the *megaterium* clade and recombinant clade 2 with the *aryabhattai* clade (the opposite arrangement to the concatenated gene tree in Figure 2-1a).

Finally, 19 gene trees (e.g. Figure 2-3d) were not useful for phylogenetic inference because too many genomes were identical in their sequences, including those from different clades. These genes included many 30S and 50S ribosomal genes, which have been previously used for phylogenetic inference in other species (121). My results show that, like the 16S rRNA gene, these genes are too conserved among *P. megaterium* and *P. aryabhattai* to be used for this purpose.

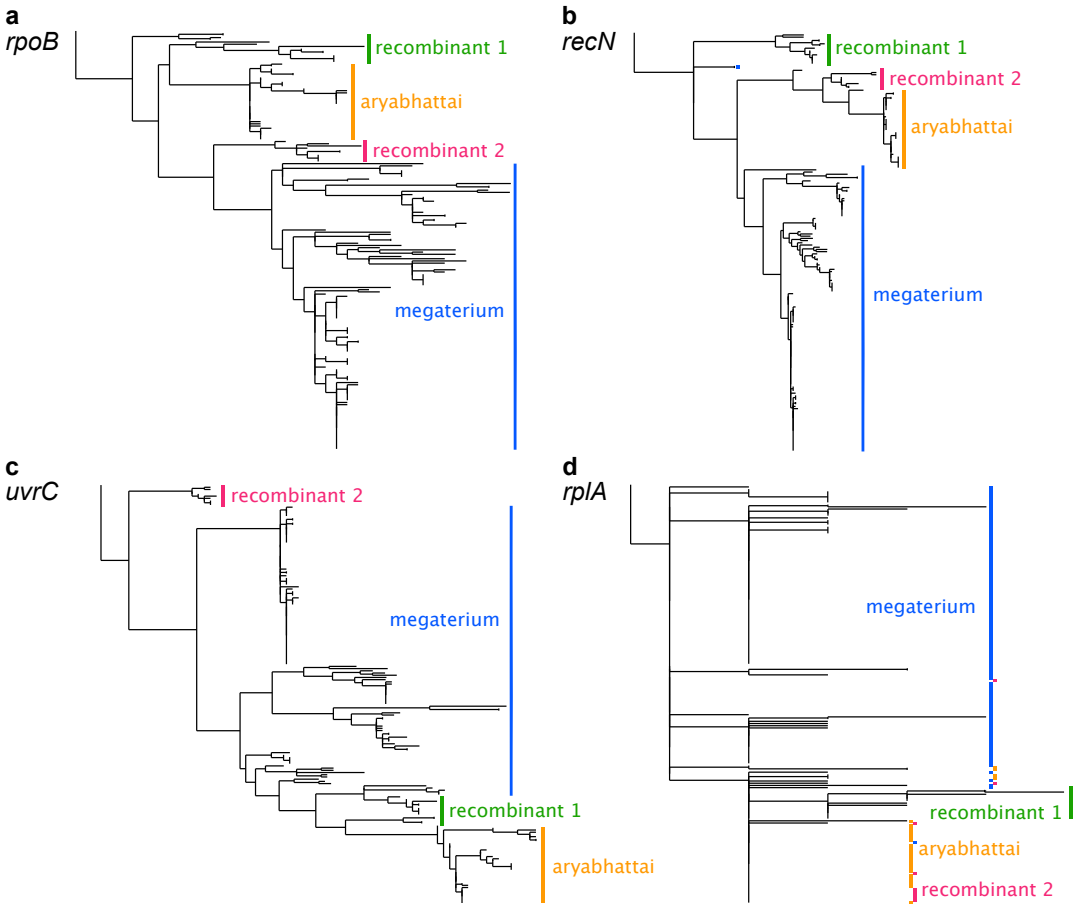


Figure 2-3: Maximum likelihood trees constructed separately using four of the 74 genes that were used in Figure 2-1a. Trees are rooted to the outgroup *Priestia flexa* GN22-4; the branch length to the outgroup has been shortened for clarity. a: gene tree of *rpoB* (RNA polymerase beta subunit), which recreates the same tree topology as Figure 2-1a. **b**: gene tree of *recN* (DNA repair protein), which gives a similar tree structure to the ANI tree in Figure 2-1b. **c**: gene tree of *uvrC* (DNA repair protein), which places the *aryabhattai* clade and recombinant clade 1 as a subgroup within the *megaterium* clade. **d**: gene tree of *rplA*, which lacks sequence substitutions between genomes and is not useful for phylogenetic analysis.

Clonal genome phylogeny

A clonal genome phylogeny of the *P. aryabhattai/megaterium* group was constructed in order to determine the ancestry of the recombinant clades. The results of the clonal genome phylogeny showed extensive putative recombination events throughout the tree. The tree of the clonal genome recreated the four clades that were found in Figure 2-1. The clonal genome tree placed recombinant clade 1 in a group with the *aryabhattai*

clade, and placed recombinant clade 2 outside of both the *aryabhattai* and *megaterium* clades.

Whole-genome genetic distances

I aimed to determine whether the whole-genome genetic distance between the *megaterium* and *aryabhattai* clades is great enough to consider them as separate species. Using the methods dDDH (114) and ANI (14), genomes are conventionally defined as belonging to the same species if they are at least 70% or 95–96% similar, respectively (68,72).

The two metrics of dDDH and ANI showed inconsistent results for species identification (Figure 2-5). The pairwise dDDH values showed that the similarities between the *megaterium* clade and the *aryabhattai* clade were always below 70% (purple, lower left), with an average of 64% +/-0.08%, indicating that they are separate species. However, most of the pairwise ANI values were within the 95–96% threshold, with an average of 95.2% +/-0.02%. The majority of genome pairs (174 out of 182) were above 95% ANI, suggesting that *P. megaterium* and *P. aryabhattai* are the same species. Thus, the two distance measurements do not consistently separate the two clades as different species due to their close genetic relationships with each other.

The pairwise values of both dDDH and ANI between the recombinant clades and the other clades (green and black) were higher than the values between *P. megaterium* and *P. aryabhattai* (purple), and lower than the values within *P. megaterium* or within *P. aryabhattai* (blue and orange), supporting their intermediate position between

P. megaterium and *P. aryabhattai*. The dDDH values between the recombinant clades and the other clades were lower than the species identification threshold, while the ANI values were higher than the threshold. In addition, most recombinants (19 out of 20 genomes) are genomes that were originally isolated from outdoor environments. The high ANI between the recombinants and the other clades, and the presence of naturally occurring recombinants, suggests a recent split or ongoing speciation between *P. megaterium* and *P. aryabhattai*.

Phenotype testing

The results of 61 biochemical phenotype tests on 13 *Priestia* isolates are shown in Table A-2. No test or tests were found that could distinguish between the *aryabhattai* and *megaterium* clades that were shown in Figure 2-1. Some tests were identical for both clades, with every *P. aryabhattai* and *P. megaterium* isolate giving the same result, while other tests were inconsistent within clades, so they could not tell a *P. aryabhattai* isolate from a *P. megaterium* isolate because not all *P. aryabhattai* isolates gave the same result. APIWEB successfully identified all 13 strains as belonging to the larger *Priestia megaterium* group; of course, it was not designed to differentiate between *P. megaterium* and *P. aryabhattai.*

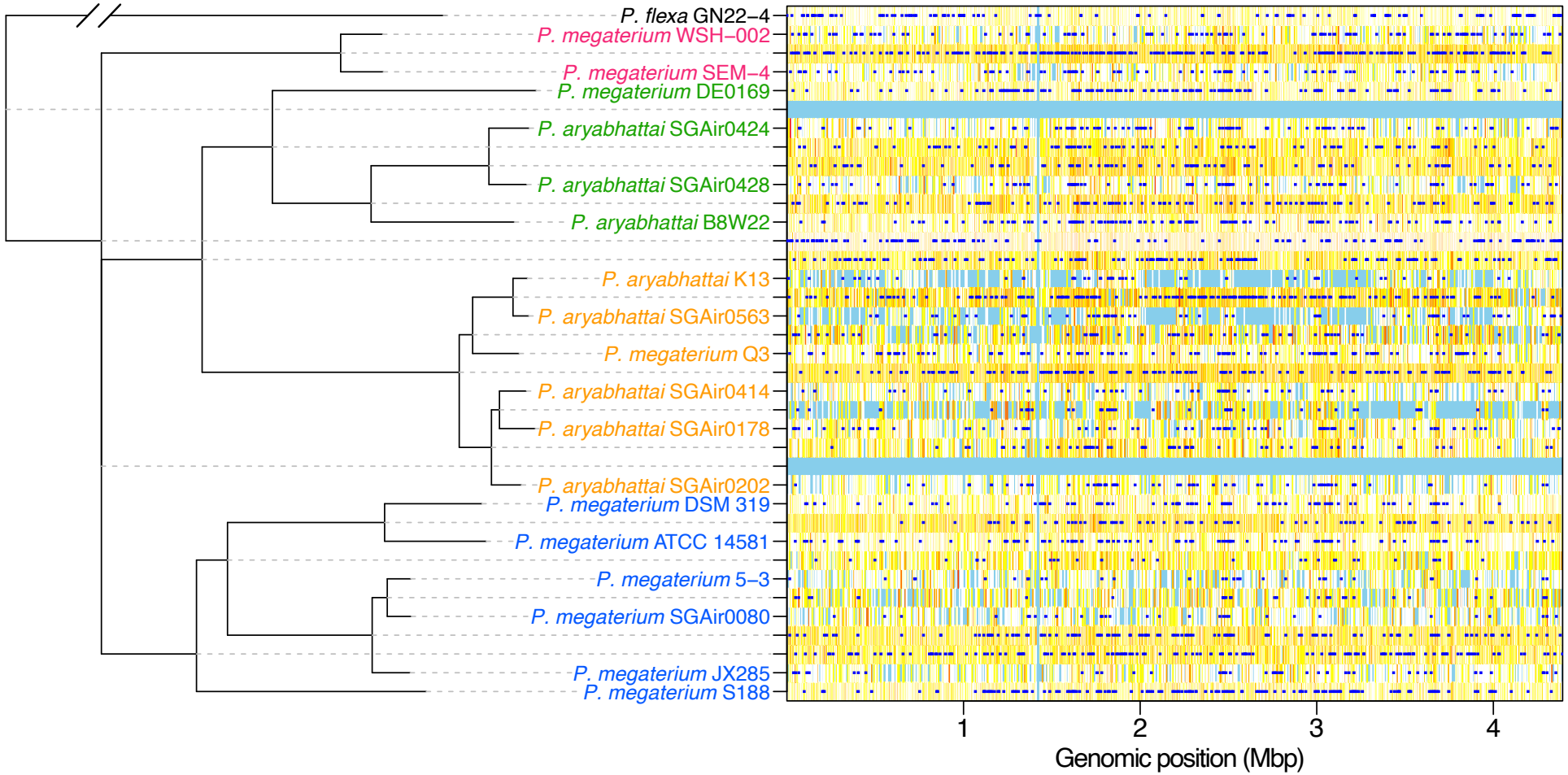


Figure 2-4: Clonal genome phylogeny of 18 genomes from the *P. megaterium/aryabhattai* group with a *P. flexa* outgroup. Genomic regions marked in dark blue are putative recombination events, and are excluded from the sequence used for building the tree. Sites in light blue have not changed since the last node up the tree. Other sites are coloured on a scale from white to yellow to orange based on the level of homoplasy shown. The tree places recombinant clade 2 outside of the *aryabhattai* and *megaterium* clades, and recombinant clade 1 in a group with the *aryabhattai* clade.

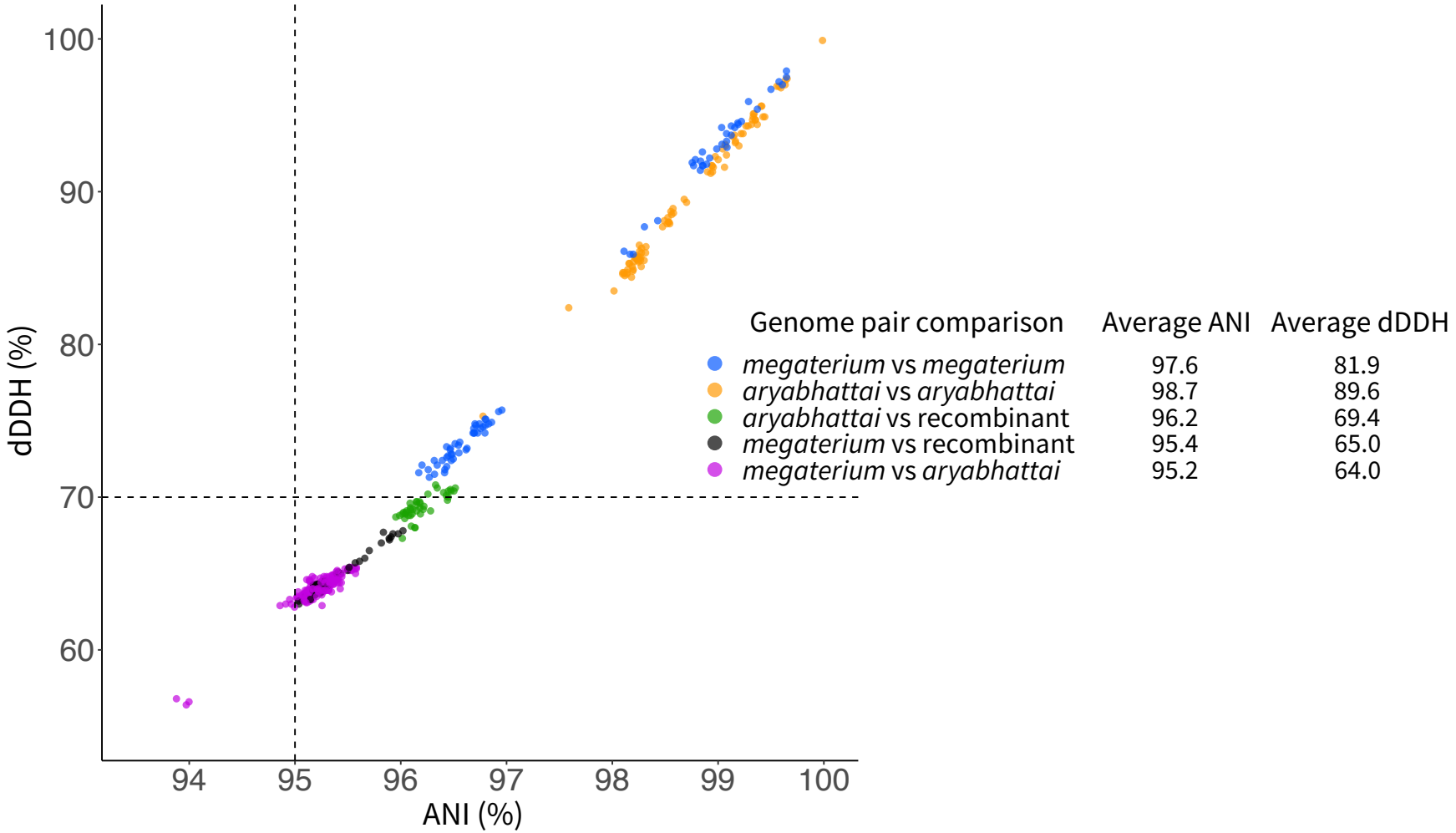


Figure 2-5: Pairwise dDDH and ANI values between *P. megaterium* and *P. aryabhattai* complete genomes (13 from *megaterium* clade, 13 from *aryabhattai* clade, 2 from recombinant clade 1, and 1 from recombinant clade 2). Dotted lines indicate the conventional same-species thresholds of 70% dDDH and 95% ANI. The dDDH values of comparisons between *P. megaterium* and *P. aryabhattai* are below 70% (purple, lower left), indicating different species, whereas the ANI results of such comparisons are ambiguously found to be usually higher but sometimes lower than the same-species threshold.

Table 2-2: Average pairwise ANI similarities and sequence alignment p-distances, within and between clades, for the 189 *P. megaterium/aryabhattai* genomes in Figure 2-1a-b.

Comparison between clades:		Average ANI value (% +/- st.dev.)	Average alignment p-distance (% +/- st.dev.)
megaterium	*megaterium*	97.33 +/- 1.23	0.78 +/- 0.48
	aryabhattai	95.25 +/- 0.18	1.92 +/- 0.07
	Recombinant 1	95.31 +/- 0.22	1.85 +/- 0.11
	Recombinant 2	95.69 +/- 0.28	1.62 +/- 0.13
aryabhattai	*aryabhattai*	98.63 +/- 0.39	0.24 +/- 0.09
	Recombinant 1	95.95 +/- 0.34	1.49 +/- 0.10
	Recombinant 2	96.24 +/- 0.15	1.89 +/- 0.05
Recombinant 1	Recombinant 1	97.46 +/- 1.20	0.87 +/- 0.51
	Recombinant 2	95.38 +/- 0.20	1.94 +/- 0.06
Recombinant 2	Recombinant 2	98.9 +/- 0.33	0.33 +/- 0.16

Discussion

I aimed to resolve the debate over the identity of *P. aryabhattai*; to determine whether it is a synonym of *P. megaterium* or a separate species. My high-resolution phylogenetic analysis using a vast whole-genome dataset demonstrates that the two bacteria names do in fact form distinct, monophyletic clades. The *aryabhattai* clade is smaller and less diverse than the *megaterium* clade, to which it connects by a long branch; these traits are suggestive of adaptation due to selection pressure. I also found that 24 out of 190 genomes appear to be recombinants between the two clades, based on their positions in the phylogenetic trees and their inconsistent gene content.

I constructed trees of the *P. megaterium/aryabhattai* group by five different combinations of data type (marker gene sequences or whole-genome ANI scores) and tree-building algorithm (maximum likelihood, neighbor-joining, and neighbor-net) in order to test the robustness of the clades shown. The results from each method were in perfect agreement on the members of each clade, if not on the branch structure within each clade, but each method comes with its own considerations.

Marker gene trees use genes that are already well-studied and are known to show enough variation on the interspecies level. The commonly used 16S rRNA gene, which has historically been used to define species as strains that are > 97% identical, has not diverged enough between *P. aryabhattai* and *P. megaterium* to reconstruct the clades on a tree. The use of this gene for phylogenetic reconstruction is also problematic when multiple non-identical copies exist in the same genome, as they do in *Priestia* genomes.

For these reasons I instead used a published set of marker genes (107) for phylogenetic tree construction. Using multiple concatenated gene sequences provides more polymorphic sites to separate the genomes on a tree, but not all marker genes were useful in discriminating the *Priestia* species (Figure 2-3d), and only 74 out of the set of 120 marker genes were present in the *Priestia* genomes. This method could thus be improved by carefully selecting the marker genes which are useful for one's species of interest; for example, by first identifying orthologs among the genomes (see Chapter 3) and then using the results to select marker genes which are common to all of the genomes under study for building trees, as in Liang et al. (122). However, using this procedure to select the marker genes for a species requires multiple whole genome

datasets for a comparative study, and the method is thus not economical or efficient for the identification of novel species with only one cultured strain.

ANI genome similarity scores have an advantage in that they include all of the sites that can be aligned between each genome pair, rather than a defined set of marker genes, thus providing far more data for the phylogenetic tree construction to distinguish between each genome. One issue that comes with the method is the difficulty of calculating bootstrap support in the same way as with smaller sequence alignments. In short, traditional bootstrapping for phylogenies from sequence alignments works by randomly sampling sites from the alignment and building the tree again from the sample to check if the tree branches are consistent when different samples are used (123). The ANI method also aligns sequences from genomes against each other in order to calculate an overall percentage similarity, but to the best of my knowledge no existing ANI software will provide the alignments as output to the user, and so the results of an ANI analysis cannot easily be bootstrapped in the same way. ANI calculations between 190 genomes are also far more computationally intensive than marker gene alignments.

Network trees are useful for the interpretation of more complex evolutionary relationships. When the data is ambiguous or contradictory, as in the case of the recombinant clades for which some marker genes are closer to *P. megaterium* sequences and some are more similar to *P. aryabhattai* sequences, a classical bifurcating tree must make some compromise and place the clade in one position or another. A network tree instead represents the conflicting data as a phylogenetic network where splits between taxa can be represented by parallel edges rather than single branches. In this study, the phylogenetic network (Figure 2-1c) shows the uncertainty of the

placement of the recombinant clades by the combined width of their parallel branches, with the distance to the *megaterium* and *aryabhattai* clades varying according to different parts of the dataset.

My results expand on the findings of the previous studies of the species identification of *P. aryabhattai* by showing that the average ANI between the two groups when considering all available genomes (95.2%) is lower than the value that was found by comparing a single pair of genomes (100). Additionally, the average dDDH value between species that was found in this study was 64%, which disagrees with Shivaji *et al.'s* DDH result of 35% (76) but is still lower than the 70% threshold. Finally, the phylogenetic trees using all available genomes showed four separate clades that were not apparent in the previous studies which compared one genome from each group (76,100).

During the writing of this thesis, a study was published that described a novel *P. megaterium* strain and also constructed a phylogenetic tree of 26 *P. aryabhattai* and *P. megaterium* genomes, and noted the presence of two separate clades (124). The present study uses all available genome assemblies – 189 *P. megaterium/aryabhattai* genomes – and shows that the complete phylogenetic tree contains four robust clades, not only two, and that the dDDH similarity between the *aryabhattai* and *megaterium* clades is below the 70% same-species threshold.

Table 2-3 lists 17 genome assemblies in GenBank which currently have species names which do not match the clade that they are placed in by this phylogenetic analysis. Most of these genomes are from large environmental sequencing projects from which many

assemblies were generated and given the same species name, when in fact some fall under the *megaterium* and some under the *aryabhattai* clades. The common practice of finding the closest database match by ANI to new isolates is much more likely to give an accurate answer for these two species than using only the 16S rRNA gene (Figure 2-1); however, knock-on errors are caused when existing database genomes are already misnamed. Our new *P. aryabhattai* genomes, whose closest database matches are strains named *P. megaterium* that belong in the *aryabhattai* clade, are prime examples (Table 2-1). This compounding problem of publishing misidentified species using a single identification method has already been discussed at least 20 years ago (70).

As an example of the confusion caused by species misidentification, a 2017 study (125) compared the gene content of eight *P. aryabhattai* strains to five *P. megaterium* strains, but used several strains in their comparison whose names do not match their clade according to my results: *P. aryabhattai* C765 was renamed to *megaterium* in 2018, *P. megaterium* Q3 consistently placed in the *aryabhattai* clade in my analysis, and *P. aryabhattai* AB211, *P. aryabhattai* B8W22, and *P. megaterium* WSH-002 were in the recombinant clades in my trees.

Table 2-3: Suggested renaming of *Priestia* GenBank assemblies whose names do not match the clade that they consistently placed in during this study.

Accession no	Current name	Suggested name
GCA_014932925.1	*Priestia aryabhattai* s1338	*Priestia megaterium* s1338
GCA_015845475.1	*Priestia aryabhattai* G25−109	*Priestia megaterium* G25−109
GCA_019748835.1	*Priestia aryabhattai* I1−B2	*Priestia megaterium* I1−B2
GCA_019748735.1	*Priestia aryabhattai* I1−P1	*Priestia megaterium* I1−P1
GCA_019748815.1	*Priestia aryabhattai* I1−P3	*Priestia megaterium* I1−P3
GCA_017743055.1	*Priestia aryabhattai* LAD	*Priestia megaterium* LAD
GCA_014138775.1	*Priestia aryabhattai* S00060	*Priestia megaterium* S00060
GCA_009497655.1	*Priestia megaterium* A	*Priestia aryabhattai* A
GCA_017086545.1	*Priestia megaterium* CDC2008724129	*Priestia aryabhattai* CDC2008724129
GCA_007678145.1	*Priestia megaterium* DE0183	*Priestia aryabhattai* DE0183
GCA_007677165.1	*Priestia megaterium* DE0260	*Priestia aryabhattai* DE0260
GCA_007674025.1	*Priestia megaterium* DE0315	*Priestia aryabhattai* DE0315
GCA_007673295.1	*Priestia megaterium* DE0377	*Priestia aryabhattai* DE0377
GCA_007672925.1	*Priestia megaterium* DE0399	*Priestia aryabhattai* DE0399
GCA_007672525.1	*Priestia megaterium* DE0420	*Priestia aryabhattai* DE0420
GCA_001050455.1	*Priestia megaterium* Q3	*Priestia aryabhattai* Q3
GCA_003072605.2	*Priestia megaterium* YC4−R4	*Priestia aryabhattai* YC4−R4

Another study from 2015 (126) isolated 395 strains from 55 species including seven *P. megaterium* and four *P. aryabhattai*, identified by 16S rRNA gene sequencing, and tested each species for plant growth promoting traits such as ion solubilisation and nitrogen fixation. However, my results have found that the 16S rRNA gene cannot be used to differentiate between strains of the two clades because there is no clustering of the two clades on the 16S tree (neither by GenBank assigned names nor by the clade names assigned in this study); the sequences of the two clades often have no nucleotide

differences between them, despite the differences across the rest of the genome. The study did not name the closest matching 16S rRNA sequence in the database that each strain was matched to. Further, the results of the plant growth promotion (PGP) traits for each strain were summarised using one representative strain for each species – the study does not describe whether the species were internally consistent or how accurately the representative strains can represent their species. This means that it is likely that the study used some *P. megaterium* and *P. aryabhattai* strains which were misidentified, and so the pattern of results in PGP phenotypes that they found between the two species may not be robust. Future studies comparing the two species would benefit from species identification using whole-genome sequencing or PCR primers such as the ones presented in this study (see Chapter 2).

This study echoes the call by previous authors for a modern polyphasic approach to bacterial taxonomy (127,128). In a recent example, a 16S rRNA gene tree of the *Rhodobacteraceae* group showed inadequate resolution to delineate most species because the 16S sequences were too conserved; instead, a polyphasic approach including protein alignments, similarities at each codon position, ANI, dDDH, and *in silico* predicted phenotypes provided enough resolution to split the group into two families and reclassify 327 species (122). Crucially, if an analysis is to decide if two groups are the same species or not, then the analysis must include multiple genomes from each group so that the tree structure can be assessed and the monophyly of each group can be determined.

It has also been previously reported that using the 16S rRNA gene alone for phylogenetic inference can be problematic because it is often present in multiple copies

with non-identical sequences in the same genome (129). For example, our new *P. aryabhattai* genomes have between 13 and 17 copies each, as annotated by Prokka (130). The use of other single copy marker genes avoids the issue of having to arbitrarily choose which 16S copy to use for analysis.

Whole-genome methods such as ANI should be included in a polyphasic taxonomy analysis, but they should not necessarily replace older methods such as phenotypic tests and marker gene alignment, because a more complete taxonomy of one's bacteria of interest is gained by using multiple techniques to complement each other. However, the genes and phenotypes that work well as taxonomic markers are likely to be different in other groups of bacteria (131) and should therefore be chosen specifically. For example, the *rpoB* gene has often been used in preference to the 16S rRNA gene for distinguishing *Bacillus* species (132,133) and also showed a strong phylogenetic signal in my results.

Problematically, there seems to be no standard methodology for comparing the phenotypes of *Priestia* isolates. The previous studies that compared *P. megaterium* to *P. aryabhattai* used different phenotype tests to each other and reached opposite conclusions on whether the two species are synonyms or not (76,100). We used an array of 61 phenotype tests to compare 13 isolates, but the test kit was unable to identify which clade any strain belonged to because no single test gave a consistent result for all members of one clade and a different result for the other clade. This may be because *P. megaterium* is a species with a large genome and a generalist lifestyle, that lives in diverse environments and may be capable of switching on or off the expression of some of the functions which have been tested for. Identifying the biochemical tests and gene

sequences that are useful for a particular bacterial genus may require deeper investigation into the traits that differ between the relevant species.

Since the dDDH scores between genome assemblies of *P. megaterium* and *P. aryabhattai* are always well below the threshold of 70%, and the two groups form robust, separate, monophyletic clades, they can be considered as separate species, but the low ANI distance between the clades and the existence of recombinant strains in natural samples isolated from various environments suggests that the speciation may still be ongoing. Nevertheless, the results of this study support the preservation of the two species names for the two clades, and I will therefore refer to both *P. megaterium* and *P. aryabhattai* as species throughout the rest of this thesis.

Chapter 3 — Gene content comparison of *Priestia megaterium* and *Priestia aryabhattai*

Introduction

From the results of Chapter 2, it can be seen that *Priestia aryabhattai* and *Priestia megaterium* appear to be undergoing a speciation event, but the cause of this ongoing divergence is unknown. Speciation may be caused by natural selection operating on particular genes of strains under different selective pressures, which can be related to different habitats or environmental changes, or by random genetic drift (42). Drift may occur in species that have limited dispersal capabilities between subpopulations that consequently show strong biogeographic patterns – for example, obligate pathogens and symbionts which can only survive within a host and which experience population bottlenecks during transmission between hosts (134). *Priestia* species are able to produce resilient spores, and their dispersal appears to be relatively unconstrained. The phylogenetic trees of *Priestia* genome assemblies (Chapter 2) showed a lack of biogeographic structure, with strains sampled from different global regions placed on adjacent tree tips, and no clades that corresponded to specific countries or regions (Figure 3-1). Therefore, the speciation between *P. aryabhattai* and *P. megaterium* may be driven by natural selection to adapt to the environment.

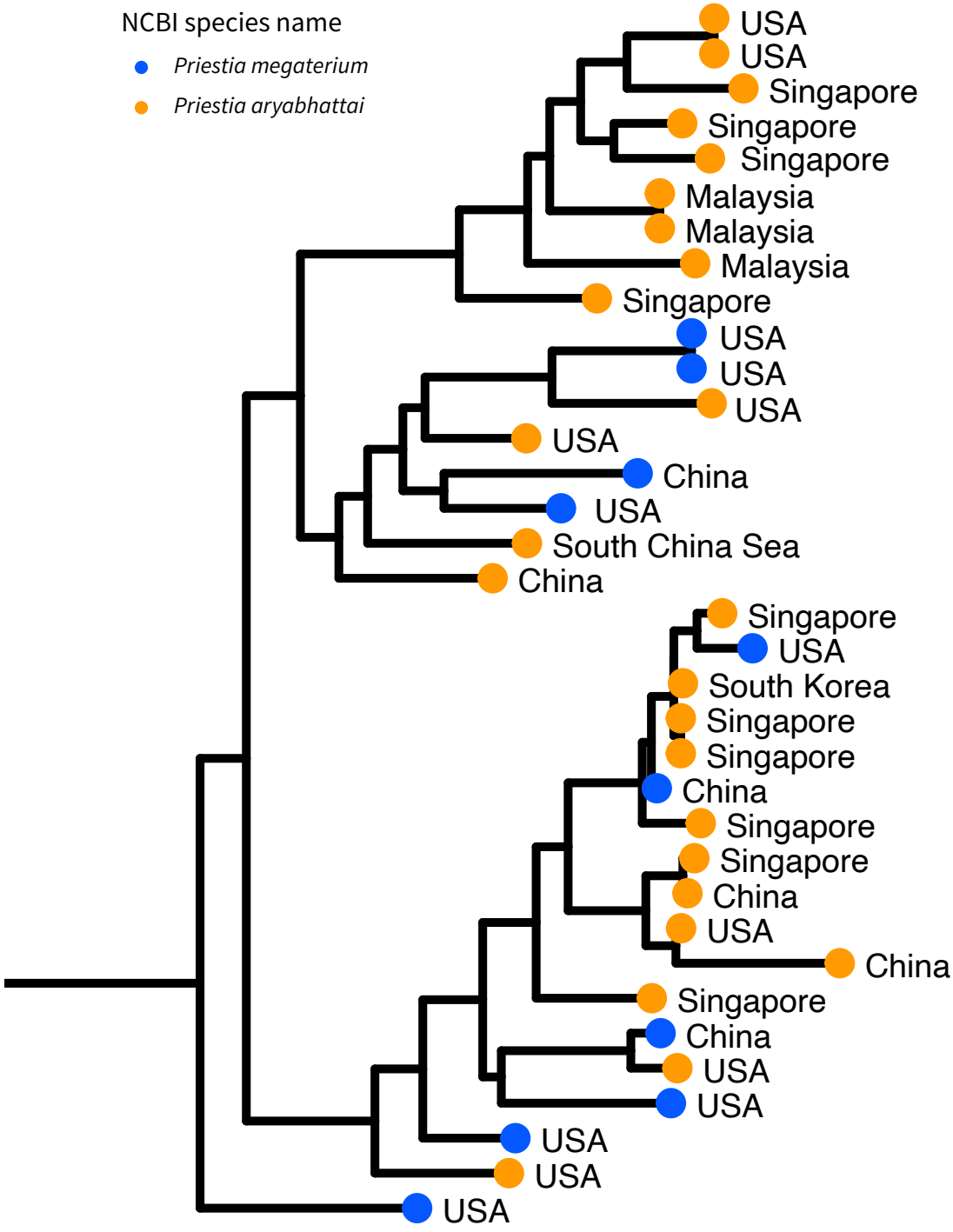


Figure 3-1: Zoomed view of the *aryabhattai* clade from Figure 2-1. Tips are labelled with the country from which the sample was taken. A biogeographic clustering effect, in which strains from nearby locations are more genetically similar to each other than to strains from other locations, cannot be discerned. Strains on the tree are globally mixed, with Singapore, China, and USA samples spread throughout all subgroups.

The available biochemical tests on the strains' phenotypes (Table A-2) were unable to identify any consistent functional differences between the two species. If the speciation is due to adaptation to a new ecological niche, rather than random drift, then identifying the genes which have undergone the most changes between the two species can provide clues as to what that new niche is. To investigate the evolutionary split further, I therefore looked for differences in gene content between the two groups.

Although 95% ANI is considered the standard species threshold, there is precedent for bacterial groups which are more than 95% similar to nonetheless be considered as separate species if they contain group-specific genes with important functions. A good example is *Bacillus cereus* sensu lato, another member of the family *Bacillaceae*, which contains several closely related groups that are up to 98% similar to each other by ANI (Table 3-1), but with each having unique toxin genes which define their roles as pathogens of different animal hosts (135).

It is the genes that are shared but different among the two species that are the actual cause of the lower ANI values between the species (Table 2-2), since the ANI method works by first aligning the genome sections which are homologous and then counting the number of sites which are identical and different for each alignment. The genes which are only found in one species have no effect on these genome similarity scores. It is therefore possible for bacteria which are extremely closely related by ANI to have unique, species-defining functions.

Table 3-1: Whole-genome ANI scores (%) between pairs of *Bacillus cereus* sensu lato strains, which have important functional differences between the *anthracis*, *cereus* and *thuringiensis* groups despite their high genome similarity. Three genomes of each species were downloaded from GenBank and compared using FastANI.

		Bacillus cereus anthracis			*Bacillus cereus cereus*			*Bacillus cereus thuringiensis*		
		CMF9	Ames ancestor	Vollum	30075	ATCC 14579	FORC 047	IMBL-B9	KF1	LX43
Bacillus cereus anthracis	CMF9	-								
	Ames ancestor	97.40	-							
	Vollum	97.39	99.98	-						
Bacillus cereus cereus	30075	97.23	97.14	97.19	-					
	ATCC 14579	91.72	91.82	91.80	91.79	-				
	FORC 047	91.44	91.54	91.57	91.40	96.91	-			
Bacillus cereus thuringiensis	IMBL-B9	91.32	91.23	91.28	91.35	96.97	96.02	-		
	KF1	94.03	94.26	94.32	93.94	91.81	91.74	91.45	-	
	LX43	91.75	91.92	91.96	91.84	98.39	96.46	96.50	91.98	-

The key question that motivates this Chapter is whether there are functional differences between *P. aryabhattai* and *P. megaterium*. I therefore analysed the gene content of the two species to look for species-specific genes and changes in protein sequences. The total set of genes that are present in a group of genomes are described as part of the core genome or accessory genome, depending on how many genomes each gene is present in.

The core genome consists of the genes which are found in all strains in a group (136). Some authors also consider a relaxed core genome of genes found in at least 95% of the genomes in a group, rather than only the strict core genome of genes in 100% of genomes (137). In this study I examined the strict core genome because the objective was to find differences between all *P. megaterium* strains and all *P. aryabhattai* strains. The accessory genome is made up of the genes which are present in at least one genome but are not in the core genome (138). Genes which are found in exactly one genome are also known as singletons (139,140). The entire set of genes found in the group of genomes, including the core genome and the accessory genome, is known as the pangenome (141).

This Chapter focuses on the core genome of the two species, assuming that a gene driving the speciation would logically be present in all members of at least one of the species. Identifying the core genome requires first identifying the genes from each strain's genome which are homologous to each other – genes which were present in the common ancestor of two strains and so are now found in both descendants.

Homologs are classified as orthologs or paralogs depending on how they originated (142). Orthologs are copies of a gene in different genomes that were separated by a speciation event. Paralogs are copies of a gene, in the same or different genomes, that were created by a gene duplication event. In the example diagram below (Figure 3-2), each species has two copies of the gene. We can see that the pair of red copies (or equally, the pair of green copies) are orthologs by tracing their lineages back to when they split during the speciation. When we compare a red copy to a green copy, we can see that they split from each other as a gene duplication within the same genome and are therefore paralogs. Identifying orthologs and differentiating them from paralogs is important for comparing the gene content of the two species. The presence or absence of orthologs in modern genomes shows whether the same ancestral function is still present in both species, or whether one species has lost the gene over time. In contrast, paralogs that have been created by gene duplication are more likely to diverge in function over time (143).

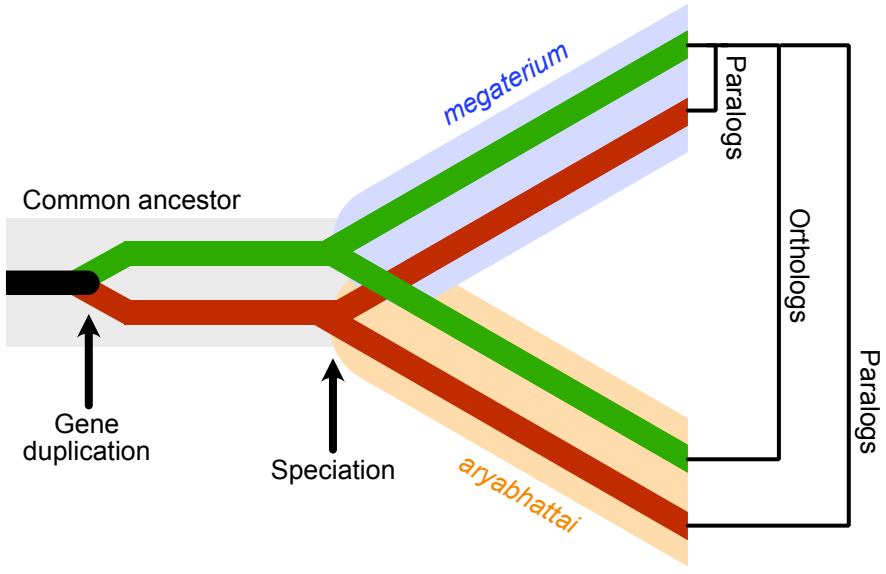


Figure 3-2: Illustration of paralogous and orthologous genes. Paralogs result from the duplication of an ancestral gene within the same genome, whereas orthologs originate from the divergence of two lineages, such as speciation events.

To investigate the differences in gene content between the *P. megaterium* and *P. aryabhattai* species, I annotated the genes present in each individual genome and then clustered the genes from all genomes into orthologous groups using a reciprocal best hits method. I showed differences between the two species both in the presence/absence of certain orthologs in each species, and also in the sequences of orthologs which are present in both species but with significant changes in protein sequence.

As an additional benefit of this analysis, the identification of genes which are specific to each species will also allow for the design of PCR primers which can be used to quickly identify new isolates as being in the *P. megaterium* or *P. aryabhattai* species. Currently, researchers who culture new strains of *P. megaterium* or *P. aryabhattai* cannot be certain which species their isolate belongs to without investing the time and money to generate a whole-genome sequence. As discussed in Chapter 1, the common method of sequencing only the 16S rRNA gene is unreliable for these closely related bacteria. Instead, as a result of the analyses in this Chapter, I will identify species-specific genes and use them as a target for PCR primers for identifying these species.

Methods

Orthologous gene clustering

I first identified the genes which are orthologous between the *P. megaterium* and *P. aryabhattai* species, and then analysed the differences in gene content between the two species using orthologs under two cases:

1. Orthologs which are found in genomes of both species, but with differences in the amino acid sequences between the two species.
2. Species-specific orthologs, that are present in genomes from one species but are not found in genomes of the other species.

The orthologs in case 2, which were found in all members of one species but none of the other, were also used to design PCR primers for the identification of the two species.

Identifying the genes which are orthologous between genomes is complicated by the fact that a single gene can be present in more than one copy in a genome. To overcome this issue, a reciprocal best hits method can be used to cluster the genes of each genome into orthologous groups. This method compares the genes of each genome to every other gene in each other genome, in an all-vs-all manner, to find the sets of closely matching genes from different genomes which are more closely related to each other than to any other genes. Each set, or orthologous gene group, is a single protein-coding gene which was present in an ancestral genome and then copied into multiple descendants – the genomes which are now under study.

From the 190 genomes used for the phylogenetic trees in Chapter 2, I selected the 18 genomes which were listed in the GenBank database as assembled at the 'Complete genome' level – no sequence gaps in the bacterial chromosome, no more than nine consecutive unknown bases, and no unplaced scaffolds (144) – plus the 12 *Priestia* genomes that were generated from Singapore air samples. The resulting dataset contained 30 high-quality genomes, of which 13 were *P. megaterium*, 14 were

P. aryabhattai, two were from recombinant clade 1, and one was from recombinant clade 2, according to the phylogenetic trees (Figure 2-1). Each genome was annotated using Prokka (130) to find the protein-coding sequences. Having found the set of genes from each genomes, SwiftOrtho (145) was used to compare the genes in an all-vs-all manner and group them into sets of orthologous genes from different genomes, using a reciprocal best hits method. A custom python script was used to compare the output of SwiftOrtho against the complete annotated genomes and identify the unique genes which were only present in one genome and had no homologs.

Pan-genome accumulation curve

The number of genes (orthologs and singleton genes unique to individual genomes) in the pan-genome and core genome was calculated for the 30 genomes above using a custom R script. In this process, the 30 genomes were added to the dataset one by one in a random order and the pan/core genome was counted at each step. This was repeated 100 times to calculate the mean number of genes in the pan/core genome for each number of included genomes from one to 30. The pangenome curve was also calculated for only the 14 *P. aryabhattai* genomes and for only the 13 *P. megaterium* genomes. Exponential models of the form $y = ax^b$ were fitted to the pangenome curves by using the optim function in R to minimise the residual sum of squares error.

Calculation of dN and dS

After identifying the sets of orthologs among the 30 *Priestia* genomes, I created an analysis pipeline using python scripts that compares the gene sequences within each ortholog group and calculates the degree of sequence divergence between each genome for that ortholog.

In the first step, the genes in each ortholog group were aligned against each other using MUSCLE (108). Alignments of both DNA and amino acid sequences were made and then used to create codon alignments for each ortholog group using Biopython (146). Codon alignments use the protein sequence to identify the position within a codon of each base in the DNA alignment, so that substitutions between sequences can be identified as synonymous or non-synonymous. Maximum-likelihood trees were then constructed from the alignments using FastTree 2.1.1 (147).

The codon alignments were then used to calculate dN and dS, which are the number of nonsynonymous substitutions per nonsynonymous site (the proportion of substitutions that change the protein, out of those that could be made), and the number of synonymous substitutions per synonymous site (those that would not change the protein) (148). These were calculated by running MEGA version 10.1.7 (149) by command line from within the python scripts, using the Nei-Gojobori proportion method with uniform rates among sites and pairwise deletion of gaps.

In order to identify the orthologs with the greatest sequence divergence between *P. megaterium* and *P. aryabhattai*, the dN and dS results for each ortholog were

calculated for ortholog pairs from the same species and for ortholog pairs from opposite species. The average dN for between-species pairs was used to rank orthologs by their divergence between species, whilst the dN and dS for within-species pairs were used to filter for genes which are relatively conserved within each species. This is because an ortholog which evolves quickly, with conspecific sequences showing many substitutions, would show a high dN between species but would not be useful for finding differences between the two species.

Each strain was identified as either *P. megaterium* or *P. aryabhattai*, or as belonging to one of the recombinant clades, according to the results of the phylogenetic analysis in Chapter 1. For each alignment of orthologous genes, the genetic distance metrics listed above were used to rank the orthologs by their divergence between species as follows:

1. Orthologs which were found in fewer than 95% of genomes of either species were excluded.
2. Orthologs where the average pairwise dN within either of the two species was equal to or greater than 0.15, or the average pairwise dS within either species was equal to or greater than 0.175, were excluded. The gene trees made from orthologs with higher dN or dS values than these thresholds were unable to recreate the two species clades (as shown in Chapter 1) due to lack of sequence conservation within species.

3. The remaining 4197 orthologs were then ranked by a dN score:

$$dN\ score = \mathrm{dN(betw)} - (\mathrm{dN(mega)} + \mathrm{dN(arya)})$$

Where, for an ortholog with sequences present in both species:

dN(betw) is the average dN of pairs of sequences from opposite species

dN(mega) is the average dN of pairs of sequences both from *P. megaterium*

dN(arya*)* is the average dN of pairs of sequences both from *P. aryabhattai*

Each orthologous gene group was also annotated using the Clusters of Orthologous Groups (COG) database (150) and the KEGG Orthology database (151), providing additional information on gene functions on top of the Prokka annotation.

Whole genome synteny plots

One genome from each of the four clades identified in Figure 2-1 was aligned against the reference genome assemblies for *P. aryabhattai* and *P. megaterium*. The reference genome assemblies used were the *P. aryabhattai* K13 and *P. megaterium* ATCC 14581. Genomes were selected for having few, well assembled contigs and a high N50. The whole genome alignments were performed using progressiveMauve 2.4.0 (115) and the results were visualised with the R package genoPlotR (152).

Clustering of genomes by gene content

The gene content of the 30 genomes was used to create phylogenetic trees, in order to compare the tree structure of the gene content trees to those created using sequence data (Figure 2-1). Using R 4.3.0 (153), the presence/absence of each of the 9,338

orthologs in each of the 30 genomes was used to calculate the Euclidean distance between each genome. Hierarchical clustering was then used with the complete linkage, single linkage, UPGMA, Ward's clustering, and McQuitty clustering methods to group the genomes into clusters using the intergenomic distances. The same Euclidean distance matrix was also used to construct a midpoint-rooted neighbor-joining tree with the package ape v5.7-1 (111).

PCR

The PCR primers were designed with Primer3Plus (154) using sequences from adjacent species-specific genes. The target sequences were searched for in all available *P. megaterium* and *P. aryabhattai* assemblies in the GenBank database using BLASTN (155) to confirm their specificity. The sample preparation and PCR were performed by collaborators in the Singapore Centre for Life Sciences Engineering.

Table 3-2: PCR primers used to detect strains of the species *P. megaterium* and *P. aryabhattai*. The primers for *P. megaterium* target a polyamine aminopropyltransferase gene followed by two hypothetical proteins. The *P. aryabhattai* primers target a small, acid-soluble spore protein gamma type and the alcohol dehydrogenase gene and unknown protein that flank it.

Species	Orientation	Primer sequence
Priestia megaterium	Forward	TCCGTGCATCATTTGTTTTACCT
	Reverse	TGGATCCATTTTAAAGCCTCCT
Priestia aryabhattai	Forward	GTCTTTGGTGCTAAAGTTATTGCA
	Reverse	GCTGGACCAACAAGTGATAACT

Bacterial culture and DNA template preparation

The isolate strains identified as *P. aryabhattai* (SGAir0178, 0179, 0202, 0257, 0265, 0269, 0414, 0425, 0427, 0563) were cultured in 5 mL tryptic soy broth at 30 ˚C overnight. The resulting cultures (5 mL) were centrifuged at 6000 x g for 10 min, and the pellets were suspended in the lysis buffer supplied in the Dneasy PowerWater DNA isolation kit (Qiagen, Germany). Genomic DNA was then extracted as described by Gusareva *et al.* (156). Strain SGAir0427 genomic DNA was extracted using the Wizard Genomic DNA Purification kit (Promega, USA). Following quantitation using Qubit (Invitrogen, USA), all genomic DNA was diluted to 10 ng/µL in ultrapure water for PCR.

PCR conditions

The primers used for this experiment are listed in Table 3-2. The reaction mixture for PCR (25 µL) consisted of 1x KAPA HiFi ready mix, 0.3 µM forward primer, 0.3 µM reverse primer, and 2.5 µL of diluted DNA (25 ng DNA). The PCR conditions were as follows: initial denaturation at 94 ˚C for 3 min, 30 cycles of denaturation at 98 ˚C for 30 sec; annealing at 50 or 58 ˚C for 20 sec; extension at 72 ˚C for 1 min, and final extension at 72 ˚C for 7 min. Annealing temperatures of 50 ˚C were sufficient for the *P. aryabhattai* primers. The annealing temperature was increased to 58 ˚C for the *P. megaterium* primers to remove off-target amplifications in the *aryabhattai* strains.

PCR product purification

The PCR products were purified with PureLink PCR Purification Kit (Invitrogen, Germany) by following company's suggested protocol. All the purified PCR product was

submitted to Sanger sequencing. The sequenced PCR products were aligned against the targeted genomic regions using MEGAX (149) to confirm that the correct region had been successfully amplified.

Results

Pan-genome accumulation curve

After identifying the orthologs among the *Priestia* genomes, the pangenome curves for each species were modelled in order to compare the genetic diversity of *P. megaterium* and *P. aryabhattai*. The combined 30 *Priestia* genomes had an average of 5,730 genes each. As each of the 30 genomes were added, the core genome decreased in size to a minimum of 4,055 genes, and the pan-genome increased to a total of 14,037 genes, including 9,338 ortholog groups and 4,699 strain-specific genes. These singleton genes, which were unique to individual genomes, were thus a substantial fraction (33%) of the pangenome for these 30 strains. The number of singletons per genome ranged from 42 to 351 with a mean of 156.6.

Genomes in the *megaterium* clade had between 5,048 and 6,567 orthologs, with a mean of 5,577. The gene content of the *aryabhattai* clade ranged from 5,134 to 6,446 with a mean of 5,548. No phylogenetic clustering of high or low gene content genomes was observed.

The curves that were fitted to the pangenome sizes for different numbers of genomes had exponential terms between 0.2 and 0.3, and showed no sign of reaching an asymptote in either species, indicating that both species have open pangenomes with unsampled diversity (157). The *P. megaterium* pan-genome showed a steeper and higher curve than the *P. aryabhattai* pangenome, showing that the *P. megaterium* species is not only more diverse in the number of sequenced genomes and in nucleotide substitutions (Chapter 2), but also in gene content.

Core genome analysis

The pangenome curve showed that the combined core genome of the two species contained 4,117 orthologs (Figure 3-3). The Venn diagram in Figure 3-4 shows the number of orthologs (excluding singleton genes, which are unique to one genome) that are unique to each species and shared by both species. The top row includes all core and accessory genes, with a total pangenome size of 9,270 excluding the recombinant clades. The middle row is the relaxed core genome of genes which are found in at least 95% of genomes. The bottom row is the strict core genome of orthologs which are found in every genome in each species: 4,117 orthologs were found in every genome, with 46 *megaterium*-specific and 21 *aryabhattai*-specific orthologs.

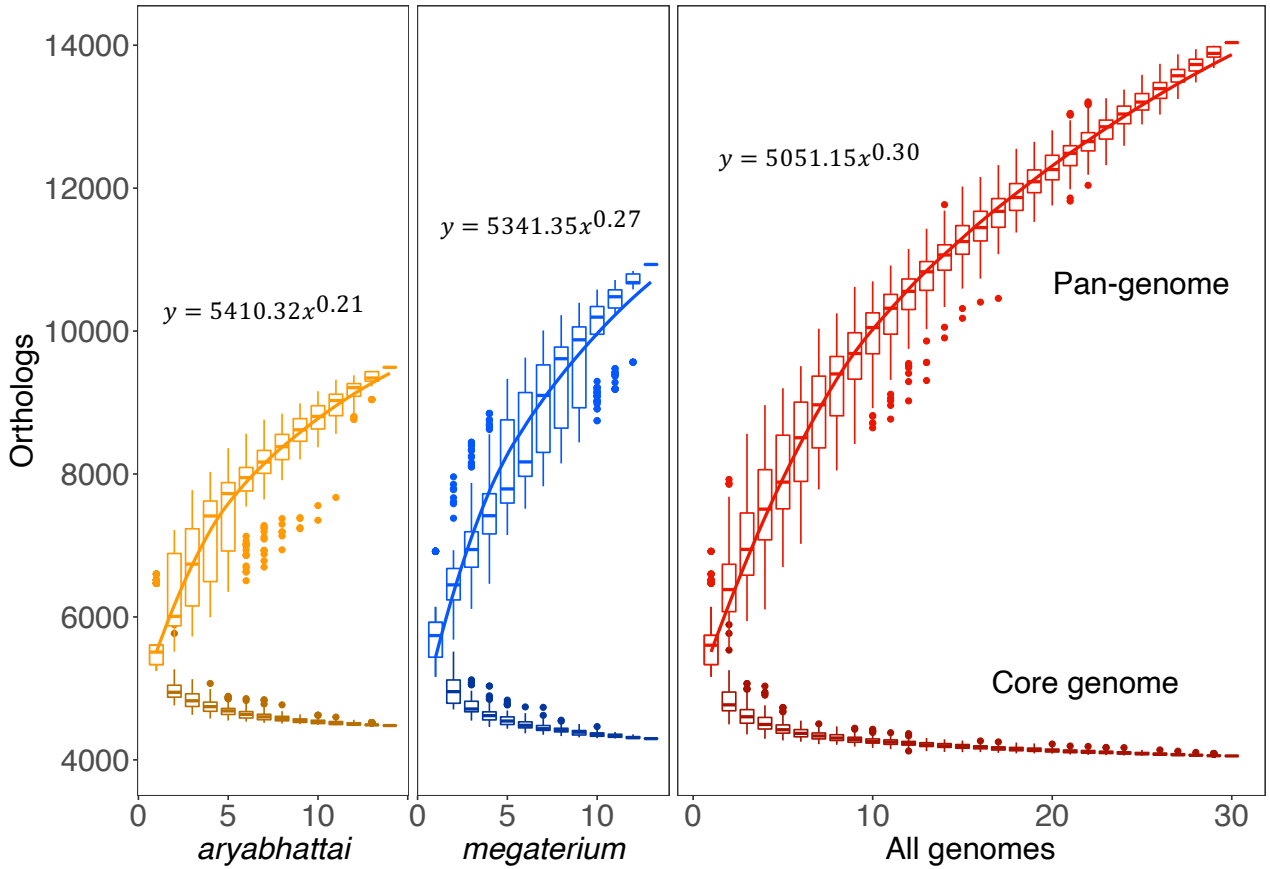


Figure 3-3: Pan-genome accumulation curves for 14 *P. aryabhattai* genomes (orange), 13 *P. megaterium* genomes (blue), and 30 combined *Priestia* genomes (red), including the previous 27 genomes plus two from recombinant clade 1 and one from recombinant clade 2 (Figure 2-1). The plots show the changing number of shared orthologs as genomes are added in a random order, with 100 repeats. The upper plots are the pangenome, including all orthologs present in at least one genome. The lower plots show the core genome of orthologs which are common to all genomes. The exponents of the fitted curves, being greater than 0, indicate an open pangenome. The steeper and higher pangenome curve for *P. megaterium* than for *P. aryabhattai* shows the greater diversity in gene content of the megaterium species per genome included.

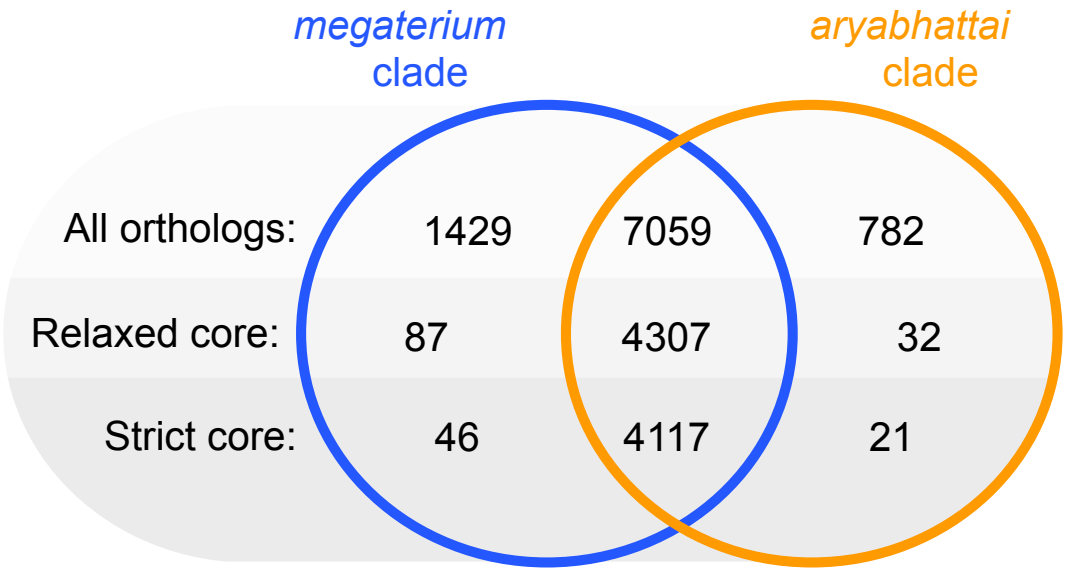


Figure 3-4: Core genome and pangenome sizes of 30 genomes from *P. megaterium* and *P. aryabhattai*, excluding singleton genes which are unique to one genome. Top row: the total number of core and accessory genes which are unique to each clade and shared between them. Middle row: the number of genes that are present in 95% of genomes in each clade. Bottom row: the number of genes which are found in every genome (13 and 14 respectively) of each clade.

Orthologs exclusive to P. megaterium and to P. aryabhattai

From the ortholog clustering results I identified 46 orthologous genes which were present in every *P. megaterium* genome but absent from all *P. aryabhattai* genomes. Similarly, 21 genes that were found in all *P. aryabhattai* genomes were missing from every *P. megaterium* genome (Table A-4). About half of these species-specific genes are found adjacent to each other in the genome in several blocks of genes up to 10kb in length (Table A-4). The consistent presence and absence of the gene blocks in the *P. megaterium* and *P. aryabhattai* strains was confirmed by BLASTN searches for each gene block in all 190 *Priestia* genomes that were used in the phylogenetic study. Genomes in the 'recombinant' groups often contained some of these gene blocks from both the *megaterium* clade and the *aryabhattai* clade, with no clear pattern.

The *megaterium*-exclusive orthologs included some that were present in multiple copies per genome, with each *P. megaterium* genome having one more ortholog group than *P. aryabhattai*, such as *speE* aminopropyltransferase, *murF* ligase, *ilvB* synthase, *aroD*/*aroQ* Shikimate kinases, *aroE* Shikimate dehydrogenase, *garP* galactarate transporter, *yngG* lyrase, *yedA* transporter, *lysN* transaminase, *sspH* small acid-soluble spore protein H, and *yofA* transcriptional regulator. Of particular note was *iolG* inositol 2-dehydrogenase, which was present in an average of six copies in *megaterium* and only two in *aryabhattai*. Three genes were unique to the megaterium group: *phnW* 2-aminoethylphosphonate--pyruvate transaminase, which has been linked to protection from hydroperoxide in *Pseudomonas aeruginosa* (158); *safD* sulfoacetaldehyde dehydrogenase, which is used for the metabolism of taurine as a nitrogen source (159); and phosphate-starvation-induced *psiE,* which is a transmembrane protein with unknown function.

Similarly, the *aryabhattai*-exclusive orthologs included some genes with homologs in *P. megaterium*. The *sspE* gene – small acid-soluble spore protein gamma type – was present in only one copy in most *P. megaterium* genomes, but the *P. aryabhattai* genomes each contained two or three orthologs of this gene. Small acid-soluble spore proteins are transcribed during spore formation to protect the spore's DNA from damage by UV radiation and other environmental stresses (160), and create a stockpile of mRNA molecules to be broken down for new RNA synthesis when the spore later germinates (161). The six *sspE* ortholog groups found in the 30 *Priestia* genomes were each very different in sequence, in accordance with previous studies that found that the gamma type SASP genes are not well conserved across species (162). Other orthologs with higher copy numbers in *P. aryabhattai* genomes were membrane protein insertase

misCA (synonym: *yidC*) and cadmium/cobalt/zinc antiporter *czcD*, which protects the cell by exporting heavy metals (163) including iron (164).

Additionally, several orthologs that are associated with iron import had a higher copy number in *P. aryabhattai* genomes than in *P. megaterium* genomes. These were identified by Prokka and COG as iron permease *efeU*, iron transporter *feoA*, ABC siderophore transport permease *yfiZ*/*fepD*, iron dicitrate permease *fecD*, iron citrate binding protein *yfmC*/*fecB*, and iron transporter *feoB*. Several of these genes have previously been reported to be associated with plant growth promotion by rhizosphere bacteria due to the benefit of increasing the host plant's iron uptake (165–171).

PCR test for discriminating P. megaterium and P. aryabhattai

To facilitate the identification and research of strains from these species, I aimed to provide PCR primers that can unambiguously place any new strain into one of the two similar species, without the need for whole-genome sequencing. The adjacent gene blocks found by homologous gene clustering are opportune targets for such a PCR test (Figure 3-5). By using primers that straddle two or more of the adjacent genes, it can be determined that the PCR reaction correctly amplifies the adjacent gene block that is exclusive to the relevant *Priestia* species, rather than any single gene in the block that may have homologs elsewhere in the genomes of the other species.

Here I provide two pairs of PCR primers (Table 3-2); one pair for identifying *P. megaterium*, and another pair for identifying *P. aryabhattai.* Researchers can identify

which species their strain belongs to by running two PCRs, once with each pair of primers. Strains of each species should not react to the other species' primers.

Three blocks of adjacent genes were identified which are specific to the *aryabhattai* species, and eight such blocks that are only found in *P. megaterium* genomes (Table A-4). I used the largest gene block from each species as PCR targets (Figure 3-5). The *P. megaterium* primers cover an unknown gene (predicted as a hypothetical protein by Prokka) and approximately half of the *speE* (polyamine aminopropyltransferase) gene and another unknown gene that flank it. The *P. aryabhattai* primers target the spore coat gene *sasP-B* and the surrounding genes – *adhT* alcohol dehydrogenase and an unknown gene. The gel electrophoresis image shows the successful identification of four *P. megaterium* and ten *P. aryabhattai* strains using these PCR primers (Figure 3-5). The strains tested for the amplifications included three strains from culture collections and 13 isolates from Singapore.

The *P. aryabhattai* primers produced a faint band when tested on strains from the *megaterium* clade (Figure 3-5d). This PCR product was re-amplified to create enough material for Sanger sequencing, which showed that they were non-specific amplifications that could not be aligned to the target genomic region. In the same photograph, the isolate SGAir0427 (*aryabhattai* clade) produced a slightly larger PCR product than the other *aryabhattai* clade isolates. When sequenced, this PCR product had a 127 bp insertion in the intergenic region between *sasP-B* and the unidentified gene that was not present in the other isolates. The PCR primers were therefore accurate in discriminating isolates from the *aryabhattai* and *megaterium* clades.

The PCR primers were also tested on two isolates from recombinant clade 1: SGAir0424 and SGAir0428. SGAir0424 showed a clear band with only the *aryabhattai* primers whilst SGAir0428 produced a positive result with both the *aryabhattai* primers and the *megaterium* primers. The sequenced product for both of these isolates had the same 127 bp insertion as SGAir0427.

Orthologs with sequence divergence between P. *megaterium and* P. *aryabhattai*

In order to identify the set of orthologs with the greatest sequence divergence between species, I developed a metric, designated as the dN score, that used the average pairwise dN between the two species minus the average pairwise dN within each of the two species. The top 100 identified orthologs by this dN score are given in Table A-5, with names as identified by Prokka and UniProt. A large number of the top 100 are unidentified proteins with closest database matches of less than 90% identity.

The list of orthologs that have diverged between the species includes multiple which are related to the synthesis and use of cobalamin, a.k.a. vitamin B12. These include the *cobS*, *cobD*, and *cobU* genes from the cobalamin synthesis pathway (Figure 3-13) and a vitamin B12-dependent ribonucleotide reductase. Other orthologs that have diverged between the species include seven GNAT family acetyltransferases, three components of the PTS sugar transport system, three genes related to spore formation and germination, two prophages, two Shikimate kinases, and three genes related to flagellum formation. The details of these genes can be found in the Discussion section of this Chapter.

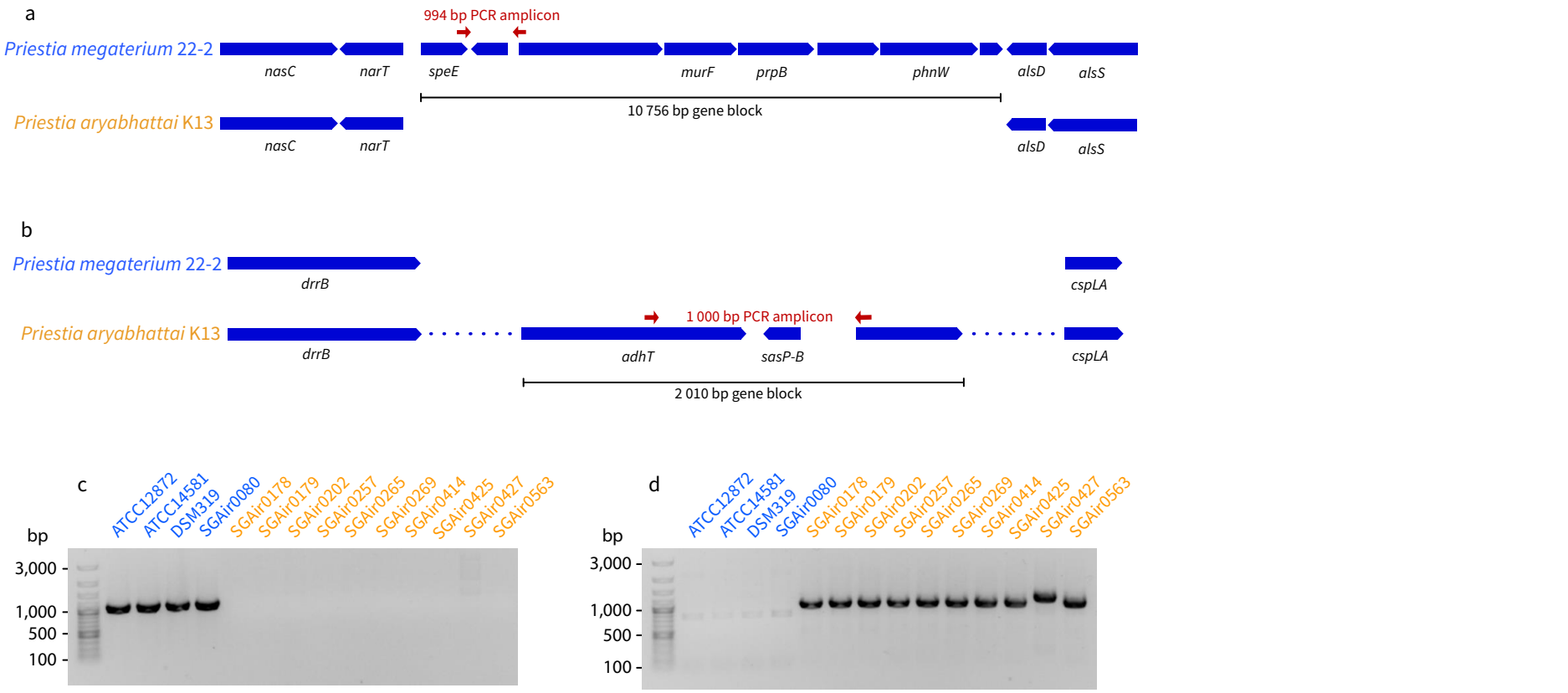


Figure 3-5: PCR test to detect strains of *P. megaterium* and *P. aryabhattai*, targeting clade-specific orthologs. **a, b:** Examples of blocks of adjacent genes which are exclusive to *P. megaterium* and *P. aryabhattai* (see Table A-4 for full list). Red arrows indicate PCR primers (Table 3-2) which target these genes in order to identify which species a genome belongs to. Unlabelled genes are those with unknown function. **a:** 10.8kb sequence only found in *P. megaterium* genomes, containing eight genes. The *narT* and *alsD* genes are consistently found on either side of the block. **b:** 2kb sequence found only in *P. aryabhattai* genomes, containing three genes. The genomic region in which the block is found seems prone to rearrangement, with the flanking genes being inconsistent. **c, d:** results of PCR experiments using the primers from **a** and **b** to successfully amplify only the genomes from the relevant species. **c:** primers targeting *P. megaterium* but not *P. aryabhattai*. **d:** primers targeting only *P. aryabhattai* and not *P. megaterium*. Blue genome names are designated *P. megaterium* on GenBank, orange names are called *P. aryabhattai*.

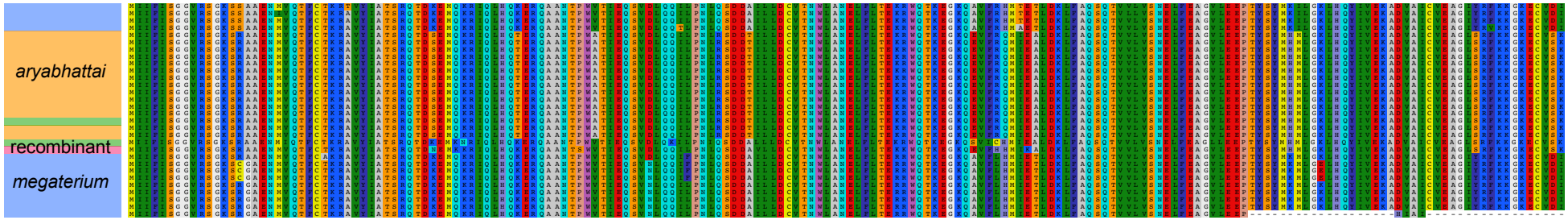


Figure 3-6: Amino acid sequence alignment of the cobalamin synthesis gene cobU from 13 *P. megaterium* genomes, 14 *P. aryabhattai* genomes, and three genomes designated as recombinant genomes between the two clades. The orange, blue and green blocks to the left of the alignment show the species that each sequence belongs to; sequences are grouped vertically by a Neighbor-Joining tree (tree structure not shown). The average dN between *P. aryabhattai* and *P. megaterium* sequences is 0.049.

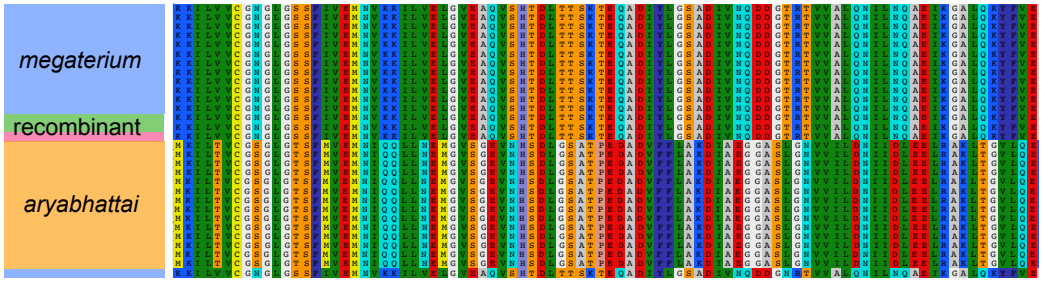


Figure 3-7: Amino acid sequence alignment of the ascorbate-specific PTS system EIIB subunit from 13 *P. megaterium* genomes, 14 *P. aryabhattai* genomes, and three genomes designated as recombinant genomes between the two clades. The orange, blue and green blocks to the left of the alignment show the species that each sequence belongs to; sequences are grouped vertically by a Neighbor-Joining tree (tree structure not shown). The average dN between *P. aryabhattai* and *P. megaterium* sequences is 0.040.

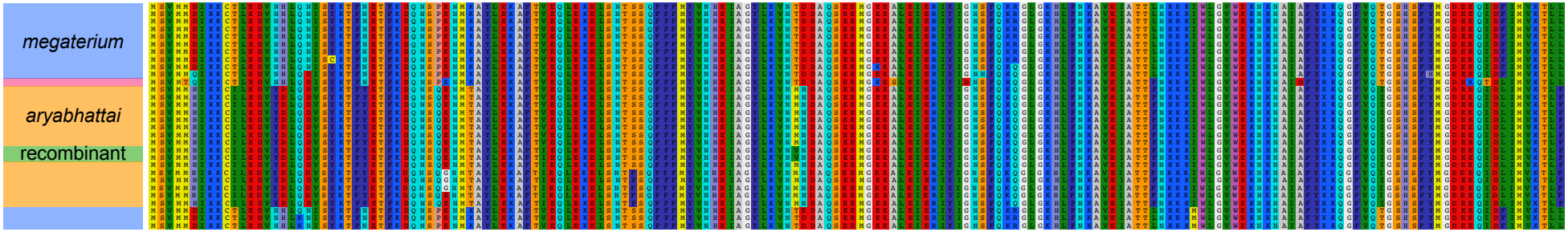


Figure 3-8: Amino acid sequence alignment of paiA spermidine/spermine-N1-acetyltransferase and sporulation negative regulatory gene from 13 *P. megaterium* genomes, 14 *P. aryabhattai* genomes, and three genomes designated as recombinant genomes between the two clades. The orange, blue and green blocks to the left of the alignment show the species that each sequence belongs to; sequences are grouped vertically by a Neighbor-Joining tree (tree structure not shown). The average dN between *P. aryabhattai* and *P. megaterium* sequences is 0.043.

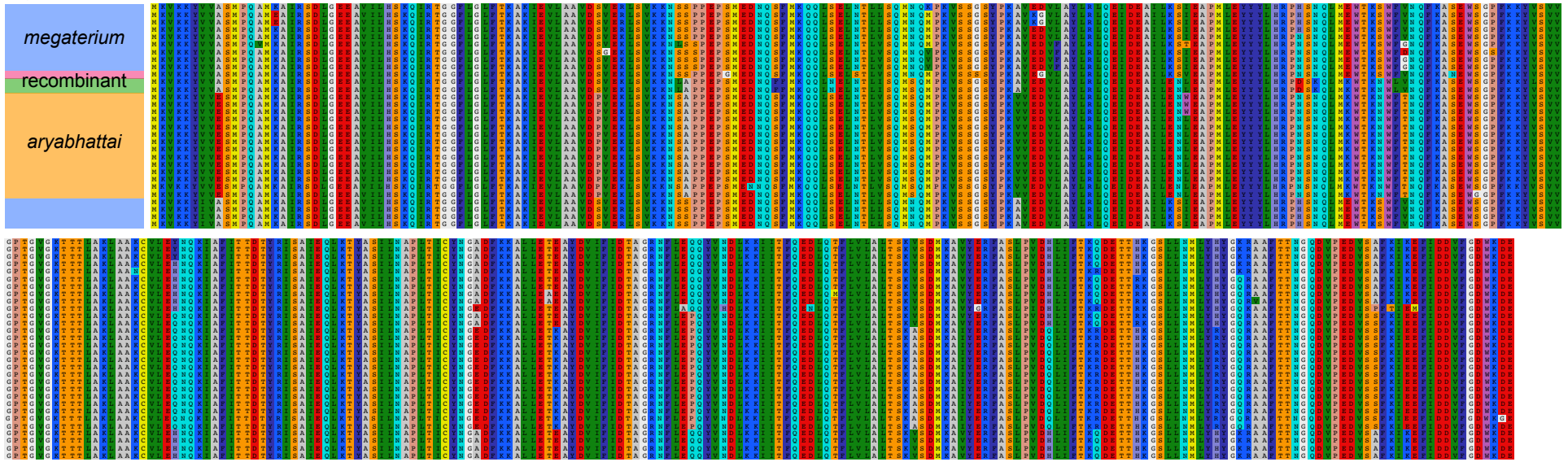


Figure 3-9: Amino acid sequence alignment of flagellar biosynthesis gene flhF from 13 *P. megaterium* genomes, 14 *P. aryabhattai* genomes, and three genomes designated as recombinant genomes between the two clades. The orange, blue and green blocks to the left of the alignment show the species that each sequence belongs to; sequences are grouped vertically by a Neighbor-Joining tree (tree structure not shown). The average dN between *P. aryabhattai* and *P. megaterium* sequences is 0.035.

Alignment of whole genomes to reference genomes

A genome from each of the four clades was aligned against a reference genome of *P. aryabhattai* and of *P. megaterium* in order to assess the level of synteny between clades and the similarity of the recombinant clades to the *aryabhattai* and *megaterium* clades. The results (Figure 3-11) showed high identity between genomes within the *megaterium* clade and within the *aryabhattai* clade, as expected, and also a high level of synteny between these two clades. The recombinant clades both showed a loss of this synteny, suggesting that they cannot be ancestral to the larger group. These clades also showed a mosaic pattern of genomic segments with higher identity to either the *megaterium* or *aryabhattai* reference genome, suggesting extensive horizontal gene transfer between these groups.

Nucleotide identity of orthologs to reference genomes

The common orthologs between one genome from each clade were aligned against those from the reference genomes of *P. megaterium* and *P. aryabhattai*, in order to determine whether the coding sequences from the recombinant clades were more similar to those of the *megaterium* or *aryabhattai* clade. The graph (Figure 3-10) shows the number of orthologs that were closer to each reference for each clade. As expected, most of the orthologs from the *megaterium* and *aryabhattai* genomes showed higher similarity to their respective reference genome. The recombinant clade genomes showed a more even ratio of orthologs, with over one third showing closer identity to *megaterium* and the rest being closer to *aryabhattai*.

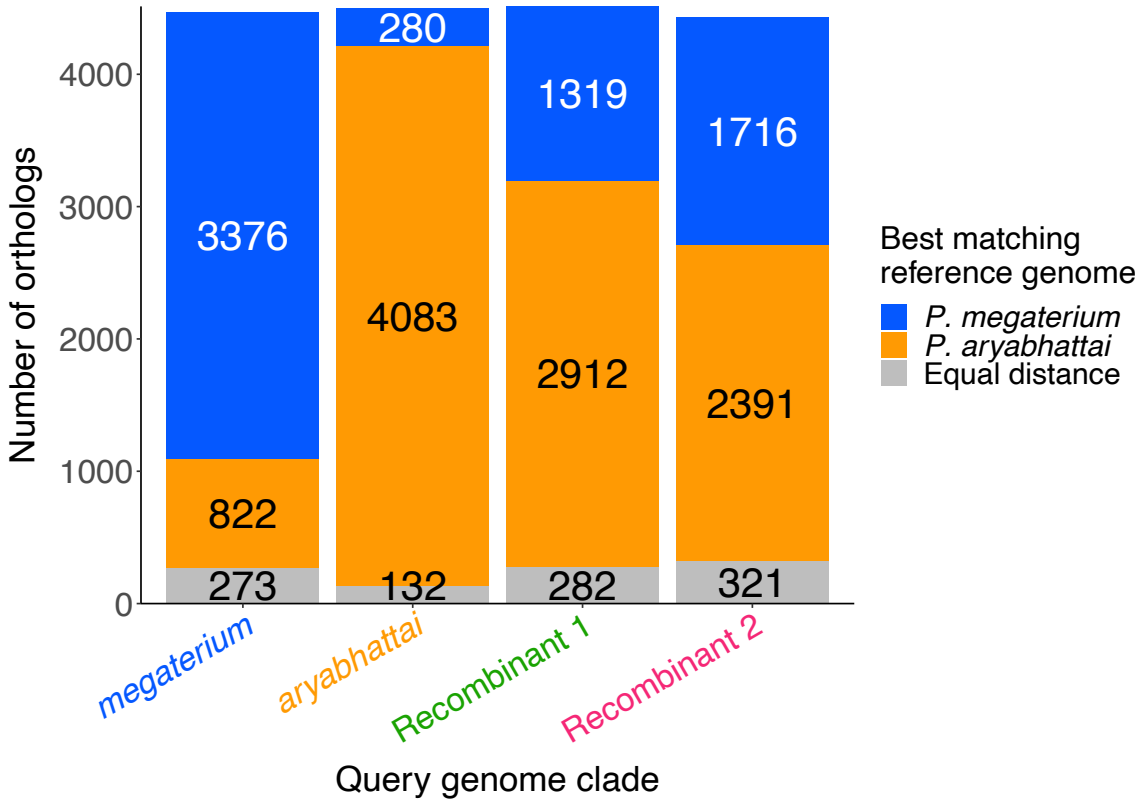


Figure 3-10: Nucleotide identity of orthologs from each of the four clades identified in Figure 2-1 to the reference genomes of *P. aryabhattai* and *P. megaterium*. For each clade, the stacked bars show the number of orthologs from one genome which had higher similarity to the *P. megaterium* reference or to the *P. aryabhattai* reference, or equal similarity to both references. As expected, most orthologs from the *megaterium* and *aryabhattai* clades are more similar to their corresponding reference genome. The recombinant clades showed more even proportions of ortholog identities, with over half of orthologs being closer to *P. aryabhattai* but also a large proportion that matches more closely to *P. aryabhattai*.

Clustering of genomes by gene content

In order to test for internal consistency of the gene content within the clades, cladograms were recreated by clustering algorithms on gene content data (Figure 3-12). The distance measure used to determine distance between genomes had no effect on the clustering found. Six different clustering algorithms produced minor variations in the trees. Each method successfully recreated the larger *megaterium* and *aryabhattai* clades, showing that these clades have their own core genomes which are more internally similar than they are to those of other clades. However, three *aryabhattai* genomes (*P. aryabhattai* SGAIr0202, SGAir0425, SGAir0563) and one *megaterium*

genome (*P. megaterium* FDU301) were consistently placed in an outside clade due to their unusual gene content, despite their robust position within their respective clades when building trees from sequence data (Figure 2-1).

The results also showed that the gene content of the genome from recombinant clade 2 was more similar to the *aryabhattai* clade, placing the genome consistently within this clade. The gene content of the genomes from recombinant clade 1 was more difficult to define by similarity, and was placed either within the *megaterium* clade or between the *megaterium* and *aryabhattai* clades by the different clustering algorithms.

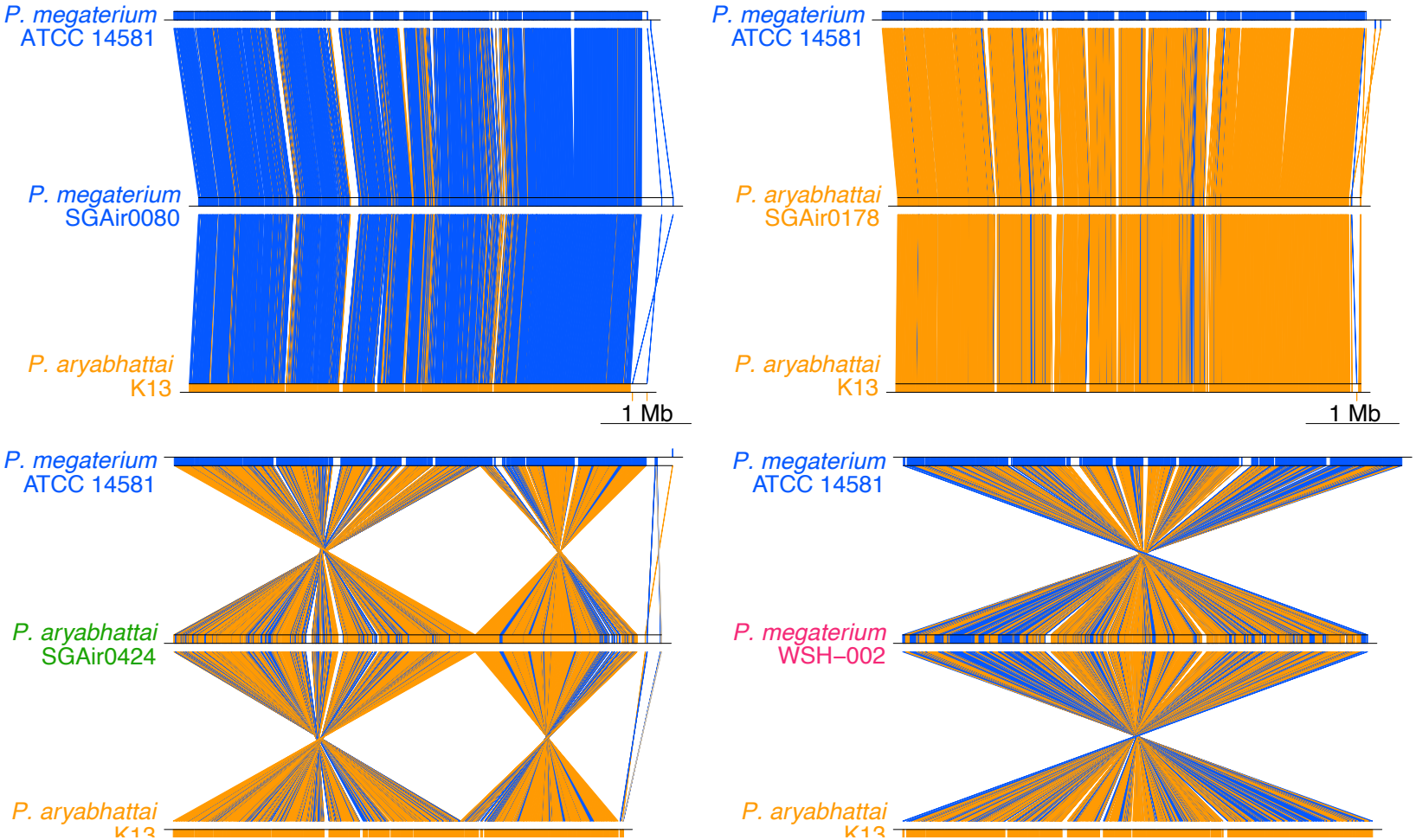


Figure 3-11: Whole genome alignments of one genome from each of the four clades identified in Figure 2-1 (centre of each plot) to reference genomes of *P. aryabhattai* and *P. megaterium* (top and bottom of each plot). Blue genome segments and connecting lines have higher similarity to the *P. megaterium* reference; orange segments have higher similarity to the *P. aryabhattai* reference. The recombinant clade genomes show mosaics of sections with higher identity to both reference genomes, and loss of the clear synteny that is present between the *megaterium* and *aryabhattai* clades.

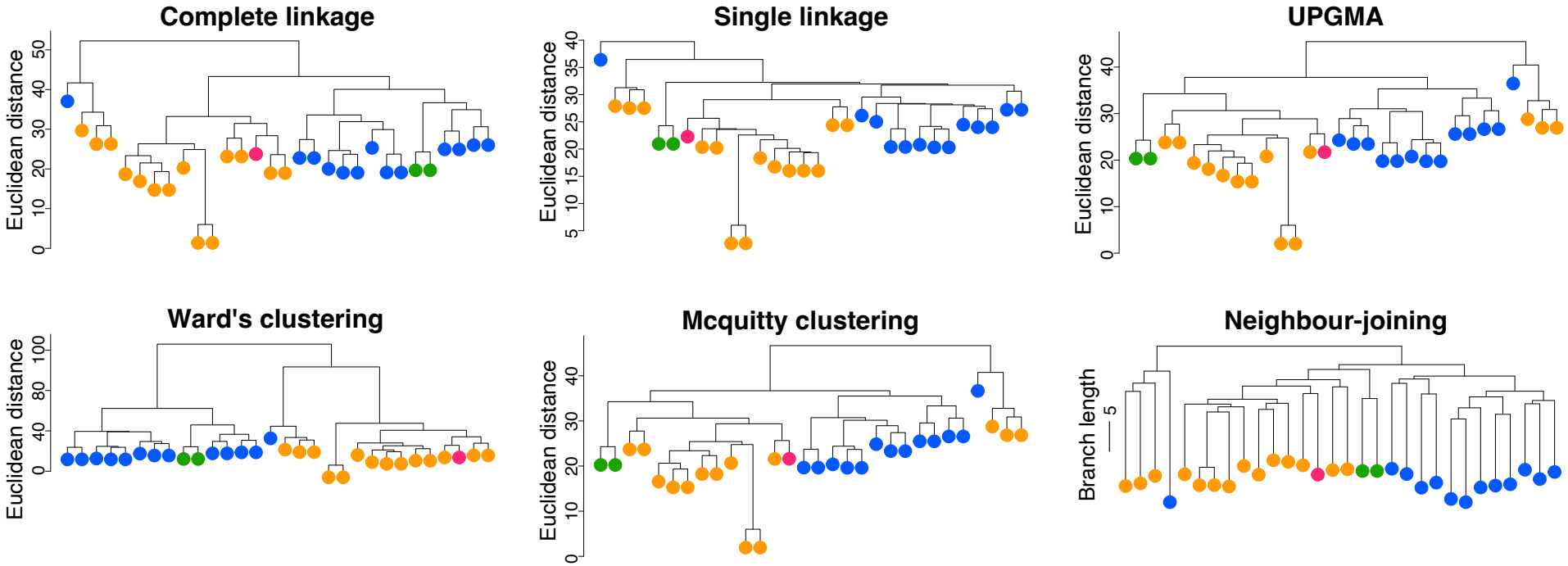


Figure 3-12: Hierarchical clustering on the presence/absence of 9,338 orthologs in 30 *P. aryabhattai/megaterium* genomes. The results of six different clustering algorithms are shown. Genomes are coloured according to the clade in which they were found in Figure 2-1: blue, *megaterium*; orange, *aryabhattai*, green, recombinant clade 1, pink, recombinant clade 2. The gene content tree largely recreates the *megaterium* and *aryabhattai* clades, but some strains with unusual gene content form a group outside the larger clades. Each clustering method places the genome from recombinant clade 2 within the *aryabhattai* clade, but the position of the recombinant clade 1 genomes varies.

Discussion

Ortholog clustering by the reciprocal best hits approach

I used a reciprocal best hits (RBH) analysis to identify sets of orthologous genes among the *Priestia* genomes. This method is used ubiquitously in the field of comparative genomics but there has been some debate as to whether it can reliably find orthologs rather than paralogs, another type of homologous gene.

The RBH method is purported to identify sets of orthologs and not paralogs. This is because orthologs are thought to evolve more slowly and conserve their function more than paralogs do, because when a gene duplication occurs, a redundant copy of the gene is created which is free to accumulate mutations and even changes of function without affecting the function of the other copy (172). Thus, paralogs are expected to diverge from each other, but orthologs between closely related genomes should be more conserved.

Modern orthology inference tools, such as SwiftOrtho (145) which was used in this study, extend the RBH method to identify not just ortholog relationships but also paralogs. After identifying genes from different genomes as orthologs by RBH, each individual gene in the ortholog group recruits other genes from their own genome into the ortholog group as inparalogs. If a gene in the ortholog group has another gene in its own genome that it matches to better than to its orthologs in other genomes, the new gene is added to the ortholog group. It is inferred to be an inparalog to the closely

matched gene from the same genome, and a co-ortholog to any such inparalogs in the other genomes. This extended RBH method allows for genes with complex evolutionary histories and multiple copies to be classified using three different relationship types, improving on the simple RBH method, and the relationships inferred are in strong agreement with reference gene phylogenies that were determined with classical tree-based methods (145).

The present study focuses on functional comparisons between *P. megaterium* and *P. aryabhattai* rather than the phylogenetic history of each ortholog group. Genes of note were investigated along with their inparalogs in order to account for the presence of copy number variation between the two species, without making conclusions on the evolutionary history of each ortholog.

Pangenome curve modelling

The models fitted to the pangenome accumulation curves for the 30 available high-quality *P. aryabhattai* and *P. megaterium* genomes had exponents from 0.21 to 0.3, depending on the strains included, and showed no sign of flattening out at 30 genomes (Figure 3-3). These results show clearly that each new strain adds previously unseen genes and that the species possesses yet more diverse orthologs and singleton genes that were not seen with our dataset of the 30 complete genomes. To give perspective, the number of ortholog groups was 9,338, and the number of singleton genes was 4,699, meaning that for the 30 genomes, over half of the pangenome consisted of genes that were unique to one genome only. The analysis of singleton genes was outside the scope of this study but researchers who work on individual strains of *P. megaterium* or

P. aryabhattai would do well to identify their strain's unique genes and how they are of benefit to the strain in its local environment. The core genomes and pangenomes that were identified in this study can contribute to the identification of such singletons.

The idea of finding the exponent of a model $y = ax^b$ where $0 < b < 1$ to define an open or closed pangenome comes from Heap's Law, which has previously been used to explore the increasing number of unique words as a text document's length increases (173). The concept was adapted to describe pangenomes by Tettelin *et al.* (157), who suggested that the pangenome is closed when $b \approx 0$ and the rate of discovery of new genes as more genomes is added becomes insignificant.

In addition to this model of increasing pangenome size as genomes are added, their paper also described an equivalent model which gives the number of additional genes that are added per new genome as $y = x^{(\gamma-1)} = x^{-\alpha}$ where $\alpha = 1 - \gamma$. With this model, the pangenome is considered closed if $\alpha > 1$ as the curve flattens out, or open if $\alpha \leq 1$ where the number of new genes discovered per genome is still decreasing. Such models can also be extrapolated to predict the expected size of the final pangenome when enough genomes have been sequenced that the rate of new gene discoveries becomes close to zero, but the low number of high-quality *Priestia* genomes and the steep curve at 30 genomes would make such predictions difficult to perform accurately for these species.

Clustering of genomes by gene content

An open question in bacterial genomics is whether the pangenome can be used to measure the genetic cohesion of a species, and hence whether it is possible to construct accurate phylogenies using gene content data (21). By definition, a gene that is a singleton cannot be used for the clustering of strains, since it is not shared by any set of genomes, but the presence or absence of genes in the core and accessory genomes can be used to construct cladograms. The results of constructing the *megaterium/aryabhattai* group phylogeny from gene content data (Figure 3-12) showed that the method was less accurate than sequence-based methods for this group of organisms; some genomes which were reliably shown to be within the *megaterium* and *aryabhattai* clades by gene sequences and whole-genome distances were instead placed outside of these clades when clustering by gene content. The positions of the recombinant clades in the tree were also different from those that were found using the clonal genome phylogeny (Figure 2-4).

A recent study on *Bacillus* and Streptomycetaceae genomes that despite frequent horizontal gene transfer, family-level phylogenies constructed from gene presence/absence were in broad agreement with core gene sequences and ANI, but not 16S rRNA (174). However, a similar study found no correlation between ANI and percentage of shared genes for genomes within the same species (1). The method of delineating species by shared gene content may be more accurate for well-separated species than for fuzzy species with mosaic genomes of mixed gene content, but is also limited by the need for abundant high-quality whole genome sequences.

Whole-genome alignment

The results of Chapter 2 showed that the two minor clades in the *megaterium/aryabhattai* group had intermediate positions in the group's phylogeny, which were inconsistent between the gene sequences that were used for the analysis. When calculating whole-genome distances, these minor clades were also intermediate between the *aryabhattai* and *megaterium* clades, at around 95.5—96.5% ANI. The clonal genome analysis indicated that strains in the *megaterium/aryabhattai* group have regions that have been potentially affected by recombination throughout their genomes. Further, the genomic regions which were present in only one of the *megaterium* or *aryabhattai* clades were not consistent in the minor clades, with some but not all of these genomes containing clade-specific orthologs from either *megaterium* or *aryabhattai*. Based on these results, I hypothesised that the small, intermediate clades were likely to be mosaic genomes that had been produced by extensive recombination between the clades.

To check for mosaic genomes as evidence for recombination, I tested the regional similarity across genomes from different clades, by whole-genome alignments and also by alignment of coding sequences. These analyses showed that the genomic regions (Figure 3-11) and coding sequences (Figure 3-10) from the *megaterium* and *aryabhattai* groups had higher similarity to the reference genomes from the same clades, as expected, and that these two clades had a high degree of synteny to each other. However, the genomes from the recombinant clades showed a more even proportion of regions and coding sequences that were more similar to each reference genome. This result would also be consistent with the hypothesis that the recombinant clades are

closer to the ancestral genomes of the larger group, causing them to be equally distant to both larger clades which diverged from them. However, this idea is contradicted by the loss of synteny in the recombinant clades that was present in the *megaterium* and *aryabhattai* clades. This suggests that the *megaterium* and *aryabhattai* clades resulted from an earlier split, and that the recombinant clades were formed later by horizontal transfer between the larger clades.

However, the specific regions which have been affected by recombination between these clades remain unknown. Future investigators may use explicit methods of horizontal gene transfer detection in order to identify specific events and the genes transferred within this lineage and from external sources (175–177). The extensive recombination within these clades also caused difficulties in designing PCR primers which react specifically to one clade's strains, as genomic regions that are thought to be exclusive to a clade may be found in a genome from the recombinant clades. A better understanding of the genomic regions which are more prone to recombination in this group would aid primer design by allowing a search for regions which are free of horizontal transfers, but still contain clade-specific genes, if such regions exist.

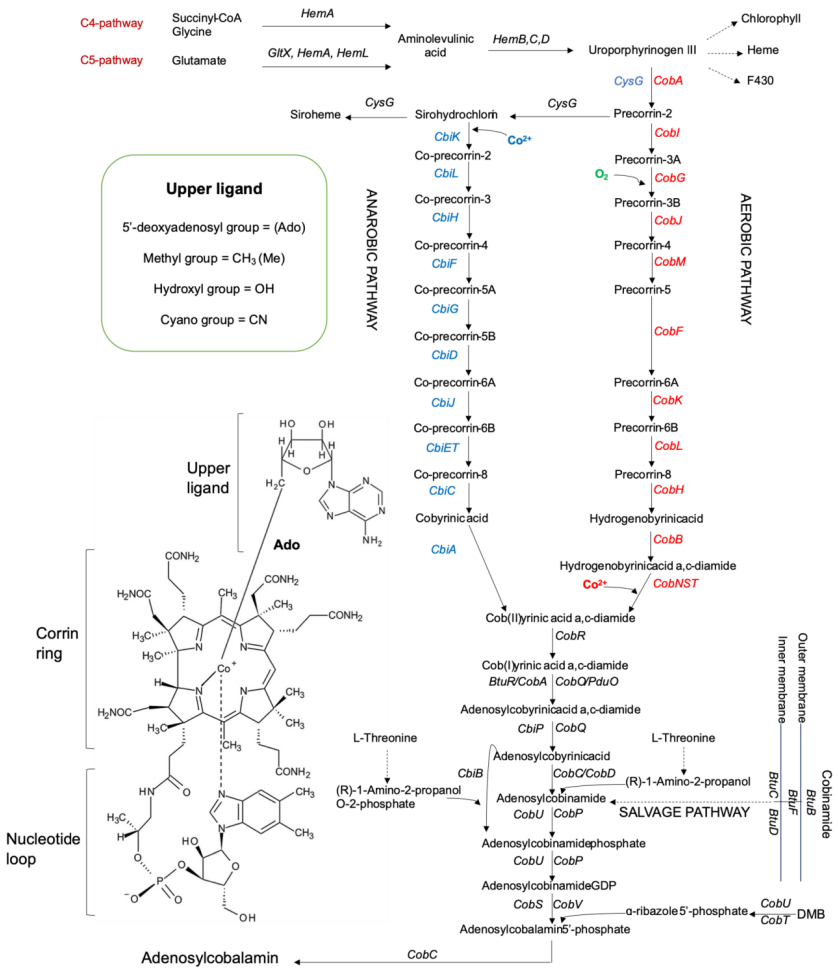


Figure 3-13: Adenosylcobalamin synthesis pathway, reproduced from Balabanova et al. 2021. The red genes (top right) show the aerobic pathway of *Paracoccus denitrificans*; the blue genes (top centre) show the anaerobic pathway of *P. megaterium*. Several genes near the end of the pathway show amino acid substitutions between *P. megaterium* and *P. aryabhattai*: *cobD*, *cobU*, and *cobS*.

Orthologs with high dN score between species

Cobalamin synthesis pathway genes

The results of calculating the dN score for the orthologs in the shared core genome of the two species names showed that, among the orthologs that are most highly divergent between the species, several are involved in the synthesis of adenosylcobalamin a.k.a vitamin B12. Strains of *P. megaterium* and *P. aryabhattai* use the anaerobic pathway for cobalamin synthesis (Figure 3-13).

Four consecutive steps near the end of the pathway are catalysed by the three proteins CobD, CobU (Figure 3-6), and CobS. When the 4,197 core genome orthologs were ranked by the average dN between species, the genes encoding these proteins were ranked 374th (dN score = 0.011), 31st (dN score = 0.049), and 17th (dN score = 0.049). *P. megaterium* has in fact become an important species for research on the synthesis of the complex B12 molecule and for producing it on an industrial scale (86,178,179); if manufactured abiotically, vitamin B12 would require a chemical synthesis pathway of about 70 different steps (180).

The selective benefits of changes to the B12 synthesis pathway are difficult to speculate on because vitamin B12 has many uses in nature, acting as a coenzyme for numerous multipurpose enzymes (87). The functions of B12-dependent enzymes range from essential methionine synthesis (181), to inducing gene expression in response to light (182), to the bioremediation of anthropogenic estrogen pollution (183). As a further complication, B12 is a community resource that is produced *de novo* by just over a third of prokaryotes but is needed by almost all (184), as well as by eukaryotes including algae (185) and humans (186). Thus, the benefit to *P. aryabhattai* of the substitutions in the B12 synthesis genes could be related to an interaction with another organism and may not become apparent when studying *P. aryabhattai* alone.

The sequence substitutions in multiple B12 synthesis genes in *P. megaterium* and *P. aryabhattai* warrant further experiments to compare the structure of the cobalamin produced by each. Alternatively, rather than changing the physical product of the pathway, the effects of these genetic changes may instead relate to the efficiency of B12

production in different environments. A comparison of the growth of each species in various locations, conditions, and climates would be a good starting point for this question.

Additionally, the 4th highest ortholog by dN score was identified as a vitamin B12-dependent ribonucleotide reductase (RNR) (dN score = 0.320). These class II RNR enzymes reduce ribonucleotides to deoxyribonucleotides, which are in turn used for forming DNA, by breaking the bond between the central cobalt ion and the upper ligand of adenosylcobalamin (187).

Class II (B12-dependent) RNRs have been well studied in *Pseudomonas aeruginosa.* A previous study investigated the differential expression of B12-independent class I RNR and B12-dependent class II RNR at different stages of the bacteria's growth cycle (188). It was found that the exponential growth phase had high expression of class I and low expression of class II, but that the expression pattern switched upon entry to the stationary phase, with a 6-fold reduction in class I expression and a 6-fold increase in class II expression. The authors posited that the role of the class II RNR is to provide enough deoxyribonucleotides for DNA repair and a low rate of replication during low oxygen stress. Subsequent studies confirmed that under anaerobic conditions, B12-dependent RNR was essential for rescuing impaired growth, and for modulating the switch from the planktonic to the biofilm lifestyle (189,190). Experiments to compare the tolerance of *P. megaterium* and *P. aryabhattai* to low oxygen availability, including the ability of each to form biofilms under such conditions, may help to explain the sequence divergence in this gene.

PTS sugar transport genes

The orthologs with the 2nd and 3rd highest dN scores were the genes for the EIIB (dN score = 0.404, Figure 3-7) and EIIC (dN score = 0.282) subunits of the ascorbate-specific PTS sugar transport system. PTS systems transport different sugars across the cell membrane whilst also phosphorylating them to prevent the sugar from passing back out of the cell. This process is performed by enzyme EII, which is composed of subunits EIIA, EIIB, and EIIC, the latter of which is the transmembrane component with the sugar binding site (191). The phosphate is removed from a PEP molecule by EI, then transferred to HPr, then EIIA, then EIIB, and then to the imported sugar. The first proteins in the sequence, EI and HPr, are shared by multiple PTS systems, but the EII subunits A, B and C have different genes for each PTS system that transports one or several specific sugars (192).

For the ortholog with the highest dN score (0.496), it was difficult to identify a gene function due to its low alignment identity (less than 48%) to proteins from species other than *Priestia megaterium,* but it seems to be a regulator of the PTS system for transporting either mannose or beta-glucosides. PTS systems are often regulated by proteins that contain domains resembling EIIA, EIIB, or EIIC of the PTS protein. The domain and method of regulation differs between PTS systems and between species, but often involves one of the PTS subunits changing the phosphorylation state of the regulator's domain in response to a change in the concentration of the target sugar outside the cell, thereby regulating its own expression as needed (193). The protein identification results from Prokka and UniProt indicate that this ortholog contains an

EIIA domain, and most closely resembles the ManR transcriptional regulator of the mannitol/fructose PTS (194), or BglG, which regulates the beta-glucosides PTS (195).

PTS genes can also have additional functions such as gene regulation of other sugar transport systems (196,197). In *E. coli,* EIIA communicates with the chemotaxis pathway to move the cell toward higher concentrations of sugars (198). The PTS system can also be a target for antimicrobial peptides from other bacteria (199), for viral DNA insertion (200), and for antibiotics (201). In order to investigate any changes in function of these PTS orthologs between *P. megaterium* and *P. aryabhattai,* strains of the two should be compared for their rate of utilisation of different sugars, and for their susceptibility to a range of viruses and antibiotics.

Prophage proteins

The 10th (dN score = 0.050) and 36th (dN score = 0.033) orthologs by dN were endogenous phage proteins. The integration of phage DNA into the bacterial genome is sometimes advantageous for the bacteria when the gene product has a useful function, such as a toxin that enables the bacteria to act as a pathogen (202). In *E. coli*, a prophage sequence encodes a protein which binds to the mannose PTS system to prevent infection by external viruses, and regulates the cell cycle (200). As another example, in *Shewanella oneidensis*, a prophage sequence controls the formation of biofilm in response to cold temperatures (203,204). Since these prophage orthologs seem to have been kept in the *Priestia* core genome since the common ancestor of *P. megaterium* and *P. aryabhattai*, they may also have useful functions that could be investigated in the future.

Sporulation-related genes

Three orthologs were related to spore formation and germination. The rank 6th ortholog was identified as a sporulation-associated protein with an unknown function (dN score = 0.076), and number 33 as a sporulation negative regulatory protein (dN score = 0.034, Figure 3-8). The 47th ortholog's best match was the stage V sporulation protein E (dN score = 0.030), which is expressed about two hours into the sporulation process (205), but this gene had a high BLASTP e-value of 0.069 and so was not reliably identified. These proteins all had low identity to their closest matches outside of *P. megaterium/aryabhattai* – 31.2%, 75.1%, and 51.7% – which reflects the large variation in sporulation gene sequences between species (206).

The 33rd ortholog showed a high sequence similarity to the *Bacillus subtilis* gene *paiA*, which encodes an N-acetyltransferase that acetylates spermine and spermidine, and may have a role in the bacterial stress response by regulating the binding affinities of other gene regulators (207). This gene has also been shown to be an inhibitor of the sporulation process (208). In *Enterococcus faecalis*, a biofilm-forming human pathogen, the level of expression of *paiA* changed when the cells were in a biofilm versus when they were planktonic, but since *E. faecalis* is not a spore-forming organism, this is likely related to another function of *paiA* (209).

Flagellar assembly genes

The 38th, 39th and 81st orthologs in the top 100 were the genes for flagellar hook-length control protein, *fliK* (dN score = 0.032); flagellar hook-associated protein, *fliT* (dN score = 0.032); and and flagellar biosynthesis protein, *flhF* (dN score = 0.024, Figure 3-9).

The 38th ortholog *fliK* has been studied for decades due to its important role in controlling the length of the bacterial flagellum as it is assembled, but it is not itself part of the flagellar structure (210). The flagellar hook is a short, extracellular piece of the flagellum between the motor and the propeller-like filament, and *fliK*'s role in hook length control was noticed in the 70s when it was found that *fliK* mutants produced 'superhook' flagella with long chains of conjoined hook protein (211). *fliK*'s role in the process is bifunctional, acting as a molecular ruler to measure hook length and also controlling the length of the assembling flagellum hook by switching the substrate-specificity of the type III secretion system (T3SS). When the hook length reaches approximately 55 nm, *fliK* causes the T3SS to stop secreting hook proteins and to instead secrete filament proteins (212).

In addition to FliK, several other proteins are required for the T3SS substrate switching, one of which is the FlhA transmembrane protein (213). The cytoplasmic domain of FlhA also contains the binding site for several flagellar export chaperones, including the protein encoded by the 39th ortholog, *fliT* (214). FliT is the chaperone protein for the FliD flagellar filament cap, protecting it from being degraded in the cytoplasm before it reaches the growing flagellum (215).

Species of *Bacillus* and *Priestia* have peritrichous flagella (78,216), meaning they have many flagella around the cell rather than one or a few at the cell pole (217). The average *B. subtilis* cell forms 26 flagella which are positioned in a symmetrical pattern that is determined by the 81st ortholog, *flhF* (218). The role of *flhF* in flagellar assembly is currently less well understood (219) but it is known to be essential to the process, since the deletion of *flhf* produces cells with swimming defects due to absent or mislocalised flagella (220). This gene seems to determine the locations where flagella will be assembled by localising at the cell membrane and then recruiting early flagellar assembly proteins such as FliF (221). FlhF is also a DNA-binding protein that regulates the expression of several other flagellar genes (219).

In addition to their roles in flagella formation, *fliK* and *flhF* have been shown to have functions related to pathogenicity in the species group *Bacillus cereus* sensu lato, which is closely related to the Priestia genus (222). In *Bacillus cereus*, FlhF is required for exporting toxins and thus also for pathogenicity (223), and FliK provides resistance to antimicrobial peptides in the insect pathogen *Bacillus thuringiensis*, independently of whether flagella were formed properly (224).

Several *Bacillaceae* species that were previously considered to be non-pathogenic have been increasingly recognised as infrequent, opportunistic pathogens (225–228), including *P. megaterium* (229–233), and so the sequences changes in these multifunctional flagellar genes in *P. aryabhattai* could be related to either flagella formation or to pathogenicity. Future studies should examine the flagella of *P. megaterium* and *P. aryabhattai* strains for differences in location and shape, and also compare the pathogenic capabilities of the two.

Implications for bacterial life cycle

Clade-specific sequence changes in the orthologs discussed above may also have knock-on implications for biofilm formation. Biofilms are dense communities of bacteria surrounded by an extracellular matrix, in which cells show an increased rate of horizontal gene transfer (234) as well as greater resistance to environmental stresses, including antibiotics (235). Biofilm formation has been well studied in *Bacillus subtilis* (236) and *Bacillus cereus* (237) due to its importance in the colonisation of soil, food, plants (238), animal tissues (239). Because of the resilience of biofilms and spores to environmental stress and deliberate cleaning, these Bacilli (especially *Bacillus cereus* sensu lato) are notorious contaminants of food (240,241) and hospital surfaces (242), causing food poisoning and nosocomial infections.

Flagellar motility is important for the transition from the planktonic to the biofilm lifestyle. A reduction in motility and in biofilm formation has been observed in cells with mutations in flagellar genes (243) and in cells whose flagella are damaged by turbulence (244). This is likely due to the need for motility in order to reach the site of biofilm formation; often a surface, including air-liquid interfaces (245). As the cell makes contact with such a surface, the physical interruption of flagellar rotation acts as a mechanical signal for the production of exopolysaccharides that adhere the cell to the surface (246). The flagellum is not absolutely necessary for biofilm formation, as cells with flagellar mutations may still form biofilm if they reach the air-liquid interface by Brownian movement (247).

Sporulation and biofilm formation may seem to be mutually exclusive pathways in the genetic sense, with a few key genes controlling the decision to enter into either pathway depending on environmental cues (248–250). But on the population level, these are complementary phenotypes. Biofilms, in addition to colonising an environment, also act as platforms for dispersal. Sporulation occurs from cells within the biofilm within 24 hours, and in the later stages of biofilm formation, up to 90% of the cells within the biofilm can be endospores (251).

Bacteriophages may at first appear to be simple predators of bacteria, but lysogenic infection can have complex effects on bacterial ecology (252). Prophage encoded genes may have beneficial effects on host physiology (253), and can also affect the transitions between developmental stages in the life cycles of *Bacillus* species (237,254). In *B. subtilis*, prophages have been found to inhibit the processes of replication, sporulation, and biofilm formation by interacting with cell cycle regulators (255). The expression of the polysaccharide synthesis gene *spsM*, which makes the spore coat hydrophilic and thus facilitates dispersal in water, is controlled by the timed excision of a prophage sequence from *spsM* during sporulation (256). Prophage insertion may also have negative effects, such as reducing the fitness of the biofilm community by disrupting antibiotic resistance genes (257). In *B. anthracis*, prophages can inhibit or promote sporulation, or induce exopolysaccharide expression and biofilm formation, allowing the colonisation of new environments (258). In *B. thuringiensis*, phage lysogeny can inhibit sporulation and biofilm formation in favour of swarming motility, allowing the colony to spread into nutrient-rich environments (259).

P. aryabhattai as a plant growth promoting bacterium

I also searched for genes that have diverged between the two species with functions related to plant growth promotion (PGP). The intense interest in *P. aryabhattai* as a plant growth promoting bacteria (99,260–263) seems to rely on the implicit assumption that it is better suited than other species for this task, however this does not seem to have been tested.

A swathe of recent papers have each described a single isolate of *P. aryabhattai* or *P. megaterium*, its gene content, and its effect on the rate of plant growth versus a control without bacterial inoculation (e.g. 14,17,18,41,49). A few more studies have compared the effects of several strains within the same species – within *P. aryabhattai* (266), within *P. megaterium* (91), and within the closely related *Bacillus cereus* (267). However, the results of these studies cannot be compared because each used a different species of plant in different climates and conditions.

Studies comparing the plant growth performance of multiple species are less common. Verma *et al.* (126) conducted the only study comparing *P. megaterium* with *P. aryabhattai*, as well as 53 other bacteria isolated from wheat rhizosphere, by testing for the chemical activity of 15 PGP-associated functions. Their results showed that, of the two species, only *P. megaterium* was able to solubilise potassium and zinc and showed siderophore activity, whereas *P. aryabhattai* was the only one to produce HCN and fix nitrogen. They also found differences in the geographic distribution of *P. aryabhattai* and *P. megaterium*: *P. megaterium* was present in roughly equal abundances in all six regions of India that were sampled, whereas *P. aryabhattai* was

present in only four regions. The two regions where *P. aryabhattai* was absent had the most extreme conditions of temperature and pH. Both species were widely distributed compared to the majority of other species which were only found in one or two regions. However, as discussed earlier, the study identified the species of each isolate using 16S rRNA sequences, which have now been shown to be inaccurate in distinguishing *P. megaterium* and *P. aryabhattai* (Figure 2-1). The potential misidentification of the *Priestia* isolates and the lack of genomic data provided make the results of the study difficult to interpret conclusively in terms of the difference between these two species.

Several recent studies have compared the plant yield of maize and sweetcorn following inoculation with *P. megaterium* and other species, minus *P. aryabhattai* (268–270). The results have been inconsistent, with *P. megaterium* performing equally (268) or better (269,270) than the related *Bacillus* species. However, these studies also used 16S rRNA gene sequencing for species identification, which cannot distinguish *P. megaterium* and *P. aryabhattai* (Figure 2-1).

The results of the ortholog clustering in this Chapter identified six genes involved in iron import, a function known to be important for plant growth promotion, that had a higher copy number in all *P. aryabhattai* genomes than in the *P. megaterium* genomes. This result was the only robust difference found between all genomes of both species in a plant growth promoting function. However, the method of *in silico* core genome comparative genomics does not show whether these additional orthologs are expressed, resulting in a higher rate of iron transport to the bacterial cell and higher iron availability to the plant, or if these genes have been silenced, or if the total expression of the orthologs with identical functions is maintained at a constant level by the cell (136).

Future studies need to compare the expression levels of these iron transport genes between the two species and study the effects of the expression levels on plant growth. Transcriptomics experiments of the two species growing in the rhizosphere may identify other orthologs with differences in expression level when the bacteria are in close association with a plant host.

The bacterial species concept often relies on functional and behavioural differences between strains to decide where the boundary between species should be (271), and so I aimed to reveal the differences in gene content that have occurred during the evolutionary split between the two species *P. megaterium* and *P. aryabhattai*. I also used the species-specific orthologs to develop PCR primers that researchers may use to quickly identify whether their new isolates belong to *P. megaterium* or *P.* a*ryabhattai*, without the need for whole genome sequencing. For all of the orthologs that were found to differ between the species, further work needs to be done to identify the domains and active sites of the orthologs' proteins and whether the sequence substitutions shown have changed those domains.

Chapter 4 — Metagenomics study of the global distributions of *Priestia megaterium* and *Priestia aryabhattai*

Introduction

Chapter 3 examined the whole-genome sequences of strains from the two species *Priestia aryabhattai* and *Priestia megaterium*, in order to identify differences in gene content between them. Chapter 4 now focuses on identifying differences in the abundance of each species in different habitats and environmental conditions.

The hypothesis that *P. aryabhattai* has adapted to survival at high altitudes was first suggested by the authors who first named *P. aryabhattai*; after isolating it from the air at 41 km altitude, they also found that their strain was more resistant to UV radiation – a key trait for surviving for long periods while airborne – than its nearest phylogenetic neighbour, a *P. megaterium* strain (76). The authors suggested that the *P. aryabhattai* strain may have been lifted from the ground to the air by updrafts, acknowledging in effect that it presumably grows in habitats on the ground and disperses through the air. However, the hypothesis of adaption to the high-altitude air environment was backed up by their results that showed a difference in survival between the two species after exposure to UV radiation. The number of *P. megaterium* colony forming units dropped quickly at increasing levels of UV radiation, with none surviving at 0.3 J cm^{-2} or above, whereas *P. aryabhattai* continued to produce small numbers of colonies at 0.8 J cm^{-2}.

In Singapore, the strains cultured from air samples include twelve *P. aryabhattai* strains (Table 2-1) to one *P. megaterium* strain (83), a bias which may appear to be in

concordance with the idea that *P. aryabhattai* is better able to survive while airborne. However, rather than 41 km altitude, these samples were taken from 1.5 m above ground, where they would have been less exposed to UV radiation. In Chapter 3 it was found that *P. aryabhattai* genomes have diverged from *P. megaterium* in the sequences and copy numbers of several genes related to the sporulation process, which could be linked to adaptations for the formation of spores which are more resistant to the environmental stresses of airborne dispersal. Thus, the suggestion that *P. aryabhattai* is better adapted to high altitude air than *P. megaterium* remains plausible and requires testing.

Another question is whether there is any difference of the two species in their ability to survive under different conditions. Although the first *P. aryabhattai* strain was isolated above India in 2005 (76), GenBank now contains genome sequences of both species from locations across the world, including Asia (accession, GCA_024434365), Africa (GCA_024581075), North America (GCA_019748975), South America (GCA_020251185), Oceania (GCA_020179075), and Europe (GCA_019193015). This shows that both species are globally dispersed. However, it is not known if there are any differences in their abundances or growth rates between different environments and climates.

A series of recent papers have described the effects of the changing environment on the bacterial diversity of the air microbiome, highlighting in particular the importance of the air temperature on community dynamics. A diel cycle was found to occur near the ground, where bacteria increase in relative abundance in comparison to fungi during the daytime (156). In the same study, the species richness in Singapore air increased with

temperature and decreased with atmospheric CO_2 concentration, with some species abundances correlating strongly to changes in these environmental factors. For example, the phylum Bacillota, which includes the *Priestia* genus, changed in abundance in response to temperature, but its relative abundance changed less in response to the day/night cycle than other groups. Temperature, relative humidity, and CO_2 concentration all changed regularly with the time of day, but the concentrations of nitrogen oxides and sulphur oxides in the air did not, and so no response in the microbiome to these air pollutants could be observed. Rainfall events correlated with the relative abundances of fungi, proteobacteria and cyanobacteria, but the community composition did not respond to the changing Singapore monsoon seasons over the 13 months of sampling (272).

A subsequent study confirmed the presence of the diel cycle in a temperate climate, and further found that the diel effect diminished and disappeared when sampling was conducted at higher altitudes, where the temperatures at day and night converged and the community composition also changed (273). The changes in DNA yield and community composition at different heights were explained better by the air potential temperature rather than actual temperature. In one experiment in which sampling was performed onboard an aircraft, the potential temperature increased with altitude, leading to vertically stable air with little mixing between the near-ground air and the high-altitude air. The high-altitude air thus had lower DNA concentrations and a distinct microbial community composition which contained higher abundances of reads that mapped to radiation-tolerant bacteria and to the DNA repair genes *uvrABC* and *phrB*.

The same study contrasted the results from the aircraft sampling with the results of sampling atop a 200 m tower. In that experiment, the potential temperature decreased as the altitude increased, leading to upwards convection that mixed the ground-level and tower-level air. This led to the ground and tower-top having similar DNA concentrations and community compositions.

The diel cycle was also observed year-round near the equator and also in Siberia in the summer, but not in the Siberian winter, although there was a similarly diverse (but taxonomically distinct) air microbiome present in the winter (274). The Bacillota showed an especially large response to the seasons, with higher relative abundance in the winter. Given these results, we can expect the relative abundances of both *Priestia* species to change across different altitudes and across the day/night cycle. However, these studies took a big-picture approach and compared abundances at the phylum level; a species-level analysis in search of differing responses of *P. megaterium* and *P. aryabhattai* to the environment has not been done.

In order to investigate the potential ecological differences between *P. megaterium* and *P. aryabhattai*, this Chapter compares the spatial distribution of the two species across the world. This was done using a metagenomic approach, in which DNA was sequenced from the various organisms that are present in an environmental sample, to calculate the relative abundances of the two species in each location. Any differences or patterns found in the geographic distributions of the two species may help to infer the different environmental conditions or ecological niches that favour one group over the other, and further, the cause of the speciation between them.

I investigated two specific questions using metagenomics data from air samples. The first was to test the hypothesis that *P. aryabhattai* is a species that has adapted for surviving at high altitudes better than *P. megaterium*. The second was to analyse the global distributions of the two bacteria and identify any environmental conditions or locations that favour one over the other. Using data from air samples is obviously required to answer the first question but is also advantageous for the second question. *P. megaterium* and *P. aryabhattai* are generalist bacteria that can grow in a variety of environments such as soil and water. Sampling from a single environment such as soil would give an incomplete picture of the diversity of a region, whereas air samples can capture spores that have dispersed from multiple environments, providing a snapshot of the overall diversity of each species is growing in a local environment.

In addition to the local environmental measurements that were taken during sampling such as temperature and humidity, other publicly available data may be useful in finding environmental conditions that correlate with the relative abundances of *P. megaterium* and *P. aryabhattai*. If there is an unknown environmental condition that explains the changing ratio of the two species between samples, it should be something that changes over time and may plausibly affect the abundances of bacteria in the air. For the analysis of samples taken in Singapore, I checked for effects of the monsoon seasons, and of air quality.

Singapore experiences the Northeast monsoon season in December to early March, and the Southwest monsoon season in June to September. Between these monsoon seasons are two inter-monsoon periods. The study discussed previously found no

significant changes in phylum-level community composition across monsoon seasons (272) but the effects on individual species are unknown.

The effect of air pollution on the air microbiome community composition have received more attention recently, with several studies showing changes in the relative abundances of different taxa during changes in pollution levels having been published: Cao *et al.* (275) observed changes in the relative abundances of bacteria associated with terrestrial, marine, freshwater, and faecal habitats over the course of a smog event as the PM2.5 and PM10 levels changed. A later study compared the community composition between levels of pollution and showed that the air microbiome species diversity was correlated with levels of PM, NO_2, and CO, with the phylum Bacillota (which contains the genus *Priestia*) becoming more abundant at medium levels of pollution (276). Finally, an air pollution study in Beijing found that air samples with higher levels of PM2.5 and PM10 pollution had higher species diversity, with a peak in abundance of Bacillota during major smog events (277). These studies demonstrate that the abundance of Bacillota responds to air pollution, but it is unknown whether individual species within the phylum are affected equally.

Methods

Sample collection, metagenomic data generation, and Kaiju species identification were performed through a collaboration with the research team from SCELSE NTU. The details are described in previous publications (156,273).

Metagenomics sampling and sequencing

To investigate any differences in distribution between *P. megaterium* and *P. aryabhattai*, I used three metagenomics datasets that were sequenced from outdoor air samples. The samples were collected across the world between 19th Feb 2017 and 27th Feb 2020, as part of two vertically stratified experiments (273) and one global ground level sampling project.

The first two datasets ('tower' and 'aircraft') are those that were used in the published studies on vertical sampling that were described above (see Introduction). They were collected during experiments in Germany in 2018, and both aimed to compare the abundances of microorganisms at different heights. In the first experiment, 120 samples were taken at the top and bottom of a 200 m high meteorological tower at the Karlsruhe Institute for Technology (49° 5' 33'' N, 8° 25' 33'' E) between 8th and 13th Oct 2018. In the second, 114 samples were taken from an aircraft as it flew at different heights between 0 m and 3,500 m above Brunswick, onboard a research aircraft operated by Technische Universität Braunschweig, between 9th Oct and 12th Oct 2018. The data from these experiments allowed us to compare the distributions of *P. megaterium* and *P. aryabhattai* in the air at different altitudes above the same two locations.

The third dataset ('global') contains 1,180 air samples taken from 1 m above the ground in locations across the world between 2017 and 2019. This global dataset allows for comparisons of the two groups between different locations and climates using

metadata on temperature, humidity, location, time of year, and air quality where available.

The collection and processing of the air samples have been described in Gusareva *et al.* (156). Briefly, airborne biomass was collected onto filters using SASS3100 air samplers with an airflow of 300 L/m for 2 hours, 1.5 m above the floor, placed in outdoor air adjacent to residential or university buildings; for example on balconies. During the aircraft experiment, the outside air was instead piped into a chamber containing the air samplers.

After sample collection, the filters were washed in PBS/Triton X-100 and DNA was extracted with the DNeasy PowerWater Kit (Qiagen). DNA sequencing was performed using the Illumina HiSeq 2500 to generate 251 bp paired end reads. The reads were aligned to the NCBI nonredundant protein database using Kaiju v1.7.2 (278). The sample collection, metagenomics sequencing, and first-pass species identification for each dataset were performed by the authors of those studies.

Environmental data access

Measurements of latitude, longitude, temperature, humidity, time, and altitude were taken at the time of sampling. Sample locations were classified into six biogeographic realms – Afrotropical, Australasian, Indomalayan, Nearctic, Neotropical, and Palearctic (279). For Singapore samples, historical AQI data taken by the five domestic recording stations were downloaded from aqicn.org and matched to the metagenomics samples

by closest geographic distance. Singapore air samples were classified as being taken during or between the monsoon seasons according to the dates in Table 4-1.

Table 4-1: Typical dates of the monsoon seasons of Singapore.

Dates	**Monsoon season**
1st Dec – 15th Jan	Early Northeast monsoon
16th Jan – 15th Mar	Late Northeast monsoon
16th Mar – 31st May	First inter-monsoon period
1st Jun – 30th Sep	Southwest monsoon
1st Oct – 30th Nov	Second inter-monsoon period

Air quality in Singapore is recorded by the National Environment Agency using five monitoring stations in the North, East, South, West, and Central regions of the island. Individual readings of PM2.5, PM10, O_3, NO_2, SO_2, and CO are each converted to AQI values that approximate their public health risk, with the overall reported AQI value being the highest of the six component AQI values (280).

As a small island country with frequent and publicly available AQI data from across the island, Singapore would be an ideal place to study the effect of air pollution on the relative abundances of the *Priestia* groups. However, the air quality in Singapore during the air sampling for this dataset remained mostly within the 'Good' category, with a maximum recorded value of 60 in the low end of 'Moderate'. For reference, AQI is in the 'good' range when it is under 50, with 60 being in the low end of 'moderate' (Table 4-2). This means that the global dataset is missing the more extreme AQI values that might

have more of an influence on the air microbiome, and the effects of AQI changes within the 0–60 range could be minor.

Table 4-2: Scale of air quality index values and their health implications, defined by the US EPA, from aqicn.org. The air quality at the time of sampling for the global metagenomics dataset was usually 'Good' and occasionally 'Moderate', up to an AQI value of 60.

AQI	Air Pollution Level	Health Implications	Cautionary Statement (for PM2.5)
0–50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk	None
51–100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
101–150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
151–200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion
201–300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
300+	Hazardous	Health alert: everyone may experience more serious health effects	Everyone should avoid all outdoor exertion

Species reclassification of reads

The species identification issue between *P. megaterium* and *P. aryabhattai*, and the presence of genomes of these species with misassigned names in databases, meant that the species identifications made by Kaiju using the protein database could be more accurate if confirmed by DNA alignment to properly identified reference genomes. In order to improve the accuracy of the assignment of reads to species, and to increase the number of *Priestia* reads available for analysis, I analysed every read that Kaiju had identified as *Priestia* or *Bacillus* (prior to the renaming of the *Priestia* genus (99)) in the three metagenomics datasets. Each read was aligned against the DNA sequences of the GenBank representative genome assemblies of the two species (*Priestia megaterium* 22-2 and *Priestia aryabhattai* K13) using BWA (281) and samtools (282).

Reads were only considered for reassignment if they had at least 95% identity (identical DNA bases) and 80% alignment coverage (aligned bases) to at least one of the representative genomes. The two alignments to the reference genomes were then compared by their sequence identity and alignment coverage. If one alignment was better by both metrics, or better by one metric with the other being equal, then the read was (re)assigned to the species with the better alignment. If one alignment had higher identity but the other alignment had higher coverage, the read was not reassigned, and the species name given by the Kaiju amino acid alignment was used. In other words, reads were only reassigned when it was clear that they aligned better to the other species.

Calculation of species' relative abundance

To be able to compare the numbers of reads of each species between samples, the numbers of reads of each group in each sample were normalised using the formula:

$$N = C \times S \times m$$

Where N is the relative abundance of a species, C is the number of reads of the species in the sample, S is the total number of reads in the sample, and m is the total number of reads in the sample with the fewest reads.

For the Germany tower and aircraft datasets, the number of reassigned, normalised reads at each height were then compared. For the global dataset, an FAMD (Factor Analysis of Mixed Data) ordination was performed on the environmental metadata using the R package FactoMineR v2.4 (283), in order to search for patterns of abundance of the two species using multiple categories of metadata at once.

Modelling of species majorities

Random forest and support vector machine (SVM) classifiers were used to model the samples plotted on the FAMD ordination and predict which species had the higher relative abundance in each sample, in order to test whether any clusters shown could be reliably predicted from the same metadata used for the ordination. These were done using the R packages randomForest v4.7.1.1 (284) and e1071 v1.7.11 (285) respectively. Since the number of *megaterium*-majority samples was considerably greater than the number of *aryabhattai*-majority samples, the former were randomly sampled to be

equal in number to the latter. The samples were then split 80:20 into train and test data for the random forest and SVM classifiers.

Results

Changes in relative abundances after reclassification of reads

I developed a pipeline to correct the species identification of metagenome sequencing data using DNA-DNA alignments. This pipeline can be applied to any species; here I describe the reclassification of the two closely related species *P. megaterium* and *P. aryabhattai*. The reclassification method, of identifying reads using Kaiju and then reclassifying the *Priestia* reads using nucleotide alignments, greatly increased the number of reads that were identified as *P. megaterium* and *P. aryabhattai*. The mean abundance of *P. megaterium* increased by 2.4x and 2.9x in the tower and aircraft experiments respectively, whereas the *P. aryabhattai* abundance increased by 10x and 10.9x (Figure 4-1).

Most of the newly identified *P. megaterium/aryabhattai* reads were those that Kaiju had identified as unknown species of the genus *Bacillus/Priestia*. Additionally, reads were reassigned from *P. megaterium* to *P. aryabhattai* much more frequently than the reverse, which may be due to a lack of *aryabhattai*-specific protein sequences in the database used by Kaiju and the existence of genes with similar amino acid sequences but differing nucleotide sequences between the two species.

The global dataset showed a smaller increase in the number of *P. aryabhattai* reads from reclassification than the two vertical datasets. This is likely due to the difference in databases used by Kaiju to identify reads: the reads in the tower and aircraft data were identified by Kaiju using the NCBI non-redundant protein database downloaded 22 Nov 2019, whereas the global analysis used the database from 30th Mar 2021. There may be a better representation of *P. megaterium* and *P. aryabhattai* in the updated database which allowed more of the reads of the two species to be identified on the first pass by Kaiju, with less need for the second pass of identification by DNA sequences.

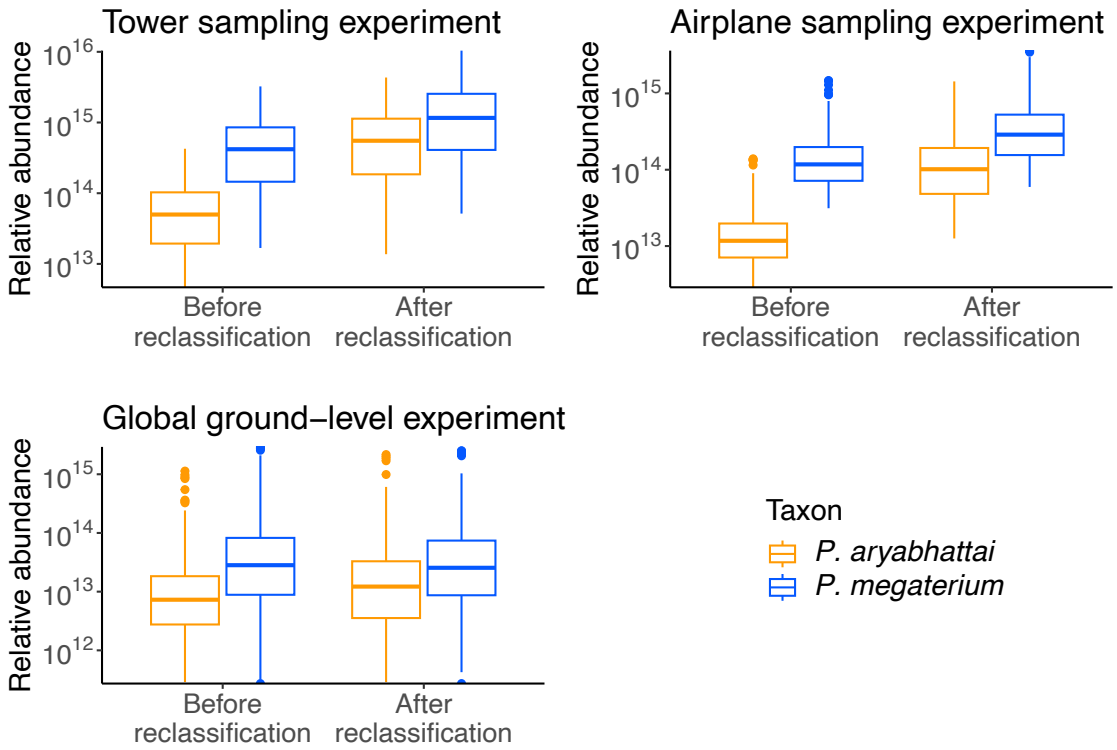


Figure 4-1: Changes in relative abundance of *P. megaterium* and *P. aryabhattai* after reclassifying reads by aligning them to reference genomes (see Methods). For both vertical experiments (top), the relative abundance of the two species increased after reclassifying reads. For the global ground-level air sampling (bottom), reclassification increased the average relative abundance of *P. aryabhattai* but decreased that of *P. megaterium*. The before and after abundances for both species in all three experiments were significantly different ($p < 0.05$, paired Wilcoxon test).

Vertical distribution of P. *megaterium and* P. *aryabhattai*

In order to investigate the hypothesis that *P. aryabhattai* is better at surviving than *P. megaterium* while airborne at higher altitudes, I analysed the distribution of the two species in the metagenome sequencing data from air samples collected at different heights – ground level (0m), on a meteorological tower (200m), and onboard an aircraft (300–3500 m).

At all heights for both the tower and aircraft experiments, the relative abundance of megaterium was significantly greater than that of aryabhattai at the $p < 0.01$ level using the Mann-Whitney U test (Figure 4-2). This result is evidence against the hypothesis *P. aryabhattai* is adapted to survive for longer than *P. megaterium* in high-altitude air.

The two vertical datasets did differ in the trend of the overall abundance vs altitude; the tower data had higher abundances of both species at 200 m than at 0 m ($p < 1 \times 10^{-6}$, Mann-Whitney U test), whereas the aircraft data showed lower abundances of both species at increasing altitudes ($p < 3 \times 10^{-15}$, Mann-Whitney U test) — note however that the tower's height was smaller than the lowest airborne aircraft sample (300 m).

Figure 4-3 shows the ratio of *P. aryabhattai* to *P. megaterium* abundances in each sample of the two vertical experiments. In the two experiments, the ratio of the two species did not change significantly at different heights ($p > 0.01$, Mann-Whitney U test). The aircraft experiment did show a small drop in the ratio of *P. aryabhattai* to *P. megaterium* at the highest level of 3500 m.

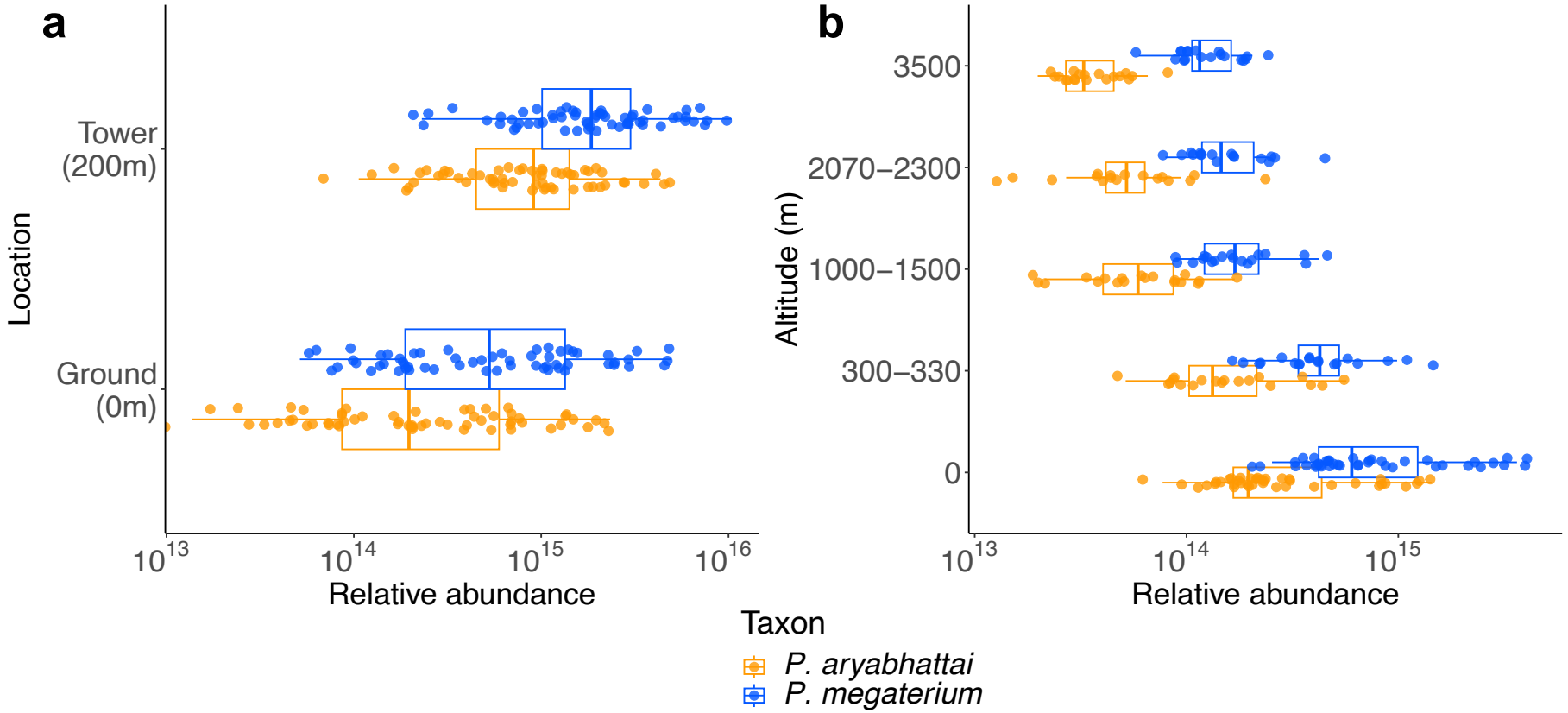


Figure 4-2: Relative abundances of *P. megaterium* and *P. aryabhattai* in two vertically stratified metagenomics experiments. The read counts of the two experiments were normalised separately and the abundances cannot be directly compared between experiments. Each sample taken has one blue point and one orange point, showing the abundance of each species in that sample. The boxplots show the summary statistics for all abundances of each species at a height stratum.

a: Species abundances in outdoor air samples taken at the top and bottom of a 200 m tower. Both species increased in relative abundance at the top, with *P. megaterium* being the more abundant on average.

b: Species abundances in outdoor air samples taken from an aircraft at heights up to 3500 m. Both species fall in relative abundance as height increases, with *P. megaterium* consistently more abundant than *P. aryabhattai*.

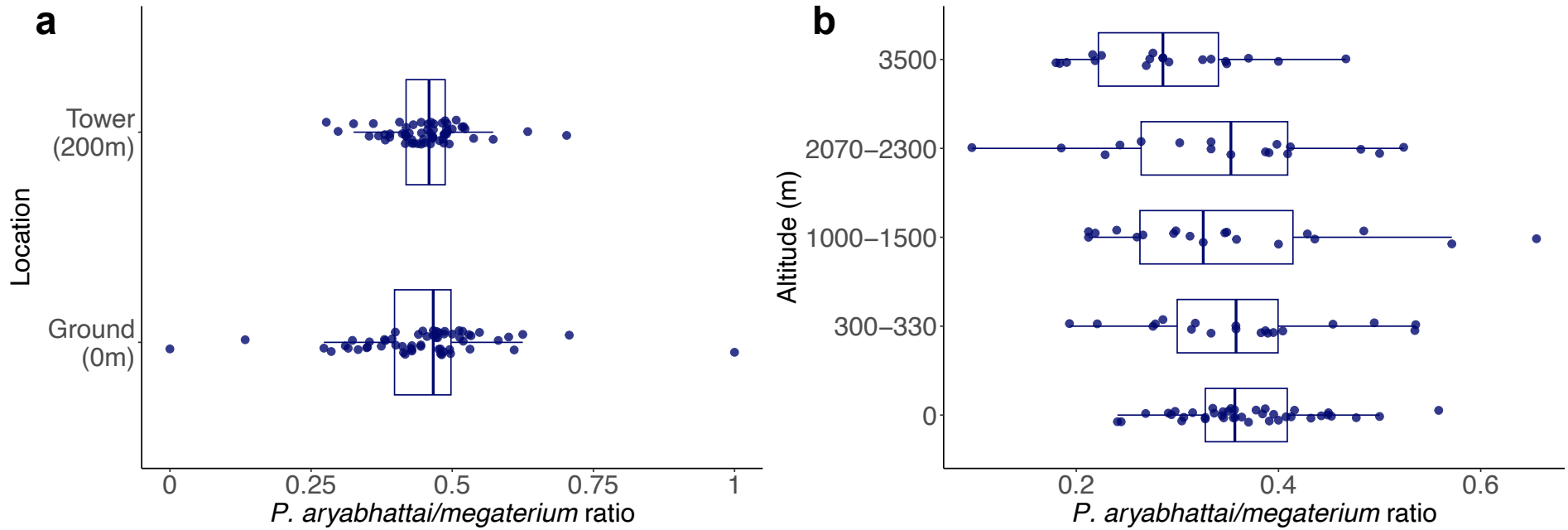


Figure 4-3: Ratios of the relative abundance of *P. aryabhattai* to the relative abundance of *P. megaterium* in each sample, for the same experiments as Figure 4-2 above. Each sample has one point representing the species ratio for that sample. **a:** The ratio between the two species did not significantly change between the tower and ground ($p > 0.05$, Kruskal-Wallis test) because both species followed the same trend of increasing abundance with height. **b:** The ratio between the two species changed between altitudes ($p < 0.05$, Kruskal-Wallis test), with the proportion of *aryabhattai* decreasing slightly but inconsistently.

Global distribution of P. megaterium and P. aryabhattai

The analysis of the vertically stratified datasets showed that *P. megaterium* had generally higher relative abundance than *P. aryabhattai*, with both species showing decreased relative abundance with increasing altitude above 300 m. The global dataset showed the same general prevalence of *P. megaterium* over *P. aryabhattai,* but with a cluster of samples in Southeast Asia that had higher than normal ratios of *P. aryabhattai* to *P. megaterium*.

Onsite vs weather station metadata

For the global dataset, the temperature and humidity were recorded using both on-site HOBO sensors and using public data from nearby weather stations. The HOBO sensors gave data at the exact time and location that the air samples were taken, whereas the weather stations were not always nearby the sample site. Samples taken in rural areas were especially less likely to have a weather station in the immediate area, and the available data from different countries is often available only for timepoints every few hours apart, or as a daily average.

However, several issues with the HOBO sensors make their readings unreliable, with generally poor agreement between the HOBO measurements and the weather station data (Figure 4-4). The humidity measurements were sometimes extremely high or low, for example reading at 1% while the closest weather station recorded 96%. The linear model of the weather station data from the HOBO data diverged significantly from the

y=x identity line. The two data sources were correlated much better on the temperature data, but the relationship between HOBO temperature and weather station temperature became non-linear at higher temperatures where the HOBO sensors gave values much higher than the weather station values. For example, weather station recordings of 20–30 °C often had corresponding HOBO readings of 50–56 °C. Anecdotally, team members who were involved in the sample collection have commented that this can happen when the HOBO sensor is placed in direct sunlight. For these reasons, the weather station data can be considered to be more reliable, and was used instead of the HOBO data during the global metagenomics analysis.

Distributions and correlations of environmental variables

The geographic locations of the samples were well distributed over the northern hemisphere as well as South, Southeast and East Asia, with a lack of samples from South America (Brazil and Ecuador), Africa (only São Tomé), and Oceania (only New Zealand). Because of this, the environmental data had a large range of temperature and humidity but were skewed toward the high temperatures which were represented mainly by Southeast Asia. Out of 1,180 samples, 173 were collected in Germany and 274 were collected in Southeast Asia, which is the reason for the bimodal distributions of latitude, longitude and biogeographic realms (Figure 4-5). The different biogeographic realms and seasons were naturally associated with certain locations and temperature and humidity ranges (Figure 4-6). The Singapore samples captured a range of air quality values up to 60, and from each monsoon season, but without samples taken between the monsoon seasons (Figure 4-8).

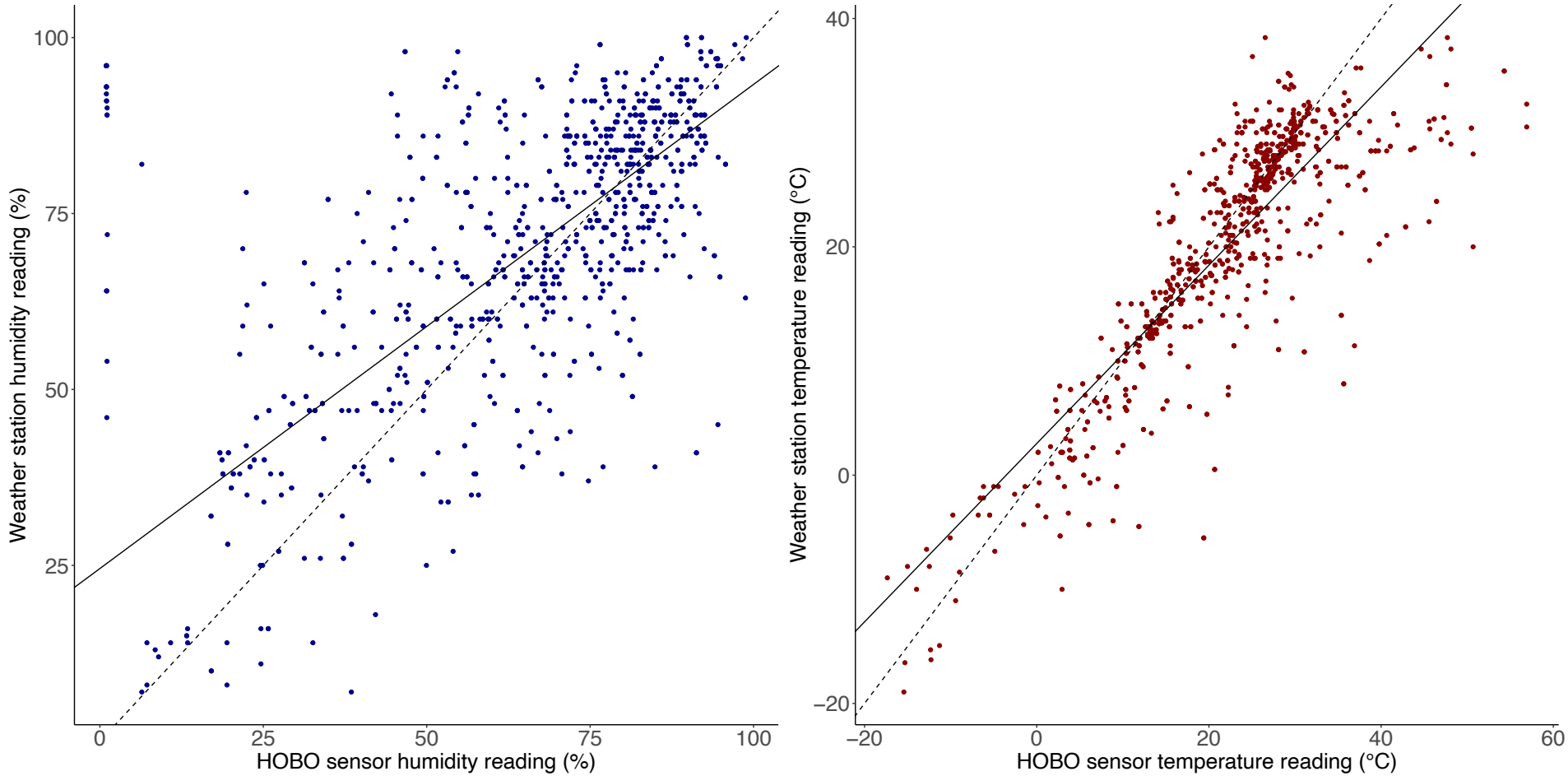


Figure 4-4: Temperature and humidity readings taken during global air sampling, as measured by HOBO sensors and by the nearest weather stations. The solid lines are linear models between HOBO sensor and weather station data (the linear model on the humidity plot excludes the outliers where HOBO humidity = 1). The dashed lines are 1:1, where the two data sources would be in agreement.

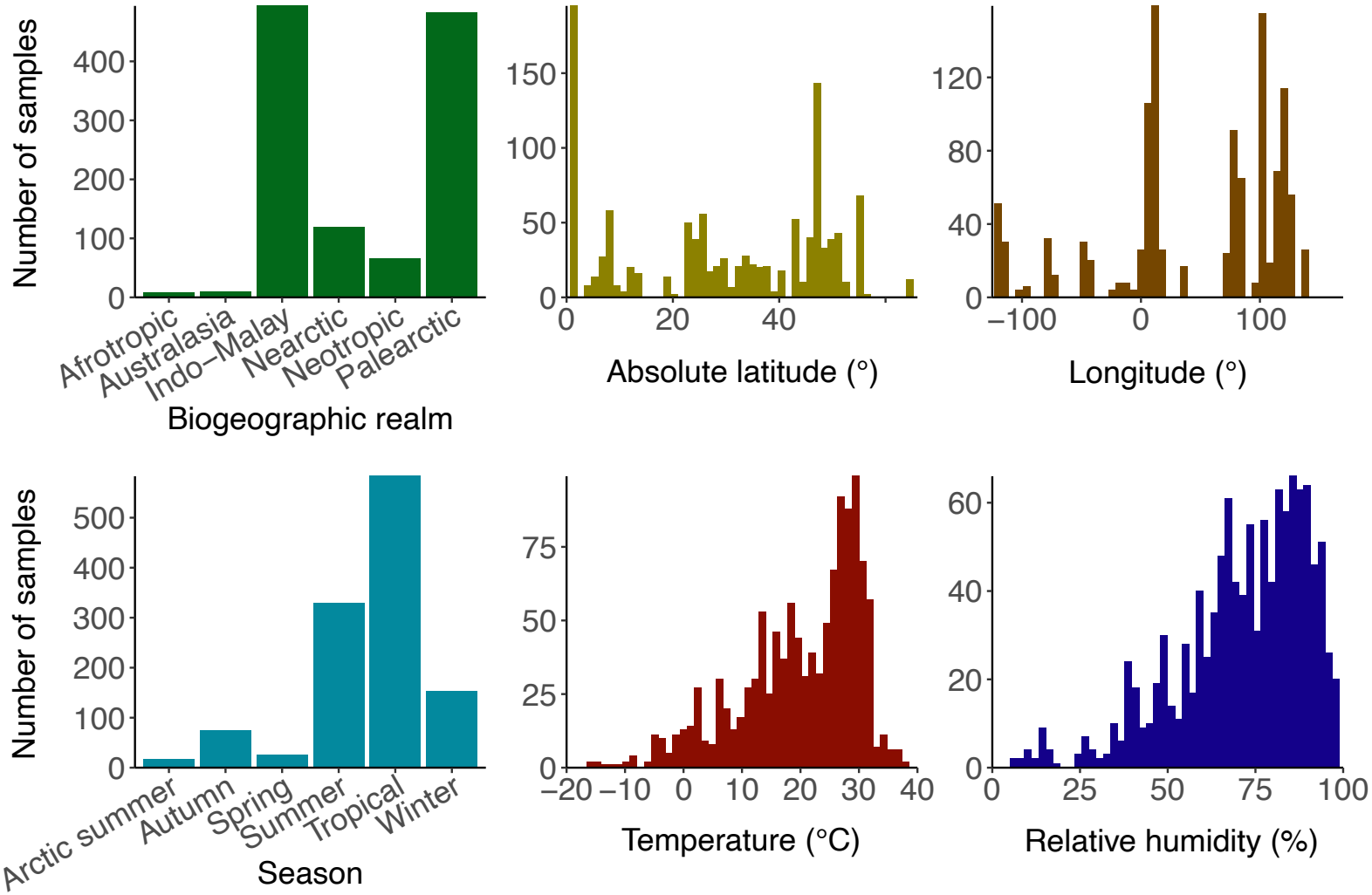


Figure 4-5: Distributions of environmental metadata for 1,180 metagenomics air samples. A large portion of samples were taken in Southeast Asia (Indomalayan realm, ~5° latitude, ~100° longitude, tropical season, high temperature and humidity) and in the summer in Germany (Palearctic realm, ~50° latitude, ~10° longitude).

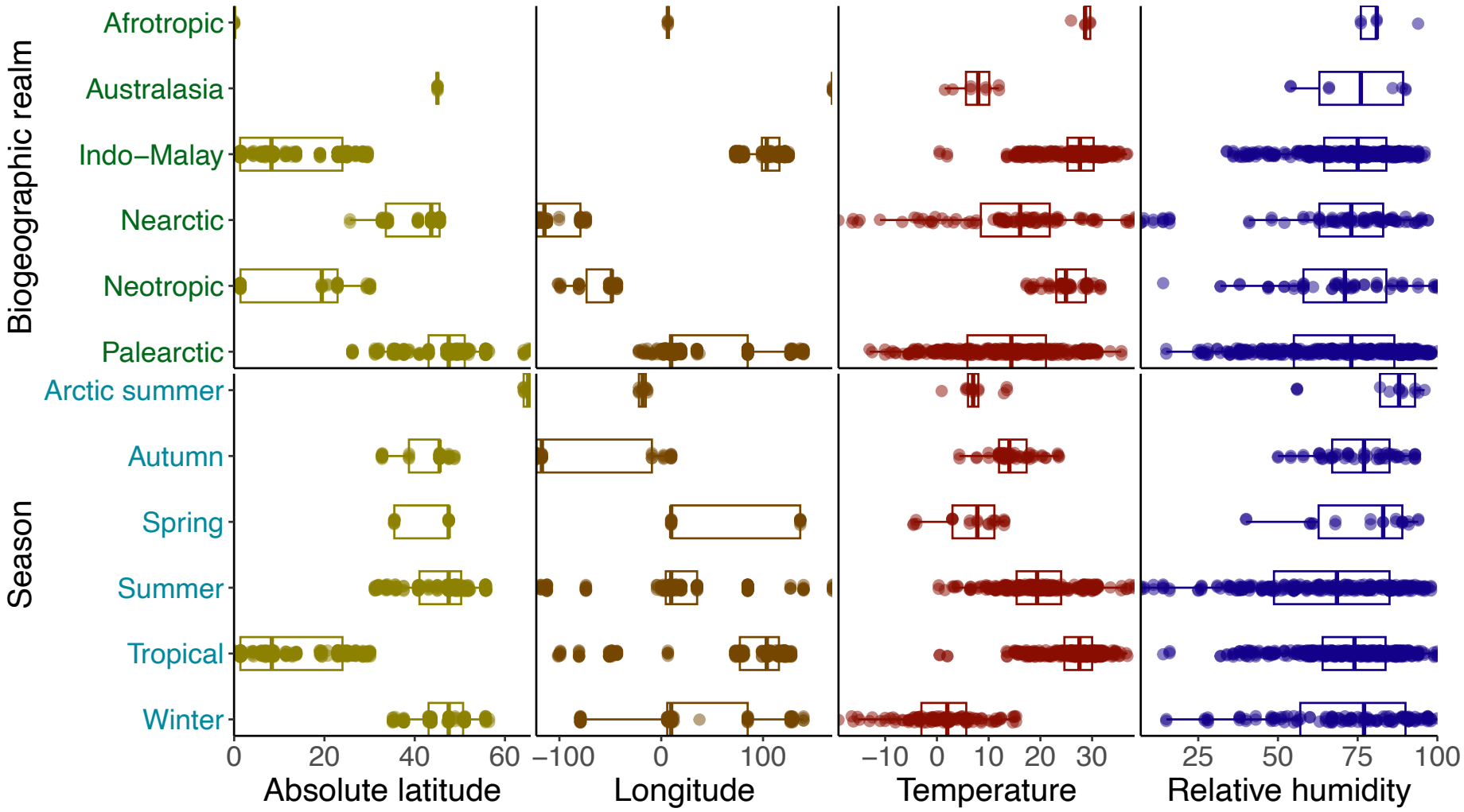


Figure 4-6: Distributions of the four continuous variables above across different levels of the two categorical variables above (biogeographic realm and season). High temperature samples in this dataset were associated with tropical climates but humidity was not clearly associated with any season or realm.

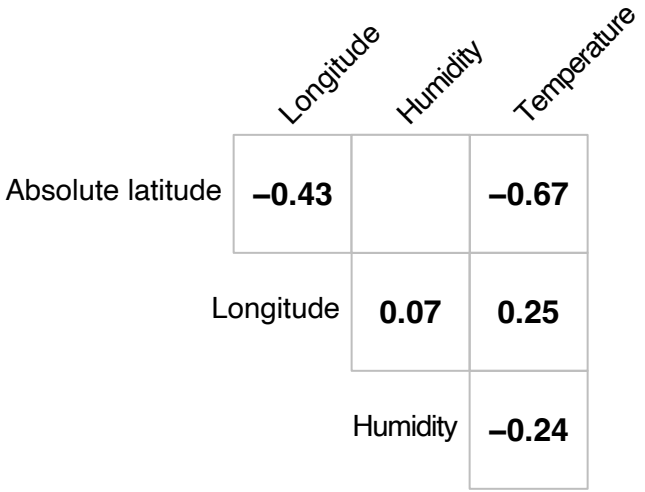


Figure 4-7: Pearson correlation coefficients between the four continuous variables in the global metagenomics dataset. Values are shown for correlations that were significant at the p = 0.05 level. Absolute latitude had no significant correlation with humidity.

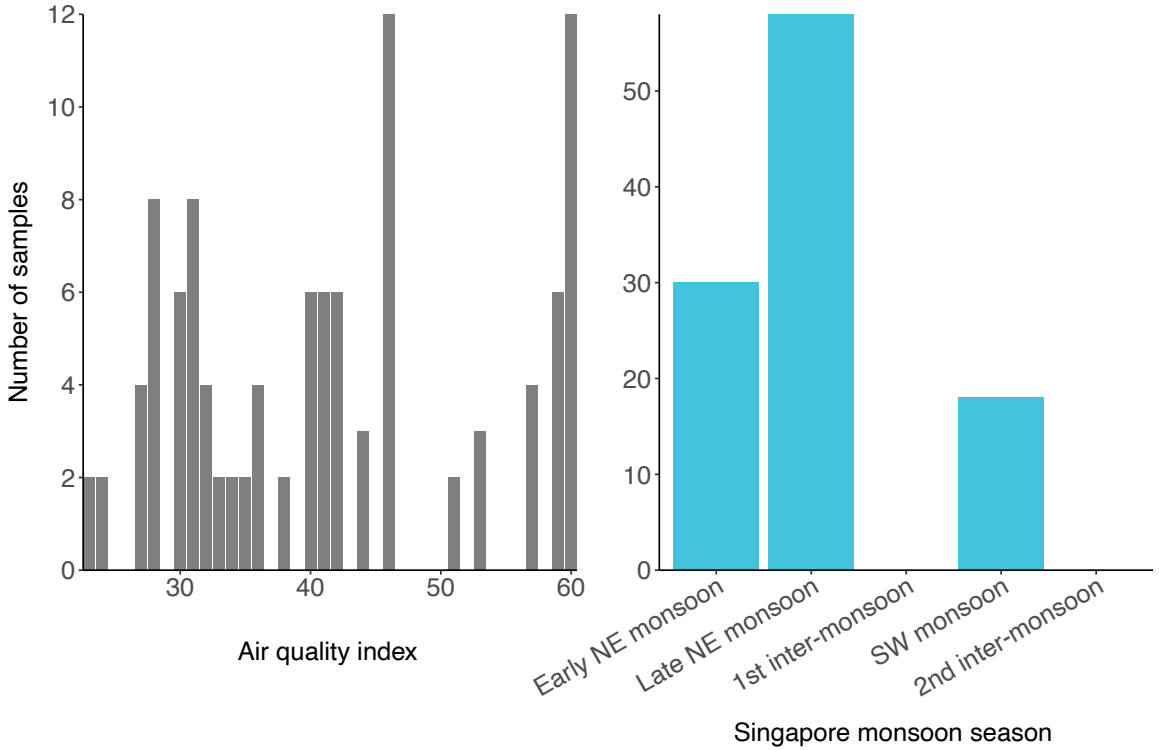


Figure 4-8: Distributions of air quality index and monsoon season for 106 air samples taken in Singapore.

Ordination of global metadata and the global distribution of Priestia species

To examine the geographical distribution of the two *Priestia* species, I calculated the ratio between the two species rather than the individual species abundances, because the abundances obtained from metagenome data are always relative abundances that are dependent on the other species in the sample. Any increase in the individual relative abundance of each species could be either due to the species being more numerous or some of the many other species in the sample being more numerous. Instead, a change in the ratio of *P. aryabhattai* and *P. megaterium* is evidence of an environment that benefits one over the other, which is the motivation behind this study.

The geographic map of the global dataset (Figure 4-9) shows the sums of the *P. aryabhattai* and *P. megaterium* relative abundances in each location. This projection based only on latitude and longitude shows a trend toward lower *aryabhattai/megaterium* ratios in Europe and higher ratios (but still less than 50%) in tropical regions, especially Southeast Asia. The smaller number of samples from other tropical regions makes it difficult to discern how robust this relationship is for Africa and South America. Since the geographic map includes only latitude and longitude information, I created an ordination using the other environmental data to show the ratio of the two species (Figure 4-10). The ordination can be understood in basic terms as an extension of the map, but using the temperature, humidity, biogeographic realm, and season data rather than only latitude and longitude.

The ordination on the environmental data of the global dataset showed that most (960 out of 1,180) of the samples had higher abundances of *P. megaterium* than of

P. aryabhattai (Figure 4-10). However, the samples which had more *P. aryabhattai* than *P. megaterium* were mostly in tropical areas (68 out of 80), especially in Singapore, Malaysia, and Indonesia. Among other tropical regions, São Tomé (eight samples) had a high *aryabhattai/megaterium* ratio, as did Brazil (50 samples), but Ecuador (eight samples) and the Maldives (eight samples) did not follow this pattern. The inconsistency in the high *aryabhattai/megaterium* ratio among the South American regions that were sampled, and the low sample count in São Tomé, make it unclear whether *P. aryabhattai* is more abundant in all tropical regions or only in Southeast Asia.

This result suggests that the two species may have different selection constraints to environmental condition(s) which are more common in the tropical areas, especially in Southeast Asia. However, the trend of higher *P. aryabhattai* abundances in tropical areas was not robust. There were still many individual Southeast Asia samples that had more *P. megaterium* than *P. aryabhattai* (Figure 4-12). These results indicate that the environmental condition which determines *P. aryabhattai's* preference to tropical areas is not clear and needs to be investigated by a different approach.

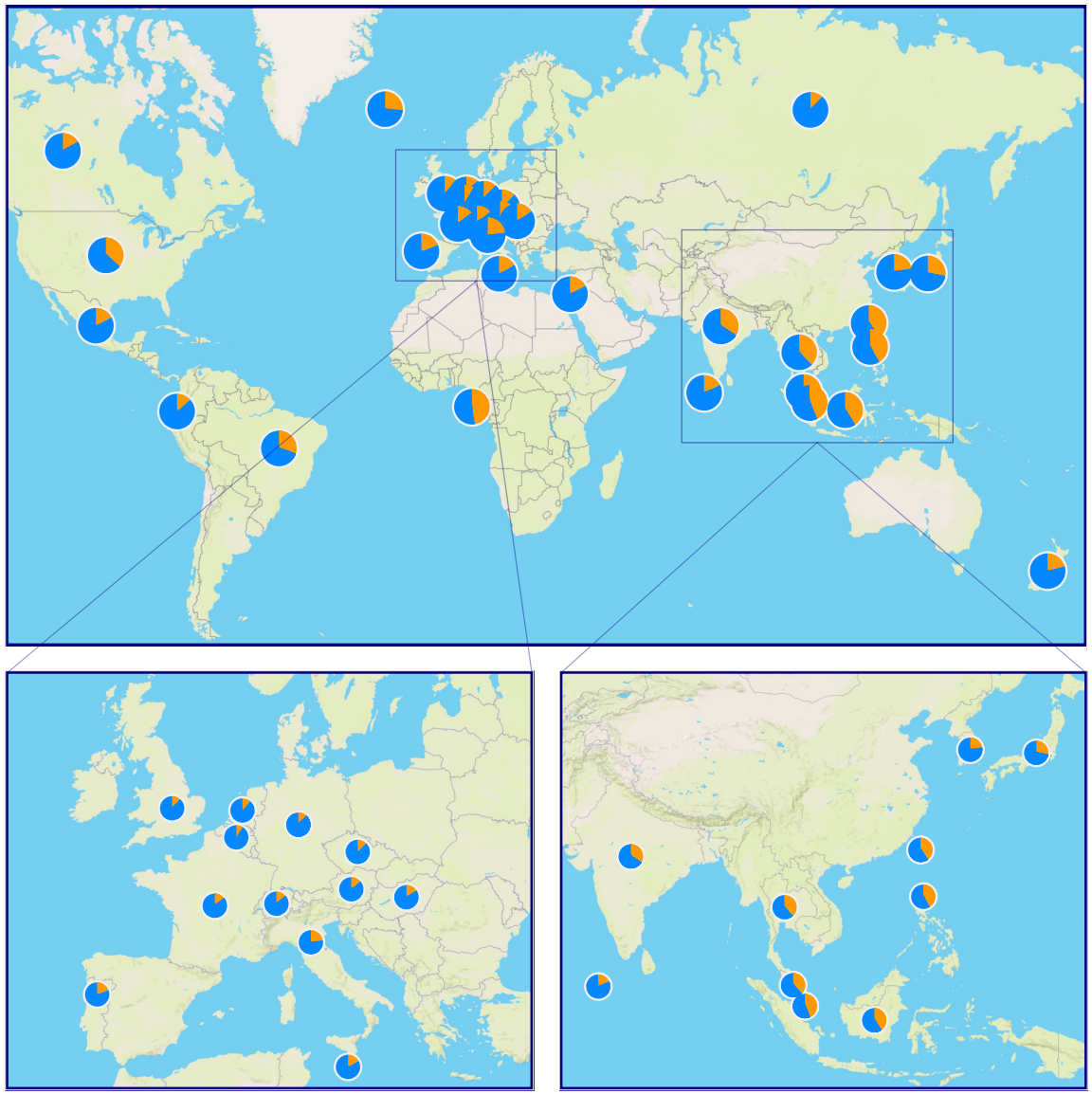
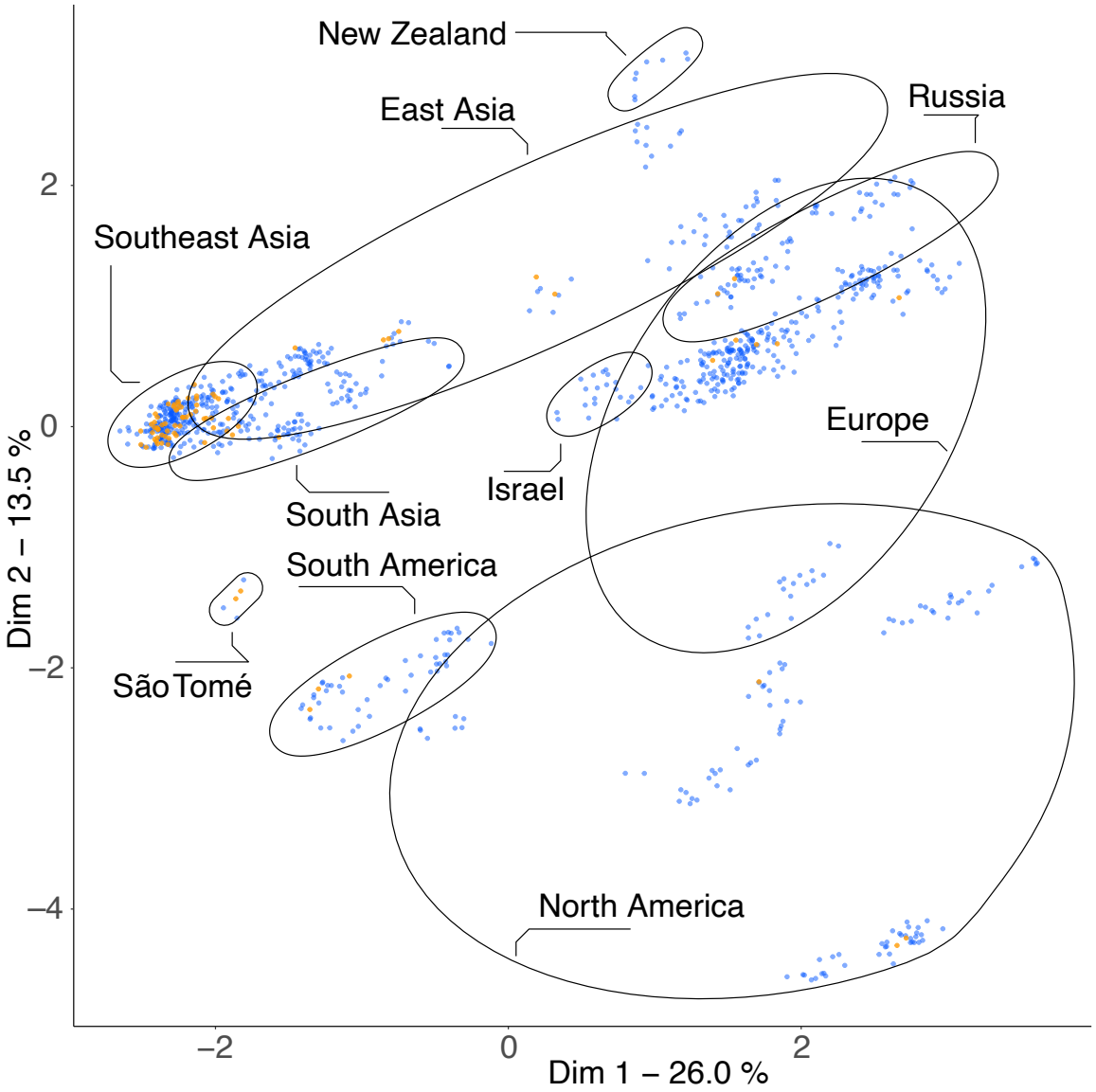


Figure 4-9: Proportions of relative abundance of *P. megaterium* (blue) vs *P. aryabhattai* (orange) in each country of a global air sampling dataset, with zoomed panels showing the frequently sampled regions of Europe and Southeast Asia. One pie chart is shown for each country, with the pie segments being the sum of each sample's *P. megaterium* relative abundance and the sum of each sample's *P. aryabhattai* abundance. *P. megaterium* was generally the more abundant species, but the average proportion of *P. aryabhattai* was higher in Southeast Asia than in other regions.



Most abundant species

- *Priestia aryabhattai*
- *Priestia megaterium*

Figure 4-10: Ordination of environmental metadata of 1,180 air samples from around the world. Samples are coloured blue if the relative abundance of *P. megaterium* was higher than that of *P. aryabhattai*; orange points had more *P. aryabhattai* than *P. megaterium*. The ordination uses absolute latitude, longitude, temperature, humidity, season, and biogeographic realms. The samples where the abundance of *P. aryabhattai* was higher than *P. megaterium* were almost all found in Southeast Asia. Points are randomly jittered by up to 0.1 on both axes in order to show overlapping points.

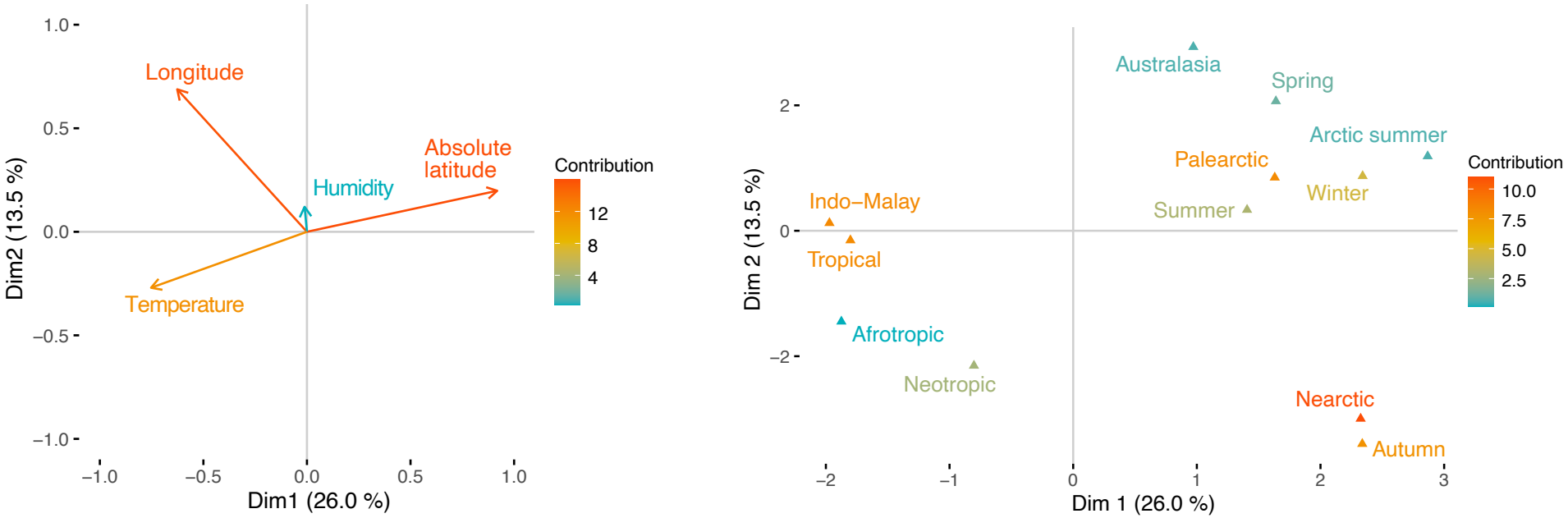


Figure 4-11: Effects of individual quantitative (left) and qualitative (right) variables in the ordination of environmental metadata of 1,180 air samples from around the world (Figure 4-10). The Southeast Asia cluster in the ordination is associated with higher temperatures. Temperate seasons (spring, summer, autumn, winter) had similar frequencies of high-*aryabhattai* samples. Humidity had little effect on the ordination.

For the global dataset, the random forest (RF) and SVM classifiers attempted to predict which of the two species had the highest relative abundance, based on the environmental data, using 80% training data and 20% testing data. The random forest classifier was run with data on latitude, longitude, temperature, humidity, and season. Season was removed for the SVM classifier because SVM is designed for numerical data only. The RF and SVM predictions had 83% and 73% accuracy respectively (Table 4-3).

Table 4-3: Confusion matrices of Random Forest and Support Vector Machine classifiers that aimed to predict whether *P. megaterium* or *P. aryabhattai* had the higher relative abundance in global metagenomic samples, based on local environmental data.

RF		Prediction	
		P. megaterium	*P. aryabhattai*
Observation	*P. megaterium*	29	3
	P. aryabhattai	5	11

SVM		Prediction	
		P. megaterium	*P. aryabhattai*
Observation	*P. megaterium*	24	8
	P. aryabhattai	5	12

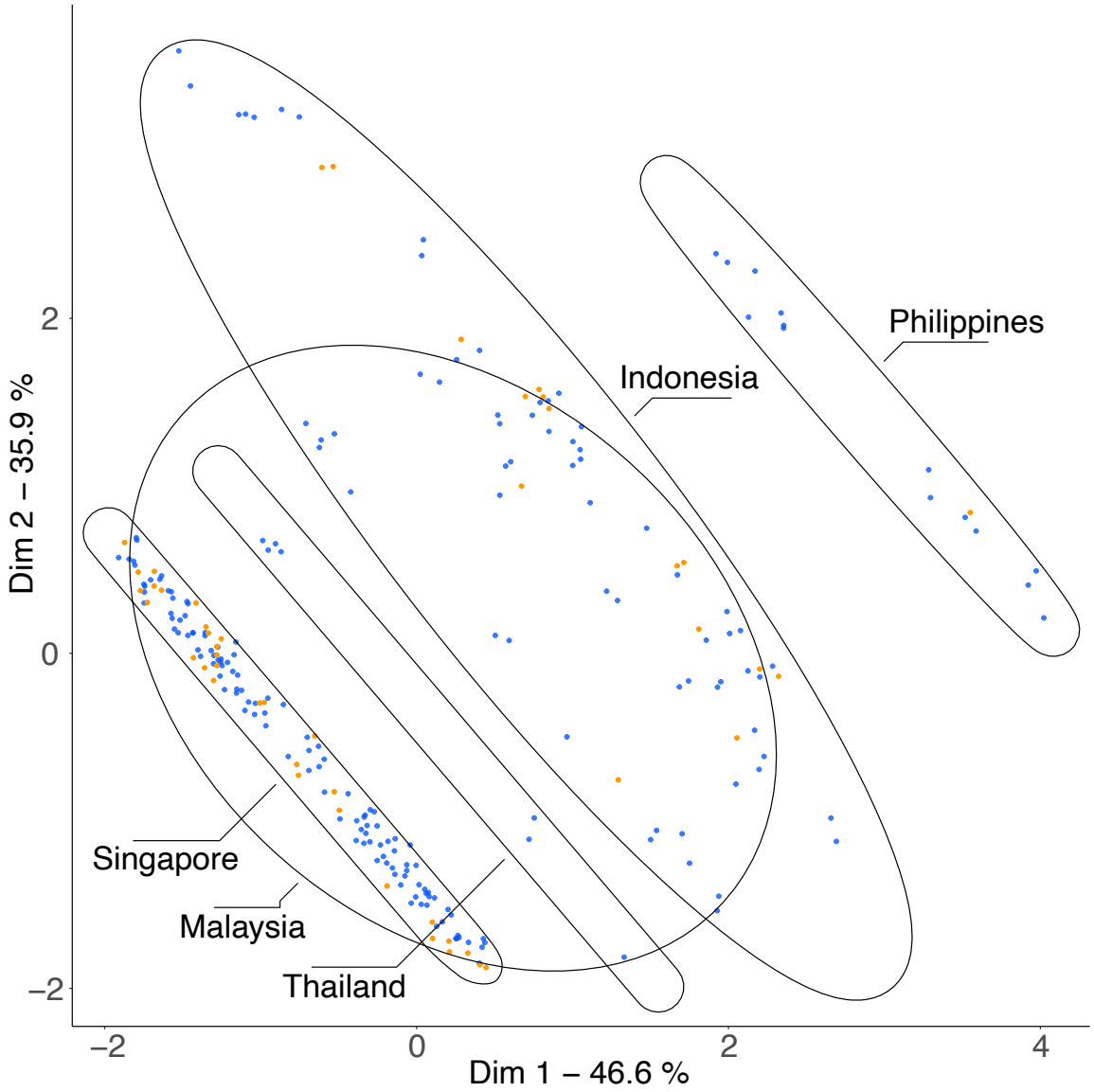


Figure 4-12: Ordination of environmental metadata of 274 air samples from Southeast Asia (a subset of the global samples shown in Figure 4-10. Samples are coloured blue if the relative abundance of *P. megaterium* was higher than that of *P. aryabhattai*; orange points had more *P. aryabhattai* than *P. megaterium*. The ordination uses absolute latitude, longitude, temperature, and humidity. Compared to the rest of the world, Southeast Asia has a much larger number of samples in which *P. aryabhattai* was more abundant than *P. megaterium*, but the environmental conditions that cause this difference are unknown, as shown by the *aryabhattai*–majority samples being mixed among the *megaterium*-majority samples, rather than separated from them.

The ordination using only samples from Southeast Asia showed no obvious tendency on the ratio of the two species (Figure 4-12). This result suggests that the environmental condition which is preferred by *P. aryabhattai* is common across the Southeast Asian area included in this study, but that it may be inconsistent between locations or over time.

The RF and SVM classifiers gave 69% and 62% accuracy on the Southeast Asia samples (Table 4-4). The falling performance of the models compared to the global samples reflects their reliance on the dichotomy between Southeast Asian samples with higher probability of *P. aryabhattai* excess and samples outside Southeast Asia with a lower probability.

Table 4-4: Confusion matrices of Random Forest and Support Vector Machine classifiers that aimed to predict whether *P. megaterium* or *P. aryabhattai* had the higher relative abundance in Southeast Asia metagenomic samples, based on local environmental data.

RF		Prediction	
		P. megaterium	*P. aryabhattai*
Observation	*P. megaterium*	17	6
	P. aryabhattai	3	3

SVM		Prediction	
		P. megaterium	*P. aryabhattai*
Observation	*P. megaterium*	18	2
	P. aryabhattai	9	0

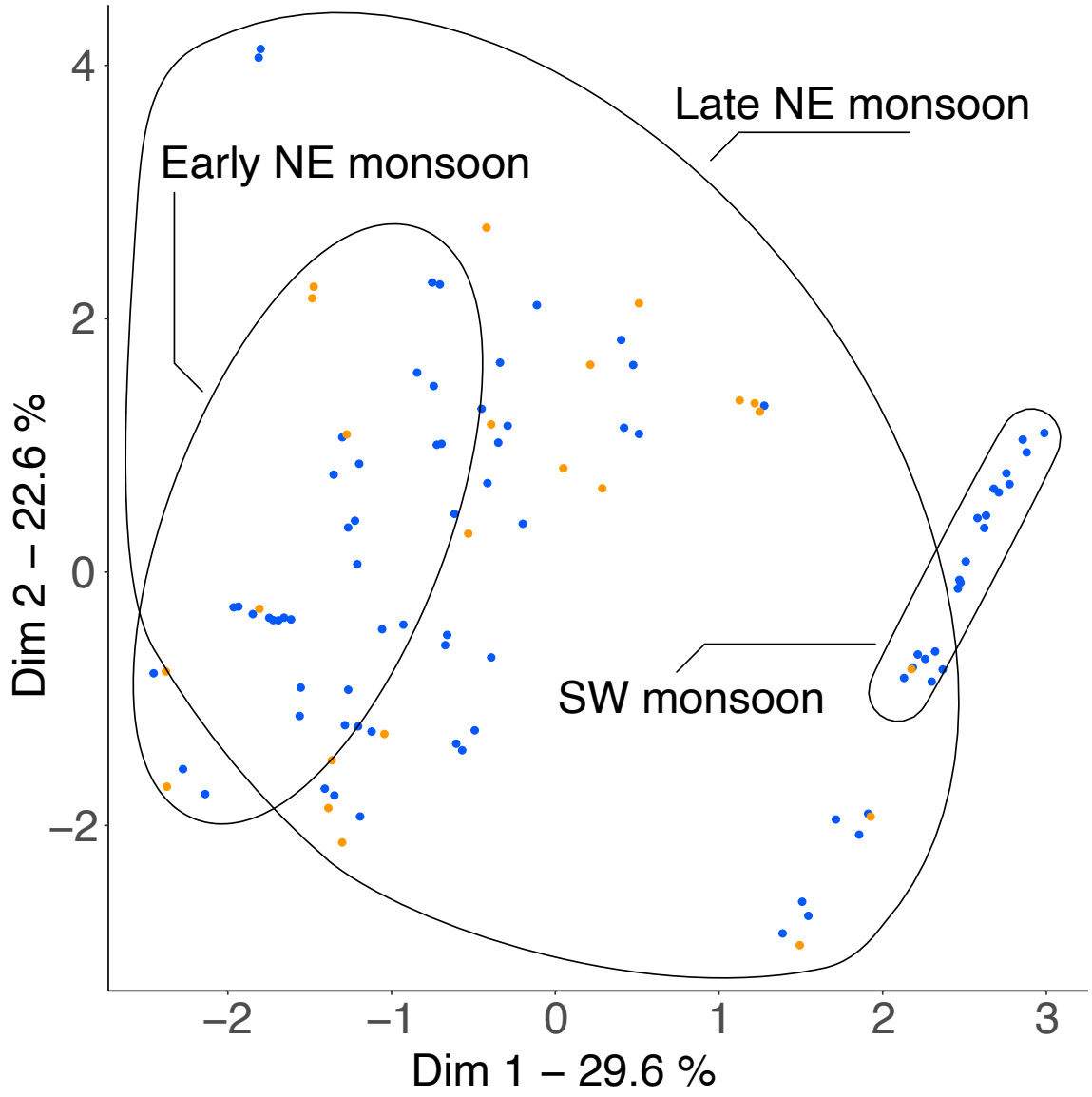


Figure 4-13: Ordination of environmental metadata of 106 air samples from Singapore (a subset of the global samples shown in Figure 4-10. Samples are coloured blue if the relative abundance of *P. megaterium* was higher than that of *P. aryabhattai*; orange points had more *P. aryabhattai* than *P. megaterium*. The ordination uses, temperature, humidity, monsoon season, and five measures of air quality: PM2.5, PM10, O_3, SO_2, and CO. The addition of monsoon and air quality data to the ordination did not separate the majority-*aryabhattai* samples from the majority-*megaterium* samples.

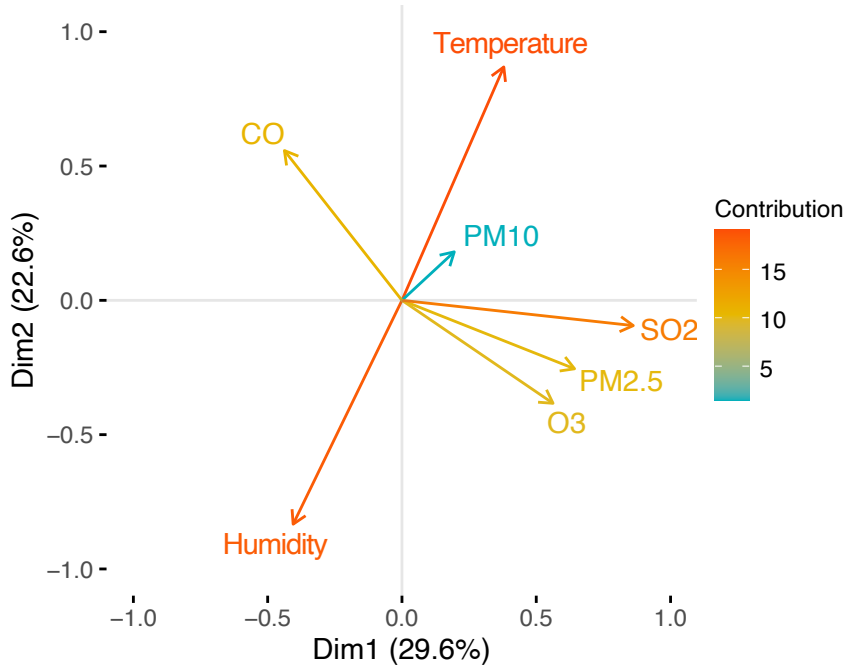


Figure 4-14: Effects of individual variables in the ordination of environmental metadata of 106 air samples from Singapore (Figure 4-13). The only categorical variable, monsoon season, is shown in Figure 4-8. The large range of temperature, humidity, and air pollution readings did not affect the prominent species in each sample.

For the Singapore samples it was possible to add in more environmental data pertaining to air quality (PM2.5, PM10, O3, SO2, CO) and monsoon season. However, the ordination showed no pattern of species ratios (Figure 4-13) and the machine learning classifiers were still unable to show or predict the prevalent species between *P. aryabhattai* and *P. megaterium*. The RF and SVM classifiers gave 64% and 64% accuracy (Table 4-5). The categorical variable of monsoon season was removed for SVM.

To investigate how the species abundances and ratios differed with individual environmental variables, I plotted the relative abundances of megaterium and aryabhattai in each sample of the global dataset, broken down by the geographic region, season, temperature, and humidity (Figure 4-15). Within each plot, the levels of the variable on the y axis are sorted in descending order by the average ratio of

P. aryabhattai to *P. megaterium* in each sample. The geographic region and season plots (A and B) clearly show the bias toward Southeast Asia and the other tropical areas in the high-aryabhattai samples. The temperature plot (C) shows a trend of higher temperature samples having higher average proportions of *P. aryabhattai*, with an exception at –20 to –10 °C where the small number of samples at this temperature range had a higher average *aryabhattai/megaterium* ratio than –0 to 10 °C and –10 to 0 °C. The plot of humidity ranges (D) illustrates the lack of relationship in the global dataset between humidity and species ratio.

Table 4-5: Confusion matrices of Random Forest and Support Vector Machine classifiers that aimed to predict whether *P. megaterium* or *P. aryabhattai* had the higher relative abundance in Singapore metagenomic samples, based on local environmental data.

RF		Prediction	
		P. megaterium	*P. aryabhattai*
Observation	*P. megaterium*	8	3
	P. aryabhattai	2	1

SVM		Prediction	
		P. megaterium	*P. aryabhattai*
Observation	*P. megaterium*	8	0
	P. aryabhattai	5	1

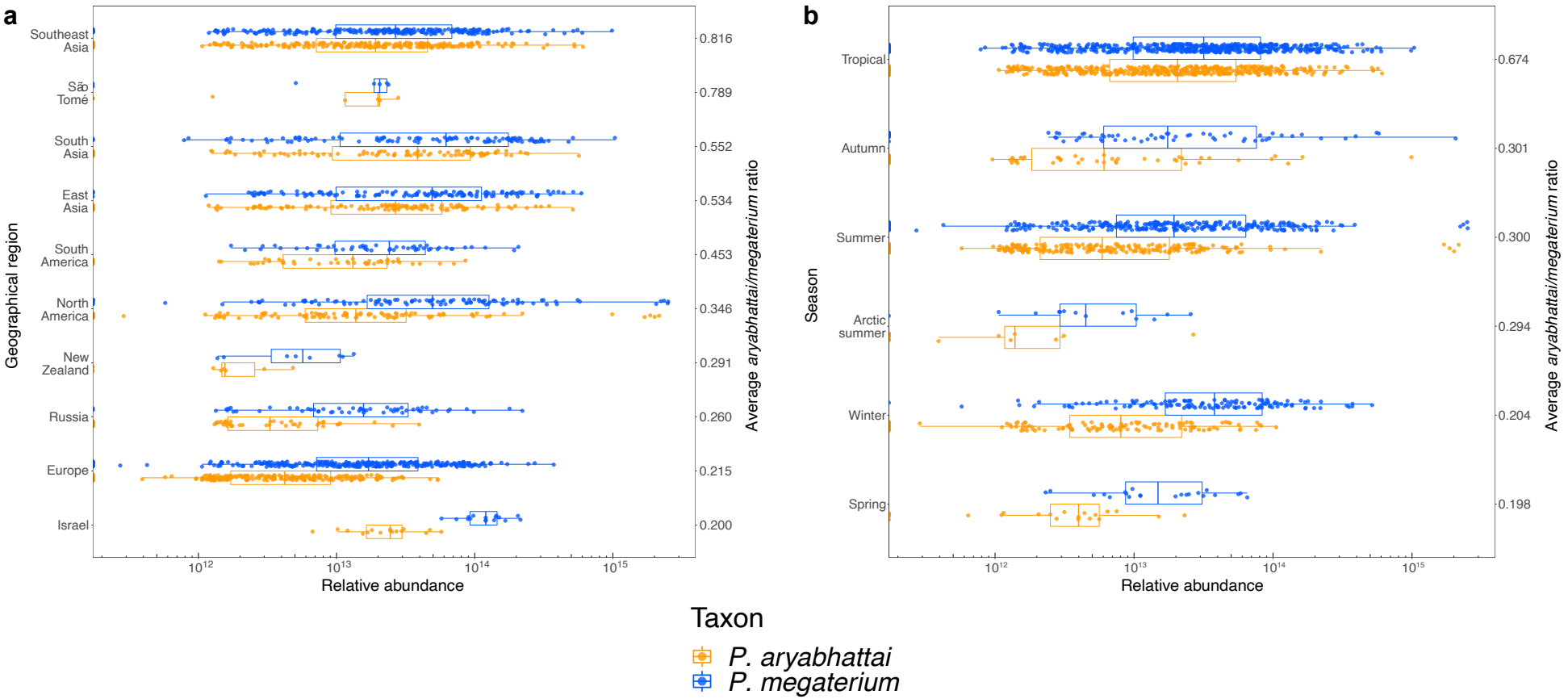


Figure 4-15 (continued next page): Boxplots of *P. aryabhattai* and *P. megaterium* relative abundances in samples from different geographical regions (A), seasons (B), temperature ranges (C), and humidity ranges (D) in a global metagenomics sampling project. Each sample is represented by one orange and one blue point giving the abundances of the two species for that sample. The variable levels on the y-axes are ranked by the average ratio of *P. aryabhattai* to *P. megaterium* in each sample, with the categories that had the highest ratios at the top of the plot.

a: The average ratio of *P. aryabhattai* to *P. megaterium* in air samples was highest in Southeast Asia, followed by nearby regions and other areas at low latitude.

b: Samples taken in tropical climates had higher average *P. aryabhattai* to *P. megaterium* ratios.

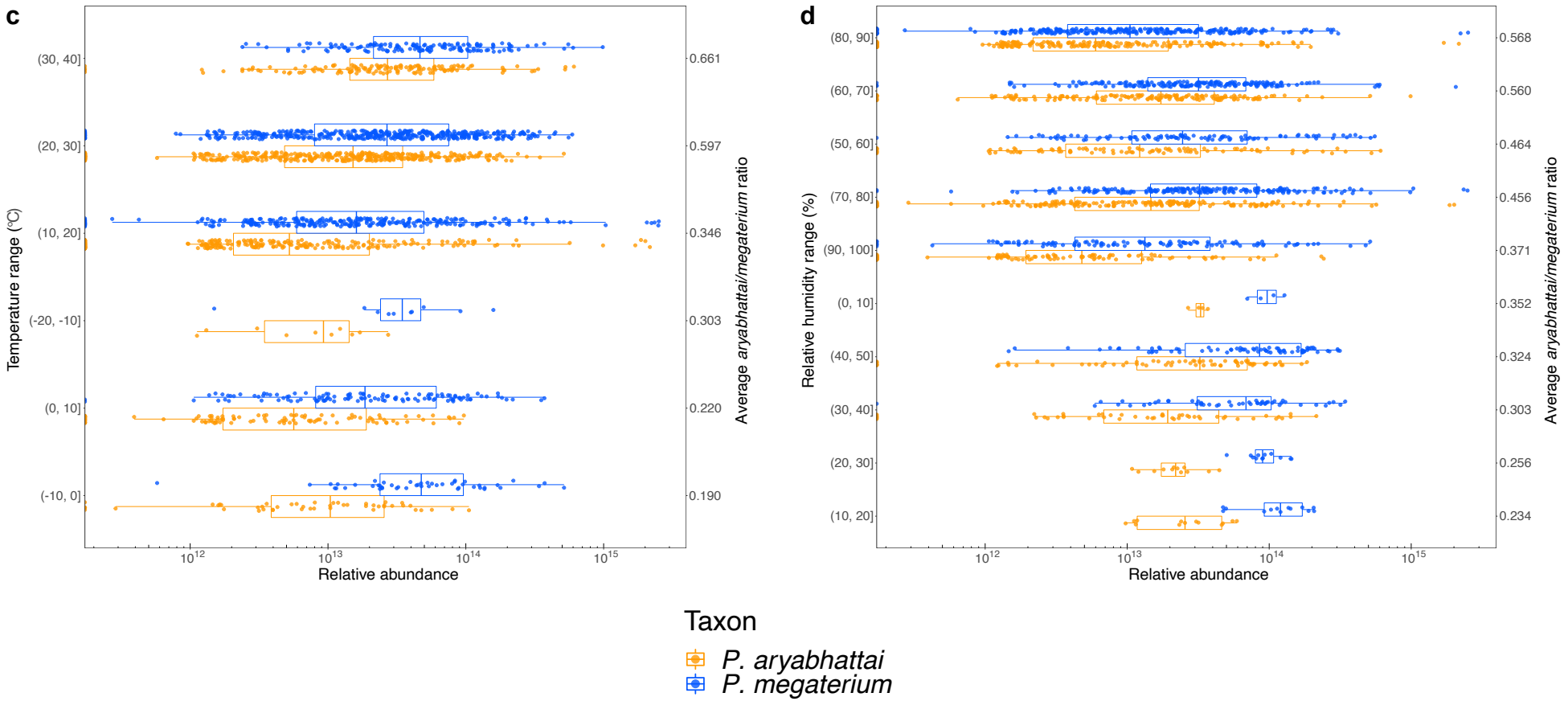


Figure 4-15 continued:

c: The average *P. aryabhattai* to *P. megaterium* ratio increased with higher temperature ranges for temperatures above 10 °C.

The abundance of both species, as well as the ratio between them, were significantly different across geographic regions, seasons, and temperature ranges ($p < 0.01$, Kruskal-Wallis test).

d: Relative humidity correlated poorly with the species ratio.

Discussion

The results of the metagenomics analysis did not identify a clear niche in which *P. aryabhattai* was more abundant than *P. megaterium,* since all three experiments showed that *P. megaterium* was almost always the more abundant species. I made DNA alignments of all *Priestia* reads to the reference genomes of both species, which increased the number of *P. aryabhattai* reads in the sample, in order to account for the possibility that *P. aryabhattai*, as a more recently discovered species that was not previously accepted as being distinct from *P. megaterium*, had less representation in the databases used for species identification, which could lead to a *P. megaterium* bias in the species identification of the reads. It seems that the bias shown in most samples toward *P. megaterium* was due to a real bias in the number of cells present in the environment rather than a systematic error in the identification of reads. These results suggest that *P. megaterium* may have a higher rate of growth, spore dispersal, or survival in the environments that were sampled, except in certain unknown conditions in Southeast Asia.

Upon its discovery in an air sample from 41 km altitude, *P. aryabhattai* was suggested to be uniquely adapted to survival at high altitudes due to its superior UV resistance over its nearest phylogenetic neighbour, *P. megaterium* (76). The results of that study are not directly contradicted by those described in the present study because we did not sample up to 41 km, where the UV radiation experienced by the bacteria would be much higher (286), and the colony forming ability of the *P. aryabhattai* and *P. megaterium* in the samples could not be tested. However, our results showed that *P. aryabhattai* was

significantly less abundant than *megaterium* at heights of up to 3500m, and that both species decreased in relative abundance at similar rates as altitude increased, leading to similar species ratios across altitude levels. In the absence of evidence to suggest that this trend would reverse at higher altitudes, the findings of this study suggest that *P. aryabhattai* is not more tolerant to high-altitude life than *megaterium*. However, there may also be variation of UV resistance within *P. aryabhattai,* and future studies can check for vertical stratification of UV resistance gene variants.

In the tower experiment (Figure 4-2), the relative abundance of both species actually increased at 200 m compared to ground level. The previously published study on this dataset noted that a change in the potential temperature gradient between 0m and 200 m caused air convection and vertical mixing of the airborne microbial community (273). If this effect is responsible for the increased abundances of the two *Priestia* species at 200 m, then these species must be benefitting from the convection more than the other species present in order for their relative abundances to increase, possibly because they can produce spores that are smaller and lighter than the cells of other species. As a follow-up study, it would be interesting to compare the changes in relative abundances of spore-forming species in general versus other species as height increases.

I used an additional metagenomics dataset, consisting of air samples taken at 1.5 m above the ground at locations around the world, to search for differences in the global distribution of the two species. Most samples had a higher abundance of *P. megaterium* than of *P. aryabhattai*, with samples taken at higher temperatures tending to have increased ratios of *P. aryabhattai* to *P. megaterium.* The results showed that *P. aryabhattai* was more likely to exceed *P. megaterium* in relative abundance in

samples from Southeast Asia than in any other region, although *P. megaterium* was still more abundant in most samples. This effect could not be linked to changes in temperature, humidity, location, air quality, or monsoon season, and so it is unknown whether the large fluctuations in the ratio of airborne *P. aryabhattai* and *P. megaterium* between samples are correlated to an unexamined factor or whether they are simply unrelated to the local environmental conditions. Given that the two species are very closely related, with a maximum whole-genome distance of just over 6% ANI, it is possible that they simply do not differ in this regard. The addition of more samples from other tropical regions, such as South America and Africa, would show if the occurrence of higher *P. aryabhattai* to *P. megaterium* ratios is a Southeast Asian phenomenon or a general effect of tropical climates.

Another limitation of this study was the exclusive use of air sample metagenomics, rather than samples from other environments in which *Priestia* are known to grow, such as the rhizosphere. Given the finding of plant-growth-promotion associated genes with divergent sequences between the two species (Chapter 3), an interesting avenue of research would be to compare the two species' abundances in rhizosphere samples with different soil conditions and plant species.

The motivation of this study was to compare the genomes and ecology of the two *Priestia* species in order to assess the claim that they represent distinct groups. Because of that goal, the focus of the metagenomics analysis was limited to using the species ratio to search for differences between the distributions of the two species under all possible conditions, in order to uncover any environmental factors which benefit one

species over the other. However, future research can move past this comparative approach and examine each species on its own merits.

Conclusions

I have shown that *P. megaterium* and *P. aryabhattai* form separate, robust, monophyletic clades, and that the whole genome distances by ANI and dDDH are sufficient to classify them as separate species. This study resolves the taxonomic uncertainty regarding this group, and provides a practical and accessible tool for species identification: PCR primers targeting the species-specific genomic regions.

The results of this study also show that the two species are good examples of Hanage's concept of fuzzy species (44). In addition to the separate *megaterium* and *aryabhattai* clades, there also exist two small clades on intermediate branches. These clades appear to have mosaic genomes due to recombination between the larger *megaterium* and *aryabhattai* clades. Most of these genomes are sequenced from environmental samples, not culture collections or industrial strains, suggesting an extensive overlap in the ecological occurrence of the two species. It is not known whether this fuzzy species state is due to an ongoing but incomplete speciation, or a despeciation due to reintroduction of the two species to the same habitat (47), or if the partial separation of two species can be a stable and lasting state. A deeper analysis of the rates of recombination between the clades may help to answer this question (47).

Bacterial species concepts that make use of the 70% DNA hybridisation cut-off to delineate species were previously criticised for using an arbitrary threshold (31), but later research using whole-genome sequence data showed a natural discontinuity in genome similarity values. Genome pairs from the same species are usually over 95% similar by ANI, and genomes from different species are usually 83% or lower, but few

genome pairs are observed between these values (14). The results of the whole-genome distance comparison in Chapter 2 showed a clear but narrow genetic discontinuity between the *megaterium* and *aryabhattai* clades, with the recombinant clades showing intermediate ANI values (Figure 2-5). Genome pairs from the *megaterium* clade had a minimum of 96.17% ANI, and genome pairs from the *aryabhattai* clade had a minimum of 96.78% ANI, but genome pairs from opposite clades had a maximum of 95.58% ANI. The next closest species, *Priestia flexa*, has a maximum of 80.94% ANI to genomes of the *P. aryabhattai/megaterium* group (Figure A-1). These results support the concept of bacterial species delineation by ANI discontinuity, given the understanding that some species groups can be more diverse than others and that closely related groups may have indistinct boundaries.

It has been debated whether groups of bacteria with high rates of horizontal gene transfer can show a consistent phylogenetic signal, given that different genes within a genome will have different evolutionary histories (287). In practice, phylogenetic studies commonly use bifurcating tree models that do not account for recombination, although alternate models for reticulate evolution have also been developed (288). The inherent assumption behind the use of bifurcating trees is that, despite the occurrence of horizontal gene transfer, there remain some parts of the genome that have been undisrupted and can show an internally consistent phylogenetic signal — a clonal frame (134). Core housekeeping genes have been argued to be good candidates for phylogenetic studies because the cell would incur fitness costs if these genes were to be horizontally transferred (32), but multiple housekeeping genes may not give exactly the same phylogeny (289). It was uncertain if a small number of genes could accurately represent the phylogeny of the whole genome (290).

The results of this study showed a broad agreement between the tree structure produced by housekeeping gene sequences and whole-genome similarities (Figure 2-1), suggesting that the latter method was robust against conflicting signals from recombination in other parts of the genome within the *megaterium* and *aryabhattai* clades. However, these methods could not reliably place the interior clades of mosaic genomes, for which reticulate models (Figure 2-1c) or the filtering out of recombinant regions (Figure 2-4) may be more suitable. Thus, it cannot always be assumed that a selection of core genes will give an accurate phylogenetic tree, and conflicting phylogenetic signals must be investigated.

In this study I developed two useful procedures. First, a dN score which can identify orthologs which have greatly diverged between species while also being conserved within species. This method is useful for examining the functional differentiation between two groups of sequences. Second, an algorithm for species reassignment of metagenomic reads using DNA sequence data. Since metagenomics analysis uses amino acid sequences for species identification (278), a second-pass species assignment using DNA sequences is useful for the fine resolution of closely-related species, for analyses of various purposes. Both procedures can be applied to any species.

I used ortholog clustering and calculations of dN and dS between orthologs to identify the genes which show the greatest divergence between the two species, providing candidate genes with potential functional differences that can be investigated further by future research. Future work should aim to determine whether the sequence changes have affected these gene functions. I found that calculations of dN-dS, a test to show whether sequences are evolving neutrally or under positive or purifying selection, were

unable to provide such candidate genes showing divergence between the two species. For the top 100 orthologs which clearly showed extensive non-synonymous substitutions, no genes were detected under positive selection between the species based on dN-dS values with the Z-test of selection because the number of synonymous substitutions was higher than the number of non-synonymous substitutions.

Further investigation can be done in a more applied manner, starting with a bioinformatics approach using these candidate genes. For those genes which have been previously well characterised, with the protein's active sites being known, the substitutions which are in active sites can be identified. Cultures of both species can then be tested *in vivo* for observable changes in protein structure, expression level, protein-protein interactions, and cell growth rate and survival between the two species under different environmental conditions.

As for which orthologs are most deserving of further research, the top 100 orthologs by dN score in my results identified several themes in which multiple orthologs associated with a particular ortholog all showed substantial sequence changes. These included genes involved in cobalamin (vitamin B12) synthesis and use, the components of the PTS sugar transport systems, the sporulation process and its regulation, and flagellar assembly. Experiments to verify the functional differences in the two species should therefore focus on the roles of these systems as well as their less obvious effects, such as the vulnerabilities of PTS systems to antimicrobial molecules and viruses.

The genes related to sporulation that differed between species included a gene with amino acid substitutions that regulates the sporulation process (Figure 3-8), and a

protective spore protein (Table A-4) that had a higher gene copy number in the *aryabhattai* genome assemblies. Given that mutations in other genes have been shown to affect the sporulation timing of a related species (291,292), future investigations into differences in the dispersal of *P. megaterium* and *P. aryabhattai* should begin with an experimental comparison of the conditions required to induce sporulation and its timing. These orthologs with extensive substitutions between groups could also affect the timing of bacterial lifecycle transitions, such as the conditions under which biofilm formation or motile flagellated cells are favoured.

I found that *P. aryabhattai* had an increase in gene copy number of several genes involved in iron import, a function that is related to plant growth promotion by rhizospheric bacteria. Given that the results of my metagenomics analysis did not support the niche of *P. aryabhattai* as a high-altitude species, but that the comparative genomics showed these changes in plant growth promoting genes, I offer the alternate hypothesis that *P. aryabhattai* may be preferentially associated over *P. megaterium* with one or more tropical plant species that are local to Southeast Asia. This hypothesis aims to also explain the finding from this study that metagenomic samples from Southeast Asia were more likely than those from other regions to show a higher abundance of *P. aryabhattai* than of *P. megaterium.* As discussed earlier, both of these species have received research attention for their plant growth promotion and cobalamin production (86,91,266), but they have not been directly compared on these functions. The hypothesis of a Southeast Asian plant niche can be tested by further metagenomics experiments which compare the abundances of the two species in samples of rhizospheric soil and plant roots from Southeast Asia and other tropical and temperate regions.

The results of Chapter 3 suggest that two smaller clades within the *P. megaterium/aryabhattai* group consist of mosaic genomes that are likely to have formed by frequent recombination between the larger clades. This recombination would have required strains from the *megaterium* and *aryabhattai* clades to coexist in close proximity, and so an extensive overlap in the ecological and geographical distributions of the two groups was to be expected. It is unclear whether the lower general abundance of *P. aryabhattai* in the air metagenomics data was the result of the clade having less diversity than *P. megaterium*, or if there are more unsampled environments which would show higher abundances of *P. aryabhattai*.

The methodology of using species abundances in air samples to hypothesise differences in adaptation to niches on the ground assumes that the two species are sporulating at equal rates, but the observed substitutions in sporulation genes may affect the rate or conditions for sporulation. Differences in the ability of each species to colonise an environment, such as soil or a plant host, may also affect their ability to form biofilms from which large numbers of spores can be created. Further metagenomics studies can be used to compare the abundances of the two species in tropical and non-tropical soils, and assess the correlation between *Priestia* abundances in soil and in the surrounding air. Using the updated taxonomy presented here, culture-based studies of plant-associated *Priestia* strains can also compare the two species' colonisation and growth effects on plants from different climates.

Appendix

Table A-1: Genome assembly statistics of new *P. aryabhattai* genomes announced in this study, plus the GenBank representative strain K13 for comparison.

Genome name	*Priestia* clade	GenBank accession	Number of contigs	Genome length (bp)	N50 (bp)	GC content (%)	CheckM completeness	CheckM contamination	CheckM heterogeneity	Prokka CDS	Prokka rRNA (5S, 16S, 23S)	Prokka tRNA
SGAir0178	*aryabhattai*	CP028074-CP028080	7	5,473,470	5,001,814	38.03	98.85	0.03	0	5579	43	148
SGAir0179	*aryabhattai*	CP025620-CP025623	4	5,305,001	5,053,919	38.14	99.43	0.07	0	5363	43	133
SGAir0202	*aryabhattai*	CP028043-CP028049	7	6,362,169	5,077,550	37.22	99.43	7.42	0	6569	43	138
SGAir0257	*aryabhattai*	CP028019-CP028030	13	5,576,818	5,017,599	37.93	99.43	0.1	0	5679	46	145
SGAir0265	*aryabhattai*	CP027997-CP028008	12	5,571,671	5,011,965	37.93	99.43	0.69	60	5678	46	134
SGAir0269	*aryabhattai*	CP027989-CP027996	8	5,516,307	5,028,224	37.94	99.43	0.1	0	5603	41	119
SGAir0414	*aryabhattai*	CP027914-CP027919	7	5,390,002	5,159,113	38.01	99.43	0.09	0	5492	42	130
SGAir0424	Recombinant 1	CP027900-CP027906	8	5,659,875	5,262,388	37.95	99.43	0.03	0	5669	46	156
SGAir0425	*aryabhattai*	CP027889-CP027899	14	6,481,834	5,183,207	37.18	99.43	7.53	0	6604	42	158
SGAir0427	*aryabhattai*	CP027876-CP027885	10	5,628,789	5,103,336	37.94	99.43	0.68	37.5	5700	51	150
SGAir0428	Recombinant 1	CP027870-CP027875	6	5,617,367	5,226,421	37.93	99.43	0.09	0	5687	44	146
SGAir0563	*aryabhattai*	CP027931-CP027939	10	6,572,477	5,017,311	37.19	99.43	7.7	0	6782	46	155
K13	*aryabhattai*	CP024035-CP024037	3	5,254,250	5,035,815	38.17	-	-	-	5310	42	130

Table A-2: Phenotype tests on 12 *P. aryabhattai* isolates and one *P. megaterium* isolate (continued next page). Strains in orange are in the *aryabhattai* clade; blue, in the *megaterium* clade; green, in recombinant clade 1 (see Figure 2-1). There is no test that has been shown to distinguish between these clades. The *P. aryabhattai* and recombinant strains are the new genomes described in this study; the *P. megaterium* strain SGAir0080 is from Kalsi et al. 2019.

Strain name	0 CTRL	1 GLY glycerol	2 ERY erythritol	3 DARA d-arabinose	4 LARA l-arabinose	5 RIB d-ribose	6 DXYL d-xylose	7 LXYL l-xylose	8 ADO d-adonitol	9 MDX methyl-betaD-xylopyranoside	10 GAL d-galactose	11 GLU d-glucose	12 FRU d-fructose	13 MNE d-mannose	14 SBE l-sorbose	15 RHA l-rhamnose	16 DUL dulcitol	17 INO inositol	18 MAN d-mannitol	19 SOR d-sorbitol
SGAir0178	-	+	-	-	+	+	-	-	-	-	+	+	+	+	-	-	-	+	+	-
SGAir0179	-	+	-	-	-	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0202	-	+	-	-	+	+	-	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0257	-	+	-	-	+	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0265	-	+	-	-	+	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0269	-	+	-	-	+	+	+	-	-	-	+	+	+	+	-	-	-	+	+	-
SGAir0414	-	+	-	-	+	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0425	-	+	-	-	+	+	-	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0427	-	+	-	-	+	+	-	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0563	-	+	-	-	+	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-
SGAir0424	-	+	-	-	+	+	+	-	-	-	+	+	+	+	-	-	-	+	+	-
SGAir0428	-	+	-	-	+	+	+	-	-	-	+	+	+	+	-	-	-	+	+	-
SGAir0080	-	+	-	-	+	+	+	-	-	-	+	+	+	+	-	-	-	+	+	-

20 MDM methyl-alphaD-mannopyranoside	21 MDG methyl-alphaD-glucopyranoside	22 NAG n-acetylglucosamine	23 AMY amygdaline	24 ARB arbutin	25 ESC esculin ferric citrate	26 SAL salicin	27 CEL d-cellobiose	28 MAL d-maltose	29 LAC d-lactose	30 MEL d-melibiose	31 SAC d-saccharose sucrose	32 TRE d-trehalose	33 INU inulin	34 MLZ d-melezitose	35 RAF d-raffinose	36 AMD amidon starch	37 GLYG glycogen	38 XLT xylitol	39 GEN gentiobiose	40 TUR d-turanose
-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+

Table A-2 continued from previous page.

41 LYX d-lyxose	42 TAG d-tagatose	43 DFUC d-fucose	44 LFUC l-fucose	45 DARL d-arabitol	46 LARL l-arabitol	47 GNT potassium gluconate	48 2KG potassium 2-ketogluconate	49 5KG potassium 5-ketogluconate	50 ONPG beta galactosidase	51 ADH arginine dihydrolase	52 LDC lysine decarboxylase	53 ODC ornithine decarboxylase	54 CIT citrate utilisation	55 H2S production	56 URE urease	57 TDA tryptophan deaminase	58 IND indole production	59 VP acetoin production	60 GEL gelatinase	61 NIT
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	+	-
-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	+	+	-
-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-

Table A-3: Housekeeping genes used in the multilocus sequence analysis of *P. aryabhattai* and *P. megaterium* (Figure 2-1a).

Function	Genes
Ribosome assembly	rbfA, rimM, rplA, rplB, rplD, rplM, rplO, rplP, rplT, rplV, rplX, rpsA, rpsB, rpsC, rpsD, rpsE, rpsG, rpsH, rpsI, rpsK, rsmA, rsmG, rsmH, typA, ybeY
Translation	frr, infC, lepA, prfA, prfB, rsfS, smpB
Transcription	nusB, nusG
tRNA synthesis	alaS, aspS, cysS, metG, pheS, serS, tilS, trmD, tsaD
DNA topology	gyrA, gyrB, recG
DNA repair	mfd, recA, recN, recR, ruvA, ruvB, uvrB, uvrC
DNA replication	dnaA, dnaN, dnaX
RNA polymerases	rpoB, rpoC
Protein transport	secA, secE, secG, secY
Other	atpD, clpX, coaD, dnaK, era, ffh, ftsY, guaB, pnp, purB, rseP

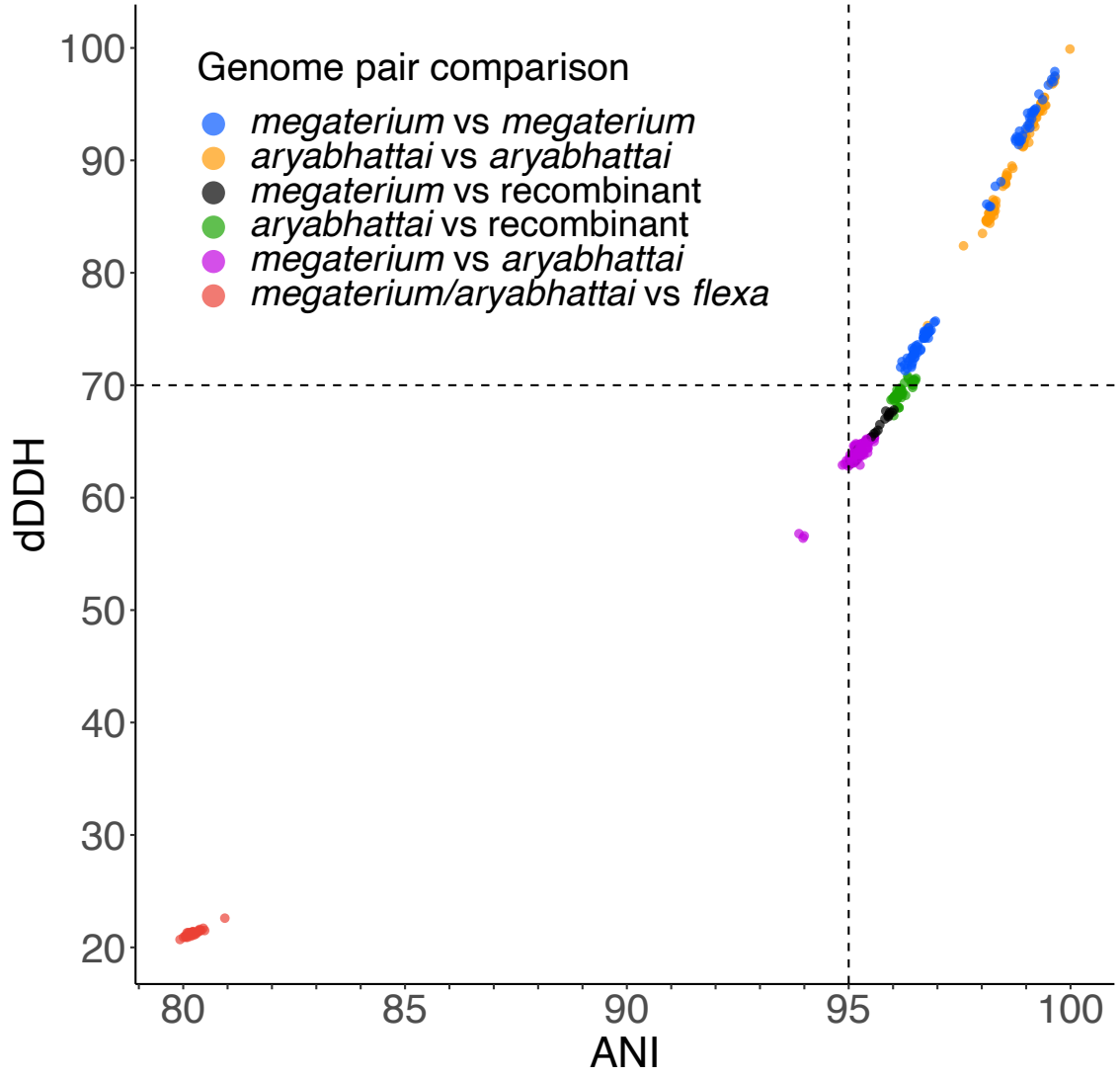


Figure A-1: Pairwise dDDH and ANI values between *Priestia* complete genomes, as in Figure 2-5, but also including two genomes from the nearest other species *Priestia flexa*. Dotted lines indicate the conventional same-species thresholds of 70% dDDH and 95% ANI. The discontinuity in genetic distance between species is clearly visible in comparisons between *P. flexa* and the *P. aryabhattai/megaterium* group (red). Comparisons between *P. megaterium* and *P. aryabhatta* are around the threshold of different species (purple), but could be considered as 'fuzzy species' due to the occurrence of recombinant genomes between them (green, black).

Table A-4: Details of the blocks of adjacent species-specific core genes for *P. megaterium* and *P. aryabhattai*. Gene functions were predicted using Prokka 1.14.6. Reference genomes (column 4) are *P. megaterium* 22-2 (genome accession GCA_009935415.1, genes are found on contig NZ_NKAQ01000003.1) and *P. aryabhattai* (genome GCA_002688605.1, contig NZ_CP024035.1).

Species	Block number and size, reference genome contig	Gene name	Location in reference genome contig
Priestia megaterium	1 10,756 bp NZ_NKAQ01000002.1	*speE* polyamine aminopropyltransferase	658387..659274
		Hypothetical protein	657709..658383
		Hypothetical protein	654842..657463
		murF UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	653436..654845
		prpB 2-methylisocitrate lyase	651905..653422
		Hypothetical protein	650782..651912
		phnW 2-aminoethylphosphonate—pyruvate transaminase	648948..650789
		Hypothetical protein	648518..648940
	2 6,252 bp NZ_NKAQ01000002.1	*degA* HTH-type transcriptional regulator	475064..476095
		aroK Shikimate kinase	474482..475039
		safD Sulfoacetaldehyde dehydrogenase	472504..473934
		ilvB Acetolactate synthase large subunit	470695..472452
		aroD 3-dehydroquinate dehydratase	470149..470628
	3 7,681 bp NZ_NKAQ01000002.1	*aroE* Shikimate dehydrogenase	452325..453221
		garP putative galactarate transporter	450378..451682
		smtB succinyl-CoA-L--malate CoA-transferase beta subunit	449058..450266

		yngG hydroxymethylglutaryl-CoA lyrase	448102..449058
		safD sulfoacetaldehyde dehydrogenase	446502..447950
		Hypothetical protein	445539..446414
	4 394 bp	Hypothetical protein	25035..25214
		Hypothetical protein	24820..25008
	5 2,495 bp NZ_NKAQ01000003.1	*yedA* putative inner membrane transporter	1681432..1682349
		lysN 2-aminoadipate transaminase	1679854..1681287
	6 1,478 bp	*sspH* Small, acid soluble spore protein H	1254644..1254823
		Hypothetical protein	1253857..1254117
		Hypothetical protein	1253345..1253500
	7 2,107 bp	*iolG* inositol 2-dehydrogenase / D-chiro-inositol 3-dehydrogenase	1160037..1161146
		yofA HTH-type transcriptional regulator	1159039..1159896
	8 778 bp	Hypothetical protein	1037478..1037831
		psiE Protein PsiE	1037053..1037475
Priestia aryabhattai	1 2,010 bp	*adhT* alcohol dehydrogenase	3441256..3442296
		sasP-B small acid soluble spore protein gamma type	3441037..3441204
		Hypothetical protein	3440286..3440777
	2 1,059 bp	*misCA* membrane protein insertase	3307377..3308141
		Hypothetical protein	3307082..3307321
	3 1,418 bp	*czcD* cadmium/cobalt/zinc antiporter	1729542..1730477
		Hypothetical protein	1729059..1729208

Table A-5: Top 100 orthologs by average dN between *P. aryabhattai* and *P. megaterium* sequences. Only genes which were present in 95% of genomes from both species are shown. Genes were included only if the average dN within each species was < 0.15 and the average dS within each species was < 0.175. Genes are ranked by the average dN of sequences from different species minus the average dN of sequences from the same species. Gene length is the length of the trimmed alignment of the individual sequences of the ortholog. Gene names are as annotated by Prokka; the best match for each ortholog in the UniProtKB database was found by BLASTP search using the ortholog's sequence from the genome assembly of the strain *P. megaterium* A.

dN score rank	Gene name (Prokka)	Average dN between species	Gene length (aa)	UniProt best match	UniProt best match sequence identity (%)	UniProt best match e–value
1	Transcriptional regulator ManR	0.495552675	667	Transcription antiterminator BglG	49.4	0
2	Ascorbate-specific PTS system EIIB component	0.403587048	89	PTS lactose transporter subunit IIB	75.6	2.1e-40
3	Ascorbate-specific PTS system EIIC component	0.282406341	420	PTS ascorbate transporter subunit IIC	74.9	0
4	Vitamin B12-dependent ribonucleoside-diphosphate reductase	0.319603351	439	Vitamin B12-dependent ribonucleotide reductase	86.9	0
5	Maltose O-acetyltransferase	0.195696175	185	Acetyltransferase	71.9	3.1e-96
6	Hypothetical protein 1370	0.106122214	119	Sporulation protein	81.5	7.3e-61
7	Phosphoribosyl-dephospho-CoA-transferase	0.081093714	202	Holo-ACP synthase, malonate decarboxylase-specific	84.6	6.2e-118
8	Shikimate kinase	0.068419279	183	Shikimate kinase aroK	90.7	8.8e-116
9	Hypothetical protein 2239	0.074971472	133	Flagellar assembly protein FliT	83.9	5.1e-61

10	Hypothetical protein 1788	0.054407855	64	Phage protein	87.5	2.9e-39
11	Hypothetical protein 869	0.07570106	241	Zinc transporter, ZIP family	83.4	7.3e-143
12	Hypothetical protein 1020	0.060292666	181	DUF4825 domain-containing protein	86.7	4.9e-109
13	Hypothetical protein 1704	0.069987894	209	CAAX amino terminal protease family protein	86.1	1.1e-128
14	Hypothetical protein 2238	0.066361603	134	Acetyltransferase, GNAT family	85.8	6.9e-82
15	Hypothetical protein 1857	0.103817638	224	DUF1573 domain-containing protein	79.5	1.2e-126
16	Hypothetical protein 1216	0.065213201	325	Multidrug resistance efflux transporter family protein	87.7	0
17	Adenosylcobinamide-GDP ribazoletransferase	0.049378508	261	Adenosylcobinamide-GDP ribazoletransferase, Cobalamin synthase, Cobalamin-5'-phospate synthase	90.7	3e-167
18	Hypothetical protein 1164	0.080779198	56	DUF4366 domain-containing protein	87.5	6.9e-31
19	Hypothetical protein 2222	0.054339369	204	Iron reductase	89.7	1.8e-130
20	Hypothetical protein 501	0.045981425	57	DUF3970 domain-containing protein	92.9	6.2e-37
21	Hypothetical protein 893	0.052612799	127	Putative membrane protein	89.8	9e-77
22	Hypothetical protein 1456	0.061066545	152	Acetyltransferase, GNAT family	88.1	3.3e-91
23	IS1595 family transposase ISBsu3	0.057199026	148	Acetyltransferase family protein	87.8	4e-90
24	Hypothetical protein 499	0.056383434	30	Lmo0937 family membrane protein	56	0.0012
25	Hypothetical protein 1077	0.086347355	209	Alpha/beta hydrolase	82.9	5.4e-118
26	Hypothetical protein 1145	0.047426278	191	Acetyltransferase, GNAT family	90	9.9e-128

27	Putative serine threonine-protein kinase YbdM	0.04810806	273	Serine/threonine protein kinase	89.7	6.1e-175
28	Hypothetical protein 1248	0.061239166	298	Dot/Icm system substrate protein LidA	86.4	0
29	Hypothetical protein 1708	0.061551338	76	DUF3953 domain-containing protein	85.1	1.8e-43
30	Hypothetical protein 1614	0.054760246	166	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	88.6	7.8e-108
31	Bifunctional adenosylcobalamin biosynthesis protein CobP, Bifunctional adenosylcobalamin biosynthesis protein CobU	0.048545594	186	Adenosylcobinamide kinase, Adenosylcobinamide-phosphate guanylyltransferase	90.2	2.4e-144
32	Hypothetical protein 1977	0.050300362	385	Transporter (Major facilitator Superfamily)	89.9	0
33	Spermidine/spermine N1-acetyltransferase	0.043105066	174	Protease synthase and sporulation negative regulatory protein PAI 1	91.1	2.9e-106
34	Hypothetical protein 318	0.051341078	445	HATPase_c_4 domain-containing protein	89.7	0
35	Hypothetical protein 2150	0.040725526	162	dUTPase	90.1	4.4e-103
36	Hypothetical protein 1166	0.056957546	54	Prophage protein	88.7	2.8e-37
37	Shikimate kinase	0.069869651	186	Shikimate kinase	88	5.1e-113
38	Hypothetical protein 2012	0.085342329	746	Flagellar hook-length control protein FliK	81	0
39	Hypothetical protein 1927	0.053694004	112	Flagellar protein FliT	88.4	3.3e-65
40	Beta-phosphoglucomutase	0.046089327	235	Beta-phosphoglucomutase	89.9	3e-148
41	Copper homeostasis protein CutC	0.043989708	209	PF03932 family protein CutC	92.3	3.8e-136
42	Hypothetical protein 1362	0.044857525	110	DUF4199 domain-containing protein	90	5e-66
43	Hypothetical protein 1030	0.043869853	119	KleE	89.9	3.1e-71

44	Foldase protein PrsA	0.038675577	300	peptidylprolyl isomerase	91.7	0
45	Hypothetical protein 1026	0.037391532	58	50S ribosomal protein L29	91.4	3.2e-35
46	Hypothetical protein 927	0.038338163	89	Helix-turn-helix domain of resolvase	91	4.1e-53
47	Hypothetical protein 1462	0.037307587	33	Stage V sporulation protein E	87.5	0.069
48	Hypothetical protein 1233	0.05252934	122	Oligoendopeptidase F	25.5	0.44
49	Hypothetical protein 1914	0.054582836	117	Cupin domain protein	88.9	4.8e-74
50	Molybdenum cofactor guanylyltransferase	0.034218831	189	Molybdenum cofactor guanylyltransferase MobA	92.6	1.2e-127
51	Hydroxyacylglutathione hydrolase, putative metallo-hydrolase YflN	0.060658698	219	Metal-dependent hydrolase	92.3	3.5e-165
52	Mycothiol acetyltransferase	0.040870211	276	Acetyltransferase, GNAT family	90.9	0
53	Hypothetical protein 288	0.047937173	396	DUF58 domain-containing protein	92.5	0
54	L-2,4-diaminobutyrate decarboxylase	0.041668081	478	L-2,4-diaminobutyrate decarboxylase	91.4	0
55	Histidine biosynthesis bifunctional protein HisB, D-glycero-beta-D-manno-heptose-1 7-bisphosphate 7-phosphatase	0.045650251	184	D,D-heptose 1,7-bisphosphate phosphatase	91.8	3.9e-119
56	RNA 2’,3’-cyclic phosphodiesterase	0.03801614	184	RNA 2',3'-cyclic phosphodiesterase, RNA 2',3'-CPDase	91.3	5.8e-121
57	Hypothetical protein 1324	0.033100002	75	DUF2536 family protein	93.2	1.7e-53
58	Hypothetical protein 1710	0.045326776	65	CYCLIN domain-containing protein	39.3	0.028
59	Hypothetical protein 1253	0.040925032	113	DUF3147 family protein	90.5	2.5e-59

60	Urease subunit beta	0.031264032	108	Urease subunit beta, Urea amidohydrolase subunit beta	93.5	1.8e-67
61	HTH-type transcriptional regulator SutR	0.0454336	185	Helix-turn-helix DNA-binding protein	90.8	3.6e-120
62	Molybdate-binding protein ModA	0.037474211	265	Molybdate ABC transporter, molybdate-binding protein ModA	92.5	6.5e-174
63	ATP-dependent dethiobiotin synthetase BioD	0.032227416	240	ATP-dependent dethiobiotin synthetase BioD, DTB synthetase, DTBS	93.8	1e-160
64	IS1595 family transposase ISSpgl1	0.046122732	179	Acetyltransferase, GNAT family	89.9	2e-114
65	Hypothetical protein 352	0.037979863	178	Movement protein BC1	91.6	2.1e-115
66	Hypothetical protein 352	0.035945639	149	Putative lipoprotein	90.6	3.3e-94
67	L-amino acid N-acetyltransferase AaaT	0.068164933	180	Acetyltransferase, GNAT family	88.3	4.9e-111
68	Hypothetical protein 1706	0.041453023	166	Acetyltransferase, GNAT family	92.2	3.3e-109
69	Hypothetical protein 772	0.061969776	215	Conserved membrane protein	86.4	2.2e-131
70	Intracellular serine protease	0.037851538	319	Intracellular serine protease	92.8	0
71	Hypothetical protein 727	0.044003045	306	Putative oxidoreductase, short chain dehydrogenase/reductase family protein	93.8	0
72	Leucine efflux protein, Homoserine/homoserine lactone efflux protein	0.036703128	209	Translocator protein, LysE family	93.3	9.4e-137
73	Hypothetical protein 1427	0.045667183	247	Xylose isomerase domain protein TIM barrel	90.3	3.5e-160

74	Peptidoglycan L-alanyl-D-glutamate endopeptidase CwlK	0.041352763	173	Cell wall carboxypeptidase	91.9	1.8e-114
75	Hypothetical protein 1188	0.036495105	61	DUF4083 domain-containing protein	93.4	9.4e-40
76	Hypothetical protein 359	0.033019362	57	Uncharacterized protein	94.5	1.3e-35
77	Hypothetical protein 1800	0.045984016	127	Lipoprotein	89	3.9e-61
78	4,4’-diaponeurosporenoate glycosyltransferase	0.033500905	368	Glycosyl transferase domain protein, group 2 family protein	93.2	0
79	Bacillibactin exporter	0.033718536	403	Transporter, major facilitator family	93.1	0
80	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	0.030346123	204	Alpha-ribazole phosphatase CobC	93.1	2.1e-142
81	Flagellar biosynthesis protein FlhF	0.034581789	365	Flagellar biosynthesis protein FlhF, Flagella-associated GTP-binding protein	92.6	0
82	Hypothetical protein 900	0.044388005	174	DinB 2 domain-containing protein	90.8	5.4e-112
83	Hypothetical protein 1232	0.044924217	472	Ethanolamine utilization protein EutA	92.8	0
84	Hypothetical protein 1646	0.032684504	56	KTSC domain-containing protein	91.1	1.2e-34
85	Fructokinase	0.043965153	316	Carbohydrate kinase, PfkB family protein	91.1	0
86	Hypothetical protein 2489	0.050304647	133	PA14 domain-containing protein	88.7	3.7e-81
87	Sensor histidine kinase gras	0.034479454	344	Histidine kinase	92.7	0
88	Hypothetical protein 928	0.032420056	214	Fibronectin-binding protein	92.1	3.e-140
89	Putative FMNH2-dependent monooxygenase SfnC	0.031186187	395	Acyl-CoA dehydrogenase	92.9	0

90	Hypothetical protein 2133	0.106309372	64	Two-component sensor histidine kinase	100	7.9e-36
91	Hypothetical protein 1897	0.030613383	152	Carboxymuconolactone decarboxylase family protein	91.4	5.3e-97
92	Hypothetical protein 1254	0.031191396	135	Transcriptional regulator, MarR family	93.3	5.7e-86
93	Hypothetical protein 678	0.03288705	265	Hydrolase	93.6	1.7e-176
94	Hypothetical protein 319	0.042377522	298	Amidase	89.9	0
95	Hypothetical protein 1717	0.03683765	266	DNA-binding protein	93.9	1.2e-42
96	Putative protein YpbG	0.031491121	255	Phosphoesterase	92.9	7.9e-170
97	Hypothetical protein 1519	0.079977442	50	Ovule protein	84	1.9e-27
98	Hypothetical protein 2145	0.030873432	108	ABC transporter permease	94.4	1.9e-6
99	Hypothetical protein 1571	0.083224991	174	Lipoprotein, putative	81	8.1e-98
100	Hypothetical protein 223	0.061772329	69	Endonuclease	93.8	5.6e-42

References

1. Ford Doolittle W, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009 May 1;19(5):744–56.
2. Brigandt I. Species Pluralism Does Not Imply Species Eliminativism. Philos Sci. 2003 Dec;70(5):1305–16.
3. Cohan FM. What are Bacterial Species? Annu Rev Microbiol. 2002 Nov 28;56:457–87.
4. Abe K, Nomura N, Suzuki S. Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. FEMS Microbiol Ecol. 2020 May 1;96(5):31.
5. Abby SS, Tannier E, Gouy M, Daubin V. Lateral gene transfer as a support for the tree of life. Proc Natl Acad Sci U S A. 2012 Mar 27;109(13):4962–7.
6. Knöppel A, Lind PA, Lustig U, Näsvall J, Andersson DI. Minor Fitness Costs in an Experimental Model of Horizontal Gene Transfer in Bacteria. Mol Biol Evol. 2014 May 1;31(5):1220–7.
7. Raz Y, Tannenbaum E. The Influence of Horizontal Gene Transfer on the Mean Fitness of Unicellular Populations in Static Environments. Genetics. 2010 May 1;185(1):327–37.
8. Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends Microbiol. 2015 Oct 1;23(10):598–605.
9. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007 Jan 26;315(5811):476–80.
10. Baltrus DA. Exploring the costs of horizontal gene transfer. Trends Ecol Evol. 2013

Aug 1;28(8):489–95.

11. Majewski J. Sexual isolation in bacteria. FEMS Microbiol Lett. 2001 May 1;199(2):161–9.

12. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012 Feb 1;14(2):347–55.

13. Melendrez MC, Becraft ED, Wood JM, Olsen MT, Bryant DA, Heidelberg JF, et al. Recombination does not hinder formation or detection of ecological species of *Synechococcus* inhabiting a hot spring cyanobacterial mat. Front Microbiol. 2016 Jan 14;6(JAN):166934.

14. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018 Dec 30;9(1):5114.

15. Deloger M, El Karoui M, Petit MA. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol. 2009 Jan;91(1):91–9.

16. Rodriguez-R LM, Conrad RE, Viver T, Feistel DJ, Lindner BG, Venter F, et al. An ANI gap within bacterial species that advances the definitions of intra-species units. bioRxiv [Preprint]. 2023 Jul 5;2022.06.27.497766.

17. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015 Aug 18;43(14):6761–71.

18. de Albuquerque NRM, Haag KL. Using average nucleotide identity (ANI) to evaluate microsporidia species boundaries based on their genetic relatedness. J Eukaryot Microbiol. 2023 Mar 1;70(2):e12944.

19. Lachance MA, Lee DK, Hsiang T. Delineating yeast species with genome average

nucleotide identity: a calibration of ANI with haplontic, heterothallic *Metschnikowia* species. Antonie Van Leeuwenhoek. 2020 Dec 1;113(12):2097–106.

20. Accetto T, Janež N. The lytic *Myoviridae* of Enterobacteriaceae form tight recombining assemblages separated by discontinuities in genome average nucleotide identity and lateral gene flow. Microb genomics. 2018 Mar 1;4(3):e000169.
21. Bobay L-M. The Prokaryotic Species Concept and Challenges. In: Tettelin H, Medini D, editors. The Pangenome. Springer; 2020. p. 21–49.
22. Bobay L-M, Ochman H. Biological Species Are Universal across Life's Domains. Genome Biol Evol. 2017 Mar 1;9(3):491–501.
23. Dykhuizen DE, Green L. Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol. 1991;173(22):7257–68.
24. Rossellö -Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev. 2001 Jan 1;25(1):39–67.
25. De Queiroz K. Different species problems and their resolution. BioEssays. 2005 Dec 1;27(12):1263–9.
26. De Queiroz K. Species Concepts and Species Delimitation. Syst Biol. 2007 Dec 1;56(6):879–86.
27. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 2008 May 7;6(6):431–40.
28. Moldovan MA, Gelfand MS. Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. Front Microbiol. 2018 Mar 12;9(MAR):339098.
29. Ereshefsky M. Darwin's solution to the species problem. Synthese. 2010 Apr

11;175(3):405–25.

30. Lawrence JG, Retchless AC. The myth of bacterial species and speciation. Biol Philos. 2010 May 5;25(4):569–88.

31. Ereshefsky M. Microbiology and the species problem. Biol Philos. 2010 May 4;25(4):553–68.

32. Lan R, Reeves PR. When does a clone deserve a name? A perspective on bacterial species based on population genetics. Trends Microbiol. 2001 Sep 1;9(9):419–24.

33. Bapteste E, Boucher Y. Epistemological impacts of horizontal gene transfer on classification in microbiology. In: Gogarten MB, Gogarten JP, Olendzenski LC, editors. Horizontal Gene Transfer. Humana Press; 2009. p. 55–72.

34. Ereshefsky M. Eliminative Pluralism. Philos Sci. 1992 Dec;59(4):671–90.

35. Cohan FM. Bacterial Species and Speciation. Syst Biol. 2001 Aug 1;50(4):513–24.

36. Shapiro BJ, Polz MF. Microbial Speciation. Cold Spring Harb Perspect Biol. 2015 Oct 1;7(10):a018143.

37. Doolittle WF. Speciation without Species: A Final Word. Philos Theory, Pract Biol. 2019 Jul 10;11(20220112).

38. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, et al. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Bacteriol. 1987;37(4):463–4.

39. Cohan FM. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos Trans R Soc B Biol Sci. 2006 Nov 29;361(1475):1985–96.

40. Mallet J. Speciation in the 21st century. Heredity (Edinb). 2005 Jun 27;95(1):105–9.

41. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria

and archaea. ISME J. 2009 Oct 2;3(2):199–208.

42. Straub TJ, Zhaxybayeva O. A null model for microbial diversification. Proc Natl Acad Sci U S A. 2017 Jul 3;114(27):E5414–23.

43. Hanage WP, Fraser C, Spratt BG. Fuzzy species among recombinogenic bacteria. BMC Biol. 2005 Mar 7;3(1):1–7.

44. Hanage WP. Fuzzy species revisited. BMC Biol. 2013 Apr 15;11(1):1–3.

45. Hanage WP, Fraser C, Spratt BG. Sequences, sequence clusters and bacterial species. Philos Trans R Soc B Biol Sci. 2006 Nov 29;361(1475):1917–27.

46. Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A, Syrjänen R, Spratt BG. Using multilocus sequence data to define the pneumococcus. J Bacteriol. 2005 Sep;187(17):6223–30.

47. Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. Convergence of Campylobacter Species: Implications for Bacterial Evolution. Science. 2008;320(5873):237–9.

48. Rosselló-Móra R, Amann R. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol. 2015 Jun 1;38(4):209–16.

49. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, et al. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol. 2008 Sep 1;31(4):241–50.

50. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998 Mar 17;95(6):3140–5.

51. Ibal JC, Pham HQ, Park CE, Shin JH. Information about variations in multiple copies of bacterial 16S rRNA genes may aid in species identification. PLoS One. 2019 Feb

1;14(2):e0212090.

52. Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, et al. Use of 16S rRNA Gene for Identification of a Broad Range of Clinically Relevant Bacterial Pathogens. PLoS One. 2015 Feb 6;10(2):e0117617.

53. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov 6;10(1):1–11.

54. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol. 2005 Mar;71(3):1501–6.

55. Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol. 2006 Mar 1;6(2):97–112.

56. Pérez-Losada M, Porter ML, Viscidi RP, Crandall KA. Multilocus Sequence Typing of Pathogens. Genet Evol Infect Dis. 2011 Jan 1;503–21.

57. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect Genet Evol. 2013 Jun 1;16:38–53.

58. Ahmad Y, Gertz RE, Li Z, Sakota V, Broyles LN, Van Beneden C, et al. Genetic relationships deduced from emm and multilocus sequence typing of invasive *Streptococcus dysgalactiae* subsp. *equisimilis* and *S. canis* recovered from isolates collected in the United States. J Clin Microbiol. 2009 Jul;47(7):2046–54.

59. Passerini D, Beltramo C, Coddeville M, Quentin Y, Ritzenthaler P, Daveran-Mingot ML, et al. Genes but Not Genomes Reveal Bacterial Domestication of *Lactococcus lactis*. PLoS One. 2010;5(12):e15306.

60. Vanlaere E, Baldwin A, Gevers D, Henry D, De Brandt E, LiPuma JJ, et al. Taxon K, a complex within the *Burkholderia cepacia* complex, comprises at least two novel species, *Burkholderia contaminans* sp. nov. and *Burkholderia lata* sp. nov. Int J Syst Evol Microbiol. 2009 Jan 1;59(1):102–11.

61. Vanlaera E, LiPuma JJ, Baldwin A, Henry D, Brandt E De, Mahenthiralingam E, et al. *Burkholderia latens* sp. nov., *Burkholderia diffusa* sp. nov., *Burkholderia arboris* sp. nov., *Burkholderia seminalis* sp. nov., and *Burkholderia metallica* sp. nov., novel species within the *Burkholderia cepacia* complex. Int J Syst Evol Microbiol. 2008 Jul 1;58(7):1580–90.

62. Guinebretière MH, Auger S, Galleron N, Contzen M, de Sarrau B, de Buyser ML, et al. *Bacillus cytotoxicus* sp. nov. is a novel thermotolerant species of the *Bacillus cereus* group occasionally associated with food poisoning. Int J Syst Evol Microbiol. 2013 Jan 1;63(1):31–40.

63. Chaloner GL, Ventosilla P, Birtles RJ. Multi-Locus Sequence Analysis Reveals Profound Genetic Diversity among Isolates of the Human Pathogen *Bartonella bacilliformis*. PLoS Negl Trop Dis. 2011 Jul;5(7):e1248.

64. Tanigawa K, Watanabe K. Multilocus sequence typing reveals a novel subspeciation of *Lactobacillus delbrueckii*. Microbiology. 2011 Mar 1;157(3):727–38.

65. Do T, Jolley KA, Maiden CJ, Gilbert SC, Clark D, Wade WG, et al. Population structure of *Streptococcus oralis*. Microbiology. 2009 Aug 1;155(8):2593–602.

66. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, et al. Re-evaluating prokaryotic species. Nat Rev Microbiol. 2005 Aug 10;3(9):733–9.

67. Meier-Kolthoff JP, Klenk HP, Göker M. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. Int J Syst Evol Microbiol. 2014 Feb

1;64(PART 2):352–6.

68. Klappenbach JA, Goris J, Vandamme P, Coenye T, Konstantinidis KT, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007 Jan 1;57(1):81–91.

69. Brenner DJ. Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. Int J Syst Bacteriol. 1973 Oct 1;23(4):298–307.

70. Janda JM, Abbott SL. Bacterial Identification for Publication: When Is Enough Enough? J Clin Microbiol. 2002;40(6):1887–91.

71. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005 Feb 15;102(7):2567–72.

72. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009 Nov 10;106(45):19126–31.

73. Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci. 2010 Feb 28;2(1):117–34.

74. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013 Feb 21;14(1):1–14.

75. Sentausa E, Fournier PE. Advantages and limitations of genomics in prokaryotic taxonomy. Clin Microbiol Infect. 2013 Sep 1;19(9):790–5.

76. Shivaji S, Chaturvedi P, Begum Z, Pindi PK, Manorama R, Padmanaban DA, et al. *Janibacter hoylei* sp. nov., *Bacillus isronensis* sp. nov. and *Bacillus aryabhattai* sp. nov., isolated from cryotubes used for collecting air from the upper atmosphere. Int J Syst Evol Microbiol. 2009 Dec 1;59(12):2977–86.

77. de Bary A. Vergleichende Morphologie und Biologie der Pilze, Mycetozoen und Bacterien. Leipzig: Wilhelm Engelmann; 1884.

78. Boone DR, Castenholz RW, Brenner DJ, Krieg NR, Staley JT, De Vos P, et al. Bergey's Manual of Systematic Bacteriology. 2nd ed. Springer; 2012.

79. Vary PS. Prime time for *Bacillus megaterium*. Microbiology. 1994;140:1001–13.

80. Davidson D, Beheshti B, Mittelman MW. Effects of *Arthrobacter* sp., *Acidovorax delafieldii*, and *Bacillus megaterium* colonisation on copper solvency in a laboratory reactor. http://dx.doi.org/101080/08927019609378310. 2009;9(4):279–92.

81. Huang FL, Zhang Y, Zhang LP, Wang S, Feng Y, Rong NH. Complete genome sequence of *Bacillus megaterium* JX285 isolated from *Camellia oleifera* rhizosphere. Comput Biol Chem. 2019 Apr 1;79:1–5.

82. Wahhab BHA, Samsulrizal NH, Edbeib MF, Wahab RA, Al-Nimer MSM, Hamid AAA, et al. Genomic analysis of a functional haloacid-degrading gene of *Bacillus megaterium* strain BHS1 isolated from Blue Lake (Mavi Gölü, Turkey). Ann Microbiol. 2021 Dec 1;71(1):1–11.

83. Kalsi N, Uchida A, Purbojati RW, Houghton JNI, Chénard C, Wong A, et al. Whole-Genome Sequence of *Bacillus megaterium* strain SGAir0080, Isolated from an Indoor Air Sample. Microbiol Resour Announc. 2019 Dec 12;8(50).

84. Polter SJ, Caraballo AA, Lee YP, Wilhelm WW, Gan HM, Wheatley MS, et al. Isolation, Identification, Whole-Genome Sequencing, and Annotation of Four *Bacillus* Species, *B. anthracis* RIT375, *B. circulans* RIT379, *B. altitudinis* RIT380, and *B. megaterium* RIT381, from Internal Stem Tissue of the Insulin Plant *Costus igneus*. Genome Announc. 2015;3(4).

85. Shwed PS, Crosthwait J, Weedmark K, Hoover E, Dussault F. Complete Genome

Sequences of *Priestia megaterium* Type and Clinical Strains Feature Complex Plasmid Arrays. Microbiol Resour Announc. 2021 Jul 8;10(27).

86. Biedendieck R, Knuuti T, Moore SJ, Jahn D. The “beauty in the beast”—the multiple uses of *Priestia megaterium* in biotechnology. Appl Microbiol Biotechnol. 2021 Jul 15;105(14):5719–37.

87. Balabanova L, Averianova L, Marchenok M, Son O, Tekutyeva L. Microbial and Genetic Resources for Cobalamin (Vitamin B12) Biosynthesis: From Ecosystems to Industrial Biotechnology. Int J Mol Sci. 2021 Apr 26;22(9):4522.

88. Moore SJ, Lawrence AD, Biedendieck R, Deery E, Frank S, Howard MJ, et al. Elucidation of the anaerobic pathway for the corrin component of cobalamin (vitamin B12). Proc Natl Acad Sci U S A. 2013 Sep 10;110(37):14906–11.

89. Collins HF, Biedendieck R, Leech HK, Gray M, Escalante-Semerena JC, McLean KJ, et al. *Bacillus megaterium* Has Both a Functional BluB Protein Required for DMB Synthesis and a Related Flavoprotein That Forms a Stable Radical Species. PLoS One. 2013 Feb 14;8(2):e55708.

90. Chakraborty U, Chakraborty B, Basnet M. Plant growth promotion and induction of resistance in *Camellia sinensis* by *Bacillus megaterium*. J Basic Microbiol. 2006 Jun 1;46(3):186–95.

91. Bhatt K, Maheshwari DK. Zinc solubilizing bacteria (*Bacillus megaterium*) with multifarious plant growth promoting activities alleviates growth in *Capsicum annuum* L. 3 Biotech. 2020 Jan 7;10(2):1–10.

92. Dahmani MA, Desrut A, Moumen B, Verdon J, Mermouri L, Kacem M, et al. Unearthing the Plant Growth-Promoting Traits of *Bacillus megaterium* RmBm31, an Endophytic Bacterium Isolated From Root Nodules of *Retama monosperma*. Front Plant Sci. 2020 Feb 27;11:124.

93. Kumari WMNH, Thiruchittampalam S, Weerasinghe MSS, Chandrasekharan NV, Wijayarathna CD. Characterization of a *Bacillus megaterium* strain with metal bioremediation potential and in silico discovery of novel cadmium binding motifs in the regulator, CadC. Appl Microbiol Biotechnol. 2021 Mar 2;105(6):2573–86.

94. Park YG, Mun BG, Kang SM, Hussain A, Shahzad R, Seo CW, et al. *Bacillus aryabhattai* SRB02 tolerates oxidative and nitrosative stress and promotes the growth of soybean by modulating the production of phytohormones. PLoS One. 2017;12(3):e0173203.

95. Ghosh PK, Maiti TK, Pramanik K, Ghosh SK, Mitra S, De TK. The role of arsenic resistant *Bacillus aryabhattai* MCC3374 in promotion of rice seedlings growth and alleviation of arsenic phytotoxicity. Chemosphere. 2018 Nov 1;211:407–19.

96. Balakrishna Pillai A, Jaya Kumar A, Kumarapillai H. Biosynthesis of poly(3-hydroxybutyrate-co-3-hydroxyvalerate) (PHBV) in *Bacillus aryabhattai* and cytotoxicity evaluation of PHBV/poly(ethylene glycol) blends. 3 Biotech. 2020 Jan 7;10(2):1–10.

97. Ash C, Farrow JAE, Wallbanks S, Collins MD. Phylogenetic heterogeneity of the genus *Bacillus* revealed by comparative analysis of small-subunit-ribosomal RNA sequences. Lett Appl Microbiol. 1991 Oct 1;13(4):202–6.

98. Logan NA, De Vos P. Genus *Bacillus* Cohn 1872. In: De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, et al., editors. Bergey's Manual of Systematic Bacteriology. New York: Springer; 2009. p. 21–128.

99. Gupta RS, Patel S, Saini N, Chen S. Robust demarcation of 17 distinct *Bacillus* species clades, proposed as novel Bacillaceae genera, by phylogenomics and comparative genomic analyses: Description of *Robertmurraya kyonggiensis* sp. nov. and proposal for an emended genus *Bacillus* limiting it only to the members

of the *subtilis* and *cereus* clades of species. Int J Syst Evol Microbiol. 2020 Oct 28;70(11):5753–98.

100. Narsing Rao MP, Dong ZY, Liu GH, Li L, Xiao M, Li WJ. Reclassification of *Bacillus aryabhattai* Shivaji et al. 2009 as a later heterotypic synonym of *Bacillus megaterium* de Bary 1884 (Approved Lists 1980). FEMS Microbiol Lett. 2019 Nov 1;366(22).

101. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 May 5;10(6):563–9.

102. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10–2.

103. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One. 2014 Nov 19;9(11):e112963.

104. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.

105. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.

106. Hunt M, Silva N De, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: Automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015 Dec 29;16(1):1–10.

107. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017 Nov 11;2(11):1533–42.

108. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 8;32(5):1792–7.

109. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.

110. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.

111. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–290.

112. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. McInerny G, editor. Methods Ecol Evol. 2017 Jan 1;8(1):28–36.

113. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Vol. 23, Molecular Biology and Evolution. Oxford Academic; 2006. p. 254–67.

114. Auch AF, Klenk HP, Göker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci. 2010 Feb 28;2(1):142–8.

115. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. Stajich JE, editor. PLoS One. 2010 Jun 25;5(6):e11147.

116. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Comput Biol. 2015;11(2):e1004041.

117. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul 1;25(7):1043–55.

118. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021 Sep 27;38(10):4647–54.

119. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007 Jan 1;73(1):278–88.

120. Gaballa A, Cheng RA, Trmcic A, Kovac J, Kent DJ, Martin NH, et al. Development of a database and standardized approach for rpoB sequence-based subtyping and identification of aerobic spore-forming Bacillales. J Microbiol Methods. 2021 Dec 1;191:106350.

121. Bennett JS, Watkins ER, Jolley KA, Harrison OB, Maiden MCJ. Identifying *Neisseria* species by use of the 50S Ribosomal protein L6 (*rplF*) gene. J Clin Microbiol. 2014;52(5):1375–81.

122. Liang KYH, Orata FD, Boucher YF, Case RJ. Roseobacters in a Sea of Poly- and Paraphyly: Whole Genome-Based Taxonomy of the Family Rhodobacteraceae and the Proposal for the Split of the “Roseobacter Clade” Into a Novel Family, Roseobacteraceae fam. nov. Front Microbiol. 2021 Jun 25;12:1635.

123. Felsenstein J. Phylogenies and the Comparative Method. Am Nat. 1985 Jan 15;125(1):1–15.

124. Guzmán-Moreno J, García-Ortega LF, Torres-Saucedo L, Rivas-Noriega P, Ramírez-Santoyo RM, Sánchez-Calderón L, et al. *Bacillus megaterium* HgT21: a Promising Metal Multiresistant Plant Growth-Promoting Bacteria for Soil Biorestoration. Microbiol Spectr. 2022 Oct 26;10(5).

125. Bhattacharyya C, Bakshi U, Mallick I, Mukherji S, Bera B, Ghosh A. Genome-guided

insights into the plant growth promotion capabilities of the physiologically versatile *Bacillus aryabhattai* strain AB211. Front Microbiol. 2017 Mar 21;8(MAR):411.

126. Verma P, Yadav AN, Khannam KS, Kumar S, Saxena AK, Suman A. Molecular diversity and multifarious plant growth promoting attributes of Bacilli associated with wheat (*Triticum aestivum* L.) rhizosphere from six diverse agro-ecological zones of India. J Basic Microbiol. 2016 Jan 1;56(1):44–58.

127. Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol. 2014 Feb 1;64(PART 2):384–91.

128. Gillis M, Vandamme P, Vos P De, Swings J, Kersters K. Polyphasic Taxonomy. Bergey's Man Syst Archaea Bact. 2015 Sep 14;1–10.

129. Hassler HB, Probert B, Moore C, Lawson E, Jackson RW, Russell BT, et al. Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. Microbiome. 2022 Jul 8;10(1):1–18.

130. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

131. Vandamme P, Peeters C. Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek. 2014 Mar 15;106(1):57–65.

132. Blackwood KS, Turenne CY, Harmsen D, Kabani AM. Reassessment of Sequence-Based Targets for Identification of *Bacillus* Species. J Clin Microbiol. 2004 Apr;42(4):1626–30.

133. Ki JS, Zhang W, Qian PY. Discovery of marine *Bacillus* species by 16S rRNA and *rpoB* comparisons and their usefulness for species identification. J Microbiol Methods. 2009 Apr 1;77(1):48–57.

134. Shapiro BJ, Leducq JB, Mallet J. What Is Speciation? PLOS Genet. 2016 Mar 1;12(3):e1005860.

135. Ehling-Schulz M, Lereclus D, Koehler TM. The *Bacillus cereus* Group: *Bacillus* Species with Pathogenic Potential. Microbiol Spectr. 2019 May 31;7(3).

136. Callister SJ, McCue LA, Turse JE, Monroe ME, Auberry KJ, Smith RD, et al. Comparative Bacterial Proteomics: Analysis of the Core Genome Concept. PLoS One. 2008 Feb 6;3(2):e1542.

137. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. Genome Res. 2020 Nov 1;30(11):1667–79.

138. Segerman B. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Front Cell Infect Microbiol. 2012 Sep 6;2:116.

139. Chávez-Luzanía RA, Montoya-Martínez AC, Parra-Cota FI, de los Santos-Villalobos S. Pangenomes-identified singletons for designing specific primers to identify bacterial strains in a plant growth-promoting consortium. Mol Biol Rep. 2022 Sep 20;1:1–10.

140. Carlos Guimaraes L, Benevides de Jesus L, Vinicius Canario Viana M, Silva A, Thiago Juca Ramos R, de Castro Soares S, et al. Inside the Pan-genome - Methods and Software Overview. Curr Genomics. 2015 May 11;16(4):245–52.

141. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015 Feb 1;23:148–54.

142. Jensen RA. Orthologs and paralogs - we need to get it right. Genome Biol 2001 28. 2001 Aug 3;2(8):1–3.

143. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar

in Function than Paralogs. PLOS Comput Biol. 2012 May;8(5):e1002514.

144. NCBI. Help for Assembly [Internet]. [cited 2022 Dec 22]. Available from: https://www.ncbi.nlm.nih.gov/assembly/help/

145. Hu X, Friedberg I. SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. Gigascience. 2019 Oct 1;8(10):1–12.

146. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422–3.

147. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. PLoS One. 2010 Mar 10;5(3):e9490.

148. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. PLOS Genet. 2008 Dec;4(12):e1000304.

149. Stecher G, Tamura K, Kumar S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Mol Biol Evol. 2020 Apr 1;37(4):1237–9.

150. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D261–9.

151. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D457–62.

152. Guy L, Kultima JR, Andersson SGE, Quackenbush J. genoPlotR: comparative gene and genome visualization in R. Bioinformatics. 2010 Sep 15;26(18):2334–5.

153. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

154. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al.

Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012 Aug 1;40(15):e115–e115.

155. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

156. Gusareva ES, Acerbi E, Lau KJX, Luhung I, Premkrishnan BN V., Kolundžija S, et al. Microbial communities in the tropical air ecosystem follow a precise diel cycle. Proc Natl Acad Sci. 2019 Oct 28;116(46):23299–308.

157. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008 Oct 1;11(5):472–7.

158. Panmanee W, Charoenlap N, Atichartpongkul S, Mahavihakanont A, Whiteside MD, Winsor G, et al. The OxyR-regulated *phnW* gene encoding 2-aminoethylphosphonate:pyruvate aminotransferase helps protect *Pseudomonas aeruginosa* from tert-butyl hydroperoxide. PLoS One. 2017 Dec 1;12(12):e0189066.

159. Krejčík Z, Denger K, Weinitschke S, Hollemeyer K, Pačes V, Cook AM, et al. Sulfoacetate released during the assimilation of taurine-nitrogen by *Neptuniibacter caesariensis*: Purification of sulfoacetaldehyde dehydrogenase. Arch Microbiol. 2008 Aug 28;190(2):159–68.

160. Moeller R, Raguse M, Reitz G, Okayasu R, Li Z, Klein S, et al. Resistance of *Bacillus subtilis* spore DNA to lethal ionizing radiation damage relies primarily on spore core components and DNA repair, with minor effects of oxygen radical detoxification. Appl Environ Microbiol. 2014 Jan;80(1):104–9.

161. Byrd B, Camilleri E, Korza G, Craft DL, Green J, Rocha Granados M, et al. Levels and Characteristics of mRNAs in Spores of Firmicute Species. J Bacteriol. 2021 Jun 22;203(14).

162. Vyas J, Cox J, Setlow B, Coleman WH, Setlow P. Extremely variable conservation of γ-type small, acid-soluble proteins from spores of some species in the bacterial order Bacillales. J Bacteriol. 2011 Apr;193(8):1884–92.

163. Anton A, Große C, Reißmann J, Pribyl T, Nies DH. CzcD is a heavy metal ion transporter involved in regulation of heavy metal resistance in *Ralstonia* sp. strain CH34. J Bacteriol. 1999 Nov 15;181(22):6876–81.

164. Xu J, Cotruvo JA. The *czcD* (NiCo) Riboswitch Responds to Iron(II). Biochemistry. 2020 Apr 21;59(15):1508–16.

165. Olanrewaju OS, Ayilara MS, Ayangbenro AS, Babalola OO. Genome Mining of Three Plant Growth-Promoting *Bacillus* Species from Maize Rhizosphere. Appl Biochem Biotechnol. 2021 Dec 1;193(12):3949–69.

166. Abdullahi S, Haris H, Zarkasi KZ, Amir HG. Complete genome sequence of plant growth-promoting and heavy metal-tolerant *Enterobacter tabaci* 4M9 (CCB-MBL 5004). J Basic Microbiol. 2021 Apr 1;61(4):293–304.

167. Gaete A, Andreani-Gerard C, Maldonado JE, Muñoz-Torres PA, Sepúlveda-Chavera GF, González M. Bioprospecting of Plant Growth-Promoting Traits of *Pseudomonas* sp. Strain C3 Isolated from the Atacama Desert: Molecular and Culture-Based Analysis. Diversity. 2022 May 1;14(5):388.

168. Igiehon NO, Babalola OO, Aremu BR. Genomic insights into plant growth promoting rhizobia capable of enhancing soybean germination under drought stress. BMC Microbiol. 2019 Jul 11;19(1):1–22.

169. Nascimento FX, Hernández AG, Glick BR, Rossi MJ. Plant growth-promoting activities and genomic analysis of the stress-resistant *Bacillus megaterium* STB1, a bacterium of agricultural and biotechnological interest. Biotechnol Reports. 2020 Mar 1;25:e00406.

170. Maymon M, Martínez-Hidalgo P, Tran SS, Ice T, Craemer K, Anbarchian T, et al. Mining the phytomicrobiome to understand how bacterial coinoculations enhance plant growth. Front Plant Sci. 2015 Sep 24;6(September):784.

171. Ludueña LM, Anzuay MS, Angelini JG, McIntosh M, Becker A, Rupp O, et al. Genome sequence of the endophytic strain *Enterobacter* sp. J49, a potential biofertilizer for peanut and maize. Genomics. 2019 Jul 1;111(4):913–20.

172. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938–50.

173. Van Leijenhorst DC, Van Der Weide TP. A formal derivation of Heaps' Law. Inf Sci (Ny). 2005 Feb 25;170(2–4):263–72.

174. Wright ES, Baum DA. Exclusivity offers a sound yet practical species criterion for bacteria despite abundant gene flow. BMC Genomics. 2018 Oct 3;19(1):1–12.

175. Didelot X, Lawson D, Darling A, Falush D. Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences. Genetics. 2010 Dec 1;186(4):1435–49.

176. Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. Virus Evol. 2021 Jan 20;7(1):87.

177. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. Mol Biol Evol. 2017 May 1;34(5):1167–82.

178. Mohammed Y, Lee B, Kang Z, Du G. Development of a two-step cultivation strategy for the production of vitamin B12 by *Bacillus megaterium*. Microb Cell Fact. 2014 Jul 15;13(1):1–10.

179. Biedendieck R, Malten M, Barg H, Bunk B, Martens JH, Deery E, et al. Metabolic engineering of cobalamin (vitamin B12) production in *Bacillus megaterium*. Microb Biotechnol. 2010 Jan 1;3(1):24–37.

180. Martens JH, Barg H, Warren M, Jahn D. Microbial production of vitamin B12. 2002;58(3):275–85.

181. Deobald D, Hanna R, Shahryari S, Layer G, Adrian L. Identification and characterization of a bacterial core methionine synthase. Sci Reports 2020 101. 2020 Feb 7;10(1):1–13.

182. Cervantes M, Murillo FJ. Role for vitamin B12 in light induction of gene expression in the bacterium *Myxococcus xanthus*. J Bacteriol. 2002;184(8):2215–24.

183. Chiang YR, Wei STS, Wang PH, Wu PH, Yu CP. Microbial degradation of steroid sex hormones: implications for environmental and ecological studies. Microb Biotechnol. 2020 Jul 1;13(4):926–49.

184. Shelton AN, Seth EC, Mok KC, Han AW, Jackson SN, Haft DR, et al. Uneven distribution of cobamide biosynthesis and dependence in bacteria predicted by comparative genomics. ISME J. 2018 Nov 14;13(3):789–804.

185. Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature. 2005 Nov 3;438(7064):90–3.

186. Froese DS, Fowler B, Baumgartner MR. Vitamin B12, folate, and the methionine remethylation cycle—biochemistry, pathways, and regulation. J Inherit Metab Dis. 2019 Jul 1;42(4):673–85.

187. Larsson KM, Logan DT, Nordlund P. Structural basis for adenosylcobalamin activation in adocbl-dependent ribonucleotide reductases. ACS Chem Biol. 2010 Oct 15;5(10):933–42.

188. Torrents E, Poplawski A, Sjöberg B-M. Two Proteins Mediate Class II Ribonucleotide Reductase Activity in *Pseudomonas aeruginosa*. J Biol Chem. 2005 Apr 29;280(17):16571–8.

189. Lee KM, Go J, Yoon MY, Park Y, Kim SC, Yong DE, et al. Vitamin B12-Mediated restoration of defective anaerobic growth leads to reduced biofilm formation in *Pseudomonas aeruginosa*. Infect Immun. 2012 May;80(5):1639–49.

190. Crespo A, Pedraz L, Astola J, Torrents E. *Pseudomonas aeruginos*a exhibits deficient biofilm formation in the absence of class II and III ribonucleotide reductases due to hindered anaerobic growth. Front Microbiol. 2016 May 9;7(MAY):688.

191. Jeckelmann JM, Erni B. Transporters of glucose and other carbohydrates in bacteria. Pflügers Arch - Eur J Physiol. 2020 May 6;472(9):1129–53.

192. McCoy JG, Levin EJ, Zhou M. Structural insight into the PTS sugar transporter EIIC. Biochim Biophys Acta - Gen Subj. 2015 Mar 1;1850(3):577–85.

193. Joyet P, Bouraoui H, Aké FMD, Derkaoui M, Zébré AC, Cao TN, et al. Transcription regulators controlled by interaction with enzyme IIB components of the phosphoenolpyruvate:sugar phosphotransferase system. Biochim Biophys Acta - Proteins Proteomics. 2013 Jul 1;1834(7):1415–24.

194. Wenzel M, Altenbuchner J. The *Bacillus subtilis* mannose regulator, ManR, a DNA-binding protein regulated by HPr and its cognate PTS transporter ManP. Mol Microbiol. 2013 May 1;88(3):562–76.

195. Gulati A, Mahadevan S. The *Escherichia coli* antiterminator protein BglG stabilizes the 5’ region of the bgl mRNA. J Biosci. 2001;26(2):193–203.

196. Dean DA, Reizer J, Nikaido H, Saier MH. Regulation of the maltose transport system of *Escherichia coli* by the glucose-specific enzyme III of the

phosphoenolpyruvate-sugar phosphotransferase system. Characterization of inducer exclusion-resistant mutants and reconstitution of inducer exclusion in proteoliposomes. J Biol Chem. 1990 Dec 5;265(34):21005–10.

197. Sondej M, Weinglass AB, Peterkofsky A, Kaback HR. Binding of Enzyme IIA^{Glc}, a Component of the Phosphoenolpyruvate:Sugar Phosphotransferase System, to the *Escherichia coli* Lactose Permease. Biochemistry. 2002 Apr 30;41(17):5556–65.

198. Somavanshi R, Ghosh B, Sourjik V. Sugar Influx Sensing by the Phosphotransferase System of *Escherichia coli*. PLOS Biol. 2016 Aug 24;14(8):e2000074.

199. Bieler S, Silva F, Soto C, Belin D. Bactericidal activity of both secreted and nonsecreted microcin E492 requires the mannose permease. J Bacteriol. 2006 Oct;188(20):7049–61.

200. Ragunathan PT, Vanderpool CK. Cryptic-prophage-encoded small protein DicB protects *Escherichia coli* from phage infection by inhibiting inner membrane receptor proteins. J Bacteriol. 2019 Dec 1;201(23).

201. Chopra I. Molecular mechanisms involved in the transport of antibiotics into bacteria. Parasitology. 1988;96(S1):S25–44.

202. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat Rev Microbiol. 2015 Sep 16;13(10):641–50.

203. Zeng Z, Liu X, Yao J, Guo Y, Li B, Li Y, et al. Cold adaptation regulated by cryptic prophage excision in *Shewanella oneidensis*. ISME J. 2016 Aug 2;10(12):2787–800.

204. Gödeke J, Paul K, Lassak J, Thormann KM. Phage-induced lysis enhances biofilm formation in *Shewanella oneidensis* MR-1. ISME J. 2010 Oct 21;5(4):613–26.

205. Theeragool G, Miyao A, Yamada K, Sato T, Kobayashi Y. In vivo expression of the *Bacillus subtilis* spoVE gene. J Bacteriol. 1993;175(13):4071–80.

206. Driks A. *Bacillus subtilis* Spore Coat. Microbiol Mol Biol Rev. 1999 Mar;63(1):1–20.

207. Forouhar F, Lee IS, Vujcic J, Vujcic S, Shen J, Vorobiev SM, et al. Structural and functional evidence for *Bacillus subtilis* PaiA as a novel N1-spermidine/spermine acetyltransferase. J Biol Chem. 2005 Dec 2;280(48):40328–36.

208. Honjo M, Nakayama A, Fukazawa K, Kawamura K, Ando K, Hori M, et al. A novel *Bacillus subtilis* gene involved in negative control of sporulation and degradative-enzyme production. J Bacteriol. 1990;172(4):1783–90.

209. Seneviratne CJ, Suriyanarayanan T, Swarup S, Chia KHB, Nagarajan N, Zhang C. Transcriptomics Analysis Reveals Putative Genes Involved in Biofilm Formation and Biofilm-associated Drug Resistance of *Enterococcus faecalis*. J Endod. 2017 Jun 1;43(6):949–55.

210. Aizawa SI. Mystery of FliK in length control of the Flagellar hook. J Bacteriol. 2012 Sep;194(18):4798–800.

211. Patterson-Delafield J, Martinez RJ, Stocker BAD, Yamaguchi S. A new *fla* gene in *Salmonella typhimurium-flaR*-and its mutant phenotype-superhooks. Arch Mikrobiol. 1973 Jun;90(2):107–20.

212. Kinoshita M, Hara N, Imada K, Namba K, Minamino T. Interactions of bacterial flagellar chaperone-substrate complexes with FlhA contribute to co-ordinating assembly of the flagellar filament. Mol Microbiol. 2013 Dec 1;90(6):1249–61.

213. Hirano T, Mizuno S, Aizawa SI, Hughes KT. Mutations in Flk, FlgG, FlhA, and FlhE that affect the flagellar type III secretion specificity switch in *Salmonella enterica*. J Bacteriol. 2009 Jun;191(12):3938–47.

214. Bange G, Kümmerer N, Engel C, Bozkurt G, Wild K, Sinning I. FlhA provides the adaptor for coordinated delivery of late flagella building blocks to the type III secretion system. Proc Natl Acad Sci U S A. 2010 Jun 22;107(25):11295–300.

215. Bennett JCQ, Thomas J, Fraser GM, Hughes C. Substrate complexes and domain organization of the *Salmonella* flagellar export chaperones FlgN and FliT. Mol Microbiol. 2001 Feb 1;39(3):781–91.

216. Eppinger M, Bunk B, Johns MA, Edirisinghe JN, Kutumbaka KK, Koenig SSK, et al. Genome sequences of the biotechnologically important *Bacillus megaterium* strains QM B1551 and DSM319. J Bacteriol. 2011 Aug;193(16):4199–213.

217. Fujii M, Shibata S, Aizawa SI. Polar, Peritrichous, and Lateral Flagella Belong to Three Distinguishable Flagellar Families. J Mol Biol. 2008 May 29;379(2):273–83.

218. Guttenplan SB, Shaw S, Kearns DB. The cell biology of peritrichous flagella in *Bacillus subtilis*. Mol Microbiol. 2013 Jan 1;87(1):211–29.

219. Li X, Ren F, Cai G, Huang P, Chai Q, Gundogdu O, et al. Investigating the Role of FlhF Identifies Novel Interactions With Genes Involved in Flagellar Synthesis in *Campylobacter jejuni*. Front Microbiol. 2020 Mar 24;11:460.

220. Arroyo-Pérez EE, Ringgaard S. Interdependent Polar Localization of FlhF and FlhG and Their Importance for Flagellum Formation of *Vibrio parahaemolyticus*. Front Microbiol. 2021 Mar 17;12:557.

221. Zhang K, He J, Cantalano C, Guo Y, Liu J, Li C. FlhF regulates the number and configuration of periplasmic flagella in *Borrelia burgdorferi*. Mol Microbiol. 2020 Jun 1;113(6):1122–39.

222. Okinaka RT, Keim P. The Phylogeny of *Bacillus cereus* sensu lato. Microbiol Spectr. 2016;4(1).

223. Mazzantini D, Fonnesu R, Celandroni F, Calvigioni M, Vecchione A, Mrusek D, et al. GTP-Dependent FlhF Homodimer Supports Secretion of a Hemolysin in *Bacillus cereus*. Front Microbiol. 2020 May 6;11:879.

224. Attieh Z, Mouawad C, Rejasse A, Jehanno I, Perchat S, Hegna IK, et al. The *fliK*

Gene Is Required for the Resistance of *Bacillus thuringiensis* to Antimicrobial Peptides and Virulence in *Drosophila melanogaster*. Front Microbiol. 2020 Dec 18;11:3160.

225. Idelevich EA, Pogoda CA, Ballhausen B, Wüllenweber J, Eckardt L, Baumgartner H, et al. Pacemaker lead infection and related bacteraemia caused by normal and small colony variant phenotypes of *Bacillus licheniformi*s. J Med Microbiol. 2013 Jun 1;62(PART6):940–4.

226. La Jeon Y, Yang JJ, Kim MJ, Lim G, Cho SY, Park TS, et al. Combined *Bacillus licheniformis* and *Bacillus subtilis* infection in a patient with oesophageal perforation. J Med Microbiol. 2012 Dec 1;61(PART12):1766–9.

227. Celandroni F, Salvetti S, Gueye SA, Mazzantini D, Lupetti A, Senesi S, et al. Identification and Pathogenic Potential of Clinical *Bacillus* and *Paenibacillus* Isolates. PLoS One. 2016 Mar 1;11(3):e0152831.

228. Sahu C, Kumar K, Sinha MK, Venkata A, Majji AB, Jalali S. Review of endogenous endophthalmitis during pregnancy including case series. Int Ophthalmol. 2013 Oct 24;33(5):611–8.

229. Crisafulli E, Aredano I, Valzano I, Burgazzi B, Andrani F, Chetta A. Pleuritis with pleural effusion due to a *Bacillus megaterium* infection. Respirol Case Reports. 2019 Jan 1;7(1):e00381.

230. Duncan KO, Smith TL. Primary cutaneous infection with *Bacillus megaterium* mimicking cutaneous anthrax. J Am Acad Dermatol. 2011 Aug 1;65(2):e60–1.

231. Ramos-Esteban JC, Servat JJ, Tauber S, Bia F. *Bacillus megaterium* delayed onset lamellar keratitis after LASIK. J Refract Surg. 2006;22(3):309–12.

232. Guo FP, Fan HW, Liu ZY, Yang QW, Li YJ, Li TS. Brain abscess caused by *Bacillus megaterium* in an adult patient. Chin Med J (Engl). 2015 Jan 6;128(11):1552–4.

233. Bocchi MB, Perna A, Cianni L, Vitiello R, Greco T, Maccauro G, et al. A rare case of *Bacillus megaterium* soft tissues infection. Acta Biomed Atenei Parm. 2020 Dec 30;91(14-S):e2020013–e2020013.

234. Morikawa M. Beneficial biofilm formation by industrial bacteria *Bacillus subtilis* and related species. J Biosci Bioeng. 2006 Jan 1;101(1):1–8.

235. Alvarez-Ordóñez A, Coughlan LM, Briandet R, Cotter PD. Biofilms in Food Processing Environments: Challenges and Opportunities. https://doi.org/101146/annurev-food-032818-121805. 2019 Mar 25;10:173–95.

236. Arnaouteli S, Bamford NC, Stanley-Wall NR, Kovács ÁT. *Bacillus subtilis* biofilm formation and social interactions. Nat Rev Microbiol. 2021 Apr 6;19(9):600–14.

237. Majed R, Faille C, Kallassy M, Gohar M. *Bacillus cereus* Biofilms-same, only different. Front Microbiol. 2016 Jul 7;7(JUL):203805.

238. Lin Y, Briandet R, Kovács ÁT. *Bacillus cereus* sensu lato biofilm formation and its ecological importance. Biofilm. 2022 Dec 1;4:100070.

239. Candela T, Fagerlund A, Buisson C, Gilois N, Kolstø AB, Økstad OA, et al. CalY is a major virulence factor and a biofilm matrix protein. Mol Microbiol. 2019 Jun 1;111(6):1416–29.

240. Kable ME, Srisengfa Y, Xue Z, Coates LC, Marco ML. Viable and total bacterial populations undergo equipment- and timedependent shifts during milk processing. Appl Environ Microbiol. 2019;85(13).

241. Kim YJ, Kim HS, Kim KY, Chon JW, Kim DH, Seo KH. High Occurrence Rate and Contamination Level of *Bacillus cereus* in Organic Vegetables on Sale in Retail Markets. https://home.liebertpub.com/fpd. 2016 Dec 1;13(12):656–60.

242. Kuroki R, Kawakami K, Qin L, Kaji C, Watanabe K, Kimura Y, et al. Nosocomial Bacteremia Caused by Biofilm-Forming *Bacillus cereus* and *Bacillus thuringiensis*.

Intern Med. 2009;48(10):791–6.

243. Li Y, Chen N, Wu Q, Liang X, Yuan X, Zhu Z, et al. A Flagella Hook Coding Gene *flgE* Positively Affects Biofilm Formation and Cereulide Production in Emetic *Bacillus cereus*. Front Microbiol. 2022 Jun 10;13:897836.

244. Liaqat I, Mirza SA, Iqbal R, Ali NM, Saleem G, Majid S, et al. Flagellar motility plays important role in Biofilm formation of *Bacillus cereus* and *Yersinia enterocolitica*. Pak J Pharm Sci. 2018 Sep 1;31(5(Supplementary)):2047–52.

245. Houry A, Briandet R, Aymerlch S, Gohar M. Involvement of motility and flagella in *Bacillus cereus* biofilm formation. Microbiology. 2010 Apr 1;156(4):1009–18.

246. Belas R. Biofilms, flagella, and mechanosensing of surfaces by bacteria. Trends Microbiol. 2014 Sep 1;22(9):517–27.

247. Hölscher T, Bartels B, Lin YC, Gallegos-Monterrosa R, Price-Whelan A, Kolter R, et al. Motility, Chemotaxis and Aerotaxis Contribute to Competitiveness during Bacterial Pellicle Biofilm Development. J Mol Biol. 2015 Nov 20;427(23):3695–708.

248. Carabetta VJ, Tanner AW, Greco TM, Defrancesco M, Cristea IM, Dubnau D. A complex of YlbF, YmcA and YaaT regulates sporulation, competence and biofilm formation by accelerating the phosphorylation of Spo0A. Mol Microbiol. 2013 Apr 1;88(2):283–300.

249. Huang Q, Zhang Z, Liu Q, Liu F, Liu Y, Zhang J, et al. SpoVG is an important regulator of sporulation and affects biofilm formation by regulating Spo0A transcription in *Bacillus cereus* 0–9. BMC Microbiol. 2021 Dec 1;21(1):1–17.

250. Špacapan M, Danevčič T, Štefanic P, Porter M, Stanley-Wall NR, Mandic-Mulec I. The ComX Quorum Sensing Peptide of *Bacillus subtilis* Affects Biofilm Formation Negatively and Sporulation Positively. Microorganisms. 2020 Jul 27;8(8):1131.

251. Wijman JGE, De Leeuw PPLA, Moezelaar R, Zwietering MH, Abee T. Air-liquid

interface biofilms of *Bacillus cereus*: Formation, sporulation, and dispersion. Appl Environ Microbiol. 2007 Mar;73(5):1481–8.

252. Fernández L, Rodríguez A, García P. Phage or foe: an insight into the impact of viral predation on microbial communities. ISME J 2018 125. 2018 Jan 25;12(5):1171–9.

253. Cumby N, Davidson AR, Maxwell KL. The moron comes of age. http://dx.doi.org/104161/bact23146. 2012 Oct;2(4):e23146.

254. Abe K, Yoshinari A, Aoyagi T, Hirota Y, Iwamoto K, Sato T. Regulated DNA rearrangement during sporulation in *Bacillus weihenstephanensis* KBAB4. Mol Microbiol. 2013 Oct 1;90(2):415–27.

255. Ventroux M, Noirot-Gros MF. Prophage-encoded small protein YqaH counteracts the activities of the replication initiator DnaA in *Bacillus subtilis*. Microbiol (United Kingdom). 2022 Nov 29;168(11):001268.

256. Abe K, Kawano Y, Iwamoto K, Arai K, Maruyama Y, Eichenberger P, et al. Developmentally-Regulated Excision of the SPβ Prophage Reconstitutes a Gene Required for Spore Envelope Maturation in *Bacillus subtilis*. PLOS Genet. 2014 Oct 1;10(10):e1004636.

257. Sanchez-Vizuete P, Le Coq D, Bridier A, Herry JM, Aymerich S, Briandet R. Identification of *ypqP* as a new *Bacillus subtilis* biofilm determinant that mediates the protection of *Staphylococcus aureus* against antimicrobial agents in mixed-species communities. Appl Environ Microbiol. 2015 Oct 17;81(1):109–18.

258. Schuch R, Fischetti VA. The Secret Life of the Anthrax Agent *Bacillus anthracis*: Bacteriophage-Mediated Ecological Adaptations. PLoS One. 2009 Aug 12;4(8):e6532.

259. Gillis A, Mahillon J. Influence of lysogeny of tectiviruses GIL01 and GIL16 on

Bacillus thuringiensis growth, biofilm formation, and swarming Motility. Appl Environ Microbiol. 2014 Dec 15;80(24):7620–30.

260. Pal Roy M, Datta S, Ghosh S. A novel extracellular low-temperature active phytase from *Bacillus aryabhattai* RS1 with potential application in plant growth. Biotechnol Prog. 2017 May 1;33(3):633–41.

261. Lee S, Ka JOO, Song HGG. Growth promotion of *Xanthium italicum* by application of rhizobacterial isolates of *Bacillus aryabhattai* in microcosm soil. J Microbiol. 2012 Feb 27;50(1):45–9.

262. Yoo SJ, Weon HY, Song J, Sang MK. Induced Tolerance to Salinity Stress by Halotolerant Bacteria *Bacillus aryabhattai* H19-1 and B. mesonae H20-5 in Tomato Plants. J Microbiol Biotechnol. 2019 Jul 28;29(7):1124–36.

263. Mehmood S, Khan AA, Shi F, Tahir M, Sultan T, Munis MFH, et al. Alleviation of Salt Stress in Wheat Seedlings via Multifunctional *Bacillus aryabhattai* PM34: An In-Vitro Study. Sustainability. 2021 Jul 19;13(14):8030.

264. Xu H, Gao J, Portieles R, Du L, Gao X, Borras-Hidalgo O. Endophytic bacterium *Bacillus aryabhattai* induces novel transcriptomic changes to stimulate plant growth. PLoS One. 2022 Aug 1;17(8):e0272500.

265. Bishara A, Moss EL, Kolmogorov M, Parada A, Weng Z, Sidow A, et al. Culture-free generation of microbial genomes from human and marine microbiomes. bioRxiv. 2018 Feb 11;263939.

266. Ramesh A, Sharma SK, Sharma MP, Yadav N, Joshi OP. Inoculation of zinc solubilizing *Bacillus aryabhattai* strains for improved growth, mobilization and biofortification of zinc in soybean and wheat cultivated in Vertisols of central India. Appl Soil Ecol. 2014 Jan 1;73:87–96.

267. Zeng Q, Xie J, Li Y, Gao T, Xu C, Wang Q. Comparative genomic and functional

analyses of four sequenced *Bacillus cereus* genomes reveal conservation of genes relevant to plant-growth-promoting traits. Sci Rep. 2018 Nov 19;8(1):1–10.

268. Katsenios N, Andreou V, Sparangis P, Djordjevic N, Giannoglou M, Chanioti S, et al. Assessment of plant growth promoting bacteria strains on growth, yield and quality of sweet corn. Sci Rep. 2022 Jul 8;12(1):1–13.

269. Efthimiadou A, Katsenios N, Chanioti S, Giannoglou M, Djordjevic N, Katsaros G. Effect of foliar and soil application of plant growth promoting bacteria on growth, physiology, yield and seed quality of maize under Mediterranean conditions. Sci Rep. 2020 Dec 3;10(1):1–11.

270. Lipková N, Cinkocki R, Maková J, Medo J, Javoreková S. Characterization of endophytic bacteria of the genus *Bacillus* and their influence on the growth of maize (Zea mays) in vivo.

271. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol. 2020 Jun 4;18(9):491–506.

272. Gusareva ES, Gaultier NE, Uchida A, Premkrishnan BN V, Heinle CE, Phung WJ, et al. Short-range contributions of local sources to ambient air. PNAS Nexus. 2022 Jun 27;1(2):1–10.

273. Drautz-Moses DI, Luhung I, Gusareva ES, Kee C, Gaultier NE, Premkrishnan BNV, et al. Vertical stratification of the air microbiome in the lower troposphere. Proc Natl Acad Sci U S A. 2022 Feb 15;119(7):e2117293119.

274. Gusareva ES, Gaultier NPE, Premkrishnan BNV, Kee C, Lim SBY, Heinle CE, et al. Taxonomic composition and seasonal dynamics of the air microbiome in West Siberia. Sci Rep. 2020 Dec 9;10(1):1–7.

275. Cao C, Jiang W, Wang B, Fang J, Lang J, Tian G, et al. Inhalable microorganisms in Beijing's PM2.5 and PM10 pollutants during a severe smog event. Environ Sci

Technol. 2014;48(3):1499–507.

276. Ji L, Zhang Q, Fu X, Zheng L, Dong J, Wang J, et al. Feedback of airborne bacterial consortia to haze pollution with different PM2.5 levels in typical mountainous terrain of Jinan, China. Sci Total Environ. 2019 Dec 10;695:133912.

277. Qin N, Liang P, Wu C, Wang G, Xu Q, Xiong X, et al. Longitudinal survey of microbiome associated with particulate matter in a megacity. Genome Biol. 2020 Mar 3;21(1):1–11.

278. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016 Apr 13;7(1):1–9.

279. Udvardy MDF. A classification of the biogeographical provinces of the world. Morges; 1975.

280. United States Environmental Protection Agency. Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI) [Internet]. Environmental Protection. 2018. p. 22. Available from: https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf

281. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16;

282. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Jan 29;10(2):1–4.

283. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. J Stat Softw. 2008 Mar 18;25(1):1–18.

284. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2(3):18–22.

285. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Friedrich L. e1071: Misc

Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-11. [Internet]. 2022 [cited 2022 Nov 7]. Available from: https://cran.r-project.org/web/packages/e1071/index.html

286. Pfeifer MT, Koepke P, Reuder J. Effects of altitude and aerosol on UV radiation. J Geophys Res Atmos. 2006 Jan 16;111(D1):1203.

287. Bapteste E, Boucher Y. Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol. 2008 May 1;16(5):200–7.

288. Huson DH, Scornavacca C. A Survey of Combinatorial Methods for Phylogenetic Networks. Genome Biol Evol. 2011 Jan 1;3(1):23–35.

289. Wertz JE, Goldstone C, Gordon DM, Riley MA. A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. J Evol Biol. 2003 Nov 1;16(6):1236–48.

290. Doolittle WF, Bapteste E. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci. 2007 Feb 13;104(7):2043–9.

291. Shafikhani SH, Leighton T. AbrB and Spo0E Control the Proper Timing of Sporulation in *Bacillus subtilis*. Curr Microbiol. 2004 Apr;48(4):262–9.

292. Ireton K, Grossman AD. Interactions among mutations that cause altered timing of gene expression during sporulation in *Bacillus subtilis*. J Bacteriol. 1992;174(10):3185–95.